Cross-modal Sentiment Information Expression of Voice Source Characteristics using Image Texture Features

Win Thuzar Kyaw, Yoshinori Sagisaka
Department of Pure and Applied Mathematics
Faculty of Science and Engineering
Waseda University
winthuzarkyaw@akane.waseda.jp, sagisaka@waseda.jp

Abstract

Following the successful findings of high correlations between speech and color such as F0 and Value, Loudness and Saturation and Spectrum and Hue, we analyzed the correlations between voice source characteristics and the image parameters showing textural differences in this paper for better scientific understanding of their correlations and effective use in visualization of speech information. Through sentiment association experiments, we could have observed high positive correlations between H1-H2* (amplitude difference between first and second harmonics corrected for vocal tract effects), H1-A1 (amplitude difference between first harmonic and first formant) and Contrast, high negative correlations between H1*-H2*, H1-A1, H1-A2, H1-A3, Harmonic-to-Noise Ratio (HNR) in 0 to 3500Hz frequency band and Variance, Prominence and negative correlations between H1*-A3*, HNR in 0 to 500 Hz and Prominence. These results show the possibility of direct visualization of speech characteristics which cannot be effectively carried out by conventional mapping using discrete language expressions.*

1. Introduction

In the field of speech information processing, most of research efforts have been devoted to send, generate and extract linguistic information mainly such as coding, synthesis and recognition. On the other hand, speech contains not only linguistic information but also other information quite importantly employed in daily speech communications. Though only a part of it has begun to be studied such as so-called para-linguistic information in speech synthesis, the most of it has neither yet been known nor described.

To treat undescribed information embedded in speech, we have been proposing to used its sentiment information as its descriptor and show the possibilities of their use in speech synthesis (Shao et al., 2013; Greenberg et al., 2009; Li et al., 2007; Greenberg et al., 2005; Sagisaka et al., 2005). Quite recently, we have started to generalize sentiment correlation analysis between speech and other media by replacing language medium to image medium to understand its scientific background and to apply to speech information processing (Watanabe et al., 2014) and (Watanabe et al., 2015).

To analyze and control communicative speech, we have been employing sentiment information obtained by their listening. For their description, we have been successfully applying language information expressing their perceptual impressions. Like other NLP applications, a vector space model can be applied by enumerating discrete language impressions of continuous psychological space with multiple dimensions. Applying Multi-Dimensional Scaling (MDS), we could have found communicative prosody information can be reduced to three dimensional expressions (doubtful-confident, unacceptable-allowable, positive-negative) nicely corresponding to its prosody characteristics (Shao et al., 2013). Between language and speech, we could have found that not only the above mentioned communicative prosody characteristics expressed in three dimensional psychological impressions using language, but also sentiment correlations between consonant categories and their sentiment impressions expressed by multi-dimensional space expressed by language expressions (Isonaka et al., 2015). Though discrete language expressions can be effectively employed in the above cases, they have potential problems to describe other modal information. As we use creaky voice or breathy voice for speech, texture for image, we can specify only its very vague features but we cannot exactly identify their physical properties expressing acoustic voice source characteristics or various kinds of image texture parameters.

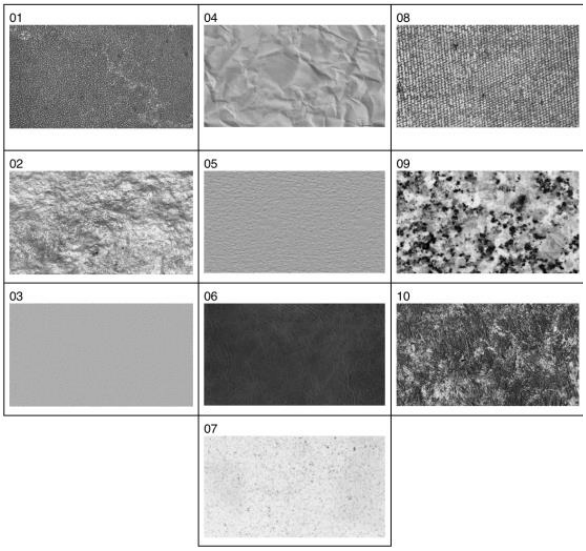

Figure 1. Texture images used in the experiment

To be free from these constraints naturally imposed by language use, we have been looking for the possibilities of mutual information transformations among language, speech and image. Sentiment correlations between cross media have been extensively studied in many academic fields. The correlations between speech and image have been studied originally in Phonetics area. The correlations between vowels and colors have been analyzed for color hearing synesthesia which is a special ability to see color while listening to sounds (Jacobsen, 1962), with Polish subjects who are English language learners (Wrembel and Rataj, 2008) and in a tone language (Cantonese Chinese) with ordinary people (Mok et al., 2015).

