2017 IEEE 31st International Conference on Advanced Information Networking and Applications

An MFCC-based Speaker Identification System

Fang-Yie Leu, Guan-Liang Lin

Computer Science Department, TungHai University, Taiwan
{leufy, g03350017}@thu.edu.tw

***Abstract*— Nowadays, many speech recognition applications have been used by people in the world. Typical examples are the SIRI of iPhone, Google speech recognition system, and mobile phones operated by voice, etc. On the contrary, speaker identification in its current stage is relatively immature. Therefore, in this paper, we study a speaker identification technique which first takes the original voice signals of a person, e.g., Bob, and then normalizes the audio energies of the signals. After that, the audio signals is converted from time domain to frequency domain by employing Fourier transformation approach. Next, a MFCC-based human auditory filtering model is utilized to identify the energy levels of different frequencies as the quantified characteristics of Bob's voice. Further, the probability density function of Gaussian mixture model is utilized to indicate the distribution of the quantified characteristics as Bob's specific acoustic model. When receiving an unknown person, e.g., x's voice, the system processes the voice with the same procedure, and compares the processing result, which is x's acoustic model, with known-people's acoustic models collected in an acoustic-model database beforehand to identify who the most possible speaker is.**

Keywords*— *speaker identification, Fourier transformation, Mel-frequency cepstral coefficients, Gaussian mixture model, acoustic model

I. Introduction

In this information era, many high-tech products gradually enter our everyday lives, and change our living habits and patterns significantly. On the other hand, the high-tech continues to evolve toward more human-oriented ones. The biometrics identification technology which provides us with easier and more convenient methods to identify specific people has gradually replaced some existing authentication techniques, that need to be learned before people can operate them properly. The face recognition systems used at airport halls [1], and the voice assistant SIRI of iPhone [2], are two examples of the biometric identification mechanisms.

On the one hand, sound has been the most direct way for people to express something, communicate with others and do something for interaction. People invented telephones, which started from the home phone, then evaluating to the next generation, called functional phone, and at last the current smart phone. No matter how their functions and shapes are changed, the fact that people use voice to deliver information and communicate with others, has not been changed. In fact, sound is easiest and most convenient way for people to transmit their messages. Therefore, identifying people's identities from user's voice dialogue and dialogue contents, and then providing the corresponding services should be a better practical way to improve and convene our everyday lives.

However, up to present, the voice recognition technology has been well developed, and the speech recognition technology [3] is relatively matured and has been applied to our living activities. But the speaker identification technology [4] is still far away from its formal practical. The reasons are that 1) there are too many parameters needed to be processed for speaker identification; 2) it is hard to collect voice features completely; 3) the identification process is complicated and takes a long time for calculation; 4) it is difficult to be applied to those applications which need immediate response. Nowadays, the studies of speaker identification are partial, rather than a whole. For example, Hidden Markov Model Toolkit (HTK) Speech Recognition Toolkit [5], Kaldi Speech Recognition Toolkit [6], and so on, individually focus on different portions of speech recognition. Therefore, in this study, we intend to implement a practical system, which will improve and integrate several existing partial techniques/subsystems to make them as a whole so as to bring more convenience to people's lives.

The rest of this paper is organized as follows. Sections 2 and 3 describe related works and background of this study. Section 4 introduces our system architecture. The system implementation and test on this paper is presented in Section 5. Section 6 concludes this study.

1550-445X/17 $31.00 © 2017 IEEE
DOI 10.1109/AINA.2017.130

II. Related Work

2.1 Voice Recognition Systems

Voice recognition technology can be roughly divided into two sub-areas: speech recognition and speaker recognition. The former is to analyze the content of the words/speech spoken by a speaker, whereas the latter is to identify who the speaker is. As mentioned above, the speech recognition has been relatively mature. Some applications can be found in the market. But the speaker recognition is far away from mature. This study focuses on the latter, which can be roughly divided into two parts: Speaker Identification and Speaker Verification [7]. The former is that when receiving voice signals from an unknown user, e.g., from u, a speaker recognition system will find out the most likely and possible speakers from a set of known speakers $S = \{x_1, x_2, x_3 \ldots x_n\}$, by comparing u and all x_is, $x_i \in S, 1 \leq i \leq n$. On the other hand, the speaker verification is to determine the probability that a speaker u is really the speaker x_i in S as a verification. In fact, the above-mentioned speaker identification system verifies u for each x_i in S, chooses the one with the highest probability, such as x_j, and concludes that the probability that u is most likely x_i.

