INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Cross-modal Analysis between Phonation Differences and Texture Images based on Sentiment Correlations

Win Thuzar Kyaw, Yoshinori Sagisaka

Graduate School of Fundamental Science and Engineering
Department of Pure and Applied Mathematics
Waseda University, Japan

winthuzarkyaw@akane.waseda.jp, sagisaka@waseda.jp

Abstract

Motivated by the success of speech characteristics representation by color attributes, we analyzed the cross-modal sentiment correlations between voice source characteristics and textural image characteristics. For the analysis, we employed vowel sounds with representative three phonation differences (modal, creaky and breathy) and 36 texture images with 36 semantic attributes (e.g., banded, cracked and scaly) annotated one semantic attribute for each texture. By asking 40 subjects to select the most fitted textures from 36 figures with different textures after listening 30 speech samples with different phonations, we measured the correlations between acoustic parameters showing voice source variations and the parameters of selected textural image differences showing coarseness, contrast, directionality, busyness, complexity and strength. From the texture classifications, voice characteristics can be roughly characterized by textural differences: modal- gauzy, banded and smeared, creaky - porous, crystalline, cracked and scaly, breathy - smeared, freckled and stained. We have also found significant correlations between voice source acoustic parameters and textural parameters. These correlations suggest the possibility of cross-modal mapping between voice source characteristics and textural parameters, which enables visualization of speech information with source variations reflecting human sentiment perception.

Index Terms: cross-modal sentiment correlation, phonation differences, voice source features, textural features

1. 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 para-linguistic 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 [7-9][12][15]. 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 [12] [15].

Quite recently, as high correlations have been observed between speech features and color parameters in image medium [18-19], we started to look for another big possibility for speech descriptions using image related information by replacing word expressions of language medium. If we can effectively describe speech showing para-linguistic differences in real fields by other media such as colors and image features, we can employ it not only in communicative prosody generation and perception, but also in more general purposes such as speech visualization and cross-modal information mapping where many useful applications can be considered. In this paper, as the next trial to describe another sentiment information embedded in speech, we focused on the voice characteristics of source differences. As a cross-modal description of them, we employed texture differences where we do not need to express their differences using linguistic terms with scores. After perceptual experiments on texture selection through the perception of speech with different voice source characteristics, we could have succeeded to roughly describe voice source differences and calculate direct correlations between voice source parameters and texture feature parameters.

In Section 2, we introduce the conventional perceptual psychology studies on the correlations between acoustic characteristics and textures [14] together with speech and texture parameters employed in this paper. In Section 3, we describe the texture selection experiment based on listening of speech with different source characteristics. In Section 4, after showing how the three different source characteristics employed in the experiment (modal, creaky and breathy) can be roughly correspond to the texture images which also gives expressions in linguistic terms, we show direct correlations between voice source parameters and texture feature parameters. In Section 5, we wrap up our findings and summarize the current understandings on the correlations between voice source characteristics and texture feature parameters. In Section 6, we conclude our paper.

2. Sentiment correlation between voice source characteristics and texture images

2.1. Previous studies on voice-source and texture correlation

For the description of voice source differences, we have been employing word expressions such as modal, creaky and breathy. The selection of these words were based on their perceptual impressions of phonation differences for the purpose of speech generation research. For the purpose of speech description of para-linguistic information, we believe that the sentiment information description is quite useful to understand perceptual impressions and their control from the success of our studies on

http://dx.doi.org/10.21437/Interspeech.2017-1236

communicative prosody control [7-9][12][15].

By consulting with studies on voice-source and texture correlation, we could find perceptual psychology studies on their correlations [14]. Up to now, since we did not find any direct correlations between voice source and texture parameters neither in speech processing nor in image processing, we decided to calculate direct correlations between voice source and image parameters. In [14], Moos et al conducted the experiment to test color and texture associations in voice-induced synesthesia by using voice recordings in ten different phonation types as speech stimuli and texture images which were selected from some databases based on the descriptive words used by the voice synesthetes (people experiencing concurrent perceptions generating by the sound of people's voices) of their textural concurrents. They parameterized the voices using four acoustic features relating to overall pitch, pitch range, vocal tract settings and vocal tract size, and vocal fold vibration. For textural parameters, they used subjective scores based on human perceptual judgements. Then, they quantitatively analyzed the influence of acoustic attributes of the different phonation types on participants' responses.

