# Diagnosing Vocal Disorders using Cobweb Clustering of the Jitter, Shimmer, and Harmonics-to-Noise Ratio

**Keonsoo Lee[1], Chanki Moon[2], Yunyoung Nam[2*]**
[1]Medical Information Communication Technology, Soonchunhyang University,
Asan, Republic of Korea
[2]Department of Computer Science and Engineering Soonchunhyang University,
Asan, Republic of Korea
[e-mail: keonsoo@sch.ac.kr, moonchanki1992@gmail.com, ynam@sch.ac.kr]
*Corresponding author: Yunyoung Nam

*Received April 10, 2018; revised May 14, 2018; accepted May 30, 2018;
published November 30, 2018*

## Abstract

A voice is one of the most significant non-verbal elements for communication. Disorders in vocal organs, or habitual muscular setting for articulatory cause vocal disorders. Therefore, by analyzing the vocal disorders, it is possible to predicate vocal diseases. In this paper, a method of predicting vocal disorders using the jitter, shimmer, and harmonics-to-noise ratio (HNR) extracted from vocal records is proposed. In order to extract jitter, shimmer, and HNR, one-second's voice signals are recorded in 44.1khz. In an experiment, 151 voice records are collected. The collected data set is clustered using cobweb clustering method. 21 classes with 12 leaves are resulted from the data set. According to the semantics of jitter, shimmer, and HNR, the class whose centroid has lowest jitter and shimmer, and highest HNR becomes the normal vocal group. The risk of vocal disorders can be predicted by measuring the distance and direction between the centroids.

**Keywords:** *Cobweb, Clustering, HNR, Jitter, Shimmer, Vocal Disorder*

This work was supported by the Soonchunhyang University Research Fund and also supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2017R1D1A3B03028309).

**http://doi.org/10.3837/tiis.2018.11.020** **ISSN : 1976-7277**

## 1. Introduction

The voice is one of the most significant elements for communication [1, 2]. As an element of non-verbal communication, the voice is responsible for two objectives. The first is explicitly to deliver the contents of the communication. The second is to represent the characteristics of the speaker. Phonetic quality indicates whether the pronunciation of each word is correct. For determining phonetic quality, the pronunciations of persons in various age, nationality, gender, and so on, are collected. The the representative pattern of pronunciations for each word is extracted from the collected set. By comparing the similarity between the representative pattern and the input voice signal, the phonetic quality is determined [3]. On the other hand, voice quality indicates whether the pronunciation is audible [4, 5]. Tho voices of those who catch a cold have lower voice quality than the voices of noraml persons. The voice quality is affected by two elements. One is an organic factor such as vocal tract anatomy or physiology. The other is a setting factor which is the habitual muscular setting for articulatory [2].

For natural language processing, it is important to recognize the semantics of the recorded voice signals. However, the semantics of speaking is not required for determining vocal disorders. Instead, the steadiness of modulating the voice, which shows the status of articular organs, is more significant feature for predicting vocal disorders. The most popular properties for describing the steadiness of voice are jitter, shimmer, and harmonics-to-noise ratio (HNR) [6, 7, 8]. Jitter shows the periodicity of voice signals. The voice of good quality should have a stable periodicity. Shimmer shows the stability of amplitude. The amplitude is calculated from the peak and valley in every period. The voice of good quality should have a low standard deviation for the amplitudes of each period. HNR shows the ratio of the additive noises in the voice signals. Noises are defined as an aperiodic part of a voice. The aperiodic parts are mostly resulted from the turbulent airflow during phonation. Improper closure of the vocal folds makes the turbulent.

From the definitions of jitter, shimmer, and HNR, it is obvious to determine whether a given voice is qualified or not. A voice, which has low jitter, low shimmer, and high HNR, will be determined as a voice of good quality. However, the criteria for separating the low and high values of jitter, shimmer, and HNR are not confirmed. Depending on the way of calculating jitter, shimmer, and HNR, the thresholds for these features can be changed [9, 10, 11, 12, 13]. Moreover, the explanation of the predicted vocal disorders needs to be provided. Some bad voices can have high jitter value, high shimmer value, or high HNR value. Therefore, it is necessary to distinguish the bad voices, which have different patterns of features.

In this paper, we propose a method of diagnosing vocal disorders with detailed explanation, which shows why the given voice is classified as a disordered voice. For diagnosing vocal disorders, cobweb clustering method is used [14, 15]. Using the cobweb, the collected voice records are grouped with similar records. From the clusters, a cluster whose centroid has lowest jitter, lowest shimmer, and highest HNR is selected as a normal cluster that includes the voice records in healthy condition. By comparing the distance and direction between the centroids of other clusters, the vocal records in other clusters are determined as disordered voices. For the experiment, 151 voice records from 115 participants are collected.

Section 2 describes the methods for calculating jitter, shimmer, and HNR from the voice signals, and the way of predicating vocal disorders from the calculated three features. In Section 3, an experiment for evaluating the proposed method is explained. Section 4 presents the discussion about the experimental results and concludes.

# 2. Method

## 2.1 Features for Determining Vocal Disorders

In general, steady voice is regarded as a good voice. In order to measure the steadiness of a voice, various criteria such as frequency measures, frequency perturbation measures, measures of perturbation intensity, voice break measures, and mute or unvoiced segments measures are used [16]. Applications that calculate such measures from voices are distributed both commercially and non-commercially. Multi-Dimensional Voice Program (MDVP) [17], Dr. Speech [18], Praat [19], CSpeech [20] are widely used application for this purpose. From the various criteria, the preferred features for determining the steadiness of a voice are jitter, shimmer, and HNR [6, 7, 8].

