EXTRACTION OF RELATIONS BETWEEN LECTURER AND STUDENTS BY USING MULTI-LAYERED NEURAL NETWORKS

Eiji Watanabe
Faculty of Intelligence and Informatics, Konan University, Kobe 658–8501, Japan

Takashi Ozeki
Faculty of Engineering, Fukuyama University, Fukuyama 729–0292, Japan

Takeshi Kohama
School of Biology-Oriented Science and Technology, Kinki University, Kinokawa 649–5493, Japan

Keywords: Image processing, Neural networks, Lecturer, Students, Behavior, Relation.

Abstract: In this paper, we discuss the extraction of relationships between lecturer and students in lectures by using multi-layered neural networks. First, a few features concerning for behaviors by lecturer and students can be extracted based on image processing. Here, we adopt the following features as behaviors by lecturer and students; the loudness of speech by lecturer, face and hand movements by lecturer, face movements by students. Next, the relations among the above features concerning on their behaviors by lecturer and students can be represented by multi-layered neural networks. Next, we use a learning method with forgetting for neural networks for the purpose of extraction of rules. Finally, we have extracted relationships between behaviors by lecturer and students based on the internal representations in multi-layered neural networks for a real lecture.

1 INTRODUCTION

Gestures and eye contact play important and strong roles on human communication. Especially, in lectures, speech and gestures by lecturer have a great influences on the interest of students. As shown in Fig. 1, the contents (words, images, figures and speech) and gestures by lecturer are communicated to students in lectures. Also, the understanding and interest of students are transformed to students' behaviors and their behaviors are communicated to lecturer.

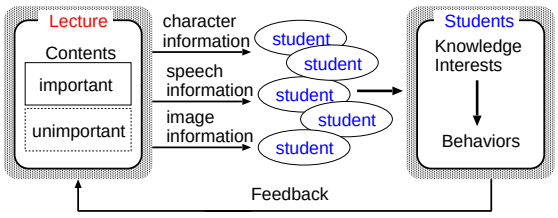


Figure 1: Information communication and interaction between lecturer and students in lectures.

Iso has shown that gestures by lecturer has strong relations with the skill of speech (Iso, 2001) Moreover, it is shown that the frequency for gestures have positive correlation with the skill of speech. On the other hand, Hatakeyama et al. have discussed a case that lecturer can not see the behavior of students (Hatakeyama and Mori, 2001) Thus, they have investigated how this case influenced with the speech and behaviors by lecturer. Therefore, in the evaluation of the interest of students, it is important to investigate the interaction between lecturer and students. Marutani et al. have proposed a method for the grasp of behavior by lecturer by using multiple cameras (Marutani et al., 2007).

In this paper, we discuss the relations between behaviors by lecture and students by using multi-layered neural networks. Moreover, we analyze the interaction based on the internal representation of neural networks. We adopt the face movement, the hand movement, and the loudness of speech by lecturer as their behaviors. Also, we adopt the face movements by students as students' behaviors. These features can be extracted by image processing methods from moving images. The time-series model can be generated by neural networks. Furthermore, the above model have complicated internal representations and it is difficult

Watanabe E., Ozeki T. and Kohama T..
EXTRACTION OF RELATIONS BETWEEN LECTURER AND STUDENTS BY USING MULTI-LAYERED NEURAL NETWORKS.
DOI: 10.5220/0003316200750080
In *Proceedings of the International Conference on Imaging Theory and Applications and International Conference on Information Visualization Theory and Applications* (IMAGAPP-2011), pages 75-80
ISBN: 978-989-8425-46-1

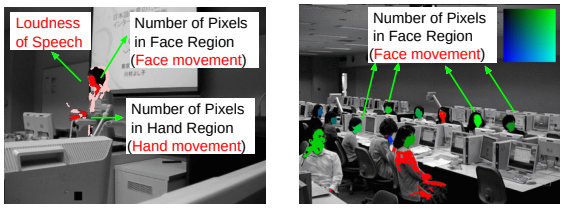
to explain the obtained rules. For this difficulty, we have introduced a structural learning algorithm with forgetting (SLF) for the purpose of clarifying internal representations of neural networks (Ishikawa, 1996). Finally, we have extracted the relations between behaviors by lecturer and students based on the internal representations in neural networks.

