Inter-aural Speech Spectrum Representation by Spatio-Temporal Masking Pattern

Tatsuya Hirahara

ATR Auditory and Visual Perception Research Laboratories

Twin 21 Bldg. MID Tower, 2-1-61 Shiromi, Higashi-ku, Osaka, 540 Japan

ABSTRACT

In this paper, several speech sounds are examined by a masking method to show typical examples of speech spectrum in the auditory pathway represented by a spatio-temporal masking pattern and to clarify differences between interaural and physical representation of speech spectrum. Three types of Japanese speech, monosyllables, continuous speech and a monosyllable reproduced time reversely, are chosen for masker sounds. Using 1/3 octave band noise bursts with 25msec. duration as maskees, simultaneous and temporal masking are measured for the whole period of each masker. Spatio-temporal masking patterns thus obtained are an inter-aural speech spectrum. Compared with the physical spectral pattern: speech onsets and the formant structure, in particular, the transition of formants are emphasized and represented prominent in the masking patterns. These spectral emphases in the auditory pathway are composed of three functions, AM/FM masking, forward/backward masking, and adaptation. Further, taking into account the considerable differences between inter-aural and physical representation of speech spectrum, the inter-aural spectrum can be implemented as better representation of speech spectrum in speech feature extraction and speech signal processing by computers.

INTRODUCTION

Spectrum analysis in the human auditory system is performed by cochlear function and neural network processing. These characteristics are assumed to be different from those of spectral analysis techniques based upon digital signal processing we usually use.

A number of psychophysical and neuro-physiological studies have been carried out to date to obtain knowledge on this auditory spectral analysis characteristics[1,2,3]. These studies indicate that the auditory system has its own signal processing functions such as, critical band filtering, lateral inhibition, adaptation, saturation, combination tones generation, masking and so on. Therefore, the inter-aural spectrum, i.e. sound spectrum representation in the auditory pathway, is different from the physical spectrum. Also, the remarkable abilities of the human auditory system to detect, separate, and recognize speech sounds are assumed to be performed using these inter-aural spectrum as input data for higher level signal processing. Therefore, inter-aural spectrum is superior to the physical spectrum representation when discussing perceptual cues of speech sounds.

From this standpoint, recent efforts have been made to develop a speech analysis method based on auditory functions. Several researchers have reported studies that simulate some auditory functions and a number of them have tried to apply their results, in part, to the field of automatic machine speech recognition [4,5,6,7,8,9,11].

Very few reports, however, have been given on studies concerned with inter-aural representation of dynamically varying and/or complex structured sound, such as speech [10,12,13,14,15]. It is the purpose of this paper to observe speech sounds from the viewpoint of spatio-temporal masking pattern, and show typical examples of speech sound representation in the auditory pathway. Differences between inter-aural spectrum and physical spectrum representation of speech are also clarified.

METHODS

Basically, two methods have been used to measure inter-aural spectral patterns. One is a neuro- physiological method, by which activities of the auditory nerve fibers measured directly correspond to sound stimuli inputs [17]; however, this method can not be used to study human auditory system. Another is a psychophysical method, by which activities of auditory system are measured indirectly. Three major psychophysical methods used to measure peripheral activity are the masking method [10,16], the pulsation threshold method [19] and the cancelling method [18]. In this paper, two traditional masking methods, temporal and simultaneous masking methods, were chosen since they are most appropriate for measuring inter-aural spectral representation of speech sound of wide range, time-varying spectral dynamics.

A masking value $M(m;t,s)$ is defined as the threshold shift of maskee signal s overlapped with masker sound m at time t. from masker onset. That is,

$$M(m;t,s) = L(m;t,s) - L(s) \quad [\mathrm{dB}] \qquad (1)$$

where $L(m;t,s)$ and $L(s)$ are the hearing threshold level of maskee signal s with and without masker sound m. present at the time t. When the maskee signal s is a function of frequency f, $M(m;t,s)$ is also a function of frequency f. Therefore, a three dimensional masking pattern for the masker sound can be obtained by measuring $L(m;t,s(f))$ at various t and f. This three dimentional masking pattern is considered to be an inter-aural spectrum representation of a masker sound after peripheral auditory processing.

EXPERIMENTS

Three experiments were carried out. Maskers were different types of speech sounds, while maskee signals and experimental procedure remained the same throughout the experiments.

