World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:3, No:8, 2009

Automatic Voice Classification System Based on Traditional Korean Medicine

Jaehwan Kang, and Haejung Lee

***Abstract*—**This paper introduces an automatic voice classification system for the diagnosis of individual constitution based on Sasang Constitutional Medicine (SCM) in Traditional Korean Medicine (TKM). For the developing of this algorithm, we used the voices of 309 female speakers and extracted a total of 134 speech features from the voice data consisting of 5 sustained vowels and one sentence. The classification system, based on a rule-based algorithm that is derived from a non parametric statistical method, presents 3 types of decisions: reserved, positive and negative decisions. In conclusion, 71.5% of the voice data were diagnosed by this system, of which 47.7% were correct positive decisions and 69.7% were correct negative decisions.

***Keywords*—**Voice Classifier, Sasang Constitution Medicine, Traditional Korean Medicine, SCM, TKM.

I. Introduction

FOUR-Constitution Medicine, also called Sasang Constitutional Medicine (SCM) [1], is a branch of Traditional Korean Medicine (TKM) initiated by Lee Je-ma, who stressed the theory of the four constitutions in healing human diseases: physiology, pathology, diagnosis, and maintenance of health. In SCM theory, a person can be divided into one of four constitutional types: Tai-yang (greater yang), So-yang (lesser yang), Tai-eum (greater yin), So-eum (lesser yin) person and different oriental medical treatments should be applied according to patient's constitution. The classification of Sasang Constitution is based on four diagnostic processes - inspection, auscultation, inquiry and palpation. The auscultation process involves collecting information for diagnosis by listening to the voice of the patient. All types of constitutional persons have a different voice characteristic as follows. Tae-yang persons have a powerful, loud and clear ringing voice. So-yang persons have a high pitched, plain voice and are apt to be talkative. Tae-eum persons have a low pitched, powerful but vague voice. So-eum persons have a weak, powerless, light and clear voice. In the field of SCM, since Tai-yang persons are just only 0.1% of the total population, the study of SCM classifier mostly focuses on the classification of 3 constitutional types. The goal of this work is to develop a speech analysis algorithm that can classify three constitutional patient types, excluding the Tai-Yang type persons, and to provide a comprehensive, integrated constitutional algorithm including appearance, skin condition and palpation, thereby providing a quantitative result of speech analysis and improving the system's precision.

Jaehwan Kang is with the Korean Institute of Oriental Medicine, Daejeon, Korea (corresponding author to provide phone: 82-42-868-9301; fax: 82-42-868-9480; e-mail: doskian@kiom.re.kr)

Haejung Lee is with the Korean Institute of Oriental Medicine, Daejeon, Korea (e-mail: haejung0614@kiom.re.kr)

This work was supported by the Korea Ministry of Knowledge Economy (10028438).

II. Material and Methods

A. Speakers and Recording

To develop the algorithm, we used 309 female speakers whose constitutional types were confirmed by a TKM doctor experienced in the area of SCM at least for 5 years. Speakers were composed of 116 So-yang persons, 111 Tae-eum persons and 82 So-eum persons. The speakers were aged between 10 and 75 (48.2±16.1). None of the speakers had any health or voice problem. Voice data were collected from the same room under the same quietness condition (sound-level between 20 – 30 dB) using the same microphone (Sennheiser e-835s) with a fixed microphone stand and the same PC environment (Sound Card: Sound Blaster Live External, Record Program: Goldwave 5.08). The microphone was held at 4-5cm from the mouth and a constant recording sound level was used. Each speaker was asked to say a series of sustained vowels (A, E, I, O, U, in the said order) and one simple sentence twice using her natural voice at a constant rate. All sustained vowels lasted about 1.5 seconds. The sampling rate was 44,100 Hz and 16 bit per samples.