2 EXTRACTION OF FEATURES

In this section, we discuss the extraction of features in lecture. For the extraction of relations between behaviors by lecturer and students, we use the following features;

- For lecturer:
  - **Loudness of Speech.** Students are sensitive to not only the explanation by the speech but also the change and the loudness of the speech.
  - **Face and Hand Movements.** Lecturer can attract attention to students by looking at all faces of students and gesturing by lecturer.
- For students:
  - **Face Movement.** Lecturer can grasp the understanding and the interest of students by monitoring their face movements.

Fig. 2 shows features detected from moving images for lecturer and students. Especially, we focus on the number of pixels in face and hand regions for the extraction of movements by lecturer and students.



(a) Lecturer (b) Students

Figure 2: Features detected by digital signal processing for lecturer and students.

2.1 Loudness of Speech

The speech by lecturer can be recorded by general digital video camera. The loudness of speech by lecturer can be represented by the amplitude of the sound wave in WAVE format file. We use the following averaged quantity $\bar{x}^{\mathrm{L}}_{\mathrm{speech}}(t)$ as a feature for the loudness of speech by lecturer defined by

$$\bar{x}^{\mathrm{L}}_{\mathrm{speech}}(t) = \frac{1}{\Delta}\sum_{k=1}^{\Delta} |x^{\mathrm{L}}_{\mathrm{speech}}(t+k)|, \qquad (1)$$

where Δ denotes the section length for the moving average processing and $x^{\mathrm{L}}_{\mathrm{speech}}(t+k)$ denotes the amplitude of the sound wave.

2.2 Face and Hand Movements

These features can be extracted by image processing and we detect the region including skin-colored pixels from images for lecturer and students. Here, we extract the skin-colored regions including face and hands based on the detection of pixels with the following conditions;

$$\begin{cases} f_{\mathrm{Red}}(x,y) > \varepsilon_{\mathrm{Red}}, \\ f_{\mathrm{Red}}(x,y) > f_{\mathrm{Green}}(x,y) + \Delta_{\mathrm{Green}}, \\ f_{\mathrm{Red}}(x,y) > f_{\mathrm{Blue}}(x,y) + \Delta_{\mathrm{Blue}}. \end{cases} \qquad (2)$$

where $\varepsilon_{\mathrm{Red}}$ denotes a threshold for the detection of red-colored pixel. Also, Δ_{Green} and Δ_{Blue} denote thresholds for the evaluation of the objective pixel.

The regions having skin-colored pixels are used to extract face and hand movements as follows;

Next, the above mentioned regions having skin-colored pixels are used to extract face and hand movements as follows;

- **Number of Skin-colored Pixels in Face Region.** The position of a digital video camera is fixed in a lecture room and this camera does not pan, tilt and zoom the objectives. Therefore, when the lecturer and students move their face, the number of skin-colored pixels in face region does change according to the position and the direction of their faces. Here, we use the number of skin-colored pixels in face region as a feature of face movements by lecturer and students.
- **Number of Skin-colored Pixels in Hand Region.** Similarly, we use **the number of skin-colored pixels in hand region** as a feature of hand movements by lecturer.
- **Discrimination between Face and Hand Region.** Face and hand regions in lecturer include the same color pixels and their regions can not be discriminated by using only Eq. (2). Here, we detect the head region in lecturer by extracting hair-colored pixels from images for lecturer and the distance between head and skin (face or hand) regions is calculated. If the distance is small, then we can decide the objective region is face one. On the other hand, if the distance is large, then we can decide the objective region is hand one.

3 EXTRACTION OF RELATIONS

First, we discuss the influence on the student movements by lecturer movement. The input-output relation between the two movements can be represented by the inputs $\bar{x}^L_{\text{speech}}(t)$ (the loudness of speech by lecturer), $x^L_{\text{face}}(t)$ (the number of skin-colored pixels in lecturer's face region), and $x^L_{\text{hand}}(t)$ (the number of skin-colored pixels in lecturer's hand region) and the output $x^S_{\text{face}}(t,p)$ (the number of skin-colored pixels in the face region for the p-th student). Moreover, the above two movements are changed according to time t and they are interacted with each other with the time delay ℓ. Therefore, we can model the two features by the following time-series model;

$$x^{\text{S}}_{\text{face}}(t,p) = f^{\text{LS}}(x^{\text{L}}_{\text{face}}(t-\ell), x^{\text{L}}_{\text{hand}}(t-\ell), x^{\text{L}}_{\text{speech}}(t-\ell); w^{\text{LS}}_{ij}(p)), \quad (3)$$

where $\ell = 1,2,\cdots,T$), T and p denotes the length of the objective section and the student number respectively. In this equation, a function $f^{\text{LS}}(\cdot)$ is unknown and the value of coefficients $w^{\text{LS}}_{ij}(p)$ are unknown. Therefore, we use a neural network model as shown in Fig. 3. Unknown coefficients $w^{\text{LS}}_{ij}(p)$ denote weights in this neural network model.

