

工學碩士 學位論文

**Detection of Boundaries between Unvoiced Consonants and Noise
using Histogram**

指導教授 辛 沃 根

2001年 2月

韓國海洋大學校 大學院

工 學 科

朴 正 任

本 論 文 朴 正 任 工 學 碩 士 學 位 論 文 認 准

委 員 長 工 學 博 士 劉 永 昊 印

委 員 工 學 博 士 朴 侗 讚 印

委 員 工 學 博 士 辛 沃 根 印

2000年 12月

韓 國 海 洋 大 學 校 大 學 院

工 學 科 朴 正 任

Abstract

1	1
2	4
2.1	(phonemes)	4
2.1.1	(vowels)	4
2.1.2	(consonants)	5
2.1.3	(semivowel)	6
2.2	(voiced and unvoiced sounds)	7
2.2.1	7
2.2.2	10
3	(unvoiced consonant) (noise)	12
3.1	12
3.1.1	12
3.1.2	13
3.2	18
3.2.1	가	18
3.2.2	19
3.3	20
4	26
4.1	28
4.1.1	28
4.1.2	29
4.1.3	(Power Spectrum)	29
4.1.4	30

4.2	30
4.2.1	30
4.2.2	32
4.3	34
4.3.1	(Smoothing)	35
4.3.2	35
4.4	38
4.4.1	39
4.4.2	42
5	가	46
5.1	47
5.2	47
5.2.1	가	47
5.2.2	49
5.2.3	49
6	56
	58

Detection of Boundaries between Unvoiced Consonants and Noise using Histogram

Jeong Im Park

Department of Computer Engineering, Korea Maritime University, Pusan, Korea

Abstract

Voice activity detection(VAD), which separates the voice region from silence or noise region of input speech signal, is one of the indispensable pre-processing steps in continuous speech recognition, speech coding and noise estimation/reduction etc. While many successful researches were conducted continuous speech in noiseless environment or for isolated words in noisy environment, there are few method of VAD for continuous speech in heavy noise environment. Since unvoiced consonant signals have very similar characteristics to those of noise signals, it may result in serious distortion of unvoiced consonants to estimate and remove the noise components if voice activity detection and thereafter noise estimation/removal is carried out without paying special attention on unvoiced consonants.

In this dissertation, assuming that the voiced sound regions are removed by a method developed in our lab, we propose a method to explicitly extract

the boundaries between unvoiced consonant region and noise region so that more exact VAD could be performed. The proposed method is based on histogram in frequency domain which was successfully used by Hirsch for noise estimation, and also on similarity measure of frequency components between adjacent frames. To evaluate the performance of the proposed method, experiments on unvoiced consonant boundary detection was carried out on noisy speech signals of 10dB and 15dB SNR. For all seven kinds of noised, the overall rate of correct extraction resulted in approximately 90%.

The proposed algorithm could be used for VAD for speech recognition and speech coding as well as for noise estimation and reduction in heavy noise environments.

1

가

가

가

가

(preprocessing)

가

[1].

가

[19, 20].

[1, 9]

Junqua[7]가

IORB (island of reliability boundary)

IORB

[2, 21]

Hirsch[9]가

(frame)

가

(SNR : signal to noise ratio)가 10dB 15dB

6

7가

가

가

90%

SNR,

, 가

2

, 3

, 4

,

.

가 5

6

.

2

2.1 (phonemes)

(phoneme)

가 [3].
 (vowel), (consonant), (semi-vowel)

2.1.1 (vowels)

(monophthong) (diphthong)

[4].

2.1

2.1

Table 2.1 Classification of Korean vowels

	ㅣ [i]	(ㅟ [y])			ㅡ [ɯ]	ㅜ [u]
	ㅑ [e]	(ㅓ [ø])	(ㅗ : [œ])			ㅛ [o]
	ㅓ [ɛ]					
			ㅏ [a]		ㅓ [ʌ]	

2.1.2 (consonants)

(fricative), (affricate), (nasal), (liquid) (plosive),

가

/ㅂ, ㅃ, ㅄ, ㅅ, ㅆ, ㅈ, ㅊ, ㅋ, ㆁ/

, /ㅅ, ㅆ, ㅈ, ㅊ, ㅋ, ㆁ/

/ㅂ, ㅃ, ㅄ, / /ㅅ, ㅆ, ㅈ, ㅊ, ㅋ, ㆁ/

/ㅅ, ㅆ, ㅈ, ㅊ, ㅋ, ㆁ/

/ㅂ, ㅃ, ㅄ, /ㅅ, ㅆ, ㅈ, ㅊ, ㅋ, ㆁ/

가

/ㅂ, ㅃ, ㅄ, /ㅅ, ㅆ, ㅈ, ㅊ, ㅋ, ㆁ/

/ㅂ, ㅃ, ㅄ, /ㅅ, ㅆ, ㅈ, ㅊ, ㅋ, ㆁ/

가

/ㅂ, ㅃ, ㅄ, /ㅅ, ㅆ, ㅈ, ㅊ, ㅋ, ㆁ/

가

1)

1) 가 가 (place of articulation), (aspiration), (manner of articulation), (tenseness)

2.2

Table 2.2 Classification of Korean consonants

		ㅂ [b] ㅃ [p ^h] ㅍ [p [̃]]	ㄷ [d] ㅌ [t ^h] ㅍ [t [̃]]		ㄱ [g] ㅋ [k ^h] ㆁ [k [̃]]	
				ㅈ [dʒ] ㅊ [tʃ ^h] ㅉ [tʃ [̃]]		
			ㅅ [s ^h] ㅆ [s [̃]]			ㅎ [h]
		ㅁ [m]	ㄴ [n]		ㅇ [ŋ]	
			ㄹ [l]			

2.1.3 (semivowel)

/j, w, ɰ/가 .

/j/,

/w/,

/ɰ/

.

/ɰ/ /j/ /w/

.

[j, ɰ, w, ɰ]

[i, y, u, ɰ]

가 가

.

2.3

2.3

Table 2.3 Classification of Korean semivowels

	j	ɰ	ɰ	w
	i	y	ɰ	u

2.2 (voiced and unvoiced sounds)

가

(voiced sound)

, , , .

가

(unvoiced sound)

, , .

2.2.1

2.1²⁾

2.2

2.1 2.2

가

(sec),

가

(sec),

(Hz),

가

2.1(a)

가,

2.1(b)

가

/ ㅏ /

()

()

(a) (b)

가

가

2.2 (a)

가,

2.2 (b)

가

/ /

/ ㅑ /

()

()

2.1

,

가

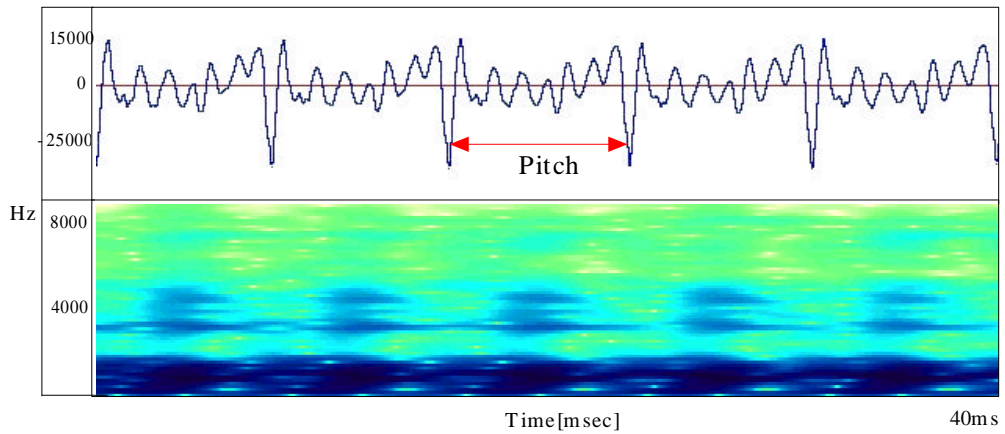
,

2)

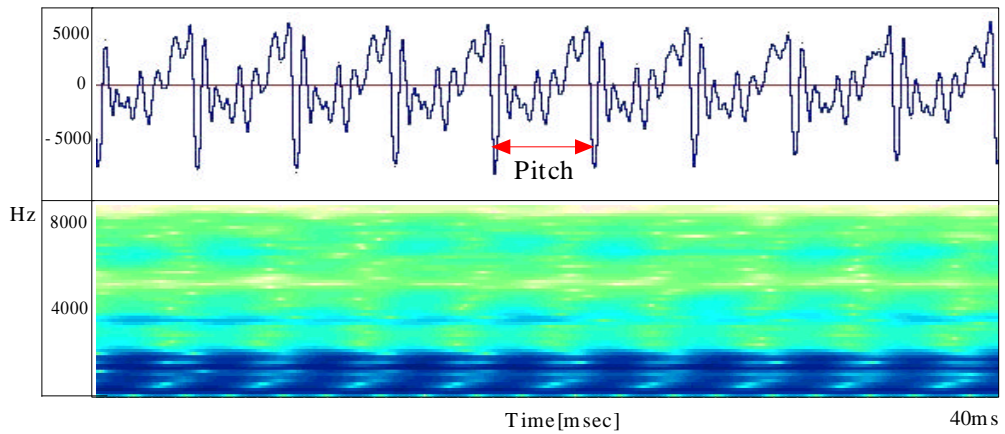
(International phonetic Association) alphabet ' ' ' ' .

