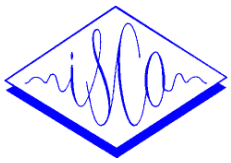

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

LISTENER ADAPTIVE CHARACTERISTICS IN DIALOGUE SPEECH - EFFECTS OF TEMPORAL ADJUSTMENT ON EMOTIONAL ASPECTS OF SPEECH -

Satoshi Imaizumi, Akiko Hayashi** and Toshisada Deguchi***

*** Research Institute of Logopedics and Phoniatrics**
Faculty of Medicine, University of Tokyo, Tokyo, JAPAN
imaizumi@tansei.cc.u-tokyo.ac.jp
**** Department of Education, Tokyo Gakugei University, Tokyo, JAPAN**

ABSTRACT

Effects of listener adaptive temporal adjustment in dialogue on emotional characteristics of speech were investigated by analyzing the speech of teachers directed to hearing-impaired (HI) or normal-hearing (NH) children. 1) Four factors were extracted which accounted for 81% of the total variance in perceptual rating scores on emotions. 2) Factor 1 represented the emotional differences between the read tokens (RD) and the dialogue speech (NH and HI), which were the contrasts between pleasant versus discomfort. 3) Factor 3 represented the differences between HI and the others (RD and NH), which was the contrast between "Slow, Stiff, Intelligible, Strong" versus "Busy, Tense, Rough, Dull". 4) The measures of the temporal structure of speech significantly accounted for the above-mentioned emotional contrasts. These results suggest that the listener-adaptive temporal adjustment of dialogue speech affects not only the segmental characteristics of speech such as devoicing but also prosody-related characteristics of speech such as the emotional profiles.

I. INTRODUCTION

Interlocutors seem to adapt their speaking style to various aspects of the situation, particularly to their partners with whom dialogue is going on. Our previous analyses of dialogue between professional teachers and normal-hearing (NH) or hearing-impaired (HI) children showed that teachers tended to use simpler and shorter sentences, inducing more responses, toward HI than toward NH. They also inserted longer pauses into phonological phrase boundaries, stretched syllable duration and reduced their speaking rate (Imaizumi et al, 1993a, b). They also showed that the teachers significantly reduced the devoicing rate in dialogue directed to HI (Imaizumi et al, 1994b).

Comparing speech read in an effort to produce clear speech for the hearing impaired to conversational speech, Picheny et al (1985, 1986) concluded; 1) The speaking rate decreases and the intelligibility increases substantially; 2) vowels are reduced to a lesser extent; 3) stop bursts, as well as all word-final consonants, are released; and 4) the RMS intensities for obstruent sounds, particularly consonants, are greater.

These reports suggested that interlocutors adapt their speaking style to their listeners' hearing/speech capacities in order to help them understand dialogue.

This paper reports effects of listener adaptive adjustment of the temporal structure of dialogue on emotional aspects of speech. The emotional aspects of dialogue speech seem to be significantly affected by the temporal structure of dialogue and to be very dependent on speaker-listener interactions, although such aspects of dialogue have not intensively been studied so far.

II. METHOD

2.1 Experiment I

In Experiment I, dialogue was recorded during a simple picture-pointing game through which a teacher attempted to assess the speech communication ability of a hearing-impaired or a normal-hearing child.

One of two panels, A or B, displaying 11 pictures of boys/girls, was named by the teacher, and then the child had to point at the specified picture as fast as possible.

The target names were three mora words having the form of /C1V1C2V2C3V3/. The target names were classified into six groups according to the manner of articulation of the component consonant (fricatives, affricates or stops) and mora position (initial or middle). Accent was put on the initial mora. The details of the material were described in other place (Imaizumi et al., 1993c, 1994b).

Seven professional teachers, seven hearing-impaired children (HI) and seven normal-hearing children (NH) participated in the test. They were speakers of the standard Tokyo dialect of Japanese.