Jitter is the temporal variation of the signal. When the pulse of the signal is delayed or ahead, the jitter of the voice increases. Therefore, a voice that has irregular pulse cycles shows high jitter value. Jitter is calculated by dividing the average absolute difference between consecutive periods with the average period. According to MDVP, a voice in pathological problems has a jitter value higher than 1.040% [21]. The numerator of the equation for calculating jitter can be also used for replacing jitter. According to MDVP, the average absolute difference between consecutive periods of a voice that has pathological problems is higher than 83.200 microseconds. Generally, when this value is divided by the average period, about 1.040 is resulted. In order to calculate jitter, the average period is calculated from the 1 second's signals. By using a window size, the relative average perturbation can be used. The window size can be 3, 5, or 11. The window size means the number of periods to be used for calculating the relative average perturbation. According to MDVP, when the window size is 5, the threshold for pathology is 0.680%. Depending on the way of calculating jitter, its accuracy and computing load are different. Moreover, jitter is one of the most sensitive for the noise. In the process of calculating jitter, finding the start and end positions in the given voice signals is important for defining the length of each period. Therefore, a complex or dynamic pronouncement is hard to find the length of a period. At the same time, noises in the voice signals easily confuse the process of detecting the length of a period. Therefore, the pronouncement of vowels whose signals are stable is preferred. /a/, /u/, or /i/ are the preferred vowels for calculating jitter, and /a/ which requires the least muscular tension and consists of relatively high frequencies is used in the experiments in Section 3. Equation 1 shows the method for calculating jitter used in this paper.

$$\text{Jitter} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} * 100 \qquad (1)$$

Shimmer is the variability of the peak-to-peak amplitude of the voice. There are various methods for calculating shimmer. The simplest method is to divide the average absolute difference between the amplitudes of consecutive period by the average amplitude. According to MDVP, a voice in pathological problems has a shimmer value higher than 3.810%. The difference between the amplitudes of consecutive periods may calculated with common logarithm. In this case, the unit for shimmer value is dB. According to MDVP, a voice in pathological problems has higher than 0.350 dB. The average absolute difference between amplitude of periods can be changed using different window size. The window size can be 3, 5, or 11. The window size is the number of periods to be used for calculating the average absolute difference between amplitude. The main difference of these methods for calculating shimmer is the numerator of the equation because the denominator is the same in all the methods.

Depending on the required accuracy and available computing power, the proper method can be selected because all the methods calculate the shimmer, which shows how the voice signals are trembling in amplitude. Equation 2 shows the method for calculating shimmer used in this paper.

$$\text{Shimmer} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} * 100 \qquad (2)$$

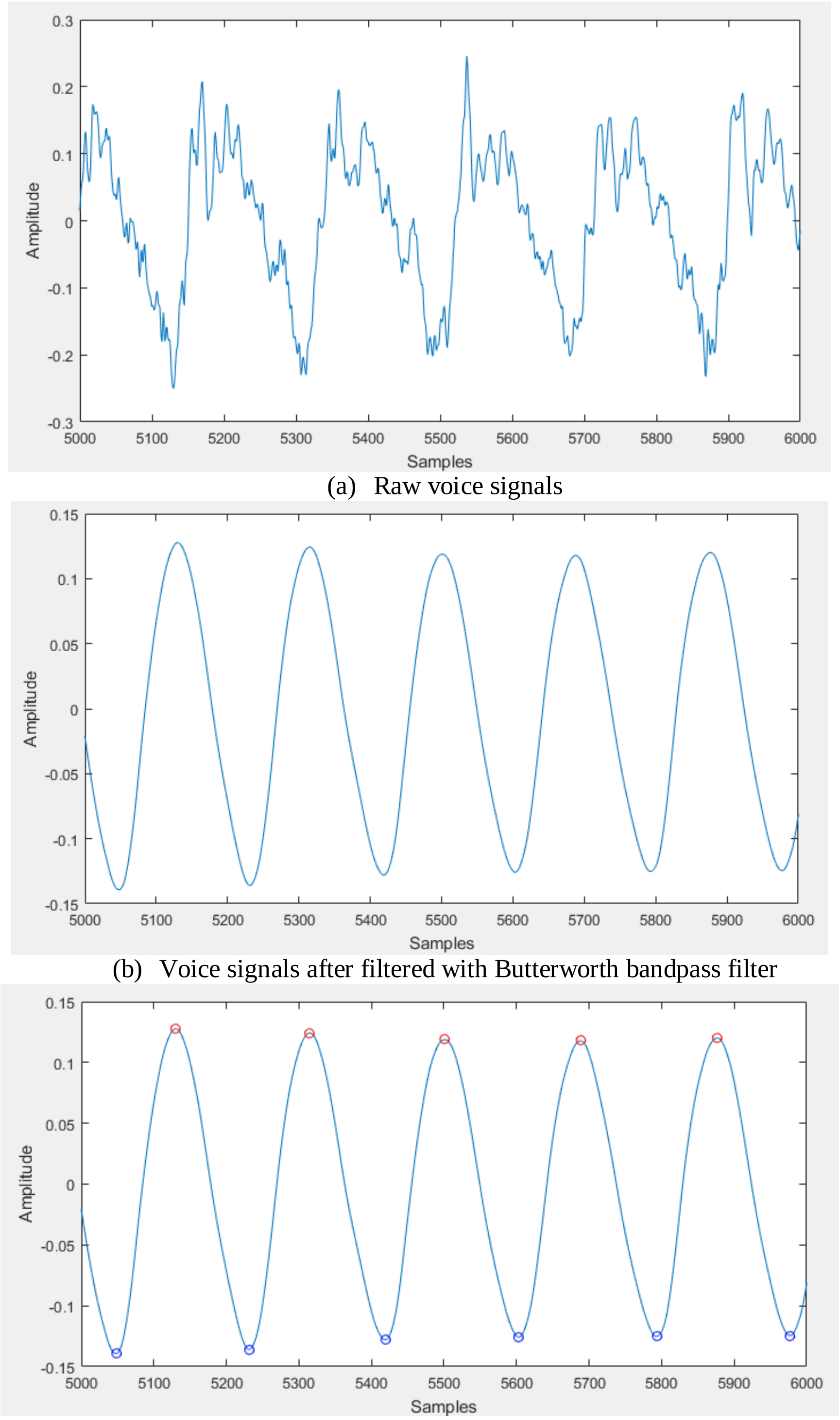
HNR is the ratio between the periodic part and aperiodic part of the voice signal. Voice with less aperiodic part has a low HNR value. However, higher HNR does not means higher voice quality. It is impossible to remove all the noises from the obtained signals practically. At the same time, it is also impossible for bionic organs to generate perfectly static movement. Depending on the pronunciation, health voices have proper range of HNR. For example, /a/ or /i/ sounds have a harmonicity of around 20 dB, which means 99% of periodic part and 1% of aperiodic part. For /u/ sound, the proper HNR is around 40 dB, which means 99.99% of periodic part and 0.01% of aperiodic part. This difference is originated from the fact that /a/ are /i/ sound are mainly composed of high frequencies and /u/ sound is mainly composed of low frequencies. Therefore, /u/ sound is easier for a larynx to pronounce with less muscle tensions. When /a/ sound with HNR value lower than 20 dB, it is heard hoarsely. When /a/ sound with HNR value higher than 40 dB, some pathological problems in vocal cords or mistakes in obtainment of the voice signals can be suspected. Equation 3 shows the method for calculating HNR used in this paper.

