

IJACT 20-9-41

A Study on Correcting Korean Pronunciation Error of Foreign Learners by Using Supporting Vector Machine Algorithm

Kyungnam Jang¹, Kwang-Bock You² and Hyungwoo Park³

¹ Department Of Korean Language & Literature, Soongsil University, Korea

² Electronic Information Engineering IT Convergence, Soongsil University, Korea

³ Quantum Dot Display Development team, Samsung Display, Korea

E-mail pphw@ssu.ac.kr

Abstract

It has experienced how difficult People with foreign language learning, it is to pronounce a new language different from the native language. The goal of various foreigners who want to learn Korean is to speak Korean as well as their native language to communicate smoothly. However, each native language's vocal habits also appear in Korean pronunciation, which prevents accurate information transmission. In this paper, the pronunciation of Chinese learners was compared with that of Korean. For comparison, the fundamental frequency and its variation of the speech signal were examined and the spectrogram was analyzed. The Formant frequencies known as the resonant frequency of the vocal tract were calculated. Based on these characteristics parameters, the classifier of the Supporting Vector Machine was found to classify the pronunciation of Koreans and the pronunciation of Chinese learners. In particular, the linguistic proposition was scientifically proved by examining the Korean pronunciation of /ㄹ/ that the Chinese people were not good at pronouncing.

Keywords: Korean Learning, Machine Learning, Support Vector Machine, and Korean Pronunciation

1. INTRODUCTION

Globally, increasing interest in Korean culture and the Korean language is increasing the number of foreigners who want to learn Korean. In accordance with this trend, Korean language education for foreigners is being actively conducted abroad. In Korea, training is conducted in the form of semi-public education, such as language schools and Korean academics in universities. Outside of regular education, private tutoring is conducted, or students learn by themselves using uploaded data on social networks. About 8% of Korean international students are increasing each year, but the share of public education is only 25%, 34% for private education, 25% for self-taught, 7% for learning through media (movie, drama), and 2% for clubs. [1] The number of international students studying Korean is increasing from 90,000 in 2015 to 140,000 in 2018 and continues to increase. [2] In order to attract international students, it is considered to be reorganized into an adult-friendly undergraduate structure, or to increase the number of international students to 200,000 by creating demand through overseas expansion of domestic universities [1][2].

As Korean culture spread and the attracting of international students from Korean universities has increased, the number of foreigners learning Korean has increased. The goal of foreigners learning Korean is to have the ability to communicate smoothly in Korean. In addition, those who have learned foreign languages know that

Received: August 17, 2020 / Revised: September 09, 2020 / Accepted: September 17, 2020

Corresponding Author: [Hyungwoo Park](mailto:pphw@ssu.ac.kr) (Ph. D. Staff Engineer) pphw@ssu.ac.kr

Tel: +82-10-7371-0049

Quantum Dot Display Development team, Samsung Display

it is complicated to accurately reproduce the similar pronunciation of the language being studied as a habit of speaking in the native language [3]. Therefore, there should be a difference in Korean language education goals, context, and foreign language education methods. For example, for Koreans and foreigners, the meaning of speaking good Korean is entirely different. It is intended for foreigners to communicate fluently or to learn or acquire skills other than in Korea (technical, scientific, and vocational skills). Verbal communication is done through speaking, listening, writing and reading functions, and most communication is divided into speech and listening by speech-language and hearing ability. The study of voice began in the middle of the 20th century and has been developed in various ways. Through advancements such as machine learning and natural language processing, computers and humans can communicate and correct pronunciation or learn languages through this.

