Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.024734
Article

Tech Science Press

Voice Response Questionnaire System for Speaker Recognition Using Biometric Authentication Interface

Chang-Yi Kao[1] and Hao-En Chueh[2,*]

[1]Soochow University, Taipei City, 100, Taiwan
[2]Chung Yuan Christian University, Taoyuan City, 32023, Taiwan
*Corresponding Author: Hao-En Chueh. Email: hechueh@cycu.edu.tw
Received: 29 October 2021; Accepted: 02 December 2021

Abstract: The use of voice to perform biometric authentication is an important technological development, because it is a non-invasive identification method and does not require special hardware, so it is less likely to arouse user disgust. This study tries to apply the voice recognition technology to the speech-driven interactive voice response questionnaire system aiming to upgrade the traditional speech system to an intelligent voice response questionnaire network so that the new device may offer enterprises more precise data for customer relationship management (CRM). The intelligence-type voice response gadget is becoming a new mobile channel at the current time, with functions of the questionnaire to be built in for the convenience of collecting information on local preferences that can be used for localized promotion and publicity. Authors of this study propose a framework using voice recognition and intelligent analysis models to identify target customers through voice messages gathered in the voice response questionnaire system; that is, transforming the traditional speech system to an intelligent voice complex. The speaker recognition system discussed here employs volume as the acoustic feature in endpoint detection as the computation load is usually low in this method. To correct two types of errors found in the endpoint detection practice because of ambient noise, this study suggests ways to improve the situation. First, to reach high accuracy, this study follows a dynamic time warping (DTW) based method to gain speaker identification. Second, it is devoted to avoiding any errors in endpoint detection by filtering noise from voice signals before getting recognition and deleting any test utterances that might negatively affect the results of recognition. It is hoped that by so doing the recognition rate is improved. According to the experimental results, the method proposed in this research has a high recognition rate, whether it is on personal-level or industrial-level computers, and can reach the practical application standard. Therefore, the voice management system in this research can be regarded as Virtual customer service staff to use.

Keywords: Biometric authentication; customer relationship management; speaker recognition; questionnaire

1 Introduction

The communication between people is mainly through the exchange of information with each other through language and body, while the communication between people and machines must be achieved through image recognition and voice recognition. According to related studies, in the way of communication between humans and machines, most users have different acceptance levels for image recognition and voice recognition. As far as customer service is concerned, in the past mainly relying on traditional customer service personnel, but when enterprises pay more and more attention to customer service, another customer service method-call center has emerged. When customer service is more and more important, call center is becoming the best way for companies to service their customers. In the past, service personnel in call centers responded to user needs via telephone, but the current trend is to introduce a system that can use buttons or voice to serve users, and provide users with further services based on user operations or voice.

In the E-questionnaire (electronic ways to do questions) [1,2] system, speaker recognition offers lots of advantages and its existence is thereby justified. Voice recognition is a kind of biometric authentication technology that does not need to be carried, contacted, nor will it be forgotten, with a low cost in introduction. Currently the recognition rate of speech identification stands at more than 90%, although it can be affected by noise, and the accuracy degree of sound print identification has reached 80%.The accuracy degree increases to over 98% when the two are multiplied: $1 - (1 - 0.9) * (1 - 0.8) = 98\%$. If a password is provided via voice, the password content is "something U Know" and the speaker recognition is "Something U Are".

2 Related Works

Research on image recognition has developed from a simple face recognition [3,4] to gender determination [5–7], age verification [8,9], or the application of services available on vehicles, like tests for dizziness, as well as the study on the evaluation of emotional response in e-learning [10]. Regardless of the accuracy rate in image recognition, data of the user in the distance can be collected through the phone voice answer, a practice that is able to expand the application of image as far as the application of the questionnaire system is concerned. Speech recognition is an important field in biometric recognition technology. It is a method of recognizing living beings by a number of intrinsic physical or behavioral characteristics.

After assessment is made, we find the following benefits in the speech interactive voice response questionnaire:

- According to reports released by the International Biometrics Group, the global revenue brought in by the biometrics techniques increased sharply, with the value in 2007 standing at US$3.7 billion and that by 2010 reaching US$5.7 billion.
- The great advantage of speech input lies in the fact that it matches human behavior.
- It's uneasy to collect and apply such private biometric data as finger print or sound print.
- Cost for iris recognition equipment is high, and the practice is uneasy for dissemination. Being too close in conducting recognition may result in uncertain identification.
- The success rate for face recognition is not high, as photos can cheat on the process.
- Angles of the camera may affect the results, and therefore introduction of the method to the system is impractical.