, 가

(IPA; international phonetic



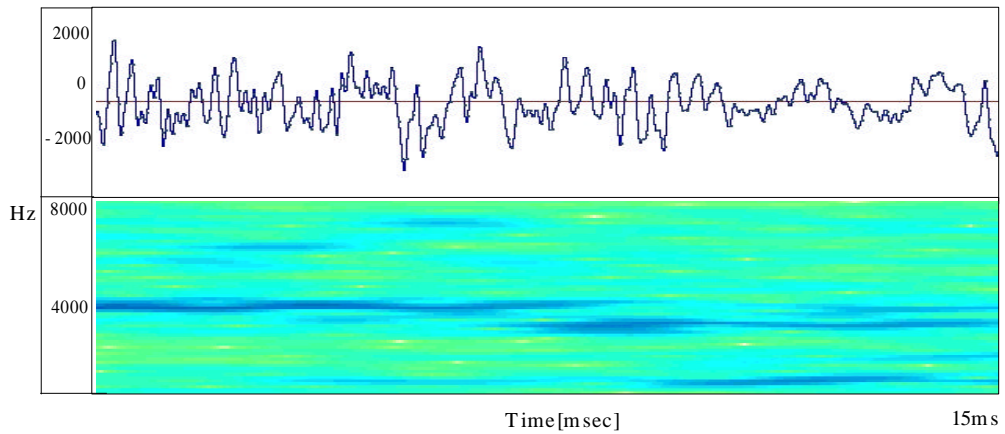
(a)



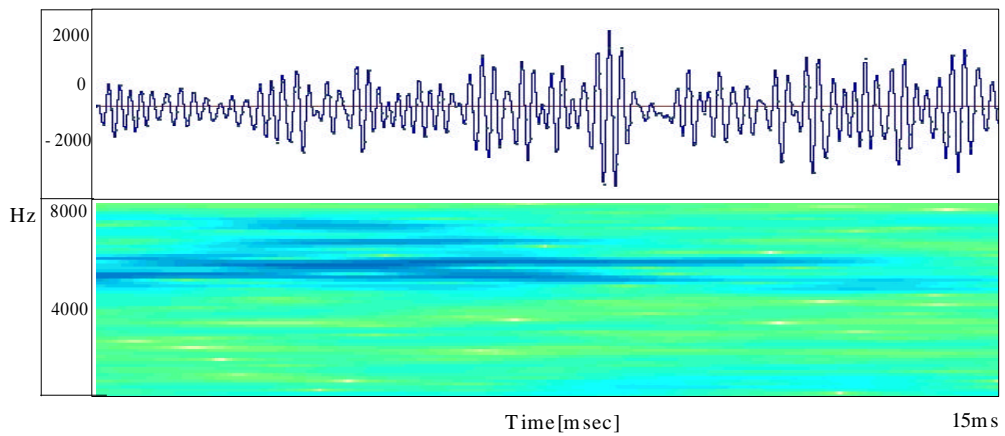
(b)

2.1 (a) 가 / ㅏ / ()
 () (b) 가 / ㅏ / () ()

Figure 2.1 (a) Speech wave (up) and spectrogram (down) of the vowel /a/ of a male speaker (b) Speech wave (up) and spectrogram (down) of the vowel /a/ of a female speaker



(a)



(b)

2.2 (a) 가 / / / ㅏ / ()
 () (b) 가 / / / ㅏ / ()
 () ()

Figure 2.2 (a) Speech wave (up) and spectrogram (down) of the unvoiced consonant /tʰ/ in a male speaker's utterance /tʰəŋ/ (b) Speech wave (up) and spectrogram (down) of the unvoiced consonant /tʰ/ in a female speaker's utterance /tʰəŋ/

가 , 가 , 가

2.2.2

2.4

2.4

Table 2.4 Composition of phonemes and noise/silence in a non-voiced sound region in Korean

N	/ (all noise/silence)
UNU	+ / + (unvoiced+noise/silence+unvoiced)
UN	+ (unvoiced+noise/silence)
NU	+ (noise/silence+unvoiced)
U	(all unvoiced)
UU	+ = (unvoiced+unvoiced = all unvoiced)

가 (silence) 가
, N(noise) , U(unvoiced)
가 . 가
UU
U .

3 (unvoiced consonant) (noise)

3

가

3.1

, 3.2

3.1 (unvoiced consonant)

3.1.1

ETRI POW (phonetically optimized word) (corpus)

POW 7khz

16khz , 16bit 40

3,848 POW

1

8 16 가 470

가가

2 3 200 stop

pause 4 150

350

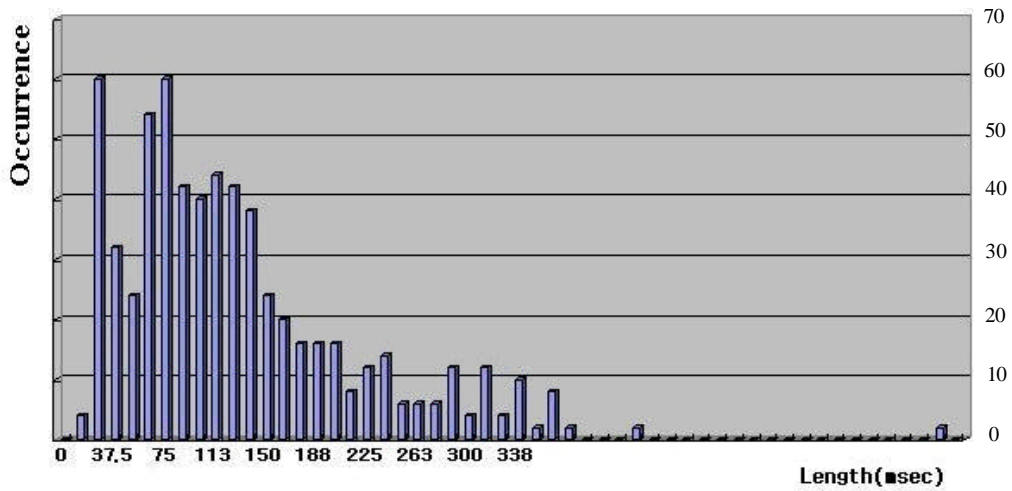
가

가

3.1

150ms 가 10ms 300ms 30ms

가



3.1

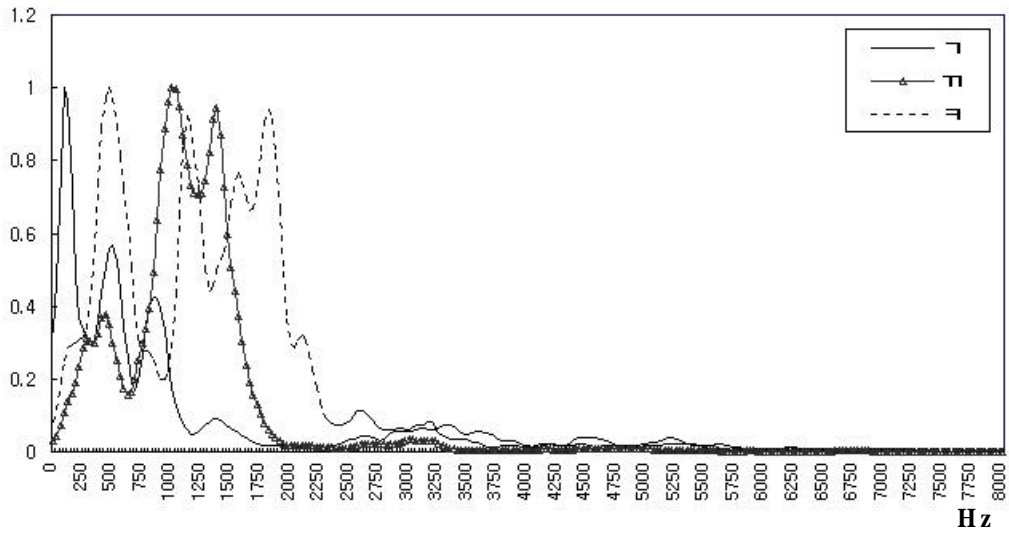
Figure 3.1 Duration distribution of unvoiced consonants

3.1.2 (unvoiced consonant)

가

가 , 가

가 가



3.2 /ㄱ/, /ㄷ/, /ㅋ/

Figure 3.2 Normalized power density spectrum of velar plosives

3.2 (velar plosive) /ㄱ/, /ㄷ/, /ㅋ/

(normalization)