$$\text{HNR} = 10 * \log_{10} \frac{n * \int_0^r f_A{}^2(r)\, dr}{\sum_{i=1}^{n} \int_0^{T_i} [f_i(r) - f_A(r)\,]^2\, dr} \qquad (3)$$

## 2.2 Data Collection and Signal Processing

In order to determine the vocal disorders, 147 records are collected from 115 participants who attend the Department of Otorhinolaryngology in Soonchunhyang Bucheon Hospital (SBH). Each participant was asked to pronounce /a/ sound for 3 seconds. As shown in Section 2.1, jitter, shimmer, and HNR are easily polluted by noises, /a/ sound is used. From the 3 seconds' length of the voice, the first and third second's signals are removed. The voice signals are recorded in 44100 sampling rates.

In order to extract jitter, shimmer, and HNR from each record, the signals are preprocessed. The preprocessing process is composed of two steps. In the first step, noises except the patient's voice are removed by using a filter. In this research, 2nd-order Butterworth bandpass filter with cutoff frequency range between 80 and 300 Hz is applied to the raw voice signal. The cutoff frequency is determined because the speech of typical adult has a fundamental frequency from 85 to 196 Hz for a male and from 155 to 334 Hz for a female [1]. In the second step, the peaks and valleys in the signals are detected. Each period in the signals are extracted based on the peaks and valleys. A period is calculated from a peak to the next peak or from a valley to the next valley. In this paper, the length of peak-to-peak is used for measuring the length of a period. An amplitude is calculated from a peak to the nearest next valley or a valley to the nearest next peak. Depending on the first point of the voice signal, valley-to-peak or peak-to-valley is determined. **Fig. 1** shows the voice signals following the processing steps.



(a) Raw voice signals

(b) Voice signals after filtered with Butterworth bandpass filter

(c) Peaks and Valleys detected from the filtered signals

**Fig. 1.** Steps for preprocessing the obtained voice signals

## 2.3 Vocal Disorder Determination

In order to determine vocal disorders, clustering method is employed. Generally, for determining the category of new observation, classification methods are used. The problem of determining vocal disorders is also a kind of classification problem because the result is to identify the category where the newly provided voice belongs. However, to employ such classification methods such as neural networks [22], support vector machine [23], decision tree [24], or k-nearest neighbors [25], the features of the voices should be preciously calculated. Without the guarantee that features of the same voice are always same, the rules or learned classifiers cannot be applied to other voices whose features are calculated with different methods, or preprocessed with different configured filters. Therefore, instead of using classification methods, we use clustering methods, which group the voices depending on the similarity [26, 27]. The methods for calculating each feature's value and preprocessing the voice signals become independent from the way of determining vocal disorders. The determination process consists of four steps. **Fig. 2** shows the process for determining vocal disorders.

