Recent Progress in Voice-Based Sasang Constitutional Medicine: Improving Stability of Diagnosis

Jun-Su Jang, Young-Su Kim, Boncho Ku, and Jong Yeol Kim
KM Health Technology Research Group, Medical Research Division, Korea Institute of Oriental Medicine, 1672 Yuseongdae-ro, Yuseong-gu, Daejeon 305-811, Republic of Korea

Correspondence should be addressed to Jong Yeol Kim; ssmed@kiom.re.kr

Received 2 May 2013; Accepted 19 July 2013

Copyright © 2013 Jun-Su Jang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sasang constitutional medicine is a unique form of tailored medicine in traditional Korean medicine. Voice features have been regarded as an important cue to diagnose Sasang constitution types. Many studies tried to extract quantitative voice features and standardize diagnosis methods; however, they had flaws, such as unstable voice features which vary a lot for the same individual, limited data collected from only few sites, and low diagnosis accuracy. In this paper, we propose a stable diagnosis model that has a good repeatability for the same individual. None of the past studies evaluated the repeatability of their diagnosis models. Although many previous studies used voice features calculated by averaging feature values from all valid frames in monotonic utterance like vowels, we analyse every single feature value from each frame of a sentence voice signal. Gaussian mixture model is employed to deal with a lot of voice features from each frame. Total 15 Gaussian models are used to represent voice characteristics for each constitution. To evaluate repeatability of the proposed diagnosis model, we introduce a test dataset consisting of 10 individuals’ voice recordings with 50 recordings per each individual. Our result shows that the proposed method has better repeatability than the previous study which used averaged features from vowels and the sentence.

1. Introduction

Sasang constitutional medicine (SCM) is a unique form of tailored medicine in traditional Korean medicine. The origin of SCM was started by Lee, a Korean medical doctor [1]. In SCM theory, human beings can be categorized into one of the four Sasang constitutional (SC) types (Tae-Yang: TY, Tae-Eum: TE, So-Yang: SY, and So-Eum: SE) according to their inherited characteristics, such as temperament profile, physiological and pathological features, susceptibility to particular diseases, and responses to drugs [2, 3]. In the principle of SCM, treatment is to recover balance between hypoactive organs and hyperactive organs, based on one’s constitution. Therefore, determining one’s SC type is important and many attempts have been made for standardization of SC diagnosis [4–6].

Diagnosis using voice is one of the important parts in SCM. The relation between SC types and voice is referred in the literatures [1, 7]. TY type’s voice is resonant, clear, and loud. It was derived from good respiratory organs. TE types have a load baritone voice, which sounds thick, heavy, and grave. SY types have clear, fast, and high-pitched voice. They are talkative, hasty, and vigorous. SE type’s voice is clam and easy. It sounds gentle, slow, and lively. Some of the voice characteristics are described by using utterance quality and style, but personality term is also used to depict voice characteristics. Since the vague linguistic representation can be understood differently to each oriental medicine doctor, an objective and standardized voice diagnosis is needed.

To establish objective and scientific diagnosis of SC types, many studies used computerized voice analysis method. Early studies focused on finding correlation between SC types and various voice features such as fundamental frequency, formants, and energy of voice signal [8–10]. In 2004, Park and Kim found a significant difference between SE and SY in formant frequency and formant bandwidth [11]. Their result showed that SY voice was clearer than SE voice, which was coincident with the SCM literatures. Kim et al. developed a voice analysis system, namely, phonetic system for Sasang constitution (PSSC), and applied it to investigate...
constitutional characteristics of Korean adult males and females [12, 13]. They used pitch, amplitude perturbation quotient, shimmer, octave, and energy as voice features. Choi et al. studied characteristics of the Korean adult male sound using PSSC with a sentence [14]. In 2009, Kang et al. analysed 144 voice features from 5 vowels and one sentence recorded from 473 people [15]. They developed a constitutional classification method using support vector machine; however, their classifier was able to apply correctly in only limited voice data [16].