Masker Experiment I : Japanese monosyllables */e/*, */re/*, */be/* and */de/* of 300 to 400 msec. duration were chosen for maskers. Experiment II : A continuous sentence speech */Are dewa eberesutoni noborenai/* (He can not climb Mt.Everest.) was chosen for the masker. This sentence was selected because it included the monosyllables */e/*, */re/*, */be/* and */de/*. The sentence duration is 1.6 seconds. Experiment III : Reversally reproduced monosyllable */re/* was chosen for the masker to investigate how the time axis, inverse of the masker, affected the masking pattern. These speech samples were uttered by a male speaker in a soundproof room. Their average fundamental frequency was about 100Hz.

Maskee Maskee signals were sixteen 1/3 octave band noise bursts of 25 msec. duration with a linear rise and fall time of 5 msec. Their center frequencies f_c covered 100Hz to 4kHz.

Setup Experimental setups and the time chart of the stimuli are shown in Fig.1. The masker and maskee were D/A converted simultaneously via different channels. Both of them were low-pass-filtered(Fc=5kHz,-96dB/Oct.), individually attenuated to a certain level, mixed together, then presented to a subject monauraly through headphones (STAX SR-5) in a soundproof room. The presented level of masker was fixed at 70dBSPL.

Procedure Every threshold value was determined by the method of limits. At the beginning of the experiment, the maskee level was set below the threshold. Subjects were instructed to judge whether or not the maskee signal could be heard with the masker sound for each presented stimulus by pushing 'Yes' or 'No' button on the switch box. Every time the 'No' button was pushed, the system increased the maskee level by 1 or 2dB automatically. The maskee level gave the threshold value when the 'Yes' button was pushed for the first time. To allow a judgement to be made correctly and easily, subjects were allowed to use two additional buttons: 'Again' to repeat the same stimuli, and 'Check' to repeat the masker sound only.

Two well trained male subjects participated in the experiments. Measurements were repeated at least 3 times for every threshold $L(m;t,s)$ and at least 10 times for every $L(s)$ on different days for each subject.

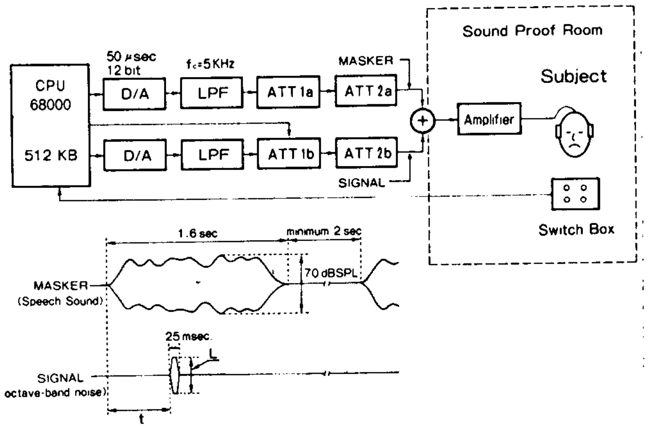

Fig.1 Experimental setups and the time chart of the stimuli. Both masker sound and maskee signal are D/A converted (20kHz, 12bits), low-pass-filtered (Fc=5kHz,-96dB/oct.), individually attenuated, mixted togather, then presented to a subject monauraly.

RESULTS

The sound spectrogram and speech waveform of the monosyllable masker /de/ are shown in Fig.2 (a) and (b), respectively. In Fig.2 (c), the spatio-temporal masking pattern measured every 25 msec. for this masker is depicted. The fine spectral structures of the masker sound, in particular, the first formant transition (e.g. at t = 125 to 175 msec.) and the vowel formant structure (e.g. at t = 175 to 300 msec.), are clearly observed in the masking pattern.

Figure 3 shows masking spectra and 1/3 octave-band spectra for */de/* at t = 250 msec. Solid lines represent masking spectra, i.e. the inter-aural spectra, and broken lines represent 1/3 octave-band power spectra, i.e. the physical spectra. Thick and thin solid lines represent differences between the two subjects. In Fig.3, the first formant (F1) in the masking spectra appears more prominent than that in the power spectra since masking values in the lower frequency region were small.