In the Korean language, the Altai family's theory is dominant, linguistically morphologically, it is a deadlock with many changes in language, and the relationship of grammar is not displayed in word order. Fourteen consonant sounds, ten vowel sounds, and 24 are used as standard. The method of analyzing Korean is the same as that of general voice analysis. Basically, voice is divided into excitation source and vocal tract parameters to analyze the occurrence characteristics and resonance characteristics to handle things such as speaker recognition, speaker identification, and voice recognition. In voice signal processing, the pitch is called the fundamental frequency and is used as a parameter to track changes in the speaker's characteristics or voice intonation. The formant is a transfer function of the vocal tract, which can be seen as the part where the sound produced by the vocal cords is resonated (resonant), and the characteristic of the pronunciation appears. In this study, the characteristics of Korean utterances of foreign speakers are divided into short sections and analyzed using their pitch, pitch change, formant characteristics, and formant change.

In the previous study, a study was conducted on the parameters that determine the accuracy of speech recognition and pronunciation for Korean speech. We compared the parameters of speech signals for pronunciation errors of Chinese Korean learners [3]. In addition, a study was conducted to extract the information of the voice through signal-to-noise ratio (SNR) estimation and enhancement, to make it possible to extract pitches and formants well, emphasizing periodicity for smooth and accurate pitch detection, and short section Previous studies were conducted to remove the pitch well in order to make accurate formant extraction [4]. And, through the long-term average spectrum (LTAS) analysis of the speech signal, the transmission characteristics were estimated by changing the parameters and vocal tract parameters for changes in phonological characteristics. We also studied [5]. To find a method for extracting the parameters from these voices, we intend to compare and analyze Korean and foreign Korean learners' correct speech characteristics to ensure correct pronunciation.

This paper consists of five sections. The introduction is presented in the section 1. In the following section, the speech production principles, analysis methods, and the SVM are described. The evaluation method for Korean pronunciation are proposed in the section 3. In the section 4, both the experiments and its results are discussed. Finally, the conclusion of this paper is stated.

2. Principles of Speech Production and Speech Analysis Methods

2.1 Speech Production Principles

The voice is pressured from the lungs of a person, and the air flows up through the airway, causing trembling of the vocal cords, and the resonance of the saints and articulation organs, spreading through the mouth and nose. As a representative example of methods for modeling the generation principle of the voice signal, there are methods in which the source-filter model and the vocal tract model are expressed in the form of a convolution sum according to time and signal flow. This model the quasi-periodic shape (pitch), size, and change that fluctuates from the gate to the fundamental frequency of vocal tremors in the form of excitation

sources. Next, the saint model can model several tubes of varying thickness in a connected form. When the vibration of air is transmitted through these tubes, the frequency spectrum is shaped according to the tube's resonance characteristics. Therefore, the reflection coefficient at the k th and $k+1$ th connections can be expressed by the following equation (1).

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \quad (1)$$

And the transfer function between these connections is expressed by the following equation (2).

$$V(z) = \frac{\frac{1}{2}(1+r_0) \prod_{k=1}^N (1+r_k) z^{-\frac{N}{2}}}{1 - \sum_{k=1}^N \alpha_k z^{-k}} \quad (2)$$

The area of the cross-section at the k -th connection is the reflection coefficient at the glottal and the prediction coefficient [6][7]. When analyzing the voice signal, if the model of the castle gate that produces tremors and the parameters of the saints to get characteristics through the sound are found and classified, the meaning of the sound and the degree of accurate pronunciation can be evaluated [6-10].

2.2 Speech Analysis Methods

Speech analysis performs the opposite process of the speech generation principle mentioned in 2.1. In general, when a person communicates through a voice in a block diagram, it appears as shown in Figure 1 below, human auditory organs recognize the tremor of the air through the inner ear, moreover the frequency change over time. And the features are transmitted to the brain by the auditory nerve. In the brain, information units grouped by phonemes and phonons are grouped into the structure of words and sentences to understand the meaning of the person speaking and the listening process completed. In this process, the human hearing organ receives pressure that changes over time, converts it into the frequency domain, judges the amount of energy in each frequency, detects the change in continuous time, and the relevant parameters have some meaning. It is possible to determine whether or not someone has spoken according to its characteristics. Judging by the key factors in speech generation, pitch, and formant are thus important parameters. In human hearing organs, the use of pitch and formant changes conveys meaning, speaker recognition, and emotions [6-10].