Although many studies existed, they were not sufficient for practical use, because the diagnosis model was made by using a limited number of data gathered in few sites. To overcome the generalization problem, a set of large data, collected from 23 different oriental clinics, were established in the study of Do et al. [4]. They developed an integrative classification tool (SCAT) using four different diagnosis components: face, body shape, voice, and questionnaire. They trained four different diagnosis models using logistic regression method, and the four models were combined into a final integrated model. They evaluated the integrated model to a test dataset which was not used in training stage, for confirming generalization ability. As a generalization test result, the diagnostic accuracies of the integrated model were 64.0% and 55.2% in the male and female patient groups, respectively. The accuracies of the voice component were 39.9% and 37.5% for male and female, respectively. Although the accuracies of voice component were low, voice information was still helpful to increase the performance of the integrated model. Another research using the same voice data of Do et al. was carried out by Kim et al. [5]. They employed linear discriminant analysis as a classifier, and the training accuracies were 51% and 47% for male and female, respectively. However, they only provided the training result, which is generally higher than generalization test result.

Most of the previous studies were tried to find voice features that were strongly correlated to SC types and developed SC classification model. A few weakly correlated features were found, and different types of classifier were developed; however, none of those studies mentioned the diagnostic stability of their method. Because of the natural variation in speech, diagnosis results can be different for the same individual speaker. To reduce a lot of unsystematic variation, speaker's utterance should be constrained by a strict standard operating procedure (SOP). Kim et al. studied about developing SOP for extracting stable voice features that can characterize individual's voice quality consistently [17] and also analysed stable voice features [18].

In this paper, we propose a method to improve stability of voice-based SC diagnosis. All of large database, strict SOP, stable voice features, and a robust classification method are required to obtain stability of diagnosis. We use the same voice data as in the previous study [4], since they are known as the largest data containing patients' SC type proved by herbal remedy [19]. The proposed method uses only sentence recordings, against many previous studies using vowels or both vowels and sentences. Most of recent speaker identification studies use words or sentences, rather than vowels, since words or sentences are better to distinguish each individual's voice characteristics [20, 21]. Our results also show that the detailed analysis using a sentence recording has better repeatability in SC classification than the previous approach [4], which used vowels and the sentence.

2. Materials and Methods

2.1. Voice Data for Sasang Constitution Diagnosis. Voice data were collected from 23 oriental medical clinics. The patient whose SC type was diagnosed to TY was excluded due to its small sample size compared with other three SC types. The total number of patients used in this study was 1,969, ranging in age 15–60 years in both genders. The patients did not suffer from any voice-related disease so that they could speak naturally with their own voice quality. Their SC type was examined by SCM practitioners, who had more than five years of experience in clinical practice. A more detailed procedure of data collection was described in Song et al. [19]. Original database included face, body shape, and questionnaire information; however, only voice data were considered in this study.

Recording environment and procedure was strictly controlled by an SOP. Environment noise kept below 40 dB for low-noise recording. Recording room temperature was controlled to 20°C ± 5°C, and humidity kept to 40% ± 5%. Sound Blaster Live 24 bit External Soundcard and Sennheiser e-835s Microphone with a microphone stand were equipped. The distance between the patient and microphone was 4–6 cm. Recording systems were controlled to avoid echoes of voice or irregular resonance. Recording was saved as a wav file with setting of mono 16 bit integer and 44.1 kHz sampling frequency.

Voice data consisted of five vowels (/a/, /e/, /i/, /o/, and /u/) and a sentence. The sentence was recorded twice. Since all the patients were Korean, the sentence was also composed of Korean words. As voice features should represent the natural characteristics of patients in short recordings, the patients were asked to pronounce their natural voice without tension as possible. Before actual recording, an operator instructed the patients about the recording contents and allowed them to rest for 1 hour. Each vowel was uttered at least 3 s. The sentence was uttered in their ordinary speed and tones. In this paper, we used only sentence part to diagnose constitution, and for comparison, the previous study [4] used both five vowels and sentence.

2.2. Voice Feature Extraction. Figure 1 shows how voice features are calculated from a voice signal. We should define the size of a frame window, which is the minimum length to process. Then, we can extract voice features in the frame. After processing the current frame window, the same feature extraction is applied to the next frame window. In this study, the size of a frame window was 46.4 ms, which mapped to 2^{11} audio samples in 44.1 kHz sampling frequency. Neighboring frame windows were overlapped by 50%.