The process of speaker recognition is generally organized into two main phases, that is, the enrollment phase (training) and the identification phase (testing).The first phase involves training in the establishment of a pattern for each individual speaker to be based on their speech traits. The second phase concerns testing of the identification through comparison between the established pattern of a speaker and the actual voice of the speaker when he or she calls. As the voice traits of a speaker may refer to the identity of that speaker, the

selection of characteristic parameters is rather important and requirements for rejection must be imposed by setting a benchmark index in accordance with the voice traits.

There are now several technologies for making the identification of a speaker [11–14], including vector quantization (VQ), DTW, artificial neural networks (ANN), hidden Markov model (HMM), nearest neighbor rule (NNR), and Gaussian mixture model (GMM). The text dependent type of the speaker recognition system needs less speech material than the text independent type of the same system in turning out an effective pattern. Therefore, the technologies involved for the most suitable system have to depend on the amount of speech material and requirements for the system.

In terms of the selection of the characteristic parameters, each speaker has its own voice traits, and the parameters able to differentiate characteristics of the speech are picked to deal with the speaker recognition practice. As human voice traits change over time, they are called time-varying signals, which cannot be analyzed by a long-term based linear non time-varying method. Therefore, acoustic information is presented only by the short-term spectrum characteristics. In the second phase of the identification, this study adopts Mel-frequency cepstral coefficients (MFCCs) [15,16], the most commonly used parameter at the present time. As to the phase of the setting of a benchmark index, this study uses four characteristic parameters–average volume, average pitch, average level of clarity, and the number of audio frames.

In MFCCs, more low frequencies than high frequencies are adopted as human ears are more sensitive to sounds of low frequencies and less sensitive to those of high frequencies. Sound volume refers to strength, power, or energy of sound. An analogy can be drawn depending on the signal amplitude within an audio frame. As far as speaker identification is concerned, the volume of test utterances may affect the accuracy degree of recognition. If the volume is too high, it will cause a sonic boom. If the volume is too low, it will not be able to verify the utterances, nor can it pinpoint the exact starting and ending points when pursuing the endpoint detection. Pitch refers to the vibration frequency of the audio signal [17]. In simple terms, the pitch is the fundamental frequency of the audio signal, which is also the reciprocal of the fundamental frequency period. When the sound signal is stable, the average person can easily observe the fundamental frequency period. The number of audio frames refers to the length of utterance fragments. As far as speaker identification is concerned, the low number of the frames is likely to make errors in recognition in that the characteristic vector is also small. Using DTW to collate this vector is therefore difficult to identify the speaker. In addition, when the test utterances are found silent and the number of audio frames is found zero, these utterances can be rejected in advance. So the reason that the benchmark index is set lower than the number of audio frames is to increase the recognition rate if the number of frames is taken as a characteristic parameter.

3 Proposed Method

3.1 Identification Model

The aim of this study is not to improve speech recognition or speaker recognition technologies, but to prove that a learning framework [18] is applicable to the classification and ranking of speaker recognition. To manipulate the voice recognition, it should be to normalize the voice signal first.

And there is noise in the voice signal; it has to take end point detection just show as Fig. 1:

Clear demarcation of the scope to the consonant parts and vowels part [19,20], and it will also enhance the recognition rate. It is shown as Fig. 2.

Then, it has to use MFCCs to get the voice feature. In the areas of speech recognition and speaker recognition, the most commonly used voice feature is the MFCCs. It takes into consideration the fact that human ears feel differently toward different pitches, thus is especially suitable for speech recognition.

