Cross-modal Correlation Analysis between Vowel Sounds and Color*

Win Thuzar Kyaw
Dept. of Pure and Applied Math.
Waseda University
Tokyo, Japan
winthuzarkyaw19@gmail.com

Atsuya Suzuki
Dept. of Pure and Applied Math.
Waseda University
Tokyo, Japan
atsuya149@gmail.com

Yoshinori Sagisaka
Dept. of Pure and Applied Math.
Waseda University
Tokyo, Japan
ysagisaka@gmail.com

***Abstract*—Vowel-color association characteristics have been studied in the field of phonetics and perception. Though it has been reported that selected color categories after listening vowel categories have similar trends in multiple languages, their sentiment correlations have not yet been thoroughly studied from the viewpoint of speech features. We tried to find sentiment association characteristics between color parameters and speech features directly to have better understanding of cross-modal correlations and to find underlying principles for multimodal applications. Vowel samples uttered by 4 male and 3 female speakers were employed to associate colors after listening them by 34 subjects. Statistical analyses showed the advantage of employing RGB color parameters and speech formants directly to conventional color category to vowel category mapping. The selected color distributions in the F1-F2 plane clearly show that the acoustic speech resonance (i.e. F1 and F2) -RGB correlations can more consistently explain their sentiment correlations. Moreover, by incorporating our sentiment association experiment results using formant-synthesized speech, their correlations can be attributed to F1 and F2 rather than vowel categories. We believe that this finding in cross-modal correlations will serve for not only scientific understanding but also further studies and applications using cross-modal information mapping.**

***Index Terms*—speech-color correlation, cross-modal information expression, cross-modal perception, sentiment information**

I. Introduction

Both in speech information processing and in phonetic science, linguistic information has been studied as a main research target for a long time. Though speech information includes non-linguistic information and so called paralinguistic information which play quite important roles in communication, they have not yet been sufficiently studied. In particular, though overall prosodic variations such as emotional variations have been started to be treated, individual utterance differences in real communications have neither yet been sufficiently described nor analyzed quantitatively to clearly specify what information is embedded by the speakers and received by the listeners. Aiming at generating natural communicative prosody, we have been studying individual utterance differences for more than a decade and succeeded their generation by increasing their applicability of output sentences [4][5][8][9][10]. In their prosody computation, constituent word attributes have been effectively employed to show their sentiment information expressing positive-negative, confident-doubtful and allowable-unacceptable through MDS (Multi-Dimensional Scaling) analyses [9][10].

Aiming at speech descriptions using image related information by replacing word expressions of language medium, we have analyzed the cross-modal sentiment correlations between speech and image such as colors and textures [11][12][14][15]. We have studied direct mapping between voice source parameters and image texture parameters reflecting human sentiment perception for the purpose of visualization of voice source differences by image textures [14]. In phonetic science, the associations among pitch differences, vowel qualities, pitch difference coded with vowel qualities with colors have been analyzed to understand cross-modal relationship between speech and color [7][13].

To understand color association after listening speech more scientifically and to describe speech variations directly using color parameters, we have carried out sentiment correlation analyses between vowels and colors selected by perceptual impression. In these studies, high correlations have been observed between F0 and Value, Sound Pressure Level (SPL) and Saturation and Spectrum and Hue in perceptual experiments by ordinary subjects who do not have any particular perceptual peculiarities such as synesthesia [11][12]. In this study, we tried to explore vowel-color association characteristics based on speech features, formants to understand the conventional color category to vowel category mapping more deeply.

In Section 2, we introduce the studies on the relationship between speech and color. In Section 3, we describe how we designed our experiment on color selection based on perceptual impressions of Japanese vowels uttered by multiple male and female speakers. Then, we present our experimental results in Section 4. These results try to relate conventional findings on categorical associations of colors after vowel listening to the associated color mapping in F1 and F2 plane showing vowel differences. In Section 5, we wrap up our findings and summarize the current understandings on the correlations between speech and color. In Section 6, we conclude our paper.