3500hz

300hz 2500hz

가 . /ㄱ/

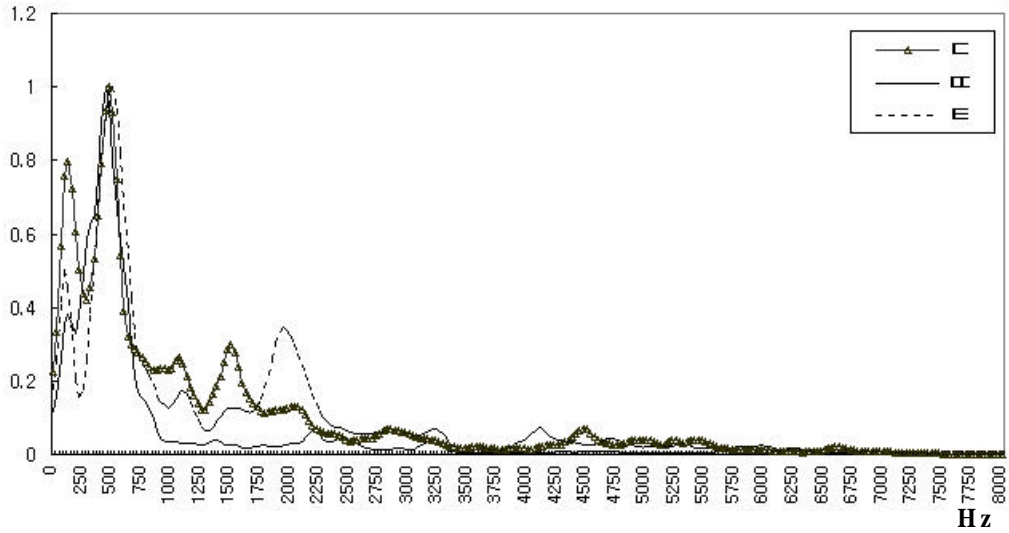
/ㄷ/, /ㅋ/

가

3.3 (alveolar plosive) /ㄷ/, /ㄸ/, /ㅌ/

가

350hz 1000hz



3.3 /ㄷ/, /ㄸ/, /ㄷ/

Figure 3.3 Normalized power density spectrum of alveolar plosives

가

가

가

3.4 (bilabial plosive) /ㅂ/, /ㅃ/, /ㅍ/

가

1000hz 3500hz

가

/ㅃ/

300hz 800hz

가

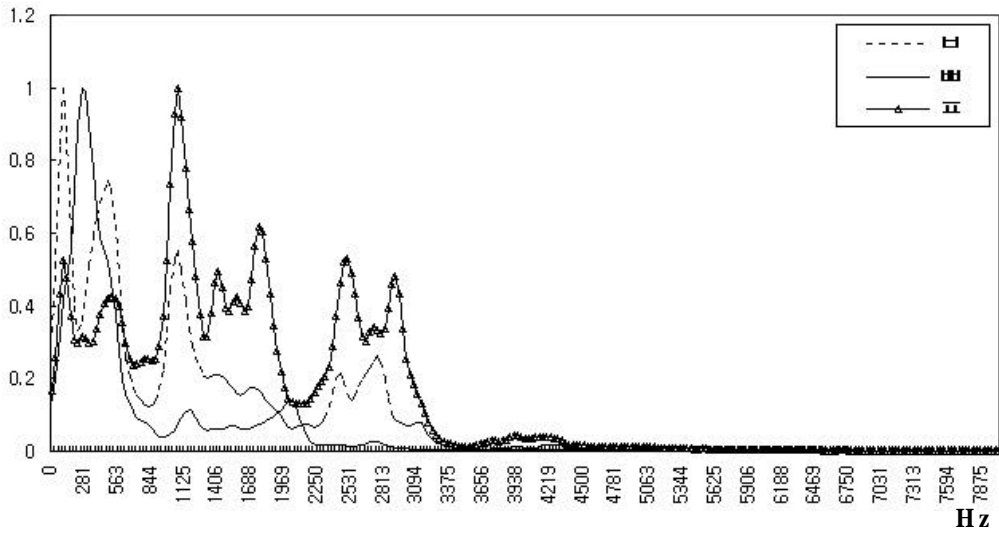
가

/ㅃ/

, /ㅃ/

1000hz 2250hz

가



3.4 /ㅂ/, /ㅃ/, /ㅍ/

Figure 3.4 Normalized power density spectrum of bilabial plosives

3.5 (fricative) /ㅅ/, /ㅆ/, /ㅎ/

(alveolar fricative) /ㅅ/, /ㅆ/

(glottal fricative) /ㅎ/

가

. /ㅆ/

3500hz 7000hz

, /ㅅ/

3500hz 6000hz

가

/ㅎ/

500hz 2200hz

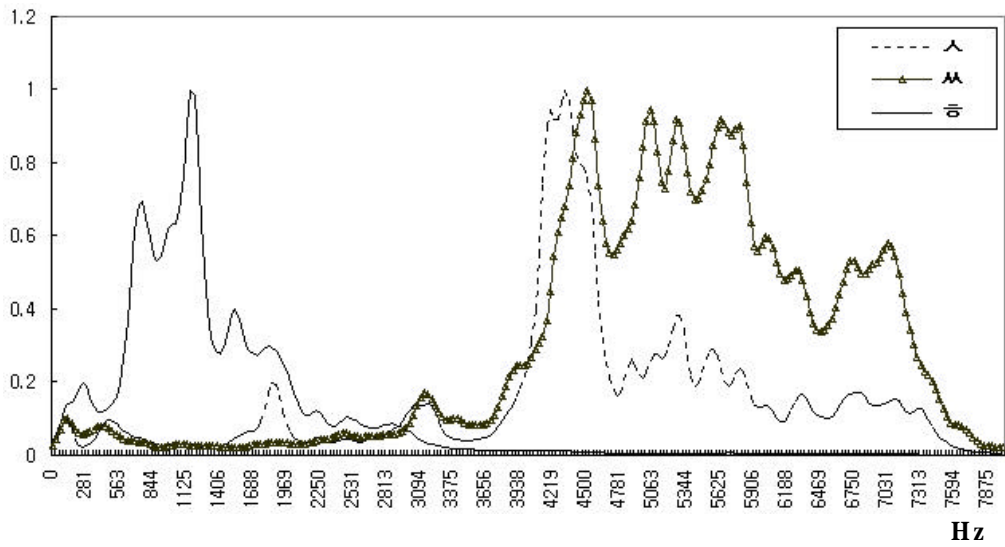
가

, /ㅎ/

가 /ㅅ/, /ㅆ/

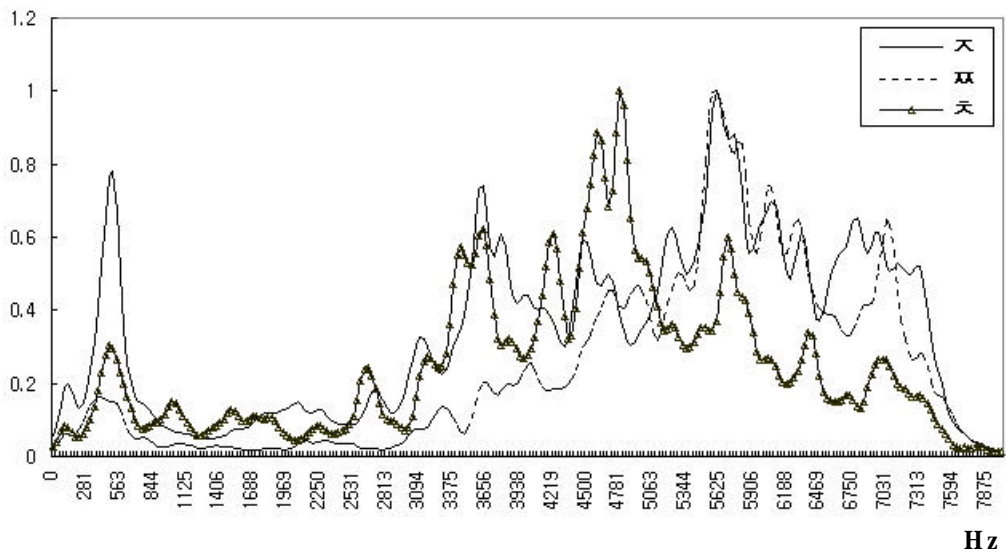
3.6 (affricate) /ㅈ/, /ㅉ/, /ㅊ/

가



3.5 /s/, /ss/, /ᄁ/

Figure 3.5 Normalized power density spectrum of fricatives



3.6 /ㅌ/, /ㅌㅌ/, /ㅌㅌ/

Figure 3.6 Normalized power density spectrum of affricates

3.2 (noise)

가 (additive noise) 가 (convolution) 가 (channel distortion)

3.2.1 가

가 가 ,

[14].

가 가

가 $n(n)$ 가 $x(n)$, 가 $y(n)$ (3.1)

$$y(n) = x(n) + n(n) \tag{3.1}$$

(3.1)

가 가 ,

가 (spectral subtraction) [15].

3.2.2 [5]

가

,

,

.

•

,

가

가

가

가

가

[5].

[16]. Hermanski

[18].

•

.

가

가

.

,

.

가

. Acero

65%

가

[17].

3.3

가

가

가

가

가

가

가

3.2

가

가

가

가

가

가

가

NoiseX-92

NoiseX-92

6

1(volvo noise),
 2(leopard noise), (babble noise), (factory
 noise), (pink noise), F16 (F16 noise) (white
 noise)

