Cross-modal description of sentiment information embedded in speech

Kanako Watanabe*, Yoko Greenberg† and Yoshinori Sagisaka*†

* Department of Pure and Applied Mathematics, Graduate School of Fundamental Science and Engineering, Waseda University, † GITI, Waseda University

E-mail: knnnnako37@gmail.com, greenberg.yoko@gmail.com, ysagisaka@gmail.com

ABSTRACT

Looking for new possibilities to describe the information embedded in speech, we have carried out sentiment correlation analysis between speech features and color attributes. Using single vowel utterances with different prosody and sound pressure level, we have asked subjects to select colors based on their perceptual impressions after listening them. By analyzing selected color attributes using Value, Saturation and Hue, we found high correlations between mean F0 and Value, sound pressure level and Saturation, and Formants and Hue. These correlations coincided with previous observations using speech and color categories, which suggests a possibility for visualization of sentiment information embedded in speech based on cross-modal sentiment correlations.

Keywords: speech-color correlation, speech description, sentiment information, cross-modal information expression, paralinguistic information

1. INTRODUCTION

In speech science, speech variations have not yet been well described. In traditional studies, variations such as colloquial speech intrinsic phenomena and quite sentimental differences have been excluded from the theoretical research targets [1]. This restriction of speech variations has nicely worked to find out underlying fundamental principles without being bothered by treating too many speech variations in a real field. However, we still do not have a good description scheme to distinguish between read speech such as "Thank you very much" or "I am very sorry" and communicative speech of them uttered in daily life.

We have been studying these differences. In particular, we have been trying to characterize and specify communicative speech prosody to enable speech synthesis and recognition for more than a decade [2]-[12]. Through these studies, we confirmed that communicative prosody variation can be formulated using its perceptual impression as a descriptor to analyze its variations and to computationally specify synthesis target. Using Multi-Dimensional Scaling, we confirmed that some of communicative speech characteristics can be described in three-dimensional psychological space representing perceptual impressions. High correlations have been found between impressions and prosodic features [3]. The positive–negative impression corresponds to high-low F0, both confident–doubtful and allowable–unacceptable impressions correspond to F0 dynamics. Using these correlations, we have succeeded in computing communicative F0 variations using impression attributes of lexicons constituting output speech [9][12].

From the previous analyses, we found that sentiment information description is the important first step to enable prosody computation. Through pursuing this idea, we noticed that we are using color as a descriptor for our sentiment impressions for speech sounds such as *vowel color, light [l]* and *dark [l]*. On the other hand, for color description, we need a supplementary word to specify color variations more precisely such as *emerald green, burgundy red* and *ivory white.* As these facts indicate, categorical classes specified by language can only provide approximated sentiment description of speech or color but cannot specify them exactly.

By considering the above insufficiencies of sentiment description of speech variations using linguistic expressions, we felt the need to scientifically understand these different information media, language, speech, and vision to fully use their advantage. In this paper, we introduce our experimental results investigating the sentiment correlations between speech and color. In the next Section 2, previous studies are introduced on the relationships between speech and color. In Section 3, we explain our two color-selection experiments to find out sentiment correlation between colors and vowels with different prosody. In Section 4, we discuss on the possibilities to describe speech variations using other information media. In section 5, we summarize the results and show further possibilities of cross-modal description for speech variations.

2. CORRELATIONS BETWEEN SPEECH AND COLOR

By looking back the history of phonetics on the correlation between speech and color, we found

Roman Jacobson's work [13]. He has pointed out the fact that people called *synesthetes* who have *color hearing* special ability to feel color when they hear sounds tend to select darker colors for /o/ and /u/, brighter colors for /e/ and /i/, and red for /a/. Recently, Magdalena et al have found that Polish subjects have a tendency to image reddish color by listening vowel /a/, yellowish and greenish for /i/, greenish and reddish for /e/, brown, black, or blue for /o/ and /u/ [14]. The subjects participating in the color-selection experiment were non-specific ordinary Polish students learning English as the second language.

