REPRESENTATION OF VOICE QUALITY FEATURES ASSOCIATED WITH TALKER INDIVIDUALITY

Hiroshi Kido[†‡] *and Hideki Kasuya*[†]

[†] Faculty of Engineering, Utsunomiya University, Utsunomiya
[‡] National Research Institute of Police Science, Tokyo

ABSTRACT

As a first step toward development of a "speech montage system", this paper attempts to derive a core set of Japanese epithets which are commonly used in an everyday life to represent voice quality features associated with talker individuality. Perceptual experiments were conducted, where subjects were asked to evaluate sentence utterances recorded from a variety of male speakers in terms of 25 epithets which were derived in another experiment [1] to be indicative of voice quality relevant to talker individuality. The evaluation scores were subjected to a statistical clustering analysis. The analysis resulted in that the 25 epithets could be grouped into either eight categories for male or seven for female subjects. These categories were basically the same as those obtained in the previous experiment [1] where subjects were required to evaluate their own voice with the same set of 25 epithets. Agreement between the results from the two experiments guarantees reliability of the core epithet categories to represent voice quality associated with talker individuality.

1. INTRODUCTION

In exploring a "speech montage system", extraction of core epithets which are commonly used in an everyday life of people to represent voice quality associated with talker individuality is a key element. Epithets expressing sound quality were extensively studied in such non-speech areas as noise control and music acoustics, where three key epithets relevant to beauty, powerfulness and metallic quality of a sound were most often extracted [2-7]. In speech areas, on the other hand, research has primarily focused on evaluation of perceived quality of pathologic voice [8-10], synthetic speech [11-13] and coded speech [14,15]. No systematic work has been made to derive a core set of Japanese epithets, which are commonly used in an everyday life of people to represent voice quality features associated with talker individuality. In this paper, we propose a systematic procedure for that purpose in which combined use of perceptual quality evaluation and statistical clustering analysis plays an important role.

Speech conveys not only linguistic but also para- and extra-linguistic information [16,17]. Perceived talker individuality is obviously included in extra-linguistic information and is a function of physical characteristics of an utterance, mental status of a listener and acoustic environment where he/she is (see Figure 1). The physical characteristics are then represented as the "product" of the voice source and the vocal tract characteristics of the talker, both of which are also functions of physical properties inherent to the talker, acoustic

Perceived talker individuality: $I = f_I(v_t, e_t, m_l)$
v_t: Physical characteristics of the utterance.
c_t: Acoustic environment.
m_l: Mental condition of the listener.
where $v_t = s_t * d_t$.
Voice source features:
$s_t = f_{st}(p_t, e_t, m_l)$
Articulatory features:
$d_t = f_{dt}(p_t, c_t, m_l)$
where
p_t: Physical characteristics inherent to the talker.
c_t: Acoustic environment.

Figure 1: Sources of perceived talker individuality.

thick (*futoi*)	frightening (*dosu-no-kiita*)	nasal (*hana-goe*)
subdued (*shibui*)	calm (*ochitsuki-no-aru*)	hoarse (*kasureta*)
strained (*dami-goe*)	dysphonic (*tsubureta*)	sonorant (*hibiki-no-aru*)
clear (*sunnda*)	powerful (*hakuryoku-no-aru*)	penetrative (*toori-no-yoi*)
high-pitched (*takai*)	rough (*garagara*)	weak (*yowayowashii*)
bright (*akarui*)	childish (*kodomoppoi*)	cute (*kawaii*)
sexy (*iroppoi*)	vivid (*ikiiki-to-shita*)	tense (*hari-no-aru*)
feminine (*joseitekina*)	youthful (*wakai-kanjino*)	refined (*hin-no-aru*)
screaming (*kanakiri-goe*)		

Table 1: 25 epithets for voice quality features relevant to talker individuality.

thick	nasal	powerful
high-pitched	calm	hoarse
tense	youthful	weak
feminine		

Table 2: Ten epithets extracted by statistical clustering analysis.

environment where he/she talks and his/her mental conditions. In this paper, we begin with controlled conditions of the environmental and mental situations that respectively restrict to direct speech communication in a soundproof room and neutral mentality. Individuality information is therefore conveyed primarily by physical characteristics of the talker's speech organ and their usage.