II. Studies on speech-to-color association

Relating to speech and color mapping, Jakobson [1] first proposed that synesthetes, people experiencing concurrent perceptions such as colors or textures when they hear the

sound of voices, tend to select red for /a/, darker colors for /o/ and /u/, brighter colors for /e/ and /i/. Ward et al [7] has shown that there is a general tendency to associate high pitch sounds with light colors and low pitch with dark colors not only by sound-color synesthetes but also by non-synesthetes. Thus, cross-modal mappings between sound and color might be common to all population. Moreover, Mok et al [13] presented non-random associations between Cantonese vowel sounds and colors by general population including both native and non-native listeners. In addition, they proposed the correlations of lexical tones with brightness, high tone with lighter color and low tone with darker color. In phonetic science, it could be found that pitch differences or vowel qualities or pitch differences coded with vowel qualities were associated with colors in cross-modal association.

To have better understanding of speech-color correlations for the description of speech variations by color, sentiment correlation analysis was carried out between color parameters using HSV(Hue, Saturation, Value) and speech features [11][12]. High correlations were found between F0 and Value(0.90), Sound Pressure Level and Saturation (0.85) and Spectrum and Hue. For the correlations between Spectrum and Hue, /a/ tends to associate with red or orange, /i/ with yellow or green, /u/ with blue, green or purple, /e/ with green or orange and /o/ with blue, green or purple. We can observe similar color categories to vowel categories mapping in multiple languages from the above studies. However, vowels to colors association characteristics have not yet been thoroughly studied from the viewpoint of speech features. In this study, we try to analyze the sentiment correlations between vowels and colors based on acoustic speech resonance.

III. Experimental setup

We conducted the experiment on color selection by listening Japanese vowel sounds with 34 Japanese subjects (22 males and 12 females) aged ranging from 18 to 24 years old. Each subject was asked to listen to individual speech stimulus presented in random order by sitting in front of a computer in a quiet room. After listening to the speech, the subject was asked to choose the most suitable color among 153 colors at a time based on his/her perceptual impression.

A. Speech stimuli

To measure the acoustic speech resonance characteristics, we collected speech samples of many speakers. Five Japanese single vowels (/a/, /i/, /u/, /e/, and /o/) were uttered by seven native speakers including four males and three females. Thus, total thirty five speech stimuli (5 Japanese vowels $\times$ (4 male speakers + 3 female speakers)) were recored with sampling frequency of 44kHz for our experiment.

B. Color stimuli

For the colors to be selected by listening Japanese vowel sounds, we employed Practical Color Co-ordinate System (PCCS) as previous studies on speech-color correlations [6][11][12]. It consists of 153 color tips (12 hues of color

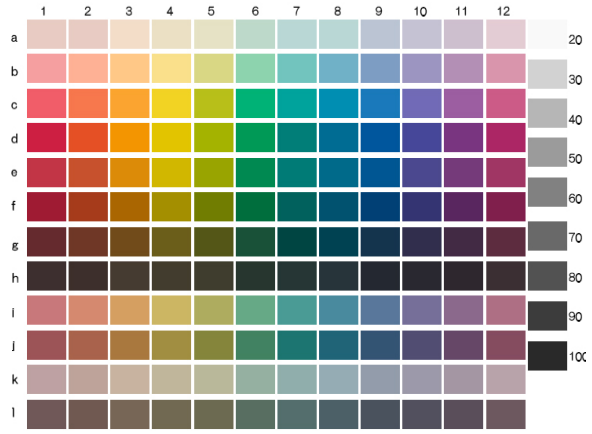

Fig. 1: Color palette used in the color selection by listening vowel sounds experiment

$\times$ 12 kinds of tone + 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”. The color palette used in our perceptual experiment is as shown in Figure 1.

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.

PCCS classifies and displays colors by three attributes of color, Hue, Saturation, and value. 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. Hues are described by a number that specifies the position of the corresponding pure color on the color wheel, as a fraction between 0 and 1. Value 0 refers to red; 1/6 is yellow; 1/3 is green; and so forth around the color wheel. Saturation relates to the degree of unmixed, clear color has high saturation and somber one has low. Value expresses the lightness.