In speech information processing, we have shown its quantitative correlations thoroughly using speech parameters (F0, sound pressure level and formants) and color parameters (Hue, Saturation and Value) (Watanabe et al., 2014; Watanabe et al., 2015). High correlations between F0 and Value (brightness) 0.85, formants and Hue 0.91, loudness and saturation 0.85 have been obtained using these continuous parameters instead of categorical values.

For further sentiment information embedded in speech, we can think of the information given by its source characteristics. In this paper, we tried to get more exact sentiment correlation between voice source parameters and image textural parameters. To look for quantitative high correlations between voice source and texture, we have newly tested voice source related parameters and eight textural features widely employed in image processing. In the following sections, we describe the perceptual experiment for speech and texture sentiment correlations and parameters used in the correlation analysis (voice source parameters and image textural parameters) in Section 2 and 3 respectively. Then, we will show the experimental results and discuss on our findings in Section 4. Finally, we will conclude in Section 5.

2. Experiment on Textural Parameters Derived from Perceptual Impressions of Japanese Vowels in Three Different Phonation Types

Following the success of direct mapping between speech and color (Watanabe et al., 2014; Watanabe et al., 2015), we conducted cross-modal sentiment analysis between speech and texture.

2.1. Speech Stimuli

For the speech stimuli used in this experiment, an adult male who can deliver three different phonation types (modal, creaky and breathy) recorded thirty speech samples (2 recordings × 3 different phonation × 5 Japanese vowels) in a quiet room. The spectral acoustic measures were calculated from five Japanese vowels with three different phonation types to search voice characteristics differences by using VoiceSauce application (Shue et al., 2011) with default settings.

2.2. Texture Stimuli

For image textures, we selected ten image textures from the web showing uniform abstract structure in the entire image and no change in the impression even any part is cut out. And then, they were converted into 256 (8bit) gray scale images by the image editing software GIMP (ver.2.8.6) in order to eliminate the influence of by color effects for visual impression and were cut into 1920 × 1080 pixel. For feature analysis of image textures, the left, center and right areas of each image is cut out 1024 × 1024 pixels in size and Image processing software ImageJ ver. 1.49 was used to calculate the eight GLCM textural feature values. Ten texture images used in the experiment are described in Figure 1.

2.3. Speech-Texture Association Experiment

In our experiment, we investigate how we feel textures by listening to the Japanese vowels in three different phonation types. Totally 40

Myanmar students aged ranging from 20 to 40 years participated. Each participant was asked to listen to randomized speech stimuli for every person in front of a computer by external speaker or a headphone in a quiet room and then requested to select one texture image and up to three textures that were perceived well match to the voice among ten textures. In this experiment, they had to listen to the speech one at a time and they were allowed to replay the speech as often as they want.

3. Feature Parameters for Correlation Analysis

In our study, we consider five Japanese vowels (/a/, /i/, /u/, /e/, /o/) in three different phonation types (modal, creaky and breathy) to know the voice source characteristics differences among these phonation types. We calculate spectral acoustic measures related to voice source. For human visual impressions, we examine gray-scale texture images not to be influenced by color effects. We employ textural features relating to uniformity, homogeneity, heterogeneity, smoothness, randomness and asymmetry driven from gray-level textures.

3.1. Acoustic Parameters for Voice Source Differences

Three common different phonation types (modal, creaky and breathy) can be determined by Open Quotient (OQ) representing the amount of time the vocal folds are open during the pitch period compared to the amount of time they are closed (Johnson, 1997). Compared to three different phonation types, the most constricted glottis is creaky, the greatest open glottis is breathy and modal voice is produced with regular vibration of vocal folds. The acoustic measure H1-H2 meaning the intensity differences taken between the first and second harmonics correlating with the open quotient is a good measure of phonation (Holmberg et al., 1995). Since OQ is related to breathiness, H1*-H2* which is the difference between the first two spectral harmonic magnitudes inversely filtered to get rid of vocal tract resonance characteristics (represented by asterisks) can be thought of as an acoustic measure of breathiness. Generally, the value of H1*-H2* is usually close to zero for modal voice, strongly positive for breathy voice representing strong H1 and usually negative for creaky voice meaning that H1 is weaker than H2. In addition to H1*-H2*, we explore more voice source or voice quality related acoustic parameters (F0, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*, H2- H4, H1-A1, H1-A2, H1-A3, Cepstral Peak Prominence (CPP), Harmonics-to-Noise Ratio (HNR)) in this analysis to measure voice source characteristics differences.