3.7

가

3.8

가

1000Hz

가

3.9

3.10

100 500Hz

가

3.11 F16

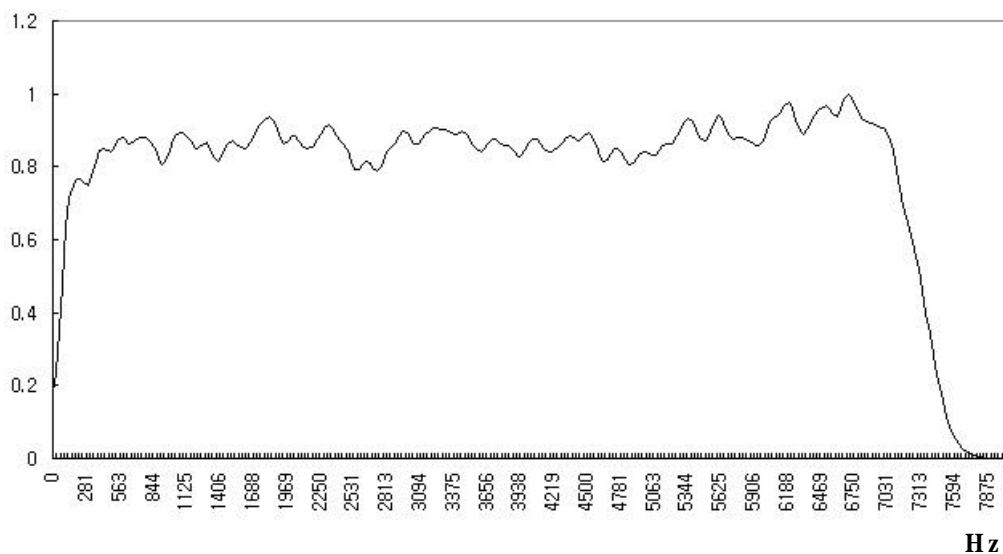
가

3.11 3.12

300Hz

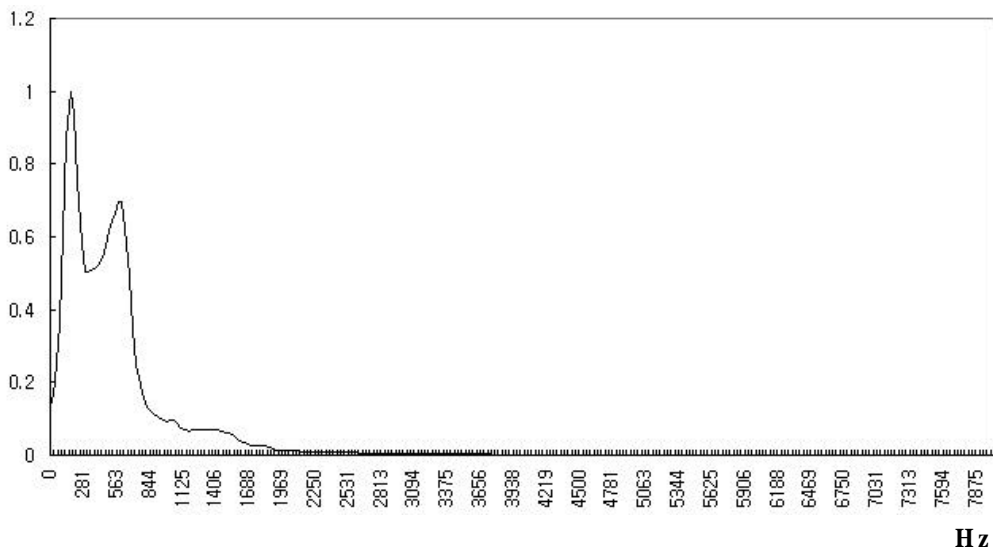
가

가



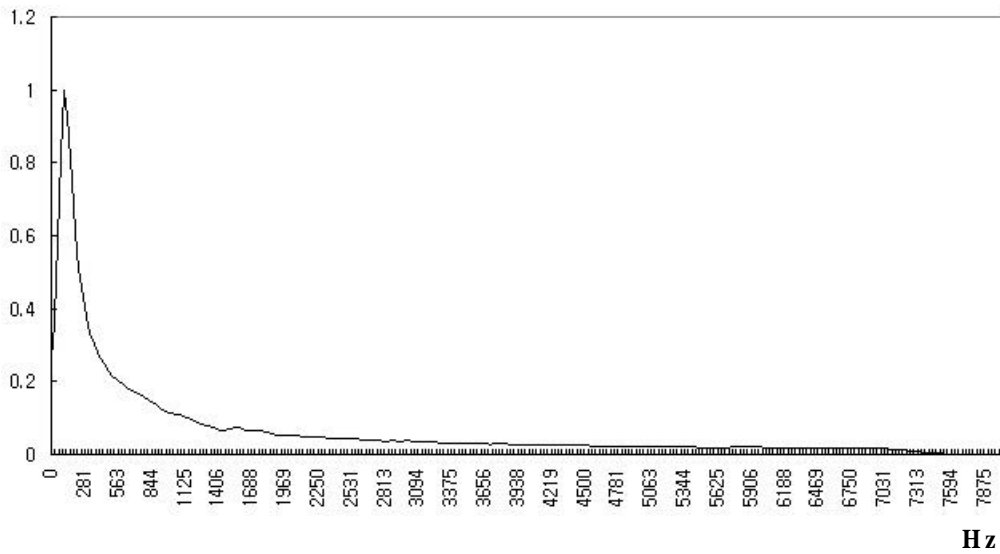
3.7

Figure 3.7 Normalized power density spectrum of white noise



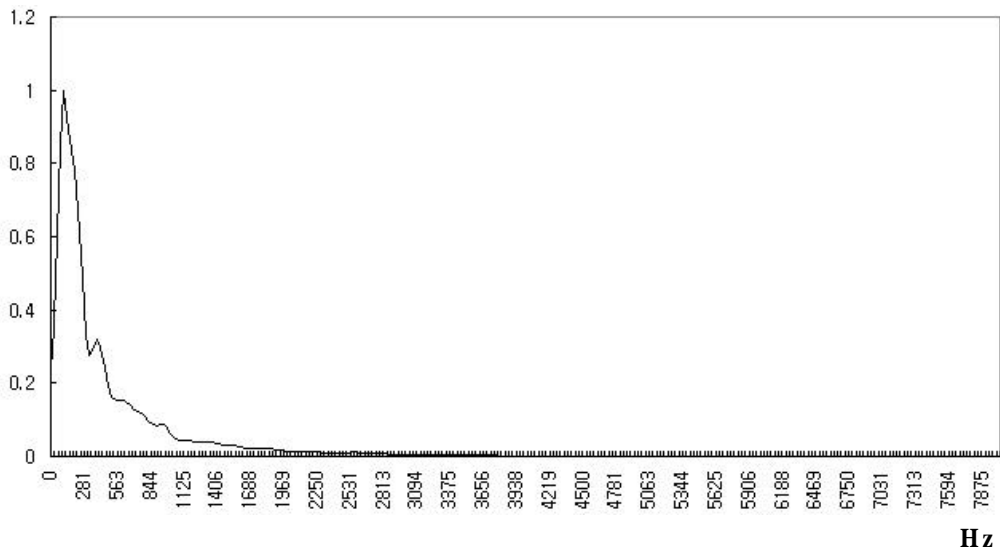
3.8 babble

Figure 3.8 Normalized power density spectrum of babble noise



3.9

Figure 3.9 Normalized power density spectrum of pink noise



3.10

Figure 3.10 Normalized power density spectrum of factory noise

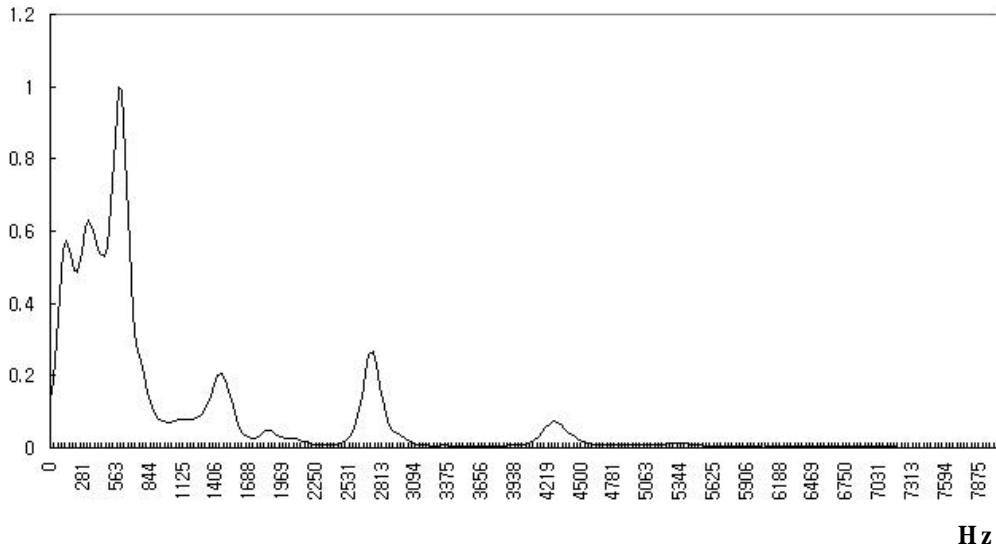


Figure 3.11 Normalized power density spectrum of F16 noise

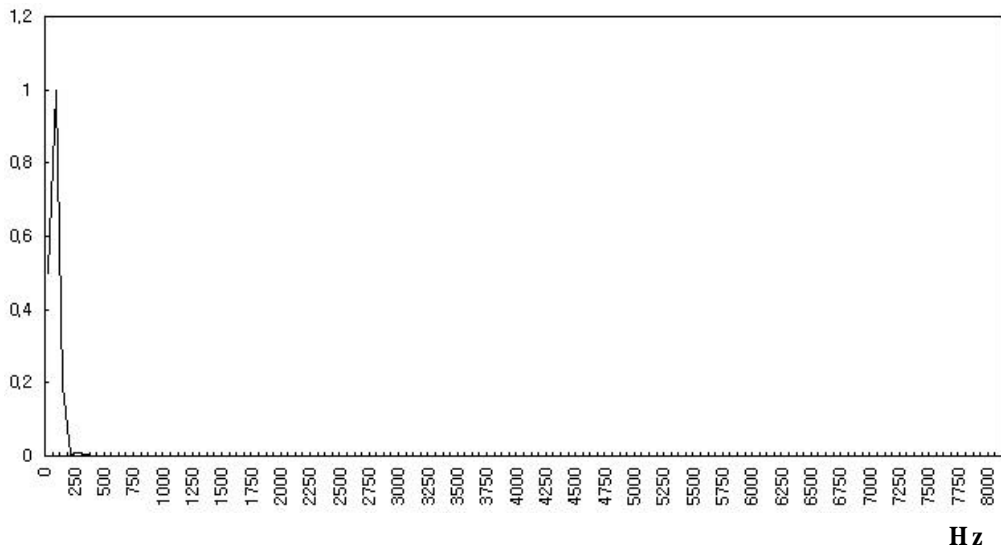
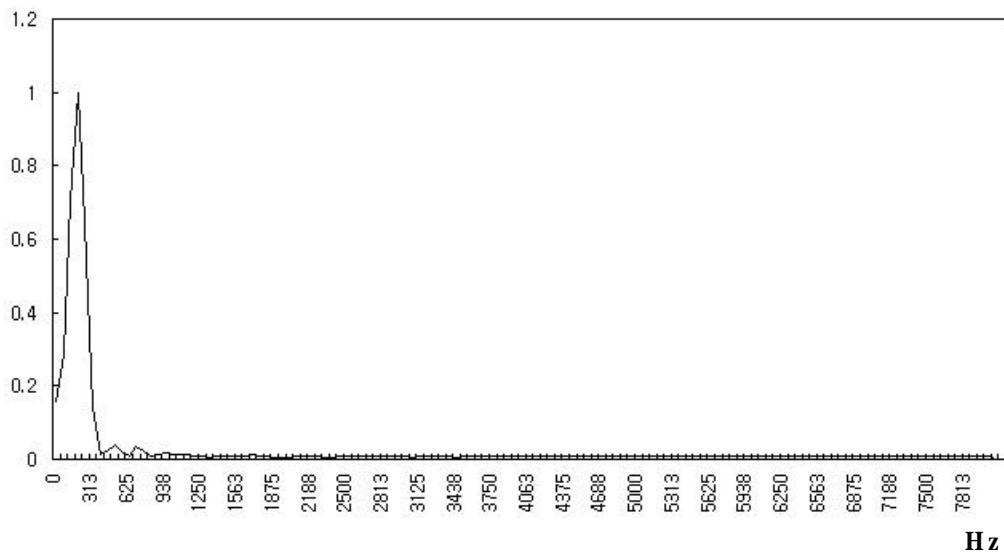


Figure 3.12 Normalized power density spectrum of car(volvo) noise



3.13 (leopard)

Figure 3.13 Normalized power density spectrum of car(leopard) noise

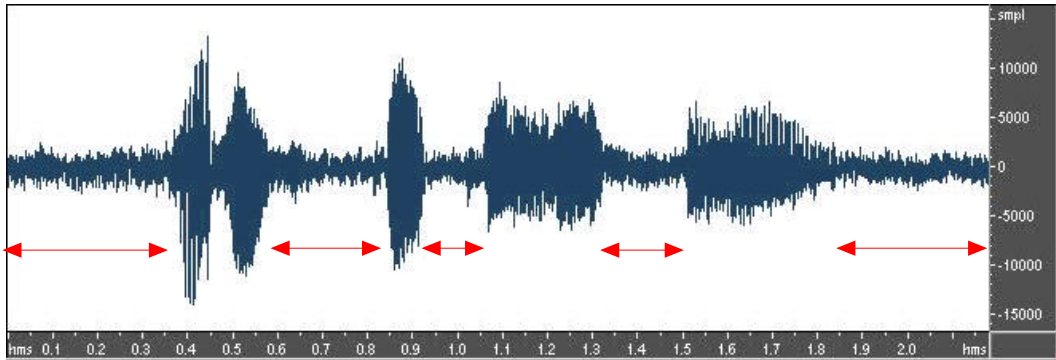
4

4.1 SNR 10dB F16 가

(a) 가
, (b) 가
가

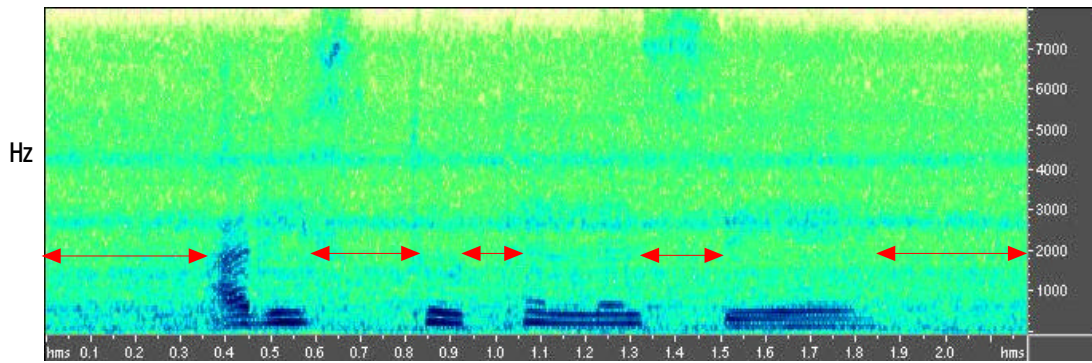
4.2 가
4.1 , 4.2

4.3 4.4 .



Time(sec)

(a)



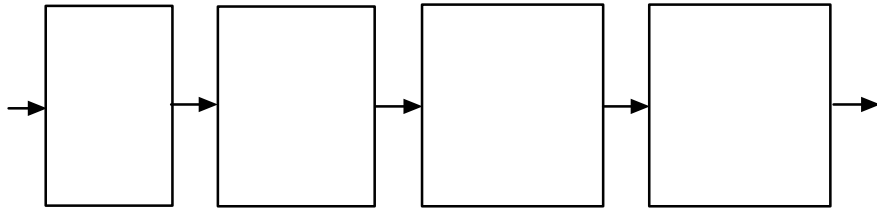
Time(sec)

(b)

4.1 SNR 10dB F16 가 (a)

(b)

Figure 4.1 (a)Signal wave and (b)spectrogram of the speech signal contaminated by F16 noise of SNR 10dB

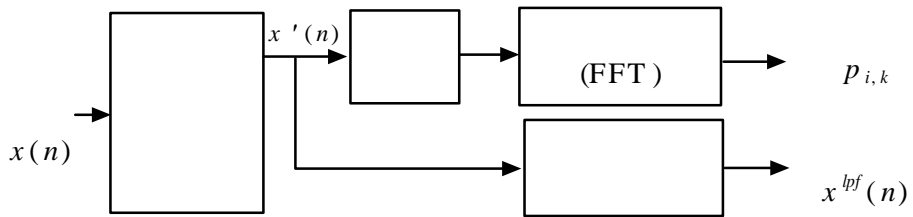


4.2

Figure 4.2 Overall block diagram of speech region detection

4.1

4.3



4.3

Figure 4.3 Block diagram of pre-processing

4.1.1

(bias)

가

가 가 . (4.1)

, (4.2) . N

, n .

$$x_{mean} = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \quad (4.1)$$

$$x'(n) = x(n) - x_{mean}, \quad \text{for all } n \quad (4.2)$$

4.1.2

(10ms 30ms)

.