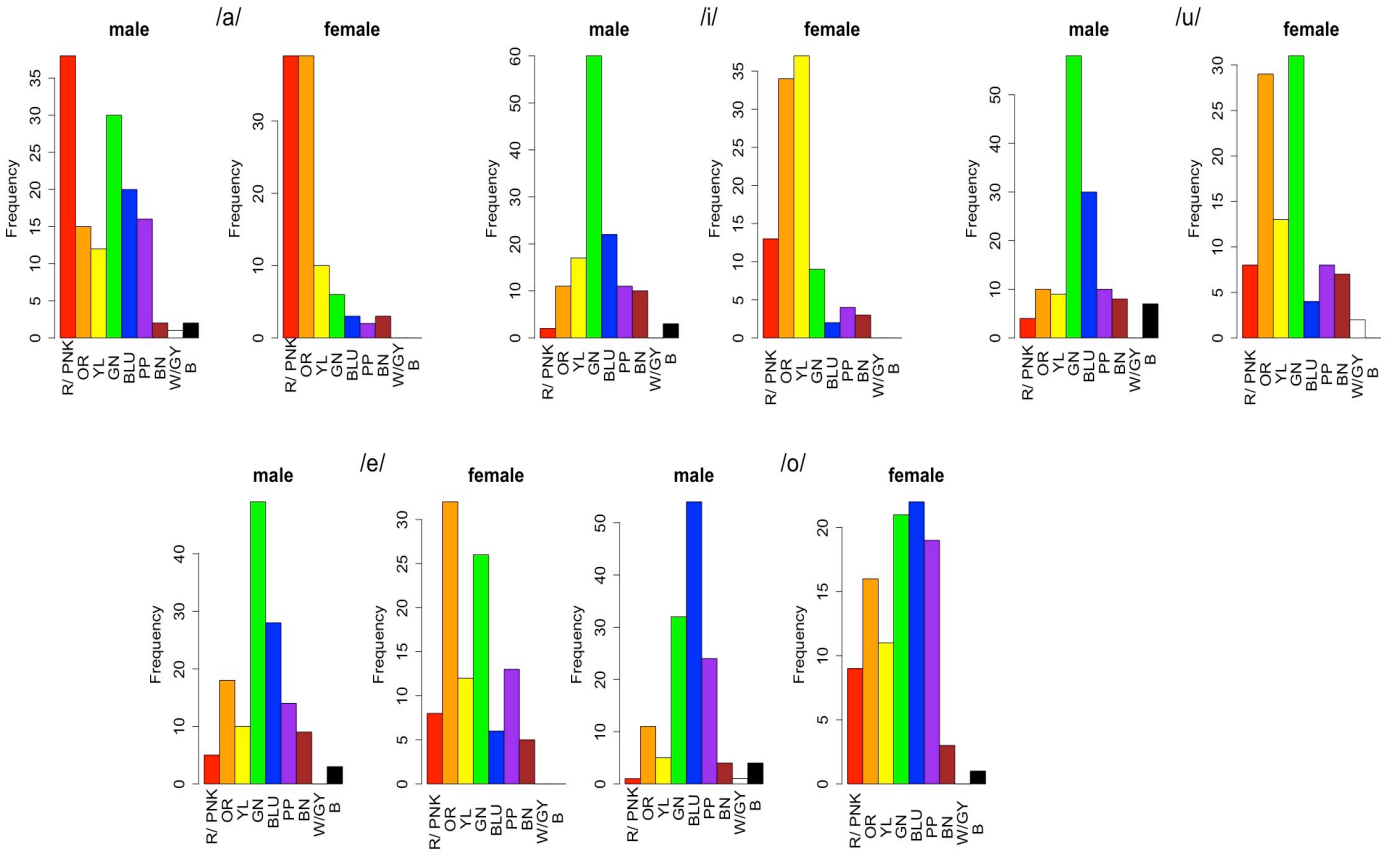

Fig. 2: Color selection differences by listening to male and female Japanese vowel speech.
(R=Red, PNK=Pink, OR=Orange, YL=Yellow, GN=Green, BLU=Blue, PP=Purple, BN=Brown, W=White, GY=Grey and B=Black)

IV. Results

A. Mapping between vowel and color categories

We investigated a categorical mapping from vowels to colors to see whether their associations are similar or different for male and female speech.