The speaker identification systems can be divided into two types according to the words pronounced or speech given, i.e., the context of the voice itself, including text-dependent [8] and text-independent [9]. The former is the case in which the context is fixed to specified words or a specific speech/article. All speakers read the same words or sentences, then the identification system records the voice signals of these words/sentences, and extract their features with which to perform its identification. The design of such a system is relatively simple. The latter is the case in which a speaker can say something that he/she likes. The words/sentences read are unlimited. This type of system extracts speaker's pronunciation features for modeling and later similarity comparison so as to identify who the speaker is. Because the scope of words/sentences involved is wider, the design of such a system is relatively complex, and the implementation is relatively difficult. But the system flexibility is high. It is useful to the real world, and the space of its future development is wide. In fact, this type of system is more helpful to people than the type of systems with which the context they process is fixed.

2.2 The environment of a voice recognition system

Voice recognition systems can be applied to a variety of domains. But voice recognition systems are very susceptible by noise, often resulting in poor recognition rate. In the real world, different environments will generate different types of noise of different features. Generally, the sounds collected, no matter whether outdoors or indoors, usually have a certain degree of environmental noise. Researchers are considering how to reduce the effects due to environmental noise, i.e., how to increase the degree of anti-interference, so as to correctly recognize voice signals. This is also one of the key topics in the research of voice recognition. Their common feature is Spectral Subtraction(SS) [10], which superimpositions a small background sound as noises over the original voice signals. In the original voice signals, those voice components which are the same as those of the noises will be removed, so as to achieve the purpose of noise reduction. However, based on the un-recoverability of the voice signal superimposition, spectral subtraction may also destroy some spectral details in the original signal, leading to the loss of some useful information. In order to improve this deficiency, the Support Vector Machine(SVM) [11], which classifies voice features into different classes, aiming to reduce the difference between voice features of the same class to improve recognition accuracy. But this method often requires a lot of training voice, and is not conducive to a timely response system.

III. Background

3.1 Feature extraction

Voice signals in time domain change very fast and sharply. But if we transform the voice signals from time domain to frequency domain, the corresponding spectrum can be clearly shown. The spectrum is the connotative characteristics of the voice signals. On the other hand, the voice signals also have a characteristic of short time stationary [12], meaning it is stable in a short time period without changing seriously. Therefore, we can also observe the instantaneous frequency [13] of the signals from the spectrum.

To extract features from voice signals, researchers often divide the voice signals into units, each of which consists of continuous signals. A unit is a very short time period, e.g., T, the length of which is fixed. Generally, the signals in a unit is called a frame, from which we will extract features by using a feature technique. In this study, the technique is Mel-Frequency Cepstral Coefficients (MFCC) [14] which is designed based on the characteristics of human ears which have different acoustical sensitivities to the sounds of different frequencies. Mel scale, as shown in Figure 1, is a non-linear frequency scale following sensitivities of human ear on heard sounds. It is proposed by Stevens et al. [14] in 1937. As shown, the human ears are not sensitive on high-frequency sound, but are relatively sensitive on low-frequency part. The equation which converts frequency f to Mel scales is as follows [14].

$$fmel = 2595 \mathrm{x} \log(1 + f / 700) \qquad (1)$$

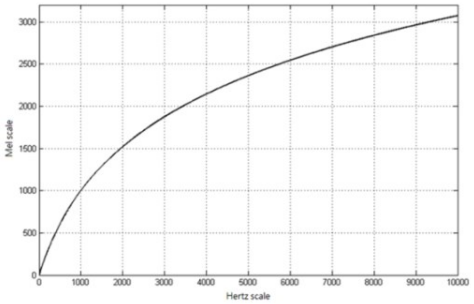

Figure 1. The Mel scale [14].

3.2 Building speaker models

After the feature extraction, the voice signals in fact are converted to a large number of feature parameters, from which we must find a suitable statistical model to describe the distribution of these feature parameters. With this model, we can compare the voice features among different persons.

In recent years, studies indicate that the energy distribution of human voice signals follows a Gaussian Model. Therefore, this paper chooses the Gaussian Mixture Model (GMM) [15] as the statistics model of these energy distribution when the text-independent speaker identification system is developed. In other words, GMM is utilized to build the feature model of a speaker, also called acoustic model of the speaker.