16ms (256samples/frame)

.

8ms (128samples/frame)

.

4.1.3 (Power Spectrum)

$x_i'(n)$ (Hanning window) $w(n)$ N-point FFT

$$X_i(k) p_{i,k} \quad (4.6) \quad . \quad i,$$

k , .

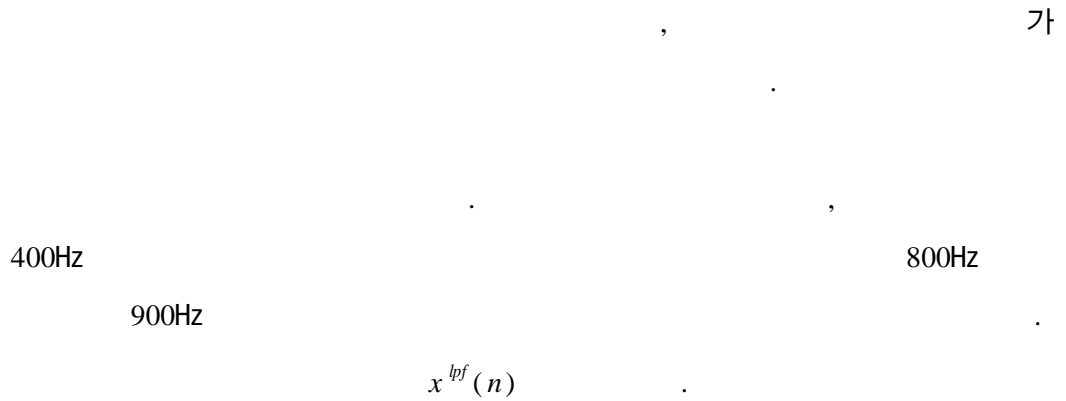
$$x_i''(n) = x_i'(n) w(n) \quad (4.3)$$

$$w(n) = \frac{1}{2} \left(1 - \cos \left(2\pi \frac{n}{N-1} \right) \right), \quad 0 \leq n \leq N-1 \quad (4.4)$$

$$X_i(k) = \sum_{n=0}^{N-1} x_i(n) e^{-j\frac{2\pi nk}{N}}, \quad k = 0, 1, 2, \dots, N-1 \quad (4.5)$$

$$p_{i,k} = |X_i(k)|^2 \quad (4.6)$$

4.1.4



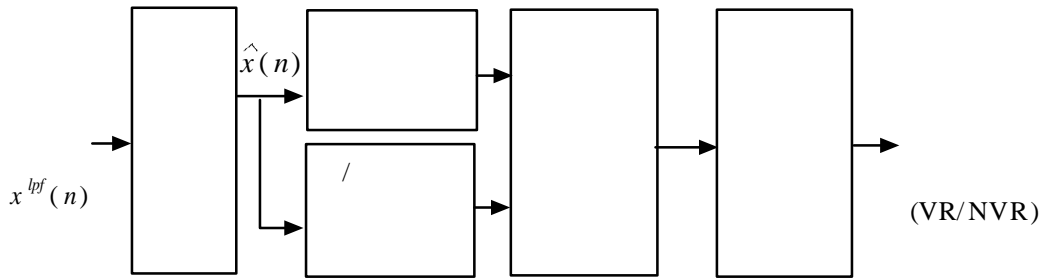
4.2

, [2, 21]

4.4

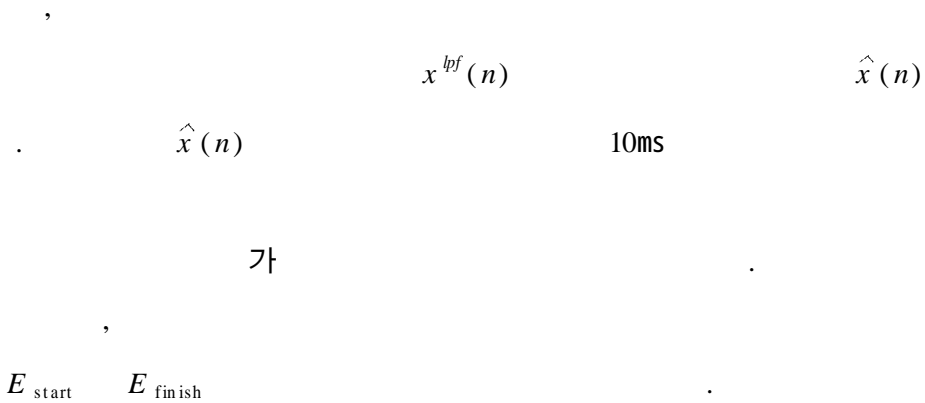
4.2.1

가



4.4

Figure 4.4 Block diagram for the detection of boundaries between voiced region and non voiced region



$$E(m) = \sum_{n=n_s}^{n_E} \hat{x}(n)^2, \quad n_s = ((m+1) \cdot 10)\text{ms}, \quad n_E = (n_s + 10)\text{ms} \quad (4.7)$$

$$E_{\text{silence}} = \frac{1}{M} \sum_{m=1}^M E(m) \quad (4.8)$$

$$E_{\text{start}} = c_S \cdot E_{\text{silence}} \quad (4.9)$$

$$E_{\text{finish}} = c_F \cdot E_{\text{silence}} \quad (4.10)$$

$E(m)$ m
 M 20ms 200ms
 c_S c_F
 2.3 2.0 E_{start}
 E_{finish}
 (pitch)

4.2.2

4.2.2

(4.11)

$$E_i^{lpf}$$

$$E_i^{lpf} = \frac{1}{L} \sum_{k=0}^L \hat{p}_{i,k} \quad (4.11)$$

L 600Hz i
 200 ms k
 E_i^{lpf} (4.3.2)
 $E^{lpf(peak)}$ E_{NTH}

$$E_{N_{TH}} = E^{lpf(peak)} \cdot (1 + \alpha_E) \quad (4.12)$$

$$\alpha_E = E_{max} \text{ dB} - E^{lpf(peak)} \text{ dB} \quad (4.13)$$

$$E^{lpf(peak)} \text{ dB} = 10 \cdot \log_{10} E^{lpf(peak)} \quad (4.14)$$

, $E_{max} \text{ dB}$ 가 가 dB
 96 . α_E $E^{lpf(peak)}$
, $E^{lpf(peak)}$ 가 가
 $E_{N_{TH}}$, $E^{lpf(peak)}$ 가 가
 $E_{N_{TH}}$.

. , 4.2.1

(over - exception)

$$(4.15)$$

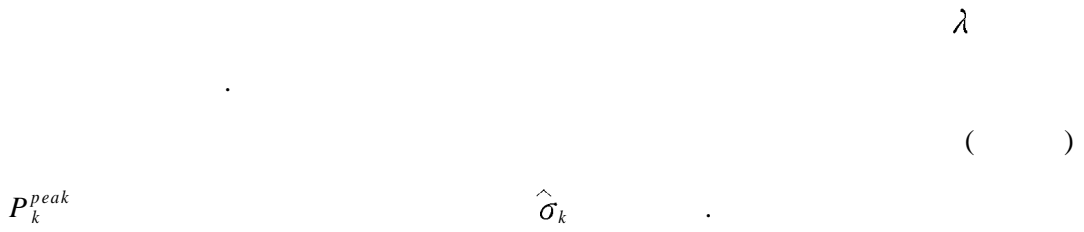
3

$$\begin{aligned} N VR_L' &= N VR_L - 3 \\ N VR_R' &= N VR_R + 3 \end{aligned} \quad (4.15)$$

, $N VR_L$ $N VR_R$.

$$E_{N_{TH}} \quad E_i^{lpf}$$

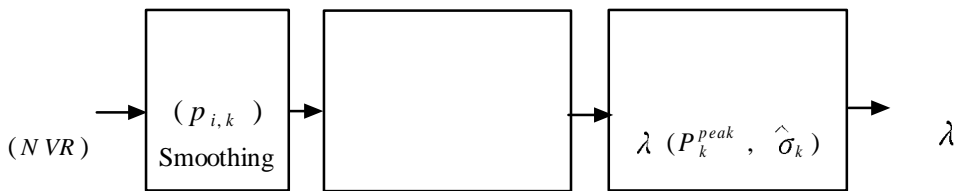
4.3



Hirsch [9]

가

λ
4.5



4.5

Figure 4.5 Block diagram for the detection of noise signal parameter

4.3.1 (Smoothing)

가
 (4.16) moving
 window filter $l + 1$
 (smoothing)

$$\hat{p}_{i,k} = \frac{1}{2l+1} \sum_{m=i-l}^{i+l} p_{m,k} \quad (4.16)$$

4.3.2

(200 ms)

$$\lambda(P_k^{peak}, \hat{\sigma}_k)$$

4.6 k

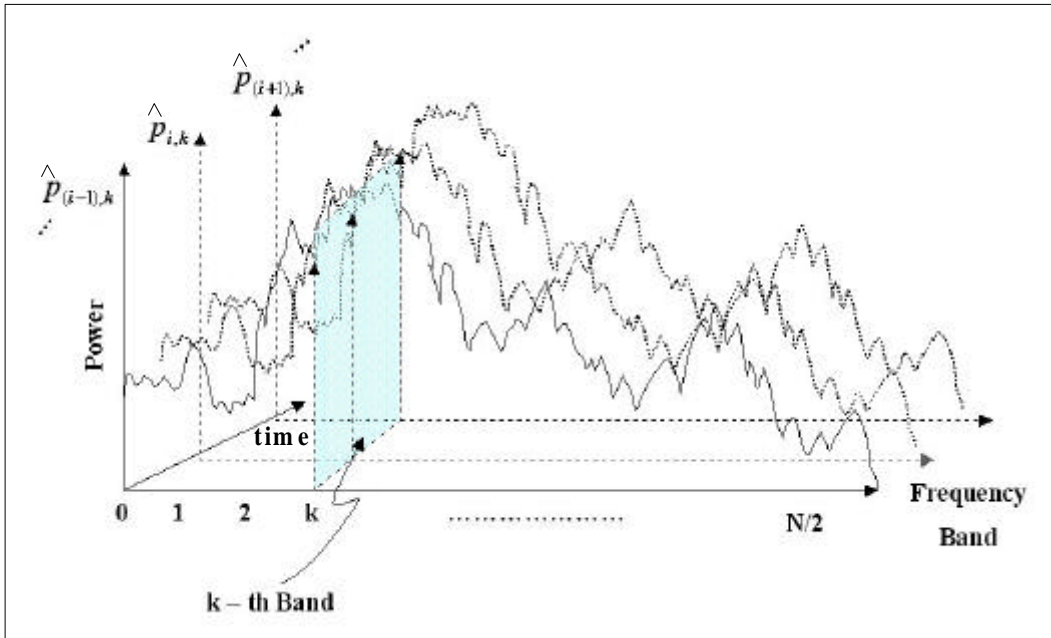
(NR : noise region)

k

$$\lambda(P_k^{peak}, \hat{\sigma}_k) \quad k$$

$$\hat{p}_{i,k} \quad (i \in NR)$$

i , k , nF rames



4.6

Figure 4.6 Power spectrums of the frames used in Histogram

(4.17), (4.18)