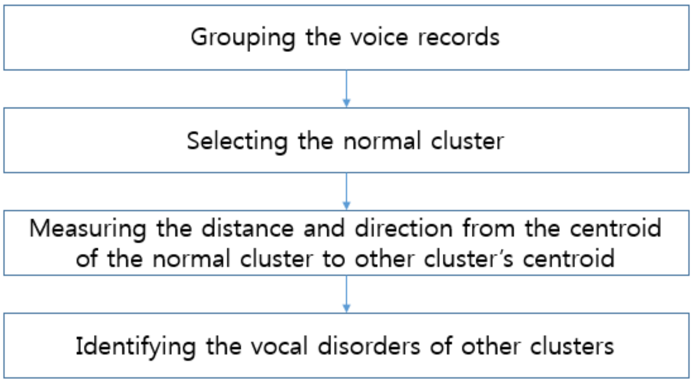

Grouping the voice records
Selecting the normal cluster
Measuring the distance and direction from the centroid of the normal cluster to other cluster's centroid
Identifying the vocal disorders of other clusters

**Fig. 2.** The process for determining vocal disorders

In the first step, clusters are made from the collected voice records. As each voice record has three features which are jitter, shimmer, and HNR, three dimensional space is used for describing the clusters. In the second step, normal cluster of voice records is selected. The normal cluster means that the voice records in the cluster are in healthy condition. Depending on the definition of three features, a cluster whose centroid has lowest jitter, lowest shimmer, and highest HNR becomes the normal cluster. Two heuristic rules are added for selecting the normal cluster. The first rule is for configuration of the weights of three features. *Rank* is a function that calculates rank of the parameter in ascending order and the inverse *Rank* is a function that calculates rank of the parameter in descending order. This rule is expressed in Equation 4.

$$\text{RankingValue} = \alpha \cdot Rank^{-1}(\text{jitter}) + \beta \cdot Rank^{-1}(\text{shimmer}) + \gamma \cdot Rank(\text{HNR}) \qquad (4)$$

From the clustering step, multiple clusters are made depend on the similarity among the voice records. Some clusters may have only a few elements. As long as the training data set is not biased, the size of the normal cluster should be bigger than the threshold value. According to the central limit theorem, if the size of the training set is big enough and the normal distribution is assumed, the normal cluster should have at least 68.27% of the total data set.

Therefore, when the size of normal cluster, which is selected by using the first rule, is not big enough, the selected cluster should be merged with near clusters or replaced by the next ranked cluster. In this paper, we use a cobweb clustering method for grouping the voice records. Cobweb is a hierarchical and conceptual clustering method. Using the category utility, which is a way of evaluating the quality of the classification, similar groups are merged and distinguishing group is split. Therefore, when a selected cluster is not sufficient for the normal cluster, it can be easily extended by rising the hierarchy. At the same time, the vocal disorder is the complementary set of normal voice, cobweb, whose root cluster covers the whole problem space, is suitable for determining vocal disorder.

The third step is to measure the distance and direction from the centroid of the normal cluster to other cluster's centroid. This step is relatively simple because the centroid of each cluster is a vector with three elements. The distance is calculated in Euclidean distance and the direction is calculated in the direction cosines. In order to calculate the distance and direction, the feature values of the centroids are normalized and multiplied with the weight for each feature as defined in Equation 4. The methods for calculating distance and direction are shown in Equation 5.

$$\begin{aligned}
\text{Distance} &= \sqrt{(\alpha\cdot\text{Norm(jitter)})^2+(\beta\cdot\text{Norm(shimmer)})^2+(\gamma\cdot\text{Norm(HNR)})^2}\\
\text{Direction}_{\text{jitter}} &= \frac{\alpha\cdot\text{Norm(jitter)}}{\sqrt{(\alpha\cdot\text{Norm(jitter)})^2+(\beta\cdot\text{Norm(shimmer)})^2+(\gamma\cdot\text{Norm(HNR)})^2}}\\
\text{Direction}_{\text{shimmer}} &= \frac{\beta\cdot\text{Norm(shimmer)}}{\sqrt{(\alpha\cdot\text{Norm(jitter)})^2+(\beta\cdot\text{Norm(shimmer)})^2+(\gamma\cdot\text{Norm(HNR)})^2}}\\
\text{Direction}_{\text{HNR}} &= \frac{\gamma\cdot\text{Norm(HNR)}}{\sqrt{(\alpha\cdot\text{Norm(jitter)})^2+(\beta\cdot\text{Norm(shimmer)})^2+(\gamma\cdot\text{Norm(HNR)})^2}}
\end{aligned} \quad (5)$$

These calculated distance and direction are used in the fourth step for identifying the vocal disorders of other clusters. The distance indicates the degree of the vocal disorders. The longer distance from the normal cluster means more significant disorders. The vocal disorders are explained by using the directions of each feature. When a specific direction is close to 0, the feature can be regarded as a neutral cause for the vocal disorders.

# 3. Experimental Result

## 3.1 Clusters of Voice Records

**Fig. 3** shows the scatter plot of the 151 voice records, which are collected and preprocessed as described in Section 2.