We investigate the frequency of the lowest harmonic (F0) (Kawahara et al., 1999). Acoustic changes including an increase in aspiration noise level (AH) and spectral slope are cues for the perception of breathy voice. Spectral slope means the amount of decrease in intensity (in dBs) as frequency increases in the spectrum. It can be expressed by the difference of amplitude of the first harmonic to that of a higher frequency harmonic or the amplitude of the formants such as the amplitude values of F1 (first formant), F2 (second formant) and F3 (third formant). It is used as a simple approximation of the harshness vs. softness of the voice quality. Hi refers to the i^{th} source spectral harmonic magnitude, and Ai refers to the magnitude of the source spectrum at the i^{th} formant. For example, H1-A3 is the amplitude of the first harmonic (H1) relative to that of the third-formant spectral peak (A3). Asterisks (*) mean that the harmonic amplitudes were corrected for the vocal tract effects. H1-A3 has been characterized as spectral tilt (Hanson, 1997) and it is concerned with the speed of closing of vocal folds (Menezes et al., 2006). Spectral tilts are most steeply positive for creaky vowels which is stronger energy in high frequency region and most steeply negative for breathy vowels meaning weaker energy in the high frequency region (Gordon and Ladgefoged, 2011).

Another cue for the perception of breathy voice is an increase in aspiration noise level. We calculate harmonic-to-noise-ratio (HNR) which is the ratio between the periodic (harmonic part) and aperiodic (noise) components to know the degree of acoustic periodicity (Krom, 1993). HNR is related to the perception of vocal roughness and hoarseness. Using a variable window length equal to 5 pitch periods by default, the HNR measurements are found by liftering the pitch component of the cepstrum and comparing the energy of the harmonics with the noise floor. HNR05 means the measure of HNR between 0-500Hz, HNR15 (HNR between 0-1500Hz), HNR25 (HNR between 0-2500Hz) and HNR35 (HNR between 0-3500Hz). Cepstral Peak Prominence (CPP), the harmonic with the greatest amplitude in the cepstrum domain (Hillenbrand et al., 1994) is used to evaluate the changes in aspiration noise of

five Japanese vowels in three different phonation types.

3.2. Image Texture Parameters

Gray Level Co-occurrence Matrix (GLCM) is one of the most well-known texture descriptors. We consider Haralick textural features computed from Gray Level Co-occurrence Matrix (GLCM) method (Haralick et al., 1973) for image texture parameters of this study. GLCM is a tabulation of the occurrences of different combinations of pixel brightness values (gray levels) in an image where the number of rows and columns is equal to the number of quantized gray levels. Thus, the matrix element P_{ij} is the set of second order statistical probability values for changes between gray level i and j at a particular displacement distance (d) and angle (θ). It can be computed by dividing the frequency of a pixel with value i is adjacent to a pixel with value j by the total frequency of such comparisons made. For clear understanding, calculation of GLCM can be found in Figure 2 as an example.

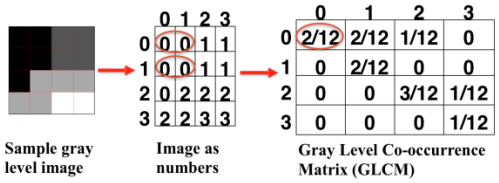

Figure 1. Sample GLCM calculation

In Figure 2, the leftmost image is the sample 4-by-4 grayscale image, the middle one is the gray scale image as a matrix of numbers where gray level values range from number 0 to 3 (0 for black and 3 for white) and the rightmost image is the GLCM matrix. The GLCM matrix in this example is calculated by considering the pixel combinations which are displacement distance 1 and horizontal direction meaning angle is 90 degree. For the first element of GLCM matrix, element [0, 0], we need to calculate how often gray level 0 and 0 at distance 1 in horizontal direction occur in an image. According to our sample figure, the matrix element P_{ij} is two divided by twelve because there are two combinations for gray level pair (0, 0) and there are total twelve combinations in the matrix. By equation, the GLCM matrix elements can be described as follows:

$$Pro(x) = \{P_{ij}|(d,\theta)\}$$

where

$$P_{ij} = \frac{C_{ij}}{\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} C_{ij}}$$