$mean_k$ σ_k .

$$mean_k = \frac{1}{nFrames} \sum_{i \in NR} \hat{p}_{i,k} \quad (4.17)$$

$$\sigma_k = \sqrt{\sum_{i \in NR} (\hat{p}_{i,k} - mean_k)^2 / nFrames} \quad (4.18)$$

$$p_k^{\min} = \min_{i \in NR} (\hat{p}_{i,k}), \quad \text{for } |\hat{p}_{i,k} - \text{mean}_k| < 2.5\sigma_k \quad (4.19)$$

$$p_k^{\max} = \max_{i \in NR} (\hat{p}_{i,k}), \quad \text{for } |\hat{p}_{i,k} - \text{mean}_k| < 2.5\sigma_k$$

(quantization)

Nq 가
(4.17), (4.18)

$$(4.19) \quad \text{가} \quad 2.5\sigma_k \quad p_k^{\min}, p_k^{\max}$$

$$\text{mean}_k \quad 2.5\sigma_k \quad p_k^{\min}, p_k^{\max}$$

$$p_k^{\min} \quad Nq \quad p_k^{\min}$$

가 가

$$Iq \quad (4.20)$$

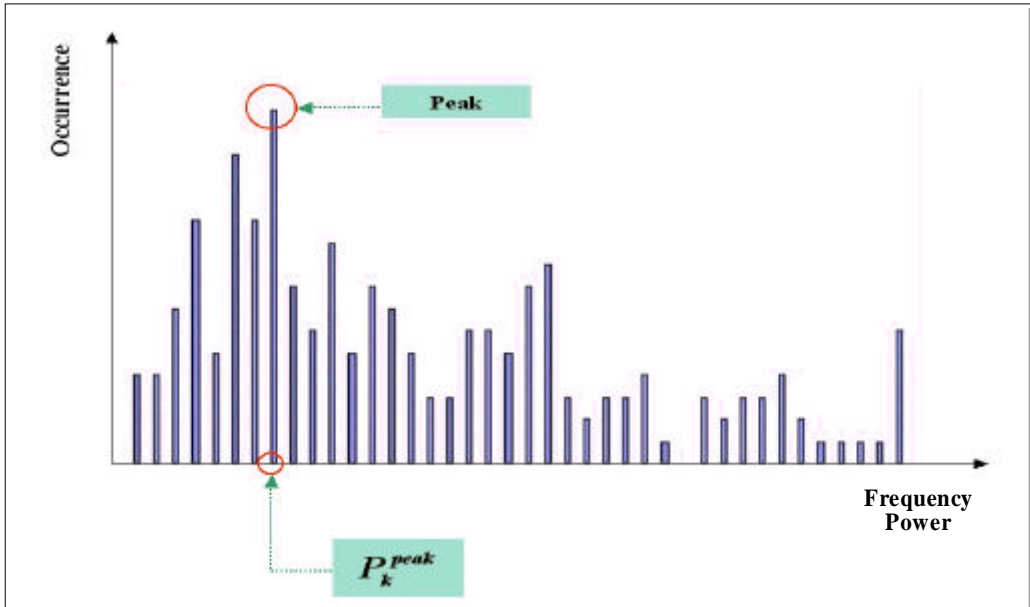
$$Nq = \frac{p_k^{\max} - p_k^{\min}}{p_k^{\min}} \quad (4.20a)$$

$$Iq = \frac{p_k^{\max} - p_k^{\min}}{10} \quad (4.20b)$$

4.7 k

가 가 k P_k^{peak} ,

$$P_k^{peak} \quad \hat{\sigma}_k \quad (4.21)$$



4.7 k (P_k^{peak})

Figure 4.7 Peak power P_k^{peak} of the k -th frequency band histogram

$$\hat{\sigma}_k = \sqrt{\sum_{i \in NR} (\hat{p}_{i,k} - P_k^{peak})^2 / nFrames} \quad (4.21)$$

$$P_k^{peak} \quad \hat{\sigma}_k \quad k \quad \lambda(P_k^{peak}, \hat{\sigma}_k)$$

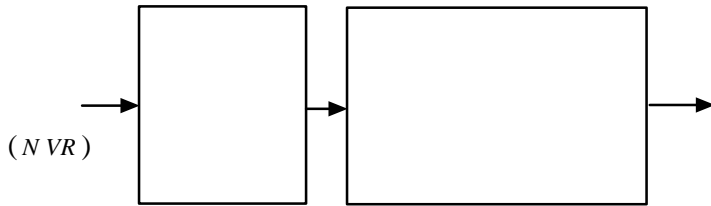
4.4

4.4

4.8

4.3.2

$$\lambda(P_k^{peak}, \hat{\sigma}_k)$$



4.8

Figure 4.8 Block diagram for between unvoiced region and noise region in non voiced region

4.4.1

(4.22)

$$\begin{aligned}
 & \text{Cnt}_i \\
 & N \text{ FFT} \quad , i \quad , k \\
 & k \quad \hat{p}_{i,k} \quad P_k^{peak} \quad \text{가} \quad \hat{\sigma}_k \\
 & flag_{i,k} \quad 1 \quad 0 \\
 & flag_{i,k} \quad 1 \quad , 0
 \end{aligned}$$

$$Cnt_i = \sum_{k=0}^{N/2} flag_{i,k} \quad (4.22)$$

$$flag_{i,k} = \begin{cases} 1, & |\hat{p}_{i,k} - P_k^{peak}| > \hat{\sigma}_k \\ 0, & otherwise \end{cases}$$

$$\Delta Cnt_i = \sum_{k=2}^{N/2} \Delta flag_{i,k} \quad (4.23)$$

$$\Delta flag_{i,k} = \begin{cases} 1, & |p_{i,k} - p_{i-2,k}| > \hat{\sigma}_k \\ 0, & otherwise \end{cases}$$

$$Cnt_{i,k} \quad Cnt_{TH}^{unv} \quad Cnt_{TH}^{nois} \quad (4.24)$$

$$Cnt_{TH} = \max(Cnt_{TH}^{min}, \frac{1}{L} \sum_{k=0}^L Cnt_{i,k}) \quad (4.24a)$$

$$Cnt_{TH}^{min} = \frac{N}{\beta}$$

$$Cnt_{TH}^{unv} = Cnt_{TH} \cdot C_{unv} \quad (4.24b)$$

$$Cnt_{TH}^{nois} = Cnt_{TH} \quad (4.24c)$$

L , N FFT

β 4.5 C_{unv} Cnt_{TH}^{unv}

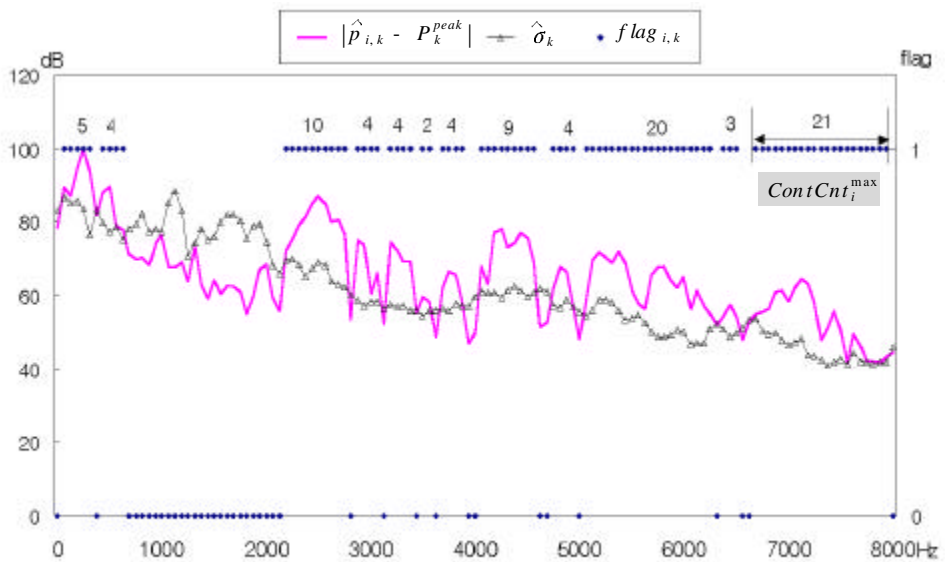
1.1

3.2

$ContCnt_i^{max}$ $ContCnt_i^{max}$

4.9 $flag_{i,k}$ 1

가



4.9 $flag_{i,k}$ i $ContCnt_i^{max}$

Figure 4.9 $ContCnt_i^{max}$ of the i -th frame by $flag_{i,k}$

$ContCnt_i^{max}$

$ContCnt_{TH}^{unv}$ $ContCnt_{TH}^{nois}$. 2

3 7 4

4.4.2

4.10

NVR

2.2.2

2.4

VR NVR

NVR

NVR

NVR

(rightward detection) , NVR

(leftward detection) .