There exist more than a hundred valid frames in each vowel or sentence. In the past studies, there were many candidate voice features, such as fundamental frequency
Most of the features were calculated by averaging the feature values obtained from the all valid frames in the voice signal. An averaged feature could be good for monotonic utterance like a vowel. For example, $F_0$ for a vowel is calculated by averaging all $F_0$ values from the valid frames. The vowel utterance is monotonic, which means that every single $F_0$ value from each frame should be similar. Therefore, averaging is appropriate for characterizing $F_0$ for a vowel. Most of the previously used features from a vowel are averaged features, which may represent quality of the vowel properly.

However, features from the vowel are not proper to characterize individual's stable voice quality. It can be easily verified that words or sentences are more suitable for characterizing individual's stable voice quality in the most of recent speaker identification application [20, 21]. To make a stable diagnosis model, which has a good repeatability for the same individual, we focus on the sentence rather than vowel. In contrast to the features for vowel, averaged features are hardly used for the sentence since the features vary a lot in many frames. When averaging, we may lose useful information of individual's voice characteristics. Therefore, we analyse the features from every single frame.

Let $y_i$ be the feature vector extracted from the $i$th valid frame. Then, we define the feature vector as follows:

$$y_i = \left[ \frac{x_i}{t_{s_i}} \right],$$  \hspace{1cm} (1)

where $t_{s_i}$ indicates the relative position of the $i$th frame in all valid frames, and it has a range between 0 and 1. When the feature vector comes from the first frame of the sentence, the value of $t_{s_i}$ is 0. The vector $x_i$ is a column vector of candidate voice features. We use Mel-frequency cepstral coefficients (MFCCs) as the candidate voice features. MFCCs are coefficients of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency [22]. The mel scale approximates the human auditory system's response more closely than the linearly spaced frequency bands. MFCCs are widely used in speech and speaker recognition systems [23]. Total 12 MFCCs were used in our study.

Finally, feature vector $y_i$ is a 13-dimensional vector, and the total number of $y_i$ is the same as the number of all valid frames, which is usually more than a hundred for a sentence. By defining the feature vector including relative position of the frame, each feature vector contains not only voice characteristics of a frame signal, but also the information where the voice features come from. The feature vector was extracted using C++ program combined with HTK [24].

Since the voice features showed nonlinear fluctuation according to the age in general, a process to reduce the effect of age was required. To eliminate the age effect to the voice features, a standardization process was performed the same as Do et al. [4]. All MFCCs were standardized by using their moving averages and standard deviations derived from the data within the length of age ±5 years for the specific age.

2.3. Gaussian Mixture Model-Based Classification. The problem of SC diagnosis is treated as three-class classification, since TY type is usually excluded due to its rareness. Many statistical pattern recognition methods were applied to SC classification, such as logistic regression [4], linear discriminant analysis [5], and support vector machine [16]. Although many attempts were made in the past, none of them analysed the features from each frame separately. We propose a detailed analysis method to deal with the features of each frame using Gaussian mixture model (GMM). It is widely used for speech analysis area, since it can effectively model voice characteristics. GMM is denoted as

$$p(y | \lambda) = \sum_{j=1}^{M} w_j \cdot g(y | \mu_j, \Sigma_j),$$  \hspace{1cm} (2)

where $\lambda$ is one of the SC classes, $M$ is the total number of Gaussian models, $w_j$ is the weight for the $j$th Gaussian model, and $g(y | \mu_j, \Sigma_j)$ is Gaussian function with mean $\mu_j$ and covariance matrix $\Sigma_j$.

The basic idea of using GMM for SC classification is that we put each Gaussian model along the time axis to cover each part of the voice signal. Since we have to deal with a few feature vectors from all valid frames, it is important to use the relative position of each feature vector. In other words, someone's voice features extracted from a certain part of the sentence should be compared to a model that is generated by voice features extracted from similar parts. The feature vector $y_i$ has 12-dimensional MFCCs and relative position information as the 13th element. Hence, each Gaussian model in GMM is 13-dimensional, and we can display the location of each Gaussian model along the time axis by using the 13th element. Figure 2 shows that GMM covers each part of the voice signal.