Figure 4 (a)-(c) show masking patterns and a 1/3 octave band power spectral patterns for the masker sound */de/* as a function of time. When compared with the power spectral pattern, three distinctive characteristics are observed in the masking pattern. (1) Masking does not take the value of 0 dB at the time before the beginning (t = -25 msec.) and after the end (t = 400 msec.) of masker speech. (2) Masking value increases remarkably at speech onset (t = 25 msec.) and at the transitional part of the formant. (3) Masking value decreases gradually in the vowel part. These characteristics were commonly observed in each masking pattern measured with respect to other monosyllable masker sounds.

The spectrogram and the speech waveform of the continuous speech masker are shown in Fig.5 (a) and (b). A spatio-temporal masking pattern measured every 25 msec. for this masker is depicted in Fig.5 (c). Formant structures and formant transitions are clearly represented in Fig.5 (c) as well as Fig.2(c).

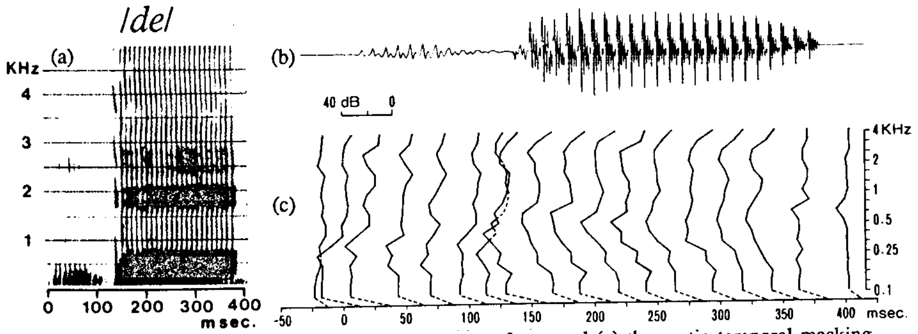

Fig.2 (a) Wide band spectrogram, (b) speech waveform, and (c) the spatio-temporal masking pattern measured every 25 msec. for the monosyllable masker /de/.

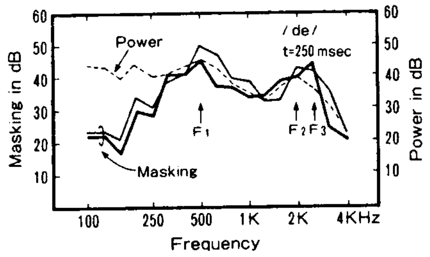

Fig.3 The masking spectra (solid lines) and 1/3 octave band power spectra (broken line) for /de/ at t=250 msec. Thick and thin lines represent differences between the two subjects.

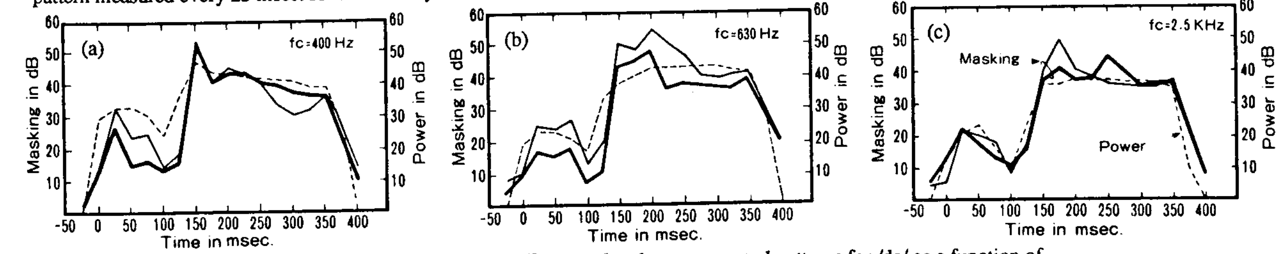

Fig. 4 Masking patterns and 1/3 octave band power spectral patterns for /de/ as a function of time at three frequency bands: (a) fc=400Hz, (b) fc=630Hz and (c) fc=2.5kHz.