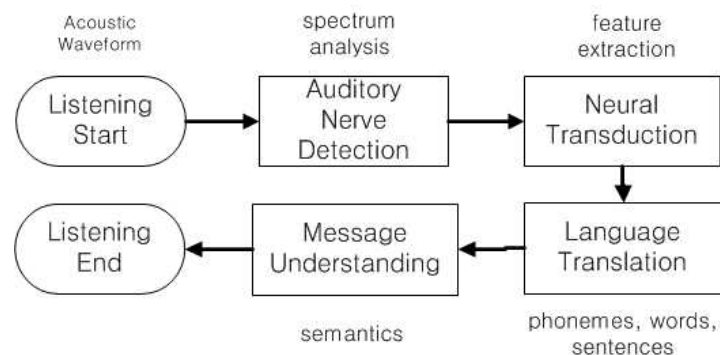


Fig 1. Block diagram of information transmission process via voice[6]

Pitch is the fundamental frequency spoken in the field of voice signal processing. Pitch refers to the period that the gates open and close, and expressed as tremors during a unit time, becomes the fundamental frequency. Pitch frequency and pitch period are parameters that vary depending on the air pressure and the amount from the lungs, and the elasticity, size, and size of the vocal cords vary depending on individual characteristics. When the pitch is calculated correctly, it helps to reduce the influence of the speaker's change in voice analysis. In addition, when synthesizing or analyzing speech, easy maintenance, naturalness, and human uniqueness can be changed. Each person has a pitch range for each person controlled by the larynx structure, usually 50-250 Hz for men and 150-300 Hz for women [6].

Formant refers to a frequency that resonates and amplifies when sound passes through the vocal cords. The shape is expressed by the position (frequency value), and the size (energy) of the bud in frequency analysis and the subscripts such as F1, F2, and F3 are expressed at a high level. The trembling of the air generated through the vocal cords may be a quasi-periodic similar sound column, but it resonates on the vocal track, causing the entire voice signal to change. In other words, a precise pronunciation is produced according to the frequency and the change of resonance (formant parameter, change) in the vocal track. When analyzing the voice, in general, the phoneme characteristics of the phoneme are determined by F1 and F2 in the vowel. And it appears that personal characteristics are expressed in F3, F4, and F5. However, F1 and F2 in the voiced section play an important role, but in the case of unvoiced, friction, and burst sounds, the formant frequency range is different from voiced, and phonological information and language information are expressed even in higher order formants [7-9].

2.3 Results of Previous Studies

In previous studies, parametric studies have been conducted to determine the accuracy of voice recognition and pronunciation. In [3], the characteristics of voice signals for Korean pronunciation errors of Chinese learners were analyzed. Differences between analysis characteristics and results were studied through comparison with normal pronunciation (Korean standard pronunciation). In the voice analysis, a study such as [3] was conducted to determine the change in pitch and intonation. Based on this, Korean learners who can speak Korean with standard pronunciation and intonation in their native language can generate it. The characteristics of the error were also studied. In addition, a study was conducted to extract parameters of voice analysis methods and techniques that can analyze elements of verbal and non-verbal information (emotion, health, etc.) possessed by voice.

The following is the result of analyzing Korean learners using native Chinese language and voices using standard Korean pronunciation. Data were recorded in a quiet room for each voice, and the signal was improved. It contains most of the audio signals, sampled at 16khz to reduce the size of the data, quantized to 16bit and digitized. After that, general pretreatment was performed. The analysis frame was analyzed by overlapping 25% in 30 ms increments. Figure 2 shows the results of waveform and energy contour spectrogram analysis of the same vocal section between Koreans and Chinese. It can be seen that the waveform, energy contour, and spectrogram have no similarity to each other. In addition, it is possible to determine whether the pronunciation is correct by using the extracted parameters [3][10].