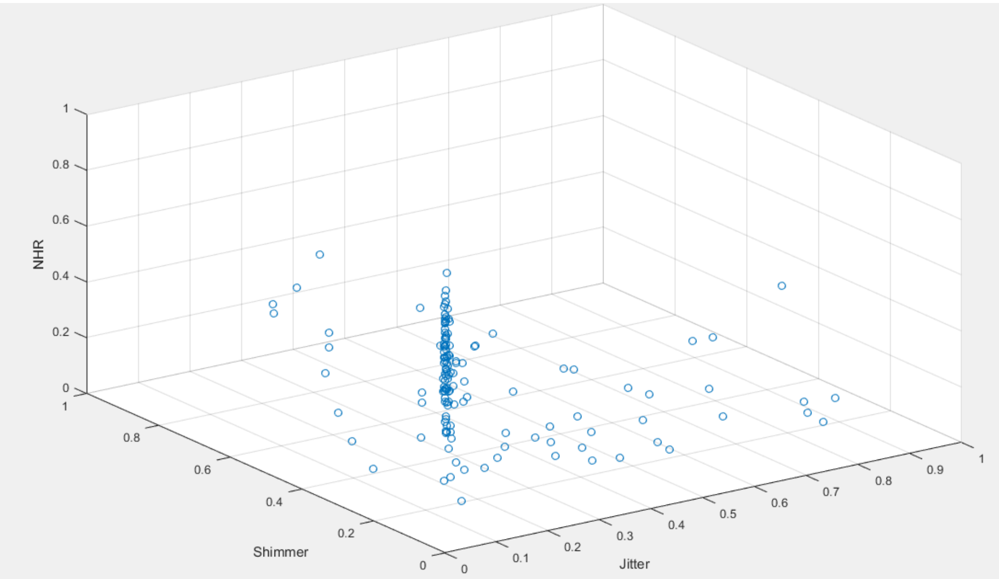


**Fig. 3.** The scatter plot of collected voice records with three features, which are jitter, shimmer, and HNR

**Table 1** shows the statistical characteristic of each features. The values of three features are different depending on the methods of signal processing and feature calculation. Depending on the window size for feature extraction, the configuration of bandpass filter, and the sampling rates of recording the voice can affect the result. In this experiment, jitter, shimmer and HNR are calculated using the equation 1, 2, and 3 shown in Section 2.1.

**Table 1.** Statistical characteristics of three features

| | Min | Max | Mean | StdDev |
|---|---|---|---|---|
| Jitter | 0.27 | 67.63 | 9.979 | 17.16 |
| Shimmer | 0.60 | 288.88 | 34.79 | 58.85 |
| HNR | 5.00 | 37.81 | 21.31 | 9.026 |
| Normalized_Jitter | 0 | 1 | 0.144 | 0.255 |
| Normalized_Shimmer | 0 | 1 | 0.119 | 0.204 |
| Normalized_HNR | 0 | 1 | 0.497 | 0.275 |

### 3.2 Vocal Disorder Determination

In order to determine vocal disorders, the voice records are clustered using cobweb. The configuration for cobweb is made with 0.1 for the acuity, 0.003 for cutoff, and 10 as a seed number for randomization. **Fig. 4** shows the result of the cobweb clustering. A cluster, which has sub clusters, has a name starting ‘Node’. A cluster, which has no sub clusters, has a name starting ‘Leaf’. As the clusters have hierarchical relationship with others, a node can represent its children by combining all the elements in its sub clusters. For example, *Node 5* can be described by combining *Leaf 6* and *Leaf 7*. Some clusters can have no elements. For example, *Leaf 20* is an empty cluster. In the process of training the cobweb, *Node 18* was divided into *Leaf 19* and *Leaf 20*. However, in the process of validation, no element is assigned to *Leaf 20*. In such case, empty clusters can be made.

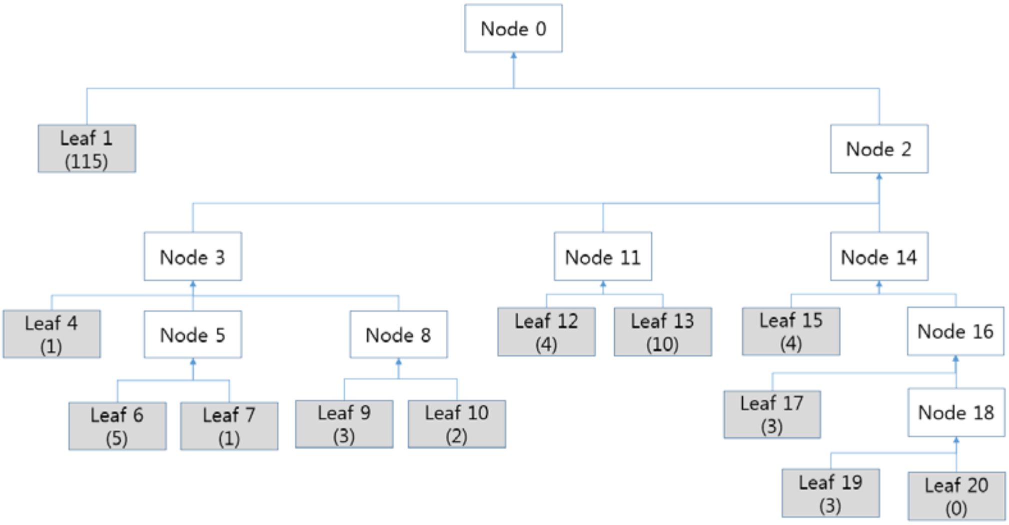


**Fig. 4.** The result of voice records after applying cobweb clustering method

The elements of each cluster resulted from the cobweb clustering are shown in **Fig. 5**. *Leaf 1* is shown as a group of red dots. The scatter plot with two features are shown in **Fig. 6**.

