to develop better tools in support of facial motion enhancement as part of speech rehabilitation.

Acknowledgements

This research was supported in part, by the Canada Research Chairs Program awarded to the second author. The authors gratefully acknowledge the assistance of Aravind Namasivayam with data collection and analysis.

References

Cheang, H. S., & Pell, M. D. (2007). An acoustic investigation of parkinsonian speech in linguistic and emotional contexts. *Journal of Neurolinguistics, 20*(3), 221-241.

Craig, M., van Lieshout, P. H. H. M., & Wong, W. (2007). Suitability of a UV-based video recording system for the analysis of small facial motions during speech. *Speech Communication, 49*(9), 679-686.

Craig, M. S., van Lieshout, P. H. H. M., & Wong, W. (Under Review). A linear model of acoustic-to-facial mapping: Model parameters, data set size and generalization across speakers. *The Journal of the Acoustical Society of America*

Goldstein, L. M., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In M. A. Arbib (Ed.), *From action to language: The mirror neuron system* (pp. 215-249). Cambridge, UK: Cambridge University Press.

Hintze, J. (2007). *NCSS, PASS, and GESS*. Kaysville, Utah, USA.

Miller, N., Allcock, L., Jones, D., Noble, E., Hildreth, A. J., & Burn, D. J. (2007). Prevalence and pattern of perceived intelligibility changes in parkinson's disease. *Journal of Neurology, Neurosurgery and Psychiatry, 78*(11), 1188-1190.

Namasivayam, A. K., van Lieshout, P., & De Nil, L. (2008). Bite-block perturbation in people who stutter: Immediate compensatory and delayed adaptive processes. *Journal of Communication Disorders, 41*(4), 372-394.

Pinto, S., Ozsancak, C., Tripoliti, E., Thobois, S., Limousin-Dowsey, P., & Auzou, P. (2004). Treatments for dysarthria in parkinson's disease. *Lancet Neurology, 3*(9), 547-556.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects: Research report. *Psychological Science, 18*(5), 392-396.

Smith, M. C., Ellgring, H., & Smith, M. K. (1996). Spontaneous and posed facial expression in parkinson's disease. *Journal of the International Neuropsychological Society, 2*(5), 383-391.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*(1-2), 3-45.

Tickle-Degnen, L., & Lyons, K. D. (2004). Practitioners' impressions of patients with parkinson's disease: The social ecology of the expressive mask. *Social Science and Medicine, 58*(3), 603-614.

van Lieshout, P. H. H. M., Bose, A., Square, P., & Steele, C. (2007). Speech motor control in fluent and dysfluent speech production of an individual with apraxia of speech and broca's aphasia. *Clinical Linguistics and Phonetics, 21*(3), 159-188.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1-2), 23-43.

FURTHER EVIDENCE OF TOP-DOWN GAIN CONTROL IN THE AUDITORY SYSTEM

Bruce A. Schneider[1] and Scott Parker[2]
[1]Centre for Research on Biological Communication Systems
University of Toronto Mississauga
Bruce.Schneider@utoronto.ca
[2]Department of Psychology
The American University
sparker@american.edu

Abstract

We have repeatedly proposed that a non-linear gain-control system under top-down governance regulates the encoding of auditory intensity. Here we examine the effects of individual differences and continued training on the identification of tones differing only in intensity. In a baseline condition, participants were asked to identify which one of four 1-kHztones (25, 30, 35, or 40 dB) was presented on a trial. In two other conditions an additional tone(50 or 80 dB) was added to the baseline set. Each of the three conditions was tested over the course of 20 sessions, each stimulus occurring 50 times per session. Performance improved over sessions with large individual differences among participants. However, in all cases, the pattern of change across conditions was consistent with the notion that the encoding of intensity along the decision axis is regulated by a non-linear gain control mechanism.