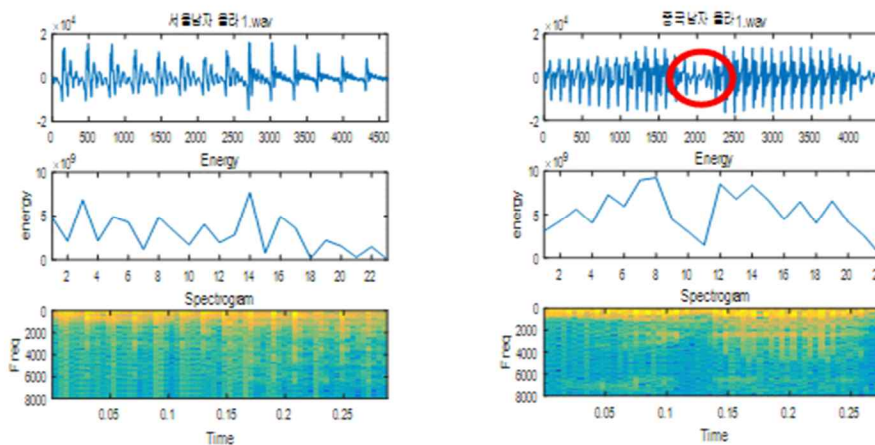


Fig 2. Korean Male (Left) and Chinese Male (Right): waveform in time (Top), energy (Middle), spectrogram (Bottom)

2.4 Support Vector Machine

SVM (Support Vector Machine) is one of the algorithms used as a discriminator in machine learning. SVM was developed based on statistical analysis, and the results through the algorithm are called dependent variables, and the factors affecting the results are called independent variables [11]. Since it is a discriminator based on statistical analysis, it is a method of predicting dependent variables using statistical similarity of independent variables in a large number of data. It is a method of analyzing various conditions and preparing judgment criteria based on statistically significant and small numbers [12]. In this SVM, the one that corresponds to the standard is called a hyperplane, and finding an independent variable that can be judged from a lot of data is called finding a hyperplane, which is not complicated [13]. SVM is machine learning that includes related learning to analyze data used for regression analysis, learning, and finding a hyperplane. SVM can distinguish what is in the dependent variable group through hyperplane. For example, it is an algorithm that distinguishes between ripe and unripe apples. SVM's classification algorithm is often used for machine learning due to its high accuracy, and its discriminator is mainly used when it is necessary to determine the correctness quickly. SVM is a kind of supervised learning field and learns the discriminator using data that knows the dependent variable, so the discriminator has high accuracy. In addition, sufficient standards can be made even with a small amount of learning data, and the time required for discrimination is short. In general, predictors are learning how to reduce errors, in which case they are overfitting and SVM is less over-fitting [12]. You can also improve the predictive performance by changing the dimensions of the data called kernel functions, which is easy to use if the characteristics of the data are well known [13]. In this study, the hyperplane was constructed by distinguishing the Korean pronunciation from the spoken voice generated by the standard Korean pronunciation. And the SVM can determine the audio sample input by this standard. The parameters obtained from the speech analysis mentioned in Section 2.1 were used as independent variables. The parameters used for judgment can be utilized by parameterizing the fundamental frequency, pitch, change in pitch, change in formant position, change in formant position, the slope of 1,2 formants, speech rate (speech rate), and energy change.

3. Proposed Evaluation Method of Korean Pronunciation

The proposed Korean pronunciation discrimination system is as follows. The voice of a student learning Korean is input through the microphone. Then, the noise quality of the surrounding environment is processed, and the voice quality is always maintained. Then, the parameter to be used as an independent variable to be

used for the SVM is extracted from the voice. In addition, additional parameters that can be used separately or utilized (elements that cannot be judged) are stored separately and classified as new independent variables. Then, the extracted parameters are input to the SVM discriminator to express the dependent results. Then, Korean learners can know whether their pronunciation is the same as the standard Korean pronunciation, listen to the standard pronunciation, and practice it repeatedly to make the correct pronunciation. The proposed method is shown in the block diagram in Figure 3. The voice enhancement was used to remove the normal background noise through the spectral-subtraction method. The recorded signal was sampled at 16khz considering the speed of processing and the amount of data, quantized to 16bit, and analyzed in a short section in 30ms increments and processed by a 25% overlap to extract parameters.