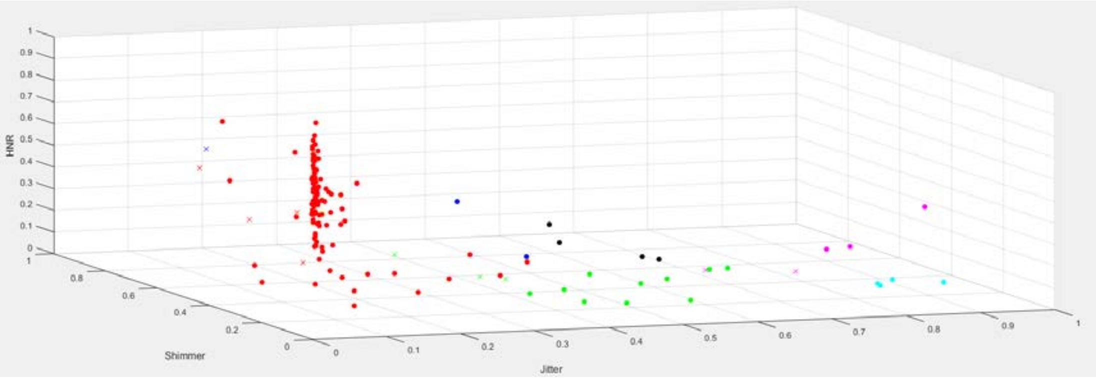
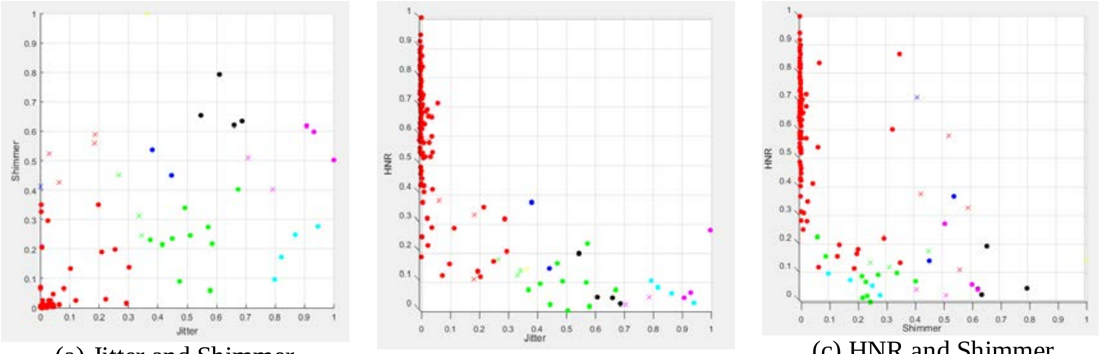


**Fig. 5.** The scatter plot of collected voice records after cobweb clustering process



(a) Jitter and Shimmer

(b) HNR and Jitter

(c) HNR and Shimmer

**Fig. 6.** The scatter plot of collected voice records with two axis

The centroid of each cluster is shown in **Table 2**. According to the rule for determining the normal cluster which is described in Section 2.3, *Leaf 1* is selected as a normal cluster. The weight for each feature is set as 0.33 because in this experiment we assumed that all the features are equally significant. As shown in **Fig. 4**, *Leaf 12* and *Leaf 13* have the same super cluster. Therefore, they can be combined into *Node 11*. Even though *Leaf 7* and *Leaf 9* are siblings, they do not have a common super cluster. Therefore, *Leaf 7* and *Leaf 9* cannot be combined directly. *Leaf 9* can be combined with *Leaf 10* because they have the same super cluster. *Leaf 7* can be combined with *Leaf 6*. After that, *Node 5* and *Node 8* can be combined. The rule for combining is based on the fact that only the most similar clusters can be combined.

**Table 2.** Centroid of each cluster made by cobweb

| | # of instances | Jitter | Shimmer | HNR |
|---|---|---|---|---|
| Leaf 1 | 115 | 0.025 | 0.025 | 0.607 |
| Leaf 4 | 1 | 0.362 | 1.000 | 0.144 |
| Leaf 6 | 5 | 0.128 | 0.531 | 0.300 |
| Leaf 7 | 1 | 0.000 | 0.412 | 0.714 |
| Leaf 9 | 3 | 0.316 | 0.335 | 0.143 |
| Leaf 10 | 2 | 0.413 | 0.494 | 0.256 |
| Leaf 12 | 4 | 0.858 | 0.197 | 0.065 |
| Leaf 13 | 10 | 0.512 | 0.230 | 0.084 |
| Leaf 15 | 4 | 0.625 | 0.674 | 0.078 |
| Leaf 17 | 3 | 0.946 | 0.572 | 0.127 |
| Leaf 19 | 3 | 0.734 | 0.473 | 0.030 |
| Leaf 20 | 0 | 0 | 0 | 0 |

As *Leaf 1* is determined as the normal cluster, other clusters can be explained why the elements of each cluster is classified as a disordered voice. **Table 3** shows the distances between the centroid of *Leaf 1* and the centroid of other clusters. The voice records in *Leaf 17*, which has the longest distance are the riskiest voices. The voice records in *Leaf 7*, which has the shortest distance, are the least risky voices. According to the direction of centroid of this cluster, the rehabilitative plan for the voices in this cluster needs to be focused on keeping the pronouncing amplitude stable.