As we have found that the scientific understanding has guided us better description and modeling in speech prosody variations in the previous studies [18-19], we started our scientific analysis on the direct correlations between voice source characteristics and texture feature parameters. Though we have already started our analyses by our own framework, we thought it is useful to adopt the idea of employing texture images with descriptive words in our experiment as the previous study [14] for mutual understanding.

2.2. Voice parameters describing source characteristics

As seen in high direct correlations between speech features and color parameters [18-19], we think that direct correlation calculations between modality driven parameters will provide scientifically better understandings and application possibilities than conventional indirect correlations using perceptual impressions employing categorical language expressions. As the first step of direct source-texture analysis, we decided to employ three representative phonation types (modal, creaky and breathy) which have been widely studied in speech production field. In the study of voice-texture associations in voice-induced synesthesia [14], they used four acoustic features measuring the overall pitch of the voice, vocal tract settings and vocal tract size, vocal fold vibration and variability of F0 in a speaker. To measure acoustic characteristics relating to vocal fold vibration, they employed a spectral tilt feature, H1*-A3* (the corrected first harmonic minus the corrected amplitude of third formant).

For our direct correlation calculations between voice sources and textures, we used voice source related acoustic features measuring a spectral slope, periodicity and a spectral noise level of the speech signal which are the important cues for phonation differences.

The spectral slope can be measured by comparing the amplitude of the first two harmonics or by comparing the amplitude of the first harmonic to the amplitude of the spectral peak of a formant. The amplitude difference between the first two harmonics corrected for vocal tract effects (H1*-H2*) and its uncorrected version, H1-H2 can be thought of as the measures of breathiness [4][10]. H1*-Ai* representing the amplitude difference of the first harmonic and the spectral peak of the first or second or third formant corrected for the effects of the formants (H1*-A1*, H1*-A2*, H1*-A3*) and their uncorrected versions H1-A1, H1-A2 and H1-A3 was shown correlated to the source spectral tilt [5][10]. The spectral slope is the most steeply positive for creaky vowels because there are stronger energy in high frequency region and the most steeply negative for breathy vowels because there are weaker energy in the high frequency region.

To estimate periodicity, we used Cepstral Peak Prominence (CPP) which is a measure of cepstral peak amplitude normalized by overall amplitude [4]. The peaks in the cepstral domain are larger for a well defined periodic source such as modal voice than a less periodic one like breathy voice. Estimation of spectral noise by separating the harmonics component from the noise component can be done using Harmonic-to-Noise Ratio (HNR) [3]. Increased noise in the breathy signal results in lower HNR value. In addition, we also measured F0 and Energy of the speech signal.

All these parameters employed in this study were obtained from five Japanese vowels with three different phonation types by using VoiceSauce application [13] with default settings.

2.3. Texture parameters for correlation analysis

In psychological study [14], texture images were quantified by participants' ratings along eight semantic scales of the perceptual space of textures with descriptive words. Same as acoustic parameters introduced in the previous Section 2.2, we employed the following textural parameters extracted from a texture image for the replacement of indirect parameters of perceptual impressions employing categorical language expressions. Thus, we adopted successful computational textural parameters embedding human visual impressions from image processing area.

Three successful features based on the co-occurrence matrix from the work of Tamura et al. [1] called coarseness, contrast and directionality and three features based on Neighborhood Gray-Tone Difference Matrix (NGTDM) from the work of Amadasun et al. [2], namely busyness, complexity and strength were considered as the first step in this study. The conceptual definitions of each of these features are described as follows.

Coarseness: Coarseness refers to the size and number of texture primitives. A coarse texture has the characteristics of containing large-sized and repetitive texture primitives and a high degree of local uniformity of gray levels. Conversely, a fine texture is composed of texture elements which are small-sized and less-repetitive texture primitives and a high degree of local variations of gray levels. Sometimes, busy or close can be used in place of fine.