3.3 The speaker identification method

When the system receives an unknown speaker, e.g., x's voice signals, the speaker identification system performs the above-described procedure to process the voice signals, and compares the likelihood of x's voice features with those of all the registered and completed-learning users' GMMs collected in the acoustic-model database. The purpose is to find out the registered user, e.g., u, whose voice features are the most similar to x's. Because the likelihood between two acoustic models in signal feature space is reflected on the distance between x's and u's acoustic models. In other words, in an ideal case, the distance between two different speakers' acoustic models, e.g., λ_x and λ_u, in the signal feature space, denoted by $|\lambda_x - \lambda_u|$, will be longer than the distance between the two acoustic models established for the same speaker, i.e., $|\lambda_x - \lambda_x'|$, and $|\lambda_x - \lambda_x'| \leq |\lambda_x - \lambda_u|$, where λ_x' is an acoustical model pre-established in the acoustic-model database for x. Therefore, theoretically $|\lambda_x - \lambda_x'| = \min_{1 \leq j \leq M} \{|\lambda_x - \lambda_j|\}$ where M is the number of trained users.

IV. The system Architecture

4.1 The system introduction

Because a speaker identification system contains wide areas of techniques, it is not easy to integrate them together. The identification procedure generally comprises three steps, i.e., Feature extraction, Acoustic model establishment and Acoustic model matching. However, the steps of the identification system may be slightly changed depending on the environments in which they work on, i.e., voice transmitted through the telephones or collected in a noisy place, etc. Even this, basically they still follow the three main steps.

4.2 MFCC

In this study, we choose MFCC as the feature extraction tool to extract voice features for a speaker. Its process flow is shown in Figure 2. After the signals are received, the system partitions the signals into frames, invokes a window function to increase the continuity of voice signals in a frame, utilizes the fast Fourier transform to convert the digital signals into spectrum data, and employs the Triangular band-pass filter designed to simulate the spectral data of the human hearing. Finally, the DCT is used to quantify the spectral energy data into units of data which are able to be analyzed by the MFCC.

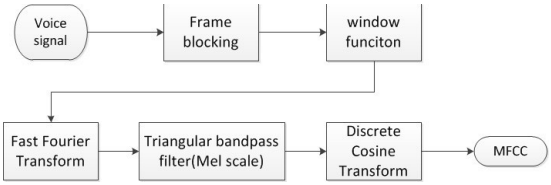

Figure 2. MFCC process

A. Framing

Framing is a voice slicing method that divides a chosen file into several time periods of fixed length, e.g., T, since the voice file F is often long. But T is very short, in which the voice signals usually are regular and continuous. That is the purpose of dividing F into many Ts is to avoid discontinuity of the signals, since signal discontinuity may lead to incorrect parameter values extraction during analysis. In general, the signals in T is called a frame and T often ranges between 20ms to 30ms. In this study, as shown in Figure 3, we choose 26ms, i.e., T = 26ms. Also, in this study, to avoid discontinuity of two consecutive frames, every two consecutive frames are mutually overlapped 13ms. That is, the signals in the second half of frame i is also the signals of the first half of frame i+1, for all is, $i = 1, 2, \ldots, n-1$ where n is the number of frames produced when dividing F. In other words, the first half of frame i and the second half of frame n of the voice file are processed only once by our scheme. The rest of the signals are processed twice since they individually appear in two consecutive frames.

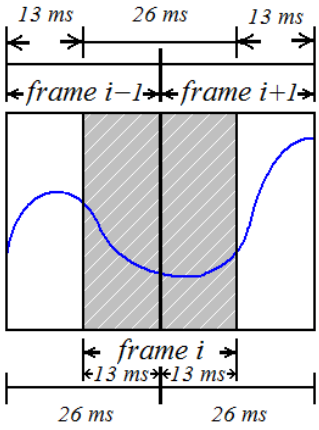

Figure 3. After Framing, each pair of consecutive frames will overlap 13ms.

B. Window function

In this study, we use the Hamming window [16] to process a frame to make all voice signals in the frame more ideally continuous where the Hamming window is a window function which is able to change the phases of voice signals to a designated range to make the voice signals more continuous. Assume that a frame of voice signals is $S(n),\ n = 0, 1, \ldots, N-1$, and $S'(n)$ is the result of Hamming window processing, where N is the total number of frames after a voice file is divided into frames. Then,

$$S'(n) = S(n) * W(n, a) \qquad (2)$$

where $W(n,a) = (1-a) - a\ \cos\frac{2\pi n}{N-1}, 0 \le n \le N-1$, is the Hamming window function, which will smooth the voice signal waveform of a frame. The head and tail of the frame will be given a greater degree of reduction so as to increase the overall signal continuity of the frame. A larger value of a causes stronger signal connectivity in a frame. Of course, the smaller the value of a, the weaker the continuity of the signals in the frame, but more signal details will be retained.