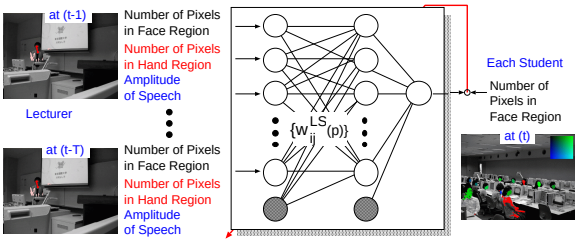


Figure 3: Neural network model for the learning of the influence on students movements by lecturer movements.

Neural network models have been widely applied to the numerous fields by their non-linear mapping ability. Rumelhart et al. (Rumelhart and McClelland, 1986) have proposed back-propagation (BP) learning algorithm for neural networks. However, since neural networks trained by BP algorithm have distributed representations in all hidden units, it is difficult to explain clearly the obtained rules and knowledge.

For this difficulty, Ishikawa (Ishikawa, 1996) has proposed a structural learning algorithm with forgetting (SLF) for the purpose of clarifying internal representations of neural networks. In SLF learning algorithm, weights in neural networks can be updated so as to minimize the following error function including the additive term $\sum|w_{ij}|$;

$$E_F = \sum_k (t(k) - o(k))^2 + \varepsilon \sum |w_{ij}|, \quad (4)$$

where $t(k)$, $o(k)$ denote teaching signal and outputs of neural network for the k-th pattern respectively. Moreover, ε denotes the amount of forgetting for weights and SLF algorithm can reduce the value of redundant weights by setting adequate ε. As one index for the clarifying for neural network model, we introduce the entropy H defined by

$$H = -\sum_{i,j} p_{ij} \log p_{ij},$$

where $p_{ij} = |w_{ij}| / \sum_{ij} |w_{ij}|$. When H_w becomes small, the number of redundant weights becomes small and we can evaluate that the objective neural network model has "clear" internal representation.

In the following section, we discuss the interaction between lecturer and students in an actual lecture by focusing on weights in neural network models.

4 ANALYSIS RESULTS

We have recorded images and speech for lecturer and students in a lecture (Title: "Application of Internet technology to teaching Japanese as a foreign language,". In this lecture, the lecturer explained the outline by speech during the first 10 minutes. We adopt the lecture scene during the first 10 minutes. Fig.4 shows the layout of the lecture room and the images for lecturer and students were recorded by two digital video cameras as shown in this figure respectively. These images were recorded by the rate 10 [fps] and the size of 640×480 [pixels].

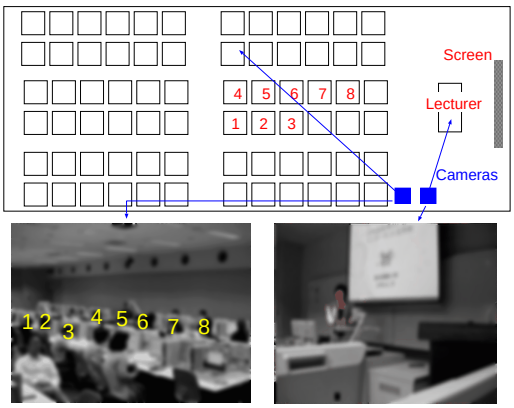
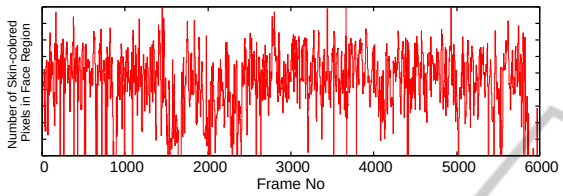


Figure 4: Layout of the lecture room.