If we expand the sentiment correlations between *sound* and color, Ward et al has shown that there is a general tendency to associate high pitch sounds with light colors and low pitch with dark colors [15]. Nagata et al have also studied correlations between sound and color for *synesthetes* [16] [17]. Through their studies, they found that the same correlation could be observed in the responses of ordinary people.

As above facts indicate, we may be able to expect color as a new descriptor of speech variations. Aiming at finding a new possibility to directly describe speech sentiment information using color, we designed correlation analyses between vowels with various prosodies and selected colors corresponding to their perceptual impressions. If we can find any consistent correlations between them, we will have a new possibility of multimodal information expression for speech variations.

3. EXPERIMNTS ON COLOR SELECTION BASED ON PERCEPTUAL IMPRESSIONS ON VOWELS WITH DIFFERENT PROSODY

3.1. Design of two sets of speech stimuli

As the first set of speech stimuli to find correlations between their variations and colors selected by perceptual impressions, we decided to employ five Japanese single vowels /a/, /i/, /u/, /e/, and /o/ with prosody variations uttered by a native Japanese male. For prosody variations, we adopted typical communicative prosody variations consisting of three tone heights (high, middle, low) and four dynamics (flat, rising, falling, rising followed by falling) which have been frequently observed in interjections such as "uhm" of daily communications and employed in our previous studies on communicative prosody [3][4]. For this experiment, we used flat sound pressure level of 60dB for all speech samples.

In addition to the first the speech set, we collected the second one to supply speech variations in sound pressure level (SPL). After preliminary analysis [18], we noticed that loudness changes might affect perceptual impressions. For speech samples, we collected the same three F0 average heights (high, middle, low) used in the first set with three SPLs. These speech samples were uttered by another native Japanese male different from the first set.

3.2. Color system for evaluation

For color selection experiment, we adopted Practical Color Co-ordinate System (PCCS) employed in the study on correlations between sound and color [16]. PCCS classifies and displays colors by three attributes of color, Hue, Saturation, and value. It consists of 153 color tips, 144 chromatic colors and 9 achromatic colors. All these tips are selected in psychologically same interval and ordered in a plane. Horizontal axis and vertical axis respectively correspond to Hue and Tone, a mixed concept of Saturation and Value. Hue consists of four primary colors (red, blue, green, and yellow) and the complementary colors of them, supplemented by additional four Hues to arrange them in the same sentiment intervals. Tone shows 12 particular conditions or atmosphere of color like ”vivid”, “soft”, and “deep”.

To present impartial and comprehensive color choices, PCCS is supposed to be the most suitable color system since it consists of almost all colors at same psychological intervals fitting to human perceptions for color. In addition, two-dimensional color display is easier for examinees to evaluate than other three-dimensional color systems like Munsell color system.

Numerical color systems can be divided into two groups by which color style it represents, self-luminous color or object color. PCCS represents object color. Among systems for expressing self-luminous color, Hue-Saturation-Value color system (HSV) is very common and convenient when operating visual media numerically. This system expresses color also by three attributes of color. Hue shows types of color including red, orange, yellow, green, blue, and purple. Saturation relates to the degree of unmixed, clear color has high saturation and somber one has low. Value expresses the lightness. While PCCS expresses object color with some words and numbers, HSV represents luminous color numerically in ratio scale. To obtain correlations between speech parameters and color parameters quantitatively, we converted selected PCCS tips to HSV scale and calculated average of them for each speech sample.

3.3. Experimental setup

Experiments were conducted independently by each subject in quiet listening conditions.

Participants were asked to be seated in front of a computer screen and were instructed to listen to individual speech stimuli presented in random order. Then subjects were asked to choose the most suitable color for their imagined-color from 153 colors at a time after listening one vowel sound. We allowed subjects to select at most three colors and to repeat a vowel sound listening before selecting colors. Color selection experiments were carried out for two speech sets independently by 63 Japanese natives (31 males and 32 females) and 10 Japanese natives (5 males and 5 females) respectively.

3.4. Results on correlation analyses between speech features and color attributes

Throughout color selection of these two speech sets, we found very clear high correlations between the following speech features and color attributes.