C. Triangular band-pass filter and Discrete cosine transform

This study uses a set of M triangular band-pass filters to filter voice signals after the signals are transformed into frequency-domain signals. The purpose is to make the signals follow the attenuation characteristics of the Mel scale (see Figure 1). In Figure 4, the frequency band is between 0 and 8000 (Hz). A total of 10 triangular band-pass filters is given with the low-frequency part which is given a denser band-pass filters, meaning after the filter, the energies of low-frequency voice signals are still strong, and those of the high-frequency part are relatively weaker. The sparser the triangle-band-pass filters are used, the weaker the energy of the converted signals are, and it is more sensitive (less sensitive) to the human ear when the frequency is lower (high).

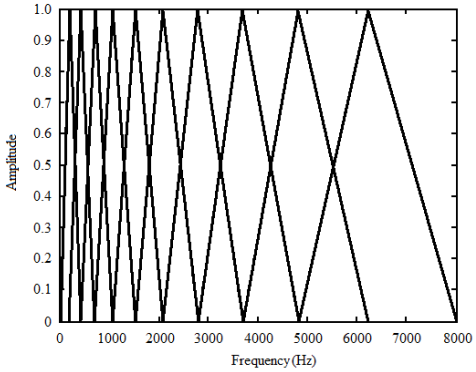

Figure 4. 10 triangular band-pass filters between 0 and 8000 (Hz) bands.

Using the Triangular band-pass filter, each filter shows its outputs with logarithm scales to represent the energy level distribution of the frequencies, and the M logarithmic energy E_k s in a frame are transformed into C_m s by using the discrete cosine function [17].

$$C_m = \sum_{k=1}^{M} E_k\ \cos\left[m\left(k - \frac{1}{2}\right)\frac{\pi}{M}\right], m = 1, \ldots, L \qquad (3)$$

Discrete cosine transform (DCT) is a transform associated with Fourier transform, similar to discrete Fourier transform, but uses only real numbers where E_k is the spectral energy value computed by one of the triangular filters in the previous step, and M is the number of triangular filters. The use of DCT conversion is expected to turn frequency-domain signals back to their time-domain ones, i.e., the cepstrum. In this study, L = 12, because the 12-dimensional feature parameter is sufficient to represent the voice feature of a frame [12].

4.3 Gaussian mixture model

In this study, the time consumed to establish the corresponding Gaussian mixture model by using training data is much longer than that of the test phase. To speed up the process of acquiring the best parameters for Gaussian mixture model, it is necessary to estimate the initial parameters fast and precisely. We use the K-means clustering to cluster the original feature vectors due to its fast, simple and good effect on clustering data. This can shorten the processing time for optimizing the parameters.

With K-means clustering approach, all given initial parameter vectors are clustered 128 groups as

the 128 Gaussian models. We compute a weight for each group and combine these 128 Gaussian models as a Gaussian mixture model, and these 128 parameter groups are the initial parameters of the Gaussian mixture model. In order to obtain the best Gaussian mixture model parameter λ, that is, the distribution of the voice features has the greatest similarity with the distribution of the model parameter λ, it is necessary to estimate the most suitable model parameter λ. Its probability density can be expressed as the follow.

$$P(X|\lambda) = \prod_{i=1}^{K} P(x_i|\lambda) \quad (4)$$

where x_i is 128 Gaussian models after clustering. X as the feature distribution of GMM is deterministic. In order to find the model parameter λ' which makes the likelihood function value of the Gaussian mixture model maximum, the EM algorithm is employed which repeatedly iterates the updating algorithm to find the best model parameters λ'. The EM algorithm will re-estimate the new model parameter λ' using the initialization parameter λ obtained in the previous K-means clustering so that $P(X|\lambda') \geq P(X|\lambda)$. Let $\lambda = \lambda'$. We continue to iteratively update the new λ until $P(X|\lambda)$ converges or reaches the upper limit of the number of times set by the system. The EM algorithm is divided into two steps, as illustrated in Figure 5, calculating the E-step of the likelihood function and updating the parameters of the model in M-step [15].