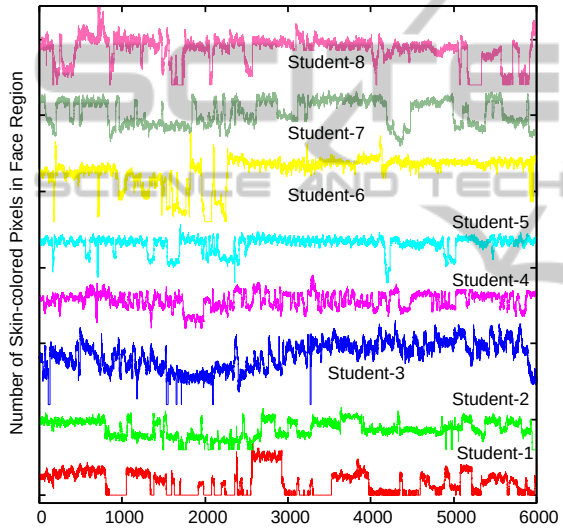
Moreover, we have set a few parameters as follows; the number of students: 8, Δ (section length for the moving average of speech): 0.5 [sec], ε_{Red} (threshold for the detection of red-colored pixel): 80, Δ_{Green} and Δ_{Blue}: 15, 15 and T (the length of the objective section): 10[sec].

4.1 Features for Lecturer and Students

Fig. 5 (a) and (b) show the number of skin-colored pixels in face regions for lecturer and all students respectively. The number of skin-colored pixels for each student changes according to the changes of features by lecturer complicatedly.



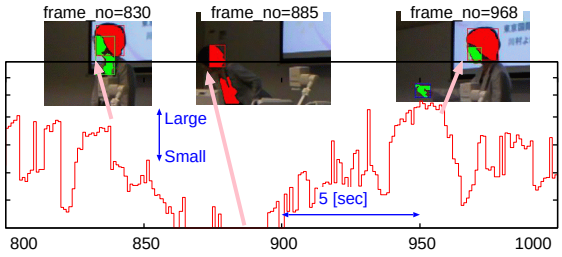
(a) For a lecturer.



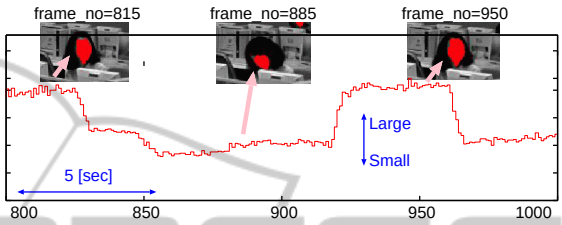
(b) For all students.

Figure 5: Changes of the number of skin-colored pixels in face region for lecturer and students.

Fig. 6 shows changes of the number of pixels in face region detected in this lecture. In Fig. 6 (a), it is shown that the number of skin-colored pixels changes according to the position and direction of face of lecturer. Similarly, in Fig. 6 (b), it is shown that the number of skin-colored pixels changes according to the position and direction for faces of students. Especially, from the 880-th frame, lecturer looks at hand for the manipulation of the mouse and the number of skin-colored pixels becomes small. Then, the number of skin-colored pixels for the 7-th student decreases. Here, when this student looks at a lecturer, the number of skin-colored pixels becomes large.



(a) For lecturer.

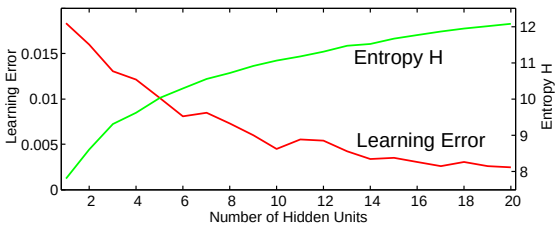


(b) For the 7-th student.

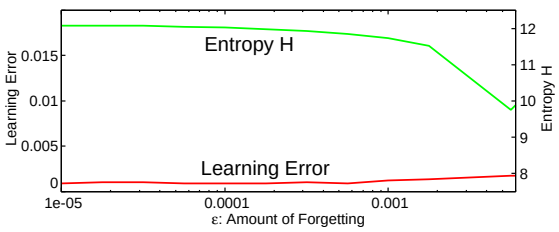
Figure 6: Changes of the number of skin-colored pixels in face region for lecturer and the 7-th student.

4.2 Learning Results

Fig. 7 shows learning error and entropy by BP and SLF algorithms for Student-7. In Fig. 7 (a), it is shown that entropy increases according to the increase of hidden units. Therefore, neural networks trained by BP algorithm have complicated internal representations. On the other hand, in Fig. 7 (b), it is shown that entropy decreases according to the increase of the amount of forgetting.