Since GMM is needed for each gender and SC type, total 6 groups of feature vectors extracted from training data are
prepared for making 6 GMMs. GMM is trained for each feature vector group using EM algorithm [25]. In training process, the feature vectors need to be assigned to one of the Gaussian models as an initial assignment. The initial assignment can be done by uniformly dividing the feature vectors using the \( t_s^i \) value. For example, when the number of Gaussian model is 15 (\( M = 15 \)), the feature vectors are extracted from the first 1/15 part of the sentence assigned to the first Gaussian model. This is just for initial assignment, and the assignment is automatically changed in learning process.

After GMM for each SC class is trained, we follow the basic maximum likelihood-based classification. Probability that input recording falls into class of \( \lambda \) is calculated as follows:

\[
\bar{p}(y | \lambda) = \prod_{i=1}^{N} p(y_i | \lambda).
\]  

(3)

The total number of feature vectors is \( N \), and the \( p(y_i | \lambda) \) is calculated using (2) for each feature vector. Then, the estimated probability for being in each SC type \( \pi_k \) can be denoted as

\[
\pi_k = \frac{\bar{p}(y | \lambda_k)}{\sum_{i=1}^{3} \bar{p}(y | \lambda_i)}. \tag{4}
\]

The index \( k \in \{1, 2, 3\} \) indicates SC types, TE, SE, and SY, respectively. Finally, let \( h \in \{1, 2, 3\} \) be the predicted SC types; then the classification rule for SC types using \( \pi_k \) is given by

\[
h = \arg\max_{k \in \{1, 2, 3\}} (\pi_k). \tag{5}
\]

This means that simply taking the class of maximum probability is the final decision.

### 3. Experimental Results

#### 3.1. GMM-Based Classifier

Total 1,969 samples (1,263 females and 706 males) were used to train GMMs. Since the voice characteristics were different according to gender, GMMs were separately trained for each gender. Total 6 GMMs, representing 3 types of SC for each gender, were obtained. Each model had 15 Gaussians to cover the voice signal of a sentence. Table 1 shows the location of each Gaussian in time axis by summarizing the mean value of \( t_s^j \). The value of \( t_s^j \) indicates the relative position of the \( j \)th Gaussian model.

**Table 1: Mean values of \( t_s^j \) in each Gaussian model.**

<table>
<thead>
<tr>
<th>No. of Gaussian</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TE SE SY</td>
<td>TE SE SY</td>
</tr>
<tr>
<td>1</td>
<td>0.050 0.046 0.035</td>
<td>0.062 0.064 0.065</td>
</tr>
<tr>
<td>2</td>
<td>0.133 0.149 0.127</td>
<td>0.116 0.109 0.109</td>
</tr>
<tr>
<td>3</td>
<td>0.172 0.179 0.177</td>
<td>0.170 0.178 0.187</td>
</tr>
<tr>
<td>4</td>
<td>0.279 0.268 0.235</td>
<td>0.225 0.227 0.214</td>
</tr>
<tr>
<td>5</td>
<td>0.332 0.283 0.307</td>
<td>0.296 0.279 0.279</td>
</tr>
<tr>
<td>6</td>
<td>0.421 0.338 0.396</td>
<td>0.407 0.332 0.339</td>
</tr>
<tr>
<td>7</td>
<td>0.502 0.412 0.432</td>
<td>0.459 0.417 0.416</td>
</tr>
<tr>
<td>8</td>
<td>0.526 0.524 0.582</td>
<td>0.529 0.488 0.478</td>
</tr>
<tr>
<td>9</td>
<td>0.590 0.592 0.590</td>
<td>0.590 0.591 0.549</td>
</tr>
<tr>
<td>10</td>
<td>0.660 0.654 0.655</td>
<td>0.691 0.645 0.637</td>
</tr>
<tr>
<td>11</td>
<td>0.687 0.694 0.721</td>
<td>0.773 0.649 0.700</td>
</tr>
<tr>
<td>12</td>
<td>0.694 0.731 0.794</td>
<td>0.798 0.817 0.774</td>
</tr>
<tr>
<td>13</td>
<td>0.820 0.819 0.814</td>
<td>0.832 0.837 0.823</td>
</tr>
<tr>
<td>14</td>
<td>0.926 0.922 0.923</td>
<td>0.885 0.932 0.896</td>
</tr>
<tr>
<td>15</td>
<td>0.981 0.978 0.979</td>
<td>0.894 0.980 0.937</td>
</tr>
</tbody>
</table>

For example, the first Gaussian in female TE has the relative position value of 0.05. This means that the first Gaussian is located in the first 5% part of the sentence, and input feature vectors are extracted around 5% part of the sentence are mainly judged by the first Gaussian.