In previous papers (Gordon & Schneider, 2007; Parker, Murphy & Schneider, 2002; Parker & Schneider, 1994; Schneider & Parker, 1990) we have argued for the existence of a nonlinear gain control mechanism in the auditory system that is under top down control. This gain control mechanism has two functions: (1) to protect the auditory system from sensory overload; and (2) to maximize the discriminability of the acoustic stimuli in the listener's current soundscape (the range of sounds that the listener has recently or is currently experiencing, or expects to hear in the near future). Figure 1 presents a model of how this system might operate.

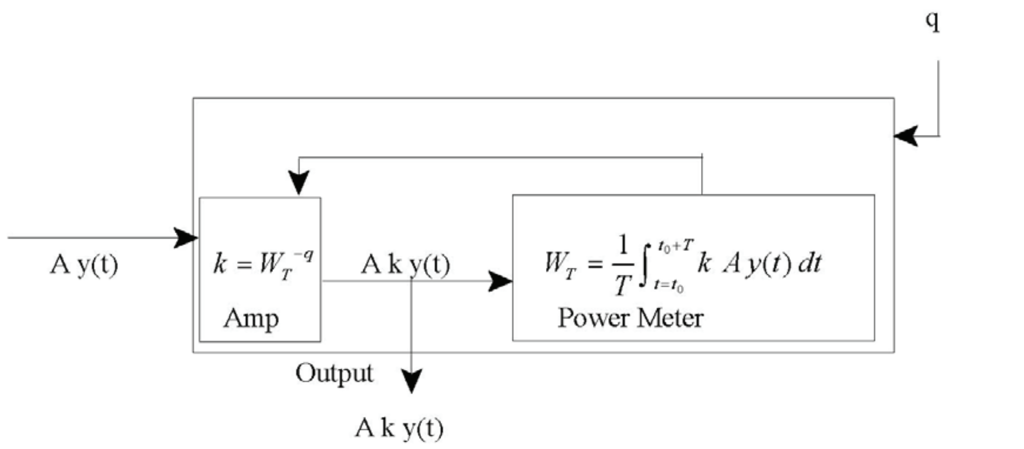

Figure 1. A non-linear gain control mechanism based on a feedback loop. The input signal is amplified by a factor k, whose value is controlled by the power at the output of the amplifier.

In this model the auditory signal is assumed to first pass through a bank of linear auditory filters with the output of each filter going to a nonlinear amplifier of the type shown in Figure 1 whose gain is subject to top-down control. Let *k* represent the current gain of the amplifier. The output of the amplifier goes to a power meter that computes the average power from the amplifier over the last *T* seconds where $T = 1/f_c$ where f_c is the center frequency of the auditory filter in question. The average power then feeds back on the input to control the gain. Parker and Schneider showed that when the input is a steady-state sinusoid [$A*Cos(2\pi f_c t + \theta)$] the steady-state output is a power function of the input, i.e., the output is

$$\sqrt{2}\left(\frac{A}{\sqrt{2}}\right)^{\frac{1}{1+2q}} Cos\left(2\pi f_c t + \theta\right) \qquad (1)$$

where the exponent of the power function is determined by the parameter *q*, which is assumed to be controlled in a top-down fashion. In particular, as *q* becomes larger, the exponent of the power function becomes smaller, thereby increasing the compression on the input.