(a) By BP algorithm.

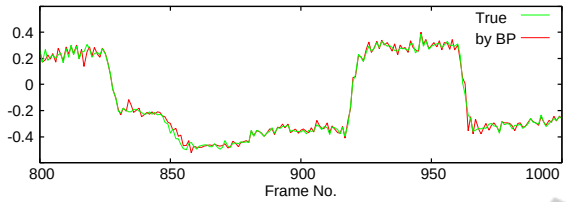


(b) By SLF algorithm.

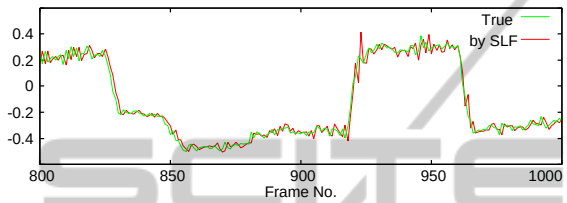
Figure 7: Learning error and Entropy by BP and SLF algorithms (Student-7).

Fig. 8 shows estimated outputs (the number of

skin-colored pixels) by the neural network model shown in Fig. 3. From this figure, we can see that BP and SLF algorithms could obtain good estimation ability.



(a) By BP algorithm ($n_2 = 20$).



(b) By SLF algorithm ($\varepsilon = 10^{-2.25}$).

Figure 8: Estimated results of the number of skin-colored pixels in face of Student-7.

4.3 Analysis of the Internal Representation in Neural Networks

In multi-layered neural networks, weights and hidden units have important roles in extracting rules through learning.

First, we focus on outputs for hidden units in multi-layered neural networks. S.D. for outputs $h_j^{\mathrm{LS}}(t)$ of hidden units by BP and SLF algorithms are shown in Table 1. Here, outputs $h_j^{\mathrm{LS}}(t)$ of hidden units can be obtained by the following equation;

$$\begin{aligned} h_j^{\mathrm{LS}}(t) \quad = \quad & f^{\mathrm{LS}}\Big(\sum_{\ell=1}^{T} w_{j\ell,\mathrm{face}}^{\mathrm{LS}} x_{\mathrm{face}}^{\mathrm{LS}}(t-\ell) \\ & + \sum_{\ell=1}^{T} w_{j\ell,\mathrm{hand}}^{\mathrm{LS}} x_{\mathrm{hand}}^{\mathrm{LS}}(t-\ell) \\ & + \sum_{\ell=1}^{T} w_{j\ell,\mathrm{speech}}^{\mathrm{LS}} x_{\mathrm{speech}}^{\mathrm{LS}}(t-\ell)\Big), \end{aligned} \quad (5)$$

In Table 1, standard deviation for all hidden units trained by BP algorithm have non-zero value and neural network shown in Fig. 3 has complicated internal representation. In the case of SLF algorithm, only two hidden units (the 12-th and 19-th hidden units) contribute to the modeling by neural network.

Next, we focus on weights between input-hidden units in multi-layered neural networks. Fig. 9 shows weights ($\{w_{j\ell,\mathrm{face}}^{\mathrm{LS}}\}$, $\{w_{j\ell,\mathrm{hand}}^{\mathrm{LS}}\}$, and $\{w_{j\ell,\mathrm{speech}}^{\mathrm{LS}}\}$ in Eq. (3)) between input and hidden layers of neural networks for Student-7 trained by BP and SLF algorithms. In Fig. 9 (a), weights trained by BP algorithm are shown. Since almost all weights have non-zero value, it is difficult to analyze the internal representations of neural network. On the other hands, Figs. 9 (b-1), (b-2), and (b-3) show weights of neural networks trained by SLF algorithm.

Table 1: S.D. for outputs of hidden units for features by BP and SLF algorithms (Student-7, $n_2 = 20$, $\varepsilon = 1 \times 10^{-2.25}$).