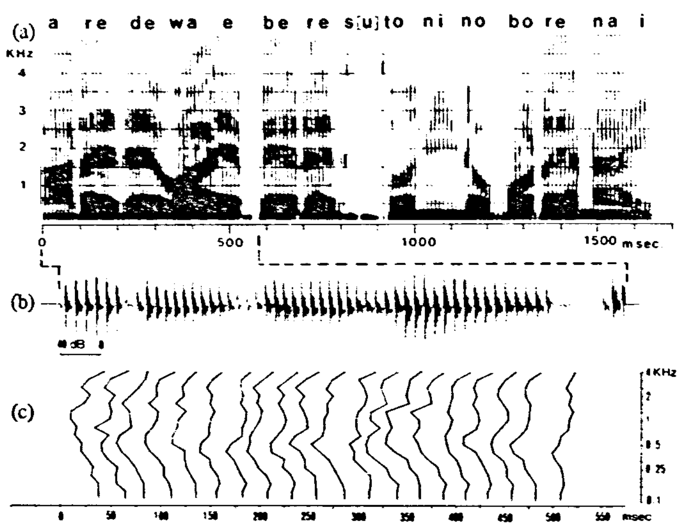

Fig.5 (a) Wide band spectrogram , (b) speech waveform and (c) the spatio-temporal masking pattern measured every 25 msec. for the continuous sentence speech */aredewa eberesutoni noborenai/*.

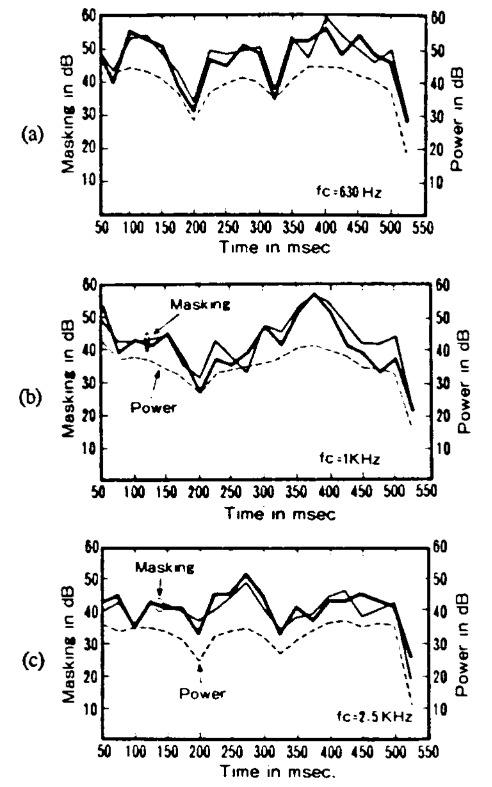

Fig.6 Masking patterns (solid lines) and 1/3 octave band power spectral patterns (bloken lines) for the continuous speech as a function of time at three frequency bands: (a) fc=630Hz, (b) fc=1kHz and (c) fc=2.5kHz. Thick and thin lines represent differences between two subjects.

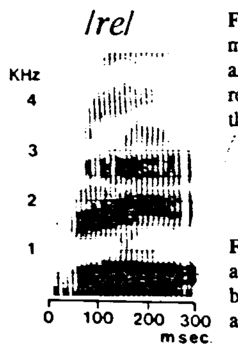

Fig.7 Wide band spectrogram of the monosyllable masker sound /re/ in normal time axis. This monosyllable and a time reversally reproduced one (reversal /re/) are the maskers in the third experiment.

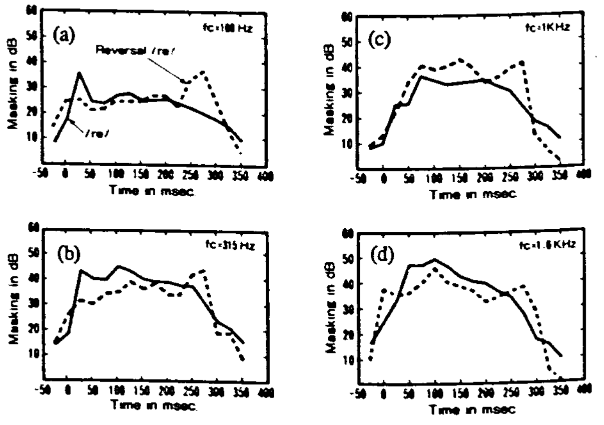

Fig.8 Masking patterns for /re/ (solid lines) and reversal /re/ (bloken lines) at four frequency bands: (a) fc=160Hz, (b) fc=315Hz, (c) fc=1kHz and (d) fc=1.6kHz

Figure 6 (a) - (c) represent masking patterns and power spectral patterns for continuous speech as a function of time. Dips seen in the masking patterns at each syllabic boundary, around t = 75, 200, 300, 475 msec., are deeper and more noticeable than those in the power spectral patterns. One reason for the dip depths in the masking patterns being prominant is that the masking values proceeding and succeeding the dips are large.