Recordings were carried out for each of 14 teacher-child pairs in a sound-treated testing room. The teacher in each pair explained the task to the child first and began the game using Panel A (Game A), and then used Panel B (Game B) after a short pause. Using DAT and VCR recorders, all the dialogue and gestures were recorded, including the explanation of the task and the two sessions, A and B.

The teachers were instructed to explain the game using their own words and sentences. Only the question "Donokoga /CVCVCV/ desuka? " which meant "Who is /CVCVCV/ ?" was fixed.

Using an object-oriented acoustic analysis system (Imaizumi et al, 1993c, 1994a), the temporal structure of the dialogue, turn taking, grammatical structure of sentences and the temporal characteristics of the speech wave form were analyzed.

The speech directed to the normal-hearing (NH) and the hearing-impaired (HI) children will be referred to as the NH and the HI tokens, respectively, in the remaining of this paper.

2.2 Experiment II

In Experiment II, in order to clarify the differences between dialogue and read speech, a reading passage was recorded and analyzed. Seven teachers who participated in Experiment I read the target sentences "Donokoga /CVCVCV/ desuka? " five times as fast and as clearly as

possible. Using the acoustic analysis system, the temporal characteristics of the speech wave form were analyzed. The read speech samples will be referred to as the RD tokens below.

2.3 Experiment III

In order to clarify perceptual effects of temporal adjustment in dialogue, perceptual characteristics of the RD, NH and HI tokens were analyzed through the semantic differential method. As shown in Fig. 1, 24 pairs of adjectives were used as 9-point dipole rating scales. The listening subjects were 8 normal hearing students. The tokens used were 21 samples of "Donokoga hikita desuka? " spoken by the 7 teachers in three mode of RD, NH and HI. Totally 168 times 24 rating scores were analyzed by a principal factor analysis, and then regression analysis was carried out to extract any significant correlation with the temporal structure of speech.

III. RESULTS

3.1 Perceptual differences

In order to test the significance of effects of mode (RD, NH and HI) and teacher (A, B,..., F), an analysis of variance (ANOVA) was carried out for the rating scores on the 24 dipole scales. For most of the scales used, the main effects of mode, teacher and their interaction were significant at the 1% level. Fisher's PLSD test revealed the differences between RD and NH or HI (read tokens versus dialogue speech) were significant for all the rating scales, but the differences between NH and HI were significant only for some of the rating scales.

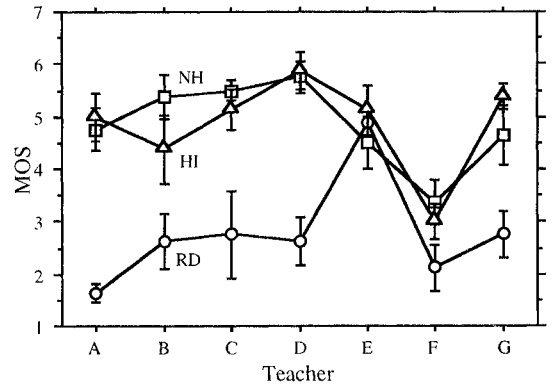

Fig. 1 (a) Interaction line plot for the mean rating scores (MOS) of "Teinei (Polite)" with standard error bars.

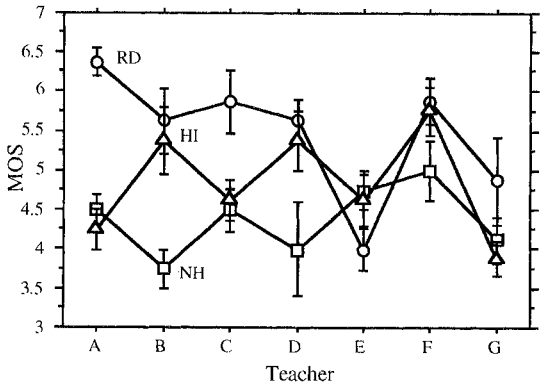

Fig. 1 (b) Interaction line plot for the mean rating scores (MOS) of "Tsuyoi (Strong)" with standard error bars.