We first analyzed vowels and colors association for all speech samples uttered by both male and female speakers by grouping the 153 color tips into 11 basic color terms (black, blue, brown, green, grey, orange, pink, purple, red, yellow and white) [2] for categorical color terms. We found that /a/ is tended to be associated with red or orange (32.35% and 22.69% respectively), /i/ with green, yellow or orange (29%, 22.69% and18.91% respectively) , /u/ with green (37.39%), /e/ with green or orange (31.51% and 21% respectively), and /o/ with blue, green, or purple (31.93%, 22.27% and 18.07% respectively). These percentages were calculated based on the total number of selections 238 (7 speakers (4 males + 3 females) $\times$ 34 listeners) for each vowel sound. We observed similar vowel to color mapping as Watanabe's work [11][12] in which /a/ is tended to be associated with reddish colors like red or orange, /i/ with yellow or green, /u/ with blue, green or purple, /e/ with green or orange and /o/ with green, blue or purple.

However, when we separately examined color selections based on male or female speech samples, we found differences in vowel to color mapping as shown in Figure 2. In this figure, the horizontal axis shows 11 color terms and the vertical axis describes the number of selections for each color by listening to each vowel sound uttered by male or female speaker. For male speakers, /a/ is tended to be associated with red or green (27.94% and 22.06% respectively), /i/ with green (44.12%) , /u/ with green or blue (42.65% and 22.06% respectively), /e/ with green or blue (36.03% and 20.59% respectively), and /o/ with blue, green, or purple (39.71%, 23.53% and 17.65% respectively). For female speakers, /a/ is tended to be associated with red or orange (38.24% and 38.24% respectively), /i/ with yellow or orange (36.27% and 33.33% respectively) , /u/ with green or orange (30.39% and 28.43% respectively), /e/ with orange or green (31.37% and 25.49% respectively), and /o/ with blue, green, or purple (21.57%, 20.59% and 18.63% respectively). The percentages for male speech were calculated based on the total number of selections 136 (4 male speakers $\times$ 34 listeners). For calculating

the percentages for female speech, it was based on the total number of selections 102 (3 female speakers × 34 listeners).

B. Mapping between vocal tract characteristics and color

When we investigated the color selections by listening to vowel sounds produced by many speakers, we could see that the subjects tended to choose different colors for the same vowel categories depending on male or female speech. From this fact, there are possibilities that color selections are dependent on acoustic characteristics rather than categories. To clearly understand and confirm color selection characteristics, we compared selected vowel categories and speech resonance characteristics by plotting selected color distributions in the F1-F2 plane. To extract the first two formant frequencies F1 and F2, we carried out LPC analysis and employed Burg's method [3] in Praat. Formant extraction was carried out from LPC spectrum and manually adjusted.

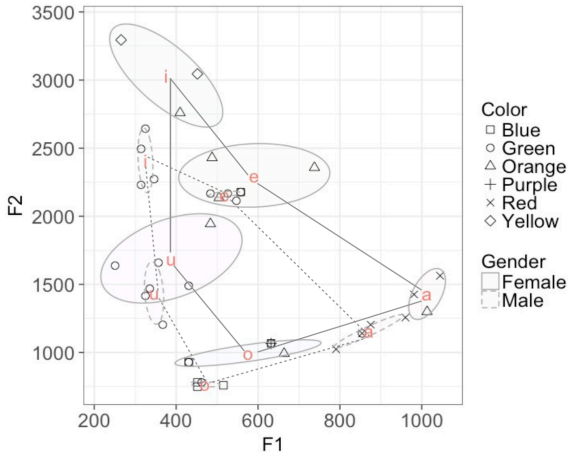

Fig. 3: Color distributions in F1-F2 plane using natural speech (The points represent the most selected color for each speech sample)

