A Study on Noise Robust Acoustic Analysis for Automatic Speech Recognition

Shoji Kajita

Abstract

A major difficulty encountered in a current automatic speech recognizer (ASR) is that the recognition performance degrades rapidly in the presence of noise and distortion, due primarily to the acoustic mismatches in training and recognizing conditions. Considerable effort has been made to overcome this problem. In a broad context, the research has been focused primarily on three areas; (1) noise robust feature extraction, (2) speech enhancement and (3) speech model compensation for noise. Among these areas, this dissertation focuses on (1) noise robust feature extraction, and describes the development of a new acoustic analysis technique, subband-autocorrelation (SBCOR) analysis and its extensions.

At first, the basic investigation of SBCOR analysis is described to show how to apply it to speech recognition under noisy conditions. SBCOR analysis is an acoustic analysis technique based on filter bank and periodicity extraction associated with the inverse of center frequency $f_{cf_i}^{-1}$. In experiments, SBCOR is evaluated for five types of filter bank and three autocorrelation detectors using a speaker-dependent DTW word recognition system under noisy conditions degraded by multiplicative white noise. The experimental results showed that SBCOR performs equally as well as smoothed group delay spectrum (SGDS) under clean conditions, and much better than it under noisy conditions. The results indicate that the most suitable configuration of SBCOR under noisy conditions is (1) the filter bank is a fixed Q filter bank whose center frequencies are equally spaced on the Bark scale, and (2) the periodicity extraction method is a conventional autocorrelation analysis without compensation for weak signals. An analysis example of speech is also shown under these conditions.

Secondly, another implementation of SBCOR is introduced based on the above results. Using this implementation, the robustness of SBCOR analysis is investigated in more realistic adverse environments; the existence of additive noise and waveform distortion. First, in DTW word recognition and HMM phoneme recognition, it is shown that SBCOR is more robust against Gaussian white noise than SGDS and mel-filterbank cepstral coefficient (MFCC), and it is comparable against computer room noise and human speech-like noise which is a kind of the bubble noise described in Appendix A. Second, it is shown that SBCOR is robust against severe waveform distortions such as infinite peak clipping. Since the information of an original speech wave is contained in the temporal pattern of the zero-crossings, it is anticipated that SBCOR analysis which extracts periodic information would be robust against zero-crossing distortion. Actually, the experimental results in DTW word recognition showed that the performance of SBCOR in the best case is about 19% higher than that of SGDS.

Thirdly, multi-delay weighting (MDW) processing is proposed to improve the robustness of SBCOR. MDW is used to extract periodicity using not only the autocorrelation coefficient at $f_{cf_i}^{-1}$, but also a weighted sum of autocorrelation coefficients at integral multiples of $f_{cf_i}^{-1}$. At first, it is shown that SBCOR with MDW processing results in the weighting processing of the power spectrum of speech by a lateral inhibitive weighting function. Then, the effectiveness of MDW for speech recognition is investigated under noisy conditions using a DTW word recognizer. The results showed SBCOR with MDW performs better than SGDS and MFCC under noisy conditions. Furthermore, the lateral inhibitive weighting of the power spectrum is specially focused to interpret the robustness of SBCOR, and it is shown that (1) spectral tilt elimination and (2) noise variability elimination would be the essence of lateral inhibitive weighting, and lead to a more robust recognition under noisy conditions.

Finally, as a straightforward extension of SBCOR, subband-crosscorrelation analysis (SBXCOR) using two input channel signals is described. In experiments, the noise robustness of SBXCOR is evaluated using a DTW word recognizer under (1) a simulated acoustic condition with white noise and (2) a real acoustic condition in a sound proof room with human speech-like noise. As the results show, under the simulated acoustic condition, SBXCOR is more robust than conventional one-channel SBCOR, but less robust than SBCOR extracted from the two-channel-summed signal. Furthermore, by applying MDW processing, the performance of SBXCOR improved about 2% at SNR 0dB. The resultant performance of SBXCOR with MDW processing was much better than those of smoothed group delay spectrum (SGDS) and mel-filterbank cepstral coefficient (MFCC) below SNR 10dB. The results under the real acoustic condition were almost the same as the simulated acoustic condition.

Through this dissertation, it is clarified that the periodicity information associated with the inverse of the center frequency included in speech signals plays the significant role in the noise robust acoustic analysis.

Acknowledgements

I would like to thank my advisor, Professor Fumitada Itakura of Nagoya University, for guiding me through this research. On numerous occasions during almost nine years I had joined in his research group, he gave me a lot of valuable suggestions and strict guidances to drive this research. He always taught me what is the original work and what I should do to pursue it. Although the path was sometimes seemed to be tough and difficult for me, I could not conclude this dissertation without his constant encouragements. I am confident that this experience will make my future more fruitful.

I also have a special debt to Professor Noboru Ohnishi who reviewed and commented on the whole parts of the first draft of this dissertation.

Special thanks belong to Associate Professor Kazuya Takeda of Nagoya University for his suggestions and encouragements in the final stage to conclude this dissertation. Also, I specially thank Mr. Daisuke Kobayashi of Matsushita Electric Industrial Co.,Ltd, who was a graduate student at Nagoya University, for his careful experiments in terms of human speech-like noise written in Appendix A.

To perform experiments and write this dissertation, a lot of people had supported me; in special, Dr. Tatsuya Hirahara of ATR International, provided me the configuration parameters of the Cochlear filter, and gave me invaluable suggestions for this research; Mr. Motohiko Yada of the technical staff, assisted me in all aspects; Mr. Mikio Ikeda and Mr. Paul Wheeler for their proof reading of the first draft of this dissertation. I would like to thank for their suggestions and supports.

I appreciate to study together with all of students who had been in our research group.

Specially, Dr. Hani Camille Yehia of ATR Human Information Processing Research Laboratories, who was a foreign student from Brazil at Nagoya University, shared with me all happy and tough moments for almost five years in the Ph.D course; Dr. Hong Wang of PictureTel kindly taught me the basis of digital signal processing as a senior student; Mr. Hideaki Yamada of KDD R&D Laboratories, and Mr. Toshiyuki Nomura of NEC C&C Laboratories spent many valuable hours with me in the Master course.

Finally, I am grateful to all members of my family. In special, I would like to thank my parents, Suguru and Hifuko, and my elder sister, Yuko, for their loves and supports in my life since I was born in 1967. I am supremely happy to have the chance of writing this dissertation. Also, I thank my fiance, Ayako for her support on numerous occasions in my private life. I hope this dissertation will be one of the best present for her in our future life.

Contents

List of Symbols

f_{cf_i}	The center frequency of i th subband.
τ_{cf_i}	The inverse of f_{cf_i}.
$R_n^i(\tau)$	Autocorrelation function of i th subband signal for n th analysis frame.
$R_n(\tau)$	Autocorrelation function of n th analysis frame signal.
$S_n(i)$	i th SBCOR coefficient in n th analysis frame.
$H_i(f)$	Transfer function of i th band pass filter
$X_n(f)$	Power spectrum of speech signal at n th analysis frame.
$P(\cdot)$	Autocorrelation detection function.
c_i	i th channel's controlling term for weak signals.
Q	Q value of subband filters.
f	The linear frequency in Hz.
$B(f)$	The Bark scaled frequency.
$x_n^i(t)$	i th subband signal.
$E\{\cdot\}$	Expectation operator.
p	Order of LPC analysis.
$H_a(z)$	All pole filter in LPC analysis.
a_i	i th LPC coefficient.
$\hat{a}_i$	i th LPC coefficient smoothed by γ.
γ	Smoothing parameter used in smoothed group delay spectrum analysis.
$\hat{S}_n(i)$	i th SBCOR coefficient using MDW in n th analysis frame.
α	MDW weight coefficient.
$W_i(f)$	Weighting function for power spectrum in i th channel.

$\hat{W}_i(f)$	MDW weighting function for power spectrum in i th channel.
$x_n(t)$	n th analysis frame signal of x(t).
$X_{x_n x_n}$	Power spectrum of $x_n(t)$.
$y_n(t)$	n th analysis frame signal of y(t).
$X_{y_n y_n}$	Power spectrum of $y_n(t)$.
$Sc_n(i)$	i th SBXCOR coefficient in n th analysis frame.
$R^i_{x_n x_n}(\tau)$	the autocorrelation function of i th subband signal for $x_n(t)$.
$R^i_{y_n y_n}(\tau)$	the autocorrelation function of i th subband signal for $y_n(t)$.
$R^i_{x_n y_n}(\tau)$	the crosscorrelation function of i th subband signal for $x_n(t)$ and $y_n(t)$.
$\hat{S}c_n(i)$	i th SBXCOR coefficient using MDW in n th analysis frame.
$\hat{W}c_i(f)$	MDW weighting function for cross spectrum in i th channel.

List of Figures

List of Tables

Chapter 1

Prologue

1.1 Introduction

The aim of automatic speech recognition (ASR) is to recognize speech spoken by humans using machines, especially computers. As a man-machine interface, speech is a more natural input method for humans than the keyboard or mouse input used in current computer systems, and the development of ASR system has been continuing for more than four decades. One might think that ASR by computers is not difficult because of ones own ability or experience of recognizing speech. ASR researchers at an early stage also thought that an ASR system would be available in near future. However, the difficulty of ASR was recognized as research moved ahead, and in spite of more than four decades passing, the intelligent ASR system such as the computer HAL in Stanley Kubrick's movie "*2001 - A Space Odyssey*" has not yet been constructed.

In the following sections, the long path to ASR is reviewed briefly, and the current status of ASR is described. In addition, problems in current ASR systems and some efforts to overcome them are also described.

1.2 A Brief History of Research Area for Automatic Speech Recognition [1][2]

Most ASR systems usually consist of (1) an acoustic analysis part as the front-end

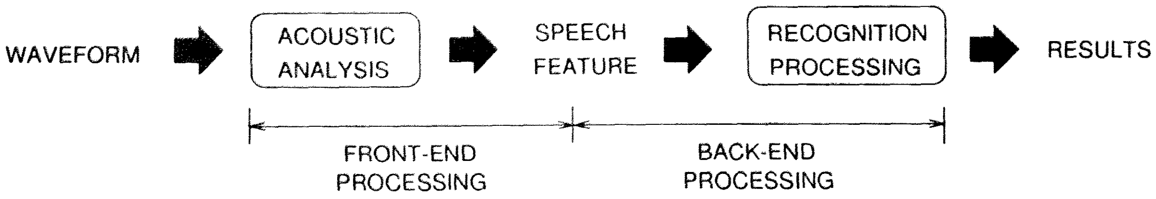


Figure 1.1: Flow-diagram of general ASR system.

and (2) a recognition processing part as the back-end, as shown in Figure 1.1. In the acoustic analysis part, speech features are extracted from waveforms of speech. Using these features, the decision making of what has been spoken is carried out in the recognition processing part. In this section, the history of developing ASR systems is briefly reviewed, focusing on the front-end and back-end processing systems.

The earliest attempts to create ASR systems were made in the 1950s. In 1952, a speaker-dependent isolated digit recognition system was built by Davis *et al* [3]. In this system, spectral resonances in the vowel region of each digit were used as feature parameters. As an independent effort, Olson and Belar tried to recognize 10 distinct syllables of a single talker in 1956 [4]. In the system, spectral measurements provided by an analog filter bank were used. In 1959, Forgie and Forgie constructed a vowel recognizer in which 10 vowels embedded in a /b/-vowel-/t/ format were recognized in a speaker-independent manner [5]. Again a filter bank analyzer was used to provide spectral information.

In the 1960s, several special-purpose hardware machines were built in Japan. One was a hardware vowel recognizer by Suzuki and Nakata using an elaborate filter bank spectrum analyzer [6]. Another was a hardware phoneme recognizer by Sakai and Doshita [7]. In this system, a zero-crossing analysis was used for acoustic analysis method.

In the 1970s, there were two significant break-throughs in the work of speech recognition research. These works affect the field of speech recognition even now. One is the dynamic programming (DP) method as a recognition method by Sakoe and Chiba [8]. The DP method is a technique to align the time axis for a pair of speech utterances, and is also referred as dynamic time warping (DTW). As an independent work, Vintsyuk in the Soviet Union proposed the use of DP method for time aligning in 1968 [9], but it was unknown in the West until the early 1980s. The other significant work is linear predictive coding (LPC) by Itakura, which had been successfully used in low-bit-rate speech coding [10]. At AT&T Bell Laboratories, he built a speech recognition system using LPC analysis for acoustic analysis method and used the DTW technique for recognition processing [11].

In the 1980s, there was a shift in the paradigm of the recognition processing from template-based approaches to statistical modeling methods, i.e. the hidden Markov model approach. Although the introduction of the HMM into speech recognition field was generally attributed to the independent work of Baker [12] and Jelinek *et al* [13], it did not become widely applied until the mid-1980s. Nowadays, most speech recognition systems use the HMM for recognition processing. As another statistical modeling method, the artificial neural network (ANN) was also applied to speech recognition in the late 1980s. Recently, researchers prefer HMM to ANN, since it is considered that HMM includes ANN as a statistical approach. In addition, another reason is that the behavior of HMM is more well-defined using statistical parameters than that of ANN.

In the 1990s, speech recognition in controlled situations has reached very high levels of performance based on the above efforts. For example, vocabulary size is tens of thousands of words and fast decoding algorithms allow continuous-speech recognition systems to provide real-time response. In word error rates, less than 0.1% is obtained in speaker-dependent isolated word recognition using a 20 thousand word vocabulary[16], and about 5% error is obtained in speaker-independent continuous speech recognition of a thousand word vocabulary in the ARPA Resource Management task [17].

1.3 Problems of current ASR systems

In spite of these efforts, however, error rates of machines are often more than an order of magnitude greater than those of a human for quiet, wide band, real speech [15]. Furthermore, the most crucial problem in current ASR systems is that the performance of

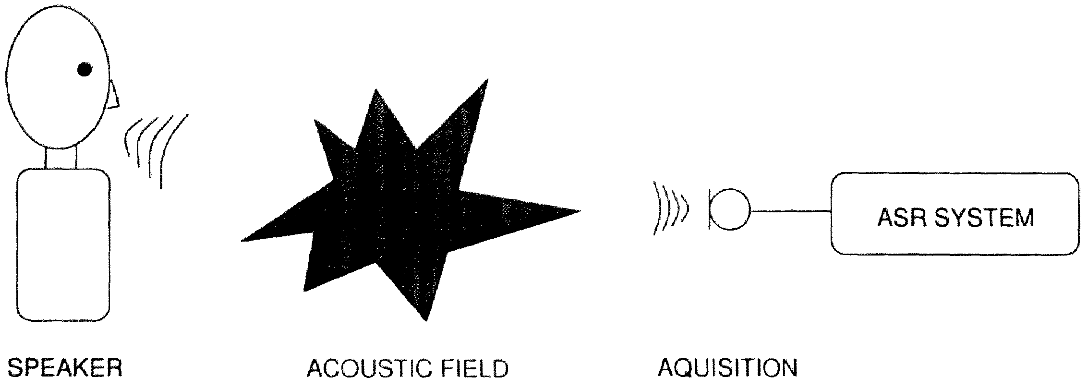


Figure 1.2: Source of variability.

machines degrades to lower than that of humans in real environments. This problem restricts the application of current ASR systems to a few constrained application areas.

The main source of variability that ASR systems are affected can be separated into three parts; (1) the speaker, (2) the acoustic field and (3) the acquisition of speech as shown in Figure 1.2 (See also Table 1.1) [18][19]. A brief review of research efforts is provided in the following section, focusing on the robustness against noise in the acoustic field.

1.4 A Brief Review of Approaches for Noise Robust Speech Recognition [14][15][18][19][20]

Considerable effort has been made to overcome the noise robust problem. Research has been focused primarily on three areas; (1) noise robust feature extraction, (2) speech enhancement and (3) speech model compensation for noise.

In the research field of noise robust feature extraction, a number of authors have developed new approaches based on psychoacoustics and neurophysiology findings in the auditory system. The most auditory-like approach is the modeling of the auditory system proposed by Seneff [21] and Ghitza [22]. These approaches will be precisely described in the following section. There is an auditory inspired approach in which several character-

Table 1.1: Main variability which ASR systems are affected.

Speaker	speaking style, stress, individuality caused by the difference of articulation system such as vocal tract length, vocal folds and so on.
Acoustic field	reverberation, environmental noise.
Acquisition	positioning and direction of microphone, characteristics of microphone, limitation and distortion in electric transmission channel such as the band width, electric noise and echo.

istics of the auditory system are incorporated in conventional acoustic analyses. In the simplest case, the auditory-like frequency axis, such as the Mel scale and the Bark scale, are employed. The most famous feature parameter of this case is Mel-filterbank cepstral coefficient (MFCC) by Davis *et al* [23]. In recent research, there is a report that the noise robustness of MFCC is almost the same as the auditory models proposed by Seneff and Ghitza [24]. A few complicated cases which incorporate the masking phenomenon of the auditory system are RASTA-PLP by Hermansky and Morgan [25] and dynamic cepstral coefficients by Aikawa *et al* [26]. In another approach, smoothed group delay spectrum (SGDS) has been proposed by Itakura and Umezaki [27].

Speech enhancement techniques, which are the second approach to noise robust speech recognition are intended to recover either the waveform or the parameters of noisy speech. Most useful techniques in this area are (1) spectral subtraction [28] and (2) feature mapping from the noisy feature to a clean one [29]. The spectral subtraction technique was originally developed for speech quality improvement rather than recognition.

Thirdly, the speech model compensation technique is mainly used for HMM. For example, HMM decomposition and HMM composition techniques are applied to HMM to model noisy speech [30] [31] [32].

Here, we briefly described several approaches for noise robust speech recognition. Further references and excellent reviews can be found in Gong [14], Acero [20] and Nakagawa in Japanese [18].

1.5 Basic Idea and Objective of This Dissertation

As described in the previous section, one of the possible approaches to robust acoustic analysis is to simulate the human auditory process from a physiological point of view, as human listeners have a good performance in recognizing speech under noisy conditions. This is what we call auditory modeling.

In recent studies of auditory modeling, the synchronous response of the auditory nerve firing and its periodicity extraction have attracted much attention. For example, Seneff has proposed a joint synchrony/mean-rate model, and shown that distinct formant peaks can be extracted by a generalized synchrony detector (GSD) from the output of the auditory nerve firing model [21][33]. This GSD detects the periodicity that is associated with the inverse of the center frequency of the cochlear filter. Some reports in which Seneff's model were applied to speech recognition have shown that the acoustic features extracted by the auditory model are robust against noise [34][35]. Moreover, Ghitza has proposed an ensemble interval histogram (EIH) computational model, and shown that the EIH encodes the relevant phonetic information and outperforms the Fourier power spectrum in the presence of high levels of background noise [22]. This EIH is defined as the ensemble histogram of the inverse of several level-crossing intervals in the auditory nerve firing.

A key to their success seems to be that GSD and EIH capture the extent of the dominance of periodicities in the auditory nerve firing, as noise that has no correlation with speech does not influence the periodicity to any extent. For example, suppose that the periodicity is expressed by the autocorrelation function. If the noise is white, it does not influence the autocorrelation function except for the zero order. Hence, the utilization of periodicities is a key to robustness against noise.

If the periodicity extraction processing is focused on, Seneff and Ghitza's models can be separated into an auditory nerve firing model and a periodicity-extraction model. As for Seneff's model, the former is the critical band filter bank and the hair cell synapse

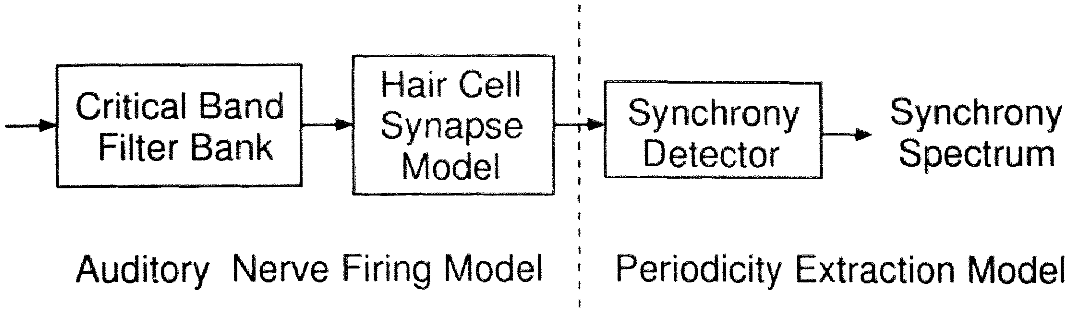

Critical Band Filter Bank
Hair Cell Synapse Model
Synchrony Detector
Synchrony Spectrum
Auditory Nerve Firing Model
Periodicity Extraction Model

Figure 1.3: Seneff's auditory model. It consists of a critical band filter bank, a hair cell synapse model and a synchrony detector.

model, and the latter is the synchrony detector (Figure 1.3). The parameters of the former are adjusted to match existing experimental results of the physiology of the auditory periphery.

If such a precise auditory model is used as the front-end of a speech recognition system, however, would it be the most suitable front-end? Of course, it would be if the post-processor were a physiological model of the human recognition process; however, in the case of using a pattern-matching approach such as Dynamic Time Warping (DTW), or a statistical approach such as Hidden Markov Model (HMM), as the post-processor, it can not be concluded because it is not known whether such a post-processor corresponds to a physiological model of the human recognition process. Such post-processors require intra-category variance-minimization and inter-category variance-maximization for the front end [36]. Since we are interested in developing a practical signal processing model as the front-end of a speech recognition system, we will only focus on its subband processing instead of using a precise auditory nerve firing model.

In this research, considering the standpoints described above, a new signal processing model based on subband processing and periodicity extraction, i.e., subband-autocorrelation (SBCOR) analysis technique is proposed. Although SBCOR is motivated by the auditory model as described above, we will develop it as a signal processing technique.

1.6 Outline of This Dissertation

The following Chapter 2 gives a basic investigation of subband-autocorrelation analysis focused on filterbank analysis and periodicity extraction associated with the inverse of the center frequency f_{cf}^{-1} in a subband, under noisy conditions degraded by multiplicative white noise.

Chapter 3 introduces another implementation of SBCOR based on the results obtained in Chapter 2. Using this implementation, we clarify the robustness of SBCOR analysis in more realistic adverse environments; the existence of three additive noises and waveform distortion.

In Chapter 4, to improve the robustness of SBCOR, multi-delay weighting (MDW) processing is described. MDW is used to extract periodicity using not only the autocorrelation coefficient at f_{cf}^{-1}, but also a weighted sum of autocorrelation coefficients at integral multiples of f_{cf}^{-1}.

As a straightforward extension of SBCOR for multichannel signal processing, Chapter 5 describes subband-crosscorrelation analysis (SBXCOR) using two input channel signals.

Chapter 6 concludes the whole dissertation.

Chapter 2

Basic Investigation of Subband-Autocorrelation Analysis

2.1 Introduction

In this chapter, subband-autocorrelation (SBCOR) analysis technique is investigated for the choice of both the filter bank (subband processing) and the periodicity detection in order to extract acoustic features with robustness against noise for speech recognition.

This chapter is constructed as follows. The following section describes the methodology of the proposed SBCOR analysis technique in detail. Sections 2.3 and 2.4 give the experimental conditions and results respectively. Section 2.5 shows an analysis example of speech using SBCOR and Section 2.6 concludes the whole chapter.

2.2 SBCOR Analysis

2.2.1 Principle of SBCOR analysis

SBCOR analysis is based on filter bank and periodicity detection associated with the inverse of the center frequency. The generalized formula is defined as follows:

$$S_n(i) \quad = \quad P\left[R_n^i(0), R_n^i(\tau_{cf_i}), c_i\right], \ \tau_{cf_i} = f_{cf_i}^{-1}, \qquad (2.1)$$

where,

$S_n(i)$: SBCOR coefficient of ith channel at nth analysis frame

$P(\cdot)$: Periodicity detection function

$R_n^i(\tau)$: Autocorrelation function of ith subband signal

c_i : ith channel's controlling term for weak signals

f_{cf_i} : Center frequency of ith subband signal

$X_n(f)$: Power spectrum of speech signal

Details of the band pass filter $H_i(f)$ and the periodicity detection function $P(\cdot)$ used in this research will be described in Section 2.3.

The flow diagram of SBCOR analysis is shown in Figure 2.1. First, the input speech signal is passed through the filter bank $\{H_i(f)\}$. Then, in the periodicity detector of each channel, the periodicity for the time delay τ_{cf_i} associated with the inverse of the center frequency f_{cf_i}, is calculated. The periodicity detection is performed every analysis frame. The array $\{S_n(i)\}$ of the output of each periodicity detector is interpreted as a "spectrum" and we refer to it as the "SBCOR spectrum".

2.2.2 Difference between Conventional Filter Bank Analysis and SBCOR Analysis

As we defined above, SBCOR analysis is a modified version of filter bank analysis. The principal difference between the conventional filter bank analysis and SBCOR analysis is the kind of feature that is extracted from the output of the filter bank; the conventional filter bank analysis calculates the subband power, i.e. $R_n^i(0)$, whereas SBCOR analysis detects the periodicity of the subband signal, which is associated with the inverse of the center frequency of the subband. For example, if $P(x, y, z) = y/x$, SBCOR calculates $R_n^i(\tau_{cf_i})/R_n^i(0)$, i.e. the subband autocorrelation coefficient at lag τ_{cf_i}. Namely, utilizing not only $R_n^i(0)$ but also $R_n^i(\tau_{cf_i})$ is the peculiarity of SBCOR.

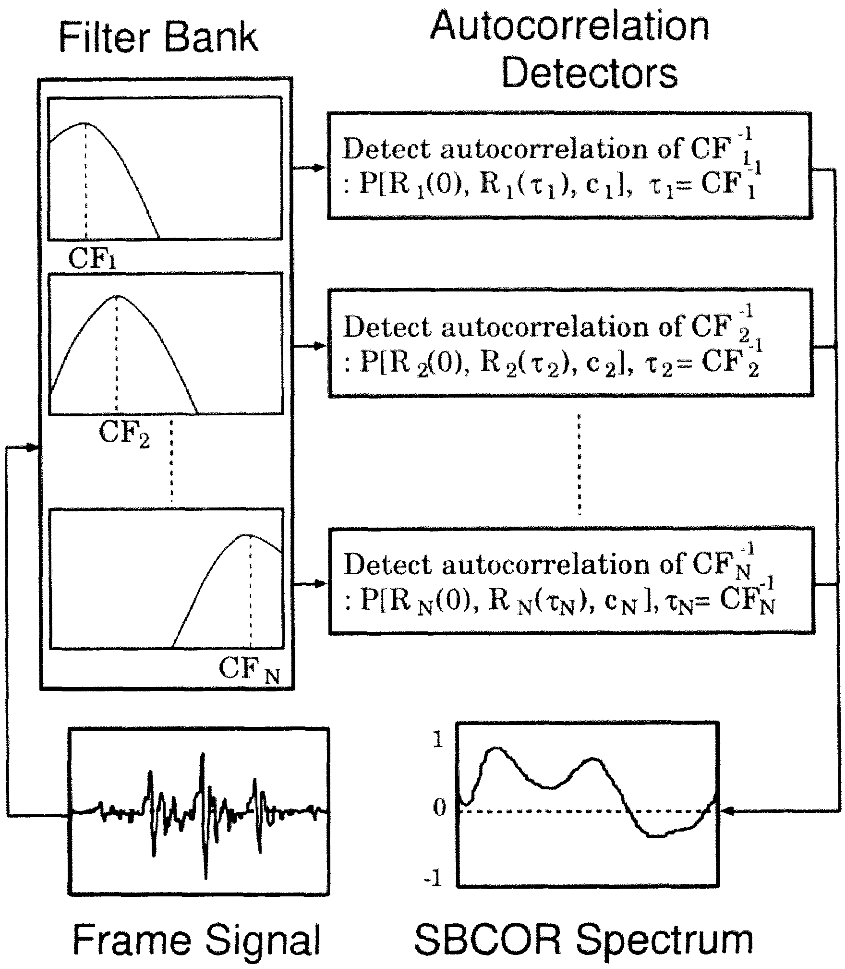

Filter Bank
Autocorrelation Detectors
CF_1
CF_2
CF_N
Detect autocorrelation of CF_1^{-1}
: $P[R_1(0), R_1(\tau_1), c_1]$, $\tau_1 = CF_1^{-1}$
Detect autocorrelation of CF_2^{-1}
: $P[R_2(0), R_2(\tau_2), c_2]$, $\tau_2 = CF_2^{-1}$
Detect autocorrelation of CF_N^{-1}
: $P[R_N(0), R_N(\tau_N), c_N]$, $\tau_N = CF_N^{-1}$
1
0
-1
Frame Signal
SBCOR Spectrum

Figure 2.1: Flow diagram of subband-autocorrelation analysis.

2.2.3 Filter Bank Design and Periodicity Detection

Since we have interests to develop a practical system for the front-end of speech recognizer, we do not always stick to the cochlear filter. Therefore, we will investigate what kind of characteristics of subband filter are appropriate for speech recognition in Sections 2.3 and 2.4. Especially, its investigation will be focused on the band width (or Q value), the spacing of center frequencies (the linear frequency spacing or the Bark scale spacing) and the shape (Cochlear filter shape or not) of each subband filter, which mainly characterize the filter bank.

In the periodicity detector, as we described above, the periodicity associated with the inverse of the center frequency τ_{cf_i} is detected. Here, the word "periodicity" means to what extent a signal is periodic. In this research, the autocorrelation coefficient is used to extract the periodicity. However, since the autocorrelation coefficient is normalized by the power, some compensation for weak signals are necessary to avoid the detection of needless periodicities. In Sections 2.3 and 2.4, the periodicity detection and the compensation techniques are investigated.

2.3 Experimental Conditions

In this and next sections, we demonstrate how to configure the filter bank and the periodicity detector for speech recognition under clean and noisy conditions, using SBCOR analysis as a front-end of a speech recognizer.

2.3.1 Recognizer and Database

A standard Dynamic Time Warping (DTW) speaker-dependent isolated word recognizer is used. The local path constraint is a symmetric one shown in Figure 2.2 and performs DP matching with fixed starting and ending points[8].

The basic database consists of two sets of 550 Japanese city names recorded twice by five Japanese male speakers. The sampling frequency is 10kHz. The first set is used as

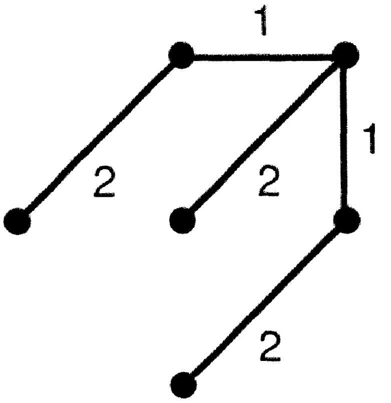


Figure 2.2: Weighed DP path used in the recognition.

