Speaking Style Variation and Speaker Personality

Akiko Mokhtari[1] *and Nick Campbell*[1,2]

[1]Kobe University Graduate Institute
[2]National Institute of Information and Comunications Technology
& ATR Spoken Language Communication Research Labs
Keihanna Science City, Kyoto 619-0288, Japan
akiko.mokhtari@gmail.com, nick@nict.go.jp

Abstract

This paper describes two experiments that were carried out to determine the relationships between speaking style (so-called 'tone-of-voice') and perceived speaker characteristics such as age and personality. We found that listeners can consistently distinguish different tones of voice from one speaker and that they showed a high degree of consistency in associating these tones with different speaker personality characteristics.

1. INTRODUCTION

A large number of studies have shown the correlation between features of speaking style and speaker's internal state or emotion [1, 2, 3]. It is also well known that acoustic features such as duration, pitch, vowel formant and voice quality are commonly used as a vehicle for conveying paralinguistic and extralinguistic information, which is to be clearly distinguished from uncontrolled displays of emotion [4, 5]. In normal interactive social communication speakers have a choice of speaking style and we show in this paper that they consistently match their tone of voice to a specific partner in order to display particular characteristics that may be pertinent to the given relationship. For example, a person will typically exhibit different behaviour, including different speaking-style characteristics, when carrying out the role of young mother than when performing other roles such as disco-dancer, hostess, daughter of a judge, or wife of a professor. We take it for granted that she will act differently in each of these roles, but are concerned here to determine the extent that her voice qualities and speaking style characteristics also change.

2. EXPERIMENT 1

We performed a series of experiments using speech data collected from the JST/ATR ESP Expressive Speech Processing Corpus [6]. In one subset of this corpus, there are speech samples collected from a series of telephone conversations between adult volunteers who were originally strangers, but who became friends throughout the period of the recordings. These samples vary considerably in expressivity and speaking style as each speaker interacts with different conversational partners.

We selected a semantically neutral set of 40 utterances from the recordings of one female speaker, consisting solely of the word "hai" (which functions similarly to "yes" in English), and prepared an experiment in which listeners were first asked to classify these utterances according to speaking style and then to identify the common characteristics of each style by means of a questionnaire. The speech samples were selected to be representative of 4 different speaking styles and voice characteristics, and there were ten different tokens of each type.

2.1. Classification

The participants were 10 Japanese speakers (8 males and 2 females) who were unfamiliar with the voices used in the experiment. They used the "Mover" software, a graphical interface written in the Tcl/Tk programming language using Snack (see figure 1) to listen to the individual stimuli (that were actually all produced at different times by the same one speaker) and to classify them into 4 groups according to perceived 'speaker identity'. In other words, participants were led to believe that the samples came from different speakers, and they had to group together the stimuli that they judged as being spoken by the same speaker by placing them into named boxes on the computer screen. The boxes were labelled A, B, C, and D, to represent different individual speakers. Participants were allowed to listen to the samples as many times as necessary, and to change their decisions freely until they were satisfied with the final classification. No restrictions were placed on the number of tokens for each box.

The initial state of the software interface is shown in the left part of Figure 1, with 40 movable circles representing the stimuli aligned in random order along the main diagonal, and 4 boxes (Box-A, Box-B, Box-C, Box-D) are pre-placed separately in the upper part of the screen. Participants were first required to listen to each stimulus by clicking with the computer's mouse within the circle in order to determine which stimuli were pro-

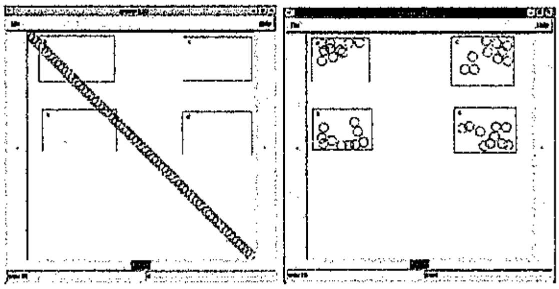

Figure 1: Mover software. Left shows initial state, right shows results after categorisation.

duced by which speaker (A – D). Each circle was then dragged (moved by the mouse) to one of the 4 boxes as shown in the right part of Figure 1. Participants were able to listen to the stimuli repeatedly without any time restrictions.