Contrast: Contrast can be generally defined as gray level differences in neighboring pixels. It is influenced by the gray-levels in the image, the ratio of white and black in the image and the intensity change frequency of gray-levels.

Directionality: Directionality value is dependent on both element shape and placement rule. In this measure, the total degree of directionality is considered. Thus, two texture patterns will result the same degree of directionality if they have only different orientations.

Busyness: The degree of busyness is the spatial rate of change in intensity from one pixel to its neighbor by suppressing the effect of contrast variations.

Complexity: A texture is considered to be complex if it consists of many texture elements which have different average intensities or many patches. In addition, a texture containing many sharp edges and/or lines can be thought of as a complex texture.

Strength: If the texture primitives including in a texture image are easily definable and clearly visible, that kind of texture

can be referred to as a strong texture. The distinctions can be made between the component primitives of a texture depending on a considerable extent upon the sizes of the primitives and the differences between their average intensities.

We used publicly available JFeatureLib (version 1.6.2) [20] to extract coarseness, contrast and directionality feature values and Image Despeckle Filtering Toolbox [17] to extract busyness, complexity and strength values of texture images.

3. Experiment on texture selection based on perceptual impression of speech with different source characteristics

3.1. Experimental setup

Forty Japanese students (20 males and 20 females) with the age ranging from 18 to 25 years participated in the experiment. The experiment was conducted in a quiet room. Each participant was asked to take a seat in front of a computer screen and to listen to one of thirty speech samples at a time by using a head phone. It was allowed to replay the speech sample if the participant felt one time was not enough for the judgement. The speech stimuli were randomized for every participant. All 36 texture images with the resolution of 128 $\times$ 128 were displayed on a computer screen. After listening to each speech sample, the participant was instructed to select the well-matched texture image with the speech sample depending on his/her perception.

3.2. Speech and texture data employed in the experiment

As the first step, we considered three of the most common phonation types: modal, creaky and breathy to understand voice source characteristics. We employed single vowel samples of five Japanese vowels /a/, /i/, /u/, /e/, and /o/. These five Japanese vowels were produced with three different phonation types by two speakers (one Japanese male and one Japanese female) and recorded in a quiet room. In total, thirty speech stimuli were recorded for our experiment. We judged the phonation differences of three phonation types uttered by two speakers perceptually. All recorded audio were sampled with 44kHz in 16 bits.

Texture images employed in the cross-modal correlation experiment are important. In the study of voice-induced synesthesia [14], their texture selection was based on the descriptive words used by the synesthetes of their textural concurrents. For our analysis on cross-modal correlations between speech and texture, we adopted a publicly available texture dataset DTD (Describable Texture Dataset) [16] from image processing filed consisting of texture images tying descriptive words based on human judgement. DTD is a collection of real-world texture images downloaded from the Internet rather than being captured with controlled settings in a laboratory. The texture images in DTD dataset are attached with one or more descriptive texture terms from 47 terms which were carefully selected from 98 texture terms of Bhusan's work [6]. This reduction was carried out by removing words independent to visual characteristics of textures and by combining some terms which have similar meanings into a single term.

By checking all texture images of DTD dataset, we removed some attributes such as pothholed, chequered, cobwebbed which we do not consider relevant to our study. Finally, we reduced 47 words to 36 words which correspond to 36 texture images from DTD dataset. As shown in Figure 1, we employed these 36 images for our texture stimuli.

Each selected texture image has only one semantic annotation and shows uniform abstract structure in the entire image and no change in the impression even any part is cut out. These images were converted into 256 (8bit) gray scale images in order to eliminate the color components of visual information and resized into 128 $\times$ 128 pixels by the image editing software GIMP (ver.2.8.6) for texture display on a computer screen.