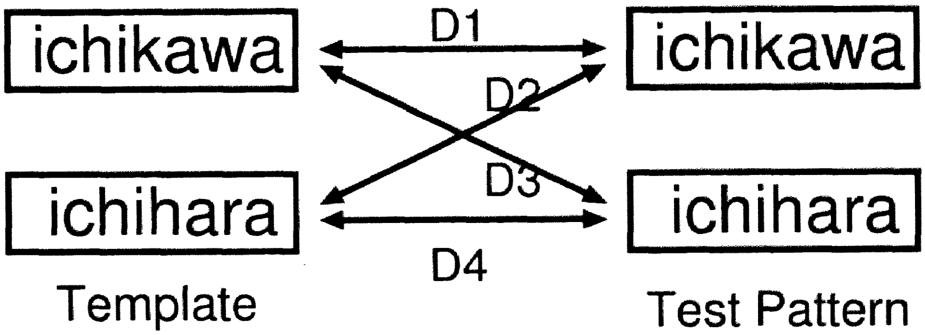


Figure 2.3: An example of word pair recognition (D1, D2, D3 and D4 represent the distance between the reference pattern and the test pattern. When D1<D2 and D4<D3, the words "ichikawa" and "ichihara" are correctly recognized).

Table 2.1: 68 Japanese city name pairs

wako:	ako:	sunagawa	sukagawa	takahama	nagahama
oga	koga	hukawaga	sakagawa	takayama	wakayama
toda	noda	yono	ono	sagae	sabae
toba	tosa	yashio	yachiyo	sanjo:	anjo:
tama	zama	toyosaka	toyonaka	hukui	tsukumi
kaga	saga	mobara	obama	hukushima	hukuchiyama
sakai	kasai	oyama	toyama	utsunomiya	ichinomiya
nara	naha	okaya	okayama	ito	mito
kuji	huji	o:da	onoda	kiryu	chiryu
uji	huji	morioka	tomioka	handa	sanda
kuji	uji	cyo:hu	ko:hu	iwatsuki	iwakuni
kitami	itami	o:tsu	go:tsu	kamogawa	kakogawa
chiba	chita	o:tsu	o:bu	ichikawa	ichihara
mutsu	huttsu	o:muta	o:mura	matsudo	matsuto:
kamaishi	takaishi	to:no	o:no	tsushima	kushima
takahagi	takasaki	gobo:	gojo:	atsugi	yasugi
hamada	yamada	kakuda	katsuta	nagaoka	takaoka
mikasa	mitaka	matsuzaka	matsubara	takaoka	kasaoka
hirakata	hirata	matsuyama	matsubara	enzan	sennan
hirata	hirara	matsuzaka	matsuyama	izumi	izumo
hino	chino	o:me	ko:be	hamamatsu	takamatsu
takikawa	tachikawa	o:date	o:take	ichinomiya	nishinomiya
kashiwara	kashihara	takayama	takahama		

the reference patterns and the second set, which was spoken a week later, is used as the test patterns. Since the following experiments are speaker-dependent word recognition, we will get a very high recognition rate. Therefore, so as to clarify the differences of the performance, we selected 68 pairs of city names with phonetically similar names (Table 1), and performed DP matching between each pair (Figure 2.3)[37, 27]. Each pair is assumed to be easily mistaken in recognition. The resulting recognition rate is averaged for five speakers.

To examine the robustness against noise, white noise is added to the test patterns. The white noise used here is the multiplicative signal-dependent white noise defined as follows:

$$\hat{s}(n) = s(n)(1 + a \cdot r(n)), \quad z = 10 \log_{10} \frac{3}{a^2} \tag{2.2}$$

where $s(n)$ is the clean speech signal, $\hat{s}(n)$ is the noisy speech signal, a is the relative noise amplitude, z is the desired signal to noise ratio (SNR) and $r(n)$ is a uniformly distributed random number between -1 to 1. Since the SNR of the noisy speech signal is constant anywhere, we can demonstrate the quantitative characteristics of the robustness. The SNR is examined for four cases, namely $z=\infty$, 20, 10 and 0dB.

2.3.2 Types of Filter Bank

Five types of filter bank shown in Figure 2.4 are used, which are different in the band widths, the center frequencies and the shape. All of the filter banks consist of 16 band pass filters.

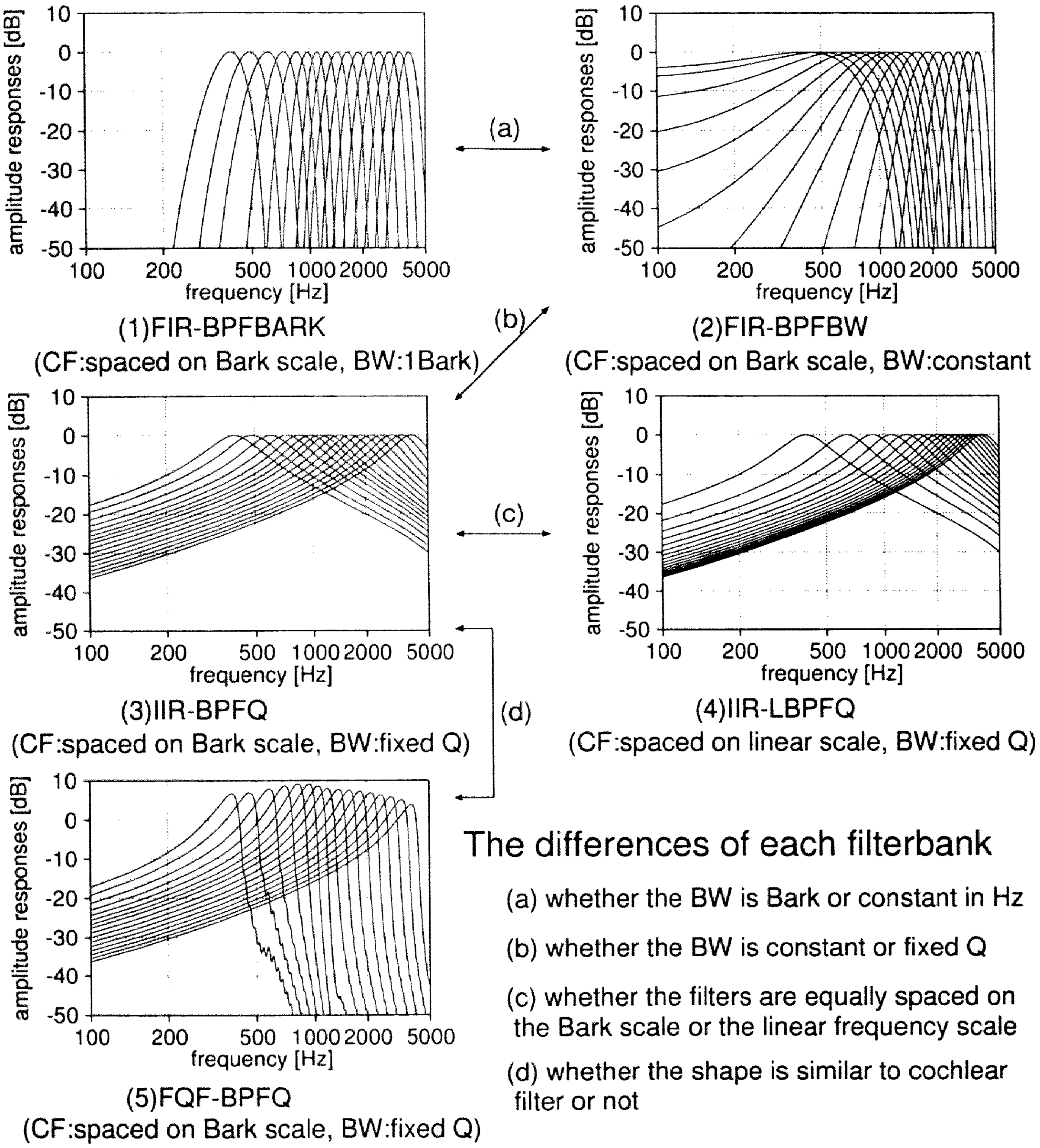
1. **FIR-BPFBARK filter bank** consists of finite impulse response (FIR) band pass filters with band width of 1 Bark. The center frequencies of the filters are equally spaced on the Bark scale between 4 and 17Bark.

2. **FIR-BPFBW filter bank** consists of FIR band pass filters with a constant band width for all filters. The center frequencies of the filters are equally spaced on the Bark scale between 4 and 17Bark. Band widths of 500, 400 and 300Hz are investigated.

3. **IIR-BPFQ filter bank** consists of second order infinite impulse response (IIR) band pass filters with fixed Q. The center frequencies are equally spaced on the Bark scale between 4 and 17Bark. The Q values of 1.0, 1.5, 2.0, 2.5 and 3.0 are investigated.

4. **IIR-LBPFQ filter bank** consists of second order IIR band pass filters with fixed Q. The center frequencies are equally spaced on the linear frequency scale between 400 and 3892Hz. The Q values of 1.0, 1.5, 2.0, 2.5 and 3.0 are investigated.

5. **FQF-BPFQ filter bank** consists of cochlear filters with fixed Q. The center frequencies are equally spaced on the Bark scale between 4 and 17Bark. This cochlear filter is the one without adaptive Q circuits proposed by Hirahara[38]. The Q values of 1.0, 1.5, 2.0, 2.5 and 3.0 are investigated.



(1)FIR-BPFBARK
(CF:spaced on Bark scale, BW:1Bark)

(2)FIR-BPFBW
(CF:spaced on Bark scale, BW:constant

(3)IIR-BPFQ
(CF:spaced on Bark scale, BW:fixed Q)

(4)IIR-LBPFQ
(CF:spaced on linear scale, BW:fixed Q)

(5)FQF-BPFQ
(CF:spaced on Bark scale, BW:fixed Q)

The differences of each filterbank

(a) whether the BW is Bark or constant in Hz

(b) whether the BW is constant or fixed Q

(c) whether the filters are equally spaced on the Bark scale or the linear frequency scale

(d) whether the shape is similar to cochlear filter or not

Figure 2.4: Five types of filter bank investigated.

The conversion to the Bark scale was defined by the following set of equations[21]:

$$B(f) = \begin{cases} 0.01f & 0 \leqq f < 500 \\ 0.007f + 1.5 & 500 \leqq f < 1220 \\ 6.0 \log f - 32.6 & 1220 \leqq f \end{cases},$$

where f is the frequency in Hertz, and B is the frequency in Bark.

2.3.3 Types of Periodicity Detector

We introduce three types of periodicity detector considering the emphasis of the autocorrelation and the compensation for weak signals, and investigate which is the most suitable for speech recognition. In the following equations, $x_n^i(t)$ represents i th subband signal. In calculating $R_n^i(\tau_{cf_i})$, we used an eight point polynomial interpolation technique.

1. **COR method** is defined as a modified autocorrelation coefficient.

$$S_n(i) \;=\; \frac{E\{x_n^i(t)x_n^i(t-\tau_{cf_i})\}}{E\{{x_n^i}^2(t)\} + c_i} = \frac{R_n^i(\tau_{cf_i})}{R_n^i(0) + c_i}. \tag{2.3}$$

   In this case, the periodicity detection function $P(\cdot)$ is

$$P(x, y, z) \;=\; \frac{y}{x+z}. \tag{2.4}$$

2. **MGSD method** is the modified generalized synchrony detector (MGSD). The MGSD is defined by the ratio of the estimated power of a summation waveform to the estimated power of a difference waveform. This MGSD method emphasizes the positive correlation.

$$S_n(i) \;=\; \frac{E\{[x_n^i(t) + x_n^i(t-\tau_{cf_i})]^2\}}{E\{[x_n^i(t) - x_n^i(t-\tau_{cf_i})]^2\} + c_i} \tag{2.5}$$

$$\;=\; \frac{1 + \dfrac{R_n^i(\tau_{cf_i})}{R_n^i(0)}}{1 - \dfrac{R_n^i(\tau_{cf_i})}{R_n^i(0)} + \dfrac{c_i}{2R_n^i(0)}}. \tag{2.6}$$

   In this case, the periodicity detection function $P(\cdot)$ is

$$P(x, y, z) \;=\; \frac{1 + \dfrac{y}{x}}{1 - \dfrac{y}{x} + \dfrac{z}{2x}}. \tag{2.7}$$

3. **FISHER method** is based on the Fisher transformation used in statistics. The FISHER method emphasizes strong correlation.

$$S_n(i) = \frac{1}{2}\log\left(\frac{1+\dfrac{R_n^i(\tau_{cf_i})}{R_n^i(0)}}{1-\dfrac{R_n^i(\tau_{cf_i})}{R_n^i(0)}}\right). \tag{2.8}$$

In this case, the periodicity detection function $P(\cdot)$ is

$$P(x,y,z) = \frac{1}{2}\log\left(\frac{1+\dfrac{y}{x}}{1-\dfrac{y}{x}}\right). \tag{2.9}$$

2.3.4 Evaluation Methods

The experiments are performed in following three steps.

Experiment 1

First, we evaluate what type of filter bank is suitable in speech recognition in the case of using the COR method as the periodicity detector. The controlling term c_i for weak signals is determined by the ratio of c_i to the averaged frame power for each channel.

Experiment 2

Second, we evaluate what type of periodicity detector is suitable in the case of using the IIR-BPFQ filter bank. The term c_i is set to 0 in accordance with the results of the first experiment.

Experiment 3

Finally, we compare the performance of SBCOR spectrum with those of the other two speech-analysis techniques.

In the experiment, firstly, to investigate the effect for detecting the periodicity at f_{cf}^{-1}, the best performance of the SBCOR spectrum is compared with that of the power detection version, which is referred to as SBPOWER, in the case of using the same filter bank of SBCOR analysis. In another word, SBPOWER detects the average power of the subband signal instead of the autocorrelation at $f_{cf_i}^{-1}$.

Second, to show the robustness under noisy conditions, the performance of SBCOR spectrum is compared with that of the smoothed group delay spectrum (SGDS), already shown to be robust[37, 27, 39]. The SGDS is the speech representation based on the group delay characteristic of the speech signal, and defined as the derivative of the smoothed phase spectrum of a p th order all pole filter;

$$\begin{aligned} \hat{a}_i &= \gamma^i a_i \quad 0 < \gamma < 1 \quad \text{for} \quad i = 1, \cdots, p \\ H_{\hat{a}}(z) &= \frac{1}{1 + \sum_{i=1}^{p} \hat{a}_i z^{-i}}, \end{aligned}$$

where,

$H_{\hat{a}}(z)$: transfer function of the smoothed all pole filter

a_i : i th LPC coefficient

$\hat{a}_i$: k th LPC coefficient smoothed by γ

γ : smoothing parameter

In order to compare the performance of SBCOR with that of SGDS under exactly the same conditions, the analysis frequency points of SGDS are chosen to be the same center frequencies as those of SBCOR.

2.4 Experimental Results

2.4.1 Choice of Filter Bank : Experiment 1

The best results under SNR 0dB (heavy noise conditions) are shown in Figure 2.5 and the parameters of the filter bank are shown in Table 2.2 (the conditions shown in Table 2.2 are

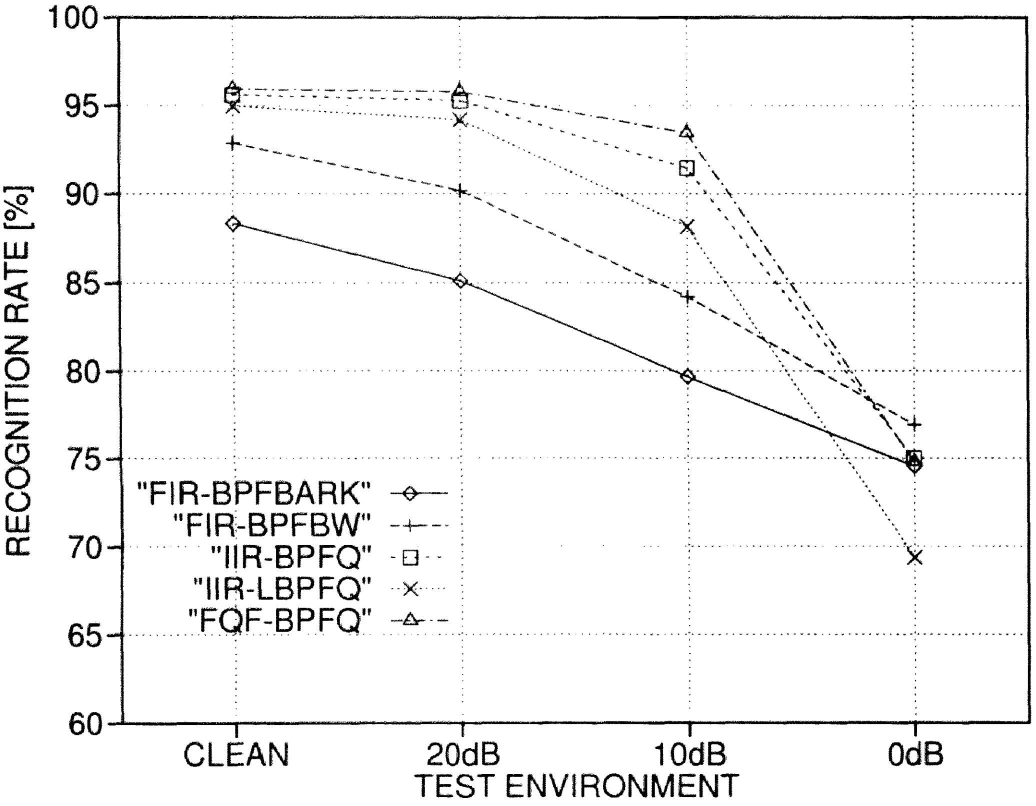


Figure 2.5: Recognition rates (%) for five SBCOR analysis. They use the same periodicity detector (COR method), but differ in the type of filter bank.

Table 2.2: Optimum conditions in Experiment 1.

FILTER BANK				PERIODICITY DETECTION
TYPE	BW	CF	SHAPE	c_k
FIR-BPFBARK	1BARK	BARK	SYMMETRIC	OPTIMIZED
FIR-BPFBW	300HZ	BARK	SYMMETRIC	OPTIMIZED
IIR-BPFQ	Q=1.0	BARK	SYMMETRIC	0.0
IIR-LBPFQ	Q=1.0	LINEAR	SYMMETRIC	0.0
FQF-BPFQ	Q=1.0	BARK	COCHLEAR	OPTIMIZED

(BW: Band Width, CF: Center Frequency)

optimum for SNR 0dB, but not necessarily optimum for other SNRs). The comparison was made as a function of SNR.

These results are summarized into three points. First, the cochlear filter bank (FQF-BPFQ) is better than any other filter bank among all conditions, but the differences of the performance between the FQF-BPFQ and IIR-BPFQ filter banks are not significant. Accordingly, it seems that the shape of the cochlear filter is not crucial in SBCOR analysis. Second, the fixed Q filter banks, such as the FQF-BPFQ and IIR-BPFQ filter banks, give a better recognition rate than the constant-band-width filter banks, such as the FIR-BPFBW and FIR-BPFBARK filter banks. Third, the filter banks whose center frequencies are equally spaced on the Bark scale, such as the IIR-BPFQ and FQF-BPFQ filter banks, are much better than those whose center frequencies are equally spaced on the linear frequency scale, such as the IIR-LBPFQ filter bank.

Thus, it is clarified that the fixed Q filter bank whose center frequencies are equally spaced on the Bark scale should be used in the SBCOR analysis. Whether the filter shape is similar to the cochlear filter or not is not crucial.

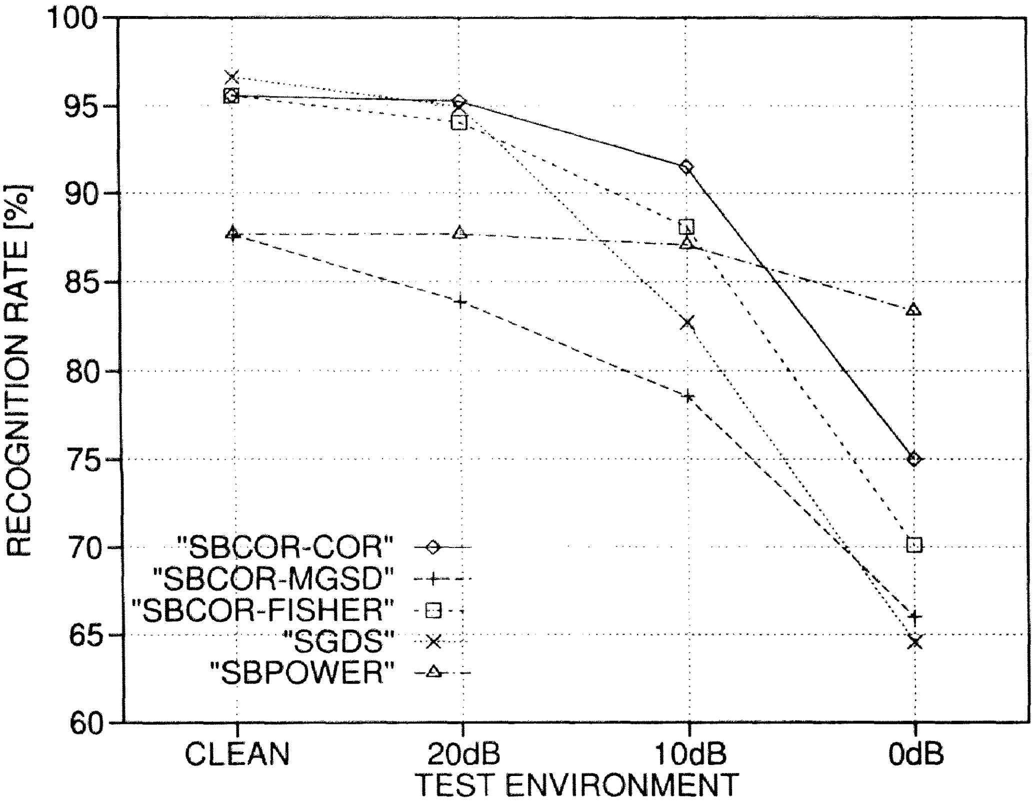


Figure 2.6: Recognition rates (%) obtained by three SBCORs, SBPOWER and SGDS (optimum $\gamma = 0.925$). The SBCORs use the same filter bank (IIR-BPFQ), but differ in their periodicity detectors.

2.4.2 Choice of Periodicity Detector: Experiment 2

The experimental results in the optimum case (Q=1.0) are shown in Figure 2.6. The comparison was made as a function of SNR. The best performance is obtained using the COR method among all conditions. Thus, it is not necessary to emphasize the difference of correlation in SBCOR analysis.

2.4.3 Comparison with SBPOWER and SGDS: Experiment 3

It is shown in Figure 2.6 that the best performance of SBCOR spectrum is comparable to SGDS under clean condition and far superior under noisy conditions. In noisy situations of SNR 0dB, the recognition rate of SBCOR is more than 10% higher than that of SGDS.

As for the comparison with SBPOWER, while the performance of SBPOWER is the worst under a high SNR because of the low frequency resolution (i.e., Q=1.0), it is the best under noisy conditions. This indicates that the patterns smeared by noise are closer to the low frequency resolution patterns by Q=1.0. However, the average performance of SBPOWER under all conditions was 86.5%, while that of SBCOR was 89.3%. Hence, the comparison with SBPOWER indicates that the periodicity detection of f_{cf}^{-1} is much better than the power detection method, when all conditions are considered.

2.5 Analysis Examples

The analysis examples of the SBCOR spectrum, smoothed group delay spectrum (SGDS) and FFT spectrum are presented in Figures 2.7 and 2.8 under clean and noisy conditions. The input speech is the utterance "bakuonga" spoken by a male speaker. The sampling frequency is 10kHz. The noisy speech of SNR 0dB was created by adding the multiplicative signal-dependent white noise to the utterance. The length and shift of the analysis frame are 20 and 5ms respectively. The number of channels of SBCOR spectrum and SGDS is 128 between 4 and 17Bark. FFT spectrum was calculated by a 256 point FFT. The other analysis conditions are shown in Table 2.3.

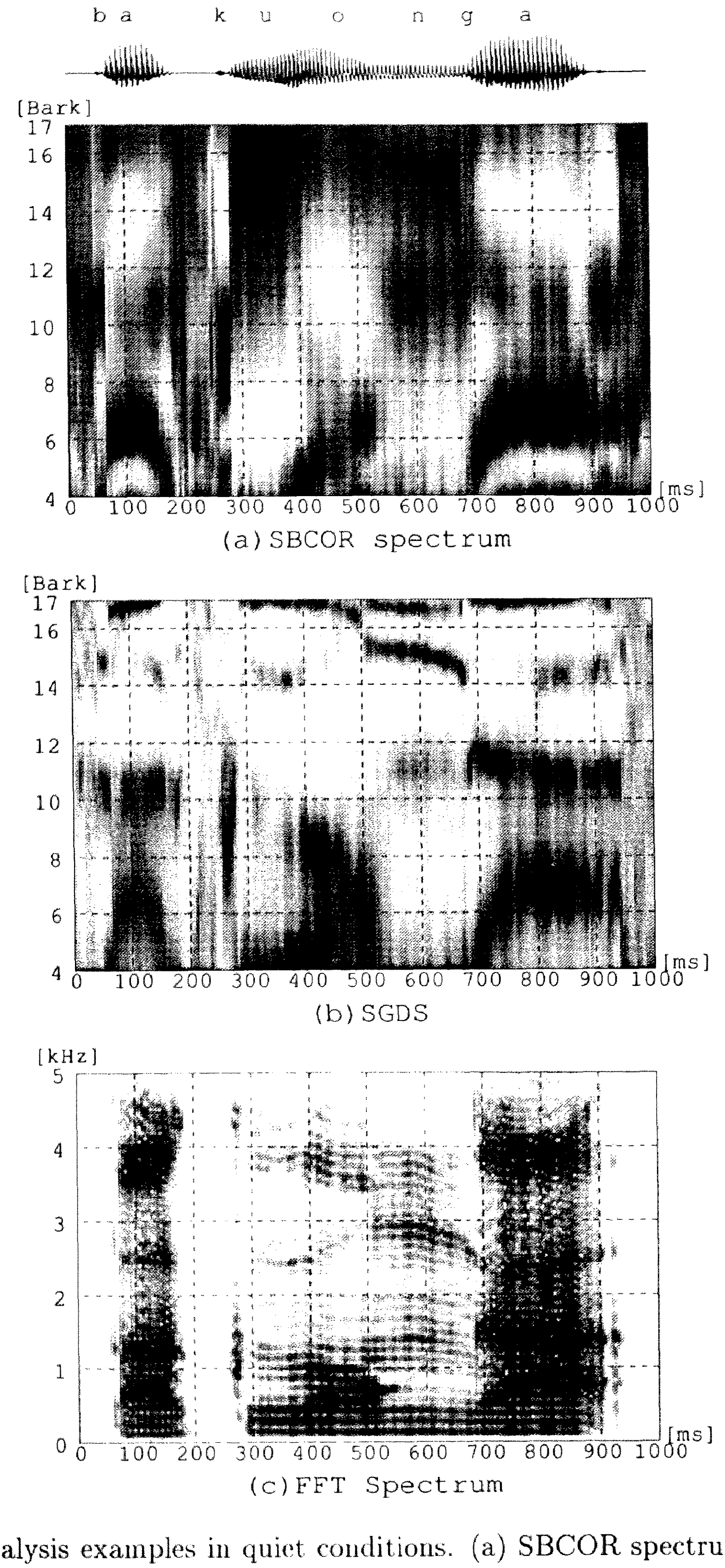
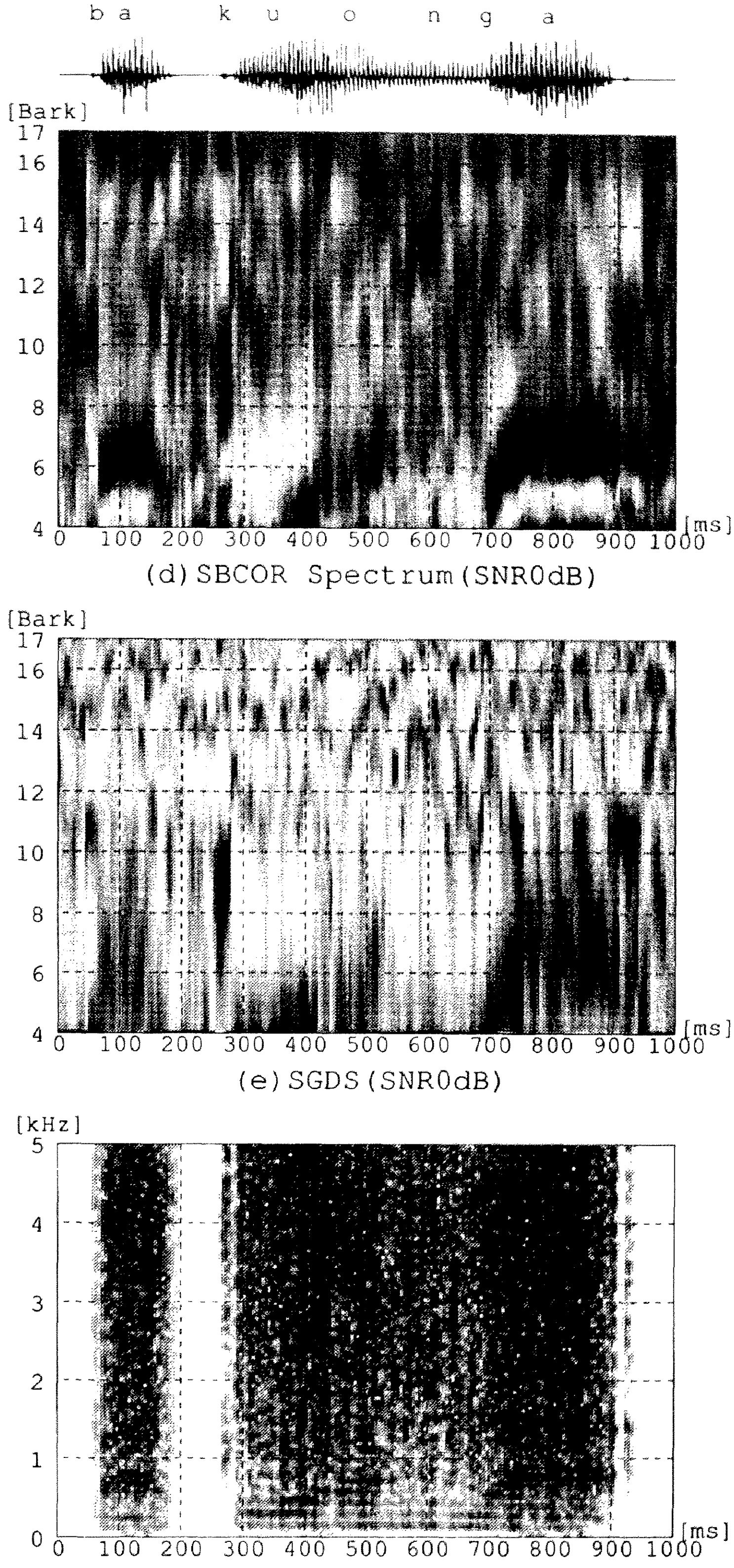


Figure 2.7: Analysis examples in quiet conditions. (a) SBCOR spectrum, (b) smoothed group delay spectrum (SGDS), (c) FFT spectrum. SBCOR spectrum and SGDS are presented with gray levels between maximum (black) and minimum (white) values in whole. FFT spectrum is presented with 50dB dynamic range.