Although the purpose of this study is to improve repeatability of diagnosis, we shortly mention the accuracy of the classifier. To make a fair comparison, we used the same test dataset of previous study conducted by Do et al. [4]. The accuracies of the GMM-based classifier were 41.3% and 39.3% for male and female, respectively, which are slightly better than Do et al. It should be noted that the accuracies of voice diagnosis are still low and have to be improved further. However, improvement of accuracy is not the main point of this study. The proposed method has better repeatability (see the next section) than the previous study while maintaining the similar accuracy.

#### 3.2. Diagnosis Repeatability Comparison

To evaluate the repeatability of the SC classification, we gathered voice data of 10 individuals. Total 50 recordings for each individual were prepared with the same way of gathering the training data explained in Section 2.1. It took more than a week to record
Table 2: Comparison results of diagnosis stability between the previous study and the proposed method.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Repeatability (%)</th>
<th>Probability mean</th>
<th>Probability standard deviation</th>
<th>Repeatability (%)</th>
<th>Probability mean</th>
<th>Probability standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>0.561</td>
<td>0.098</td>
<td>100</td>
<td>0.460</td>
<td>0.016</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>0.632</td>
<td>0.119</td>
<td>100</td>
<td>0.434</td>
<td>0.039</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>0.732</td>
<td>0.090</td>
<td>98</td>
<td>0.406</td>
<td>0.027</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>0.500</td>
<td>0.059</td>
<td>100</td>
<td>0.397</td>
<td>0.025</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>0.735</td>
<td>0.112</td>
<td>100</td>
<td>0.451</td>
<td>0.031</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0.672</td>
<td>0.090</td>
<td>70</td>
<td>0.381</td>
<td>0.020</td>
</tr>
<tr>
<td>7</td>
<td>84</td>
<td>0.603</td>
<td>0.100</td>
<td>82</td>
<td>0.362</td>
<td>0.013</td>
</tr>
<tr>
<td>8</td>
<td>56</td>
<td>0.567</td>
<td>0.082</td>
<td>76</td>
<td>0.381</td>
<td>0.013</td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>0.689</td>
<td>0.130</td>
<td>90</td>
<td>0.384</td>
<td>0.019</td>
</tr>
<tr>
<td>10</td>
<td>76</td>
<td>0.629</td>
<td>0.140</td>
<td>100</td>
<td>0.408</td>
<td>0.017</td>
</tr>
<tr>
<td>Average</td>
<td>78.2</td>
<td>0.632</td>
<td>0.102</td>
<td>91.6</td>
<td>0.406</td>
<td>0.022</td>
</tr>
</tbody>
</table>

50 times for each individual, so that the recordings contained
the intrapersonal voice variations enough. We used this data
to compare the proposed method and the previous study
conducted by Do et al.

We tested the repeatability and probability values (mean
and standard deviation) as a measure of diagnosis stability.
Repeatability was defined by the ratio of the number of majority
decision to total number of tests. In every test, we also
obtained the probability value of the current decision, that
is, the maximum probability among \( \pi_k \). Since the problem
of SC diagnosis is three-class classification, the maximum
probability is larger than 1/3. A stable classification has a high
percentage of repeatability and a low probability standard
deviation.

Table 2 shows the comparison results of diagnosis sta-
bility. For 10 individuals, the average repeatability of the
proposed method was 91.6%, which clearly showed better
repeatability than 78.2% of Do et al. And probability standard
deviation was 0.022, which was much lower than 0.102 of
Do et al. It means that the proposed method generated more
consistent results in 50 times classification test of the same
individual’s voice than the previous study. Interpretation of
the probability mean value might be controversial. However,
we think that a low probability mean value is more appropri-
ate, considering that the accuracy of the voice diagnosis is low
around 40%.