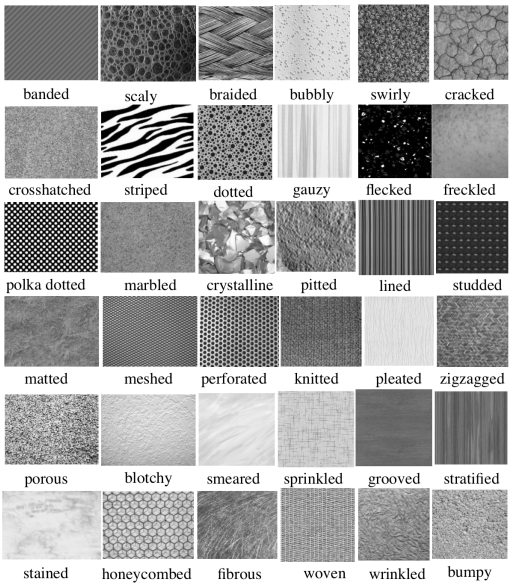

Figure 1: *Textures employed in the selection-by-listening experiment.*

4. Experimental results

We conducted a perceptual experiment asking subjects to choose texture images based on perceptual impressions on vowels with phonation differences. From our experimental results, we could find significant associations between phonation differences and texture images determined by the high percentages of selections. Moreover, we quantitatively analyzed for the direct correlations between voice source characteristics and textural characteristics using voice source related acoustic parameters and statistical textural parameters with psychological judgements.

4.1. Associations between phonation differences and textures

There were totally 400 selections for each phonation type (2 speakers $\times$ 5 vowels $\times$ 40 participants). Therefore, the average number of selections for each textural pattern was 11 (400 total selections / 36 textural patterns). By investigating the high percentages of associations for each phonation type, we could roughly characterize voice source differences by textural patterns. To visualize their associations clearly, we could represent them as in Figure 2. From our experimental results, we had the same finding of the association of creaky voice with cracked texture as in study [14] and also with porous, crystalline and scaly textures. Moreover, there were associations of modal voice with gauzy, banded and smeared textures and breathy voice with smeared, freckled and stained textures respectively.

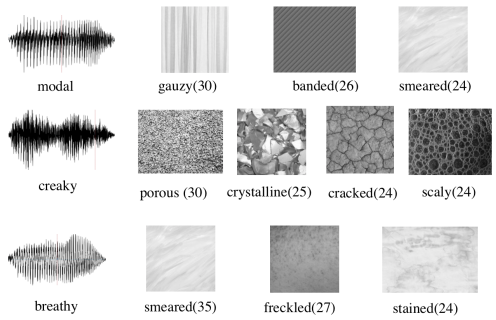

Figure 2: *Associations between phonation differences and texture patterns (specific percentages of speech and texture association are described in parenthesis).*

4.2. Sentiment correlation between voice source features and textural features embedding human visual impressions

For scientific understanding of sentiment correlations between speech and texture, we carried out quantitative correlation analysis between speech features showing voice source characteristics and textural features representing textural characteristics. The correlation scores were averaged across all subjects and were calculated based on total 1200 selections (3 phonation types $\times$ 400 selections for each phonation type). Our direct correlation scores are as shown in Table 1.

5. Discussions

By judging from the correlation results presented in the previous section, we could summarize and interpret their correlations as follows:

From the correlations between spectral slope measures and textural features, the participants tend to choose coarser, less contrast, more directional and less complex texture image if they perceive the speech signal which is less intensity in higher frequencies. From these findings, we can interpret as the participants are more likely to select the texture image constituting texture primitives which are large-sized, locally uniform and low gray level differences, directional and less inclusion of lines or edges or patches when they perceive breathy vowel which is weaker intensity in higher frequencies. On the other hand, by listening to the the creaky vowels which are stronger intensity in higher frequencies, the participants tend to choose less coarse, high contrast, less directional, more complex and stronger texture images.

Moreover, by summarizing from correlations of aspiration noise measures and textural features, when the participant listens to speech signal which is higher aspiration noise giving low HNR and less periodic having low CPP, they would like to choose texture image which is higher coarseness, less contrast, higher strength and higher busyness. We can explain that if the participant perceives breathy signal which is less periodic and includes more noise, they are likely to select texture image containing large texture primitives, low gray level differences and clearly visible. The participant would like to choose the images which are high coarseness, low contrast and low complex for the speech signal with high F0 and busy and strong images for the speech signal with high energy. From the correlations between voice source features and textural features, we can think of the possibilities to describe voice source variations using texture.