Figure 1 (a) shows the interaction line plot for the rating scores of "Teinei (Polite)". It is clearly shown that the RD tokens were perceived as the most impolite utterances and the dialogue tokens (HI and NH) as the most polite utterances, through the average rating scores differed between the teachers. The ANOVA revealed that mode, teacher and their interaction had significant effects. Fisher's PLSD test revealed significant differences between RD and NH and between RD and HI, but not between NH and HI.

Figure 1 (b) shows the interaction line plot for the rating scores of "Tsuyoi (Strong)". For all the teachers except E, the RD tokens were perceived as the strongest utterances, and the NH tokens as the weakest utterances. The HI tokens were perceived as intermediate between the RD and NH tokens, although there were large speaker-dependent differences. The ANOVA revealed that mode ($p<0.0001$), teacher ($p=0.0004$) and mode-teacher interaction ($p=0.0003$) had significant effects. Fisher's PLSD test revealed significant differences between RD and NH ($p<0.0001$), and between RD and HI ($p=0.007$), and also between NH and HI ($p=0.0115$).

A principal factor analysis was used to extract a few principal factors which account for the differences in a compact way. Four factors extracted had the accounting for rate of 72.4, 5.2, 3.7, 2.4% respectively and these factors accounted for 83.7% of the total variance in the rating scores.

Figures 2 (a) and (b) show the factor loading of the extracted four factors after the Varimax orthogonal rotation. Factor 1 (F1) represents the emotional contrast between discomfort ("Awful, Rough, Uneasy, Busy, etc.") versus pleasant ("Easy, Kind, Friendly, Restful, Polite, etc."). Factor 2 represents the contrast between "Strong" versus "Dull, Lifeless." Factor 3 represents the contrast between "Slow, Stiff, Unnatural, Intelligible, Strong" versus "Busy, Lifeless, Tense, Rough, Dull". Factor 4 represents the contrast between "Lifeless, Deep" versus "Stiff, Kind, Unnatural."

The analysis of variance (ANOVA) for the rotated factors revealed the following.

1) For Factor 1, mode ($p<0.0001$) and teacher ($p=0.0029$) and their interaction ($p<0.0001$) had significant effects. As shown in Fig. 3 (a), Factor 1 mainly represents the differences in emotional scores between the read tokens (RD) and the dialogue tokens (HI and NH). Comparing the factor loading of Factor 1 shown in Fig. 2 (a), it was found that the read tokens (RD) except those of Teacher E tended to be perceived as "Awful, Rough, Uneasy and Busy." On the other hand, the dialogue tokens (HI and NH) were perceived as "Easy, Kind, Friendly, Restful, Polite." The HI tokens had intermediate values between the RD and NH tokens, although there were some differences depending on the teachers.

2) For Factor 2, only teacher had a significant main effect ($p=0.0048$). This was also true for Factor 4, for which only teacher had a small but a significant main effect ($p=0.0149$). These two factors represent speaker-dependent differences in the emotional rating scores.

3) For Factor 3, mode ($p<0.0001$) and teacher ($p<0.0027$) had significant effects, but their interaction ($p=0.1562$) had no significant effect. Fisher's PLSD test revealed that the differences between HI and the others (RD or NH) were significant ($p<0.0001$), but the difference between RD and NH was not significant. As also shown in Fig. 3 (b), this factor represents the differences between the HI tokens versus the others (RD and NH). The contrast is between "Slow, Stiff, Unnatural, Intelligible, Strong" versus "Busy, Lifeless, Tense, Rough, Dull."