(d)SBCOR Spectrum(SNR0dB)

(e)SGDS(SNR0dB)

(f)FFT Spectrum(SNR0dB)

Figure 2.8: Analysis examples in SNR 0dB conditions. (a)SBCOR spectrum, (b) smoothed group delay spectrum (SGDS), (c) FFT spectrum.

Table 2.3: Analysis conditions for SBCOR, SGDS and FFT spectrum.

METHOD	CONDITIONS
SBCOR SPECTRUM	filter bank: IIR-BPFQ (Q=1.0); periodicity detector: COR method (c_k=0).
SGDS	LPC analysis order: 10; γ : 0.925; analysis window: hamming.
FFT SPECTRUM	analysis window: hamming; input signal: pre-emphasized.

It can be seen that SBCOR spectrum extracts important speech properties like spectral bars related to the first formant seen also in SGDS and FFT spectrum, and shows the sharpness of both onset and offset for different speech segments under clean conditions (Figure 2.7). As distinct from SGDS and FFT spectra, however, the SBCOR spectrum does not necessarily extract higher formants. Under noisy conditions, the SBCOR spectrum preserves more information up to about 10 Bark than the other spectra (Figure 2.8).

2.6 Conclusions

In this chapter, we proposed subband-autocorrelation analysis as a new signal analysis technique to be used as the front-end of speech recognition, and investigated the choice of filter bank and periodicity detector under clean and noisy conditions.

Our experimental results using a speaker-dependent DTW isolated word recognizer showed that the most suitable filter bank and periodicity detection method are a fixed Q filter bank whose center frequencies are equally spaced on the Bark scale, and a conventional autocorrelation detection, without controlling weak signals, respectively. Whether the filter shape is similar to the cochlear filter or not is not crucial. SBCOR spectrum extracted under these configurations performs equally as well as smoothed group delay spectrum under clean conditions, and much better under noisy conditions.

Although we could demonstrate experimentally that SBCOR analysis is robust against the multiplicative white noise, we should also give the reason based on the analysis characteristics. Furthermore, we should investigate the performance of SBCOR for more realistic noises, and the performance of SBCOR in the case of using HMM as the back-end. These points will be investigated in the following Chapters.

Chapter 3

Robustness of SBCOR Analysis against Additive Noises and Waveform Distortion

3.1 Introduction

To address the problem of the noise robustness in speech recognizer, we proposed subband-autocorrelation (SBCOR) analysis in Chapter 2. It is a new signal processing technique based on filter bank analysis followed by autocorrelation calculation to extract periodicities included in speech signal. In Chapter 2, various implementations of SBCOR were compared under noisy conditions affected by the multiplicative signal-dependent white noise. The experimental results showed about 10% better results than smoothed group delay spectrum (SGDS) in the best case under signal-to-noise ratio (SNR) 0dB.

In this chapter, we introduce another implementation of SBCOR based on Wiener-Khinchin Theorem. Using the implementation, we clarify the robustness of SBCOR analysis in more realistic adverse environments; the existence of three additive noises and a waveform distortion. Furthermore, the effect of back-end pattern recognition method is discussed by comparing the recognition performances of a Dynamic Time Warping (DTW) word recognizer and a Hidden Markov Model (HMM) phoneme recognizer.

This chapter is organized as follows. The following section describes the new implementation of the proposed SBCOR analysis technique in detail. Section 3.3 investigates the robustness against three additive noises; the Gaussian white noise, the human speech-like noise and a computer room noise in the context of DTW word recognition and HMM phoneme recognition. In Section 3.4, the robustness against zero-crossing distortion is investigated, as an example of severe waveform distortion. Section 3.5 summarizes and concludes this chapter.

3.2 Redefinition of SBCOR Analysis

3.2.1 Method

SBCOR analysis is on the basis of filter bank and autocorrelation analysis, and aims to extract periodicities associated with the inverse of the center frequency $f_{cf_i}^{-1}$ in a subband, as described in Chapter 2. The principal difference between the conventional filter bank analysis and SBCOR is what kind of feature is extracted from the output of the filter bank; the conventional filter bank analysis detects the power, whereas SBCOR detects the periodicity of $f_{cf_i}^{-1}$.

As shown in the previous chapter, the most suitable filter bank and periodicity detection method are a fixed Q filter bank whose center frequencies are equally spaced on the Bark scale, and a conventional autocorrelation detection, without controlling weak signals, respectively. Whether the filter shape is similar to the cochlear filter or not is not crucial. Based on these results, SBCOR used in this and the following chapters is re-defined by

$$S_n(i) = \frac{R_n^i(\tau_{cf_i})}{R_n^i(0)}, \ \tau_{cf_i} = f_{cf_i}^{-1} \tag{3.1}$$

$$R_n^i(\tau) = \int_{-\infty}^{\infty} | H_i(f) |^2 X_n(f) \cos 2\pi f \tau df. \tag{3.2}$$

The subband filterbank $\{H_i(f)\}$ consists of fixed Q Gaussian band pass filters (BPFs) whose center frequencies are equally spaced on the Bark scale. Each BPF is defined by

$$| H_i(f) |^2 = \begin{cases} e^{-2C_i(f-f_{cf_i})^2}, \ f \geq 0 \\ | H_i(-f) |^2, \ f < 0, \end{cases} \tag{3.3}$$

where,

$$C_i = \frac{2Q^2 \ln 2}{f_{cf_i}^2}. \tag{3.4}$$

Figure 3.1 shows the flow-diagram of SBCOR implemented by the fast Fourier transform (FFT) for the subband filtering and the calculation of autocorrelation coefficients based on Equations (3.1) and (3.2). For n th analysis frame, the ith subband filtering is performed by multiplying the power spectrum $X_n(f)$ by the characteristics of the subband filter $\mid H_i(f) \mid^2$. The subband autocorrelation coefficients are calculated by the inverse FFT of the subband power spectrum $\mid H_i(f) \mid^2 X_n(f)$. Then, the autocorrelation coefficient $S_n(i)$ at the lag τ_{cf_i}, which is associated with the inverse of the center frequency $f_{cf_i}^{-1}$, is picked up. The filter bank consists of 128 BPFs for analysis examples (Figures 3.14 and 3.15) and 16 BPFs for recognition experiments in Section 3.3 and 3.4 (See Figure 3.2).

Note that the implementation of subband filtering and autocorrelation calculation using FFT is computationally more efficient than the one in the time domain, as in Chapter 2.

3.2.2 Interpretation of SBCOR Analysis in Frequency Domain

From the Wiener-Khinchin theorem, SBCOR can be interpreted as the weighted sum for power spectrum $X_n(f)$ in the frequency domain. In the case, the weighting function $W_i(f)$ is

$$W_i(f) = \mid H_i(f) \mid^2 \cos 2\pi f \tau_{cf_i}. \tag{3.5}$$

As shown in Figure 3.3, the $W_i(f)$ shows that SBCOR analysis has the lateral inhibitive effect centered at f_{cf} for power spectrum of speech. Since the emphasis of spectral contrast by such lateral inhibition improves the robustness against noise[40], it indicates that the robustness of the SBCOR is due to this lateral inhibition effect.

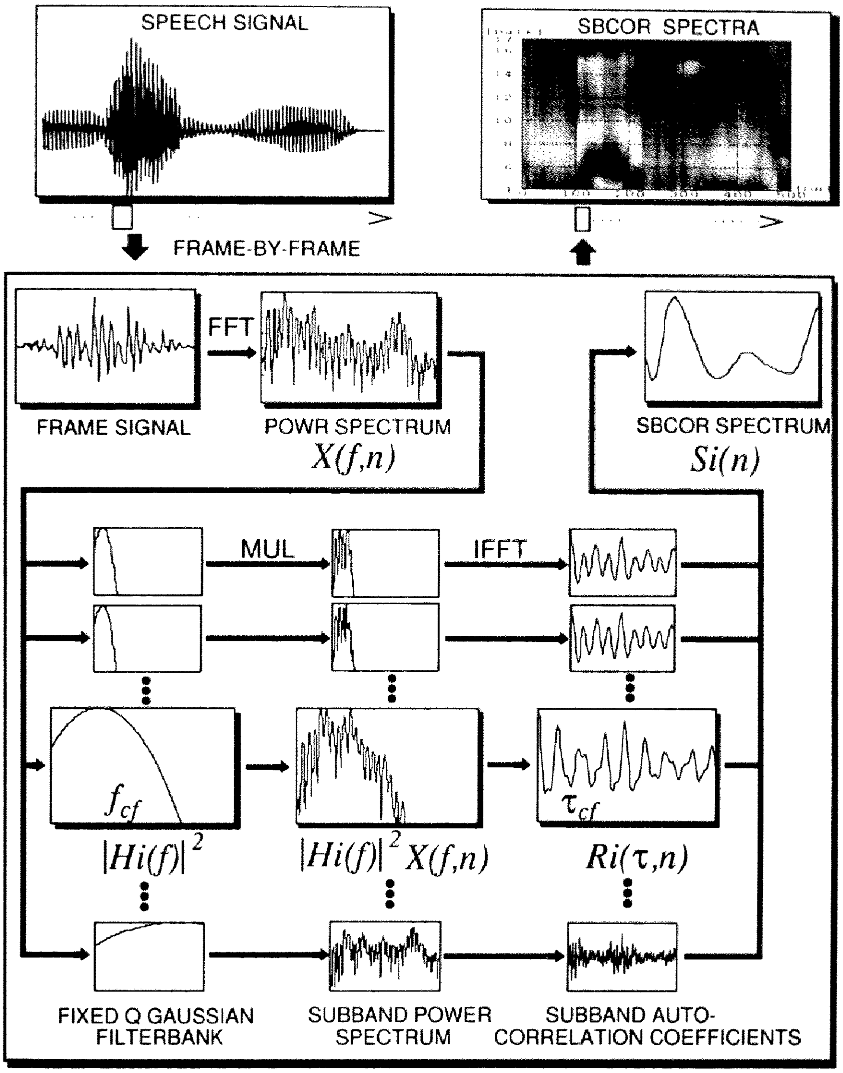


Figure 3.1: Flow diagram of SBCOR implemented by FFT.

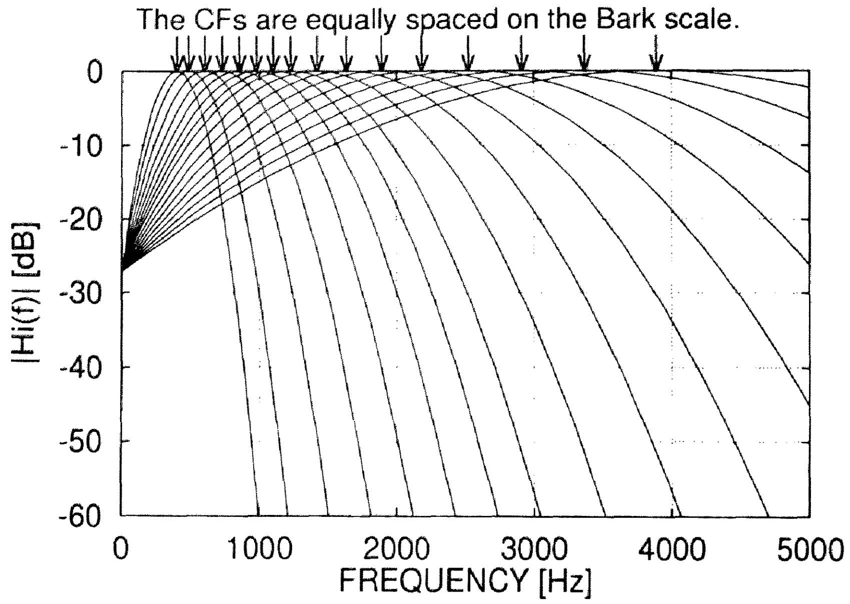


Figure 3.2: Fixed Q Gaussian filter bank whose center frequencies are equally spaced on the Bark scale.

3.3 Evaluation for Additive Noises

In this section, we will investigate the robustness against three different additive noises. The investigation is performed using both a DTW word recognizer and a HMM phoneme recognizer in order to examine the effect of back-end pattern recognition method.

3.3.1 Experimental Conditions

Three Additive Noises

The Gaussian white noise, human speech-like noise and a computer room noise are generated or recorded by the following ways:

- *Gaussian White Noise*

  The white noise was generated using a Gaussian random-number generator on computer.

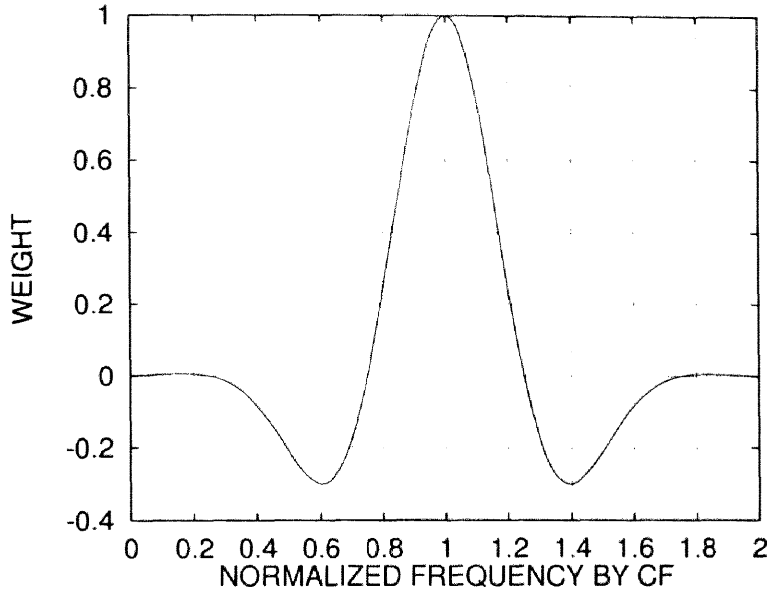


Figure 3.3: Weighting function for power spectrum. The horizontal axis is normalized by the center frequency of BPF.

- *Human Speech-like Noise (HSL noise)*

  The HSL noise is a kind of bubble noise, which simulates when a lot of people are talking simultaneously, as in the cocktail party. The HSL noise was generated by overlapping cyclically a long speech signal to a fixed-length buffer. The long speech signal was prepared by concatenating 3,200 phrases of the ASJ continuous speech corpus. The signal is about 4.6 hours long. The fixed-length buffer is three seconds long. Each phrase was normalized by the maximum absolute amplitude in the phrase. The power spectrum density has a peak at about 250Hz, and the slope in the higher frequency range is about 10 dB/Oct (Figure 3.4). It indicates that the SNRs are almost the same in any subband.

  The speech feature included in HSL noise will be investigated subjectively and objectively in Appendix A.

- *Computer Room Noise*

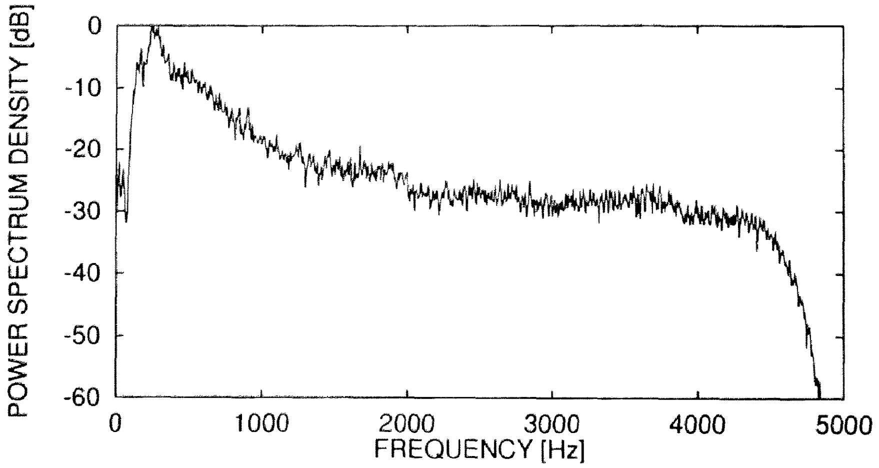


Figure 3.4: The power spectrum density of the human speech-like noise estimated by Welch's averaged periodogram method.

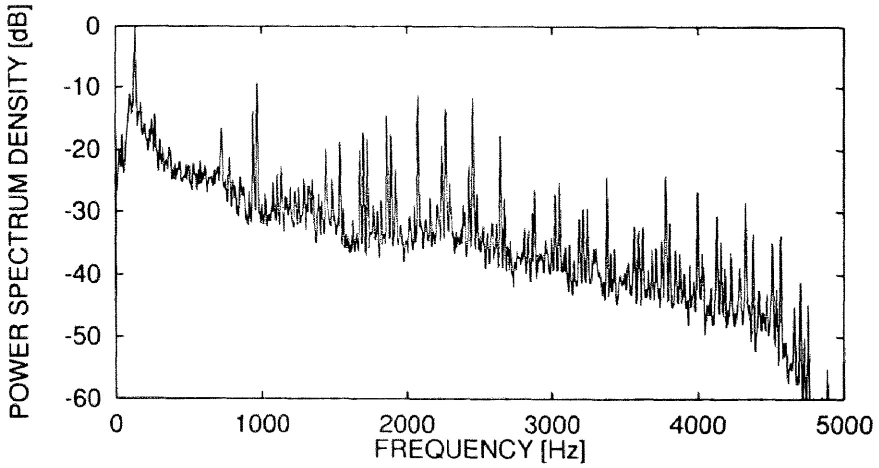


Figure 3.5: The estimated power spectrum density of a computer room noise.

The noise in a computer room was recorded using a microphone, as an example of realistic environmental noises. The power spectrum has several sharp peaks (Figure 3.5).

Back-End Recognition Methods

The robustness of SBCOR against these types of noise are evaluated by the following speaker-dependent recognition experiments using DTW and HMM.

- *DTW word recognition*

  The same standard DTW speaker-dependent isolated word recognizer used in Chapter 2 is used. The recognition task is a 68 word-pair discrimination. Each pair is a phonetically similar city name pair, selected from a 550 Japanese city name database recorded twice by five Japanese male speakers. The sampling rate is 10 kHz. The first set is used as the reference pattern and the second set, which was spoken a week later, is used as the test pattern. The recognition rate is given by the average for the five speakers.

- *HMM phoneme recognition*

  The task is the speaker-dependent 23 phoneme recognition. Each phone model has 3 states of 7 component mixture Gaussian. The parameter estimation was performed using the 2620 even-numbered words from the ATR Japanese 5240 word speech database(two male and two female speakers). The speech data for tests were collected from the odd-numbered 2620 words. The sampling rate is 10 kHz. The recognition rate is given by the average for the 4 speakers.

SBCOR and Reference Feature Parameters

Under the above experimental conditions, the robustness of SBCOR is investigated. Moreover, in order to show the effectiveness under noisy conditions, the performance of SBCOR is compared with those of the smoothed group delay spectrum (SGDS)[41] and mel filter bank cepstral coefficient (MFCC)[23].

- *SBCOR*

  The value of Q is investigated for 0.5, 1.0, 1.5, 2.0, 2.5, 3.0. The center frequency of BPFs are equally spaced on the Bark scale between 4 and 17 Bark.

- *Smoothed Group Delay Spectrum (SGDS)*

  SGDS is calculated as in Section 2.3. In order to compare the performance of SBCOR with that of SGDS under exactly the same conditions, the analysis frequency points of SGDS were chosen to be the same as the center frequencies of SBCOR. In the experiments, the value of γ is investigated for 0.85, 0.875, 0.9, 0.925, 0.95, 0.975, 1.0.

- *Mel filter bank cepstral coefficient (MFCC)*

  MFCC is commonly used speech feature in speech recognizer[23]. In recent research, the noise robustness of MFCC is almost the same as the auditory models proposed by Seneff and Ghitza[24]. In this experiment, MFCC is calculated using the triangular shape mel filter bank. The number of channels is investigated for 20, 24, 28, 32.

In extracting all feature parameters, the analysis frame length and shift are 20ms and 10ms, respectively. The dimension of each feature is 16. The global signal-to-noise ratios (SNRs) used in testing are 20, 10, 5 and 0dB.

3.3.2 Recognition Results

The best performances of SBCOR, SGDS and MFCC are shown in Figures 3.6-3.11 for the DTW word recognition and the HMM phoneme recognition. In comparing the performances among SBCOR, SGDS and MFCC, the analysis parameters, i.e., Q in SBCOR, γ in SGDS and the number of channels in MFCC, that attained the best results under all conditions are selected (Table 3.1).

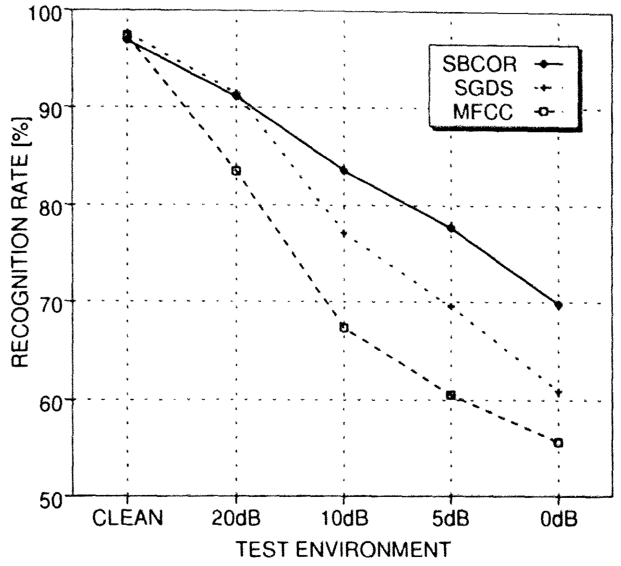


Figure 3.6: DTW word recognition rates for SBCOR, SGDS and MFCC under Gaussian white noise.

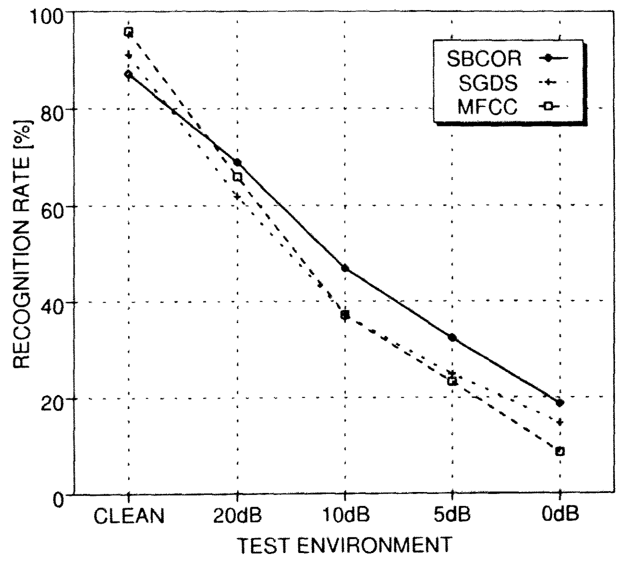


Figure 3.7: HMM phoneme recognition rates for SBCOR, SGDS and MFCC under Gaussian white noise.

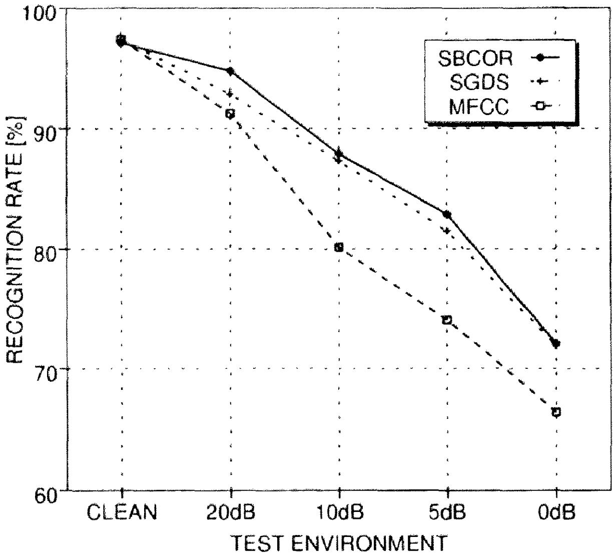


Figure 3.8: DTW word recognition rates for SBCOR, SGDS and MFCC under human speech-like noise.

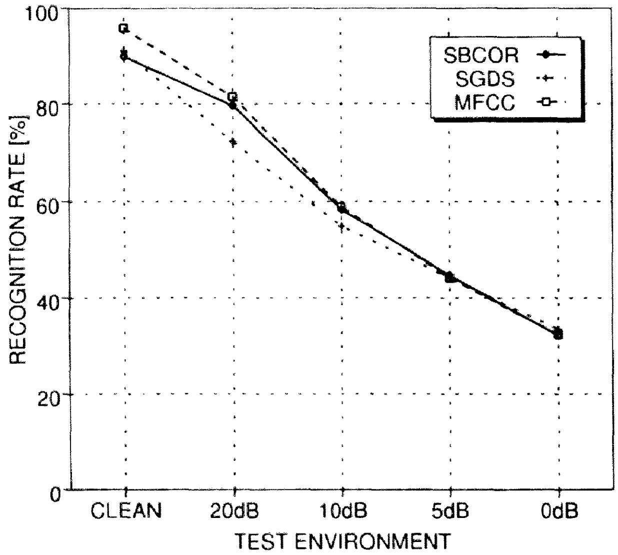


Figure 3.9: HMM phoneme recognition rates for SBCOR, SGDS and MFCC under human speech-like noise.

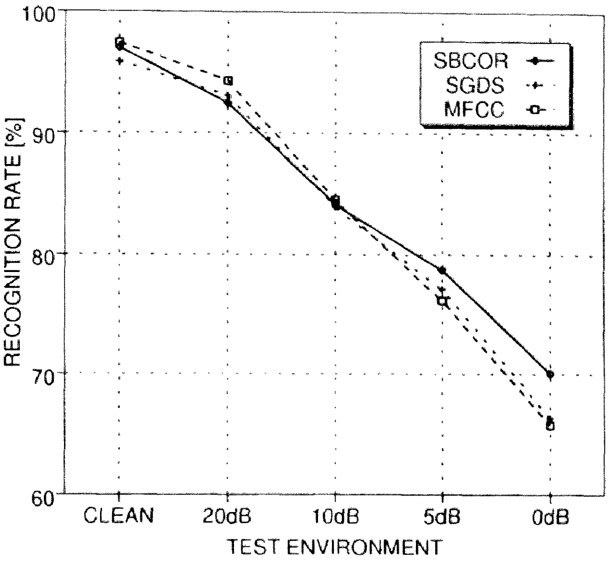


Figure 3.10: DTW word recognition rates for SBCOR, SGDS and MFCC under computer room noise.

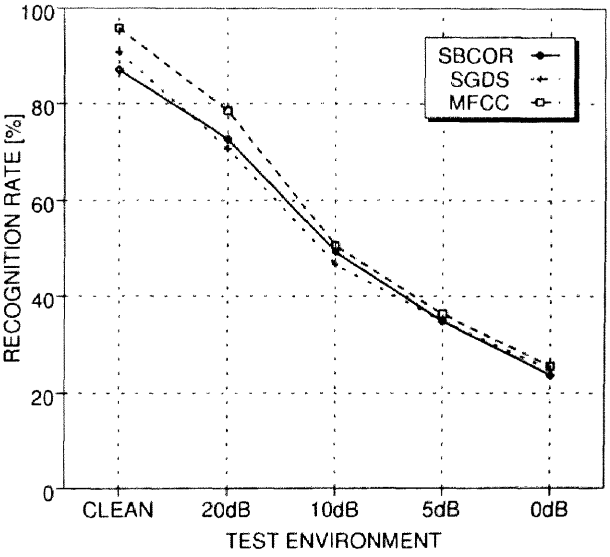


Figure 3.11: HMM phoneme recognition rates for SBCOR, SGDS and MFCC under computer room noise.

Table 3.1: The best parameters in extracting SBCOR, SGDS and MFCC for (a)DTW word recognition and (b)HMM phoneme recognition.

(a)DTW word recognition

TYPE OF NOISE	Q(SBCOR)	γ(SGDS)	# of channel(MFCC)
GAUSSIAN WHITE NOISE	1.5	0.925	28
HSL NOISE	2.0	0.95	28
COMPUTER ROOM NOISE	1.5	0.95	28

(b)HMM phoneme recognition

TYPE OF NOISE	Q(SBCOR)	γ(SGDS)	# of channel(MFCC)
GAUSSIAN WHITE NOISE	1.5	0.925	32
HSL NOISE	2.0	0.925	32
COMPUTER ROOM NOISE	1.5	0.875	32

DTW Word Recognition

For the Gaussian white noise, SBCOR performs much better than SGDS and MFCC under noisy conditions. At the SNR 10dB or lower, the recognition rates are improved about 7 to 9% for SGDS and about 14 to 16% for MFCC. For the HSL noise, since the differences of the performance between SBCOR and SGDS are about 1 to 2%, the performacen of SBCOR under HSL noise is comparable to that of SGDS. On the other hand, SBCOR outperforms MFCC considerably. For the computer room noise, although SBCOR outperforms SGDS and MFCC under SNR 0dB, the performances of them are comparable on the average.

From these results, we can summarize that the robustness of SBCOR is considerably better than those of SGDS and MFCC under the Gaussian white noise and that of MFCC under the HSL noise, while it is almost comparable to those of SGDS and MFCC under the computer room noise and SGDS under the HSL noise.

HMM Phoneme Recognition

For the Gaussian white noise, the performances of SBCOR are about 4 to 8% better than those of SGDS and MFCC. For the cases of the HSL noise and and the computer room noise, the differences of the performances among SBCOR, SGDS and MFCC are about 2% or less below SNR 10dB.

3.3.3 Short Summary

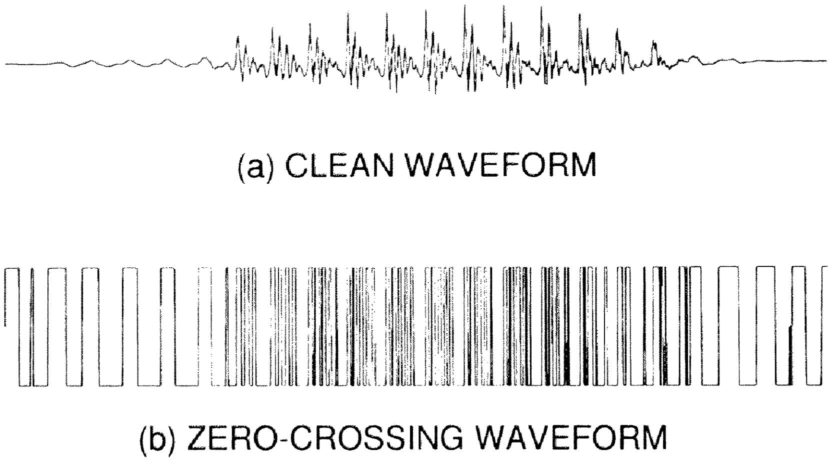
In summary of this section, we clarified that SBCOR is much more robust than SGDS and MFCC under the Gaussian white noise, but SBCOR does not always outperform SGDS and MFCC significantly under the HSL noise and the computer room noise. These indicate that SBCOR does not perform well in the real acoustic environment where noises do not have flat spectrum.