**Table 3.** Distance and Direction between clusters

| | Distance | $\mathrm{Direction}_{\mathrm{jitter}}$ | $\mathrm{Direction}_{\mathrm{shimmer}}$ | $\mathrm{Direction}_{\mathrm{HNR}}$ |
|---|---|---|---|---|
| Leaf 4 | 1.130 | 0.298 | 0.862 | -0.409 |
| Leaf 6 | 0.601 | 0.172 | 0.842 | -0.511 |
| Leaf 7 | 0.402 | -0.062 | 0.962 | 0.266 |
| Leaf 9 | 0.629 | 0.462 | 0.493 | -0.737 |

| | | | | |
|---|---|---|---|---|
| Leaf 10 | 0.703 | 0.553 | 0.667 | -0.499 |
| Leaf 12 | 1.009 | 0.826 | 0.171 | -0.538 |
| Leaf 13 | 0.743 | 0.655 | 0.276 | -0.704 |
| Leaf 15 | 1.031 | 0.583 | 0.630 | -0.513 |
| Leaf 17 | 1.174 | 0.785 | 0.466 | -0.408 |
| Leaf 19 | 1.019 | 0.697 | 0.440 | -0.567 |
| Leaf 20 | N/A | N/A | N/A | N/A |

# 4. Discussion and Conclusion

## 4.1 Discussion

The quality of sound has been analyzed in terms of the jitter, shimmer, and HNR in various academic and commercial domains. However, the values calculated for the features depend on the way the raw voice signals are preprocessed. The criteria for diagnosing vocal disorders are application-specific, as each application has its own preprocessing method, and depend on the operator’s experience and knowledge. In this paper, we have presented a clustering-based abnormality detection method. We used the cobweb method to acquire the hierarchical relations between clusters, then based our classification on the cluster whose centroid exhibited the best voice quality. The degree of vocal disorder in the other clusters was calculated based on the distance and direction from the centroid of the basis cluster.

However, three problems remain unsolved. The first problem is that the feature calculation is still influenced by the way of preprocessing the voice records. In the experiment, 1 second’s length of voice signals is used for feature extraction. The calculated features are changed depending on the size and position of the window. At the same time, the filter, which removes the noise of the signals, also influence the calculated feature values. As the main objective of this paper is not to extract the features of each record exactly but to group the records and find the relations among the cluster, more robust and accurate method for feature calculation can increase the reliability of the clusters. The second problem is that there are concealed relations among jitter, shimmer, and HNR. For the patients whose voices are in the cluster whose centroid has low jitter, high shimmer, and low HNR, a rehabilitation plan which is to help the patients to pronounce with stable amplitude can be provided. If the knowledge about the mechanism of vocal cords, which shows how a sound of low jitter and high shimmer can be made from the vocal cords, is involved, more accurate and more reliable explanation for the given vocal records can be provided. The third problem is the boundary problem. An element, which locates in the boundary, can be easily misclassified. Moreover, as the length from the centroid to the boundary of a cluster is different for each cluster, the basic rule for measuring the severity of vocal disorders is easily spoiled. Therefore, for enhancing the quality of the automated vocal disorder detection, clinical knowledge for vocal cords is necessary. As a future work, the operating principals of vocal cords and the correlation between the changes of jitter, shimmer, and HNR and status of vocal cords will be researched and used as base knowledge.

### 4.2 Conclusion

The quality of voice is affected by the status of organics and the habitual setting for speaking. Therefore, it is possible to detect vocal disorders by analyzing the vocal sound. Jitter, shimmer and HNR are the most significant features, which can be extracted from the voice records for detecting vocal disorders. In this paper, we propose a method of determining a vocal disorder relatively using cobweb based clustering. In the experiment, 21 classes and their hierarchical relations are obtained from 151 voice records. The cluster, which has the semantically finest centroid is determined as a normal cluster. The voice records in other clusters are determined to have a risk of vocal disorder. The severity is predicted by measuring the distance and direction between the centroids of base cluster and the target cluster.

However, as described in Section 4.1, the heuristic rules for selecting the normal cluster and measuring the severity of vocal disorders in other cluster are not sufficient to make accurate and reliable explanation. The explicit way of calculating the degree of the risk of vocal disorder for each cluster and validation of the resulted degree will be researched in our future work.

## Acknowledgements

The authors would like to thank Jihyun Kim, Hanna Kim, and GaRam Jung for their valuable contributions to this project.

## References

[1] Williamson, G. *Human Communication: A Linguistic Introduction.* (Speechmark, 2001).

[2] M. Tiwari, and M. Tiwari. “Voice - How humans communicate?” J Nat Sci Biol Med 3, 3–11. 2012. Article(CrossRefLink)

[3] Rose, P., “Forensic Speaker Identification,” *CRC Press*, 2003.

[4] E. Keller, “The Analysis of Voice Quality in Speech Processing,” *Nonlinear Speech Modeling and Applications 54–73 Springer*, Berlin, Heidelberg, 2005. Article(CrossRefLink)

[5] J. D. Laver, “Voice quality and indexical information,” *Br J Disord Commun 3*, 43–54. 1968. Article(CrossRefLink)

[6] J. P. Teixeira, and P. O. Fernandes, “Jitter, Shimmer and HNR Classification within Gender, Tones and Vowels in Healthy Voices,” *Procedia Technology 16*, 1228–1237. 2014. Article(CrossRefLink)

[7] P. J. Murphy, “Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals,” *The Journal of the Acoustical Society of America 107*, 978–988. 2000. Article(CrossRefLink)