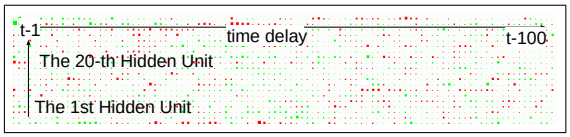
j	BP	SLF	j	BP	SLF
1	0.340	0	11	0.284	0.066
2	0.191	0	12	0.265	0.205
3	0.232	0	13	0.319	0
4	0.254	0	14	0.234	0
5	0.485	0	15	0.227	0
6	0.220	0	16	0.303	0
7	0.227	0	17	0.532	0.089
8	0.215	0	18	0.336	0
9	0.264	0	19	0.483	0.184
10	0.252	0	20	0.350	0

These weight as shown in Figs. 9 (b-1), (b-2), and (b-3) denote the influences face movement (b-1), hand movement (b-2), and speech (b-3) on Student-7. As shown in Table 1, weights between input units for behaviors by lecturer and the two hidden units (the 12-th h_{12}^{LS} and 19-th h_{19}^{LS}) dominate the interaction between lecturer and Student-7. The influences by lecturer can be summarized as follows;

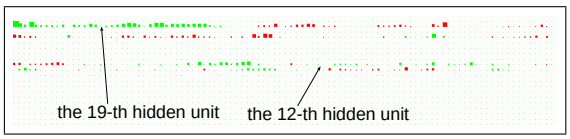
- The influence of face movement: In Fig. 9 (b-1), h_{19}^{LS} has many weights with negative and large value and their weights clustered in the former half of time-delay domain $[t-1, t-40]$.
- The influence of hand movement: In Fig. 9 (b-2), h_{19}^{LS} has many weights with negative and large value and their weights clustered in the former half of time-delay domain $[t-1, t-40]$. h_{12}^{LS} has many weights with positive and large value and their weights clustered in the latter half of time-delay domain $[t-70, t-95]$.
- The influence of the loudness of speech: In Fig. 9 (b-3), h_{12}^{LS} has many weights with positive and large value and their weights clustered in the latter half of time-delay domain $[t-60, t-90]$.

5 CONCLUSIONS

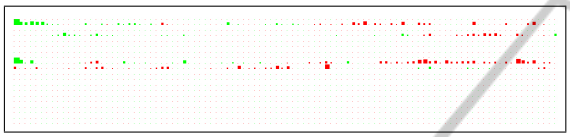
In this paper, we have constructed time-series models for the interaction in the lecturer. We have extracted a few features based on behaviors by lecturer and stu-



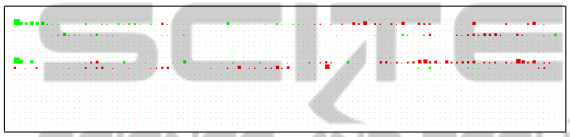
(a) By BP algorithm (for face).



(b-1) By SLF algorithm (for face).



(b-2) By SLF algorithm (for hand).



(b-3) By SLF algorithm (for speech).

Figure 9: Weights between input and hidden layers $\{w^{\mathrm{LS}}_{j\ell,\mathrm{face}}\}$, $\{w^{\mathrm{LS}}_{j\ell,\mathrm{hand}}\}$, $\{w^{\mathrm{LS}}_{j\ell,\mathrm{speech}}\}$ (Horizontal: time delay, Vertical: hidden unit, Size of square: value of weight, Green square ■ : negative weight, Red square ■ : positive weight).

dents and their features could be modeled by neural networks with SLF algorithm. Moreover, we have analyzed the interaction based on weights of neural networks. Especially, we have shown that the loudness of speech by the lecturer has influenced on behaviors by students with time-delay.

ACKNOWLEDGEMENTS

This research was partially supported by the Japanese Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 22500947, 2010. We would like to appreciate Prof. Y. Kawamura (Tokyo International University, Japan) and Prof. T. Kitamura (Konan University, Japan) for their permission of recording the lecture and their valuable comments.

REFERENCES

Hatakeyama, M. and Mori, K. (2001). Interaction between gesture and eye contact in communication. *Memoirs of the College of Education, Akita Univ.*, 56:69–75.

Ishikawa, M. (1996). Structural learning with forgetting. *Neural Networks*, 9(3):509–521.

Iso, K. (2001). Effects of nonverbal behaviors on the perception of "skillfulness of speech". *Japanese Journal of Interpersonal and Social Psychology*, 1:133–146.

Marutani, T., Sugimoto, Y., Kakusyo, K., and Minoh, M. (2007). Lecture context recognition base on statistical feature of lecture action for automatic video recording. *IEICE Trans. on Info. & Sys.*, 90–D(10):2775–2786.

Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel Distributed Processing*. MIT Press.