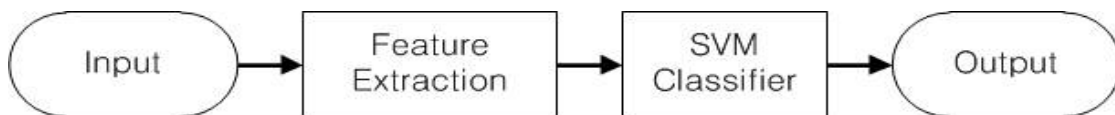


Fig 3. Block Diagram for the Proposed Evaluation Method

The formant used as the judgment parameter of SVM can utilize the point in the 2D plane using the primary and secondary formant frequency positions using the value of the section. The values shown in Table 1 are extracted from the standard Korean pronunciation of the same word and the spoken form of Chinese Korean learners. In the table, the positions of f1 and f2 are indicated in Hz, and changes can be observed for each frame. It is the result of all four speakers vocalizing the word 'Woo-Ri', but it is confirmed that the number of frames differs due to the difference in the vocal speed. The 'Woo-Ri' utterance is divided into two syllables, and it can be predicted that f1 and f2 each have two values after one transition. It is possible to extract and retain parameters for the change and rate of change of formants and normal standard vocalization by finding these features.

Table 1. The 1st and 2nd Formant frequencies of "우리" for SVM classifier: Korean Male (L-Top), Korean Female (L-Bottom), Chinese Male (R-Top) and Chinese Female (R-Bottom)

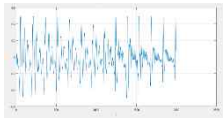
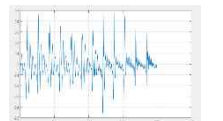
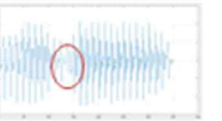
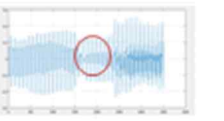
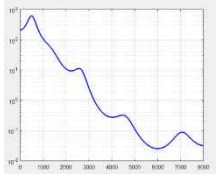
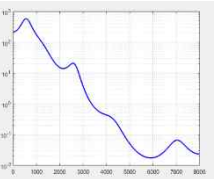
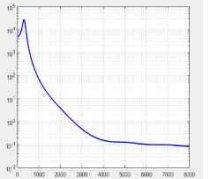
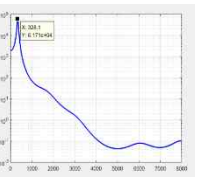
	Korean male 우리1.wav'	
#frame	f1	f2
1	218.75	4062.5
2	2015.625	5687.5
3	156.25	5546.875
4	171.875	5265.625
5	171.875	5453.125
6	187.5	1281.25
7	187.5	1515.625
8	187.5	1500
9	2500	5140.625

	Chinese male 우리1.wav'	
#frame	f1	f2
1	390.625	3421.875
2	375	3171.875
3	421.875	2687.5
4	281.25	2578.125
5	312.5	2515.625
6	312.5	2671.875

Table 1. The 1st and 2nd Formant frequencies of "우리" for SVM classifier: Korean Male (L-Top), Korean Female (L-Bottom), Chinese Male (R-Top) and Chinese Female (R-Bottom)

Korean female 우리1.wav'			Chinese female 우리1.wav'		
#frame	f1	f2	#frame	f1	f2
1	390.625	4656.25	1	375	687.5
2	421.875	4109.375	2	359.375	718.75
3	421.875	1078.125	3	343.75	3359.375
4	359.375	1390.625	4	312.5	1703.125
5	265.625	1703.125	5	359.375	2750
6	328.125	2812.5	6	390.625	2937.5
7	343.75	2765.625	7	359.375	3031.25