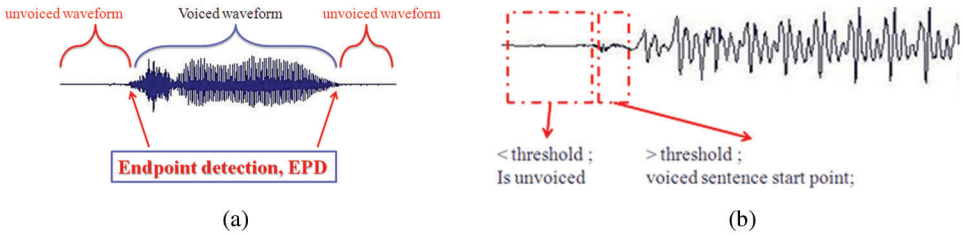

Figure 1: (a) The unvoiced feature detection; (b) the unvoiced feature detection

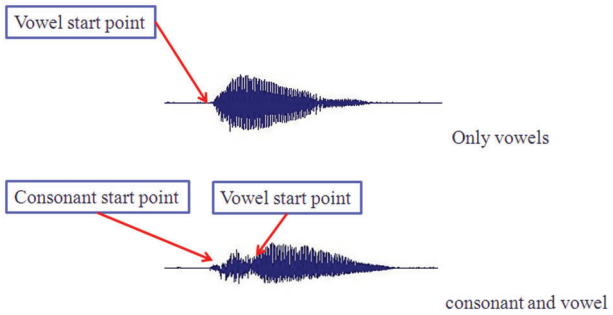

Figure 2: Consonants and vowels endpoint detection

The Mel scale shows that human ears perceive different frequencies of sound f following a logarithmic scale:

- Human ears are more sensitive to low frequency sounds.
- As the frequency increases, human ears' sensitivity decreases.

The processing procedures of MFCCs are briefly described and depicted below Fig. 3:

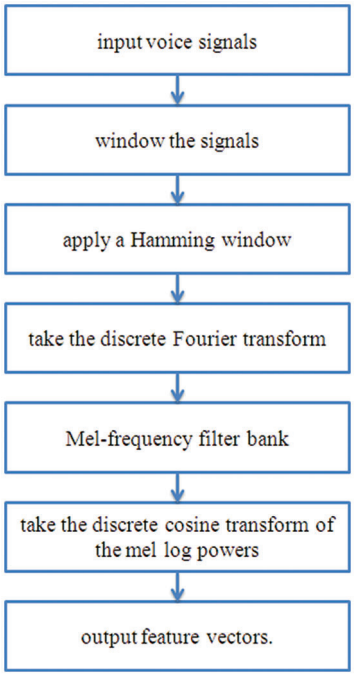

Figure 3: The processing procedures of MFCCs

We will first talk about the selection of speech characteristics and then the comparison of similarities. About the pick of speech traits, the most commonly used is MFCCs, which is also adopted for speaker identification. MFCCs can allow for better representation of human sound and is therefore easier for arithmetic of characteristics. Since the topic of MFCCs is available in many textbooks, there is no need for further discussion here. However, what should not be ignored is that during the process of MFCCs recognition results can be different because of different parameters.

DTW is a DP (Dynamic Programming) based model, which is able to shorten the time for search and comparison. Based on the theory of DP, we can describe the method in four steps:

Step1. Definition of the target function: Define $D(i, j)$ as the DTW distance between $t(1{:}i)$ and $r(1{:}j)$ with the best path in correspondence being from $(1,1)$ to (i, j).

Step2. Recurrence relations of the target function: $D(i, j) = |t(i) - r(j)| + \min\{D(i-1, j), D(i-1, j-1), D(i, j-1)\}$

Step3. Endpoint condition: $D(1, 1) = |t(1) - r(1)|$

Step4. Final solution: $D(m, n)$

In vector space, Euclidean distance is often used to estimate the distance between two points. In the case here, two MFCCs-based audio frame units are used to select the utterances whose distance is the shortest among the utterances in reference according to the shortest distance in total by a comparison of test utterances and each utterance in reference. The final step is to confirm the identity of the speaker.

The focus of this study is on the promotion of DTW for speaker identification, introducing the EDP (Endpoint Detection) method for the purpose of lowering the error degree. The objective of EDP is to pinpoint the start and end points of a sound. A false EDP may lead to the lowering of the recognition degree. The noise detected usually occurs briefly with a relatively long interval from the utterances, a characteristic that can be used to differentiate the noise and the message. The following steps are followed to rectify errors. Step 1, find the largest interval between sound fragments; Step 2, determine which one between the two has a shorter interval, with the fragment shorter less than 0.48 s being the noise; Step 3 and 4, repeat Step 1 and 2; and Step 4, continue until the interval of the fragments that is shorter than 0.48 s is located.