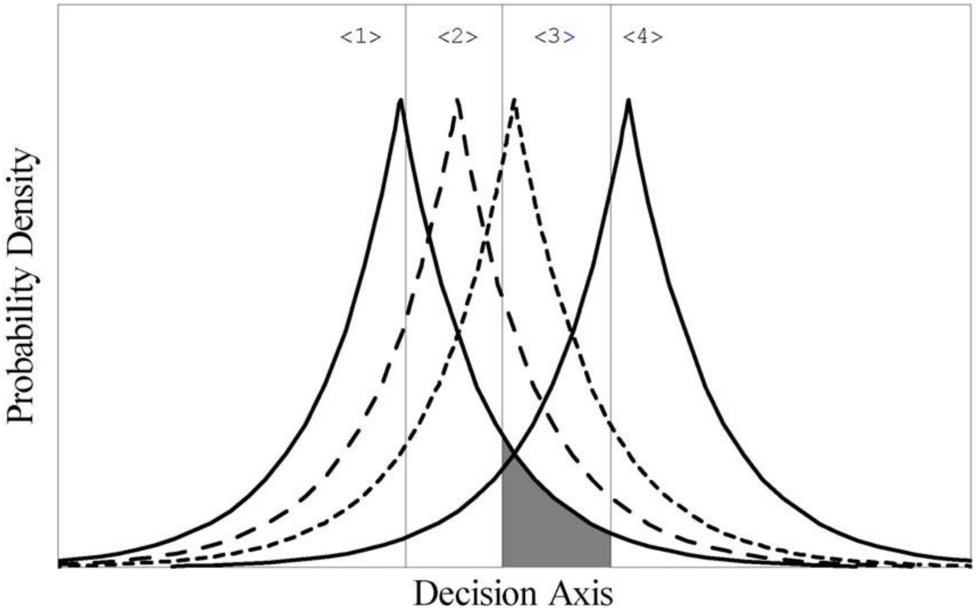

Figure 2. The EVLD SDT Model for a 4-alternative absolute identification experiment. Each stimulus, i, {1 ≤ i ≤ 4} gives rise to a Laplace distribution of events along the decision axis. The vertical lines represent the criteria which divide the decision axis into 4 response regions. The shade portion is the proportion of times stimulus 1 is identified as stimulus 3.

Part of the evidence for top-down control comes from an experiment in Parker et al. (2002) in which participants, in a baseline condition, had to identify on a trial, which one of four 1-kHz tones were presented. The four tones differed only in intensity (25, 30, 35, and 40 dB SPL) and were identified by pressing one of four buttons in an absolute identification experiment. Identification accuracy was assessed in two ways: (a) by the number of correct identifications of each of the tones; and (b) by evaluating the discriminabilities among the tones using signal-detection theory (SDT). In SDT it is assumed that the presentation of stimulus *i* gives rise to a response, r_i, along a decision axis. Because of variability in either the stimulus or in perceptual processing, the response evoked by the stimulus is assumed to vary across presentations in a random manner. Typically, it is assumed that the distribution of responses associated with a particular stimulus is Gaussian in shape, and that all distributions have the same variance. Schneider (2007), however, has shown that when data are averaged across subjects who differ in their sensitivities to stimulus differences, or when discriminability is changing over time within a subject, that the decision axis is better characterized by equal-variance, Laplace distributions (EVLDs) as shown in Figure 2. In this model the observer is assumed to place three criteria along this unidimensional decision axis, thereby dividing the decision axis into four response regions. Hence the data from such an experiment consists of a 4 X 4 stimulus response matrix, in which the entries are the probability of response *j* given stimulus *i*. Parker et al. (2002) have described a procedure in which the parameters of the model are selected so as to minimize Pearson's Chi Square, i.e., minimize

$$\chi^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(R_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \qquad (2)$$

where $R_{i,j}$ is the number of times stimulus *i* was labeled stimulus *j*, and $E_{i,j}$ is the expected number of times that stimulus *i* should be labeled as stimulus *j*, given the model shown in Figure 2.

In addition to determining stimulus discriminability among this baseline set of four stimuli, Parker et al. (2002), in a series of experiments, augmented the baseline set of stimuli by adding a fifth stimulus whose intensity was greater than those of the baseline set. If the gain on the amplifier is determined by the highest intensity stimulus that participants are likely to encounter, we would expect discriminability among the baseline set of four stimuli to progressively decrease as the intensity of the added stimulus was increased. This is exactly what they found.