Table 2. Parameters of the speech signal of Korean speakers and Chinese learners

	한국남자1 (Korean Male1)	한국남자2 (Korean Male2)	중국남자 (Chinese Male)	중국여자1 (Chinese Female1)
waveform in time				
fundamental frequency	110.3448 [Hz]	103.2256 [Hz]	197.5309 [Hz]	124.031 [Hz]
LPC estimated spectrum				
first formant frequency	500 [Hz]	531 [Hz]	328 [Hz]	328 [Hz]

The parameters of the speech signals of Korean speakers and Chinese learners are summarized in Table 2. The hyperplane of SVM was found using pitch and formant frequency from this table.

4. Experiments and Results

In this study, SVM is used as machine learning to determine the speech accuracy of Korean learners. For discrimination, the learner's voice was recorded and pre-processed, and parameters were extracted from this data to compare hyper-planes to determine whether the correct speech was achieved. For this experiment, six Korean standard speech data and four Chinese Korean learner data were used. The age group for the sample data consists of students in their 20s.

Figure 4 shows the results of SVM's evaluation and hyperplane. Related studies in linguistics have described that /ㄹ/ is the significant pronunciation in which Chinese learners have quite a difficulty speaking

Korean. Thus, in this study, an experiment was conducted to compare the / \equiv /pronunciation of Korean. As for the data /우리/ and /올라/ were used in Table 1 and Table 2, respectively. Subsequently, the pitch and 1 and 2 formant frequencies were analyzed and extracted from the data as parameters of SVM.

In the red dot in the figure, the parameters for the Korean speaker's vocalization are expressed in a two-dimensional plane, and the blue dot is the parameter point for the speaker's vocalization.

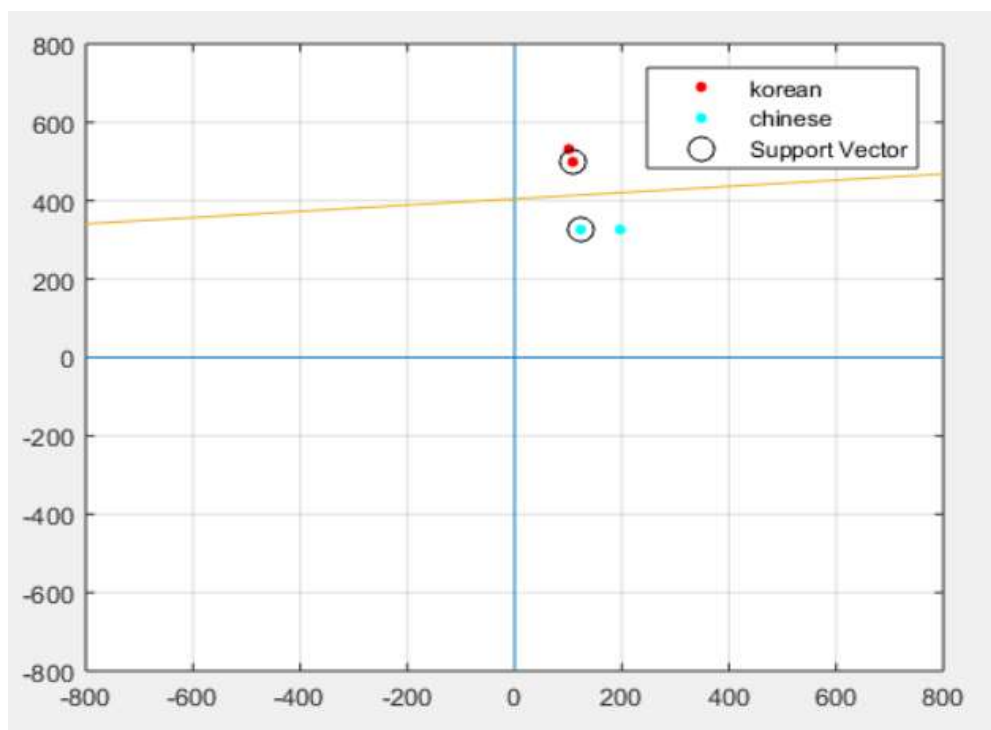


Fig 4. Results of Support Vector for the pronunciation of “ \equiv ” by Chinese learners and Korean speakers