On top of that, we reject incomplete test utterances: The utterances collected in the databank of this study include some incorrectly recorded messages, a phenomenon often found in the security gate system in the real world. Incomplete utterances are composed of missing words in some sections of the messages. In this study, we try to determine in advance if the test materials are complete by checking where the missing words are, in front or back sections of the messages. The method used here is to find if there is volume in the first audio frame or the last one. If volume is detected, it means the test utterances are incomplete; otherwise, they are complete.

Different phonemes have different energies, so energy can be regarded as a key acoustic feature. The usual approach is to combine the energy of the phoneme with other Mel scale features to create the thirteenth dimensions in the MFCCs. We get 25 features in MFCCs in every frame. And other detail has shown as the Fig. 4.

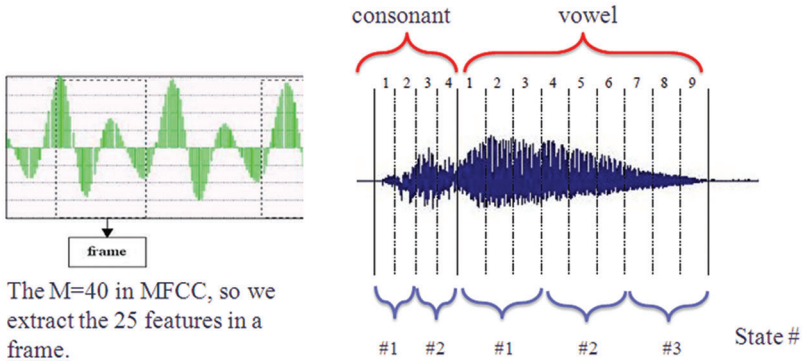

Figure 4: The consonant/vowel feature extraction

It obtains the 25 feature in a frame. It defined as the Tab. 1:

Table 1: The voice feature extraction

	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
Feature 1	$F_1^{(1)}(1)$	$F_1^{(2)}(1)$	$F_1^{(2)}(1)$	$F_1^{(2)}(1)$	$F_1^{(2)}(1)$
Feature 2	$F_1^{(1)}(2)$	$F_1^{(2)}(2)$	$F_1^{(2)}(2)$	$F_1^{(2)}(2)$	$F_1^{(2)}(2)$
Feature 3	$F_1^{(1)}(3)$	$F_1^{(2)}(3)$	$F_1^{(2)}(3)$	$F_1^{(2)}(3)$	$F_1^{(2)}(3)$
…					
Feature 25	$F_1^{(1)}(25)$	$F_1^{(2)}(25)$	$F_1^{(2)}(25)$	$F_1^{(2)}(25)$	$F_1^{(2)}(25)$

Feature 1 = (μ_1 (1), σ_1^2(1)), Feature 2 = (μ_2(2), σ_2^2(1)),…, Feature 25 = (μ_{25}(1), σ_{25}^2(1)). So, we get the two feature of the frame:

$$\mu 1\ (1) = \frac{f_1^{(1)}(1) + f_1^{(2)}(1) + f_1^{(3)}(1) + f_1^{(4)}(1) + f_1^{(5)}(1)}{5} \tag{1}$$

$$\sigma 12(1) = \frac{\left[f_1^{(1)}(1)\right]^2 + \left[f_1^{(2)}(1)\right]^2 + \left[f_1^{(3)}(1)\right]^2 + \left[f_1^{(4)}(1)\right]^2 + \left[f_1^{(5)}(1)\right]^2}{5} - [\mu 1\ (1)]^2 \tag{2}$$

We also record the relationship between states in consonant and vowel. So there are the voice features in the frame we get, the features description as follow:

State # Feature = (Feature_1, Feature_2, Feature_3, ……, Feature_25)

Inner-relation in consonant:

(State1 Features)/(State2 Features)

= (State1_Feature_1/State2_Feature_1, State1_Feature_2/State2_Feature_2, … State1_Feature_25 State2_Feature_25)

Inner-relation in vowel:

(State1 Features)/(State2 Features)

(Continued)

(continued)

State # Feature = (Feature_1, Feature_2, Feature_3, ……, Feature_25)

Inner-relation in consonant:

(State1 Features)/(State2 Features)