2.2. Questionnaire

The participants were subsequently required to answer a questionnaire about the perceived traits of each speaker (A – D) as they had been classified. Five personality traits were taken from the Big-Five Personality Inventory [7] and two extra categories were added, giving seven classes in all: (1) age group, (2) occupation, (3) "open", (4) "cooperative", (5) "sincere", (6) "sociable", (7) "calm"). Answers for (1) and (2) were categorically selected from prepared options, Participants were allowed to listen again to the stimuli as much as necessary.

2.3. Results and Discussion

Although participants were free to assign "speakers" to any of the boxes A – D, there was considerable agreement in within-box groupings so we subsequently manually grouped the boxes produced by the various participants so that they were uniformly identified by the same letter. Table 1 shows the degree of agreement in this initial classification.

2.3.1. *Classification*

Table 1 shows how how participants classified the stimuli into which box after renaming the individual boxes as described above. All numbers across any row add to ten in each quadrant. Classification results from the 10 participants were compared with the expected classification determined the authors prior to the experiment.

Participants are represented in the table by P1-P10 in the left column. . For example, P1 in the upper-left quadrant classified 8 out of 10 stimuli to Box-A, but 1 to Box-B and 1 to Box-C, indicating a good but not perfect agreement with the expected classification. Similarly, other quadrants correspond to Box-B, Box-C, and Box-D respectively. Agreement between participants was measured statistically by kappa. Moderate agreement was obtained for the total classification ($Kappa = 0.556$).

Table 1: Showing agreement in speaker categories as determined by the 10 participants

	A	B	C	D	-	A	B	C	D
	A	-	-	-		-	B	-	
P1	8	1	1	0	-	1	3	2	4
P2	9	1	0	0	-	5	5	0	0
P3	9	1	0	0	-	2	5	2	1
P4	9	1	0	0	-	0	10	0	0
P5	9	1	0	0	-	2	7	0	1
P6	9	1	0	0	-	0	10	0	0
P7	9	1	0	0	-	0	10	0	0
P8	0	1	1	8	-	4	6	0	0
P9	9	1	0	0	-	3	3	0	4
P10	6	2	2	0	-	2	8	0	0
	-	-	C	-		-	-	-	D
P1	1	5	3	0	-	0	1	3	5
P2	0	0	9	0	-	0	0	0	9
P3	0	3	6	0	-	0	0	0	9
P4	0	1	8	0	-	0	1	0	8
P5	0	0	9	0	-	0	1	0	8
P6	0	0	9	0	-	0	1	0	8
P7	0	0	9	0	-	0	1	0	8
P8	3	0	6	0	-	1	8	0	0
P9	0	0	9	0	-	1	0	0	8
P10	0	0	9	0	-	0	1	0	8

2.3.2. *Questionnaire*

Factor (1) "age group" consists of five 5 levels (15-25, 26-35, 40s, 50s, and over-60). As we can see from Figure 2, the stimuli that had actually been spoken by only one speaker were attributed to various differently aged 'speakers' by the participants. This confirms that they might actually have believed the speech samples to have been spoken by different people. Stimuli that were classified as Type-A and Type-B were attributed to relatively older speakers (40s, 50s, over 60s). On the other hand, stimuli that were classified as Type-D were attributed to a much younger speaker (16-25, and 26-35).

Factor (2) "occupation" consists of 7 items ('student', 'single/worker', 'housewife without children', 'housewife with children', 'housewife/ worker', 'other', and 'don't-know'). Although the answers varied between participants, there were tendencies that were clearly linked to factor (1) "age". For instance, "housewife with children" was selected by about half of the participants for

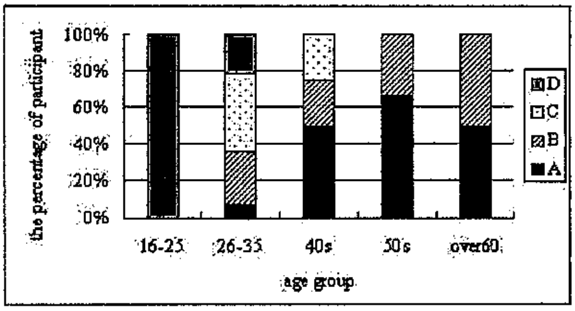

Figure 2: Speaker's perceived age.