Table 1: *Sentiment correlation between voice source features and textural features embedding human visual impressions*

Voice source features	Textural features	Correlation scores
H1-H2	coarseness	0.501
H1-H2	contrast	−0.636
H1-A1	contrast	−0.522
H1-A2	contrast	−0.652
H1-A3	directionality	0.511
H1-A3	complexity	−0.534
HNR (0 to 500 Hz)	busyness	−0.546
HNR (0 to 500 Hz)	strength	−0.531
HNR (0 to 1500 Hz)	strength	−0.586
HNR (0 to 2500 Hz)	strength	−0.520
CPP	coarseness	−0.597
CPP	contrast	0.547
F0	coarseness	0.754
F0	contrast	−0.605
F0	complexity	−0.752
Energy	busyness	0.583
Energy	strength	0.707

6. Conclusions

Aiming at visualization of voice source differences through direct mapping from voice source parameters to image texture parameters reflecting human sentiment perception, we carried out a perceptual experiment where the most fitted texture images were selected by listening vowel sounds with three different voice source characteristics (modal, creaky, breathy). By judging the participants' texture selection based on perceptual impression of speech, these voice characteristics can be roughly characterized by textural differences: modal - gauzy, banded and smeared, creaky - porous, crystalline, cracked and scaly, breathy - smeared, freckled and stained.

Moreover, quantitative analyses of sentiment correlations showed the correlations between voice source acoustic parameters and computational textural parameters embedding human visual impressions. By matching the interpretation of these underlying mappings with the visual inspection of participants' texture selections based on perceptual impression of speech with phonation differences, we can think of the possibilities to describe voice source variations using texture.

Though the current study is the only first step towards the direct mapping from voice source parameters to textural image parameters using typical voice source differences and the existing texture image database DTD, we will be able to scale up the more detailed analysis through finding new source parameters representing sentiment information and creating texture images directly associated with sentiment information. We believe this type of scientific analyses will reveal underlying sentiment correlations existing in multiple media, language, speech and image, which will be able to provide more human friendly multimodal information expressions.

7. References

[1] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 6, pp. 460–473, 1978.

[2] M. Amadasun, and R. King, "Textural features corresponding to

textural properties," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 5, pp. 1264–1274, 1989.

[3] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[4] 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, pp. 769–778, 1994.

[5] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–481, 1997.

[6] N. Bhushan, A. R. Rao, and G. L. Lohse, "The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images," *Cognitive Science*, vol. 21, no. 2, pp. 219–246, 1997.

[7] Y. Sagisaka, T. Tamashita, and Y. Kokenawa, "Generation and perception of f 0 markedness for communicative speech synthesis," *Speech Communication*, vol. 46, no. 3, pp. 376–384, 2005.

[8] Y. Greenberg, M. Tsuzaki, H. Kato, and Y. Sagisaka, "Communicative speech synthesis using constituent word attributes," in *9th European Conference on Speech Communication and Technology*, 2005, pp. 517–520.

[9] 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.

[10] J. Kreiman, M. Iseli, J. Neubauer, Y. L. Shue, B. R. Gerratt, and A. Alwan, "The relationship between open quotient and H1*-H2*," *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2495–2495, 2008.

[11] S. U. D. Khan, "An acoustic and electroglottographic study of breathy phonation in Gujarati," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2222–2222, 2009.

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

[13] Y. L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," *Energy*, vol. 1, no. H2, pp. H1–A1, 2010.

[14] A. Moos, D. Simmons, J. Simner, and R. Smith, "Color and texture associations in voice-induced synesthesia," *Frontiers in Psychology*, vol. 4, no. 568, pp. 1–12, 2013.

[15] L. Shao, Y. Greenberg, and Y. Sagisaka, "Global f0 control parameter prediction based on impressions for communicative prosody genereation," 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.

[16] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] C. P. Loizou, C. Theofanous, M. Pantziaris, and T. Kasparis, "Despeckle filtering software toolbox for ultrasound imaging of the common carotid artery," *Journal of Computer Methods and Programs in biomedicine*, vol. 114, no. 1, pp. 109–124, 2014.

[18] 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.

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

[20] F. Graf "JFeatureLib v1.6.3," *http://dx.doi.org/10.5281/zenodo.31793*, October, 2015.