(1) F0 heights and Value
(2) SPL and Saturation
(3) Vowel categories and Hue

(1) F0 heights and Value

In both of two speech sets, quite high correlations were found between average F0 heights and Value. In the first set, average correlation scores of each vowel category are 0.88, 0.91, 0.95, 0.85 and 0.95 for /a/, /i/, /u/, /e/ and /o/ respectively (0.91 in average). Figure 1 shows the individual correlation for each vowel we got from this experiment. In the second set, the correlation scores were 0.83, 0.94, 0.76, 0.77, and 0.93 for /a/, /i/, /u/, /e/ and /o/ respectively (0.85 in average). The average Values for vowels /i/ and /e/ were higher that those for /u/ and /o/, which coincides with previous findings by Jakobson [13].

(2) SPL and Saturation

In the second set, the average Saturation value highly correlates to average SPL. Figure 2 shows their correlation scores 0.81, 0.75, 0.84, 0.76, and 0.96 for /a/, /i/, /u/, /e/ and /o/ respectively (0.82 in average). The distributions of Saturation differed from each other vowel categories. The average Saturation value was ordered from front vowels to back vowels 71% (/e/), 67% (/i/), 66% (/u/), 61% (/a/), and 58%/ (o/) in descending order.

(3) Vowel categories and Hue

Though multiple Hues have been selected for each vowel, their distributions are quite vowel dependent. Table 1-(1) and Table 1-(2) show selected Hue ratio of each vowel for two speech sets respectively. For this calculation, 11 Hues based on basic color terms are employed to compare the results. These two tables show very similar selection characteristics. /a/ is tended to associated with reddish colors like red and orange, /i/ with yellow or green, /u/ with green or blue, /e/ with green or orange, and /o/ with green, blue, or purple.
These correspondences not only nicely coincide with the previous results but also more clearly show that

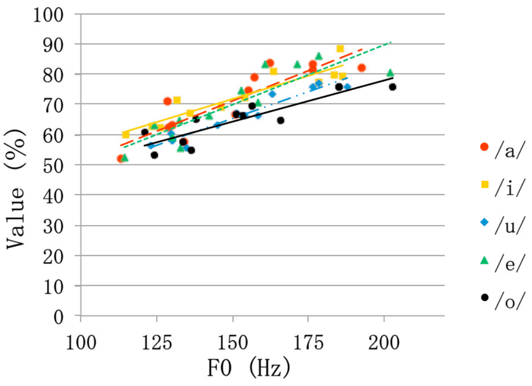

Figure 1: Correlations between F0 heights and Value.

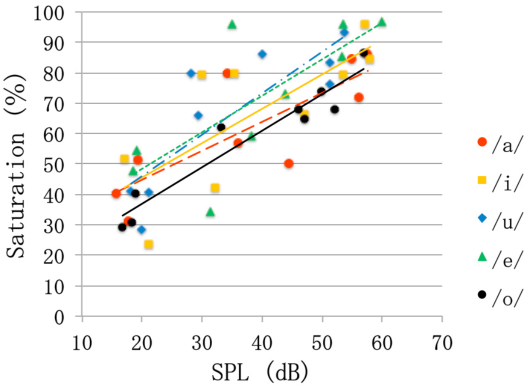

Figure 2: Correlations between SPL and Saturation.

Table 1-(1): Hue ration selected in the first experiment.

	/a/	/e/	/i/	/o/	/u/
Red / Pink	19	5	4	3	4
Orange	19	16	13	11	12
Yellow	7	13	25	8	8
Green	14	29	24	23	29
Purple	14	13	12	20	15
Blue	13	12	11	22	20
White / Grey	4	2	2	5	2
Brown	5	7	7	3	6
Black	3	2	3	4	3

Table 1-(2): Hue ration selected in the second experiment.

	/a/	/e/	/i/	/o/	/u/
Red / Pink	24	0	0	2	0
Orange	24	24	11	7	18
Yellow	4	4	18	2	4
Green	11	33	33	22	29
Purple	9	16	9	22	22
Blue	16	13	11	40	22
White / Grey	0	0	0	0	0
Brown	11	9	13	2	2
Black	0	0	4	2	2

the categorical mapping to color can be well explained by their constituting three attributes of color. That is, /e/ with green or orange is allocated in the middle of /a/ with red or orange and /i/ with yellow or green both for speech Formant plane and color mixed theory.