= (State1_Feature_1/State2_Feature_1, State1_Feature_2/State2_Feature_2, … State1_Feature_25 State2_Feature_25)

Inner-relation in vowel:

(State1 Features)/(State2 Features)

= (State1_Feature_1/State2_Feature_1, State_Feature_2/State2_Feature_2, … State1_Feature_25 State2_Feature_25)

(State2 Features)/(State2 Features)

= (State2_Feature_1/State3_Feature_1, State2_Feature_2/State3_Feature_2, … State2_Feature_25 State3_Feature_25)

$$\frac{\sum_{j=1}^{2}\sum_{i=1}^{25}\text{consonant\_State}_j(\text{feature}_i)}{\sum_{m=1}^{3}\sum_{n=1}^{25}\text{vowel\_State}_m(\text{feature}_n)} \tag{3}$$

3.2 Analysis Model

This application is for Chinese voice recognition. This study discovers the Chinese word has three types. Shown as the Fig. 5.

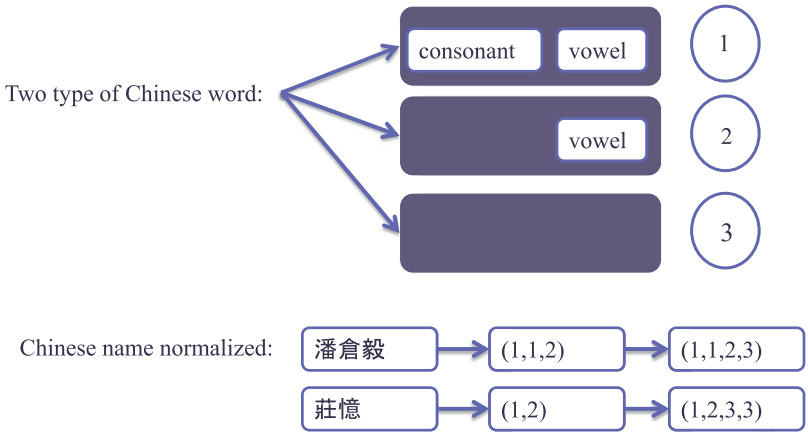

Figure 5: Chinese word types and an example

We use a learning framework for ranking to gesture recognition application. The steps of learning framework are as shown in Fig. 6:

1. Categorize data in clusters by using FCM
2. Find the center of gravity X in each cluster
3. Apply Boosting algorithm to learning to rank for each center of gravity X
4. Identify the firstly ranked cluster
5. Discard the rest of the data
6. Rewrite Boosting algorithm, and converted into a two-dimensional numerical comparison
7. Use modified Boosting algorithm to rank each point of the firstly ranked cluster
8. Rewrite the weights and error function, and deal with separately the concordant/discordant
9. And record the distance between two ranked points

Figure 6: The steps of learning framework [18]

In the existing ranking methods, we do not use learners with too strong learning ability, such as RankSVM, etc., because most learners with too strong learning ability will obtain good ranking results in a few rounds, so the boosting algorithm cannot be used. Therefore, we consider to use some learners with weak learning ability, such as some learners based on simple linear regression. This type of learner uses linear regression that minimizes errors as the ranking function, so it can be regarded as a point-by-point learning and ranking method. In addition, we adjust the weights to create a powerful learner and improve ranking accuracy.

In our proposed method, before ranking, the first stage is to cluster the data by calculating the center of gravity of each cluster. The next stage is to repeat the first stage for the cluster that is ranked first, until we can no longer classify the data. In last stage is to use the refined algorithm of boosting algorithm to rank data with RankBoost. The below Fig. 7 describes the recognition processing procedures.