B. Speech Features

First, we calculated 25 speech features consisting of 12 parameters related to pitch and formant groups and 13 MFCC (Mel Frequency Cepstral Coefficients) parameters in 5 vowel speech frame. Second, we also extracted 9 speech features which represent pitch and intensity variations in sentence speech frame. The speech features which are related to pitch and formant were calculated by our speech parameter extraction program (FvSND). The pitch extraction method in FvSND is based on the Boersma method [2] which is used for the autocorrelation and interpolation function to improve precision rate and performance. In addition, we calculated MDVP (Multi-Dimensional Voice Program) parameters like Jitter and Shimmer parameters, which are derived from the extracted pitch using equations in Table I. Because MDVP parameters represent the vocal health condition well, those parameters have become useful in the clinic part of vocal disorder diagnosis recently.

TABLE I
EQUATIONS OF MDVP

STD	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(F_i-\bar{F})^2}$		
Jita	$\frac{1}{N-1}\sum_{i=1}^{N-1}\lvert T_{0i-1}-T_{0i}\rvert$	ShdB	$\frac{1}{N-1}\sum_{i=1}^{N-1}\lvert 20\log(A_{i+1}/A_i)\rvert$
RAP	$\frac{\frac{1}{N-2}\sum_{i=1}^{N-2}\lvert T_{0i-1}+T_{0i}+T_{0i-1}-T_{0i}\rvert}{\frac{1}{N}\sum_{i=0}^{N-1}T_{0i}}\times 100$	Shim	$\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\lvert A_i-A_{i+1}\rvert}{\frac{1}{N}\sum_{i=0}^{N-1}A_i}\times 100$
PPQ	$\frac{\frac{1}{N-4}\sum_{i=2}^{N-2}\lvert T_{0i-2}+T_{0i-1}+T_{0i}+T_{0i+1}+T_{0i+2}-T_{0i}\rvert}{\frac{1}{N}\sum_{i=0}^{N-1}T_{0i}}\times 100$	APQ	$\frac{\frac{1}{N-4}\sum_{i=2}^{N-2}\lvert A_{i-2}+A_{i-1}+A_i+A_{i+1}+A_{i+2}-A_i\rvert}{\frac{1}{N}\sum_{i=1}^{N}A_i}\times 100$
sPPQ	$\frac{\frac{1}{N-54}\sum_{i=27}^{N-28}\lvert\sum_{k=i-27}^{i+27}\frac{T_{0k}}{55}-T_{0i}\rvert}{\frac{1}{N}\sum_{i=0}^{N-1}T_{0i}}\times 100$	sAPQ	$\frac{\frac{1}{N-54}\sum_{i=27}^{N-28}\lvert\sum_{k=i-27}^{i+27}\frac{A_k}{55}-A_i\rvert}{\frac{1}{N}\sum_{i=1}^{N}A_i}\times 100$

For the extraction of formant features (F1, F2), we used a 12-LPC (Linear Prediction Coding) model and a root extraction method, which is based on peak finding in an LP-derived magnitude spectrum [3]. MFCC parameters were calculated by HTK (Hidden Markov Model Toolkit) ver. 3.2 [4] and we could get one energy and 12 MFCC features. In the sentence analysis, we first extracted pitch and intensity array in the sentence speech frame by PRAAT [5] and calculated 90 percentile, 50 percentile and 10 percentile of those parameters respectively. The ratio and correlation of pitch and intensity were also calculated by the pre-calculated percentile parameters continuously. Consequently, we extracted 134 speech features from a total of 5 vowels and one sentence as shown in Fig. 1.