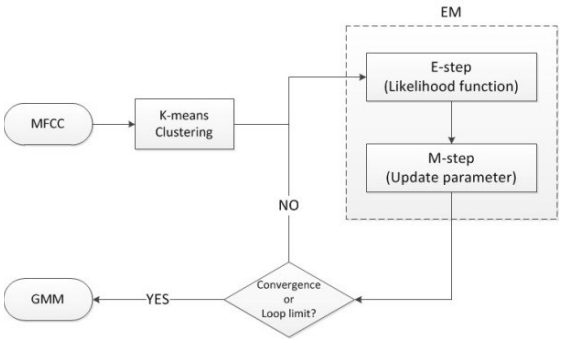

Figure 5. Gaussian mixture model of the establishment process

4.4 Bhattacharyya distance

So far, we have converted the voice signals into the GMM of the eigenvector distribution, i.e., λ, as an acoustic model of the human being. The acoustic model of each subject is established by this procedure under the assumption that a total of N user acoustic models has been collected, i.e., $U = \{\lambda_1, \lambda_1, \ldots, \lambda_N\}$, in the acoustic-model database. Now we have acquired the sound signals of a user u, process the signals with the same procedure mentioned above to obtain an acoustic model λ_u, and then calculate the Bhattacharyya distance between λ_u and λ_i, denoted by $d_{BA}(\lambda_u, \lambda_i), \lambda_i \in U, i = 1, 2, \ldots, N$. If $d_{BA}(\lambda_u, \lambda_n) = \min_{1 \leq j \leq N} \{d_{BA}(\lambda_u, \lambda_j)\}$, the probability that u is n will be the largest. But based on the conversion error and the impact of environmental noise, it is hard to ensure that u is n. In this study, we take m acoustic models with the smallest $d_{BA}(\lambda_u, \lambda_r), r = 1, 2, \ldots, m$, $m \leq N$, and sort these $d_{BA}()$s in an ascending order where

$$d_{BA}(\lambda_u, \lambda_i) = \frac{1}{8}(\mu_u - \mu_i)^T \left(\frac{\Sigma_u + \Sigma_i}{2}\right)^{-1} (\mu_u - \mu_i) + \frac{1}{2} ln \frac{\left|\frac{1}{2}(\Sigma_u + \Sigma_i)\right|}{\sqrt{|\Sigma_u| + |\Sigma_i|}} \quad (5)$$

in which $\mu_u(\mu_i)$ is the average vector of $\lambda_u(\lambda_i)$, and $\Sigma_u(\Sigma_i)$ is the covariance matrix of $\lambda_u(\lambda_i)$. We hope u will be one of the m users with the smallest $d_{BA}()$. The similarity between λ_u and λ_i is defined as arbitrary

$$\text{Similarity (\%)} = 1 - x/y \quad (6)$$

where $0 \leq \text{similarity} \leq 1$, x is the distance between λ_u and λ_i, and y is the largest distance between arbitrary two acoustic models in the acoustic-model database. The larger the similarity, the higher the probability that the person of λ_u is the person of λ_i. Also, when the similarity between λ_u and λ_i is lower than the predefined threshold, the person of λ_i will not be the person of λ_u.

V. The system Implementation and Test

Many tools are available for speech feature extraction or speech modeling. But most focus on sentence recognition. The famous HTK as a statement recognition tool is developed based on fixed sentence-voice recognition. ALIZE as a speech identification tool designed based on C++ development tools is used to build GMMs. But it lacks feature extraction and a variety of visualization capabilities. Also, the results generated by different tools are of different formats. For example, when HTK tools are invoked, the results generated by the MFCC will be stored in HTK format. The output format of the file produced by ALIZE is binary. Often the format of an output file can only be read by the respective tools. In other words, the output of these tools cannot be directly imported into existing tools that are invoked in the next stage.

This study uses the Python programming language to implement the speaker recognition system. As mentioned earlier, many existing tools or software focus on sentence recognition, or was just implemented for a sub-domain. The reason why we

choose Python programming language is that it can invoke many mathematical equations and provide many scalable libraries, e.g., SymPy, for algebra. In other words, this programming language has a certain degree of cross-platform characteristic [18]. SciPy [19] as another open-source Python library is a mathematical toolkit, with which complex mathematical operations, such as linear algebra, fast Fourier transform, etc., can be utilized. In addition, due to feature extraction, a large number of parameters are generated, Scipy is helpful. Also, the NumPy expansion library, as a high-level Python tool that supports a large number of dimensional array and matrix operations, is utilized to store the values of output parameters, so that the output formats of different steps of our system can be unified. Consequently, the characteristic parameter data generated on each step can be smoothly received by the follow-up procedures.