NVR 가 가

4.10

Ⓐ

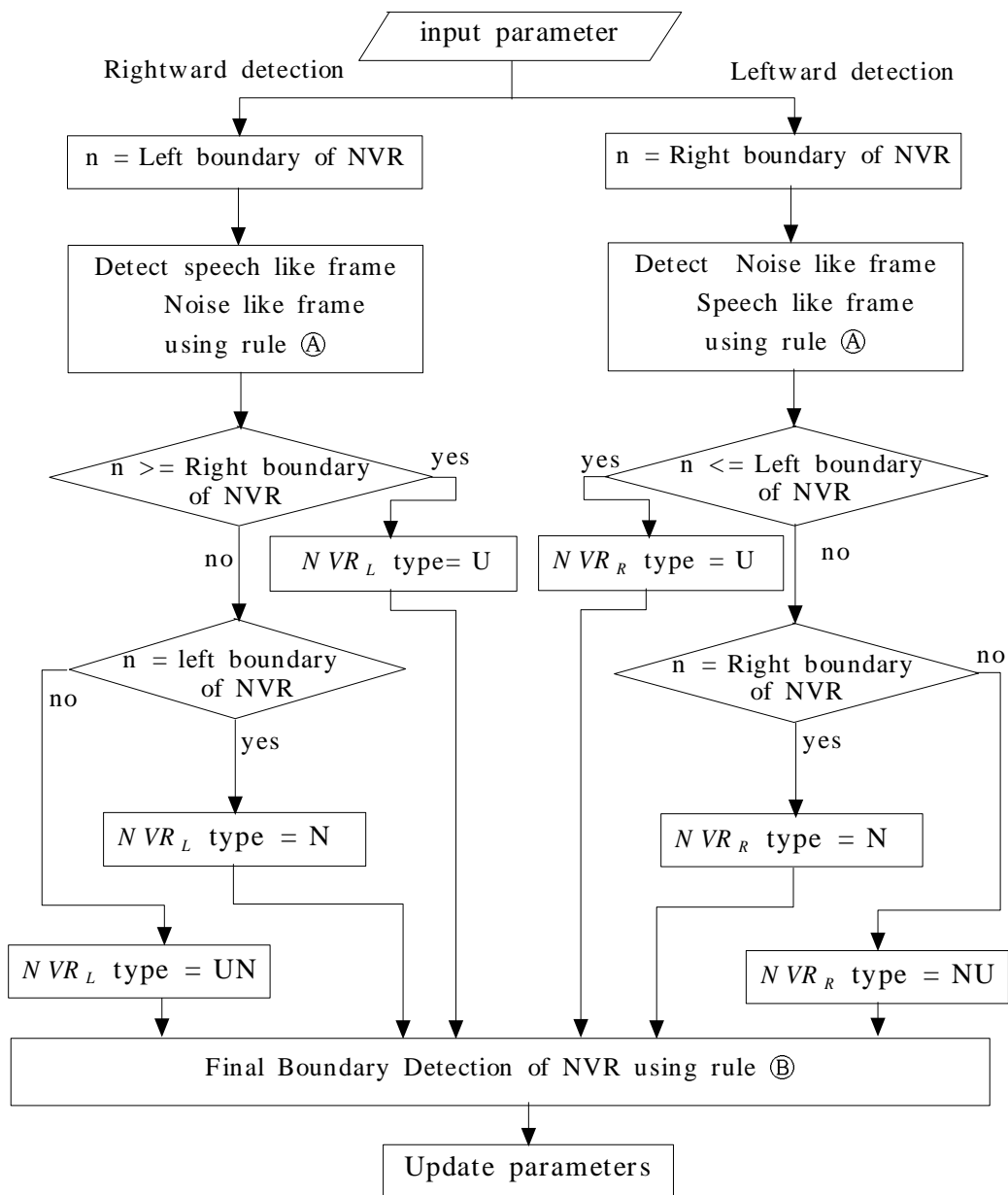
Ⓐ

$$ContCnt_n^{max} < ContCnt_{TH}^{unv}$$

$$\Delta Cnt_n < Cnt_{TH}^{unv}$$

$$|Cnt_n - Cnt_{n\pm 1}| < 15$$

$$|\Delta Cnt_n - \Delta Cnt_{n\pm 1}| < 15$$



4.10

Figure 4.10 Flowchart for the detection of boundaries between unvoiced consonant and noise region

, NVR NVR .
 . NVR 가
 N, U, NU, UN 가 .
 . NVR , 2.2.2
 2.4 NVR ⑥
 .
 NVR
 λ Cnt_{TH}^{unv} Cnt_{TH}^{nois} (update) .
 NVR 가
 가 .
 600ms
 NVR
 가

For $i = 0$ **to** N_{NR}
 $Frame_{up}[i] = Frame_{up}[N_{NR} + i]$
For $i = 0$ **to** N_{NR}
 $Frame_{up}[N_{NR}] = Frame_{NR}[i]$

N_{NR} $Frame_{NR}$ NVR
 , $Frame_{up}$.

, 4.1 ,

λ

, ,

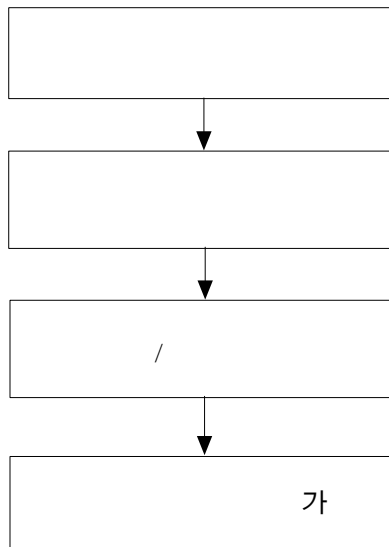
NVR

.

5 가

5.1

가 가 가 .
가
가



5.1

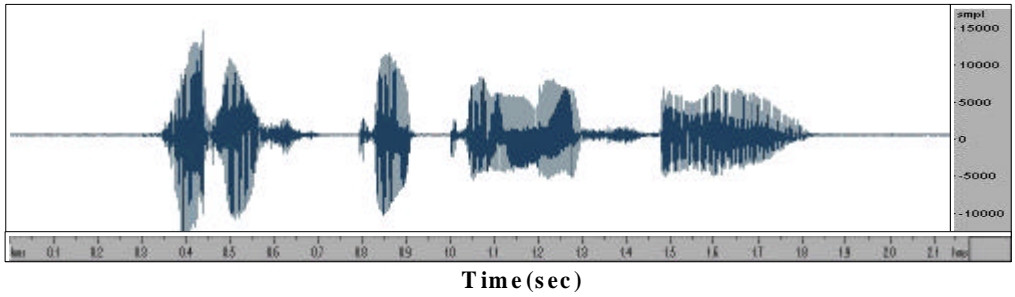
Figure 5.1 Block diagram of simulation for detecting unvoiced and noise boundaries

5.1

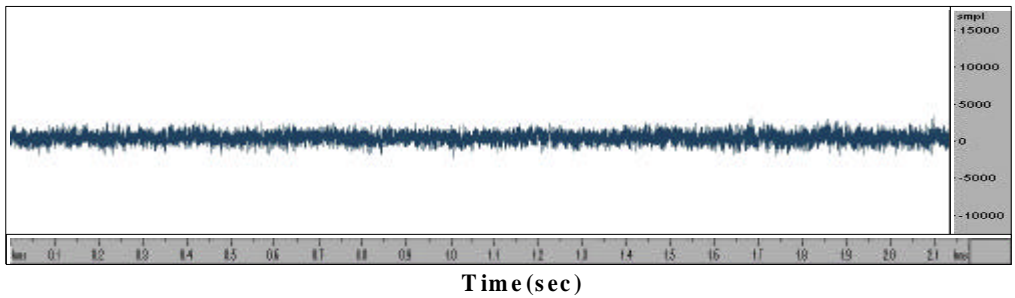
ETRI POW (phonetically
optimized word) (corpus) . 40
3,848 POW
240 .
40 가 3 .
NoiseX-92 .
(Leopard volvo), F16 (F16),
(Babble), (Factory), (Pink)
(white) 7 .
가 . SNR
10dB 15dB 가 .

5.2

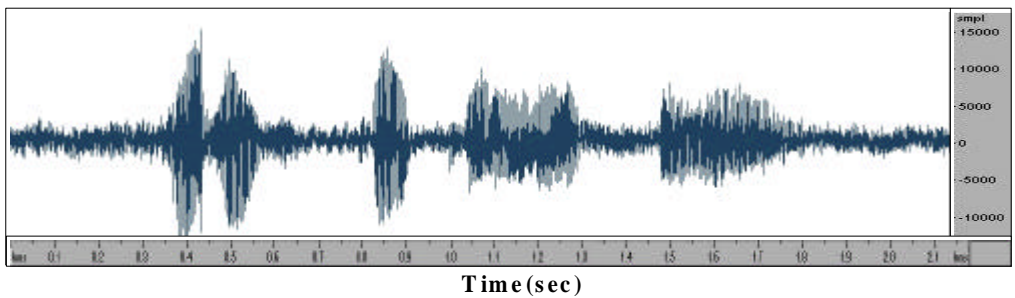
5.2.1 가
5.2 가 / /
5.4 F16 가 가 ,
5.3 F16 5.2
SNR 5.4
가



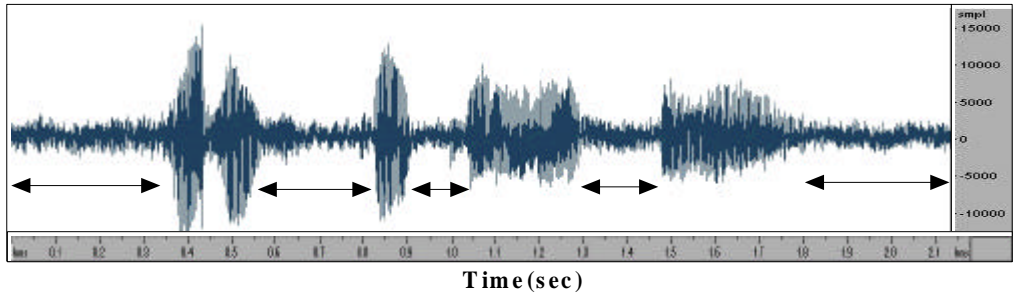
5.2 가 / /
Figure 5.2 Clean speech wave of a female speaker's utterance /ɑ-li-s-t^ho-t^hel-le-s-e/



5.3 F16
Figure 5.3 Noise wave of the F16 flight cockpit



5.4 5.2 F16 (5.3)가 가 ,
 SNR=10dB
Figure 5.4 F16 noise(fig. 5.3) added speech wave of fig. 5.2, SNR=10dB



5.5 5.4

Figure 5.5 Detected unvoiced region in fig. 5.4

5.2.2

[2, 21]

5.5 , , , , ,

5.5

50% 가

5.2.3

4.4

, FFT 256 ,

128

가

가

20ms,

가

10ms

5.6

5.10

5.5

NVR

$ContCnt_i^{\max}$, Cnt_i , Cnt_{TH}^{unv} , Cnt_{TH}^{nois} ,
 E_i^{lpf} , $E_{N_{TH}}$

가

10

“- 200,000”

, “300,000”