The Parker et al. (2002) study collapsed response-probability matrices across subjects. In the current study, six individual subjects were tested on three of the conditions in Parker et al. over the course of 20 sessions to look for individual differences in the operation of the gain control mechanism, and to see if the improvements in discriminability with practice that are often found in such experiments, could be attributed to better "tuning" of this amplifier.

Method, Stimuli, and Procedure

Six students and staff members at the University of Toronto at Mississauga (3 females, 3 males) served. Their ages ranged from 19 to 39 years. All reported normal hearing and had normal looking audiograms from 250 to 4000 Hz. The stimuli were 1-kHz tones with intensities of 25, 30, 35, 40, 50, and 80 dB SPL presented diotically via TDH-49 headphones to the participants who were seated in a sound-attenuating chamber. Tones were digitally generated at a rate of 20 kHz and converted to voltages using a Tucker Davis (T/D) sound system. All tones were produced with a 10-msec rise/fall time and were attenuated to the proper level by means of a T/D programmable attenuator. Tone durations were 500 msec.

There were 3 conditions in the experiment – baseline, baseline + 50, and baseline + 80. The stimuli in the baseline condition were the 25, 30, 35, and 40 dB tones. The other two conditions included either the 50 (baseline +50) or the 80 (baseline +80) dB tone in addition to the four used in the baseline condition. Those three conditions could be sequenced in any of 6 permutations. Each of those six permutations was used with one subject.

Subjects were informed that they would be hearing a series of tones and that they were to identify those tones by pressing the button assigned to that tone on a button-box with the left-most button corresponding to the softest tone and the right-most button corresponding to the loudest tone. In each condition, each of the stimuli was presented 50 times in each session. Thus there were 200 stimuli in a baseline session and 250 stimuli in a session of the other two conditions. Stimuli were presented in randomly sequenced blocks of 50. If for any tone the subject did not respond within 2.5 seconds of stimulus presentation, all previous data for that block were discarded and the block began anew. Feedback was provided by a 200msec flash of the light above the correct button once the subject had responded. Each session began with either 40 (baseline condition) or 50 (the other two conditions) practice trials in which each stimulus was presented 10 times. Subjects served in 20 sessions in each of the three conditions, for a total of 60 sessions. After 20 sessions in the first condition the subject began the second condition, and after that was complete served in the third condition. Sessions lasted approximately 20 minutes in the baseline condition and 25 minutes in the other two conditions.

Results and Discussion

For each of the three conditions (baseline, baseline + 50 dB SPL, and baseline + 80 dB SPL) the data from each participant was divided into four consecutive blocks, each consisting of five sessions. Because we were primarily interested in the degree to which the addition of a fifth stimulus (at either 50 or 80 dB SPL) affected the listener's ability to discriminate among the four base tones (25, 30, 35, and 40 dB), a response on either button 4 or 5 when any of the four base tones were presented, was scored as a button-4 response. Hence, for each participant, the data consisted of four, 4 X 4 response-probability matrices ($p[R_j|S_i]$, $\{1 \leq j \leq 4\}$, $\{1 \leq i \leq 4\}$) corresponding to the four blocks of sessions.

Figure 3 plots the mean percentage of times (averaged over subjects) that each stimulus was correctly identified as a function of the number of sessions (in blocks of 4) in each of the three conditions. This figure indicates that the addition of a fifth stimulus resulted in a reduction in accuracy, with the reduction being more severe the more intense the stimulus. There is also an indication that accuracy improves over time in each of the three conditions. A two-factor, within-subjects ANOVA with condition (baseline, baseline + 50, baseline + 80) as the first factor, and block number as the second factor (blocks 1 to 4 referring, respectively to sessions 1-5, 6-10, 11-15, 16-20) confirmed that there were significant main effects due to condition (F[2,10] = 32.495, MSE = 0.0054, p < .001), session number (F[3,15] = 10.343, MSE = .0015, *p* = .001), as well as a significant condition by block interaction (F[6,30] = 2.47, MSE = .0017, *p* = .046). The interaction effect is due to the fact that while percent correct increased monotonically with session block for the baseline and baseline + 80 conditions, performance in the first five sessions in the baseline + 50 condition was better than in the second five sessions.