3.4 Evaluation for Waveform Distortion

It is known well that human performance remains high with unnatural degradations caused by waveform clipping, band-reject filtering, and analog waveform scrambling[42], while machine performance degrades significantly. In this section, we investigate the robustness of SBCOR against zero-crossing distortion as an example of such unnatural degradations.

3.4.1 Zero-crossing Distortion

Zero-crossing distortion is a hard limiting distortion that is caused, for example, when a signal is clipped due to the excessive input power. It is often called infinite peak clipping distortion[43]. In the zero-crossing distorted signal, the amplitude information of the original signal is completely lost, and only the zero-crossing points are preserved, as shown in Figure 3.12.



(a) CLEAN WAVEFORM

(b) ZERO-CROSSING WAVEFORM

Figure 3.12: Waveform of a clean speech and the zero-crossing wave.

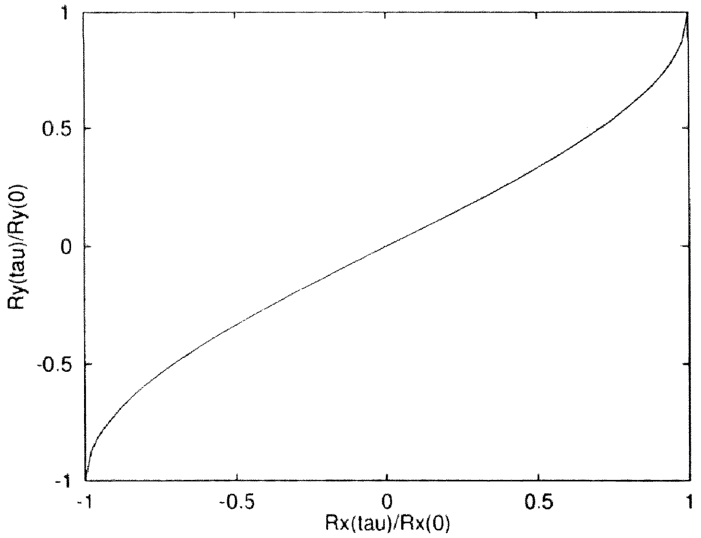


Figure 3.13: Relationship between $R_x(\tau)/R_x(0)$ and $R_y(\tau)/R_y(0)$, derived from the arcsin law.

In this paper, we define the zero-crossing distorted signal as follows:

$$y(t) = \begin{cases} a \times \mathrm{sgn}(x(t)) & \mid x(t) \mid > 0 \\ 0 & x(t) = 0, \end{cases} \tag{3.6}$$

where $x(t)$ and $y(t)$ are a signal and the zero-crossing distorted signal, respectively. The gain parameter a is determined so that the power of the input signal is preserved.

Such zero-crossing signals are still intelligible for human[43], but the performance with conventional speech recognizers deteriorates significantly, as shown later in recognition experiments. The reason can be seen in the spectrographic representations of a speech and the zero-crossing signals using FFT, SGDS in Figures 3.14 and 3.15. It can be readily seen that the zero-crossing distortion affects the formant structure significantly in FFT and SGDS spectrograms. On the other hand, The SBCOR spectrogram is stable for such a distortion as shown in Figure 3.15. The reason can be expressed in the arcsine law.

3.4.2 Influence of Zero-Crossing Distortion for SBCOR Coefficients

It is known that if $x(t)$ is a normal stationary process with zero mean, the autocorrelation of the zero-crossing distorted signal $y(t)$ equals

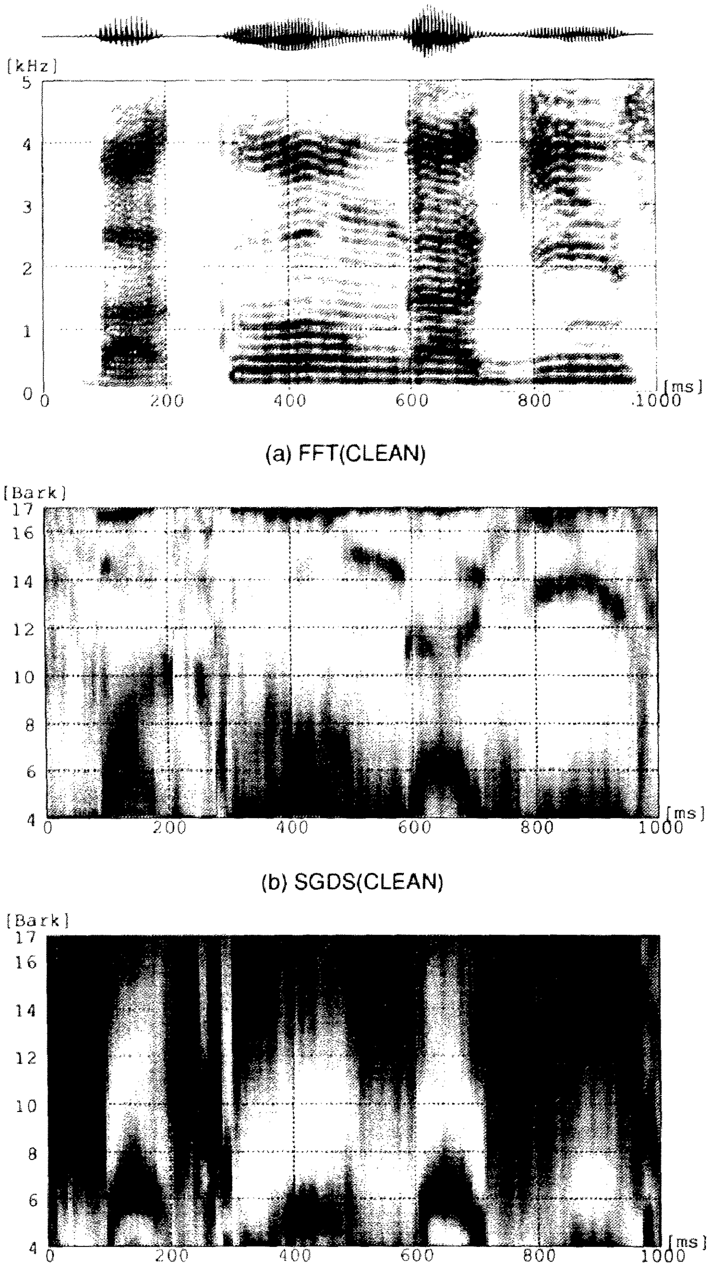
$$\frac{R_y(\tau)}{R_y(0)} = \frac{2}{\pi} \arcsin \frac{R_x(\tau)}{R_x(0)}, \tag{3.7}$$

where $R_x(\tau)$ is the autocorrelation of $x(t)$. This equation is known as the arcsine law[44].

The relationship between $R_x(\tau)/R_x(0)$ and $R_y(\tau)/R_y(0)$ is almost linear except for the regions of strong positive and negative correlations(See Figure 3.13). This indicates that the autocorrelation of the original signal would be preserved in that of the zero-crossing distorted signal. Hence, SBCOR is expected to be robust against the zero-crossing distortion.

3.4.3 Recognition Experiment

The same standard DTW speaker-dependent isolated word recognizer in the previous section is used here. The test signals are distorted by Equation (3.6). The feature extrac-

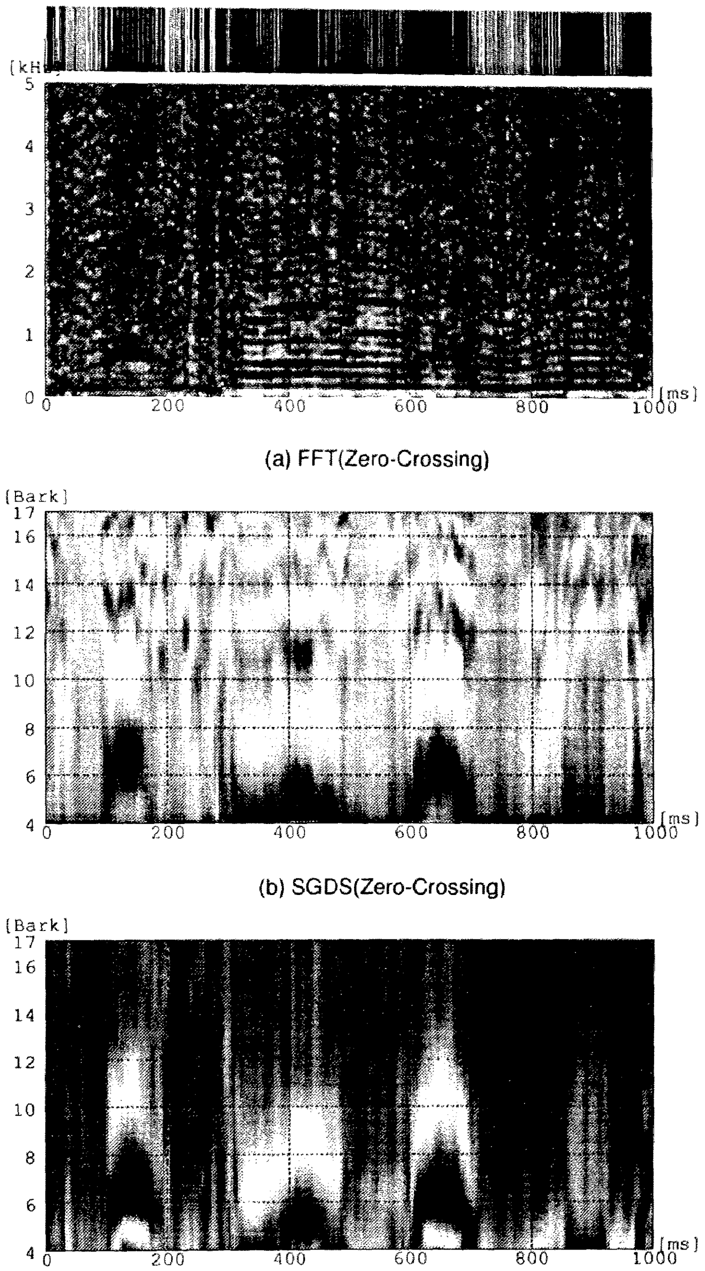


(a) FFT(CLEAN)

(b) SGDS(CLEAN)

(c) SBCOR(CLEAN)

Figure 3.14: Analysis examples of FFT spectrum, SGDS and SBCOR for clean signals. The utterance is "bakuonga" in Japanese, spoken by a female speaker. The length and shift of the analysis window are 32ms and 4ms respectively.



(a) FFT(Zero-Crossing)

(b) SGDS(Zero-Crossing)

(c) SBCOR(Zero-Crossing)

Figure 3.15: Analysis examples of FFT spectrum, SGDS and SBCOR for zero-crossing signals.

Table 3.2: Average recognition rate for zero-crossing signals.

FEATURE	CLEAN	ZERO-CROSSING
SBCOR(Q1.0)	95.6%	87.8%
SBCOR(Q1.5)	96.8%	77.6%
SGDS	97.2%	68.5%
MFCC	97.1%	65.7%

tion of SBCOR, SGDS and MFCC were performed under the same analysis conditions in section 3.

The recognition results are shown in Table 3.2. Although the performance of SGDS and MFCC for zero-crossing signals deteriorates significantly, the performance of SBCOR(Q=1.0) is about 19% and 22% higher than those of SGDS and MFCC respectively. These results indicate that the SBCOR spectrum is much more robust against the zero-crossing distortion than SGDS and MFCC.

3.5 Conclusions

In this chapter, we evaluated the robustness of SBCOR analysis technique against three types of additive noise and the zero-crossing distortion using DTW and HMM recognizers. The results can be summarized as the following four points.

1. The robustness of SBCOR against the noise whose spectrum is flat like the Gaussian white noise, is quite better than those of SGDS and MFCC.

2. The robustness of SBCOR against the human speech like noise and a computer room noise is comparable to that of SGDS. It indicates that the noise whose spectrum is not flat, which is the usual case in the real acoustic environment, affects SBCOR coefficients.

3. Significant differences between the results of DTW and HMM cannot be found.

4. SBCOR is much robust against severe waveform distortion such as the zero-crossing distortion than those of SGDS and MFCC.

These results indicate that the periodicity included in speech signal is one of effective features to improve the robustness of the front-end in adverse conditions.

Chapter 4

Multi-Delay Weighting for Improving SBCOR Performance

4.1 Introduction

In Chapters 2 and 3, we introduced subband-autocorrelation analysis (SBCOR) to extract periodicity present in a speech signal, and evaluated its robustness against noise.

In this chapter, to improve the robustness of SBCOR, multi-delay weighting (MDW) processing is proposed. MDW is used to extract periodicity using not only the autocorrelation coefficient at $f_{cf_i}^{-1}$, but also a weighted sum of autocorrelation coefficients at integral multiples of $f_{cf_i}^{-1}$. At first, we derive that SBCOR with MDW results in the lateral inhibitive weighting (LIW) processing of speech power spectrum. Then, the effectiveness for speech recognition is investigated under noisy conditions using a DTW word recognizer. The noise signals are (1) Gaussian white noise, (2) human speech like noise and (3) computer room noise. The results are compared with those of mel-filterbank cepstral coefficient (MFCC) and smoothed group delay spectrum (SGDS) under noisy conditions. Furthermore, the effectiveness of LIW is investigated by applying no LIW which does not have the negative weight of LIW, and an interpretation of LIW is discussed from the results.

This chapter is constructed as follows. The following Section 4.2 describes MDW processing in detail. Section 4.3 investigates robustness against (1) additive Gaussian noise, (2) human speech like-noise and (3) computer room noise. Section 4.4 clarifies the effectiveness of LIW. Section 4.5 concludes the chapter.

4.2 Taking Autocorrelation Coefficients at Integral Multiples of 1/CF into Account

4.2.1 Multi-Delay Weighting Processing

If a signal is periodic with period T, the autocorrelation coefficients show several peaks at the integral multiples of T. In the conventional SBCOR analysis defined by Equation (3.1), however, only one autocorrelation coefficient at T is used to extract the periodicity included in the subband signal. Therefore, we extend the SBCOR to capture the other peaks of the autocorrelation coefficients by taking a weighted sum of them with the power of α(i.e. the exponential weight) as follows:

$$\hat{S}_n(i) = \frac{\sum_{k=0}^{\infty} \alpha^k R_n^i((k+1)\tau_{cf_i})/R_n^i(0)}{\sum_{k=0}^{\infty} \alpha^k} \tag{4.1}$$

$$0 \leq \alpha < 1 \tag{4.2}$$

We refer to it as Multi-Delay Weighting(MDW) processing.

4.2.2 The Effect of MDW

When $v_n^i(\tau)$ is defined as

$$v_n^i(\tau) = \sum_{k=0}^{\infty} \alpha^k R_n^i(\tau + (k+1)\tau_{cf_i}), \tag{4.3}$$

Equation (4.1) can be calculated by

$$\hat{S}_n(i) = v_n^i(\tau)\Big|_{\tau=0} / \{A R_n^i(0)\},$$

where $A = \sum_{k=0}^{\infty} \alpha^k$. Thus, MDW processing can be seen as a linear filter whose input and output are $R_n^i(\tau)$ and $v_n^i(\tau)$ respectively. Changing the range of the summation,

$$v_n^i(\tau) = \frac{1}{\alpha} \sum_{k=1}^{\infty} \alpha^k R_n^i(\tau + k\tau_{cf_i})$$

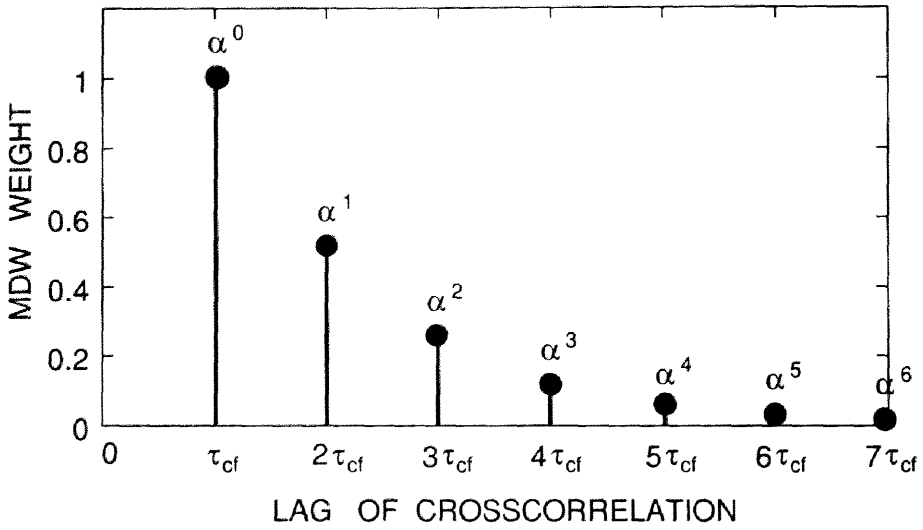


Figure 4.1: Exponentially weighted sum of autocorrelation coefficients associated with the integral multiples of the inverse of center frequencies ($\alpha = 0.5$).

$$= \frac{1}{\alpha}\left\{\sum_{k=0}^{\infty}\alpha^k R_n^i(\tau + k\tau_{cf_i}) - R_n^i(\tau)\right\}. \tag{4.4}$$

Let the Fourier transform of $R_n^i(\tau)$ and $v_n^i(\tau)$ be $X_n^i(f)$ and $V_n^i(f)$, then

$$\begin{aligned} V_n^i(f) &= \frac{1}{\alpha}\left\{\sum_{k=0}^{\infty}\alpha^k X_n^i(f)e^{j2\pi\tau_{cf_i}kf} - X_n^i(f)\right\} \\ &= \frac{X_n^i(f)}{\alpha}\left\{\sum_{k=0}^{\infty}\alpha^k e^{j2\pi\tau_{cf_i}kf} - 1\right\} \\ &= \frac{X_n^i(f)}{\alpha}\left\{\frac{1}{1-\alpha e^{j2\pi\tau_{cf_i}f}} - 1\right\} \\ &= \frac{e^{j2\pi\tau_{cf_i}f}}{1-\alpha e^{j2\pi\tau_{cf_i}f}}X_n^i(f). \end{aligned} \tag{4.5}$$

From the even property of $R_n^i(\tau)$, $V_n^i(f)$ have real components only. Hence,

$$V_n^i(f) = \mathrm{Re}\left\{\frac{e^{j2\pi\tau_{cf_i}f}}{1-\alpha e^{j2\pi\tau_{cf_i}f}}\right\}X_n^i(f) = \frac{\cos 2\pi\tau_{cf_i}f - \alpha}{1-2\alpha\cos 2\pi\tau_{cf_i}f + \alpha^2}X_n^i(f). \tag{4.6}$$

Using the inverse Fourier transform and setting $\tau = 0$, we have

$$v_n^i(\tau)\Big|_{\tau=0} = \int_{-\infty}^{\infty}\frac{\cos 2\pi\tau_{cf_i}f - \alpha}{1-2\alpha\cos 2\pi\tau_{cf_i}f + \alpha^2}X_n^i(f)df. \tag{4.7}$$

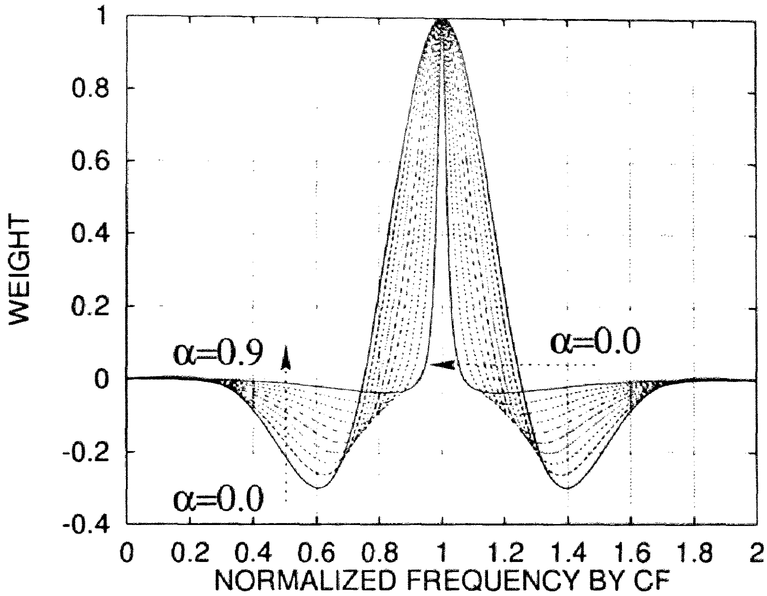


Figure 4.2: Weighting function $W_i(f)$ of the power spectrum of analysis frame signal. Q=1.5. The horizontal axis is normalized by the center frequency (CF).

Where

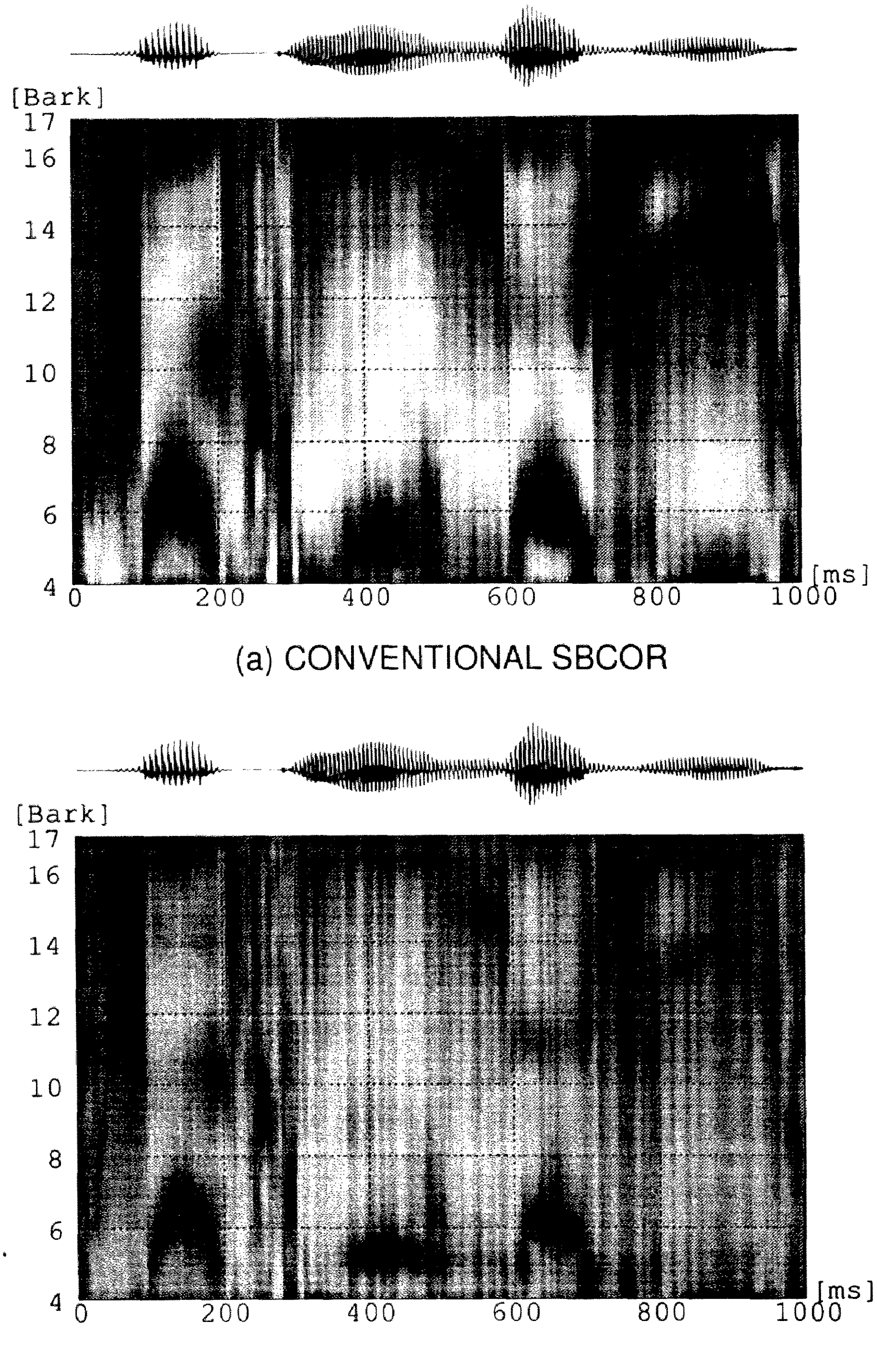
$$
\begin{aligned}
X_n^i(f) &= \mid H_i(f) \mid^2 X_n(f) && (4.8)\\
A &= \sum_{k=0}^{\infty} \alpha^k = 1/(1-\alpha), && (4.9)
\end{aligned}
$$

Equation (4.1) can be expressed in the frequency domain as follows:

$$
\begin{aligned}
\hat{S}_n(i) &= \int_{-\infty}^{\infty} W_i(f) X_n(f) df / R_n^i(0) && (4.10)\\
\hat{W}_i(f) &= \frac{(1-\alpha)(\cos 2\pi\tau_{cf_i} f - \alpha)}{1 - 2\alpha\cos 2\pi\tau_{cf_i} f + \alpha^2} \mid H_i(f) \mid^2 . && (4.11)
\end{aligned}
$$

Thus, SBCOR analysis with MDW processing results in the weighting processing of power spectrum $X_n(f)$ by the weighting function $\hat{W}_i(f)$.

Figure 4.2 shows $\hat{W}_i(f)$ normalized by center frequency. As shown in the figure, both of (1) frequency resolution and (2) spectral contrast enhancement by the lateral inhibition are controllable by MDW processing. The frequency resolution becomes high , and the spectral contrast enhancement becomes weak, as α tends closer to 1.0. (Note that $\hat{W}_i(f)$



(a) CONVENTIONAL SBCOR

(b) SBCOR WITH MDW

Figure 4.3: Examples of speech analysis using (a) the conventional SBCOR and (b) SBCOR with MDW processing ($\alpha = 0.5$). The speech is "bakuonga" uttered by a Japanese male speaker.

is equal to $W_i(f)$ derived in Section 3.2 when $\alpha = 0.0$.) The resulting SBCOR pattern with MDW processing is shown in Figure 4.3. Compared with the conventional SBCOR, it can be seen that the features extracted by SBCOR with MDW are enhanced along with the frequency scale. These effects for speech recognition performance will be investigated experimentally in the following section.

4.3 Evaluation Using DTW Word Recognition

In this section, the robustness of SBCOR with MDW processing is evaluated using DTW word recognition. Then, it will be clarified what weighting for the speech power spectrum is effective for robust speech recognition.

4.3.1 Experimental Conditions

Recognition Task

The same standard DTW speaker-dependent isolated word recognizer in Chapter 2 is used. The recognition task is a 68 pair discrimination. Each pair is a phonetically similar city name pair, selected from a 550 Japanese city name database recorded twice by 5 Japanese male speakers. The first set is used as the reference pattern and the second set, which was spoken a week later, is used as the test pattern.

Noise

Three noise types are used as in Chapter 3; (1) Gaussian white noise, (2) human speech-like noise, and (3) computer room noise.

Comparative Speech Features

The performance of SBCOR with MDW is compared with those of the smoothed group delay spectrum (SGDS)[41] and mel filter bank cepstral coefficient (MFCC)[23] (See Chapter 3 for more information). The analysis conditions are shown in Table 4.1.

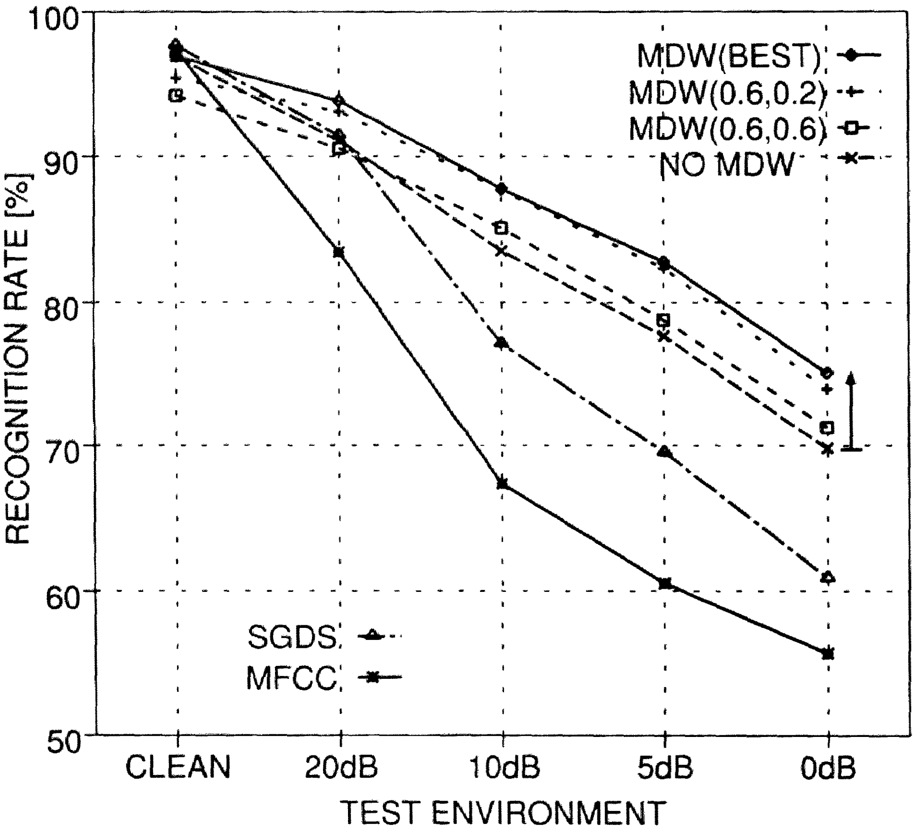


Figure 4.4: Recognition results for the Gaussian white noise. The best Q was 1.5.