Some of subjects, such as subjects 6 and 8, had low
repeatabilities in our method. The subject whose voice feature
vector is located close to decision boundary can have a low
repeatability, because a small change of the feature vector can
switch the final decision. In this case, the probability value
is also close to 1/3, which means that the decision has less
confidence. We may have a difficulty in determining one’s
constitution with high confidence when his/her voice does
not have clear constitutional characteristics. This can drop
the repeatability of diagnosis. However, even in this case, it
is good to have similar probability values in repeated tests
for the same subject. Our method shows stable diagnosis in
probability sense, having low probability standard deviations
even in low repeatability cases.

An improved repeatability is obtained from the proposed
method that analyses every single frame of the sentence. The
improvement is caused by not only analysis method but also
recording contents. Our approach uses only sentence rather
than both vowels and the sentence since the features from
vowels have relatively large variations for the same subject.
To examine the stability of each feature value itself, standard
deviations of feature values from the sentence and vowels are
summarized in Table 3. Some of the vowel features, used in
the study of Do et al., are compared to the sentence features in
this study. Table 3 shows that sentence features have smaller
standard deviations than vowels. Standard deviations of the
most sentence features are ranged from 0.2 to 0.7, while
standard deviations of many vowel features are larger than 1.
Therefore, stability of diagnosis with only sentence will out-
perform other cases, regardless of classification algorithms.

4. Discussion and Conclusions

In this study, a stable classification method for voice-based
Sasang constitutional diagnosis was proposed. In contrast
to the previous study, which used averaged features from
vowels and one sentence, the proposed method used MFCCs
extracted from the sentence only. Since we did not use
averaged features to avoid losing useful information of indi-
vidual’s voice characteristics, features from every valid frame
were required to be analysed separately. We defined a feature
vector that contained MFCCs and relative position of the
current frame in the voice signal. Therefore, the feature vector
represented both voice features and information about where
the features were extracted in time axis of the voice signal. The
technique was necessary to use the sentence instead of
the vowels, because the sentence included a variety of voice
features in each frame, which should not be averaged.

The proposed method employed GMM for detailed
analysis of every valid frame in voice signal. Each Gaussian
Table 3: Standard deviations of feature values from the sentence and vowels.

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC1</td>
<td>0.290</td>
<td>1.287</td>
<td>0.535</td>
<td>0.414</td>
<td>0.333</td>
<td>0.317</td>
<td>0.432</td>
<td>0.269</td>
<td>0.395</td>
<td>0.245</td>
</tr>
<tr>
<td>MFCC2</td>
<td>0.292</td>
<td>0.358</td>
<td>0.270</td>
<td>0.407</td>
<td>0.353</td>
<td>0.363</td>
<td>0.391</td>
<td>0.289</td>
<td>0.264</td>
<td>0.290</td>
</tr>
<tr>
<td>MFCC3</td>
<td>0.325</td>
<td>0.597</td>
<td>0.302</td>
<td>0.350</td>
<td>0.635</td>
<td>0.413</td>
<td>0.511</td>
<td>0.205</td>
<td>0.313</td>
<td>0.259</td>
</tr>
<tr>
<td>MFCC4</td>
<td>0.246</td>
<td>0.200</td>
<td>0.223</td>
<td>0.434</td>
<td>0.286</td>
<td>0.180</td>
<td>0.448</td>
<td>0.322</td>
<td>0.260</td>
<td>0.139</td>
</tr>
<tr>
<td>MFCC5</td>
<td>0.451</td>
<td>0.320</td>
<td>0.352</td>
<td>0.423</td>
<td>0.479</td>
<td>0.309</td>
<td>0.708</td>
<td>0.470</td>
<td>0.288</td>
<td>0.280</td>
</tr>
<tr>
<td>MFCC6</td>
<td>0.213</td>
<td>0.292</td>
<td>0.199</td>
<td>0.227</td>
<td>0.258</td>
<td>0.242</td>
<td>0.479</td>
<td>0.309</td>
<td>0.523</td>
<td>0.329</td>
</tr>
<tr>
<td>MFCC7</td>
<td>0.336</td>
<td>0.287</td>
<td>0.428</td>
<td>0.266</td>
<td>0.341</td>
<td>0.294</td>
<td>0.311</td>
<td>0.278</td>
<td>0.262</td>
<td>0.307</td>
</tr>
<tr>
<td>MFCC8</td>
<td>0.340</td>
<td>0.346</td>
<td>0.257</td>
<td>0.289</td>
<td>0.269</td>
<td>0.293</td>
<td>0.303</td>
<td>0.347</td>
<td>0.258</td>
<td>0.204</td>
</tr>
<tr>
<td>MFCC9</td>
<td>0.268</td>
<td>0.399</td>
<td>0.279</td>
<td>0.521</td>
<td>0.374</td>
<td>0.307</td>
<td>0.480</td>
<td>0.420</td>
<td>0.362</td>
<td>0.211</td>
</tr>
<tr>
<td>MFCC10</td>
<td>0.359</td>
<td>0.279</td>
<td>0.284</td>
<td>0.461</td>
<td>0.387</td>
<td>0.412</td>
<td>0.381</td>
<td>0.485</td>
<td>0.286</td>
<td>0.193</td>
</tr>
<tr>
<td>MFCC11</td>
<td>0.251</td>
<td>0.334</td>
<td>0.304</td>
<td>0.412</td>
<td>0.375</td>
<td>0.411</td>
<td>0.432</td>
<td>0.410</td>
<td>0.251</td>
<td>0.235</td>
</tr>
<tr>
<td>MFCC12</td>
<td>0.329</td>
<td>0.380</td>
<td>0.178</td>
<td>0.325</td>
<td>0.590</td>
<td>0.398</td>
<td>0.669</td>
<td>0.278</td>
<td>0.265</td>
<td>0.313</td>
</tr>
</tbody>
</table>