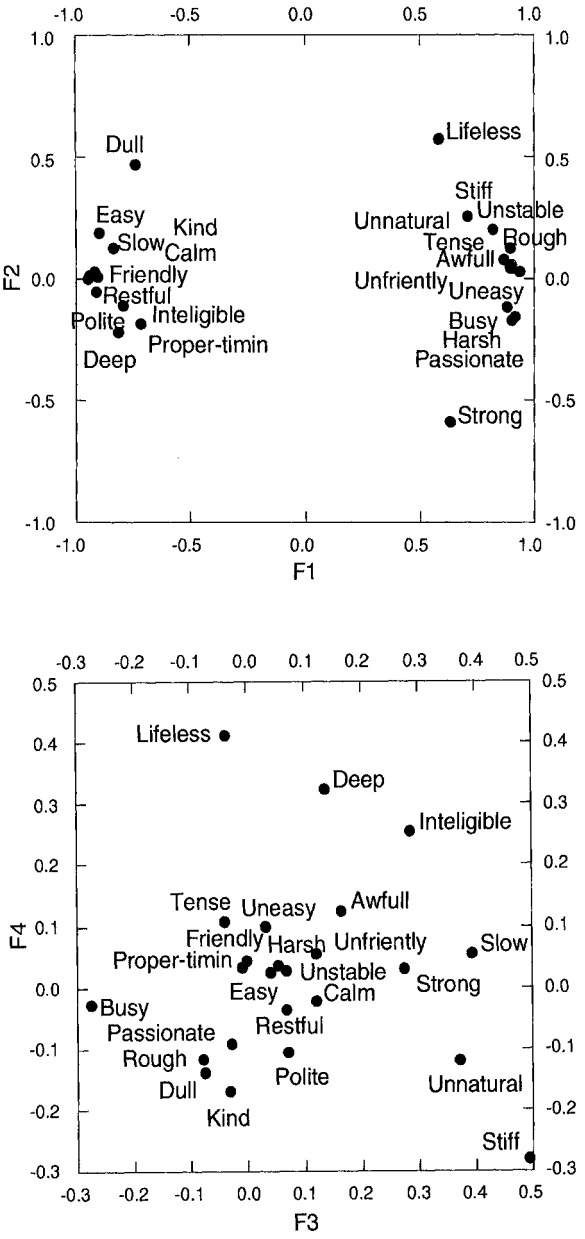

Fig. 2 (a) above and (b) below. The factor loading of the extracted four factors after the Varimax orthogonal rotation.

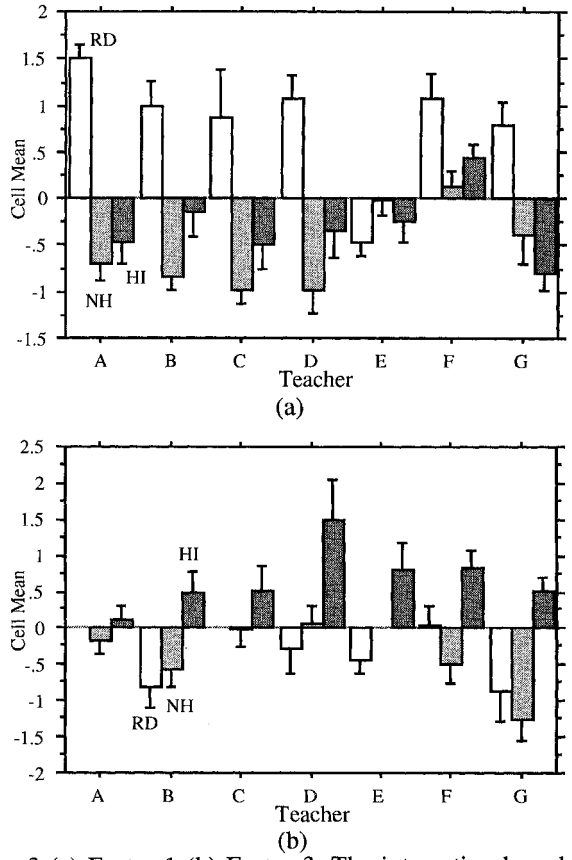

Fig. 3 (a) Factor 1 (b) Factor 3. The interaction bar plot representing the results of ANOVA with standard error bars.