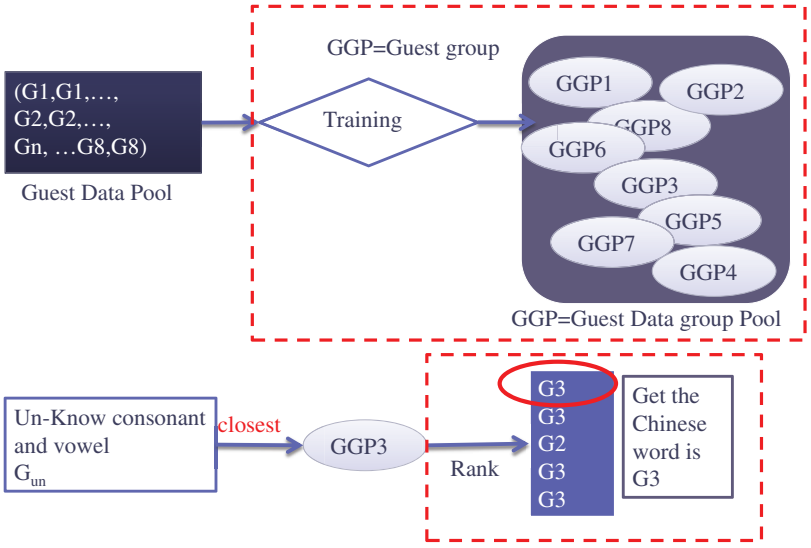

Figure 7: The questionnaires system on a multimedia machine

The learner needs to compare the ranking results of the two data sets, and try to make the ranking distance between them close to the actual distance. Therefore, in the existing ranking methods, like RankSVM etc., the learners with too strong learning ability but without the concept of cost-sensitive are

not adopted. In the next section, the experimental results of this study will be explained in detail, and the experimental results will be discussed.

4 Experience Result and Discussion

A concerned experiment has been completed in this study, with findings of which being introduced to the questionnaire system. Sound print and speech recognition is the central core of the system, thereby formulating a low cost, non-contact type of certification complex. Data used in this experiment are recorded files, and the material distribution patterns and requirements of both test and reference utterances are as follows Tab. 2:

Table 2: Recoded files in data bank of utterances

	Recorded Files	Number of People Attending Recording	Recording Time for Each Sentence	Number of Sentences for Each person	Total Number of Recorded Files
Test Utterances	First Time	80 Persons	5 Seconds	3 Sentences	1,200
Reference Utterances	Second Time	80 Persons	5 Seconds	3 Sentences	1,200

First, data must be gone through the process of characteristics selecting. Two types of trait index are determined, with one being a round sum type (can be tested on the computer) and the other a floating number type (used for speech identification on the computer). Then, an initial recognition is gained by use of DTW to calculate distance. The results: the recognition rate in the round sum type is 93% while that in the floating number type is 95.5%. The accuracy degree is different because of errors in the calculation of distance.

The following recognition results about the improvement from the false rejection in endpoint detection are from an improvement of the false acceptance in endpoint test. Be it the round sum type or the floating number type, the recognition rates are lower. Although some kept smaller segments of the test utterances are adjacent to the last utterances in the false rejection method, most of them are noise, which is not helpful to the improvement of the recognition rate. Therefore, the only choice is to adopt the false acceptance solution for the next experiment. The result is listed in the Tab. 3.

Table 3: Recognition rates from improvement

	Initial Results	Ideal Designated Endpoint	Error Lowering Rate
Round Sum Type	93.0%	94.0%	14.3%
Floating Number Type	95.5%	96.4%	20.0%

The experiment of this research is based on the multimedia navigation system installed in the exhibition hall. To conduct this research, we built a voice recognition module that can be embedded in the navigation system. This design allows visitors entering the exhibition hall to use the voice questionnaire by the way to fill in the questionnaire when using the navigation service. This built-in human-computer interaction

interface is not only convenient for visitors; it can also save a lot of manpower in the exhibition hall for issuing and filling out questionnaires.

The voice recognition system is designed to allow visitors to use a numerical code to indicate their favorite exhibits. The system will record the voice of each visitor into a voice file of less than 10 s, 16 KHz, 16-bit, and mono. After our model uses the training voice data set to learn in advance, it can compare the visitor's voice file with the training voice dataset. The comparison result is a value between 0 and 100. This value is also the confidence of the comparison result. Therefore, the larger the value, the more reliable the recognition result. In addition, the value also represents the ranking result. Therefore, the voice file with the largest value is the result identified by our model.

In order to use the voice recognition questionnaire system, visitors must first log in to the system. Therefore, the questionnaire system must also allow visitors to speak their names and identify them. In order to identify the voice file of the visitor's name, the system needs to collect at least five training voice dataset, but there is still no effective method to collect the visitor's voice data before the visitor uses the questionnaire system. In addition, although the voice recognition questionnaire system collects visitors' personal information and preference information strictly according to the Personal Information Protection Law, once the visitors log in to the questionnaire system, their identities have been exposed. Therefore, we provide an indicator information to explain and define the relationship of the exhibited items:

(Exhibited items, gender, recommended products or activities, disposable income, personal interests).

