J. Chosun Natural Sci.
Vol. 11, No. 3 (2018) pp. 131 – 138
https://doi.org/10.13160/ricns.2018.11.3.131

A Study on Image Recommendation System based on Speech Emotion Information

Tae Yeun Kim[1] and Sang Hyun Bae[2†]

Abstract

In this paper, we have implemented speeches that utilized the emotion information of the user's speech and image matching and recommendation system. To classify the user's emotional information of speech, the emotional information of speech about the user's speech is extracted and classified using the PLP algorithm. After classification, an emotional DB of speech is constructed. Moreover, emotional color and emotional vocabulary through factor analysis are matched to one space in order to classify emotional information of image. And a standardized image recommendation system based on the matching of each keyword with the BM-GA algorithm for the data of the emotional information of speech and emotional information of image according to the more appropriate emotional information of speech of the user. As a result of the performance evaluation, recognition rate of standardized vocabulary in four stages according to speech was 80.48% on average and system user satisfaction was 82.4%. Therefore, it is expected that the classification of images according to the user's speech information will be helpful for the study of emotional exchange between the user and the computer.

Keywords : Boyer-Moore Algorithm, Genetic Algorithm, Emotion recognition, PLP Algorithm, Recommendation System

1. Introduction

In order to effectively reflect the various requirements of the user, researches on recommendation techniques considering user's tendency actively conducted. An application containing a recommendation technique is used to predict the item that the user is interested in and to recommend the item[1,2]. It is necessary to understand and reflect the user's characteristics and situation such as personal preference and personal emotion in order to increase the user's satisfaction.

Human emotion means a complex state that occurs within the human body due to senses and perceptions that respond to external physical stimuli obtained through experience[3,4].

The speech, particularly, which is one of the emotion information of the user, is used not only as means for communication but also as means for conveying emotion. The emotion contained in the human speech represents psychological state of the speaker and is involved in communication with the other party. In addition, the image, which is one of the visual information, among a lot of emotion information is formed in a short time and lasts for a long time in memory, which is considered to be an important factor in successful marketing and plays a significant role in understanding and interpreting human emotion[5].

Speech-based emotion recognition is a technology that automatically recognizes emotions of a user by analyzing a user's speech signal. In particular, the speech emotion information is information indicating the current emotion state of the user, and is very useful for the cultural contents service such as music recommendation or the emotion monitoring of the user depending on the emotion state. Recently, as the smartphone has been popularized, it has become easier to collect and process user's speech data, so research on emotion recognition technology has been actively carried out[6].

There are two researches to improve the accuracy by extracting new features or applying classification methodology differently. As a new feature extraction research, it is a technique that recognizes instant emotions without determining the window size because it uses features that reflect the characteristics of each individual

[1]SW Convergence Education Institute, Chosun University, Gwangju
[2]Department of Computer Science & Statistics, Chosun University, Gwangju

[†]Corresponding author : shbae@chosun.ac.kr
(Received : August 30, 2018, Accepted : September 17, 2018)

vocalization[7]. There are techniques applying hierarchical classification methodology to apply classification methodology differently. In this study, three classifiers were used to classify similar emotional factors in speech, which showed high accuracy, but only three seconds of speech could be recognized[8]. In other researches, male and female training models are generated, and the input voice is first classified as male and female, and compared with training models suitable for gender[9].

In this paper, an image recommendation system using personal emotional information of speech is attempted to be implemented. The emotional information of speech was extracted and classified using the PLP algorithm for the human speech, and constructed the emotional DB of speech, classified the emotional information of image. In addition, a standardized image recommendation system is implemented based on emotion information of the user by matching the keyword of the emotional information of speech and the emotional information of image using the BM-GA algorithm.

This paper consists of 4 chapters. The second chapter looks into the composition and design of the suggested system, the third chapter examines the result of system implementation and describes the performance evaluation of the implemented system and the fourth chapter describes the conclusion of the study and future direction of research.

2. System Configuration and Design

This paper proposes a matching system based on the user's speech information, and the human emotions are defined as four cases: anger, hate, happiness, and sadness. And it recommends images that match the emotions of the individual using the BM-GA algorithm for recommendation.