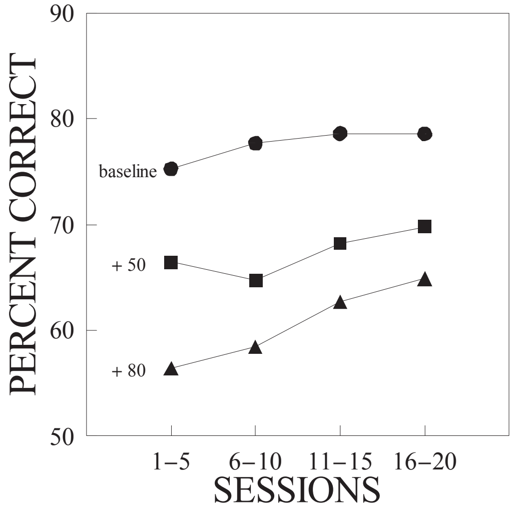

Figure 3. Percent correct identification of the 4 baseline stimuli as a function of session block in the three conditions.

In conducting a signal-detection analysis of these data, we first averaged the response-probability matrices across the six participants for each of the four session blocks to obtain four response-probability matrices for each condition. We then searched for the parameters that minimized Equation 2. In this search the standard deviation of the Laplace distributions was fixed at 1.0, and the mean of the Laplace distribution corresponding to stimulus 2 was set to 0. An iterative procedure then searched for the best fitting values of the three criteria and the remaining three means of the Laplace distributions (a total of six free parameters). These parameter values were then used to predict the probability of response *j* given stimulus *i* for all 16 combinations. Figure 4 (top row) plots the obtained probabilities the $R_{j,i}$, against the probabilities predicted by the EVLD model for the four session blocks in the baseline condition. The second and third rows of panels present the equivalent plots for the + 50 and +80 conditions, respectively. If the fit were perfect in these panels, all points would fall on the positive diagonal. Figure 4 indicates the fit of the model to the data is quite good.