In our previous work [1], 25 Japanese voice quality epithets which were commonly used to represent talker individuality were selected from 288 candidate epithets by conducting systematic questionnaires (see Table 1). Ten core categories as shown in Table 2 were finally extracted from 25 epithets through multivariate analysis of voice quality evaluations of subjects' own voice. Reason for the adoption of the self-evaluation method of subjects' own voice was to collect as many voice quality evaluation data of a variety of voices as possible and to know correlation between each of the epithets.

In this paper, on the other hand, perceptual evaluation experiments are performed on sentence utterances read by a large variety of talkers to analyze statistical tendency of the judgments made by subjects. The statistical clustering method is again used to extract core categories and the results are compared with those of the previous work.

2. PRELIMINARY EXPERIMENT

In order to obtain reasonable data from a perceptual experiment, in general, a scale of the experiment should be carefully examined, since too much burden, *e.g.* long listening session, often deprives the subjects of concentration. In a preliminary experiment, therefore, we attempted to select a necessary and sufficient number of talkers for the voice quality evaluation experiment.

2.1. Epithets to represent voice quality

As mentioned above, we have already extracted ten core epithets to represent voice quality associated with talker individuality [1]. Every epithet except for one relevant to nasality is bipolar or in pairs with its own antonym and represented in Table 3. Since subjects, who were not specialized in speech science, mostly failed to distinguish open nasality from closed nasality, a mono-polar scale was set on the nasality, thereby containing both features.

adjective(epithet) pairs
high-pitched — low-pitched
masculine — feminine
hoarse — clear
calm — excited
powerful — weak
youthful — elderly
thick — thin
tense — lax
unipolar adjective(epithet)
nasal

Table 3: The epithet pairs (bipolar) and mono-polar epithet.

2.2. Speech materials

We first recorded speech utterances from 109 males, each of whom read in a soundproof room five sentences shown in Table 4. All the sentences are declarative and rather neutral with respect to emotional expression. Out of 109 talkers, we carefully selected 36 talkers, taking the ten voice quality features shown in Table 3 into account. Most of them were thought to have one or some dominant voice quality features of the ten categories. One additional talker who was regarded as having average scores on the ten features was also selected as a reference talker.

On the desk, there are many books of a blue hard cover.
(*Tsukue no ueni aoi haadokabaa no hon ga nansatsu mo arimasu*)
On the wall, the picture of a lake was hanging.
(*Kabe niwa mizuumi no e ga kakatte imashita*)
That picture is now at sister's home in Aomori.
(*Ano e wa ima Aomori no ane no ie ni arimasu*)
Tsugaru-plane has unique climates and dialects.
(*Tsugaru-heeya wa dokutoku no huudo to hoogenn wo motte imasu*)
A little earlier, I have just drawn the Kojien dictionary.
(*Sakihodo, Koojien wo hiite shirabeta tokoro d*esu)

Table 4: Five sentences used for the perceptual experiments.

2.3. Perceptual evaluation

90 subjects participated in the experiment. They were again carefully selected so that age and sex of the subjects were balanced well. 90 subjects were grouped into four, each comprising 15 subjects. Each group evaluated the utterances of seven talkers (six test- plus one reference-talkers) for each of the nine epithets in a minute. The evaluation was made on the seven-category scale for all the epithets except for nasality for which four-category was assigned.

2.4. Results

A clustering analysis was made on the perceptual evaluation data. A dendrogram obtained by the analysis is shown in Figure 2, where the abscissa is dissimilarity distance between the talkers when the utterance is evaluated with the nine epithets and the ordinate indicates the talker identification number. Thick lines with arrows are for the dominant talkers who show the highest score on one or some of the nine epithet scales. By setting the dissimilarity distance threshold at around 40, 19 talker categories including the reference talker were obtained. A value of 40 was chosen so that all the dominant talkers are

classified into the different categories. The 19 talkers were selected from each of the 19 categories and subjected to the following experiment.