First, the measured data is extracted and analyzed by the PLP algorithm using the microphone, and then stored in the emotional speech DB in the defined four steps of standardization (anger, hate, happiness, sadness). Also, the emotional DB of image is arranged in the plane of the same two-dimensional space with emotional color and emotional vocabulary, and the distance between emotional color and emotional vocabulary is measured to determine relevant information.

The measured information is extracted from the verification and representative emotional vocabulary using factor analysis. This data is stored in the emotional DB of image in four defined levels. The composed file generates automata using BM-GA algorithm and generates personalized emotion information in order to match the user's emotion information. Such generated emotion information is matched with a corresponding image.

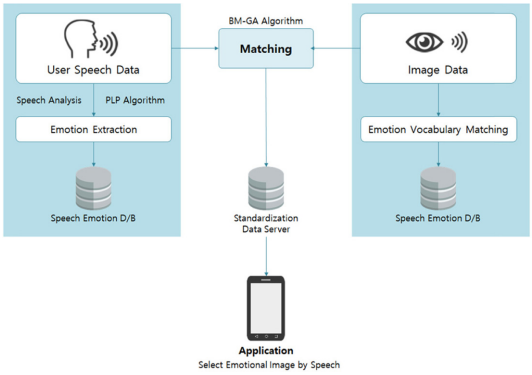


Fig. 1. The system configuration diagram.

The system structure of this paper is shown in Fig. 1.

2.1. Image Emotion Information

In this thesis, RGB, a web based color mode, is used to extract emotional information of image. 20 color emotion models defined in ';The Meaning of Color' of HP (Hewlett-Packard) are selected as the representative factors, and factor analysis and correspondence analysis were conducted through questionnaire survey of five-point scale, and created emotional space for each color.

To extract RGB from a specific point of an image, RGB of each color is stored in the database in advance, and each RGB of the color model is regarded as three-dimensional coordinates of x, y, z in order to grasp the degree of emotion according to the color distribution. The color model distribution of each image is stored in each of the 20 color model fields of the database.

The analyzed color image is transformed into a measurement value of each color and expressed as a graph, and the emotion word is matched according to the rank having the highest measurement value[10].

237 (141 males, 96 females) were randomly selected to measure the emotion of the emotional information of image, and a questionnaire on emotional vocabulary of color information' respectively. Data from questionnaires were analyzed using factor analysis and emo-

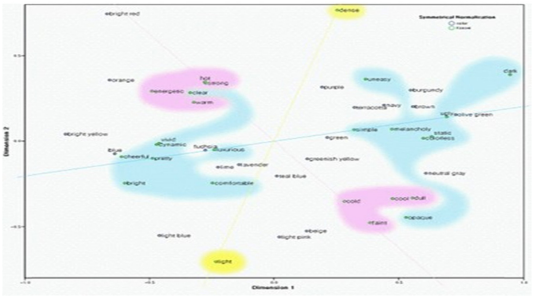


Fig. 2. Image distance measurement emotion.

tional vocabulary reading program, which programmed correspondence analysis, and the answer of the same classification is shown in Fig. 2.

As shown in Fig. 2, the emotional element and the emotional vocabulary are placed in the two-dimensional space, thereby obtaining the coordinates of the emotional vocabulary and the emotional element. The distance can be measured through these coordinates, which can be viewed as the relationship of emotional factors related to emotional vocabulary. In each coordinate, the smaller the distance, the higher the relation (in inverse proportion), the higher the color distribution in the image, the higher the meaning (in direct proportion). Equation (1) represents the equation for this[11].

$$D_{ik} = \frac{d_{ik}^{-1}}{\sum_{j=1}^{20} d_{ij}^{-1}} \tag{1}$$

The equation (1) represents the final result, which is distance between the actual emotional vocabulary (i) and the emotional element (k), and the numerator represents the distance between the emotional vocabulary (i) and the emotional element (k). And the denominator is the sum of the reciprocal of the distance to the emotional vocabulary (i) and 20 emotional vocabularies.