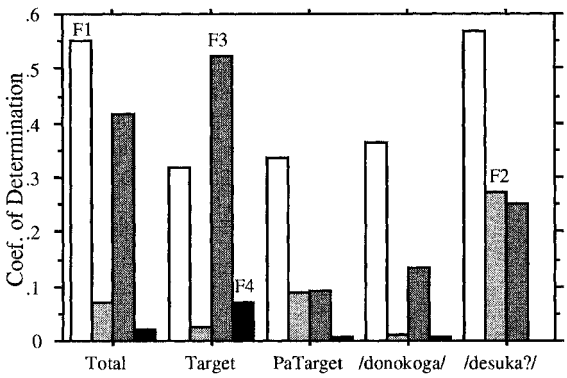

Fig. 4. The coefficients of determination obtained through the regression analyses between four factors and temporal measures. Total: length of sentence, Target: length of target name, PaTarget: pause length before the target.

3.2 Effects of temporal structure

In order to determine the effects of the temporal structure of the tokens on the emotional rating scores, lengths of various parts of the tokens, whole sentence, words, moras, pauses, voiced/unvoiced/silent segments were measured, and then regression analyses were carried out.

Fig. 4 shows the coefficients of determination (the squared correlation coefficients) obtained through the regression analyses. The results shown in Fig. 4 can be summarized as follows.

1) The total length of the tokens accounted for 55% of the total variance represented by Factor 1. The total length of the tokens is a very important measure to explain the differences between RD and the others (NH and HI) which were represented by Factor 1. The total length of the tokens accounted for 42% of the variance in Factor 3 which represents the differences between HI and the others.

2) The length of the target names accounted for 53% of the total variance in Factor 3. This measure is very important to explain the differences between HI and the others (RD and NH). This measure also accounted for 32% of the variance in Factor 1 which represents the differences between RD and the others.

3) The length of the final part of target sentences, /desuka/ in /donokoga CVCVCV desuka?/, plays a very

important role to account for Factor 1 which represents the emotional contrast between the RD tokens and the others (NH and HI). As mentioned above, Factor 1 (F1) represents the emotional contrast between discomfort ("Awful, Rough, Uneasy, Busy, etc.") versus pleasant ("Easy, Kind, Friendly, Restful, Polite, etc.").

IV. DISCUSSION

Our previous study (Imaizumi et al, 1993c, 1994b) revealed the following results. 1) The teachers significantly reduced the devoicing rate in the HI tokens comparing to the RD and NH tokens. 2) The teachers significantly lengthened the voiced and unvoiced segment lengths in the HI tokens compared to the RD and the NH tokens. 3) Mora length accounted for 53% of the total variance in the devoicing rate. 4) When the effect of mora length on the devoicing rate was excluded, the HI tokens had a 10% lower residual devoicing rate than the NH, which was statistically significant.

Based on these results, we concluded that the teachers reduced their devoicing by not only lengthening the interval of successive voicing and devoicing gestures but also by resizing component gestures when talking to the hearing-impaired (Imaizumi et al, 1994b) .

The present study revealed that emotional profiles of the tokens are represented by four factors F1, F2, F3 and F4. F1 represents the emotional contrast between discomfort ("Awful, Rough, Uneasy, Busy, etc.") versus pleasant ("Easy, Kind, Friendly, Restful, Polite, etc."), and it also represented the emotional differences between the RD and the others (NH and HI). While, F3 represents the emotional contrast between "Slow, Stiff, Unnatural, Intelligible, Strong" versus "Busy, Lifeless, Tense, Rough, Dull". F3 also represents the differences between the HI and the others (RD and NH). F2 and F4 could be interpreted as representing speaker-dependent differences among the teachers.

The present study also revealed that the listener-adaptive temporal adjustment of dialogue speech significantly affect the emotional profiles of the speech perceived by the listeners. For instance, as shown in Fig. 4, The total length of the tokens accounted for 55% of the total variance in F1 which represents the emotional differences between the RD and the others (NH and HI). It also accounted for 42% of the variance in F3 which represents the emotional differences between the HI and the others. The length of the target names accounted for 53% of the total variance in F3 which represents the emotional differences between the HI and the others (RD and NH). The length of the final part of target sentences, /desuka/ in /donokoga CVCVCV desuka?/, plays a very important role to account for F1 which represents the emotional contrast between the RD tokens and the others (NH and HI).