Some basic information of users must be entered in the customer relationship management system in advance, such as gender, age, annual income, and personal interests. After users select their favorite exhibits, the system will classify users according to the answers they choose. We believe that users classified into the same category by the system have the same preferences and interests, so we will recommend products and activities that suit them to users based on these preferences and interests, or tailor products or activities that meet their needs.

The analysis result of this intelligent questionnaire system is a sort of relevance, which can be used to recommend related products. In our research, by identifying speakers or callers to obtain their data, such as possible gender and age, questionnaire content, preferences and interests, and use these data to subgroup users, as a basis for the relative level of users or to recommend products suitable for users.

The subgroup method of the intelligent questionnaire system is FCM. Its application is briefly described hereafter. Let us assume that the data clustering is $X = \{x_1, x_2, \ldots, x_n\}$. In this study, it refers to gender, age, and the contents of each item in the questionnaire. The FCM procedure allows X to be divided into c group with the latter's clustering center being $V = \{v_1, v_2, \ldots, v_c\}$. When the subordinate function becomes u_{ij} in the data point x_j to the clustering center v_i, then the subordinate matrix can be presented as:

$$U = \begin{bmatrix} u_{11} & u_{12} & \ldots & u_{1n} \\ u_{21} & u_{22} & & u_{2n} \\ \ldots & & \ldots & \ldots \\ u_{m1} & u_{m2} & \ldots & u_{mn} \end{bmatrix} \tag{4}$$

And the subordinate group to the users is:

$$M_k = \frac{1}{N_k} . \sum_{j=1}^{N_k} X_{jk} \tag{5}$$

The foregoing analysis leads us to the answer to the group categories of the users. To believe that people in the same group have the same preferences. We recommend something similar to the same group of people.

After checking the rejecting benchmark of the individual characteristic parameters and the incomplete test utterance, the system rejects the float number type without matching the benchmark, the system identified 12 audio files incorrectly. Due to the rejection rate setting, audio files with low volume, fewer audio frames, and unclear speech fragments will be recognized incorrectly. The test results show that the recognition rate of the intelligent questionnaire system using floating-point feature parameters for speech recognition is 96.4%. After completing the test, we built an intelligent platform based on the intelligent questionnaire system. This intelligent platform adopts the voice recognition technology, and users do not need to enter personal information on the screen, nor do they need to sense any identification cards to perform identity recognition.

Finally, we surveyed the satisfaction level of users. The NPS (Net Promoter Score) of user satisfaction is more than 23.2%. An NPS greater than or equal to 10% indicates that the service is quite good.

The method we proposed in this research can be effectively applied to related identification systems. The user's data can be obtained through microphones and other related devices, and the data can be used to further identify the user. It is a low-cost and does not require identification cards or account-password method.

5 Conclusion

Except for its uniqueness and validity period, most features of speaker recognition are highly valued. These two shortcomings can be improved by the above methods. The use of human voice for biometric identification is a very futuristic technology. Since the voice can be easily obtained without the use of advanced equipment, it is not easy to cause cognitive fear of users, because it is a low-invasive method.

In practice, if we want to implement the method we proposed in this study, we recommend that the voice management system be used to identify the speakers who answer the questionnaire, that is, the system will be treated as a virtual customer service staff. The advantage of the application system based on the boost algorithm is that it can be implemented by using the rank decision graph and deal with discrete features simultaneously. Applying various methods to optimize the recognition results, the recognition accuracy rate in personal computing can be as high as 93% or more. Such a high accuracy rate can meet the standards of commercial applications. Compare the user's characteristic data with the known audio data, and cooperate with the classification method to optimize the comparison result, and then recommend the product or information according to the optimized result. The experimental results of this research confirm that the speaker recognition accuracy rate can reach 96.4% by using feature parameters of floating number type. But in practice, if it is to be used for commercial purposes, the recognition accuracy and computing power must be improved as much as possible. Although the use of human voice for biometric identification is a pioneering technology, it is also a future trend. But first of all, it is not easy to increase the average recognition accuracy rate of speech recognition to over 96.4%. Furthermore, because we use the ranking data closest to the query as the recognition result. Speech recognition and speech recognition in a noisy environment is also a big challenge. It is also the aim of the future research.