Figure 7 shows a monosyllable speech spectrogram for */re/* in normal time axis. This monosyllable and a time reversally reproduced one (reversal */re/*) are the maskers in the third experiment. Figure 8 (a) - (d) show masking patterns using */re/* (solid lines) and the reversal */re/* (broken lines) for a whole period of the masker sound. Comparing both masking patterns, masking values increase at the onset of each masker sound, whether it is reproduced reversely or not (i.e. at t = 25 msec. for */re/* and at t = 275 msec. for reversal */re/*). This phenomenon appears most remarkable at frequency band fc= 160Hz. Masking values of */re/* are larger in 5 to 10 dB than those of reversal */re/* at around t =50 to 100 msec. at frequency band fc= 315Hz and 1.6kHz. These frequency bands are those within which, F1 and F2 transition occurs, although their transition direction is different between */re/* (i.e. upward) and the reversal */re/* (i.e.downward).

DISCUSSION

Results show several important characteristics which seem to play important roles in physical to inter-aural spectral transformation by means of the non-steady part emphatic functions. Three of these characteristics found in comparing masking patterns with physical spectral patterns are discussed in this section.

First, speech onset is emphasized in masking patterns. This onset emphasis is caused by a temporal increase of the amplitude component, that is an upward amplitude modulated (AM) component. There exists a downward AM component due to temporal amplitude decrease at speech offsets. Although, the offset emphasis produced by the downward AM component is smaller than the onset emphasis.

Second, formant transitions, in particular, F1 and F2 transitions, in masking patterns are more prominent than those of the physical spectra. This is an inter-aural emphasis caused by formant movement which is composed of both AM and frequency modulated (FM) components. These AM components are produced by temporal change of each harmonics level. One of the FM components is produced by the resonance frequency movement itself as seen in broad band spectral patterns. In a strict sense, a formant transition is not a real movement of a physical existing frequency' component, such as sweep tone, but a movement of spectral envelope peaks estimated from several resonated harmonics of the fundamental frequency. However, this formant movement increases masking values as well as frequency sweep tone [20]. Another FM component included in the formant transition is fluctuation of harmonics frequencies. This fluctuation is a physically existing movement of the FM component due to fundamental frequency change.

Third, formants in middle and higher frequency ranges become prominent in masking patterns resulting from small masking value in the lower frequency range. This is due to a general masking characteristic that lower frequency components mask higher ones more effectively than higher frequency components do lower ones. In this paper, suppression effects along the frequency axis, which is seen in the results given by the pulsation threshold [12,13], are not reflected on the masking pattern since traditional masking procedures were used.

On the other hand, a decrease in masking values at the middle part of vowel is noticeable in the monosyllable masking patterns, but not so noticeable in the continuous speech patterns. This phenomenon is an adaptation effect caused by the steady state vowel part which has a several hundred milisecond duration. In a continuous speech masker, vowel part are not long enough to cause the adaptation effect. Since the adaptation decreases masking values at long steady vowel part, non-steady parts of speech (including onset and formant transitions) preceding and/or succeeding these vowels are relatively emphasized in the auditory pathway.

Furthermore, as shown in the results of the third experiment, reversing the time axis of a masker sound gives us completely different masking patterns. Two spectra with the same exact frequency structure have two different masking values. This suggests that spectral change direction and interaction between temporarily adjacent components play important roles in the physical to inter-aural spectral transformation.

To summarize, it is clear that temporal amplitude varying features, transition of formant frequencies and structures, which are considered to be important cues in speech perception, are emphasized and more prominent in the auditory pathway than those in physical spectrum patterns. It is expected that inter-aural spectral representation will bear better results than physical spectral representation when implemented in speech signal processing by computers. The physical to inter-aural spectrum transformation discussed in this paper can be described quantitatively by simulating AM/FM component emphasis, backward/forward masking, adaptation and lateral inhibition. This transformation can be implemented in a automatic speech recognition preprocessor as a better representation of speech spectrum capable of discriminating two utterances with confusable physical spectra.

CONCLUSION

In this paper, three types of speech sounds are examined by a masking method to show typical examples of inter-aural representation of speech spectrum represented by a spatio-temporal masking pattern and to clarify differences between inter-aural and physical representation of speech spectrum. Our findings are summarized as follows:

1) Compared with the physical spectral pattern: speech onsets and formant structure, in particular, formant transitions are prominent in the masking pattern.