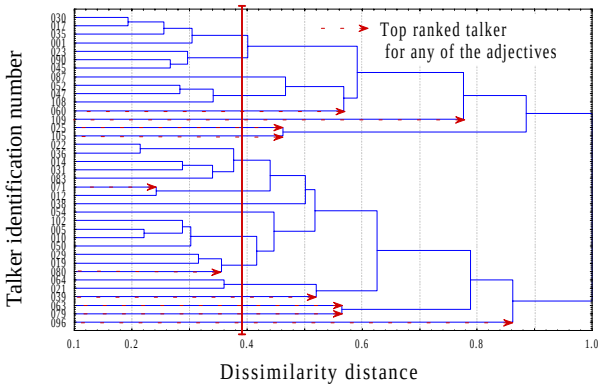

Figure 2: The dendrogram of 37 talkers.

3. EXPERIMENT

Using the utterances of the 19 talkers selected in the preliminary experiment, a similar perceptual experiment was performed for 25 epithets shown in Table 1 and a statistical clustering analysis was again made on the perceptual evaluation data.

3.1. Perceptual evaluation

90 subjects were again selected and divided into six groups, each of which contained eight males and seven females. In making the group, age of the talkers was also taken into account so as to make balance in each group. Each group evaluated the utterances of three test- plus one reference-talkers in terms of 25 epithets for the voice quality on the mono-polar seven-category scale. The subjects spent three minutes this time to complete the evaluation for 25 epithets on the scale.

3.2. Statistical analysis

3.2.1. Statistical test of difference between male and female subjects

Since the statistical U-test of Man-Whitney on the difference between the perceptual data of male and female subjects showed significant sexual difference, the clustering analysis was made separately for the male and female subjects. The female subjects also showed less variability than the males.

3.2.2. Correlation between the epithets

The Goodman-Kruskal's gamma coefficient (similar to Spearman's correlation coefficient by ranks) between the 25 epithets was computed and then converted to dissimilarity distance measure by subtracting gamma from unity. Results of the clustering analysis are illustrated in Figure 3 for the male subjects and in Figure 4 for the female subjects.

3.2.3. Usefulness of the epithet

Even if an epithet is statistically separated from the others, it would be useless without having a significant range of variations from talker to talker. We then examined significance of the difference between the talkers using Kruskal-Wallis test. As a result, "screaming" and "sexual" voice qualities were not significantly different between the talkers in case of the male subjects.

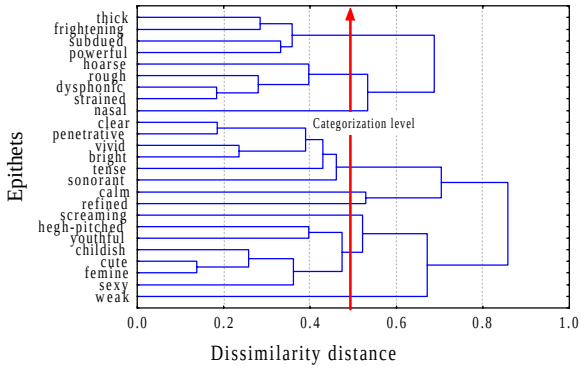

Figure 3: The dendrogram of 25 epithets (male subjects).

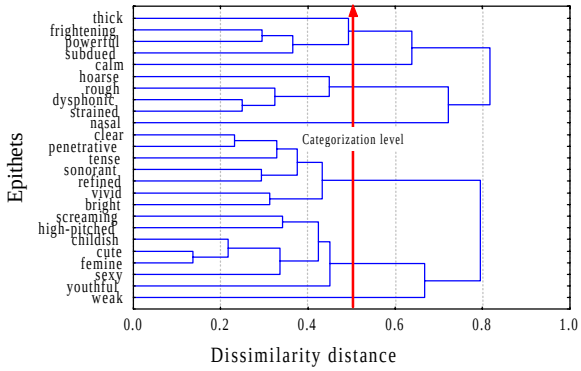

Figure 4: The dendrogram of 25 epithets (female subjects).

3.3. Discussion

If a dissimilarity distance threshold is set at 0.5, then nine- and seven-epithet categories are obtained for the male and female subjects, respectively. By excluding "screaming" from the categories in the male, we have eight- and seven-epithet categories for the male and female subjects, respectively, as shown in Table 5. Difference between the male and female subjects lies only in "refined". Table 5 agrees basically with the categories extracted in the previous work [1] except that "high-pitched" in this experiment includes both "feminine" and "youthful" in [1] and "thick" in this work corresponds to "powerful" in the previous one [1]. These are said to be rather small difference. We then can conclude that the core epithet categories extracted are quite reliable in describing voice quality associated with talker individuality.