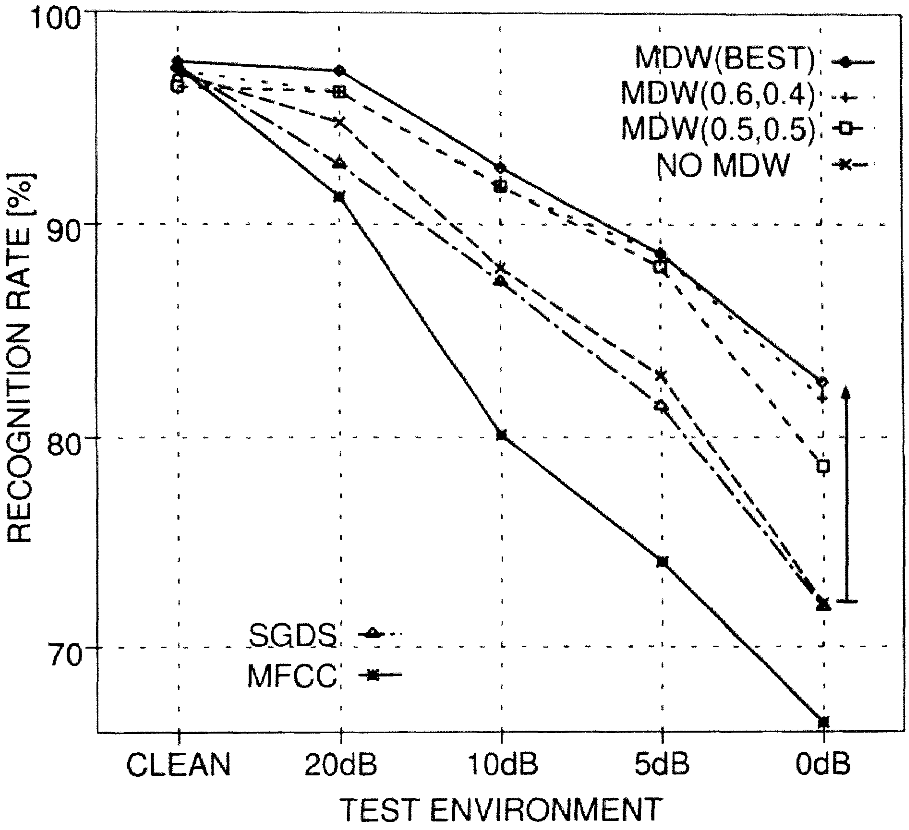


Figure 4.5: Recognition results for the human speech-like noise. The best Q was 2.0.

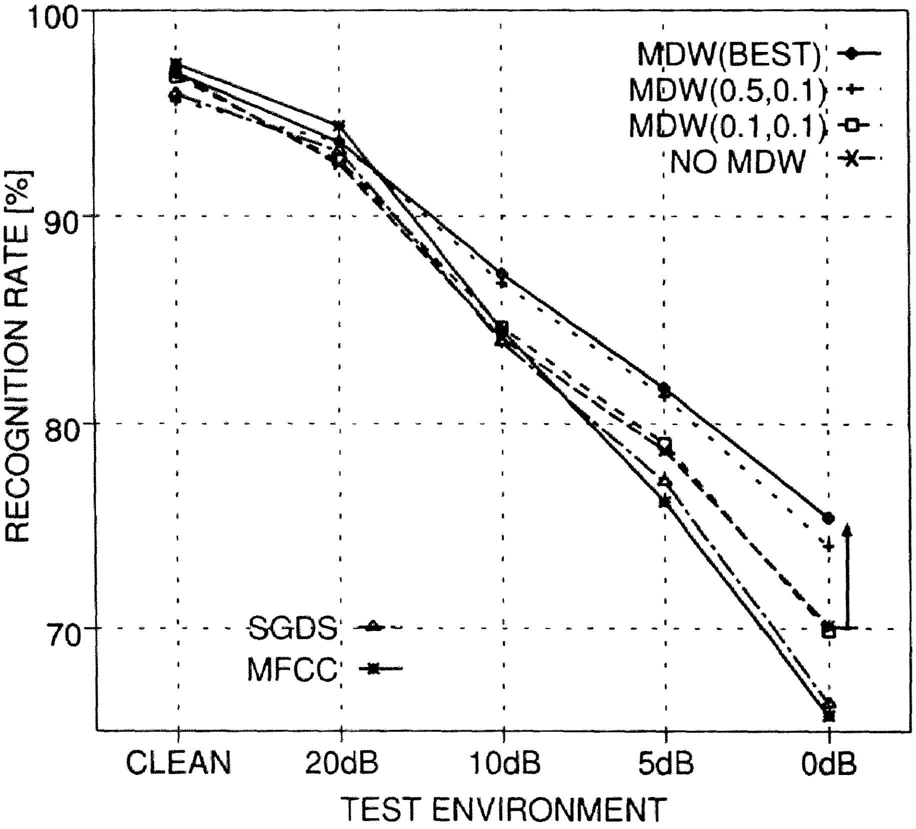


Figure 4.6: Recognition results for the computer room noise. The best Q was 1.5.

Table 4.1: Analysis conditions for SBCOR with MDW, SGDS, and MFCC.

SBCOR	The Q value is investigated for 0.5,1.0,1.5,2.0,2.5,3.0, and the center frequencies are equally spaced on the Bark scale between 4Bark and 17Bark. The α value in MDW processing is 0.0, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. In MDW processing, the summation range for autocorrelation coefficients in Equation 4.1 is from zero to eight.
SGDS	The analysis frequency points of the SGDS are chosen to be the same center frequencies of SBCOR.
MFCC	The filter bank consists of 28 triangle BPF whose center frequencies are equally spaced on the Mel scale.
COMMON	The analysis frame length and shift length is 20ms and 10ms respectively. The dimension of all features is 16.

4.3.2 Experimental Results

The recognition results are shown in Figures 4.4-4.6. The figures are summarized from the following four view points:

1. The case of the best combination of α (α_r, α_t) in extracting reference and test patterns under each SNR condition so that the best performance is attained (BEST)

2. The best case of fixed combination (α_r, α_t) under all test conditions, but it is not necessary $\alpha_r = \alpha_t$.

3. The best case of fixed combination (α_r, α_t) under all test conditions, when $\alpha_r = \alpha_t$.

4. The case of no MDW processing, i.e., $\alpha_r = \alpha_t$=0.

The optimum combinations (α_r, α_t) for the best case (BEST) are shown in Table 4.2.

Table 4.2: The optimum combination of α. The α_r and α_t stand for the α in extracting reference and test patterns respectively.

NOISE		CLEAN	20dB	10dB	5dB	0dB
WHITE NOISE	α_r	0.0	0.6	0.5	0.7	0.8
	α_t	0.0	0.3	0.1	0.3	0.4
HSL NOISE	α_r	0.3	0.4	0.6	0.6	0.7
	α_t	0.2	0.2	0.3	0.4	0.4
COMPUTER ROOM NOISE	α_r	0.0	0.5	0.3	0.5	0.8
	α_t	0.0	0.1	0.0	0.01	0.3

The effect of MDW processing.

For all types of noise, the recognition performances under noise-conditions are improved by using MDW processing. The degree of improvement is different for each noisy condition, the best one is the case adjusted α in each SNR. In that case, about 5% for Gaussian white noise and HSL noise, and about 10% for computer room noise is higher than the conventional SBCOR at SNR 0dB (the improvements are shown as arrows in Figures 4.4-4.6.

In addition, for the case of the combination $\alpha_r = \alpha_t$ or $\alpha_r \neq \alpha_t$, the former case is better than the latter case, and the α_r is greater than α_t. It indicates that the rapidly attenuated weighting is better for noisy speech since the subband autocorrelation coefficients at integral multiples of $f_{cf_i}^{-1}$ are affected by noise.

Furthermore, the recognition rates of the case $\alpha_r > \alpha_t$ under noisy conditions are as good as those of the best combination (BEST), The combinations (α_r, α_t) for Gaussian white noise, HSL noise and computer room noise are $(0.6, 0.2)$, $(0.6, 0.4)$ and $(0.5, 0.1)$ respectively. These results indicate that (1) the higher frequency resolution and the lower spectral contrast enhancement are better in extracting reference patterns, (2) the lower

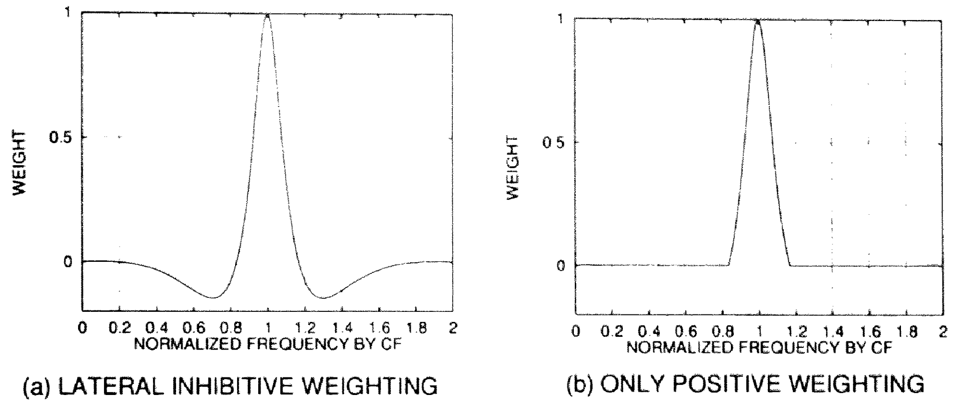


Figure 4.7: (a) Lateral inhibitive weighting, and (b) only positive weighting of (a). (Q=1.5,$\alpha = 0.5$)

frequency resolution and the higher spectral contrast enhancement are better in extracting test patterns.

Comparison with SGDS and MFCC.

The best SBCOR with MDW processing (BEST) performs better than SGDS and MFCC under all SNR conditions for Gaussian white noise and HSL noise. At SNR 0dB, the recognition rates are about 15% and 20% higher than that of SGDS and MFCC. For computer room noise, the best SBCOR outperforms SGDS and MFCC below SNR 10dB, the degrees of improvement are less than that of Gaussian white noise and HSL noise. It seems that the computer room noise has several periodic components.

4.4 Evaluation of Lateral Inhibitive Weighting

To investigate to what extent the lateral inhibitive weighting for power spectrum shown in Figure 4.2 is effective for noise robust speech recognition, DTW word recognition performs for the case of (1) the lateral inhibitive weights are removed and (2) the lateral inhibitive weights are used.

4.4.1 Experimental Conditions

Gaussian white noise and the computer room noise were added to clean speech. The power spectrum has several sharp peaks (See Figure 3.5 in Chapter 3). The same standard DTW speaker-dependent isolated word recognizer in the previous section was used.

Under the above experimental conditions, the effectiveness of LIW is investigated by removing the inhibitive (or negative) weight, as shown in Figure 4.7. The LIW is applied directly to power spectrum of each analysis frame signal calculated by FFT. Moreover, in order to show the effectiveness under noisy conditions, the performance of LIW is compared with those of SGDS[41] and MFCC[23].

4.4.2 Results

The experimental results are shown in Figure 4.8 for each noise. The results of LIW are the best case. The best pair (α_r, α_t) in figures stands for αs in extracting reference patterns and test patterns respectively. The results of "no LIW" were obtained by the same αs of the best LIW except for using positive only weights.

As shown in figures, the LIW performs better than the non LIW under all conditions. These results clarify that the robustness of SBCOR is the effect of the LIW for power spectrum.

4.4.3 Discussions

When $\hat{W}_i(f)$ is considered as an impulse response applied to power spectrum, LIW can be seen as a weighting procedure in the autocorrelation domain, like liftering procedure in the cepstrum domain. As shown in Figure 4.9, LIW suppresses low order autocorrelation unlike the case of positive only weighting. This implies that LIW is qualitatively equivalent to the spectral tilt elimination. In addition, the higher order autocorrelation is also suppressed for smaller α. The recognition results that smaller α (0.3 for white noise, 0.1 for computer room noise) is preferred under noisy conditions coincide with the fact that higher order autocorrelation is more influenced by noise. These effects, i.e., (1) spectral tilt

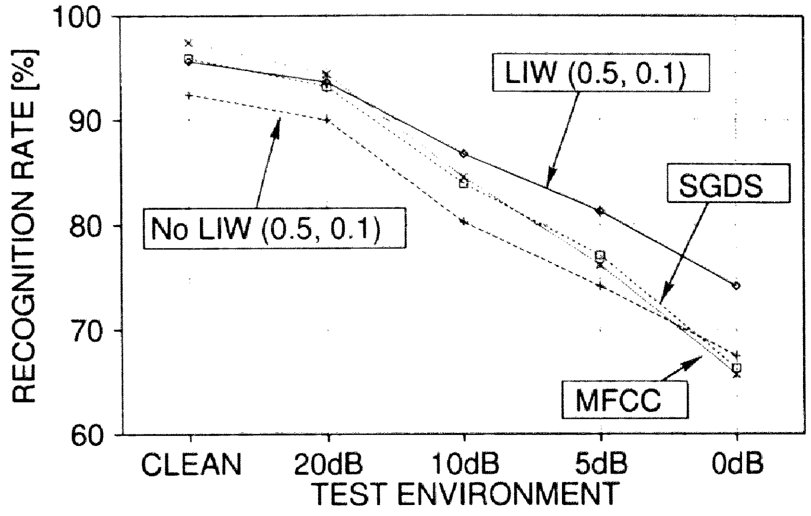


Figure 4.8: Averaged recognition rates for each features.

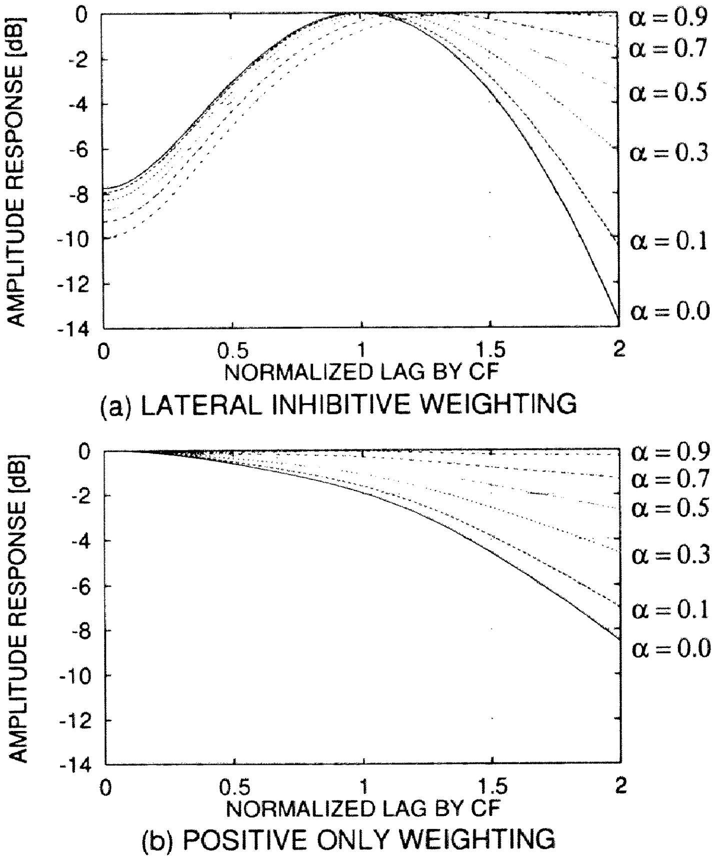


Figure 4.9: Amplitude response of LIW normalized by τ_{cf_i}.

elimination and (2) noise variability elimination, would be the essence of lateral inhibitive weighting, and lead to a more robust recognition under noisy conditions.

4.5 Conclusions

In this chapter, to improve the robustness of SBCOR, multi-delay weighting (MDW) processing was introduced, which is used to extract periodicity using not only the autocorrelation coefficient at 1/CF, but also a weighted sum of autocorrelation coefficients at integral multiples of 1/CF. The experimental results using a DTW recognizer showed SBCOR with MDW performs better than SGDS and MFCC under noisy conditions. Furthermore, the lateral inhibitive weighting of the power spectrum is specially focused to interpret the robustness of SBCOR, and it is shown that (1) spectral tilt elimination and (2) noise variability elimination would be the essence of lateral inhibitive weighting, and lead to a more robust recognition under noisy conditions.

Chapter 5

Subband-Crosscorrelation Analysis Using Two Input Channel Signals

5.1 Introduction

In this chapter, subband-crosscorrelation analysis (SBXCOR) is proposed in order to improve the robustness of SBCOR. In SBXCOR, the crosscorrelation coefficients of two input channel signals at a lag of τ_{cf_i}, which is associated with the inverse of center frequency, are used instead of the autocorrelation coefficients in SBCOR. Furthermore, to capture more periodicity information, multi-delay weighting processing (MDW) used for SBCOR in Chapter 4 is also applied for SBXCOR. The evaluation using a DTW word recognizer is performed under a simulated acoustic condition on computer and a real acoustic condition.

This chapter is constructed as follows. The following section reviews SBCOR analysis and describes the proposed SBXCOR analysis in detail. In addition, MDW processing and the interpretation in the frequency domain are presented. Sections 5.3 and 5.4 investigate the robustness under a simulated acoustic condition on computer and a real acoustic condition, respectively. Finally, Section 5.5 concludes the whole chapter.

5.2 SBXCOR Analysis and MDW Processing

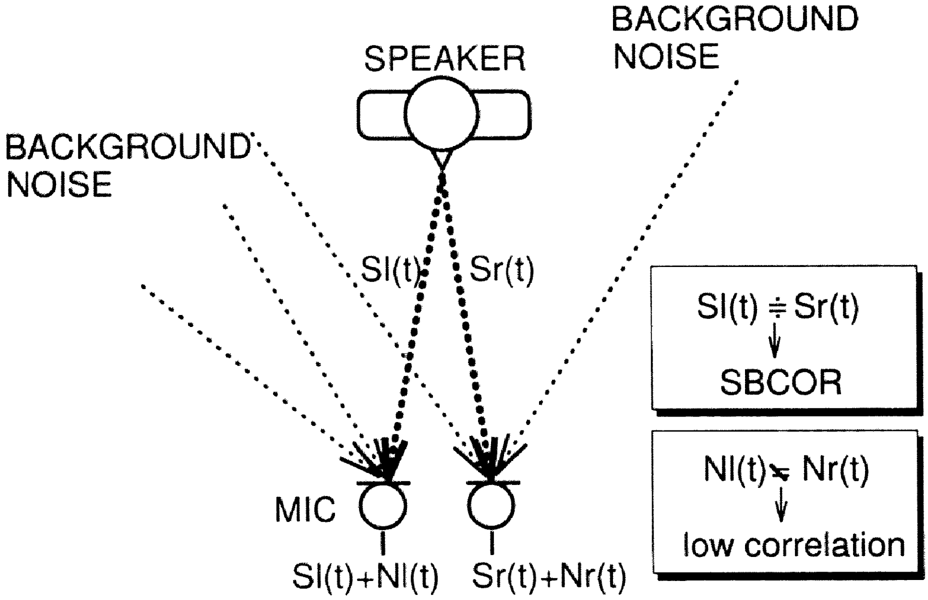


Figure 5.1: Concept of SBXCOR analysis. Since the speech components $S_l(t), S_r(t)$ in the signals recorded by two microphones, which is uttered just in front of two microphones, have the same amplitude and phase, SBXCOR extracts the same spectrum as SBCOR. On the other hand, if noise components $N_l(t), N_r(t)$ are low correlation between two channels, their influences are canceled in the processing.

5.2.1 SBXCOR Analysis

SBCOR analysis is a signal processing method based on the "autocorrelation" of a speech signal so as to extract periodicity in terms of the inverse of the center frequency. As seen in the auditory system, however, binaural signal processing seems to be more important in the real environment. Therefore, in order to improve the performance of speech recognition, we extend SBCOR analysis so that the autocorrelation analysis is replaced by crosscorrelation analysis defined as Equation (5.1), and refer to it as "subband-crosscorrelation" analysis, or SBXCOR analysis in the abbreviated form.

$$\begin{aligned} Sc_n(i) &= \frac{R^i_{x_n y_n}(\tau_{cf_i})}{\sqrt{R^i_{x_n x_n}(0) R^i_{y_n y_n}(0)}}, \quad \tau_{cf_i} = f_{cf_i}^{-1} \\ R^i_{x_n y_n}(\tau) &= \int_{-\infty}^{\infty} \mid H_i(f) \mid^2 X_{x_n y_n}(f) e^{j2\pi f\tau} df \\ R^i_{x_n x_n}(0) &= \int_{-\infty}^{\infty} \mid H_i(f) \mid^2 X_{x_n x_n}(f) df \\ R^i_{y_n y_n}(0) &= \int_{-\infty}^{\infty} \mid H_i(f) \mid^2 X_{y_n y_n}(f) df, \end{aligned} \tag{5.1}$$

where $R^i_{x_n x_n}(\tau)$, $R^i_{y_n y_n}(\tau)$ and $R^i_{x_n y_n}(\tau)$ are the autocorrelation and crosscorrelation functions of ith subband signal respectively, and $X_{x_n x_n}(f)$ and $X_{y_n y_n}(f)$ are the power spectrums of the nth analysis frame signals $x_n(t)$ and $y_n(t)$ respectively. $X_{x_n y_n}(f)$ is the cross power spectrum.

The robustness of SBXCOR against noise can be explained as shown in Figure 5.1. Since the speech signals recorded by two microphones, which is uttered just in front of two microphones, have the same amplitude and phase, SBXCOR extracts the same spectrum as SBCOR. On the other hand, if the noise has a low correlation between the two channels, their influences are canceled in the processing. In the following experiments, we will investigate the performance of SBXCOR under the assumption that speakers utter just in front of two microphones and noises are not correlated between two channels.

SBXCOR analysis is implemented using FFT in this research, as in the case of SBCOR.

5.2.2 SBXCOR with Multi-Delay Weighting (MDW) Processing

If both of the binaural signals are periodic with a period T, the crosscorrelation coefficients show several peaks at integral multiples of T. In SBXCOR analysis defined by Equation (5.1), however, only one crosscorrelation coefficient at T is used to extract the periodicity included in the subband signal. Therefore, we extend the SBXCOR to capture the other peaks of the crosscorrelation coefficients by taking a weighted sum of them with the power of α, i.e. the exponential weighting (Figure 5.2) as follows;

$$\hat{Sc}_n(i) = \frac{1}{A}\sum_{k=0}^{\infty}\alpha^k\frac{R^i_{x_n y_n}((k+1)\tau_{cf_i})}{\sqrt{R^i_{x_n x_n}(0)R^i_{y_n y_n}(0)}}, \qquad (5.2)$$

where $A = \sum_{k=0}^{\infty}\alpha^k$, $0 \leq \alpha < 1$. We have referred to it as multi-delay weighting (MDW) processing[45]. The MDW processing has been shown to be effective in SBCOR analysis[45, 46].

Equation (5.2) can be expressed in the frequency domain as follows (see Appendix 5.A):

$$\begin{aligned}\hat{Sc}_n(i) &= \frac{\int_{-\infty}^{\infty}\hat{Wc}_i(f)X_{x_n y_n}(f)df}{\sqrt{R^i_{x_n x_n}(0)R^i_{y_n y_n}(0)}}\\ \hat{Wc}_i(f) &= \frac{(1-\alpha)e^{j2\pi\tau_{cf_i}f}}{1-\alpha e^{j2\pi\tau_{cf_i}f}} \mid H_i(f)\mid^2 .\end{aligned}$$

This means that SBXCOR analysis with MDW processing results in the weighting processing of the complex cross power spectrum $X_{x_n y_n}(f)$ by the complex weighting function $\hat{Wc}_i(f)$. It is easily derived that the magnitude of $\hat{Wc}_i(f)$ is

$$\mid \hat{Wc}_i(f)\mid = \frac{(1-\alpha)\mid H_i(f)\mid^2}{\sqrt{1-2\alpha\cos 2\pi f\tau_{cf_i}+\alpha^2}.} \qquad (5.3)$$

As shown in Figure 5.3, by MDW processing, the frequency resolution of SBXCOR is controllable by α. The frequency resolution is higher as α becomes closer to one. The contribution of this effect on recognition performance will be experimentally shown in the following recognition experiments.

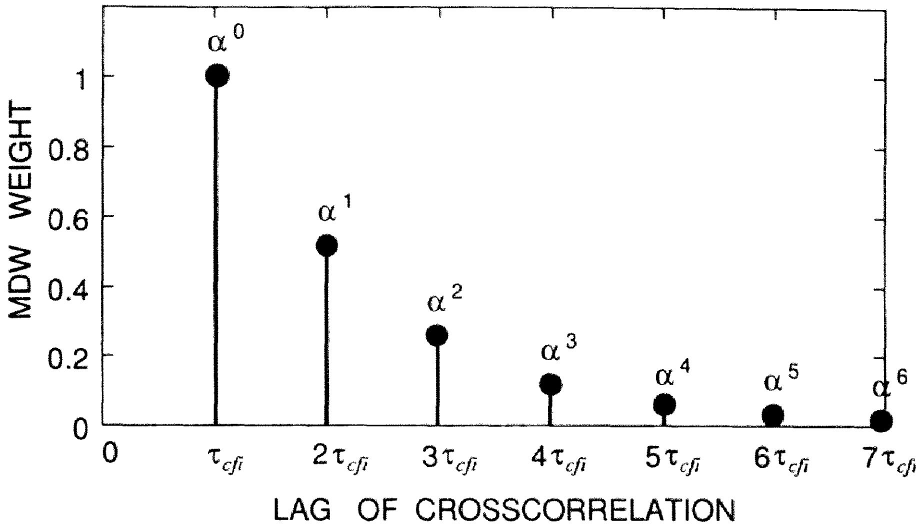


Figure 5.2: Multi-delay weighting for crosscorrelation coefficients at integral multiples of τ_{cf_i}.

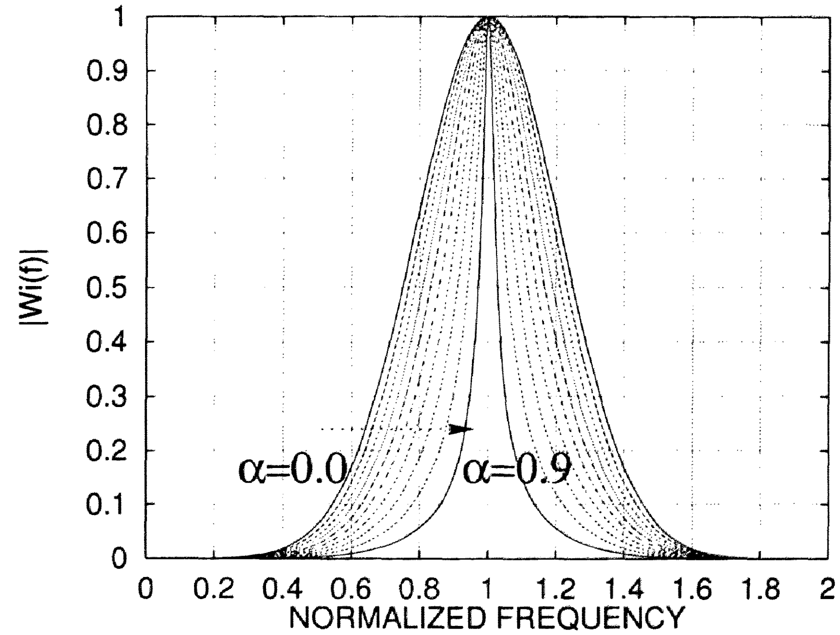


Figure 5.3: Weighting function for the magnitude of cross power spectrum (Q=2.0). The horizontal axis is the normalized frequency by center frequency (CF).

5.3 Evaluation under Simulated Acoustic Condition

In this section, in order to investigate the upper-bound performance of SBXCOR, we perform recognition experiments under a simulated acoustic condition. We assume that speakers utter just in front of the two microphones, i.e. ;

1. the speech signals recorded by two microphones are perfectly synchronized,

2. the reverberation in the real environment can be ignored,

3. received noise signals are not correlated between two microphones.

Of course, these assumptions are not realistic in the real acoustic environment where a speech recognizer is used. However, since there are a lot of uncontrollable factors in the real environment, we start the investigation under these assumptions.

5.3.1 Experimental Conditions

The above condition is implemented on computer by adding Gaussian white noise to speech signals. In the experiment, using DTW word recognition, we compare the robustness of SBXCOR analysis with that of SBCOR. As a further reference, we also compare SBXCOR with smoothed group delay spectrum (SGDS)[27, 39] and mel-filterbank cepstral coefficient (MFCC)[23] extracted from one-channel signal by simply summing the two signals.

DTW word recognizer

The same standard DTW speaker-dependent isolated word recognizer is used as in Chapter 2. The recognition task is a 68 pair discrimination. Each pair is a phonetically similar city name pair, selected from a 550 Japanese city name database recorded twice by 5 Japanese male speakers. The first set is used as the reference pattern and the second set, which was spoken a week later, is used as the test pattern.

Generation of two-channel signals

In order to simulate the situation that noise components in two channels are not correlated each other, two Gaussian white noises generated by changing the seed are added to the speech database. The global signal-to-noise ratios (SNRs) used in the test phase are 20, 10, 5 and 0dB.

Generation of one-channel signal by simply summing the two signals

The two-channel-summed signals are generated by simple synchronized summation of the above two signals. By doing this processing, the effective SNR improvement of the two-channel-summed signals is about 3dB.

Smoothed group delay spectrum (SGDS)

SGDS has been shown to be robust against noise, and it is calculated as the derivative of phase of a p th order all pole filter that has smoothed poles[27, 39]. In order to compare the performance of SBXCOR with that of SGDS under exactly the same conditions, the analysis frequency points of SGDS were chosen to be the same as the center frequencies of SBXCOR.

Mel-filterbank cepstral coefficient (MFCC)

MFCC is commonly used as speech feature in speech recognition[23]. In recent research, the noise robustness of MFCC is almost the same as the auditory models proposed by Seneff and Ghitza[24]. In this experiment, MFCC is calculated using a 28 triangular shape mel-filterbank.

SBCOR, SBXCOR and MDW Processing

The Q values of 1.0, 1.5, 2.0, 2.5 and 3.0 are investigated. FFT-point is 1024. In order to calculate coefficients of the correlation function at τ_{cf_i} precisely, two-times oversampling and polynomial interpolation were used. The center frequencies of the BPFs are equally spaced on the Bark scale between 4 and 17 Bark. In MDW processing, the range of the summation is up to $8\tau_{cf_i}$.

Common analysis conditions

The analysis frame length and shift are 20 ms and 10 ms, respectively. The dimension of each feature is 16. The sampling rate is 10 kHz.

5.3.2 Experimental Results

The recognition rates of SBXCOR, SBCOR, SGDS and MFCC are shown in Figures 5.4-5.6. In SBXCOR and SBCOR analysis, the best Qs were 2.0 and 1.5 respectively. These results are summarized to the following four points.

1. SBXCOR is more robust than the conventional one-channel SBCOR under all test conditions. However, the improvement was less than about 2% at SNR 0dB (Figure 5.4).