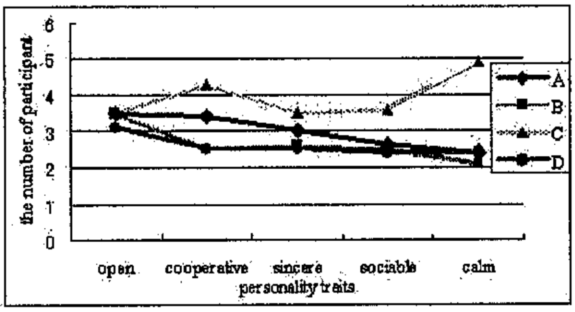

Figure 3: Speaker's perceived personality traits.

stimuli that were classified to Type-A and Type-B, while "student" and "single/worker" were selected for about half of those that were classified to Type-D. Stimuli in Type-C were positioned midway between Type-A/Type-B and Type-D. Items regarding personality trait, (3) "open", (4) "cooperative", (5) "sincere", (6) "sociable", and (7) "calm", were rated on 7-point scales as shown in Figure 3.

3. EXPERIMENT 2

Because of the variability observed in the results for voice type-B, we repeated the experiment using only three voice types, with different samples similar to those of Type-A, Type-C, and Type-D. The stimuli were of three types, ten tokens each, 30 utterances in all. The participants were 10 Japanese speakers (3 males and 7 females) who had never previously heard the speaker's voice. None of them had participated in the first experiment. Like the first experiment; this experiment also consisted of 2 parts: Classification and Questionnaire.

3.1. Classification

The same procedure was carried out, using the same software as the first experiment. In the second experiment, however, participants were required to classify each of the 30 stimuli into one of 3 types. Participants were allowed to listen to the stimuli repeatedly without any time restriction.

Table 2: Speaker's perceived lifestyle as determined by Experiment 1.

	A	B	C	D
student	0	1	0	4
single/worker	0	0	3	4
housewife/mother	0	1	1	0
housewife no children	4	6	2	0
working housewife	2	0	1	0

Table 3: Voice-quality classsification results from Experiment 2.

	E	F	G	E	F	G	E	F	G
	E	-	-	-	F	-	-	-	G
P1	10	0	0	3	7	0	1	0	9
P2	10	0	0	0	9	1	0	1	9
P3	10	0	0	1	8	1	0	1	9
P4	10	0	0	6	4	0	0	0	10
P5	5	5	0	0	9	1	0	1	9
P6	10	0	0	2	6	2	0	0	10
P7	10	0	0	1	8	1	0	0	10
P8	10	0	0	0	8	2	1	1	8
P9	9	1	0	2	7	1	0	2	8
P10	10	0	0	0	9	1	2	0	8

3.2. Questionnaire

Participants were required to answer a questionnaire about the perceived speaker-personality traits for each type which they classified. The same factors were used as in the first experiment: (1) age group, (2) occupation, and 5 personality traits: (3) "open", (4) "cooperative", (5) "sincere", (6) "sociable", (7) "calm".

3.3. Results and Discussion

3.3.1. Classification

Classification results from 10 participants were compared with the classification anticipated by the authors as shown in Table 3. Here, types A, C, and D from the first experiment were renamed as types E, F and G respectively. Agreement between participants was measured statistically by kappa. As a result, almost perfect agreement was obtained for for total classification ($kappa = 0.887$).