5. Conclusion

In this study, the voices of Korean speakers and Chinese learners were recorded, processed, and analyzed. The correct and not-correct pronunciation was classified using an algorithm of SVM classification based on the parameters analyzed for speech data. From the resulting graph in Figure 4, it was confirmed that the support vector found and distinguished between Korean and Chinese pronunciations. It has been confirmed that the pronunciation correction (measurement) device, which can correct the pronunciation of Chinese learners, can be operated by the method proposed in this study. In the future, making a better Korean pronunciation discriminator will be conducting a study by increasing the training samples and data for judging incorrect vocalization in the SVM. In addition to the Chinese learners covered in this paper, the same analysis should be applied to Vietnamese learners and English learners to accumulate good data for Korean language learning. According to the Ministry of Education's announcement in 2016, about 60% of international students in Korea are Chinese, about 7% in Vietnam, and 3-4% in Japan, Mongolia, and the United States. Therefore, it is reasonable that the research in this paper began with a Chinese learner, but future research should be expanded to other cultures.

REFERENCE

- [1] Ministry of Education, plan to expand international student attraction (plan), Ministry of Education Education Development Cooperation Team, 2015.
- [2] Ministry of Education, Support Plan for University Innovation in Response to Demographic Changes and the Fourth Industrial Revolution, "Higher Education Policy Office, Ministry of Education, 2019.
- [3] Kang-Hee Lee, Kwang-Bock You, Ha-Young Lim, 'A Comparison Study on the Speech Signal Parameters for Chinese Learners' Korean Pronunciation Errors - Focused on Korean / \equiv /, '*Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol.7, No.6, pp. 239-246, (2017).
- [4] H. W. Park, A.R. Khil, and M.J. Bae, "Pitch Detection Based on Signal-to-Noise-Ratio Estimation and Compensation for Continuous Speech Signal," *International Conference on Hybrid Information Technology*. Springer, Berlin, Heidelberg, 2012
- [5] Sue Ellen Linville and Jennifer Rens, "Vocal tract resonance analysis of aging voice using long-term average spectra ," *Journal of Voice*, Vol.15, No.3, pp.323-330, 2001
- [6] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, Pearson, 2011
- [7] H.W. Park and S. Lee, "A study of reliability parameters extracted through voice analysis ," *The Journal of the Acoustical Society of America* Vol.140, No.4 2016
- [8] J.W. Park, H.W. Park, and S.M. Lee, "A Study on the Extraction of Credit Parameters through Voice Analysis of Telephone Counseling," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol.7, No.3, 2017.
- [9] H. W. Park, An Acoustic Analysis of Noise Environments during Mobile Device Usage, *International Journal of Advanced Smart Convergence* Vol.6 No.2 16-23 (2017).
- [10] S.W. Hahm, H. Park, A Interdisciplinary Study about Voice Change of the Presidential Candidate and Cognition Change of the Voters' , *The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)*. Vol. 18, No. 3, pp.193-200, Jun. 30, 2018.
- [11] Mathworks, eBook of Matlab and machine learning, 2017.<https://kr.mathworks.com/campaigns/products/offer/machine-learning-with-matlab.html>
- [12] Fung, Glenn M., and Olvi L. Mangasarian. "Multicategory proximal support vector machine classifiers." *Machine learning* Vol. 59 No.2, 2005, pp. 77-97.
- [13] Huang, Guang-Bin, et al. "Extreme learning machine for regression and multiclass classification." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* Vol. 42, No.2, 2012, pp. 513-529.