In contrast to the above remarkable correlations, we did not find any correlations between F0 characteristics and Saturation. Little correlation scores were found between average F0 heights and average Saturation scores, 0.02, 0.03, 0.31, -0.13, -0.13 for /a/, /i/, /u/, /e/ and /o/ respectively. The average Saturation scores for each F0 dynamic pattern were almost equivalent, 67% for rising and falling and 66% for flat and rising followed by falling.

4. POSSIBILITIES OF THE DESCRIPTION OF SPEECH VARIATIONS USING COLOR

From the correlation analyses shown in the previous section, we can think of the possibilities to describe speech variations using other information media. From the current analysis results, the following description of speech characteristics can be considered based on sentiment similarities.

4.1. Fundamental frequency characteristics

In a series of our studies, we have been proposing the usefulness of perceptual impressions and their magnitude for the control of communicative prosody [2]-[12]. In these studies, F0 heights can be nicely controlled by positive-negative impressions embedded in constituent output speech. As observed in this color selection experiments, high correlations between F0 and Value strongly support this positive-negative sentiment correlation. In color psychology, Value gives a light-heavy impression which coincident with one of the positive-negative impressions in speech.

On the other hand, F0 has no correlation with Saturation in both experiments. This characteristics reflects the fact that Saturation gives impressions of strong-weak, motion-rest, or loud-subdued and do not relate to positive-negative impressions manifested in F0 characteristics. As for dynamics, correlation between F0 patterns and color attributes has not been clearly observed. We are thinking that this lower correlations of F0 dynamics result from the constraint of color selection. Dynamically changing F0 characteristics may also require dynamic color changing to share same sentiment information. We need another selection experiments.

4.2. Vowel color characteristics

The vowel specific distributional differences of Hues in two experiments shows high correlation score 0.91 in average. This correlation and the correlation characteristics discussed in Section 3.4 (3) suggest a possibility of direct mapping from speech spectral features such as formants or cepstrum parameters to Hue values. We need further experiments with large number of subjects to confirm whether this mapping can be established between vowel categories to color categories or between spectral parameters and Hue values.

We have tried to train a neural net to enable this mapping. However, it has not yet been well trained. We found that the above agreement is quite clear in global characteristics, but not in individual responses even between the responses of same listeners. This correlation score drops to 0.41 by pooling the results to each listener. In addition, as seen in the two experiments, selected Hue distribution differs between speakers. Speaker of the second set has deeper voices than the other male of the first set. For further understanding, we are planning to replicate this experiments using speech samples uttered by multiple speakers including females.

4.3. Loudness characteristics

The high correlation average score 0.82 between SPL and Saturation suggests the correlation between loudness characteristics and Saturation. It has been confirmed that Saturation has psychological effect in impact dimension like strong-weak, motion-rest, or loud-subdued, which coincides the perceptual impressions of forceful or deep voices [19]. Again, this correlation score drops to 0.57 by pooling the results to each listener. We may need multiple evaluates for each sample to understand subjects' selection variances.

5. CONCLUSIONS

To find out a new possibility to describe the information embedded in speech, we have carried out sentiment correlation analysis between vowels and colors selected by perceptual impression. High correlations were found between mean F0 and Value, sound pressure level and Saturation, and Formants and Hue. These correlations nicely coincide with observations on vowel colors presented in previous studies and provide new possibilities for the mapping across different modalities. Though more precise investigations are needed, we can expect to express the impression of speech differences directly in visible way. We are expecting to use color and other visual information to describe not only the differences between read speech and communicative one but also miss-pronunciations by L2 speakers and further detailed prosodic differences found in sports coaching where their easy-to-understand information expression plays crucial roles [20].

REFERENCES

[1] Sapir, E., "Language: An introduction to the study of speech" Harcourt, Brace and Company 1921.

[2] Sagisaka, Y., Yamashita, T., and Kokenawa, Y., “Generation and perception of F0 markedness for communicative speech synthesis”, Speech Communication, Vol.46, No.3-4, pp.376-384, 2005.