Sentence features

<table>
<thead>
<tr>
<th>Vowel features</th>
<th>aENG</th>
<th>aFI</th>
<th>aSHIM</th>
<th>eSHIM</th>
<th>iDTF0</th>
<th>iJITT</th>
<th>oDTF0</th>
<th>oPW</th>
<th>uF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>aENG</td>
<td>0.069</td>
<td>0.878</td>
<td>0.208</td>
<td>0.417</td>
<td>0.394</td>
<td>0.129</td>
<td>0.912</td>
<td>0.244</td>
<td>1.549</td>
</tr>
<tr>
<td>aFI</td>
<td>0.634</td>
<td>0.336</td>
<td>0.453</td>
<td>0.380</td>
<td>0.282</td>
<td>0.552</td>
<td>0.489</td>
<td>0.652</td>
<td>0.366</td>
</tr>
<tr>
<td>aSHIM</td>
<td>0.848</td>
<td>1.030</td>
<td>0.875</td>
<td>1.167</td>
<td>1.094</td>
<td>1.094</td>
<td>2.843</td>
<td>2.195</td>
<td>1.448</td>
</tr>
<tr>
<td>eSHIM</td>
<td>1.127</td>
<td>1.397</td>
<td>0.986</td>
<td>1.555</td>
<td>0.873</td>
<td>1.120</td>
<td>2.870</td>
<td>3.048</td>
<td>1.393</td>
</tr>
<tr>
<td>iJITT</td>
<td>0.933</td>
<td>1.066</td>
<td>0.695</td>
<td>5.042</td>
<td>1.623</td>
<td>1.154</td>
<td>2.869</td>
<td>2.789</td>
<td>1.096</td>
</tr>
<tr>
<td>oPW</td>
<td>0.046</td>
<td>0.454</td>
<td>0.206</td>
<td>0.500</td>
<td>0.294</td>
<td>0.272</td>
<td>1.133</td>
<td>0.225</td>
<td>1.198</td>
</tr>
<tr>
<td>uF1</td>
<td>0.263</td>
<td>0.286</td>
<td>0.874</td>
<td>0.580</td>
<td>0.188</td>
<td>0.356</td>
<td>0.422</td>
<td>0.273</td>
<td>0.216</td>
</tr>
</tbody>
</table>

Vowel features, xENG (energy), xFI (1st formant), xSHIM (shimmer), xDTF0 (average difference of pitch over the time interval), xJITT (jitter), and xPW (power) were used in the study of Do et al. [4] x ∈ {a, e, i, o, u}.

Acknowledgment

This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (no. 2006-2005173).

References