Acknowledgement: The authors thank to the 1,200 users of the questionnaire system who were willing to provide analysis for this study.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Zhou, Y. Goto and J. D. Cheng, "QSL: A specification language for E-questionnaire, E-testing, and E-voting systems," *IEICE TRANSACTIONS on Information and Systems*, vol. E102-D, no. 11, pp. 2159–2175, 2019.

[2] L. Matosas-Lopez and B. Garcia-Sanchez, "Benefits in the distribution of evaluation of teaching web questionnaires through SMS messaging in the university context: Participation rates, investment of time when completing the questionnaire and data collection periods," *Revista Complutense de Educación*, vol. 30, no. 3, pp. 831–845, 2019.

[3] K. Jonsson, J. Matas, J. Kittler and Y. Li, "Learning support vectors for face verification and recognition," in *Proc. Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Grenoble, France, pp. 208–213, 2000.

[4] T. Moriyama, T. Kanade, J. Xiao and J. F. Cohn, "Meticulously detailed eye region model and its application to analysis of facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 738–752, 2006.

[5] C. Shan, S. Gong and P. W. McOwan, "Fusing gait and face cues for human gender recognition," *Neurocomputing*, vol. 71, no. 10–12, pp. 1931–1938, 2008.

[6] K. Balci and V. Atalay, "PCA for gender estimation: Which eigenvectors contribute," in *Int. Conf. on Pattern Recognition*, vol. 3, pp. 363–366, 2002.

[7] R. Iga, K. Izumi and H. Hayashi, "Gender and age estimation system from face images," in *SICE Annual Conf.*, Fukui, Japan, pp. 756–761, 2003.

[8] R. H. Flin, "Age effect in children's memory for unfamiliar faces," *Developmental Psychology*, vol. 16, no. 4, pp. 373–374, 1980.

[9] R. P. Hobson, "The autistic children's recognition of age-and-sex-related characteristics of people," *Journal of Autism and Developmental Disorders*, vol. 17, no. 1, pp. 63–79, 1997.

[10] C. S. Fahn, H. M. Wu and C. Y. Kao, "Real-time facial expression recognition in image sequences using an Adaboost-based multi-classifier," in *Proc. of APSIPA, 2009 on 3D Synthesis and Expression*, Sapporo, Japan, pp. 8–17, 2009.

[11] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[12] C. Wutiwiwatchai, V. Achariyakulporn and C. Tanprasert, "Text-dependent speaker identification using LPC and DTW for Thai language," in *TENCON 99. Proc. of the IEEE Region 10 Conf.*, Cheju, Korea, pp. 674–677, 1999.

[13] H. R. Lee, C. Chen and J. S. R. Jang, "Approximate lower-bounding functions for the Speedup of DTW for melody recognition," in *The 9th IEEE Int. Workshop on Cellular Neural Networks and their Applications*, Hsinchu, Taiwan, pp. 178–181, 2005.

[14] J. S. R. Jang, C. L. Hsu and H. R. Lee, "Continuous HMM and its enhancement for singing/humming query Retrieval," in *Int. Symp. on Music Information Retrieval*, London, UK, pp. 546–551, 2005.

[15] Z. Qawaqneh, A. A. Mallouh and B. D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," *Knowledge-Based Systems*, vol. 115, pp. 5–14, 2017.

[16] Z. Tufekci and G. Disken, "Scale-invariant MFCCs for speech/speaker recognition," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, no. 5, pp. 3758–3762, 2019.

[17] P. McLeod and G. Wyvill, "A smarter way to find pitch," in *The 30th Annual Int. Computer Music Conference*, Coral Gables, Florida, pp. 138–141, 2005.

[18] C. S. Fahn and C. Y. Kao, "A multi-stage learning framework for intelligent system," *Expert System with Applications*, vol. 40, no. 9, pp. 3378–3388, 2013.

[19] J. Wang, J. R. Green, A. Samal and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech Language and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[20] T. Nazzi and A. Cutler, "How consonants and vowels shape spoken-language recognition," *Annual Review of Linguistics*, vol. 5, no. 5, pp. 25–47, 2019.