C_{ij} = the number of occurrences of gray levels i and j within the window, at a certain (d, θ) pair

d = the distance between interested neighborhood pixels

θ = the angle between interested neighborhood pixels

N_g = the number of gray levels

The means for the columns and rows of the matrix can be respectively defined as:

$$\mu_x = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} i.P_{ij}$$

$$\mu_y = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} j.P_{ij}$$

The standard deviations for the columns and rows of the matrix can be defined as:

$$\sigma_x = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} (i-\mu_x)^2.P_{ij}$$

$$\sigma_y = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} (i-\mu_y)^2.P_{ij}$$

The eight textural features used in our study are successfully employed in ultrasonic texture evaluation of parotid glands (Yang et al., 2012). The first two measures used in our study fall into orderliness group in which P_{ij} values themselves relating to commonness of occurrence are used as some form of weights. A weight that increases with commonness will yield a texture measure that increases with orderliness. A weight decreasing with commonness yields a texture measure increasing with disorder.

(1) Angular Second Moment (Uniformity)

$$\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} {P_{ij}}^2$$

Angular Second Moment (ASM) is used to measure the textural uniformity. Since this measure uses each P_{ij} as a weight for itself, the value of this measure is high when the area of the image is very orderly.

(2) Entropy (Randomness)

$$\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} P_{ij}\log(P_{ij})$$

Entropy which is negatively correlated to ASM is a measure of the disorder or randomness of the area of the image. As the average weight for each GLCM position is a negative logarithmic function of the frequency, it yields higher values

when there is a random distribution of brightness values.

The two measures relate to contrast group where weights are set depending on the distance from the GLCM diagonal.

(3) Contrast (Local heterogeneity)

$$\sum_{k=0}^{N_g-1} k^2 \left\{ \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_{ij} \| i-j \| = k \right\}$$

Contrast measures the local variations presented in an image which is negatively correlated with Inverse Difference Moment (IDM). Its weighting factor is the square of the gray level difference and it is going to increase exponentially away from the diagonal. The value of this measure is high when there is a large amount of differences in tone in an image.

(4) Inverse Difference Moment (Local homogeneity)

$$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \frac{1}{1+(i-j)^2} P_{ij}$$

Inverse Difference Moment (IDM) measures the local homogeneity of the image by the inverse of the contrast weight. Its weights decrease exponentially away from the diagonal. Thus, this measure results high values when there is little difference in gray level in an image. The next two measures are statistics derived from the GLC matrix.

(5) Variance (Global heterogeneity)

$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i-\mu_x)^2 . P_{ij} + \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i-\mu_y)^2 . P_{ij}$

Variance is a measurement of heterogeneity. It increases when the gray-scale values differ from their means because it puts high weights on the elements that differ from the average value of P_{ij}. Unlike contrast, variance has no spatial frequency.

(6) Correlation (Smoothness)

$$\frac{\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (ij) . P_{ij} - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Correlation indicates local gray-level dependency on the texture image; higher values can be obtained for similar gray-level regions.

To measure the skewness or asymmetry of the texture image, we use cluster shade and cluster prominence features.

(7) Cluster Shade (Skewness)

$$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i+j-\mu_x-\mu_y)^3 P_{ij}$$

Cluster Shade measures the skewness of the matrix which is a two-dimensional version of the gray-level histogram skewness. When the Cluster Shade is high, the image is asymmetric.

(8) Cluster Prominence (Asymmetry)

$$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i+j-\mu_x-\mu_y)^4 P_{ij}$$

Cluster Prominence is a measure of how peaked is a distribution which is two-dimensional version of the gray-level histogram kurtosis. Cluster Prominence is also a measure of asymmetry. When Cluster Prominence value is low, there is a peak in the GLCM matrix around the mean values and the image is more symmetric.

We employ eight textural features driven from GLCM showing the degrees of uniformity (ASM), homogeneity (IDM), heterogeneity (Contrast and Variance), smoothness (Correlation), randomness (Entropy), and asymmetry (Cluster Shade and Cluster Prominence) in our experiment. By finding sentiment correlations between these textural parameters and voice source parameters for three different phonation differences, we can find the possibility of the visualization of the voice source characteristics with image textural characteristics.

4. Results and Discussion of Sentiment Association Analysis

In our experiment, we cannot find any significant correlations between F0, H2-H4, H1-H2, H1*-A1*, H1*-A2*, H2*-H4*, Cepstral Peak Prominence (CPP) features and eight textural features. Observed correlation scores between other voice source related acoustic measures and image texture features are shown in Table 1.