2.2. Speech Emotion Information

In this paper, the PLP algorithm to analyze speech information is used. PLP analysis is a technique that can simultaneously analyze the time domain and the frequency domain of a speech signal in consideration of human hearing emotion[12]. Human hearing emotion indicates that the human ear has different emotions depending on the frequency when hearing the sound. In other words, to hear the sound of the same intensity, it

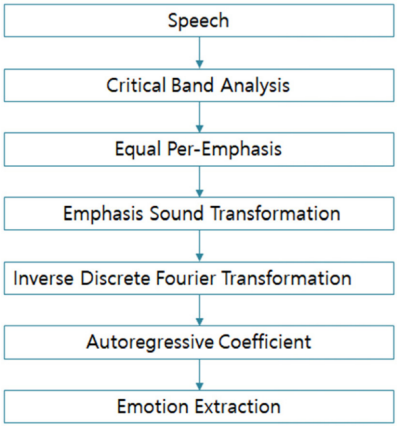


Fig. 3. PLP algorithm flow chart.

means that different intensity should be given to the frequency[13].

The PLP analysis algorithm is shown in Fig. 3. In Fig. 3, the critical analysis process is a step of obtaining the power of the Fourier-transformed signal, and the output power of the corresponding filter is obtained by passing through the band-specific filter.

If the speech value is input, the range of the center frequency of the filter can be obtained by converting the frequency as shown in equation (2). Here f_s is the sampling frequency.

$$z = 6\log\left[\frac{f_s}{600} + \sqrt{\left(\frac{f_s}{600}\right)^2 + 1}\right] \tag{2}$$

Equivalent pre-emphasis is the process of applying the intensity of the sound as the frequency changes. In other words, the change in frequency at low frequency shows a strong change in intensity, while the sensitivity to frequency change decreases with increasing frequency. By using such an equation (3) with this feature, the values generated by each filter are pre-emphasized[14].

$$E(w) = 1.151\sqrt{\frac{(w^2+144\times10^4)w^2}{(w^2+16\times10^4)(w^2+961\times10^4)}} \tag{3}$$

The change of emphasis finally performs the final processing in the frequency domain through the data compression as in equation (4). Here $X(w)$ is the speech signal converted into the frequency domain.

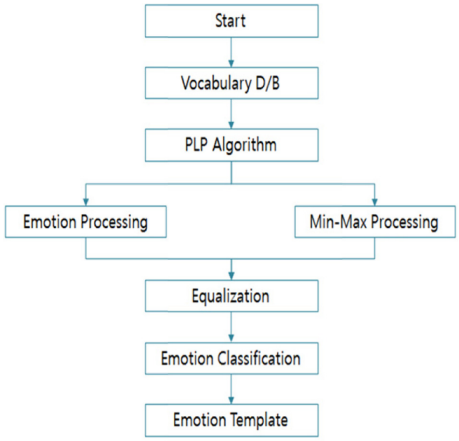


Fig. 4. Block diagram of personalized emotion-based template creation.

$$F(w) = |X(w)|^{1/3} \tag{4}$$

The above-mentioned three steps are processing in the frequency domain to find PLP parameters. This can be expressed as equation (5).

$$F_k = E(w_k)\int_0^{\pi} C_k(w)F(w)dw \tag{5}$$

The final PLP in the frequency domain is obtained by inverse Fourier transform as shown in equation (5). PLP analysis is mainly used for speaker recognition and has a characteristic of personal speech signal. In other words, the PLP analysis shows a lot of characteristics such as personal intonation, tone color, and language of a speech signal. The PLP analysis of these speech signals is considered to be an appropriate tool for personalized emotion recognition.

Then, the existing emotional template and the feature data of the currently generated speech signal are compared with all emotion templates using Euclidean distance as shown in equation (6).

$$D(T,P) = \sqrt{\sum_i^N (T_i - P_i)^2} \tag{6}$$

2.3. The Proposed Algorithm

The BM-GA algorithm proposed in this paper is used in order to recommend the data optimized for the users (anger, hate, happiness, sadness) in four steps from the emotional DB of speech and the emotional DB of image.

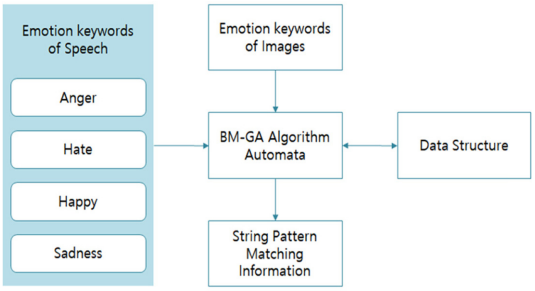


Fig. 5. String patten matching model.