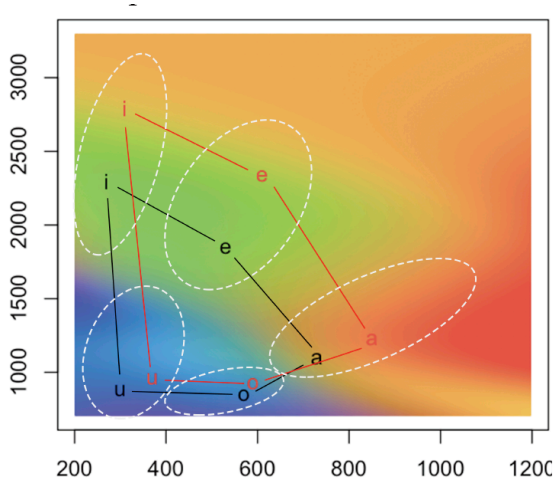

Fig. 4: Color map in F1-F2 plane using synthesized speech

In Figure 3, we show the maximum number of color selections based on first two formants (F1 and F2) and the average formant frequencies for 5 Japanese vowels of male and female speech. In the figure, we can see that yellow and orange colors are the most selected colors for female /i/ sound and green color is the most selected color for male /i/ sound in spite of being the same vowel category /i/. The differences in color selections can also be found in other vowel categories. From this fact, it can be said that different associative colors are more likely to be selected depending on the speech resonance characteristics (F1 and F2) rather than vowel qualities.

Moreover, we could have confirmed that vowel-to-color correlations can be attributed to F1 and F2 rather than vowel categories by integrating our sentiment association experiment results using formant-synthesized speech [15]. In Figure 4, we depicts the color map produced by neural network using synthesized F1, F2 and its associated RGB values as learning data. When we compare the result obtained by using natural speech and the result produced by neural network employing synthesized speech, we could observe similar color distributions in F1-F2 plane for both natural speech and synthesized speech. This finding supports that speech-color sentiment associations are better interpreted by using acoustic features F1 and F2 than vowel categories.

To understand the sentiment correlations between Spectrum and Hue, we carried out multiple regression analysis by employing RGB color parameters and speech formants. Since the Hue value represents the angle on the color cycle, it sometimes causes calculation problems. In order to simplify the calculation, we converted HSV values to RGB values. The following regressions were obtained for Red, Green and Blue.

(1) R = 145.0+ 0.13F1 + 0.04 F2 (R^2 = 0.118)

(2) G = 132.0 + 0.00006 F1 + 0.01 F2 (R^2 = 0.061)

(3) B = 87.9 - 0.003 F1 - 0.007F2 (R^2 = 0.005)

F1 ($p<0.000$) and F2 ($p<0.000$) all significantly contribute to the association with Red. F2 significantly associates with Green ($p<0.000$). F2 significantly associates with Blue ($p<0.01$). However, the relative small R square values suggest that the equations can predict only small percentages of the associations of F1 and F2 with Red, Green and Blue. There may be nonlinear relationship between F1, F2 and RGB and we have to explore their relationships by employing machine learning algorithms.

V. Discussion

In this study, we investigated color association characteristics relating to vowel qualities or spectral features. By investigating color selections of listeners by incorporating male and female speech altogether, we could find similar categorical mapping between vowels and colors as Watanabe's work [11][12]. /a/ is tended to be associated with red or orange, /i/ with green, yellow or orange, /u/ with green, /e/ with green or orange and /o/ with blue, green or purple. However, color selection characteristics for male and female speech show the differences. For male speech, /a/ is tended to be associated with red or green, /i/ with green, /u/ with green or blue, /e/ with green or blue and /o/ with blue, green or purple. While, /a/ is tended to be associated with red or orange, /i/ with yellow

or orange, /u/ with green or orange, /e/ with orange or green and /o/ with blue, green or purple for female speech. From these different color selection characteristics between male and female speech, we can speculate listeners tend to select colors based on acoustic characteristics rather than vowel sounds.

By plotting color selections in the F1-F2 plane using the extracted first two formants of vowels, we could say that colors are more likely to be associated with speech resonance characteristics than vowel categories. In addition, we could have confirmed sentiment correlations between speech parameters and color features by comparing with our sentiment association experiment results using formant-synthesized speech [15]. To find the underlying mapping scheme between speech and color parameters, we converted Hue angle values to RGB values to simplify calculations and carried our multiple linear regression. From the analysis results, we observed that F1 significantly contributed to the association with Red. And, F2 significantly contributed to the associations with Red, Green and Blue.