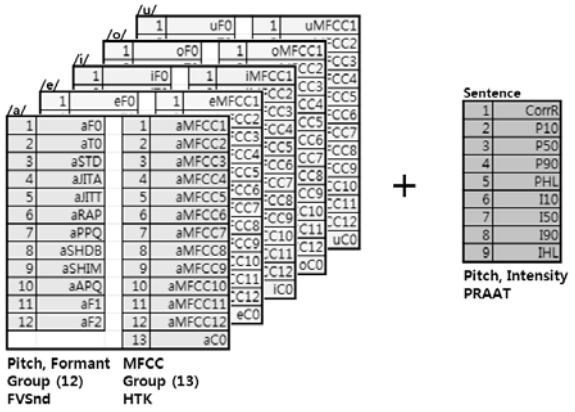

Fig. 1 Speech Features

C. Two Logical Rules in Each Speech Feature

For the purpose of constitutional classification, we developed a rule-based algorithm by the nonparametric statistical method. First, by examining all of 134 speech features derived from the voice data of 309 females thoroughly, we could find the maximum and minimum within each constitution group. In the so-obtained maxima and minima groups, *uMin* is defined as the minimum among the maxima and *lMax* as the maximum among the minima in each voice speech feature. In addition, *ui* is defined as the index pointing out the constitution group containing *uMin*, and likewise, *li* as the index of the constitution group containing *lMax*. After finding all 4 condition parameters (*uMin*, *ui*, *lMax*, *li*) in each speech feature, a constitutional score can be generated using those features as follows. First, we set the *uMin* on an upper threshold value and set the *lMax* on a lower threshold value. If there is an unknown input speech signal found above the upper threshold (*uMin*) in a specified speech feature, it is said to have a higher possibility that is not *ui*. Finally, it is clear that we can derive two logical rules in one speech feature as follows.

If (X > *uMin*), then NOT *ui*
If (X < *lMax*), then NOT *li*

For the efficient management of 4 condition parameters in all speech features, we allocated 4 by N matrix (called *gRule*) for 4 condition parameters. N refers to the number of all voice parameters.

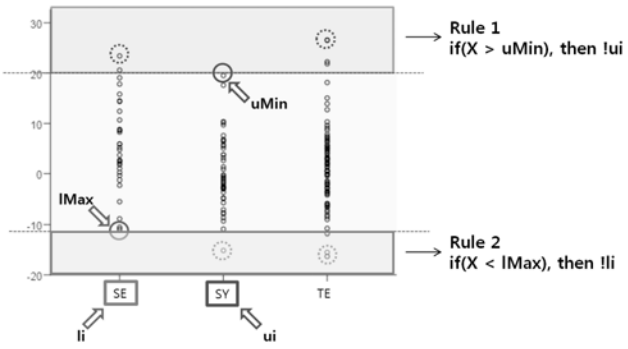

Fig. 2 Basic idea of the algorithm

D. Correlation between Voice Parameters and Age, BMI

There are some speech features which are affected by the variation of speaker's personal information such as ages, height, weight and BMI. Therefore, we scrutinized the correlation between all speech features and personal information thoroughly and found that there were 21 cases affected by age and 5 cases affected by BMI whose correlation were above 0.3 in all speech features. There is no correlation above 0.3 among other speech features such as height and weight. Those speech features which are affected by ages and BMI should be considered in designing the algorithm. For this purpose, we used the concept of sub-matrix of *gRule* and generated this as follows. First, we divided the all of the voice data into 5 groups according to the age of speakers, with each age group consecutively covering a span of 16 years, an age span considered meaningful in TKM. For those speech features affected by age, we generated a sub-matrix of *gRule* for each age group and replaced a suitable sub-matrix with *gRule* by ages. Also, all groups were divided into 4 groups in BMI and the sub-matrix was generated by the same method for different age groups.

TABLE II
CORRELATION BETWEEN SPEECH FEATURES AND AGE, HEIGHT, WEIGHT AND BMI

Feature	Age	Height	Weight	BMI
aF0	-0.394*	0.078	-0.213	-0.274
aT0	0.317*	-0.111	0.139	0.211
eF0	-0.400*	0.020	-0.263	-0.300*
eT0	0.396*	-0.057	0.240	0.293
iF0	-0.326*	0.008	-0.203	-0.232
oF0	-0.356*	0.050	-0.184	-0.228
oT0	0.327*	-0.111	0.121	0.191
uF0	-0.328*	0.033	-0.193	-0.230
P10	-0.423*	-0.021	-0.276	-0.301*
P50	-0.419*	0.024	-0.273	-0.319*
P90	-0.514*	0.133	-0.250	-0.345*
aMFCC9	0.407*	-0.079	0.202	0.250
aMFCC12	0.208	0.005	0.293	0.304*
eMFCC6	0.476*	-0.118	0.195	0.266
eMFCC11	0.425*	-0.139	0.148	0.238
eMFCC12	0.399*	-0.192	0.156	0.250
iMFCC6	0.438*	-0.074	0.182	0.230
iMFCC9	0.374*	0.000	0.200	0.209
oMFCC9	0.494*	-0.161	0.175	0.270
uMFCC7	-0.319*	0.058	-0.094	-0.138
uMFCC9	0.368*	-0.051	0.163	0.199
uMFCC12	0.322*	-0.189	0.164	0.256