4. CONCLUSION

We have extracted Japanese core epithets which are commonly used in an everyday life of people to represent voice quality associated with talker individuality, as a first step toward the development of a "speech montage system." Based on the carefully designed perceptual experiments, we derived eight and seven Japanese core epithets for male and female listeners, respectively. These are basically the same as the ones obtained in the previous work [1] where the self-evaluation was made on subjects' own voice. This agreement implies reliability of the extracted epithet categories.

Future study includes selection of antonyms to each of the core epithets and exploring of their acoustic correlates.

Male subjects	Female Subjects
thick	thick
hoarse	hoarse
nasal	nasal
tense	tense
calm	calm
weak	weak
high-pitched	high-pitched
refined	

Table 5: The adjectives extracted by perceptual evaluation experiments.

5. ACKNOWLEDGMENT

The authors would like to thank the members of Finger Print Center, Criminal Investigation Bureau, National Police Agency (NPA), for their assistance in perceptual experiments, the colleagues of Utsunomiya University and the members of National Research Institute of Police Science, NPA, for their kind cooperation in recording speech samples. The paper was in part supported by Grant-in-Aid for Scientific Research, Ministry of Education, Sports and Culture (1087051).

6. REFERENCES

1. H. Kido, H. Kasuya, "Japanese voice quality epithets associated with talker individuality," Acoust. Soc. Jpn. Tech. Report, H-97-77, 1997 (in Japanese).

2. T. Sone, K. Kido, T. Nimura, "Factor analysis of Descriptive Epithets," J. Acoust. Soc. Jpn. 18, 320-326, 1962 (in Japanese).

3. O. Kitamura, S. Namba, S. Sannohe, "Psychological evaluation of replayed sounds," Tech. Report., Electro-acoustics, IECE of Jpn.. 1962 (in Japanese).

4. O. Kitamura, S. Namba and R. Matsumoto, "Factor analytical research of tone color," Proc. the 6th ICA, A-5-11, 1968.

5. Seiichiro Namba, "Definition of timbre," J. Acoust. Soc. Jpn., 49, 823-831, 1993 (in Japanese).

6. Y. Suzuki, K. Ozawa, T. Sone, "Timbre perception of steady sound," Proc. Fall Meeting Acoust. Soc. Jpn., 623-626, 1995 (in Japanese).

7. T. Sueoka, K. Ohgushi, T. Taguti, "Subjective evaluation of piano performances and their physical correlates," J. Acoust. Soc. Jpn., 52, 333-340, 1996 (in Japanese).

8. B. Hammerberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," Acta Otolaryngol. Vol.90, 441-451, 1980.

9. H. Kasuya, "Acoustic correlates of voice qualities," Proc. Fall Meeting Acoust. Soc. Jpn., 619-622, 1993 (in Japanese).

10. S. Imaizumi, "Pathological voice quality," J. Acoust. Soc. Jpn., 51, 887-892, 1995 (in Japanese).

11. N. Higuchi, S. Yamamoto, T. Shimizu, "Evaluation of intelligibility and naturalness of the synthetic speech generated with a Japanese speech synthesizer by rule," IEICE Trans., J72-D-II, 1133-1140, 1989 (in Japanese).

12. H. Kasuya, "Assessment of speech synthesis technology," J. Acoust. Soc. Jpn., 48, 46-51, 1992 (in Japanese).

13. H. Kasuya, "Methods to evaluate synthetic speech quality," J. Acoust. Soc. Jpn., 49, 866-870, 1993 (in Japanese).

14. E. Miyasaka, "Subjective assessment of sound with small impairments," J. Acoust. Soc. Jpn., 51, 806-811, 1995 (in Japanese).

15. M. Miyahara, Y. Morita, "Investigation and analyses of words for assessment of tone quality," J. Acoust. Soc. Jpn., 52, 516 - 522, (1996) (in Japanese).

16. H. Kasuya and C.-S. Yang, "Voice quality associated with voice source," J. Acoust. Soc. Jpn., 51, 869-875, 1995 (in Japanese).

17. J. Laver, The phonetic description of voice quality, Cambridge University Press, Cambridge, 1980.