Table 1. Correlation results between voice source acoustic parameters and image texture parameters

	ASM	IDM	Contrast	Entropy	Variance	Shade	Prominence	Correlation
H1*-H2*	0.439	-0.530	0.630	-0.284	-0.633	-0.337	-0.653	0.409
H1*-A3*	0.545	-0.270	0.442	-0.396	-0.534	-0.193	-0.628	0.466
H1-A3	0.404	-0.564	0.467	-0.339	-0.678	-0.336	-0.708	0.460
H1-A1	0.384	-0.596	0.637	-0.223	-0.620	-0.294	-0.638	0.354
H1-A2	0.347	-0.563	0.497	-0.278	-0.629	-0.347	-0.643	0.361
HNR05	0.356	-0.496	0.363	-0.306	-0.597	-0.132	-0.619	0.393
HNR35	0.288	-0.520	0.345	-0.289	-0.601	-0.326	-0.628	0.365

In this paper, we can confirm that H1*-H2* and H1-A1 have positive correlations with Contrast and negative correlations with Variance and Prominence. The acoustic measures H1-A2, H1-A3 and Harmonic-to-Noise ratio in frequency band (0 to 3500 Hz) have negative correlations with textural

features Variance and Prominence. Moreover, there are also negative correlations between H1*-A3*, Harmonic-to-Noise Ratio in frequency band (0 to 500 Hz) and Prominence.

There are high positive correlations between H1*-H2* and H1-A3 and textural feature Contrast. From these correlations, we can say that the higher the degree to which intensity drops off as frequency increases, the higher the gray level difference between neighboring pixels in the texture image.

From the significant negative correlation results between acoustic features and textural feature Variance, we can conclude that the higher the degree to which intensity drops off as frequency increases and the higher in HNR decreasing the noise level, the less dispersion of gray levels from their mean in the texture. In addition, from other negative correlations between acoustic features and Prominence, we can say that the higher the degree to which intensity drops off as frequency increases and the higher in HNR decreasing the noise level, the lower the peakedness of the gray level distribution in the texture image.

5. Conclusions

To quantify sentiment correlations between speech and image, we have measured the direct correlations between voice source characteristics and texture image parameters. Through the experiments of texture sample selection by listening speech with different voicing (modal, creaky and breathy), we could have confirmed the high correlations between spectral slope features (H1*-H2*, H1*-A3*, H1-A1, H1-A2, H1-A3) and Harmonic-to-Noise Ratios (HNRs) in the frequency band (0 to 500Hz)(0 to 3500Hz) and textural features Contrast (local heterogeneity), Variance (global heterogeneity) and Prominence (asymmetry). We could have observed high positive correlations between amplitude difference between first and second harmonics corrected for vocal tract effects (H1*-H2*), amplitude difference between first harmonic and first formant (H1-A1) and Contrast (0.630 and 0.637 respectively), high negative correlations between H1*-H2*, H1-A1, H1-A2, H1-A3, Harmonic-to-Noise Ratio (HNR35) and Variance (0.633, 0.620, 0.629, 0.678 and 0.601 respectively) and Prominence (0.653, 0.638, 0.643, 0.708 and 0.628 respectively) and negative correlations between H1*-A3*, Harmonic-to-Noise Ratio (HNR) in first frequency band (0 to 500 Hz) reflecting voice source differences and textural feature Prominence (0.628 and 0.619 respectively).

These results confirm the existence of sentiment correlations between speech source and image texture more exactly than the conventional studies. Although the experiment was conducted only with Myanmar people in this study, we believe that these sentiment correlations are language universal. We expect that further specifications of related parameters and their computational correlation modeling using statistical machine learning methodology such as Deep Neural Network (DNN) will enable direct visualization of speech information. The current study gives the promising result to pursue media transform which are expected in many application fields.

This kind of cross-modal association between speech and other media including language can not only differentiate between reading speech and communicative speech but also can be applicable in hearing impaired persons as well as in L2 language learners to be easily understandable speech variations using sentiment information expressed by other media. Further applications are highly expected for coaching sports where detailed prosodic differences and onomatopoeia are important (Fujino et al., 2010).

Acknowledgements

This work was partly supported by Grand-in-aid for Science Research B, NO. 23320091 of JSPS.

References

[1] C. Menezes, K. Maekawa, and H. Kawahara, "Perception of Voice Quality in Paralinguistic Information Types: A Preliminary Study", *Proceedings of the 20th General Meeting of the PSJ*, 2006, pp.153-158.