VI. Conclusions

For the purpose of human friendly multimodal information expressions and visualization of speech using image, in this paper, we analyzed sentiment correlations between speech and color.

By employing direct mapping between speech formant parameters and image parameters, we could have found more consistent and systematic correlations than conventional mappings between vowel categories and color categories. Though conventional analyses only described the similar color selection characteristics after listening vowels, multiple color selections and the selection differences between male speech and female speech have not yet been well studied. The color selection characteristics expressed in F1-F2 plane consistently showed the varieties of selected color distributions for both male and female speech samples. Moreover, it has been shown that these color selection characteristics have same color characteristics obtained by speech samples using formant synthesizer. These facts support that speech-color sentiment associations are better interpreted by using acoustic features F1 and F2 than vowel categories. The multiple linear regression analysis expressed reasonable estimation of RGB color parameters from speech formants.

As an extension of this work, further correlations between speech source parameters and texture image parameters have been investigated. We would like to continue more detailed analyses to understand the observed complex color selection characteristics scientifically and find a color and texture generation scheme directly from speech inputs for human sentiment friendly cross-modal mapping applications.

References

[1] R. Jakobson, "Selected writings: I phonological studies," 1962.

[2] B. Berlin and P. Kay, "Basic color terms: their universality and evolution," in Berkeley and Los Angeles: University of California Press, 1969.

[3] N. Anderson, "On the calculation of filter coefficients for maximum entropy spectral analysis," in Childers: Modern Spectrum Analysis, IEEE Press, 1978, pp. 252–255.

[4] Y. Sagisaka, T. Tamashita, and Y. Kokenawa, "Generation and perception of f0 markedness for communicative speech synthesis," Speech Communication, vol.46, no.3, 2005, pp.376–384.

[5] Y. Greenberg, M. Tsuzaki, H. Kato, and Y. Sagisaka, "Communicative speech synthesis using constituent word attributes," in European Conference on Speech Communication and Technology, 2005, pp. 517–520.

[6] N. Nagata, D. Iwai, M. Tsuda, S. Wake, and S. Inokuchi, "Non-verbal mapping between sound and color - mapping derived from colored hearing synesthetes and its applications," in F. Kishino et al.(Eds.): ICEC2005, 2005.

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

[8] K.Li, Y. Greenberg, and Y. Sagisaka, "Inter-language prosodic style modification experiment using word impression vector for communicative speech generation," in InterSpeech 2007 – 8th Annual Conference of the International Speech Communication Association, 2007, pp. 1294–1297.

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

[10] L. Shao, Y. Greenberg, and Y. Sagisaka, "Global f0 control parameter prediction based on impressions for communicative prosody generation," in (O-COCOSDA/ CASLRE), 2013 – Oriental COCOSDA held jointly with 2013 conference on Asian Spoken Language Research and Evaluation, International Conference. IEEE, 2013, pp. 1–4.

[11] K. Watanabe, Y. Greenberg, and Y. Sagisaka, "Sentiment analysis of color attributes derived from vowel sound impression for multimodal expression," in Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA). IEEE, 2014, pp. 1–5.

[12] K. Watanabe, Y. Greenberg, and Y. Sagisaka, "Cross-modal description of sentiment information embedded in speech," in Proc. ICPhS 2015 A-117, (CDROM).

[13] P. Mok, Y. Yin, L. Chen, and H. Cheung, "Cross-modal association between colour, vowel and lexical tone in nonsynesthetic populations: Cantonese, Mandarin and English," in Proc. ICPhS 2015, (CDROM).

[14] W. T. Kyaw and Y. Sagisaka, "Cross-modal analysis between phonation differences and texture images based on sentiment correlations,"in Proc. InterSpeech 2017, 2017, pp. 679–683.

[15] A.Suzuki, W. T. Kyaw and Y. Sagisaka, "Sentiment analysis on associated colors by listening synthesized speech," in Proc. Fechner Day 2017, 2017, pp. 144–149.