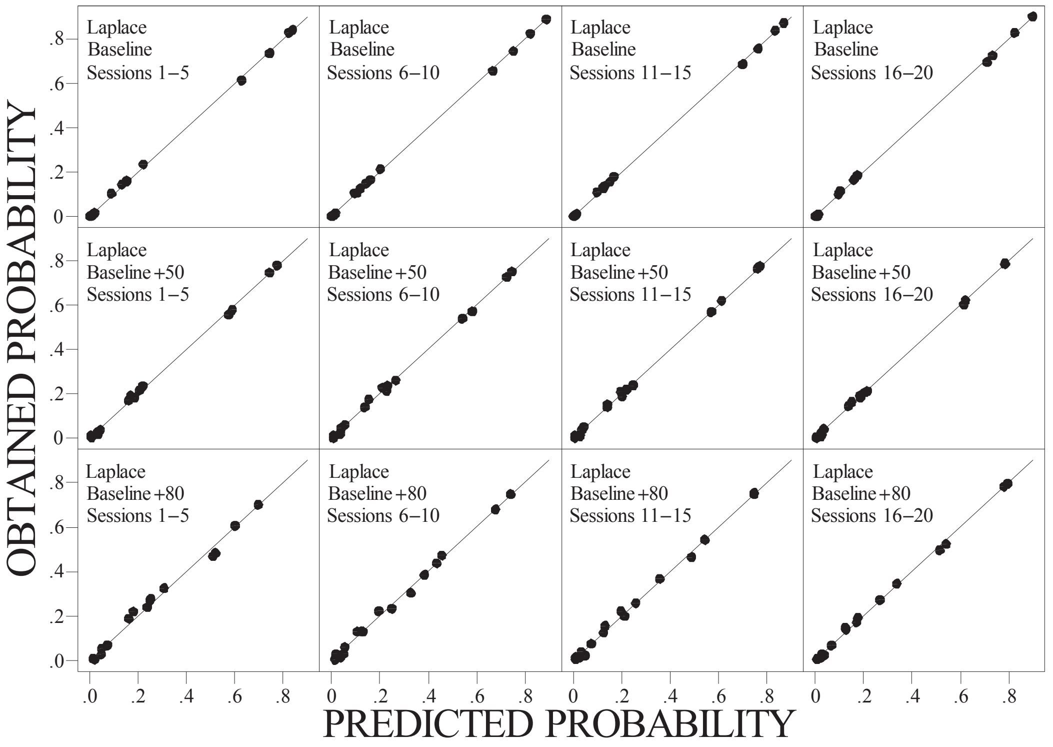

Figure 4. Obtained probability as a function of the probabilities predicted by the EVLD model for blocks of 5 sessions. Top row (baseline condition); middle row (baseline + 50); bottom row (baseline + 80).

The gain-control model predicts a reduction of discriminability among the baseline stimuli, when an additional stimulus is added to the base set, with the reduction in discriminability being greater when the intensity of the added stimulus is increased from 50 dB SPL to 80 dB SPL. An estimate of the overall discriminability in a signal-detection model can be obtained by determining the normalized distance between μ_1 and μ_4 in Figure 2. Because these distributions are Laplacian in shape, we refer to this as the Laplace dprime range, which is defined as $d'_{R,\,LP} = (\mu_4 - \mu_1)/\sigma$. Figure 5 plots the dprime range as a function of session block for each of the conditions of the experiment. Note that the same pattern emerges as was found in Figure 3, namely, that dprime range for the EVLD model increases with number of sessions, and is largest for the baseline condition, followed by the +50 and +80 conditions, respectively.

It is interesting to note that the portion of the Laplace dprime range occupied by the Laplace dprime between stimuli 1 and 2 (25 and 30 dB SPL) increases systematically as the intensity of the added stimulus is increased but does not change within a condition as a function of session block. The former is predicted by the gain control model illustrated in Figure 1 but not by other models based on auditory attention. The latter suggests that practice reduces the variability along the decision axis but does not affect the operation of the gain control mechanism.

Although not shown, the fit of the model to the data of the six individuals is also good. Hence, the pattern of results expected from a gain control model: (1) characterizes individual as well as group data; (2) indicates that discriminability improves with practice; and (3) that characteristics of the soundscape (in this case, the highest intensity expected in the soundscape) control the amount of compression exerted on the input.

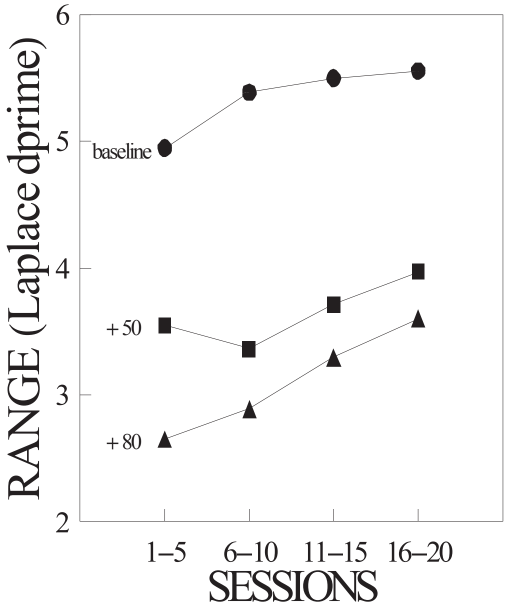

Figure 5. Laplace dprime range(distance between the means of distributions 1 and 4 in the EVLD Model) as a function of session block.

References

Gordon, M. S., & Schneider, B. A. (2007). Gain control in the auditory system. Absolute identification of intensity within and across two ears. *Perception & Psychophysics*, *69*, 232-240.