[2] E.B. Holmberg, R.E. Hillman, J.S. Perkell, P.C. Guiod, and S.L. Goldman, "Comparisons among Aerodynamic, Electroglottographic, and Acoustic Spectral Measures of Female Voice", *Journal of Speech and Hearing Research*, 38, 1995, pp. 1212-1223.

[3] G. de. Krom, "A Cepstrum-based Technique for Determining a Harmonic-to-noise Ratio in Speech Signals", *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, 1993, pp. 254-266.

[4] H. Hanson, "Glottal Characteristics of Female Speakers: Acoustic Correlates", *J. Acoust. Soc. Am*, 101, 1997, pp. 466-481.

[5] H. Kawahara, I. Masud-Katsuse and A. De. Cheveigne "Restructuring Speech Representations using a Pitch-adaptive Time Frequency Smoothing

and an Instantaneous-frequency based F0 Extraction: Possible Role of a Repetitive Structure in Sounds", *Speech Communication*, vol. 27, no. 3, 1999, pp. 187-207.

[6] J. Hillenbrand, R. A. Cleveland, and R.L. Erickson, "Acoustic Correlates of Breathy Vocal Quality", *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, 1994, pp. 769-77.

[7] K. Johnson, *Acoustic and Auditory Phonetics*, MA: Blackwell Publishers, Cambridge, 1997.

[8] K. Li, Y. Greenberg, and Y. Sagisaka, "Inter-language Prosodic Style Modification Experiment using Word Impression Vector for Communicative Speech Generation, *8th Annual Conference of the International Speech Communication Association*, InterSpeech, 2007, pp. 1294-1297.

[9] K. Watanbe, Y. Greenberg, and Y. Sagisaka, "Sentiment Analysis of Color Attributes Derived from Vowel Sound Impression for Multimodal Expression", *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA). IEEE*, 2014, pp. 1-5.

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

[11] L. Shao, Y. Greenberg and Y. Sagisaka, "Global F0 Control Parameter Prediction based on Impressions for Communicative Prosody Genereation", *Oriental COCOSDA held jointly with 2013 conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/ CASLRE)*, 2013 International Conference. IEEE, 2013, pp. 1-4.

[12] M. Gordon, and P. Ladefoged, "Phonation types: A Cross Linguistic Overview", *Journal of Phonetics*, 29(4), 2011, pp. 383-406.

[13] M. Wrembel., and K. Rataj, "Sounds like a Rainbow-Sound-Colour Mappings in Vowel Perception", in ExLing, 2008, pp. 237-240.

[14] 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", *18th International Congress of Phonetic Sciences (ICPhS)*, 2015.

[15] R. Jacobsen, *Selected Writings i: Phonological Studies*, The Hague: Mouton, 1962.

[16] R.M. Haralick, K. Shanmugam, and I.H. Dinstein, "Textural Features for Image Classification", *Systems, Man and Cybernetics*, IEEE Transactions on, no. 6, 1973, pp. 610-621.

[17] X. Yang, S. Tridanapani, J.J. Beitler, S.Y. David, E.J. Yoshida, W.J. Curran, and T. Liu, "Ultrasound Glcm Texture Analysis of Radiation-induced Parotid-gland Injury in Head-and-neck Cancer Radiotherapy: an in Vivo Study of Late Toxicity", *Medical physics*, vol. 39, no. 9, 2012, pp. 5732-5739.

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

[19] Y. Greenberg, N. Shibuya., M. Tsuzaki., H. Kato, and Y. Sagiska, "Analysis on Paralinguistic Prosody Control in Perceptual Impression Space using Nultiple Dimensional Scaling", *Speech Communication*, vol. 51, no. 7, 2009, pp. 585-593.

[20] Y. Greenberg, M. Tsuzaki., H. Kato, and Y. Sagisaka, "Communicative Speech Synthesis using Constituent Word Attributes", 9th European Conference on Speech Communication and Technology, 2005, pp. 517-520.

[21] Y. Isonaka, Y. Kanno, K. Watanabe and Y. Sagisaka, "Perceptual Impressions of Japanese Phones (in Japanese)", *Acoustic Society of Japan Spring Meeting*, 2015, pp. 919-922.

[22] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for Voice Analysis", *Proceedings of the ICPhS XVII*, 2011, pp. 1846-1849.

[23] Y. Sagisaka, T. Yamashita, and Y. Kokenawa, "Generation and perception of F0 markedness for communicative speech synthesis", *Speech Communication*, vol. 46, no. 3, 2005, pp. 376-384.