* correlation above 0.3

Fig. 3 shows an example of sub-matrix in one of the speech features which is affected by age. After generating the sub-matrix and *gRule* matrix completely, we make a constitutional score for an input speech by comparing with the appropriate logical rule of *gRule* matrix and the value of speech feature. Our constitutional classification algorithm decreases 1 point whenever each of the 134 speech features, or a voice data, enters as an input parameter and meets the IF-statement of *gRule* matrix.

gRule	#1 (aF0)	#2 (aT0)	...	#63 (p50)	...	#134 (uC0)
uMin	215.79	7.91	...	191.81	...	81.55
ui	2(SY)	3(TE)	...	2(SY)	...	3(TE)
lMax	126.36	4.63	...	143.33	...	0
li	3(TE)	2(SY)	...	1(SE)	...	1(SE)

#63 gRule_Age	Age1	Age2	Age3	Age4	Age5
uMin	241.59	214.17	227.40	233.50	191.81
ui	2(SY)	3(TE)	1(SE)	2(SY)	2(SY)
lMax	254.54	194.85	143.33	182.39	162.26
li	1(SE)	2(SY)	1(SE)	1(SE)	1(SE)

Fig. 3 Example of *gRule* Matrix and the sub-matrix

Finally, every voice data is designated a final constitutional score. An example of generating the constitution score for input speech data is shown in Fig. 4.

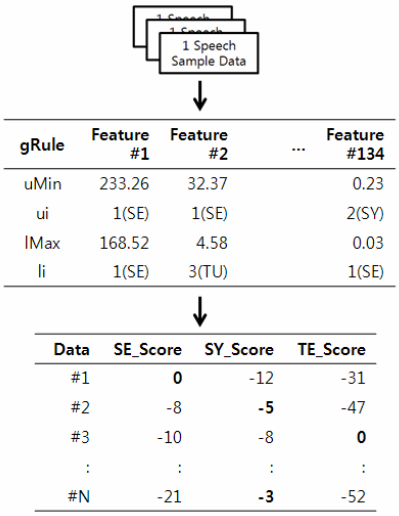

Fig. 4 Example of generating the constitutional score

E. Rule of Decision

A constitutional voice classifier in this paper finally diagnoses the input voice data as one of three decisions: reserved, positive (specific constitution) and negative decision (not specific constitution). Fig. 5 shows a detailed rule of decision flowchart. We can have a valid constitution result only if the difference between the max and min values of three constitutional scores is higher than 10%. If condition is not satisfied, we do not make our decision. In the first case, a positive decision (specific constitution) is possible when the difference between the max and median is greater than the difference between the median and the min values. If the difference between the median and min is greater than the difference between the max and median values, a negative decision (not specific constitution) is the answer.