[8] J. P. Teixeira, C. Oliveira, and C. Lopes, “Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters,” *Procedia Technology 9*, 1112–1122. 2013. Article(CrossRefLink)

[9] I. Smits, P. Ceuppens, and M. S. D. Bodt, “A Comparative Study of Acoustic Voice Measurements by Means of Dr. Speech and Computerized Speech Lab,” *Journal of Voice 19*, 187–196. 2005. Article(CrossRefLink)

[10] F. B. Núñez, R. M. González, M. G. Peláez, I. L. González, M. F. Fernández, and M. G. Morato, “Acoustic voice analysis using the Praat program: comparative study with the Dr. Speech program,” *Acta Otorrinolaringol Esp 65*, 170–176, 2014. Article(CrossRefLink)

[11] H. Oğuz, M. A. Kiliç, and M. A. Şafak, “Comparison of results in two acoustic analysis programs: Praat and MDVP,” *Turk J Med Sci 41*, 835–841, 2011. Article(CrossRefLink)

[12] Vogel, A. P. & Maruff, P. “Comparison of voice acquisition methodologies in speech research,” *Behavior Research Methods 40*, 982–987. 2008. Article(CrossRefLink)

[13] A. Lovato, W.D. Colle, L. Giacomelli, A. Piacente, L. Righetto, G. Marioni, C. Filippis, "Multi-Dimensional Voice Program (MDVP) vs Praat for Assessing Euphonic Subjects: A Preliminary Study on the Gender-discriminating Power of Acoustic Analysis Software," *Journal of Voice 30*, 765.e1-765.e5. 2016. Article(CrossRefLink)

[14] G. Biswas, J. B. Weinberg, and D. H. Fisher, "ITERATE: a conceptual clustering algorithm for data mining," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 28*, 219–230. 1998. Article(CrossRefLink)

[15] D. H. Fisher, "Knowledge Acquisition Via Incremental Conceptual Clustering," *Machine Learning 2*, 139–172. 1987. Article(CrossRefLink)

[16] M.K. Christmann, A.R. Brancalioni, C.R. Freitas, D.Z. Vargas, M. Keske-Soares, C.L. Mezzomo, and H.B. Mota, "Use of the program MDVP in different contexts: a literature review," *Revista CEFAC 17*, 1341–1349. 2015. Article(CrossRefLink)

[17] P. Campisi, T. L. Tewfik, J. J. Manoukian, M. D. Schloss, E. Pelland-Blais, and N. Sadeghi, "Computer-Assisted Voice Analysis: Establishing a Pediatric Database," *Arch Otolaryngol Head Neck Surg*, vol. 128, no. 2, pp. 156–160, Feb. 2002. Article(CrossRefLink)

[18] "Dr. Speech Software." [Online]. Available: Article(CrossRefLink). [Accessed: 25-Feb-2018].

[19] "Praat: doing Phonetics by Computer." [Online]. Available: Article(CrossRefLink). [Accessed: 25-Feb-2018].

[20] "CSpeech Analysis Software." [Online]. Available: Article(CrossRefLink). [Accessed: 25-Feb-2018].

[21] KayPentax. Software instruction manual: Multi-Dimensional Voice Program(MDVP) Model 5105. (KayPentax, 2008).

[22] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: a tutorial". Computer 29, 31–44. 1996. Article(CrossRefLink)

[23] M. M. Adankon, and M. Cheriet, "Support Vector Machine," *Encyclopedia of Biometrics*, 1303–1308. Springer, Boston, MA, 2009. Article(CrossRefLink)

[24] S. R. Safavian, and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics 21*, 660–674 1991. Article(CrossRefLink)

[25] M.-L. Zhang, and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *Proc. of 2005 IEEE International Conference on Granular Computing 2*, 718–721 Vol. 2. 2005. Article(CrossRefLink)

[26] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: A New Data Clustering Algorithm and Its Applications," *Data Mining and Knowledge Discovery 1*, 141–182. 1997. Article(CrossRefLink)

[27] P. J. Grother, G. T. Candela, and J. L. Blue, "Fast implementations of nearest neighbor classifiers," *Pattern Recognition 30*, 459–465. 1997. Article(CrossRefLink)

**Keonsoo Lee** received the M.S. and Ph.D. degrees in computer engineering from Ajou University, Korea, in 2004 and 2013, respectively. He is currently a Research Professor at the Convergence Institute of Medical Information Communication Technology and Management, Soonchunhyang University, Asan, Korea. His research area includes artificial intelligence, knowledge representation, and multi-agent system

**Chanki Moon** received the B.S. degrees in computer science engineering from Soonchunyhang University, Korea in 2017. In 2018, he is currently pursuing the M.S. degree in alma mater.

**Yunyoung Nam** received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, Korea in 2001, 2003, and 2007 respectively. He was a Senior Researcher in the Center of Excellence in Ubiquitous System (CUS) from 2007 to 2010. He was a Research Professor in Ajou University from 2010 to 2011. He also spent time as a postdoctoral researcher at Center of Excellence for Wireless & Information Technology (CEWIT), Stony Brook University, New York from 2009 to 2013. He was a Postdoctoral Fellow at Worcester Polytechnic Institute, Massachusetts from 2013 to 2014. He is a director at ICT Convergence Rehabilitation Engineering Research Center at Soonchunhyang University from 2017. He is currently an assistant professor in the Department of Computer Science and Engineering at Soonchunhyang University. His research interests include multimedia database, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare system.