3.3.2. Questionnaire

For item (1) of the questionnaire, we found that stimuli in each type were attributed to distinctively different aged speakers as shown in Figure 4. The stimuli that were classified to Type-E were judged as being spoken by a

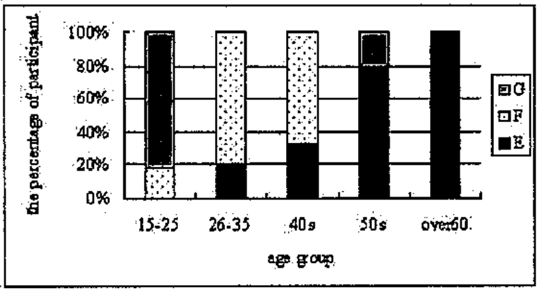

Figure 4: Speaker's perceived age, Experiment 2.

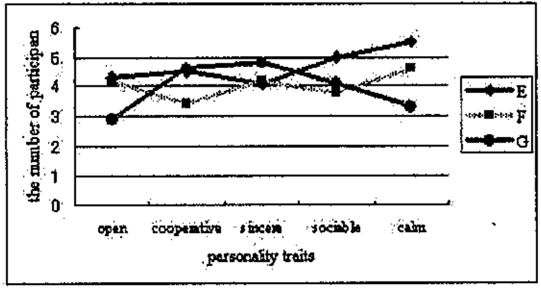

Figure 5: Speaker's personality traits, Experiment 2

relatively older speaker (40s, 50s, 60s). Note that "50s" or "over 60s" were chosen by most of participants. The stimuli of Type-G were attributed to a young speaker (16-25 yrs-old) except for one participant who selected an age of 50. Similarly for factor (2) "occupation", which consists of 7 items (student, single/worker, housewife without children, housewife with children, housewife/worker, other, unknown), the answers varied much less between participants, and there were tendencies that were more strongly linked to item (1) "age". For instance, "housewife with children" was selected by more than half of the participants for stimuli that were classified to Type-E, while more than half opted for "student" for stimuli that were classified to Type-G. Results for Type-F were not as strong as we expected and display more variability.

4. DISCUSSION

Results for experiment 2 were as expected, stronger than those for experiment 1. This may be explained statistically as being due to fewer distractors. However, we have seen that in both experiments, with different sets of participants (actually graduate students from different universities) that in each case there was stong agreement in the classification of 'speaker' based on voice quality. Furthermore, we see that participants agree strongly on (an actually mistaken) classification of speaker age and occupation. We conclude from this that it is valid to claim that speakers show different personalities when changing their speaking styles.

However, results shown in Figures 3 and 5 show that participants are not so confident in their assessment of deeper personality traits from the evidence of such very short speech samples. One word is probably not enough to judge a person's character.

Table 4: Perceived speaker lifestyle, Experiment 2

	E	F	G
student	0	2	7
single/worker	1	3	2
housewife/mother	0	2	0
housewife (no children)	7	1	1
working housewife	0	2	0

5. CONCLUSION

This study examined the listener's perception of age and personality from a small number of acoustic samples taken from spontaneous natural conversations. The listeners were given to believe that the samples were spoken by a number of different people and were required first to identify and group utterances from the same speaker, and then to describe that speaker's age and personality features. In fact, all speech samples were from the same speaker but differed considerably in speaking style.

We showed that listeners were reliably able to distinguish speech samples into groups related by speaking style, identifying them as different speakers, and then that they were consistently different in their judgements of speaker age and occupation.

Current and future work aims to extend this experiment with the voices of more than one speaker, longer samples, and the inclusion of distractors that are not clearly belonging to any one given speaking style.

6. References

[1] Williams, C., Steavens, K., 1972. Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. 52-4, 1238-1250.

[2] Sherer, K., Ladd, R., Shilverman, K. 1984. Vocal cues to speaker affect: Testing two models. J. Acoust. Soc. Am. 76-5, 1346-1356.

[3] Mozziconacci, S. 2002, Prosody and emotion. Proc. Speech Prosody, 1-9.

[4] Campbell, N. 2003. Voice quality: the 4th dimension. Proc. ICPHS, 2417-2420.

[5] Maekawa, K. 1998. Phonetic and phonological characteristics of paralinguistic information in spoken Japanese, Proc. ICSLP, 635-638.

[6] The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.atr.jp

[7] McCrae, R. R. & Costa, P. T. (1990). *Personality in adulthood.* New York: The Guildford Press.