2. The performance of SBXCOR is less than SBCOR extracted from the two-channel-summed signal (Figure 5.4).

3. By introducing MDW processing, SBXCOR performs significantly better under noisy conditions as shown in Figure 5.5. At SNR 0dB, the improvement was 5%. The recognition rates are better than those of SBCOR extracted from the two-channel-summed signal at SNR 0dB. The best combination of αs for analyzing reference and test patterns were 0.5 and 0.0. It indicates that the frequency resolution under noisy conditions should be broader than under clean conditions.

4. SBXCOR performs better than SGDS and MFCC even if two-channel-summed signals are used where SNR is below 10dB (Figure 5.6).

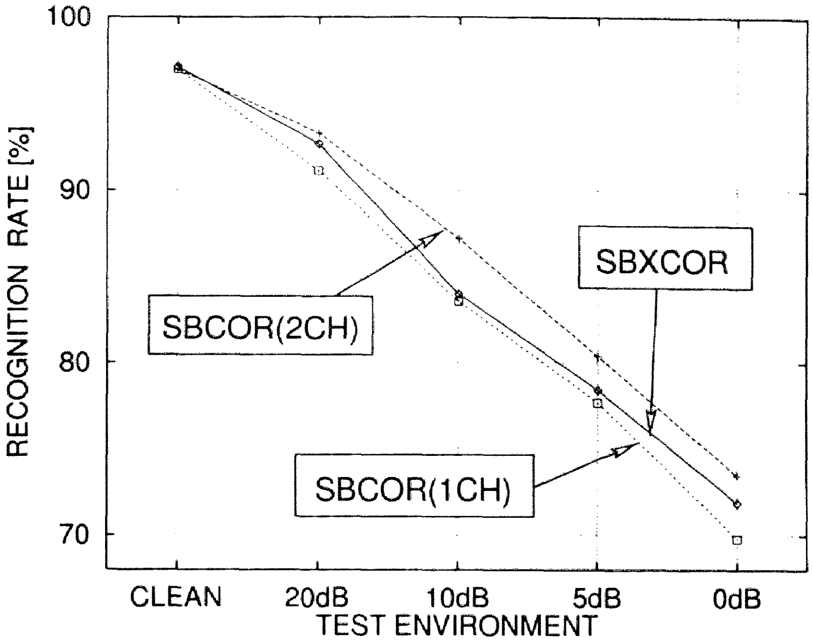


Figure 5.4: Recognition results under the simulated acoustic condition of SBXCOR, SBCOR. SBCOR(2CH) was extracted from the two-channel-summed signals.

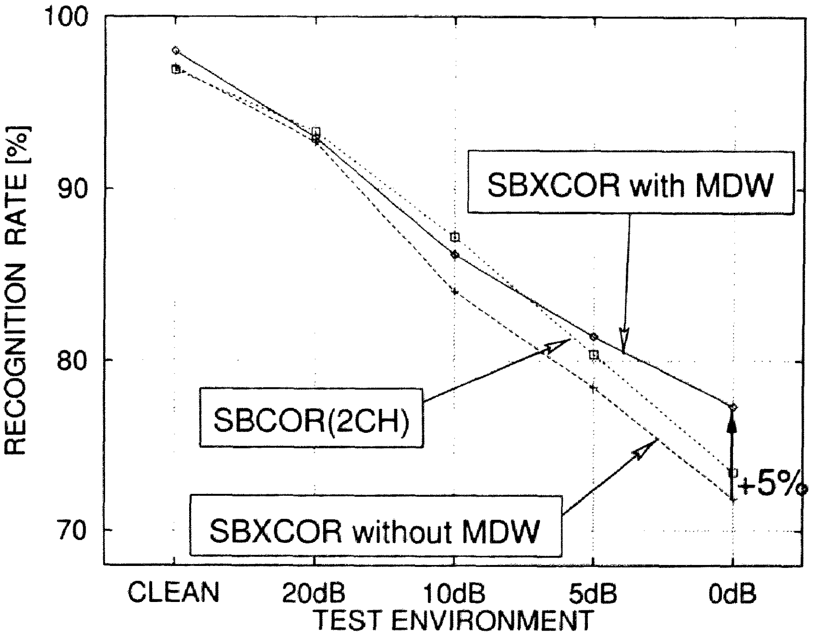


Figure 5.5: Recognition results of SBXCOR with and without MDW. The black arrow at SNR 0dB shows the improvement by using MDW.

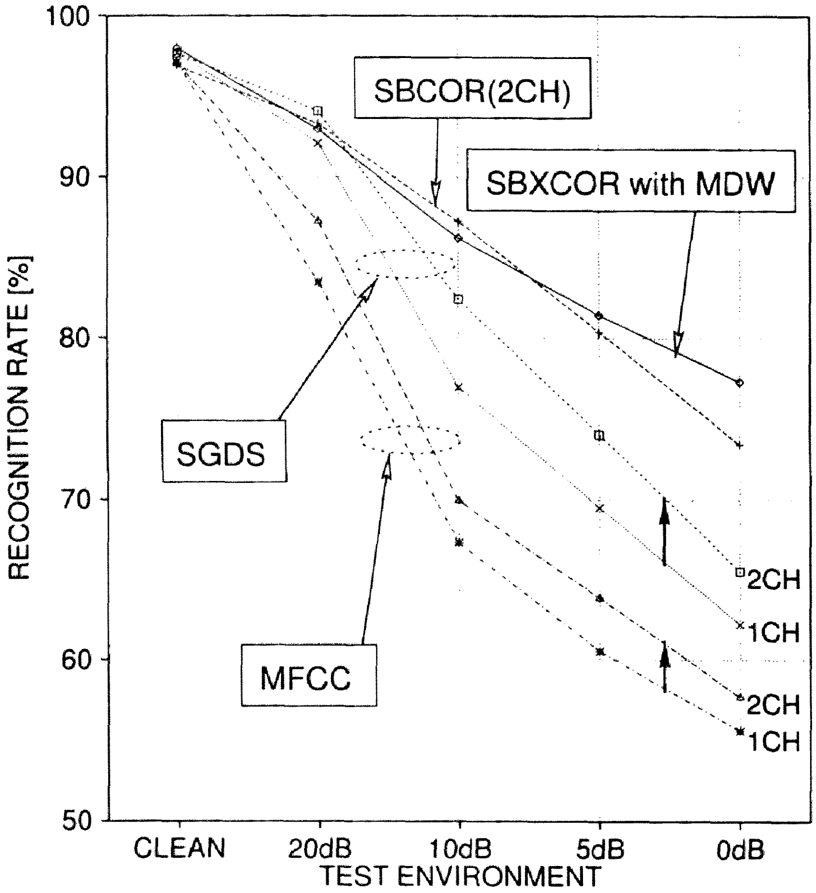


Figure 5.6: Recognition results compared with SGDS and MFCC. The black arrows show the improvements by using the two-channel-summed signal.

Table 5.1: Averaged global SNR of recorded speech data (in dB).

	CLEAN	20dB	10dB	0dB
CH0	50.2	23.7	13.9	4.0
CH1	51.7	23.4	13.5	3.7
CH2	42.5	22.7	12.9	3.0
CH3	52.7	21.6	11.8	1.9
CH4	55.0	20.5	10.6	0.8
CH5	55.2	19.7	9.8	-0.3
CH6	49.9	17.1	7.3	-2.6

5.4 Evaluation under Real Acoustic Condition

In this section, we investigate the performance of SBXCOR in a real acoustic environment. This evaluation of SBXCOR is performed by using speech data recorded by a microphone array in a sound proof room.

5.4.1 Database Recording

The 68 city name pairs used in the previous section were played from a loud speaker, and recorded using a microphone array (see Figure 5.7). The microphone array consists of seven omni-directional electret microphones (Sony ECM-77B). Spacing between microphones is 10 cm. In recording the speech database, human speech-like noise[47] was also output from the other loud speaker placed at left 90 degrees, the speech data are recorded under four different SNRs. Table 5.1 shows the averaged global SNR of speech. The averaged global SNRs clearly show that the SNR of each channel becomes lower as the microphone is close to the loud speaker for noise. For convenience, the four different environmental test conditions related to the SNRs are written by CLEAN, 20dB, 10dB and 0dB in turn.

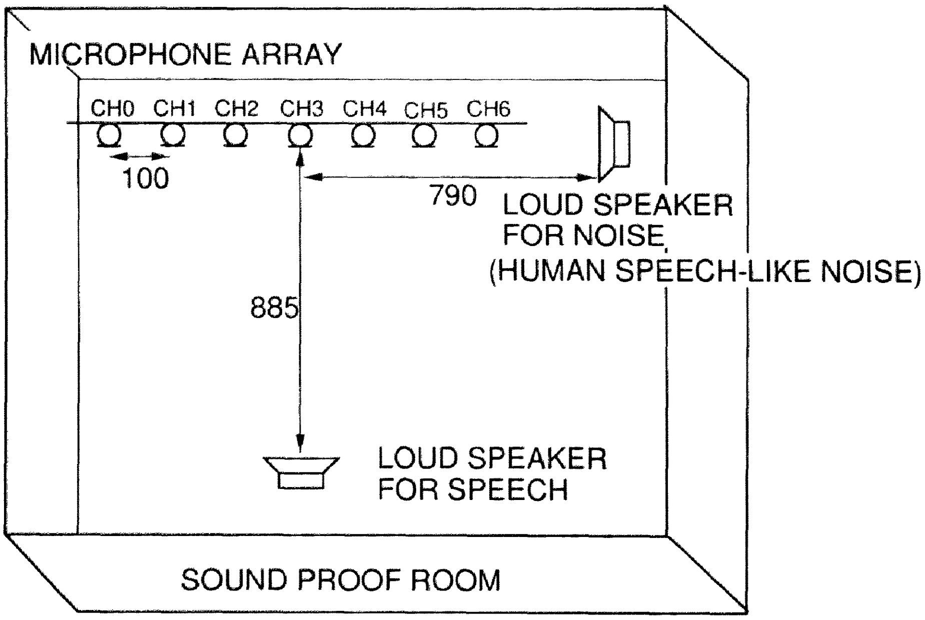


Figure 5.7: Recording system (The unit is in mm).

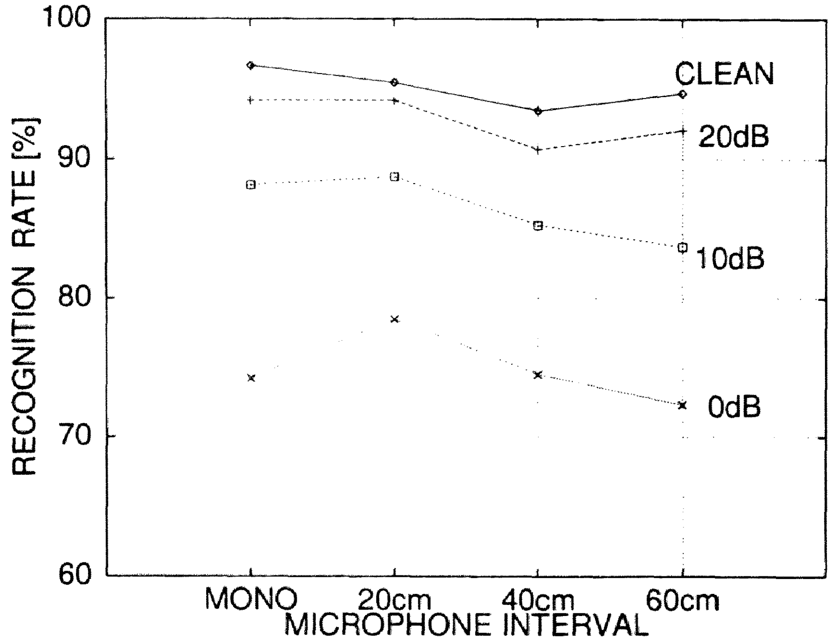


Figure 5.8: Recognition results of SBXCOR for three microphone pairs. The "MONO" stands for the results of the conventional SBCOR using CH3.

5.4.2 Experimental Conditions

The evaluation was performed by using the same DTW word recognition system in Section 3. In the experiment, the following three points were investigated:

1. the performance of SBXCOR using three microphone pairs (CH4, CH2), (CH5, CH1) and (CH6, CH0),
2. the performances of SBCOR, SGDS and MFCC using (1) CH3, which is the middle channel of the microphone array, and (2) the two-channel-summed signal generated from a channel pair (CH4, CH2) as in Section 3,
3. the effectiveness of MDW processing.

The spacing of each pairs were 20 cm for (CH4, CH2), 40 cm for (CH5, CH1) and 60 cm for (CH6, CH0).

5.4.3 Experimental Results

Figure 5.8 shows the best recognition results (Q=2.0) of SBXCOR for the three microphone pairs. As shown in the figure, SBXCOR performs equally as well as the conventional SBCOR under relatively high SNR conditions, and about 4% better than SBCOR at SNR 0dB for the best case using the (CH4,CH2) pair. In the CLEAN case, the performance of SBXCOR should be equal to that of SBCOR because the two channel signals are ideally the same. In the real acoustic environment, however, the assumption is not alway true due to the different transfer characteristics for the two channels. That is why the performance of SBXCOR deteriorates.

The best recognition rates of SBXCOR, SBCOR, SGDS and MFCC are shown in Figures 5.9-5.11. In SBXCOR and SBCOR analyses, the best Qs were 2.0. These results are summarized to four points as in the previous section.

1. The results shown in Figure 5.9 indicate that SBXCOR is more robust than the

conventional one-channel SBCOR below SNR 10dB. The best improvement was about 4% at SNR 0dB.

2. Although the performance of SBXCOR is less than SBCOR extracted from the two-channel-summed signal at SNR 10dB, SBXCOR outperforms SBCOR at SNR 0dB (See Figure 5.9 again).

3. By introducing MDW processing, SBXCOR performs better under noisy conditions as shown in Figure 5.10. At SNR 0dB, the improvement was 2%. The recognition rates are better than that of SBCOR extracted from the two-channel-summed signal at SNR 0dB. The best combination of αs in extracting reference and test patterns were 0.3 and 0.1. It indicates that the frequency resolution under noisy condition should be broader than under clean conditions, as shown in the previous section.

4. SBXCOR performs better than SGDS and MFCC below SNR 10dB even if two-channel-summed signals are used. (Figure 5.11)

5.5 Conclusions

In this chapter, we proposed subband-crosscorrelation analysis and investigated the robustness using a DTW word recognizer, under a simulated acoustic condition on computer and a real acoustic environmental condition. Under the simulated condition, we clarified that SBXCOR is more robust than the conventional one-channel SBCOR, but less robust than SBCOR extracted from the two-channel-summed signal. In addition, by applying MDW processing, the performance of SBXCOR was improved. The resultant performance of SBXCOR with MDW processing was much better than those of smoothed group delay spectrum and mel-filterbank cepstral coefficient below SNR 10dB. The results under the real acoustic condition were almost the same as the simulated acoustic condition.

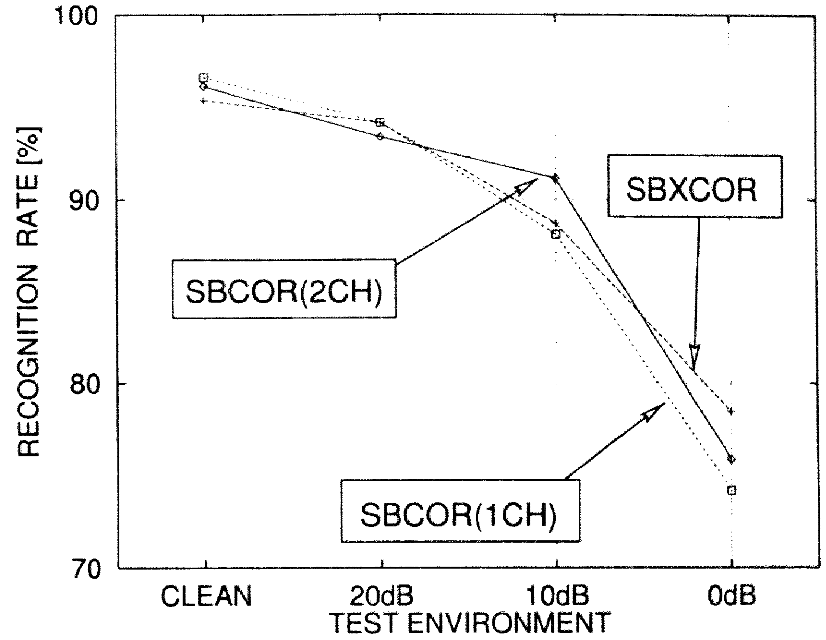


Figure 5.9: Recognition results under the real acoustic condition of SBXCOR, SBCOR. SBCOR(2CH) was extracted from the two-channel-summed signals.

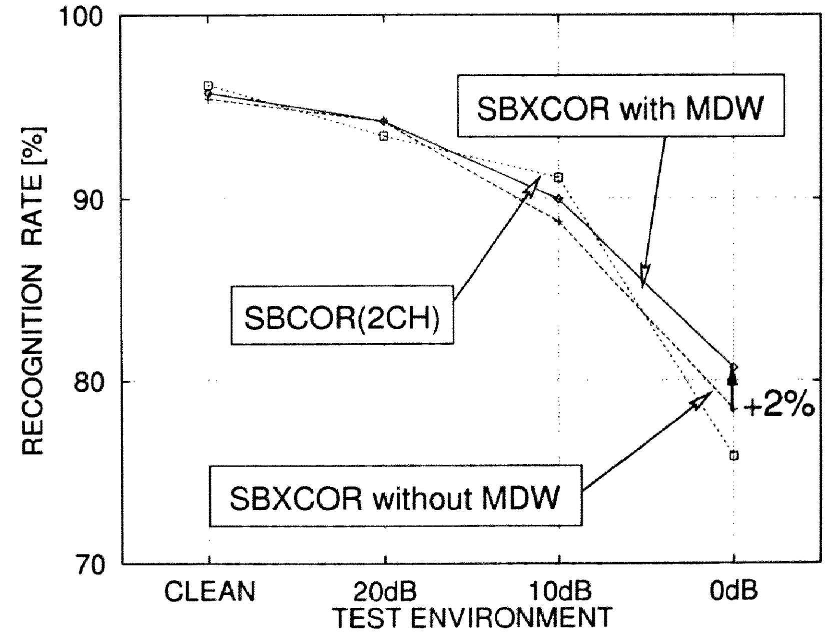


Figure 5.10: Recognition results of SBXCOR with and without MDW in the real acoustic environment. The black arrow at SNR 0dB shows the improvement by using MDW.

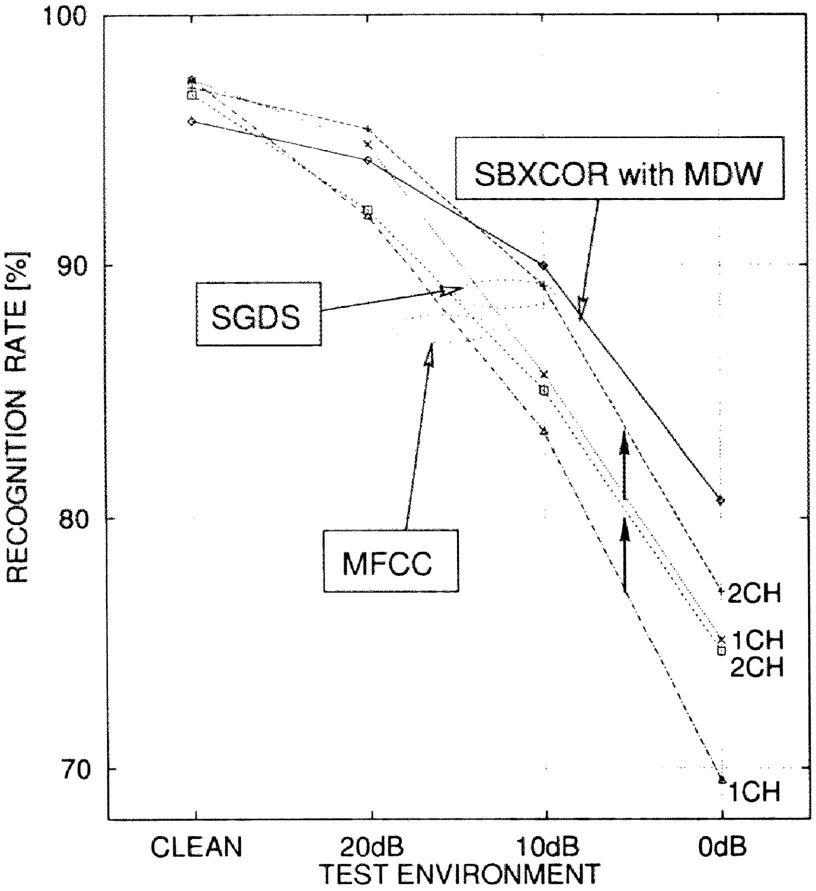


Figure 5.11: Recognition results compared with SGDS and MFCC in the real acoustic environment. The black arrows show the improvements by using the two-channel-summed signal.

5.A Appendix

When $v_n^i(\tau)$ is defined as

$$v_n^i(\tau) = \sum_{k=0}^{\infty} \alpha^k R_{x_n y_n}^i(\tau + (k+1)\tau_{cf_i}),$$

Equation (5.2) can be calculated by

$$\hat{Sc}_n(i) = v_n^i(\tau)\Big|_{\tau=0} / \{A\sqrt{R_{x_n x_n}^i(0) R_{y_n y_n}^i(0)}\}.$$

Thus, MDW processing can be seen as a linear filter whose input and output are $R_{x_n y_n}^i(\tau)$ and $v_n^i(\tau)$ respectively. Changing the range of the summation,

$$\begin{aligned} v_n^i(\tau) &= \frac{1}{\alpha}\sum_{k=1}^{\infty} \alpha^k R_{x_n y_n}^i(\tau + k\tau_{cf_i}) \\ &= \frac{1}{\alpha}\left\{\sum_{k=0}^{\infty} \alpha^k R_{x_n y_n}^i(\tau + k\tau_{cf_i}) - R_{x_n y_n}^i(\tau)\right\}. \end{aligned}$$

Let the Fourier transform of $R_{x_n y_n}^i(\tau)$ and $v_n^i(\tau)$ be $X_{x_n y_n}^i(f)$ and $V_n^i(f)$, then

$$\begin{aligned} V_n^i(f) &= \frac{1}{\alpha}\left\{\sum_{k=0}^{\infty} \alpha^k X_{x_n y_n}^i(f) e^{j2\pi\tau_{cf_i} kf} - X_{x_n y_n}^i(f)\right\} \\ &= \frac{X_{x_n y_n}^i(f)}{\alpha}\left\{\sum_{k=0}^{\infty} \alpha^k e^{j2\pi\tau_{cf_i} kf} - 1\right\} \\ &= \frac{X_{x_n y_n}^i(f)}{\alpha}\left\{\frac{1}{1-\alpha e^{j2\pi\tau_{cf_i} f}} - 1\right\} \\ &= \frac{e^{j2\pi\tau_{cf_i} f}}{1-\alpha e^{j2\pi\tau_{cf_i} f}} X_{x_n y_n}^i(f). \end{aligned}$$

By the inverse Fourier transform and $\tau = 0$, then

$$v_n^i(\tau)\Big|_{\tau=0} = \int_{-\infty}^{\infty} \frac{e^{j2\pi\tau_{cf_i} f} X_{x_n y_n}^i(f)}{1-\alpha e^{j2\pi\tau_{cf_i} f}} df.$$

Where

$$\begin{aligned} X_{x_n y_n}^i(f) &= \mid H_i(f) \mid^2 X_{x_n y_n}(f) \\ A &= \sum_{k=0}^{\infty} \alpha^k = 1/(1-\alpha), \end{aligned}$$

Equation (5.2) can be expressed in the frequency domain as follows:

$$\begin{aligned}
\hat{Sc}_n(i) &= \frac{\int_{-\infty}^{\infty} W_i(f) X_{x_n y_n}(f) df}{\sqrt{R^i_{x_n x_n}(0) R^i_{y_n y_n}(0)}} \\
W_i(f) &= \frac{(1-\alpha)e^{j2\pi\tau_{cf_i} f}}{1-\alpha e^{j2\pi\tau_{cf_i} f}} \mid H_i(f) \mid^2 .
\end{aligned}$$

Thus, SBXCOR analysis with MDW processing results in the weighting processing for cross power spectrum $X_{x_n y_n}(f)$ by the weighting function $W_i(f)$.

Chapter 6

Epilogue

To address the noise robust problem of ASR systems, subband-autocorrelation analysis and its extensions were proposed in this dissertation.

At first, various implementations of SBCOR were compared under noisy conditions affected by the multiplicative signal-dependent white noise in Chapter 2. The experimental results using a speaker-dependent DTW isolated word recognizer showed that the most suitable filter bank and periodicity detection method are a fixed Q filter bank whose center frequencies are equally spaced on the Bark scale, and a conventional autocorrelation detection, without controlling weak signals, respectively.

Chapter 3 clarified the robustness of SBCOR analysis in more realistic adverse environments; the existence of three additive noises and waveform distortion. As the results show, although the robustness of SBCOR against Gaussian white noise is much better than those of SGDS and MFCC, the robustness against human speech-like noise and computer room noise is better than that of MFCC, but only a little better than that of SGDS.

In Chapter 4, by introducing multi-delay weighting (MDW) processing, it was shown that the robustness of SBCOR improves against human speech-like noise and computer room noise. In addition, it was shown that SBCOR with MDW processing results in the weighting processing of the power spectrum of speech by a lateral inhibitive weighting function. Furthermore, it was shown that (1) spectral tilt elimination and (2) noise variability elimination would be the essence of lateral inhibitive weighting, and lead to a more robust recognition under noisy conditions.

Finally, Chapter 5 described subband-crosscorrelation analysis (SBXCOR) using two input channel signals. As the experimental results showed, under both the simulated acoustic condition and the real acoustic condition, SBXCOR is more robust than conventional one-channel SBCOR, but less robust than SBCOR extracted from the two-channel-summed signal. Furthermore, by applying MDW processing, the performance of SBXCOR improved. The resultant performance of SBXCOR with MDW processing was much better than those of smoothed group delay spectrum (SGDS) and mel-filterbank cepstral coefficient (MFCC).

Noise robust problem restricts the widespread use of ASR systems in our social life. Although the problem was not always solved perfectly through this dissertation, it was clarified that the periodicity information associated with the inverse of the center frequency included in speech signals plays the significant role in the noise robust acoustic analysis.

In the future works, we intend to investigate the robustness under the other adverse conditions, for example, reverberation, channel distortion and so on. Furthermore, the investigation in continuous speech recognition paradigm is also desired.

Bibliography

[1] L. Rabiner and B.-H. Juang: "Fundamentals of speech Recognition", Prentice Hall (1993).

[2] J. R. Deller, J. G. Proakis and J. H. L. Hansen: "Discrete-Time Processing of Speech Signals", Prentice Hall (1993).

[3] K. H. Davis, R. Biddulph and S. Balashek: "Automatic recognition of spoken digits", J. Acoust. Soc. Am., **24**, pp. 637–642 (1952).

[4] H. F. Olson and H. Belar: "Phonetic typewriter", J. Acoust. Soc. Am., **28**, pp. 1072–1081 (1956).

[5] J. W. Forgie and C. D. Forgie: "Results obtained from a vowel recognition computer program", J. Acoust. Soc. Am., **31**, pp. 1480–1489 (1959).

[6] J. Suzuki and K. Nakata: "Recognition of japanese vowels – preliminary to the recognition of speech", J. Radio Res. Lab, **37**, pp. 193–212 (1961).

[7] T. Sakai and S. Doshita: "The phonetic typewriter, information processing 1962", Proc. IFIP Congress (1962).

[8] H. Sakoe and S. Chiba: "Dynamic programming algorithm optimization for spoken word recognition", **ASSP**, **64**, pp. 43–49 (1978).

[9] T. K. Vintsyuk: "Speech discrimination by dynamic programming", Kibernetika, **4**, 2, pp. 81–88 (1968).

[10] F. Itakura and S. Saito: "Analysis synthesis telephony based on the maximum likelihood method", Reports of the 6th International Congress on Acoustics, **II**, C-5-5, pp. C17–C20 (1968).

[11] F. Itakura: "Minimum prediction residual principle applied to speech recognition", IEEE Transactions on Acoustic, Speech and Signal Processing, **23**, 1, pp. 67–72 (1975).

[12] J. K. Baker: "Stochastic modeling for automatic speech understanding", Academic Press (1975).

[13] F. Jelinek, L. R. Bahl and R. L. Mercer: "Design of a linguistic statistical decoder for the recognition of continuous speech", IEEE Transactions on Information Theory, **21**, pp. 250–256 (1975).

[14] Y. Gong: "Speech recognition in noisy environments: A survey", Speech Communication, **16**, pp. 261–291 (1995).

[15] R. P. Lippmann: "Speech recognition by machines and humans", Speech Communication, **22**, pp. 1–15 (1997).

[16] S. Das, R. Bakis, A. Nadas, D. Hahamoo and M. Picheny: "Influence of background noise and microphone on the performance of the ibm tangora speech recognition system", Proc. of ICASSP, Vol. II, pp. 71–74 (1993).

[17] X. Aubert, R. Haeb-Umbach and H. Ney: "Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models", Proc. of ICASSP, Vol. II, pp. 648–651 (1993).

[18] S. Nakagawa: "Acoustic signal processing techniques for robust speech recognition", J. Acoust. Soc. Jpn., **53**, 11, pp. 864–871 (1997).

[19] J.-C. Junqua: "Impact of the unknown communication channel on automatic speech recognition: A review", Proceedings of Eurospeech97, **2**, pp. KN29–KN32 (1997).

[20] A. Acero: "Acoustical and Environmental Robustness in Automatic Speech Recognition", Kluwer Academic Publishers (1993).

[21] S. Seneff: "A joint synchrony/mean-rate model of auditory speech processing", J. Phonetics, **16**, pp. 55–76 (1988).

[22] O. Ghitza: "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment", J. Phonetics, **16**, pp. 109–123 (1988).

[23] S. B. Davis and P. Mermelstein: "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustics, Speech and Signal Processing, **ASSP-28**, pp. 357–366 (1980).

[24] C. R. Jankowski Jr., H.-D. H. Vo and R. P. Lippmann: "A comparison of signal processing front ends for automatic word recognition", IEEE Trans. on Speech and Audio Processing, **3**, pp. 286–294 (1995).

[25] H. Hermansky and N. Morgan: "Rasta processing of speech", IEEE Trans. on Speech and Audio Processing, **2**, pp. 578–589 (1994).

[26] K. Aikawa, H. Singer, H. Kawahara and Y. Tohkura: "Cepstrum representation of speech motivated by time-frequency masking: an application to speech recognition", J. Acoust. Soc. Am., **100**, pp. 603–614 (1996).

[27] F. Itakura and T. Umezaki: "Distance measure for speech recognition based on the smoothed group delay spectrum", Proc. of ICASSP, Vol. 3, pp. 1257–1260 (1987).

[28] S. F. Boll: "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustic, Speech and Signal Processing, **27**, pp. 113–120 (1979).