Since the patterns in this thesis are normalized by expressions of emotion, GA algorithm is added to the BM algorithm among the single analysis algorithms, and the BM-GA algorithm that weights the repetitive emotion is used. The BM algorithm is fast because the comparison process starts at the end of the search pattern and processes it in the reverse direction, and when non-matching characters in the input test are not present in the search pattern, the next character can be moved to a large extent[15].

If a non-matching character occurs in the BM algorithm, the movement to the next character is determined by the following two criteria.

A) Algo1: The characters of the text at positions that do not match the characters of the pattern are compared with the characters in the pattern to determine the distance of movement of the pattern for the next comparison. If a text character does not exist in the search pattern, it can be shifted by the length from the beginning of the search pattern to the position of the non-matching character.

B) Algo2: When there is a substring that already matches the pattern in the text, it can be moved to that position if the matched substring appears at another position in the pattern. When text and pattern characters do not match, skip MAX (Algo1, Algo2) characters and perform the comparison process.

Fig. 5 defines a string pattern as a regular expression to detect various string patterns simultaneously with a specific keyword. The automata performs a function of detecting a real emotional string for a keyword which is a command stream to access the server. The repetitive emotion detected during the string pattern matching is stored in a separate data structure by the GA algorithm to reduce the repetitive matching time[16].

3. System Implementation Results and Performance Evaluation

In this paper, speech signal data, which is expressed by four emotions for 10 sentences in 5 male and 5 female sentences for personalized emotion recognition of voice, is used. And individualization of emotion recognition using speech signal is focused on lastly.

By analyzing the speech signal using PLP analysis, the values of PLP coefficients and the spectrum in the time domain can be obtained. In order to analyze the characteristics of the entire speech signal with the focus on individualization, we analyzed the characteristics of the emotion using the PLP spectral value. When the audio signal sampled at 16 KHz is subjected to PLP analysis of the twelfth order, 21 spectral values are generated in each time zone.

This spectrum value can be obtained as an average energy over all times to create a characteristic for one speech signal. Equation (7) is an equation for obtaining the average energy from the PLP analysis.

Here, the PLP spectrum value is represented by the frame number of the PLP spectrum and the total number of samples of the PLP spectrum is represented. Where P_i represents the PLP spectrum value, i is the frame number of the PLP spectrum, and N represents the total number of samples of the PLP spectrum.

$$E(i) = \frac{1}{N}\sum_{n=1}^{N} P_i(n) \tag{7}$$

Next, the difference between the minimum value and the maximum value for each spectrum for all times can be obtained as a feature of the speech signal. This can be expressed as equation (8).

$$M(i) = Max(P_i) - Min(P_i) \tag{8}$$

Experiments were applied to emotion recognition using the features of these two speech signals.

Since all speech signals are of different sizes, there is a problem in using them as general feature analysis data if they are used as they are. Therefore, it is necessary to pre-process the data through the normalization technique. In this paper, we use the Min-Max normalization method as shown in equation (9). Where P_i^* is the PLP spectral value of the min-max normalized speech signal and P is the PLP spectrum value for all frames of one speech signal.

$$P_i^* = \frac{p_i - \min(P)}{\max(P) - \min(P)} \tag{9}$$

Min-Max normalization is a normalization technique that scales a range of values to see how large a value is in that area. With this method, you can set the largest value to 1 and the remaining values to a value between 0 and 1. By performing the normalization in this manner, it becomes possible to obtain a characteristic of a speech signal that is not greatly affected by the size of the speech signal.

Table 1 shows the recognition rates for four standardized vocabularies according to speech using PLP algorithm. The average of 80.48% recognized the standardized vocabulary of anger, hate, happiness, and sadness.