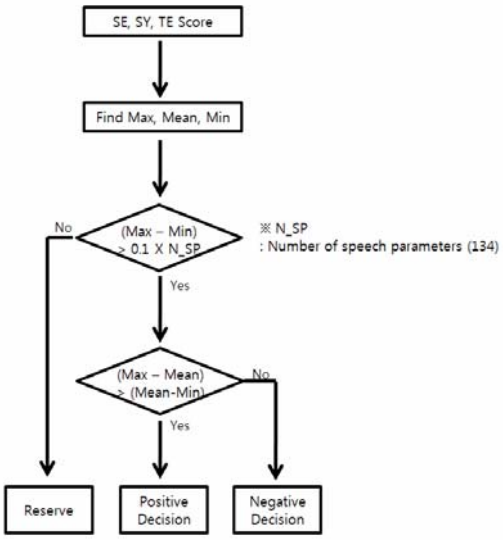

Fig. 5 The 3 decision types of the algorithm

III. Results

In order to test the performance of the proposed algorithm in this work, we used the voice data of 309 woman speakers. We grouped the total data set into two. One group consisted of 80% of total data (248 samples) and was considered to be a train set which was used to generate *gRule* matrix. The other group consisted of the remaining 20% of the total data (61 samples) and was considered to be the test set and used for testing the trained system. Each data set was chosen randomly. Once the classification result was obtained, the total data set was shuffled and chosen again in the same way. These experiments were performed 20 times and the results are listed in Table III. The result was that the possibility of arriving at a valid constitutional decision is about 71.5%, of which the positive rate constituted 22% and negative decision rate 49.5%. In the analysis of positive decisions, about 41% were proven as correct decisions when compared against the knowledge of experts. For negative decisions (not specific constitution), we had the correctness of 70%, roughly.

TABLE III
Assessment of the Algorithm

	Positive Decision	True Positive	Negative Decision	True Negative
1	0.213	0.538	0.443	0.630
2	0.164	0.400	0.590	0.750
3	0.246	0.467	0.410	0.800
4	0.246	0.333	0.508	0.613
5	0.230	0.500	0.475	0.552
6	0.311	0.421	0.443	0.778
7	0.164	0.400	0.525	0.719
8	0.197	0.583	0.525	0.688
9	0.131	0.250	0.590	0.722
10	0.246	0.400	0.475	0.690
11	0.197	0.500	0.492	0.767
12	0.180	0.636	0.492	0.700
13	0.213	0.462	0.508	0.645
14	0.311	0.474	0.459	0.786
15	0.197	0.333	0.541	0.727
16	0.180	0.455	0.557	0.706
17	0.197	0.833	0.393	0.750
18	0.279	0.529	0.459	0.750
19	0.213	0.385	0.525	0.594
20	0.213	0.615	0.492	0.600
Total	0.220	0.477	0.495	0.697

IV. Conclusion

In this paper, we classified human voices in terms of Sasang constitutions, a well-known theory in TKM. Especially, we aimed to make a proper voice classification algorithm that could integrating all aspects of Sasang constitution since Sasang constitution can be judged by a variety of factors, not only voice but also other factors like appearance, body shape and skin. The experimental results of our algorithm show that the correctness of positive decisions is about 47.7%; this means that specific constitution and the correctness of negative decision is about 69.7%, indicating that it is not a specific constitution.

Although the correctness of the constitutional decision by considering voice alone turned out a bit low, if the confidence of results could be added in the future, we believe our algorithm can contribute considerably to creating an integrated constitutional system. Our future work will investigate more significant speech features for classifying SCM and integrate this algorithm into a comprehensive SCM classifier that includes not only speech data but also appearance, skin condition and palpation.

References

[1] WHO, "WHO International Standard Terminologies on Traditional Medicine in The Western Pacific Region," Available: http://www.who.int
[2] Boersma, P, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.", *Proceeding Institute of Phonetic Sciences* 17, 1993, pp. 97-110.
[3] B. S. Atal, L. S. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave.", *J. Acout. Soc. Am.*, 1971, pp. 637-655
[4] (Program) HMM Tool Kit http://htk.eng.cam.ac.uk/
[5] Boersma, P, "Praat: doing phonetics by computer (Version 4.3.14) Retrieved May 26, 2005, from http://www.praat.org.

Jaehwan Kang received the B.S. in electronic engineering from Korea University, Yeongi, Korea in 1999, the M.S. degree in biomedical engineering from Chungbuk National University, Cheongju, Korea in 2001. From 2002 to 2006, he worked on biomedical signal processing and U-Health care system for Electronics and Telecommunications Research Institute (ETRI) in Korea. He is currently a researcher in Korean Institute of Oriental Medicine (KIOM), Daejeon, Korea. His research interests include speech signal processing and U-Healthcare system for biomedical processing.

Haejung Lee received the B.S. in statistics from Chungnam National University, Daejeon, Korea in 2005, the M.S. degree in statistics from Chungnam National University, Daejeon, Korea in 2007. Since 2005, she has been working for Korean Institute of Oriental Medicine (KIOM) in Korea. Her research interests include speech signal processing and U-Healthcare system for biomedical processing.