Parker, S., Murphy, D. R., & Schneider, B. A. (2002). Top-down gain control in the auditory system: Evidence from identification and discrimination studies. *Perception & Psychophysics*, *64*, 598-615.

Parker, S., & Schneider, B. (1994). Stimulus range effect: Evidence for top-down control of sensory intensity in audition. *Perception & Psychophysics*, *56*, 1-11.

Schneider, B. A. (2007). The shape of the underlying distributions in absolute identification experiments. In S. Mori, T. Miyaoka, & W. Wong (Eds.) *Fechner Day 2007: Proceedings of the 23rd Annual Meeting of the International Society for Psychophysics.* Tokyo, Japan: International Society for Psychophysics.

Schneider, B. A., & Parker, S. (1990). Does stimulus context affect loudness or only loudness judgments? *Perception & Psychophysics*, *48*, 409-418.

A CRITICAL-BAND-FILTER ANALYSIS OF JAPANESE SPEECH SENTENCES

Kazuo Ueda and Yoshitaka Nakajima
Perceptual Psychology Section, Kyushu University, Shiobaru, Minami-ku, Fukuoka 815-8540 Japan
ueda@design.kyushu-u.ac.jp nakajima@design.kyushu-u.ac.jp

Abstract

This investigation aims to seek common factors of speech sounds across different languages, which may be exploited widely in speech perception. Two-hundred sentences of Japanese, each uttered by 10 native speakers (5 females and 5 males), were analyzed through 20 bands of critical-band filters. Smoothed power fluctuations derived from the filters were submitted to factor analysis. The first three factors explained 33.6-34.8% of variance. These three factors, which could be related to four frequency bands, had appeared in the same way as in our previous analysis of British English speech. Intelligible noise-vocoded speech was obtained for both languages utilizing these frequency bands. The present analysis showed a common aspect of speech communication between Japanese and British English.

The present investigation focuses on frequency bands that adequately represent power fluctuation of critical-band-filtered Japanese sentences. This study is a continuation of our study on British English, presented at the last ISP meeting (Ueda & Nakajima, 2007).

The concept of critical bands (Fletcher, 1940) was proposed to model a frequency analysis function of our auditory system, more specifically the auditory periphery, in a simplified manner (Zwicker & Terhardt, 1980). The simplified model has been useful in estimating loudness of complex sounds, for example.

Critical-band filters have been also utilized in analysis of steady Dutch vowels. Plomp and his colleagues (Plomp, 1976) analyzed Dutch steady vowels with a bank of band-pass filters, which were practically equivalent to critical-band filters. They extracted principal components from the level fluctuation of the filter outputs. The first two principal components represented a vowel space very well, and the vowel configuration on a plane was well matched to the one obtained with the first two formant frequencies, the traditional way of analysis pioneered by Peterson and Barney (Peterson & Barney, 1952). However, Hillenbrand and his colleagues (Hillenbrand, Getty, Clark, & Wheeler, 1995; Hillenbrand & Nearey, 1999) clarified that spectral transition, together with static formant frequencies in steady portions of vowels, provides important cues to perceive vowels. The importance of spectral transition is also supported by other lines of investigation (Ladefoged & Broadbent, 1957; Verbrugge & Rakerd, 1986).

Studies on noise-vocoded speech suggest that the spectral transition can be rather coarse. The noise-vocoded speech (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) is a degraded speech, in which the amplitude envelope of an original speech waveform is preserved but the fine structure is replaced with noise. Usually amplitude envelopes are extracted from band-pass filtered outputs of the original speech sound. Thus, coarse mapping of spectral transition into several frequency bands is included in a process of synthesizing a noise-vocoded speech. Generally, intelligibility of noise-vocoded speech depends on the number of frequency bands: as the number increases, the speech becomes more intelligible. With extensive training--without feedback, however--noise-vocoded speech synthesized with