[29] Y. gu and J. S. Mason: "Speaker normalization via a linear transformation on a perceptual feature space and its benefits in asr adaptation", Proceedings of Eurospeech89, pp. 258–261 (1989).

[30] A. P. Varga and R. K. Moore: "Hidden markov model decomposition of speech and noise", Proc. of ICASSP, pp. 845–848 (1990).

[31] M. J. F. Gales and S. J. Young: "Robust continuous speech recognition using parallel model combination", IEEE Trans. on Speech and Audio Processing, **4**, pp. 231–239 (1996).

[32] F. Martin, K. Shikano, Y. Minami and Y. Okabe: "Recognition of noisy speech by composition of hidden markov models", IEICE Technical Report, **SP92-96**, (1992).

[33] S. Seneff: "Pitch and spectral estimation of speech based on an auditory synchrony model", PhD thesis, Massachusetts Institute of Technology (1985).

[34] M. J. Hunt and C. Lefèbvre: "Speech recognition using a cochlear model", Proc. of ICASSP, Vol. 3, pp. 1979–1982 (1986).

[35] P. Dermody, G. Raicevich and R. Katsch: "Comparative evaluations of auditory representation of speech", Visual Representations of Speech Signals (Eds. by M. Cooke and S. Beet), John Wiley & Sons, chapter 21, pp. 229–236 (1993).

[36] Y. Tohkura: "Human speech processing and its applications to engineering", J. Inst. Elec., Info. and Comm. Eng., **75**, 10, pp. 1038–1041 (1992). in Japanese.

[37] T. Umezaki and F. Itakura: "Speech analysis by group delay spectrum of all-pole filters and its application to the spectrum distance measure for speech recognition", Trans. of **IEICE, J72-D-II, no.8**, pp. 1141–1150 (1989). (in Japanese).

[38] T. Hirahara: "A nonlinear cochlear filter with adaptive Q circuits", J. Acoust. Soc. Jpn., **47**, 5, pp. 327–335 (1991). in Japanese.

[39] H. Singer, T. Umezaki and F. Itakura: "Low bit quantization of smoothed group delay spectrum for speech recognition", Proc. of ICASSP, Vol. 2, pp. 761–764 (1990).

[40] K. Obara and T. Hirahara: "Evaluation of auditory front-ends in DTW word recognition system", J. Acoust. Soc. Jpn., **50**, 6, pp. 452–464 (1994). (in Japanese).

[41] T. Umezaki, S. Harald and F. Itakura: "Evaluation of the smoothed group delay spectrum distance measure in speaker-independent speech recognition", Institute of Electronics, Information and Communication Engineers, **J74-A**, 4, pp. 610–618 (1991). in Japanese.

[42] R. P. Lippmann: "Speech perception by humans and machines", Proc. of Workshop on the Auditory Basis of Speech Perception, pp. 309–316 (1996).

[43] J. C. R. Licklider and I. Pollack: "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech", J. Acoust. Soc. Am., **20**, pp. 42–51 (1948).

[44] W. F. Sheppard: "On the application of the theory of error to the cases of normal distribution and normal correlation", Philos. Trans. R. Soc. London, **A192**, (1899).

[45] S. Kajita and F. Itakura: "SBCOR spectrum taking autocorrelation coefficients at integral multiples of 1/CF into account", Proc. of ICSLP, Vol. 3, pp. 1051–1054 (1994).

[46] S. Kajita, K. Takeda and F. Itakura: "Sbcor analysis using correlation coefficients at integral multiples of inverse of center frequency", J. Acoust. Soc. Jpn. (1997). (submitted).

[47] S. Kajita, D. Kobayashi, K. Takeda and F. Itakura: "Analysis of speech features included in human speech-like noise", J. Acoust. Soc. Jpn. (1997). in printing.

[48] T. Kobayashi, S. Itahashi, S. Hayamizu and T. Takezawa: "ASJ continuous speech corpus for research", J. Acoust. Soc. Jpn., **48**, pp. 888–893 (1992).

[49] M. D. Paez and T. H. Glisson: "Minimum mean squared-error quantization in speech ", IEEE Trans. Comm., **20**, pp. 225–230 (1972).

[50] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling: "Numerical Recipes in C", Cambridge University Press (1988).

[51] T. Miura Ed.: "Auditory System and Speech", The Institute of Electronics, Information and Communication Engineers (1980).

[52] S. Sagayama and F. Itakura: "On individuality in a dynamic measure of speech", Proceedings of the 1979 Spring Meeting of the Acoustical Society of Japan, pp. 589–590 (1979).

[53] S. Furui: "Cepstral analysis technique for automatic speaker verification", IEEE Transactions on Acoustics, Speech and Signal Processing, **ASSP-29**, pp. 254–272 (1981).

Appendix A

Analysis of Speech Features Included in Human Speech-Like Noise

A.1 Introduction

Humans listen various sounds coming from a lot of sound sources, and can distinguish speech sound only from the other sound. This suggests that the human auditory system can capture "speech features" from the acoustic signal to determine whether the sound is speech. If an objective measure of the speech features is developed, we will be able to apply it to speech interval determination, speech search from sound database and so on.

As an approach to investigate such speech features, this appendix analyzes the human speech-like (HSL) noise, which has been already introduced in Chapter 3 to evaluate the robustness of acoustic front-ends. HSL noise is a kind of the bubble noise and changes from just speech to stationary noise whose long-term averaged spectrum is the same as speech when a parameter is monotonically changed in generating HSL noise. This appendix investigates the change from speech to stationary noise through the following three points of view, (1) Gaussness of amplitude distribution, (2) temporal fluctuation of spectral fine structure, and (3) temporal fluctuation of spectral envelope.

This appendix is constructed as follows. At first, Section A.2 describes how to generate HSL noise, and quantifies the speech features included in HSL noise by subjective tests.

Section A.3 investigates the relation between speech features and Gaussness of amplitude distribution of HSL noise and its difference signal. In Section A.4, the relation between speech features and temporal fluctuation of spectral fine structure is investigated by evaluating HSL noises whose spectral envelope is flattened. Then, Section A.5 describes an objective evaluation for temporal fluctuations of spectral envelope. Finally, Section A.6 summarizes this appendix.

A.2 HSL Noise and Its Subjective Evaluations

A.2.1 How to Generate Human Speech-Like Noise

Human speech-like (HSL) noise is a superimposed signal of multiple independent speech signals. Let the number of superimposition be N, HSL noise signal $n_N[n]$ is N is defined as follows.

$$n_N[k] = \sum_{m=0}^{N-1} s[mK + k], \quad 0 \leq k \leq K - 1, \tag{A.1}$$

where

$s[n]$: a sample speech signal to be used for the generation.

k : time index

K : the length of HSL noise

This equation means that HSL noise is generated by circularly superimposing the sample speech signal $s[n]$ on the $n_N[k]$ with period K (See Figure A.1).

In this research, the sample speech signal $s[n]$ consists of the concatenation of 3,200 sentences uttered by 30 males and 34 females in the ASJ continuous speech corpus for research [48]. The sampling rate is 16kHz. The order in concatenating is random and each sentence data is normalized by its maximum amplitude (See Figure A.2).

A.2.2 Characteristics of HSL Noise

The waveforms and short-term FFT spectrograms of HSL noise are shown in Figure A.3. When the number of superimposition is small, the spectrum of HSL noise presents

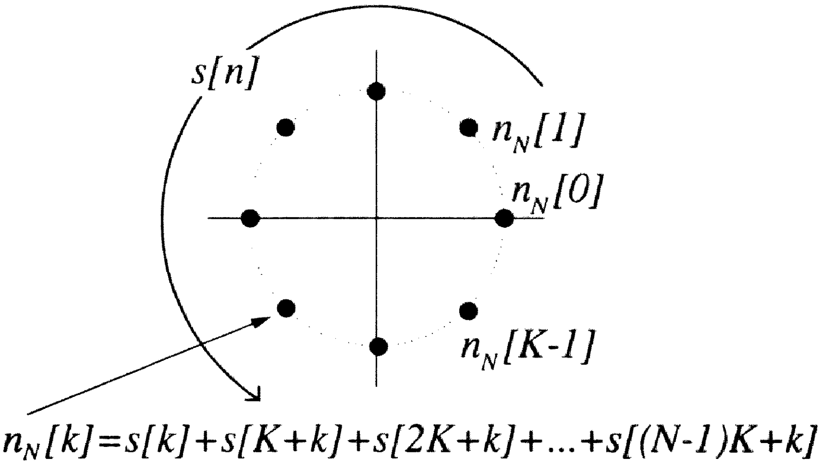


Figure A.1: The concept of HSL noise generation.

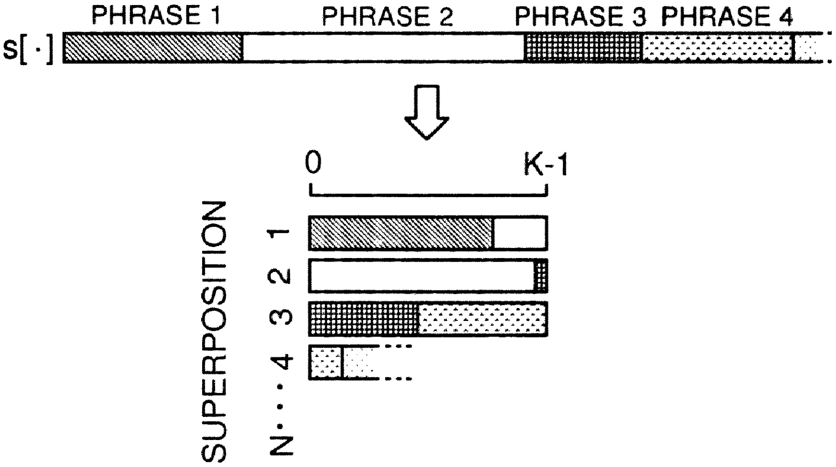


Figure A.2: The sample speech $s[n]$ is generated by the concatenation of 3,200 sentences uttered by 30 males and 34 females.

harmonics structure and formants. As the number of superimposition become large, however, such observations disappear, and at last, HSL noise becomes stationary noise. The auditory sensation of HSL noise is single speech when $N = 1$; overlapped speech of several talkers when N is moderate; and finally HSL noise becomes a stationary noise.

As described above, HSL noise has both the nature of speech and stationary noise. The peculiarity is that the speech feature included in HSL noise changes smoothly as the number of superimposition N increases.

Generally speaking, the slope of averaged long-term spectrum of male and female voice is about -10dB/oct and has a peak between 250Hz and 300Hz due to the pitch component. It suggests that HSL noise reflects the averaged long-term spectrum.

In the next section, the changes of speech feature for the number of superimposition will be quantified through a subjective evaluation experiment.

A.2.3 A Subjective Evaluation of Speech Feature in HSL Noise

Experimental Method

In the experiment, we had subjects listen to several HSL noises as shown in Table A.1 and forced them to select an answer from the criterion shown in Table A.2. In order to evaluate speech features present in each HSL noise, the obtained answers were used as the score for each HSL noise. Each presented HSL noise was normalized by the average power.

Experimental Results

The averaged scores of the subjective evaluation are shown in Figure A.5. the average scores that are close to one indicates speech-like, and those that are close to four are stationary noise-like.

As the results, the following three points are clarified:

1. HSL noise whose number of superimposition is from 2 to 10 is heard as a superim-

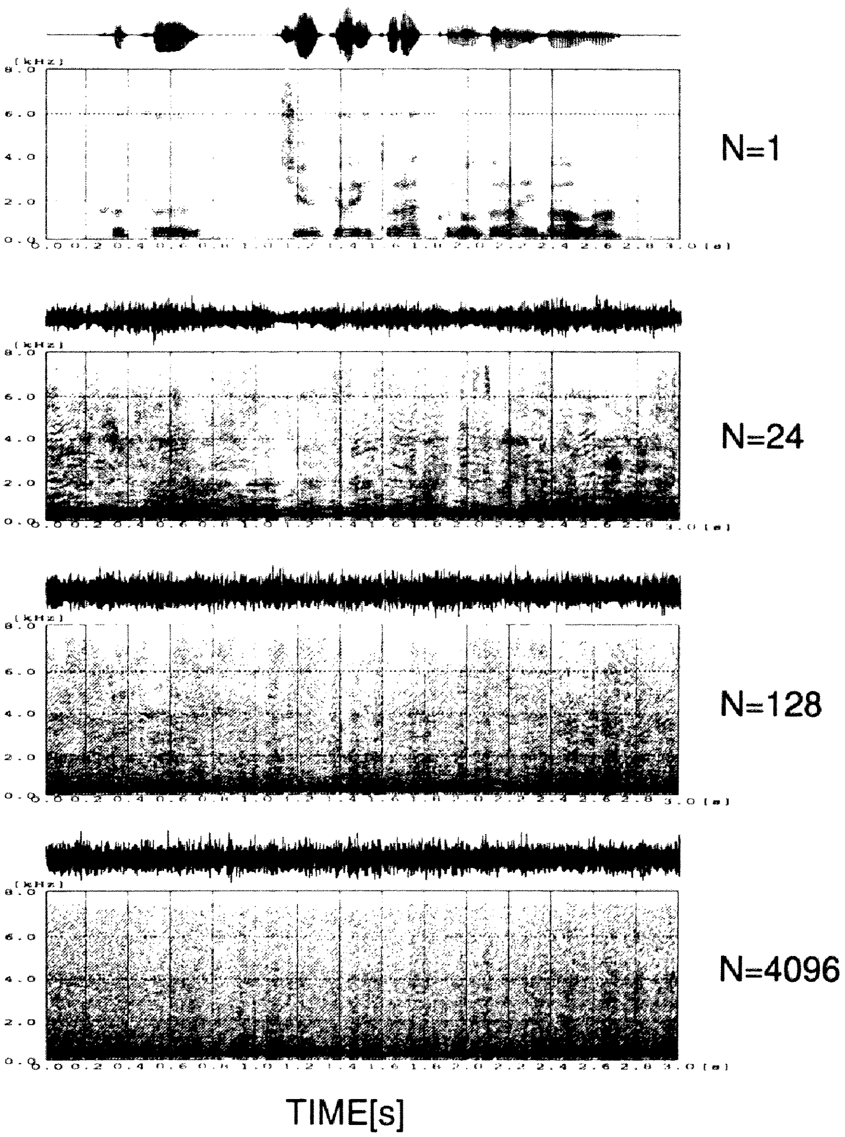


Figure A.3: The waveforms and short-term FFT spectrograms of HSL noise.

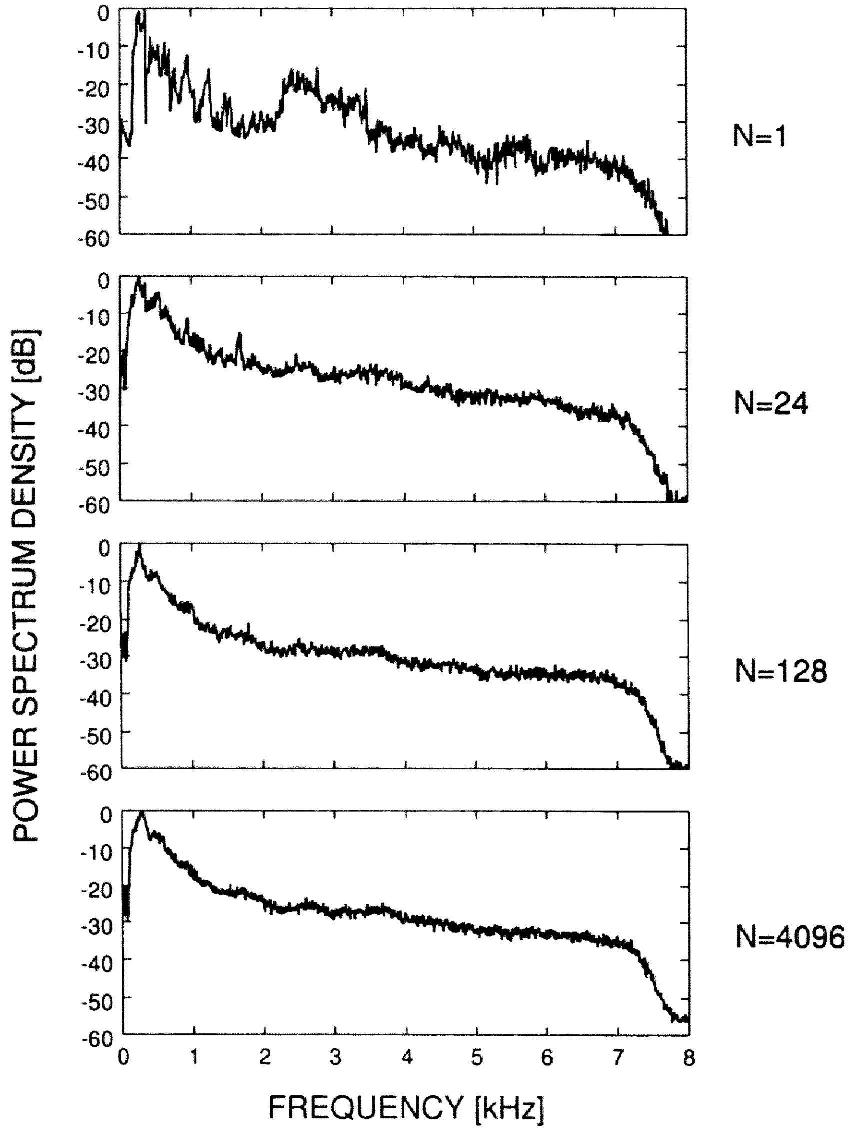


Figure A.4: long-term averaged spectrum of HSL noise estimated by Blackman-Tukey method.

Table A.1: Experimental condition for subjective test.

subject	10 person	
data	the number of superimposition (N)	2,4,6,8,10,16,24,32,64,128 256,512,1024,2048,4096
	length	1,3,10 sec
	sampling frequency	16kHz
	three signals every length and the number of superimposition	
present method	each signal is presented for left side ear using STAX SR (ATR version) headphone.	

Table A.2: Score and criterion.

score	criterion
1	perceived as superimposed speech only
2	perceived as a signal that is superimposed speech and stationary noise (mainly speech)
3	perceived as a signal that is superimposed speech and stationary noise (mainly noise)
4	perceived as a stationary noise

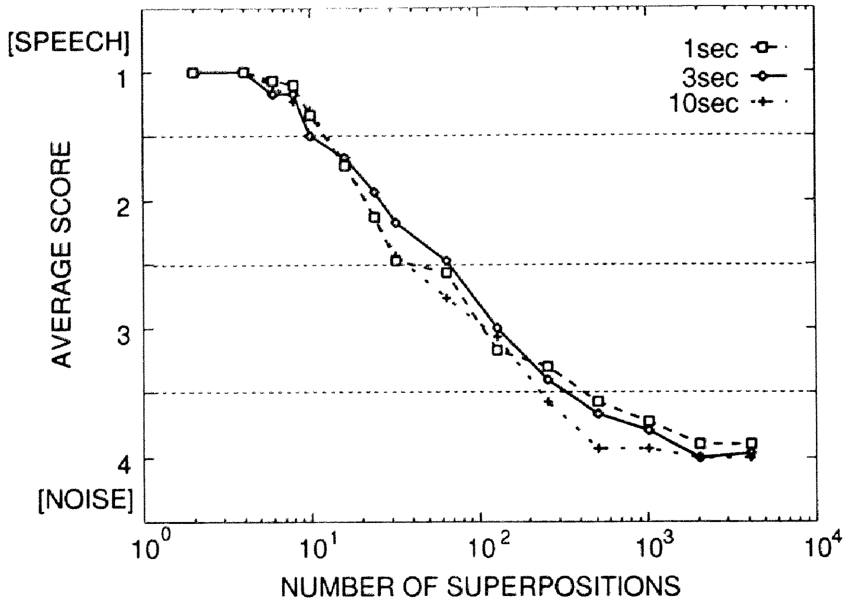


Figure A.5: Results of the subjective evaluation for HSL noise.

posed speech (score: 1.0-1.5)

2. HSL noise whose number of superimposition is from 16 to 256 has both of speech-like and stationary noise-like natures (score: 1.5-3.5)

3. HSL noise whose number of superimposition is from 512 to 4096 is almost heard as a stationary noise (score: 3.5-4.0)

From these results, we can conclude that the threshold for the stationary noise is about 256.

A.3 Gaussness for Amplitude Distribution of HSL Noise and Speech Features

Since the amplitude distribution is approximately a Gamma distribution[49], the distribution of HSL noise of a few superimpositions is also close to Gamma distribution. However, the amplitude distribution of HSL noise whose number of superimposition is large would be Gaussian distribution due to the central limit theorem (see Figure A.6). Therefore, it is reasonable to suppose that there would be some relation between the changes from Gamma to Gaussian distribution and speech features included in HSL noise.

In this section, we investigate speech features related to Gaussness of amplitude distribution using higher order statistics, i.e. skewness and kurtosis.

A.3.1 Experimental Methods

Skewness and Kurtosis

In general, given a data sequence $\boldsymbol{x} = \{x_1, x_2, \cdots, x_N\}$, the distribution of the data can be characterized by the following statistics[50]:

Skewness $\mathrm{S}(\boldsymbol{x})$ characterizes the degree of asymmetry of a distribution around its mean. The skewness is conventionally defined in such a way to make it non-dimensional. A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive $\boldsymbol{x}$; a negative value signifies a distribution whose tail extends out towards more negative $\boldsymbol{x}$. For speech signals, we calculate the absolute value in the following experiments because the sign does not have any meaning. The usual definition is

$$\mathrm{S}(\boldsymbol{x}) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{x_j - \overline{x}}{\sigma(\boldsymbol{x})}\right]^3, \tag{A.2}$$

where $\sigma(\boldsymbol{x})$ is the distribution's standard deviation.

Kurtosis $\mathrm{K}(\boldsymbol{x})$ measures the relative peakedness or flatness of a distribution related to Gaussian distribution. It is also a non-dimensional quantity. A distribution with positive kurtosis is termed leptokurtic and the one with negative kurtosis is termed platykurtic. The conventional definition of the kurtosis is

$$\mathrm{K}(\boldsymbol{x}) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{x_j - \overline{x}}{\sigma(\boldsymbol{x})}\right]^4 - 3, \tag{A.3}$$

where the -3 term makes the value zero for a Gaussian distribution.

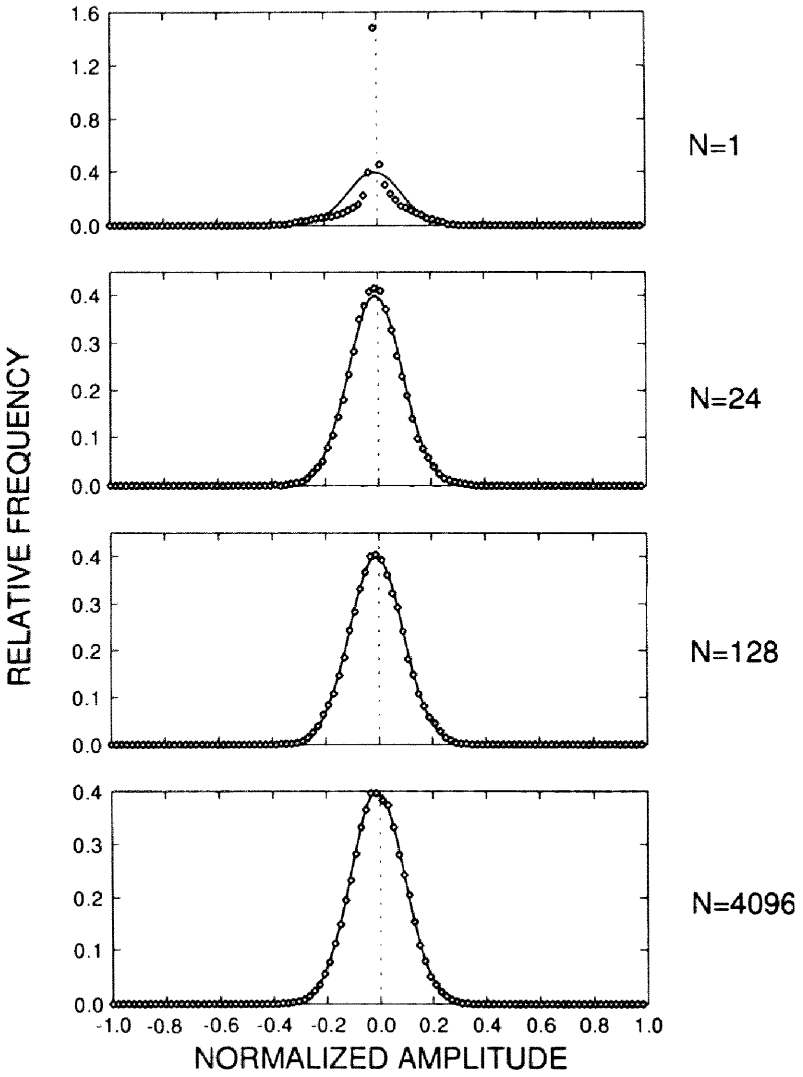


Figure A.6: Amplitude Distribution of HSL noise. The bold line represents a Gaussian distribution with the same variance of the sample data.

Table A.3: Data used in the experiment.

the number of superimposition (N)	1,2,4,8,16,32,64,128 256,512,1024,2048,4096
the length	3 sec
10 signals every the number of superimposition	

In the following experiments, the absolute value of skewness $\mid S(\boldsymbol{x}) \mid$ and the kurtosis $K(\boldsymbol{x})$ under the conditions shown in Table A.3 are calculated for HSL noises and its differenced signals.

Test method of Gaussness

For the idealized case of a Gaussian distribution, the standard deviation of $S(\boldsymbol{x})$ and $K(\boldsymbol{x})$ is approximately $\sigma(S(\boldsymbol{x})) = \sqrt{15/N}$, $\sigma(K(\boldsymbol{x})) = \sqrt{96/N}$ respectively[50]. Then, assuming that the distributions of skewness $S(\boldsymbol{x})$ and kurtosis $K(\boldsymbol{x})$ is approximately Gaussian, the hypothesis that the distribution of $\boldsymbol{x}$ is Gaussian is rejected with the significance level 0.01 if $\mid S(\boldsymbol{x}) \mid > 2.326\sigma(S(\boldsymbol{x}))$、$\mid K(\boldsymbol{x}) \mid > 2.326\sigma(K(\boldsymbol{x}))$. In this experiment, since $N = 48000$,

$$\mid S(\boldsymbol{x}) \mid \; > \; 0.0411 \tag{A.4}$$

$$\mid K(\boldsymbol{x}) \mid \; > \; 0.104. \tag{A.5}$$

When $S(\boldsymbol{x})$ and $K(\boldsymbol{x})$ are not included in this interval, we consider that the distribution is not Gaussian.

A.3.2 Experimental Results

Gaussness of Distribution for HSL noise

The values of the skewness and kurtosis for HSL noises and its differenced signals are shown in Figure A.7 and A.8 respectively. In the figures, the horizontal axis represents the superimposition number, ten points per superimposition represents each value for ten HSL noises generated from ten different speech samples $s[n]$. The bold line is the averaged value. In addition, the rejected intervals for the hypothesis that the distribution is Gaussian are illustrated by the hatched pattern. The magnified figures of Figures A.7 and A.8 are also presented in Figure A.9 to be enable easier viewing of the rejected intervals.

In the case of HSL noises, for the skewness, the hypothesis whether the distribution is Gaussian is rejected on the average for greater than 16 superimpositions. On the other hand, for the skewness, the hypothesis is rejected on an average of greater than 32 superimpositions. Therefore, it concludes that the amplitude distribution for HSL noises whose superimposition number are greater than 32 is Gaussian.

In the case of HSL noise's differenced signal, for the skewness, the hypothesis is rejected on an average of greater than 32 superimpositions. On the other hand, for the skewness, the hypothesis is rejected on an average of greater than 256 superimpositions. Therefore, it concludes that the amplitude distribution for HSL noises whose superimposition number is greater than 256 is Gaussian.

Furthermore, when both of the skewness and kurtosis are outside the rejected interval, the distribution is Gaussian. It suggests that the decision of Gaussness is enough to evaluate the kurtosis.

Changes of kurtosis and speech features

To compare with the subjective results obtained in section A.2, the averaged values of the kurtosis for HSL noises and its differenced signals are shown in Figure A.10. The changes

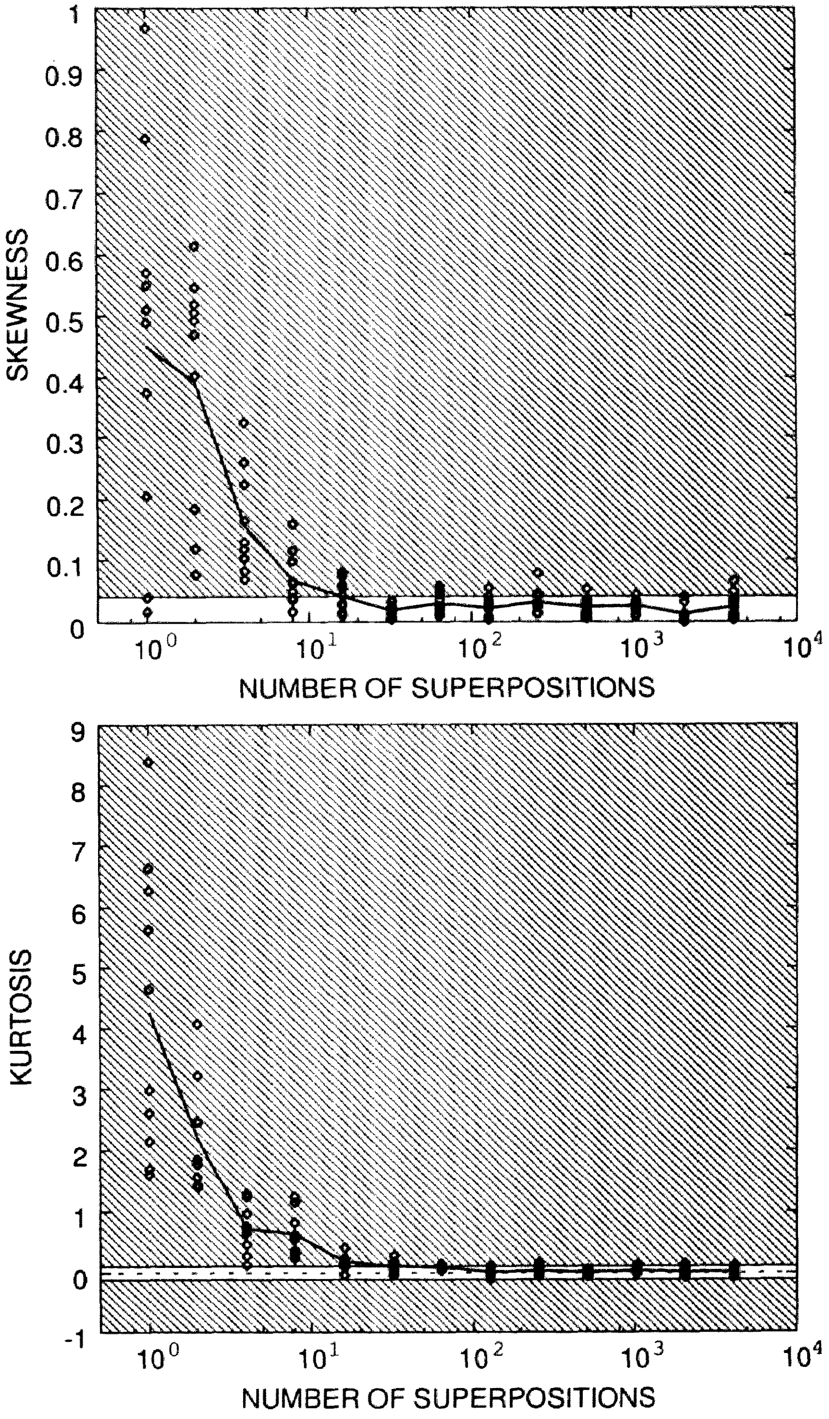


Figure A.7: The skewness (upper) and kurtosis (lower) for the amplitude distributions of HSL noises.