2) Spectral emphasis is presumably composed of three auditory functions: AM/FM components emphasis, forward/backward masking and adaptation. These play important roles in physical to inter-aural spectrum transformation.

3) The direction of AM/FM component movements in speech sounds is of great importance and strongly affects the process of producing the inter-aural spectrum pattern.

4) Taking into account the considerable differences between inter-aural and physical representation of speech spectrum, the inter-aural spectrum can be implemented as a better representation of speech spectrum in speech feature extraction and speech signal processing by computers, particularly in automatic speech recognition by machine.

ACKNOWLEDGEMENTS

This work was mainly done at NTT Integrated Communications Labolatories. The author is grateful to Dr. Ryouhei Nakatsu for his continuing guidance and encouragement, and to Dr. Noboru Sugamura for his fruitful discussions. The author also wishes to thank Dr.Yoh'ichi Tohkura for his helpful advice in writing this paper.

References

[1] E.Zwicker and E.Terhardt (Eds.) (1974) *Facts and Models in Hearing*, Springer-Verlag, New York.

[2] R.Plomp (1976) , *Aspects of Tone Sensation; A psychophysical study*, Academic Press.

[3] R.Carlson and B.Granstrom (Eds.) (1982) *The Representation of Speech in the Peripheral Auditory System*, Elsevier Biomedical,

[4] D.H.Klatt (1980), "SCRIBERand LAFS: Two new approach to speech analysis," in *Trends in Speech Recognition* ,W.A.Lea (Ed.), Prentice-Hall, pp. 529-555

[5] J.B.Allen (1985)," Cochlea Modeling," IEEE ASSP Magazine, Jan., pp. 3-29

[6] R.F.Lyon (1982), "A Computational Models of Filtering, Detection and Compression in the Cochlea," Proceedings of ICASSP, pp. 1282-1285.

[7] S.Seneff (1986)," A Computational Model for the Peripheral Auditory Sysytem Application to Speech Recognition Research," Proceedings of ICASSP, pp. 1983-1986

[8] E.Zwicker (1986), "Peripheral Preprocessing in Hearing and Psychoacoustics a Guidlines for Speech Recognition," Proceedings of the Symposium on Speech Recognition, Montreal, pp.1-4.

[9] S.A.Shamma (1986)," The auditory processing of speech," Proceedings of the Symposium on Speech Recognition, Montreal, pp.14-17

[10] T. Ifukube (1973), " Masking by Frequency Modulated Tone," J. Acoust. Soc. Japan, vol.29, No.11, pp. 679-687.

[11] T. Ifukube (1975), " Auditory Masking of Amplitude Modulated tone and its Analysis by Analog Simulation," J. Acoust. Soc. Japan, vol.31, No.4, pp.237-245

[12] T.Houtgast (1974) ," Auditory Analysis of Vowel-Like Sounds," Acoustica vol.31, pp.320-324

[13] R.S.Tyler and B.Lindbolm (1982)," Preliminary study of simultaneous-masking and pulsation-threshold pattern of vowels," J. Acoust. Soc. Am., vol.71, No.1, pp. 220-224.

[14] B.C J.Moore and B.R.Glasberg (1983), " Masking patterns for synthetic vowels in simultaneous and forward masking," J. Acoust. Soc. Am., vol.73, No.3 pp.906-917

[15] A.Sidwell and Q.Summerfield (1986), " The Auditory Representation of Symmetrical CVC Syllables," Speech Communication, vol.5, No.3,4, pp.283-297

[16] E. Miyasaka (1983), " Spatio-temporal characteristics of masking of brief test-tone pulses by a tone-burst with abrupt switching transients," J. Acoust. Soc. Japan, vol.39, No.9, pp.614-623

[17] B.Delgutte and N.Y.S.Kiang (1984), "Speech coding in the auditory nerve I-III," J. Acoust. Soc. Am., vol.75, No.3, pp. 866-896.

[18] J.L.Goldstein (1966), " Auditory Nonlinearity," J. Acoust. Soc. Am., vol.41 No.3, pp.676-689

[19] T.Houtgast (1973), "Psychophysical Experiments on Tuning Curves and Two-Tone Inhibition," Acoustica, vol.29, pp.168-179.

[20] T. Hirahara (1985)," Auditory Response by Formant Transitional Stimuli from the Viewpoint of Simultaneous Masking," Proceedings of the Autamn meeting of Acoust. Soc. of Japan, pp. 249-250