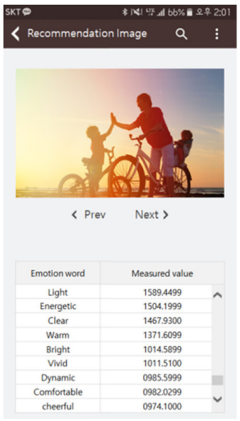
We also conducted 10-fold cross validation to minimize the influence from learning data and ensure reliability. Experimental data were divided into evaluation data and verification data, and they were used at a ratio of 7:3, and the system was optimized by leave-one-out method. Table 2 shows the accuracy (unit: %) of the verification data obtained from the experiment. Table 2

Table 1. Recognition result of speech emotion

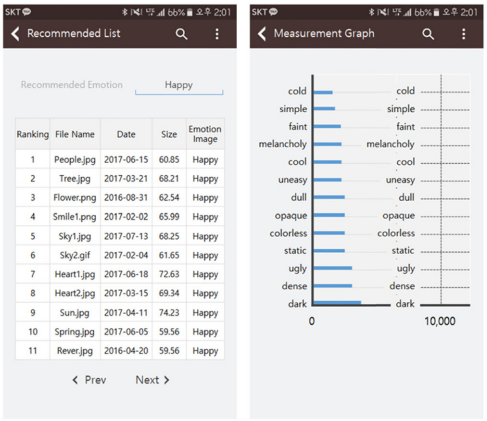
Types of emotion	Number of data (count)	Number of recognition (count)	Recognition rate (%)
Anger	60	48	80
Hate	11	9	81.82
Happy	27	22	81.48
Sadness	25	20	80
Total	123	99	80.48

Table 2. Performance evaluation of data

Fold No	Accuracy
1	83.5
2	86.7
3	88.6
4	88.2
5	84.2
6	83.7
7	84.3
8	85.6
9	85.1
10	83.1



(a) The recommended emotional image and measured values



(b) The recommendation list (c) the measured graph of emotional image

Fig. 6. Implementation result of emotional image recommendation system.

shows that the average accuracy of the proposed system in this paper is 85.3%, which shows that the accuracy of four emotional information matching for speech information is high.

In this paper, an image recommendation system using personal emotional information of speech is implemented as a mobile application. The image using the recognized emotion information value by predicting the user's preference is recommended. Fig. 6 shows the result of implementation of emotional image recommendation system based on user's emotional intelligence information implemented by the application in this paper. (a) of Fig. 6 shows the recommended emotional image and measured values, (b) is the recommendation list, and (c) is the measured graph of emotional image.

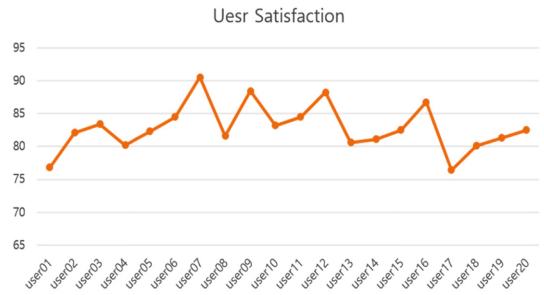


Fig. 7. User satisfaction of implemented system.

In this system, the user's emotional information of voice and emotional information of image are matched according to emotion information of the user who is standardized in four stages, and then a preferred image according to the user's selected pattern is searched in the standardized database and recommended by rank. By expressing the emotional analysis chart of the analyzed image, it was able to recommend the optimized image to the user.

The evaluation of the recommendation result of the system is measured by the user satisfaction of the degree of reliability of the users who use the system based on emotional information because of the characteristics of the subjective emotional information. To do this, 20 sample users were selected and four emotional keywords were presented, and the satisfaction of the search results was evaluated based on them.

As shown in Fig. 7, the sample user satisfaction rate of the emotion information of the retrieved images was 82.84%. This proves that the proposed emotion-based search result is relatively similar to the emotion felt by real people.

4. Conclusions

In this paper, we propose an image recommendation system using user's emotion information. We extracted and classified the speech emotion information using the PLP algorithm for the user's speech, constructed the voice emotion DB, classified the emotion color and emotional vocabulary through the factor analysis into one space and classified the image emotion information.

In addition, we tried to implement a standardized image recommendation system based on matching emotional information of users by matching each keyword by using BM-GA algorithm for speech emotion information and image emotion information.

Therefore, we standardize the user's emotional and image sensory information in four stages (anger, hate, happiness, sadness) and acquire and recognize speech and emotional information of image to extract reliable data. Algorithm, factor analysis, and correspondence analysis. By using the acquired emotional information and recommending an image suitable for each individual, we have implemented a service for recommending images according to user's emotional sensibility.