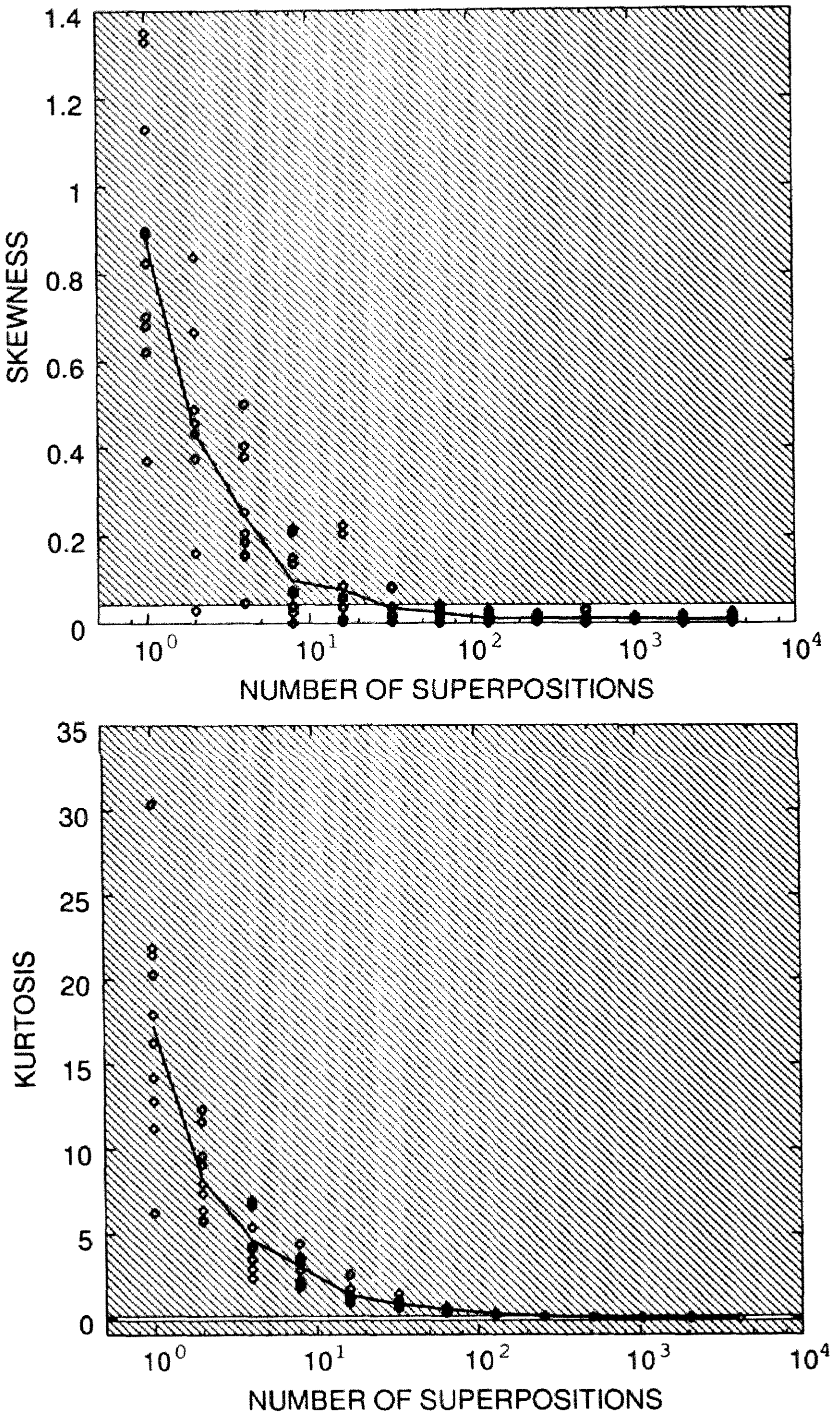


Figure A.8: The skewness (upper) and kurtosis (lower) for the amplitude distributions of HSL noise's differenced signals.

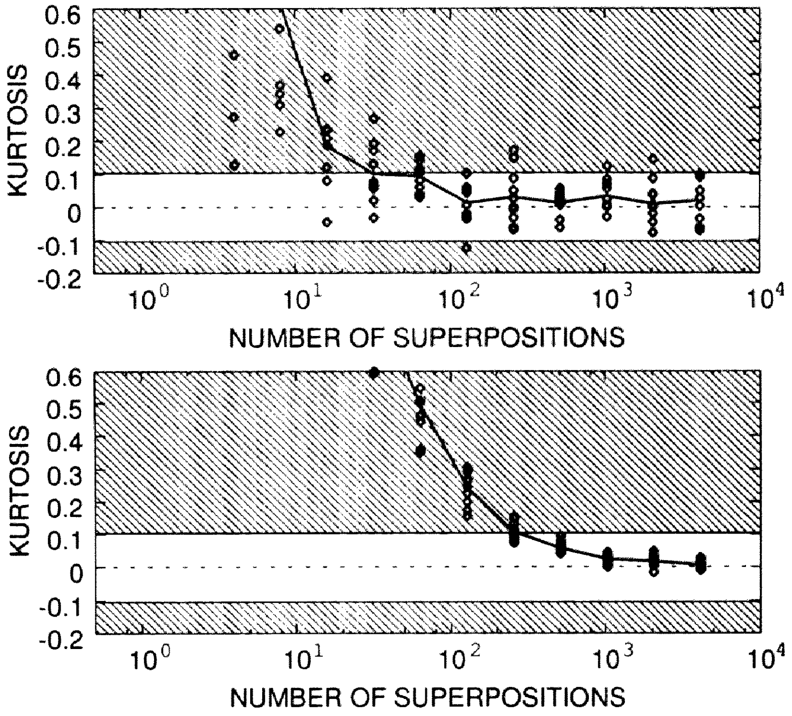


Figure A.9: The magnified kurtosis for the amplitude distributions of HSL noises (upper) and its differenced signals.

of the kurtosis for the differenced signal tend to be closer to the subjective results than the changes for HSL noises. Thus, the kurtosis for HSL noise's differenced signal can be an objective measure to evaluate speech features. This suggests that speech features could be included in the temporal changes of HSL noise.

A.4 Fine Structure of Spectrum in HSL Noise and Speech Features

As noted in section A.2, the slope of long-term averaged spectrum for male and female is about -10dB/oct and has a peak between 250Hz and 300Hz due to the pitch component [51]. The HSL noise that is superimposed by generating a lot of speech also has the long-term averaged spectrum as well as the one of typical speech. As seen in Figure A.4, the spectral envelope of HSL noise is closer to the long-term averaged spectrum of speech as the superimposition number is increased. Therefore, it seems that there could be some relation between the changes of spectral envelope and speech features.

In this section, by eliminating such spectral envelope using linear predictive (LP) analysis, we investigate the relationship between speech features and temporal fluctuation of spectral fine structure. The reason of the separation is to restrict the target of investigations for speech features.

A.4.1 Experimental Methods

The following three signals calculated from HSL noises whose number of superimposition are from 32 to 256 were used in the subjective test.

signal (1) is obtained by eliminating the long-term averaged spectral envelope using LP analysis. In other words, it is the residual signal obtained by 32 order LP analysis.

signal (2) is obtained by eliminating the short-term spectral envelope. It is an overlap-and-added signal of the residual signal of HSL noise using 32 order LP analysis in a 30 ms long frame shifted by 10 ms long.

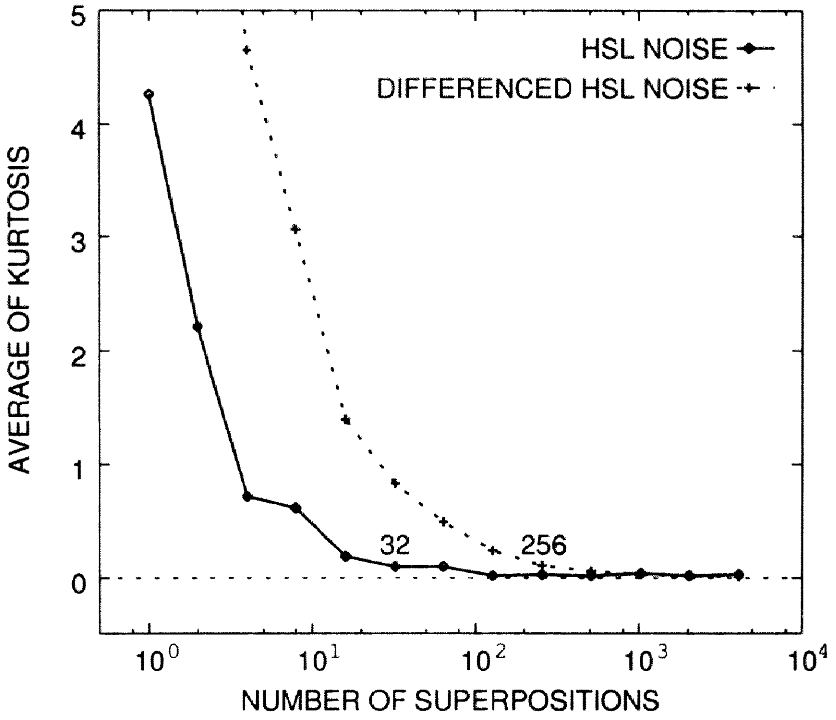


Figure A.10: The averaged kurtosis for HSL noises (bold line) and its differenced signals (broken line).

Table A.4: Data and experimental conditions.

subject	10 person	
data	the number of superimposition (N)	32, 64, 128, 256
	length	3 sec
	sampling frequency	16kHz
	three signal every (1),(2),(3)	
present method	each signal is presented for left side ear using STAX SR (ATR version) headphone in a sound proof room.	

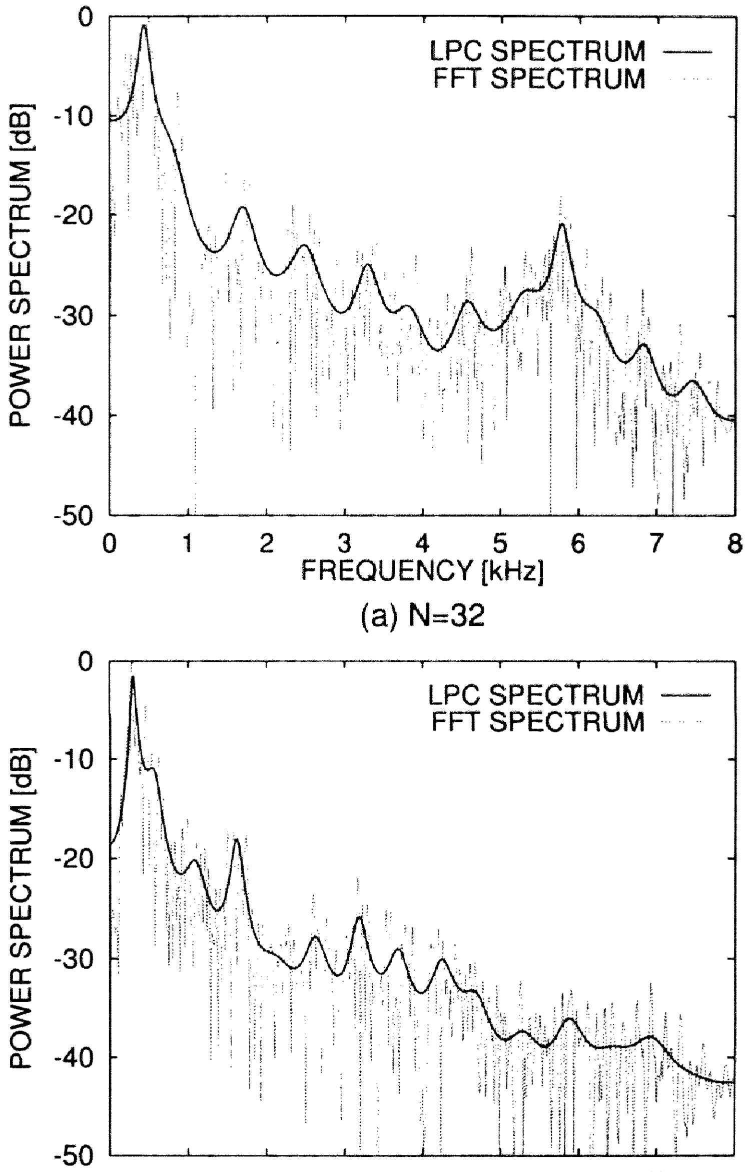
signal (3) is obtained by eliminating the changes of short-term power in order to left the short-term spectral change. It was calculated by normalizing the averaged power in 30 ms long frame every 10 ms.

For HSL noises whose superimposition numbers are greater than 32, the spectral envelope is more complicate than that of the usual speech signal due to the superimposition of a lot of speech uttered by multiple speakers (see Figure A.11). In order to fit such a complicated spectral envelope using an all-pole filter, we selected relatively higher 32 order than the one used in typical LP analysis for speech.

Each presented signal was normalized by its average power, quantized with 16 bit, and recorded in DAT. The same subjective test as in the previous section was performed. The specific experimental condition is shown in Table A.4.

A.4.2 Experimental Results

Figure A.12 shows the averaged subjective scores. It indicates that signal (1) and (3) lose a little speech feature compared with the original HSL noise, but follow the change of speech feature as in the HSL noise. On the other hand, signal (2) almost loses speech feature in spite of the number of superimposition. Therefore, speech feature in HSL noise



(a) N=32

(b) N=256

Figure A.11: FFT spectrum and LPC spectrum for HSL noises whose number of superimposition is 32 and 256 (the analysis condition for LP analysis is the same in calculating the signal (2), and the FFT points is 1024).

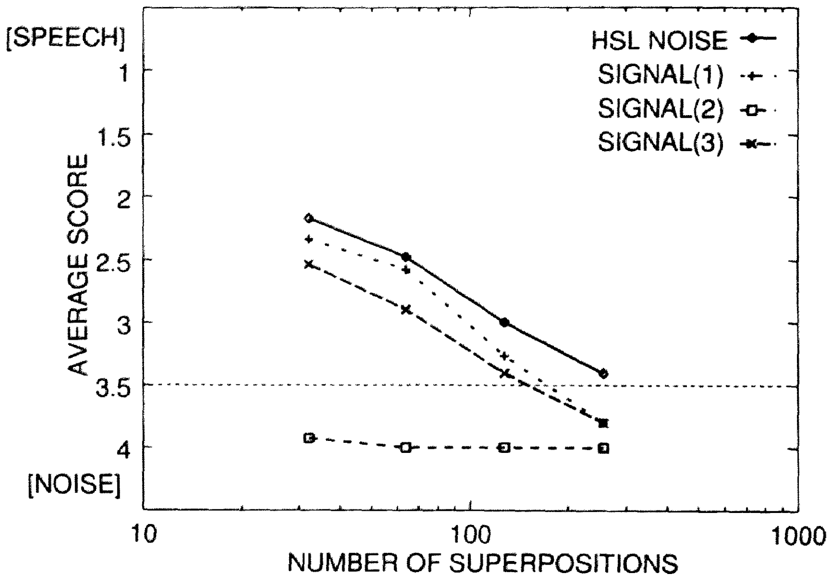


Figure A.12: Averaged subjective score. The averaged score for HSL noise obtained in Section A.2 is shown again.

is

1. not included in the averaged spectral envelope (from the results for signal (1)),
2. not related to the change of the short-term power (from the results for signal (3)),
3. is significantly characterized by the change of the short-term spectrum (from the results for signal (2)).

The significance of such change of short-term spectrum is suggested by the research for syllable perception and phoneme segmentation, and these experimental results support the fact.

A.5 Temporal Changes of Spectral Envelope in HSL Noise and Speech Features

As shown in the previous section, speech feature included in HSL noise is characterized by short-term spectral change. In this section, the short-term spectral change is quantified

Table A.5: Analysis condition.

frame length	30ms
frame shift	10ms
analysis window	Hamming window
LPC order	32 order
cepstrum order	34 order
window length in calculating delta cepstrum	3,4,5,7 frames

by dynamic measure [52] using delta cepstrum, and speech feature present in HSL noise is evaluated. Then, the results for the three signals used in the previous section are compared with the one of dynamic measure.

A.5.1 Dynamic Measure

Dynamic measure using delta cepstrum [53] is defined as follows [52]:

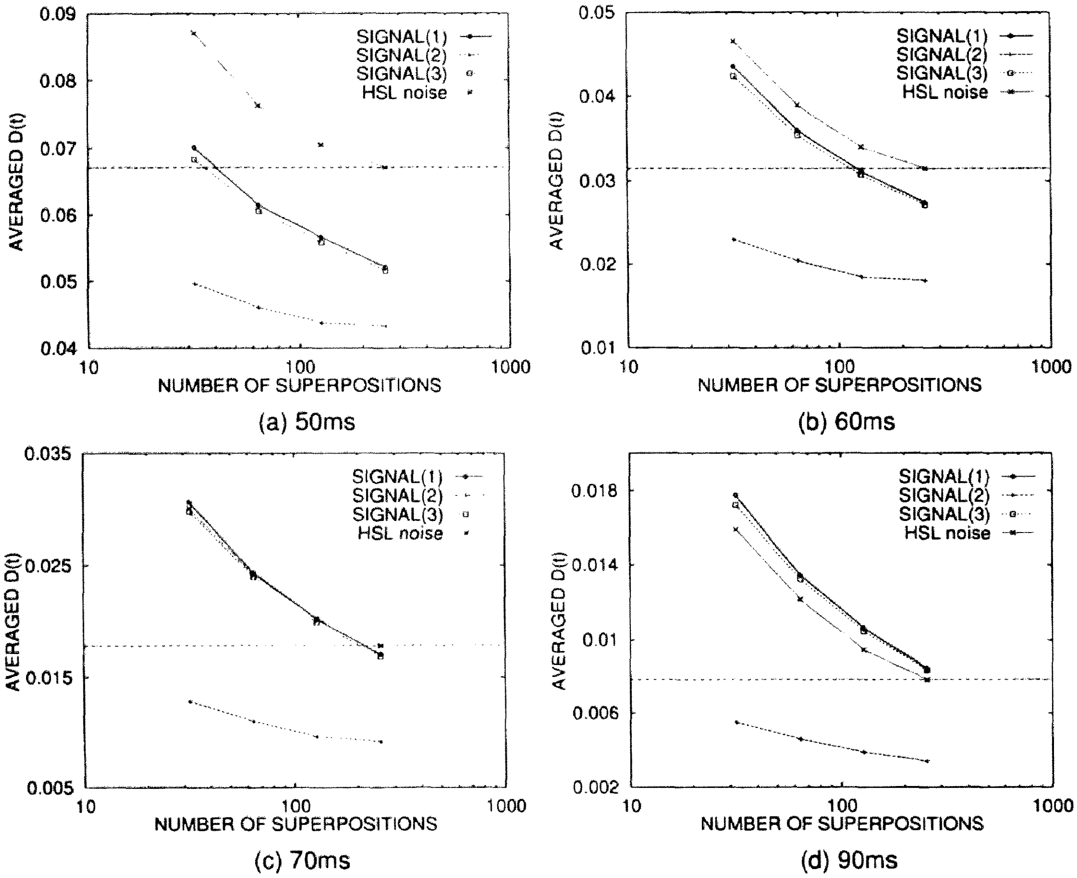
$$D[n] = \sum_{i=1}^{q} (\Delta c_i[n])^2. \tag{A.6}$$

where, $\Delta c_i[n]$ is ith delta cepstrum at nth analysis frame, and q is the order of cepstrum analysis. This dynamic measure $D[n]$ is calculated by the square sum of delta cepstrum $\Delta c_i[n]$ along with the quefrency. Thus, dynamic measure represents the change of log spectrum [52].

Using averaged dynamic measure $\overline{D[n]}$ by the total number of analysis frame, the short-term spectral change for three signals used in Section A.4 is evaluated. The analysis condition of dynamic measure is shown in Table A.5.

A.5.2 Experimental Results

The $\overline{D[n]}$s for HSL noise(N:256) and for the three signals used in Section A.4 are shown in Figure A.13.



(a) 50ms

(b) 60ms

(c) 70ms

(d) 90ms

Figure A.13: The $\overline{D[n]}$s for HSL noise (N:256) and three signals used in Section A.4

From the previous section, signals (1) and (3) whose number of superimposition is 128, and HSL noise whose number of superimposition is 256 are comparable for the subjective score (see Figure A.12). Noting this relation in evaluating the experimental results, $\overline{D[n]}$ for signals (1) and (3) whose number of superimposition is 128, and HSL noise whose number of superimposition is 256 is almost same, when the length of delta cepstrum window is 60 ms. Thus, it indicates that the change for the auditory sensation resulted in the previous section coincides with the change of short-term spectral over 60 ms interval. Therefore, we can conclude that it is possible to evaluate speech feature objectively using a metric that measures the change of short-term spectrum such as dynamic measure.

A.6 Conclusions

In this appendix, we introduced HSL noise and analyzed the speech feature present in HSL noise.

At first, we clarified that the discriminative threshold between speech and stationary noise is about 256. Then, we investigated the change from speech to stationary noise from three points of view, i.e., (1) Gaussness of amplitude distribution, (2) temporal fluctuation of spectral fine structure, and (3) temporal fluctuation of spectral envelope. These experiments clarified that the change of the kurtosis for HSL noise's differenced signal and the change of short-term spectrum in the 60 ms interval coincide with the change of sensation for HSL noise as the number of superimposition is increased. They indicate the significance of dynamic change of speech to discriminate speech and noise.

However, in order to apply these results to speech detection problem and speech/non-speech discrimination problem, the feature of spectral change for speech should be investigated in more detail, because the speech feature evaluated in this appendix is restricted for the discrimination between speech and stationary noise. There are a lot of sound whose spectrum changes dynamically in short-term.

List of Publications

Journal Papers

1. S. Kajita and F. Itakura, "Speech Analysis and Speech Recognition Using Subband-Autocorrelation Analysis", *Journal of Acoustical Society of Japan (E)*, vol.15, 5, pp.329-338 (1994)

2. S. Kajita, D. Kobayashi, K. Takeda and F. Itakura, "Analysis of Speech Features Included in Human Speech-Like Noise", *Journal of Acoustical Society of Japan*, vol.53, 5, pp.337-345 (1997) (in Japanese)

3. S. Kajita, K. Takeda and F. Itakura, "Noise Robust Speech Recognition Using Subband-Autocorrelation Analysis with Autocorrelation Coefficients at Integral Multiples of 1/CF", *Journal of Acoustical Society of Japan*, in printing (1997) (in Japanese)

4. S. Kajita, K. Takeda and F. Itakura, "Noise Robust Speech Recognition Using Subband-Crosscorrelation Analysis", submitted to *IEICE Transactions on Information and Systems* (1997)

International Conferences

1. S. Kajita and F. Itakura, "Subband-Autocorrelation Analysis and Its Application for Speech Recognition", *In Proceedings of the 1994 IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP '94)*, Adelaide, Australia, Vol.II, pp.193-196 (1994)

2. S. Kajita and F. Itakura, "SBCOR Spectrum Taking Autocorrelation Coefficients at Integral Multiples of 1/CF into Account", *Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP'94)*, Yokohama, Japan, Vol.3, pp.1051-1054 (1994)

3. S. Kajita and F. Itakura, "Robust Speech Feature Extraction Using SBCOR Analysis", *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP '95)*, Detroit, U.S.A., Vol.1, pp.421-424 (1995)

4. S. Kajita, K. Takeda and F. Itakura, "Subband-Crosscorrelation Analysis for Robust Speech Recognition", *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, U.S.A., Vol.1, pp.422-425 (1996)

5. D. Kobayashi, S. Kajita, K. Takeda and F. Itakura, "Extracting Speech Features from Human Speech Like Noise", *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, U.S.A., Vol.1, pp.418-421 (1996)

6. S. Kajita, K. Takeda and F. Itakura, "A Binaural Speech Processing Method Using Subband-Crosscorrelation Analysis for Noise Robust Recognition", *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP '97)*, Munich, Germany, Vol.II, pp.1243-1246 (1997)

7. S. Kajita, K. Takeda and F. Itakura, "Spectral Weighting of SBCOR for Noise Robust Speech Recognition", submitted to *the 1998 IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP '98)*, Seattle, USA (1998)

Annual Meetings

1. S. Kajita and F. Itakura, "Noisy speech analysis and recognition using synchrony spectrum", *Proceedings of the 1991 Autumn Meeting of the Acoustical Society of Japan*, pp.405-406 (1991) (in Japanese)

2. S. Kajita and F. Itakura, "Frequency resolution of Subband-Autocorrelation Analysis", *Proceedings of the 1992 Spring Meeting of the Acoustical Society of Japan*, pp.393-394 (1992) (in Japanese)

3. S. Kajita and F.Itakura, "Speech Recognition Using Auditory Subband-Autocorrelation Analysis", *Proceedings of the 1993 Spring Meeting of the Acoustical Society of Japan*, pp.59-60 (1993) (in Japanese)

4. S. Kajita and F. Itakura, "Investigation of SBCOR Spectrum with Robustness for Noisy Environments", *Proceedings of the 1993 Autumn Meeting of the Acoustical Society of Japan*, pp.145-146 (1993) (in Japanese)

5. S. Kajita and F. Itakura, "HMM Phoneme Recognition Using SBCOR Spectrum", *Proceedings of the 1994 Autumn Meeting of the Acoustical Society of Japan*, pp.437-438 (1994) (in Japanese)

6. S. Kajita and F. Itakura, "Evaluation of SBCOR Spectrum against Additive Noises ", *Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan*, pp.7-8 (1995) (in Japanese)

7. S. Kajita and F. Itakura, "An Investigation of SBCOR Analysis Based on Cross-correlation between Two Input Signals", *Proceedings of the 1995 Autumn Meeting of the Acoustical Society of Japan*, pp.149-150 (1995) (in Japanese)

8. D. Kobayashi, S. Kajita, K. Takeda and F. Itakura, "Extracting Speech Features from Human Speech-like Noise", *In Proceedings of the 1996 Spring Meeting of the Acoustical Society of Japan*, pp.89-90 (1996) (in Japanese)

9. S. Kajita, K. Takeda and F. Itakura, "Robust Speech Recognition Using Subband-Crosscorrelation Analysis", *In Proceedings of the 1996 Autumn Meeting of the Acoustical Society of Japan*, pp.135-136 (1996) (in Japanese)

10. S. Kajita, K. Takeda and F. Itakura, "Significance of Periodicity Information for Noise Robust Speech Analysis", *In Proceedings of the 1997 Spring Meeting of the Acoustical Society of Japan*, pp.135-136 (1997) (in Japanese)

11. S. Kajita, K. Takeda and F. Itakura, "Significance of Lateral Inhibition Processing for Noise Robust Speech Analysis", *Proceedings of the 1997 Autumn Meeting of the Acoustical Society of Japan*, pp.7-8 (1997) (in Japanese)

Technical Meetings

1. S. Kajita and F. Itakura, "Speech Recognition Using Synchrony Spectrum", *IEICE Technical Report*, **EA91-4**, pp.1-8 (1991) (in Japanese)

2. S. Kajita and F. Itakura, "Speech Processing using Subband-Autocorrelation Analysis", *IEICE Technical Report*, **SP92-41**, pp.15-22 (1992) (in Japanese)

3. S. Kajita and F. Itakura, "Speech Processing Using Auditory Subband-Autocorrelation Analysis", *IEICE Technical Report*, **SP92-148**, pp.9-16 (1993) (in Japanese)

4. S. Kajita and F. Itakura, "SBCOR Spectrum Taking Autocorrelation Coefficients at Integer Multiples of CF^{-1} into Account", *IEICE Technical Report*, **SP93-64**, pp.1-9 (1993) (in Japanese)

5. S. Kajita and F. Itakura, "Speech Recognition Using Subband-Autocorrelation Spectrum", *IEICE Technical Report*, **SP93-80**, pp.29-34 (1993)

6. S. Kajita and F. Itakura, "Robustness of SBCOR Analysis against Waveform Distortion and Additive Noises", *IEICE Technical Report*, **SP94-117**, pp.69-76 (1995) (in Japanese)

7. D. Kobayashi, S. Kajita, K. Takeda and F. Itakura, "Extracting Speech Features from Human Speech-like Noise", *IEICE Technical Report*, **SP95-105**, pp.85-92 (1995) (in Japanese)

8. S. Kajita, K. Takeda and F. Itakura, "Speech Recognition Using Binaural Subband-Crosscorrelation Analysis", *IEICE Technical Report*, **SP96-47**, pp.39-46 (1996) (in Japanese)

9. S. Kajita, K. Takeda and F. Itakura, "Speech Processing Using Subband-Crosscorrelation Analysis Analysis", *Trans. Tech. Com. Psycho. Physio. Acoust.*, **H-96-76** (1996) (in Japanese)

Other Meetings and Reports

1. S. Kajita, K. Takeda and F. Itakura, "Investigation of Subband-crosscorrelation Analysis for Speech Recognition under Noisy Conditions", *Proceedings of the 1996 Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan*, pp.1123-1128 (1996)

2. S. Kajita and F. Itakura, "Speech Recognition Using Synchrony Spectrum", *Record of 1991 Tokai-Section Joint Conference of the Six Institutes of Electrical and Related Engineers*, pp.436 (1991) (in Japanese)

3. S. Kajita, K. Takeda and F. Itakura, "SBCOR Analysis based on crosscorrelation between two channel signals", *Record of 1995 Tokai-Section Joint Conference of the Six Institutes of Electrical and Related Engineers*, pp.258 (1995) (in Japanese)

4. S. Kajita, K. Takeda and F. Itakura, "Robust Speech Recognition using SBXCOR with crosscorrelation coefficients at integral multiples of CF^{-1}", *Record of 1996 Tokai-Section Joint Conference of the Six Institutes of Electrical and Related Engineers*, pp.249 (1996) (in Japanese)