[3] Kokenawa, Y., Tsuzaki, M., Kato, H., and Sagisaka, Y., "F0 control characterization by perceptual impressions on speaking attitude using Multiple Dimensional Scaling analysis" Proc. IEEE ICASSP 2005, pp.273-275 2005.

[4] Kokenawa, Y., Tsuzaki, M., Kato, K., and Sagisaka, Y. "Communicative speech synthesis using constituent word attributes", Proc. INTERSPEECH, 517-520, 2005.

[5] Greenberg, Y., Tsuzaki, M., Kato, K., and Sagisaka, Y., “A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech”, Proc. Speech Prosody 2006, pp.37-40, 2006

[6] Li, K., Greenberg, Y., Campbell, N., and Sagisaka, Y.,” On the analysis of F0 control characteristics of nonverbal utterances and its application to communicative prosody generation” in NATO Security through Science Series E: Human and Societal Dynamics Vol.8 The Fundamentals of Verbal and Non-verbal Communication and the Biometric Issue pp.179-183 IOS Press 2007

[7] Li, K., Greenberg, Y., and Sagisaka, Y., “Inter-language prosodic style modification experiment using word impression vector for communicative speech generation”, Proc. Interspeech 2007 pp.1294 - 1297, 2007

[8] Zhu, M., Li, K., Greenberg, Y., and Sagisaka, Y., "Automatic extraction of paralinguistic information from communicative speech" Proc. the 7th Symposium on Natural Language Processing 2007 pp.207-212, Dec.2007

[9] Greenberg, Y., Shibuya, N., Tsuzaki, M., Kato, H., and Sagisaka, Y., “Analysis on paralinguistic prosody control in perceptual impression space using multiple dimensional scaling”, Speech Communication Vol.51 No.7 pp. 585-593, 2009.

[10] Greenberg, Y., Tsuzaki, M., Kato, H., and Sagisaka, Y., “Communicative prosody generation using language common features provided by input lexicons", Proc. SNLP2009, pp.101-104, 2009.

[11] Greenberg, Y., Tsuzaki, M., Kato, H., and Sagisaka, Y., "*Analysis of impression-prosody mapping in communicative speech consisting of multiple lexicons with different impressions*", Proc. O-COCOSDA, 2010

[12] Lu, Shao., Greenberg, Y., Sagisaka, Y., "Global F0 control parameter prediction based on impressions for communicative prosody generation" Proc. Oriental COCOSDA 2013, pp. 1 – 4, 2013

[13] Jakobson, R., "Selected Writings: I phonological Studies", 1962.

[14] Wrembel, M., and Rataj, K., "Sounds like a rainbow – sound-color mappings in vowel perception-", Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, pp.237-240, 2008.

[15] Ward, J., Huckstep, B., and Tsakanikos, E., "Sound-colour synaesthesia: to what extent does it use cross-modal mechanisms common to us all? ", Cortex, 42, pp.264-280, 2006

[16] Nagata, N., Iwai, D., Tsuda, M., Wake, S., Inokuchi, S., S.(2005). "Non-verbal Mapping between Sound and Color - Mapping Derived from Colored Hearing Synesthetes and Its Applications - ". F. Kishino et al. (Eds.): ICEC2005

[17] Takahashi, R., Fujisawa, T. X., Nagata, N., Sugio, T., and Inokuchi, S., "Brain Activity in Colored-hearing by Listening to Music: An fMRI Study" Proc. Second International Congress Synaesthesia, Science & art. 2007

[18] Watanabe, K., Greenberg, Y., and Sagisaka, Y., " Sentiment analysis of color attributes derived from vowel sound impression for multimodal expression" Proc.APSIPA 2014.

[19] Kido, H., and Kasuta, H., “Vocal quality expressions of speech utterance and their acoustic correlates”, IEICE2002, pp.7-12

[20] Fujino, Y., Kikkawa, M., Yamada, T. and Sagisaka, Y., "Japanese Sports Onomatopoeias”, pp.163-166 in "Computer processing of Asian languages" S. Itahashi and C. Tseng (Eds.) Consideration Books 2010