As a result of the performance evaluation, recognition rate of standardized vocabulary in four stages according to speech was 80.48% on average and system user satisfaction was 82.4%. Therefore, it is expected that the classification of images according to the user 's speech information will be helpful for the study of emotional exchange between the user and the computer.

Future research needs to analyze and study various algorithms to improve the recognition rate as shown in the emotional recognition of speech results. Therefore, a system with more stable recognition rate is attempted to be implemented by using emotion extracted through facial expression and speech.

Acknowledgments

This study was supported by research funds from Chosun University, 2017.

References

[1] H.-T. Choi and S.-B. Cho, "A Collaborative Filtering Recommendation System using ConceptNet-based Mood Classification by Genre," in Proceeding of the 38th Annual Symposium on Korea Information Science Society, Korea, pp. 216-219, Jun. 2011.

[2] B.-H. Oh, J.-H. Yang, and H.-J. Lee "A Hybrid Recommender System based on Collaborative Filtering with Selective Utilization of Content-based Predicted Ratings," Journal of KIISE, Vol. 41, No. 4, pp. 289-294, Apr. 2014.

[3] S.-Z. Lee, Y.-H. Seong, and H.-J. Kim, "Modeling and Measuring User Sensitivity for Customized Service of Music Contents," Journal of Korean Society For Computer Game, Vol. 26, No. 1, pp. 163-171, March 2013.

[4] J. Byun and D.-K. Kim, "Design and Implementation of Location Recommending Services using Personal Emotional Information based on Collaborative Filtering," Journal of the Korea Institute of Information and Communication Engineering, Vol. 20, No. 8, pp. 1407-1414, Aug. 2016.

[5] T.-Y. Kim, B.-H. Song, and S.-H. Bae, "A Design and Implementation of Music & Image Retrieval Recommendation System based on Emotion," Journal of The Institute of Electronics Engineers of Korea, Vol. 47, No. 1, pp. 73-79. 2010.

[6] J.-H. Bang and S.-Y. Lee, "Call Speech Emotion Recognition for Emotion based Services," Journal of KIISSE, Vol. 41, No. 3, pp. 208-213. 2014.

[7] A. B. Kandali, A. Routray, and T. K. Basu, "Emotion recognition from Assamese speeches using MFCC features and GMM classifier," TENCON 2008-2008 IEEE Region 10 Conference, pp.1-5, Nov, 2008.

[8] Z. Xiao, Dellandrea, L. Chen, and W. Dou, "Recognition of emotions in speech by a hierarchical approach," ACII 2009. 3rd International Conference, pp.401-408, 2009.

[9] Y.-h. Cho and K.-S. Park, "A Study on The Improvement of Emotion Recognition by Gender Discrimination," Journal of IEEK, Vol. 45, pp.401-408, 2008.

[10] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 25, No. 9, pp. 1075-1088, 2003.

[11] S.-K. Baek, "Kansei Distribution Space Creation by Visual Information Elements for Kansei-based Image Retrieval," Master's thesis, Chosun University, 2007.

[12] M. Zulkifly and N. Yahya, "Relative spectral-perceptual linear prediction (RASTA-PLP) speech signals analysis using singular value decomposition (SVD)," IEEE 3rd International Symposium in Robotics and Manufacturing Automation (ROMA), pp. 1-5, 2017.

[13] B.-W. Jung, "Personalized Emotion Recognition using PLP Analysis of Speech Signal," Master thesis Pusan National University, 2008.

[14] S. Hwangbo, S.-Y. Chun, S.-Y Gang, and C.-S. Lee, "Lighting Control using Frequency Analysis of Music," Journal of Korea Multimedia Society, Vol. 16, No. 11, pp. 1325-1337, 2013.

[15] S.-C. Hur, S.-H. Cho, and J.-S. Sim, “Compressed pattern matching for dictionary-based compressed text,” Journal of Korean Institute of Next Generation Computing, Vol. 12, No. 1, pp. 67-74, 2016.

[16] H.-Y. Noh, S.-B. Choi, and H.-C. Ahn, “Social Network-based Hybrid Collaborative Filtering using Genetic Algorithms,” Journal of KIISS, Vol. 23, No. 2, pp. 19-38, 2017.