Summarizing these results, we may conclude that the listener-adaptive temporal adjustment of dialogue speech affects not only the segmental characteristics, such as devoicing and vowel undershoot, but also prosody-related characteristics such as the emotional properties.

Our previous study (Imaizumi et al, 1993a,b) revealed that teachers tend to use simpler and shorter sentences, inducing more responses, toward the hearing impaired than toward the normal-normal hearing. These results indicate that a listener-oriented adaptation of speaking style affects grammatical encoding, prosody forming and segmental structuring processes.

V. CONCLUSION

Effects of listener-adaptive temporal adjustment in dialogue on emotional characteristics of speech were investigated by acoustically and perceptually analyzing the speech of teachers directed to hearing-impaired (HI) or normal-hearing (NH) children. Read tokens (RD) were also analyzed for the comparison. Acoustical analyses showed the following. 1) The emotional profiles of dialogue could be represented by four factors F1, F2, F3 and F4. 2) F1 represented the emotional differences between the read tokens (RD) and the dialogue speech (NH and HI), which were mainly the contrast between discomfort versus pleasant. 3) F3 represents the differences between the HI and the others (RD and NH), which could be interpreted as the emotional contrast between "Slow, Stiff, Unnatural, Intelligible, Strong" versus "Busy, Lifeless, Tense, Rough, Dull". 4) F2 and F4 could be interpreted as representing speaker-dependent differences among the teachers. 5) The measures of the temporal structure of speech significantly accounted for the above-mentioned emotional properties of the speech.

These results suggest that the listener-adaptive temporal adjustment of dialogue speech affects not only the segmental characteristics of speech such as devoicing but also prosody-related characteristics of speech such as the emotional properties.

Acknowledgments

We are sincerely grateful to all of the teachers and children at Ohji Elementary School, Tokyo, where speech recording was carried out. This research was partially supported by Grant-in-Aid for Scientific Research for Priority Areas on "Spoken Dialogue," Ministry of Education, Science and Culture, Japan.

References

Imaizumi, S, Hayashi, A. and Deguchi, T. (1993a). "Planning in speech production: Listener adaptive characteristics," Jpn J. Logoped. and Phoniatr. 34, 394-401 (in Japanese).

Imaizumi, S, Hayashi, A. and Deguchi, T. (1993b). "Listener adaptive characteristics in dialogue speech," in Proceedings of International Symposium on Spoken Dialogue, edited by K. Shirai et al, (Waseda Univ. Printing, Tokyo, Japan), 279-282.

Imaizumi, S, Hamaguchi, S. and Deguchi, T. (1993c). "Vowel devoicing in teachers' speech directed to the hearing-impaired/ normal-hearing children-Do teachers avoid devoicing to help hearing-impaired understand dialogue?-" Comm. Dis. Res., 22, 7-20 (in Japanese).

Imaizumi, S., Hartono, et al. (1994a). "Evaluation of vocal controllability by an object oriented acoustic analysis system," J. Acoust. Soc. Jpn (E) 15, 113-116.

Imaizumi, S, Hamaguchi, S. and Deguchi, T. (1994b). "Vowel devoicing in Japanese dialogue between teachers and hearing-impaired or normal-hearing children: Listener adaptive characteristics of dialogue speech spoduction," J. Acoust. Soc. Am., 95(5), 3012.

Picheny, M. A., Durlach, N. I., and Braida, L. (1985). "Speaking clearly for the hard of hearing: Intelligibility differences between clear and conversational speech," J. Speech and Hearing Res. 28, 96-103.

Picheny, M. A., Durlach, N. I., and Braida, L. (1986). "Speaking clearly for the hard of hearing: Acoustic characteristics of clear and conversational speech," J. Speech and Hearing Res. 29, 434-446.

Picheny, M. A., Durlach, N. I., and Braida, L. (1989). "Speaking clearly for the hard of hearing: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech," J. Speech and Hearing Res. 32, 600-603.