5.6 NVR

E_i^{lpf} 가

5.6

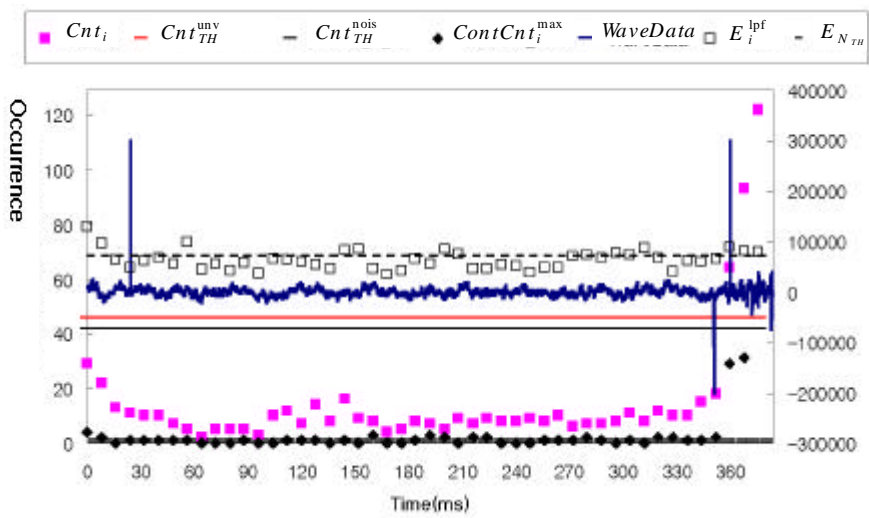
5.7 NVR 가 UNU

가

가

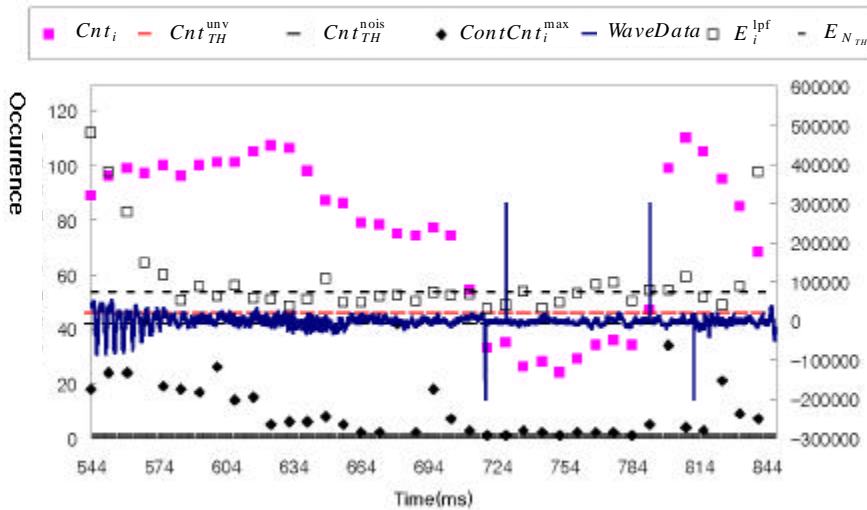
5.8 NVR 가 NU

5.9 U



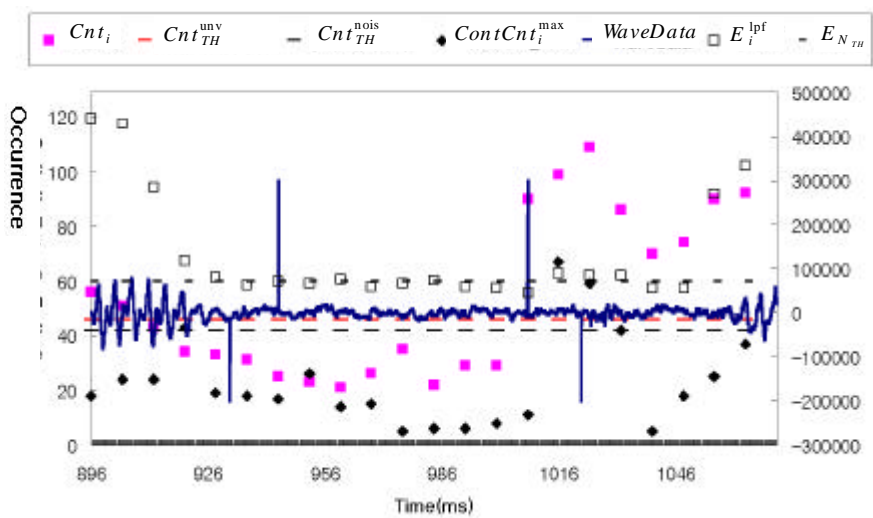
5.6 5.5 NVR

Figure 5.6 Result of boundary detection in -th NVR of fig. 5.5



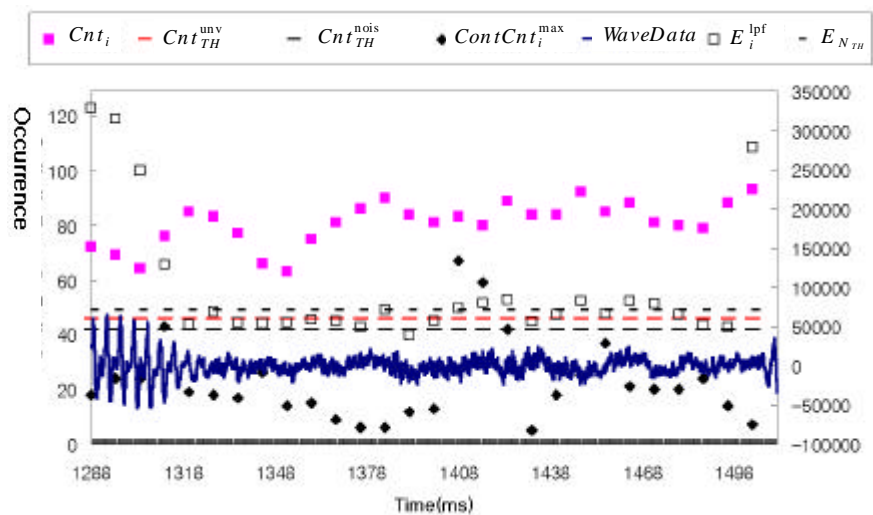
5.7 5.5 NVR

Figure 5.7 Result of boundary detection in -th NVR of fig. 5.5



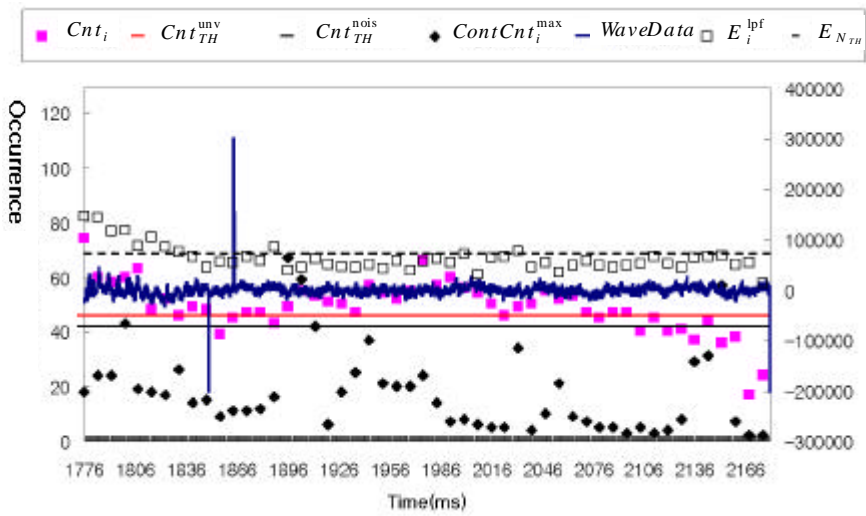
5.8 5.5 NVR

Figure 5.8 Result of boundary detection in -th NVR of fig. 5.5



5.9 5.5 NVR

Figure 5.9 Result of boundary detection in -th NVR of fig. 5.5



5.10 5.5 NVR

Figure 5.10 Result of boundary detection in -th NVR of fig. 5.5

5.10 가 NVR

N

5.1

SNR 10dB

89.7%, 15dB

91.4%

가

SNR

$$Cnt_i \quad \Delta Cnt_i$$

가

(stop)

(pause)

가

5.1 NVR

Table 5.1 Result of boundary detection between unvoiced consonant and noise in NVR by proposed method

	(Correct Detection)(%)	
	SNR=10dB	SNR=15dB
Volvo(car)	92.7	94.0
Leopard(car)	94.5	95.3
F16	87.3	91.3
Babble	88.8	92.7
Factory	86.0	86.7
Pink	94.0	95.3
white	88.8	91.4
	90.30	92.55

가

가

가

16ms 40ms

가

가

가

가

가

. 가
가 가 .

가

. 600ms

가 .

6

가 3

,

,

가

가

가

가

6가

SNR

10dB

15dB

가

가

90%

(musical tone)

가

가

- [1] J. G. Wilpon, L. R. Rabiner, and T.B. Martin, "An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints," *AT&T tech.J.*, 63 (3): 479-798, March 1984.
- [2] , , , 2000.
- [3] , , , 1996.
- [4] , , , , " / - , " , 13 , 6 , pp. 31-43, 1994.
- [5] , , , 1998.
- [6] , " , " <http://maincc.hufs.ac.kr/~kimwy/lin-in/fonologia.htm>.
- [7] Jean-Claude Junqua, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise," *IEEE Transactions on Speech & Audio Processing*, Vol. 2, No. 3, pp. 406-412, July 1994.
- [8] , , , , , " , " , 10 , 1 , pp. 257-260, 1997.
- [9] H. G. Hirsch, "Estimation of Noise Spectrum and its Application to SNR Estimation and Speech Enhancement," Technical Report TR-93-012, International Computer Science Institute, Berkeley, USA, 1993.
- [10] H. G. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for

- Robust Speech Recognition,” in Proc. IEEE Int. Conference Acoustic, Speech and Signal Processing, ICASSP-95, (Detroit, Michigan), pp. 153- 156, 1995.
- [11] S. F. Boll and D. C. Pulsipher, “Suppression of Acoustic Noise in Speech Using Two Microphone Adaptive Noise Cancellation,” IEEE. Trans on ASSP, vol. 28, pp. 752-755, 1980.
- [12] Nathalie Virag, “Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System,” IEEE. Transactions of Speech and Audio Processing, Vol. 7, No. 2, march 1999.
- [13] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proc. IEEE, vol, 67, pp. 1586- 1604, Dec. 1979.
- [14] D. C. Bateman, D. K. Bye, and M. J. Hunt, “Spectral contrast normalization and other techniques for speech recognition in noise,” in ICASSP, pp. 241-244, 1992.
- [15] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” in ICASSP, ASSP-27(2), pp. 113- 120, 1979.
- [16] C. Mokbel, J. Monne and D. Juvet, “On line adaptation of a speech recognizer to variations in telephone line conditions,” in EUROSPEECH, pp. 1247- 1250, 1993.
- [17] Acero, A. and R. Stern, “Environmental robustness in automatic speech recognition,” in ICASSP, pp. 849-852, April 1990.
- [18] P. Moreno and R. Stern, “Sources of degradation of speech recognition

in the telephone network,” in ICASSP, pp. 109- 112, 1994.

- [19] , , , “ , ” 9 , SCAS-9 , 1 , pp. 133- 137, 1992.
- [20] Rathinavelu Chengalvarayan, “Robust energy normalization using speech/nonspeech discriminator for german connected digit recognition,” in EUROSPEECH, pp. 61-63, 1999.
- [21] D. G. Ha and O. K. Shin, “Adaptation of pitch information in vowel feature extraction for speech recognition,” EALPIIT 2000, pp. 324- 329, 2000.
- [22] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

2

가

가

가

가

‘ 88’

가

가

SNR

10dB

15dB

90%

가

Hirsch