In order to store data generated at various stages in the database, we employ the HDF5 [20] data storage format, which as a system with typical data model, library and storage file format, is ideal for storing a large number of scientific observed data. As a result of using established standard data storage format to solve the problem of producing different data types and formats on different stages, the developers of a system are often allowed to carry out their research and development projects to invoke our system, i.e., to scale up or improve the function of speaker identification. Our system was tested on PC, the specifications of which are shown in Table 1.

Table 1. Tested hardware and software specifications.

Item	Description
CPU	Intel Core I5
OS	Ubuntu 14.04.5 LTS
Memory	4GB
Hard disk	128GB
Microphone	INTOPIC JAZZ-010
Sampling frequency	44.1 KHz
File format	16-bit linear PCM
Recording software	Audacity

A total of three experiments were performed in this study. In the first experiment, the words employed in the training and test phases are the same. In the second, words utilized in the two phases are different. The third experiment compares our system with the MFCC system [21].

The first experiment was performed in a quiet environment. Five students were invited to read the Pronunciation Guide of the oxford learner's dictionaries [22] as the training data so as to establish their acoustic models. The guide contains all the English word pronunciation. Table 2 lists the words for training.

Table 2. Training word list.

pen	bad	tea	did	cat	get	chain	jam
fall	van	thin	this	see	zoo	shoe	vision
hat	man	now	sing	leg	red	yes	wet
happy	sit	ten	father	got	saw	put	actual
too	cup	fur	about	say	go	my	boy
near	hair	pure					

During the test, the five students read the training words a total of 43 times. The accuracies of speaker identification are shown in Table 3.

Table 3. The average similarity and average accuracy on using the words listed in Table 2 as the training data.

Similarity: Data base / Tester	A	B	C	D	E	Accuracy
A	87.2%	76.1%	42.0%	36.1%	45.8%	95.3%
B	74.2%	84.4%	54.5%	28.1%	29.5%	91.7%
C	39.4%	50.6%	85.7%	55.0%	65.8%	92.7%
D	35.4%	30.5%	52.4%	88.7%	66.8%	93.0%
E	42.8%	22.7%	70.4%	65.0%	90.6%	97.7%

Because the words used in the test phase and the training phase are the same, the tones and modes of the testers on the same words will be almost the same, meaning the acoustic model contains phonological features. The acoustic models of the same speaker are similar enough so that the system can identify the testers more accurately. Although the similarity is relatively high, it cannot reach 100% because even the same words are pronounced by the same tester, the pronunciation on different times cannot be exactly the same. Therefore, the accuracies of the system will be reduced. But the similarity of the same tester is higher than those between two different testers.

In the second, we used different 40 training words listed in Table 4 to test the system without changing the acoustic models established in Experiment 1. Table 5 illustrates the test results.

As shown, accuracies illustrated in Table 5 are slightly lower than those shown in Table 3 owing to using different test and training words. The main difference is caused by different words of different pronunciations, resulting in lower accuracies, even the speaker is the same one. However, since the distribution of speech features in the feature space is actually resulted from different frequency distributions in human voice signals, when some words are not included in the training phase, our

identification system can still identify the identity of the subjects among the acoustic models.

Table 4. 40 test words which are different from those listed in Table 2.

able	advice	beauty	boot	careful
credit	cushion	date	develop	discuss
export	fear	force	grand	highligh t
itself	lamb	lively	margin	milk
out	pay	pink	positive	roof
conventio n	environmen t	increase	not	solve
often	swap	evidenc e	instructio n	send
chapter	down	idea	must	cousin

Table 5. The average similarity and average accuracy obtained given 40 words different from those listed in Table 2 during the training phase.

Similarity / Data base / Tester	A	B	C	D	E	Accuracy
A	80.2%	70.8%	52.0%	26.1%	35.8%	82.5%
B	71.2%	83.4%	64.3%	38.1%	36.5%	85.0%
C	49.4%	55.6%	75.7%	50.1%	54.8%	80.0%
D	20.5%	40.5%	49.9%	85.7%	67.9%	89.5%
E	22.8%	37.5%	51.4%	65.2%	84.6%	87.5%

In the third experiment, we compare our system with the MFCC system [21], by redoing the first and second experiments. Table 6 (Table 7) shows the results of the test using the same (different) words during the training and test phases.

Table 6. The accuracies of recognition as using the same words during the training and test phases.

system / tester	Our	MFCC
A	95.3%	80.8%
B	91.7%	81.4%
C	92.7%	82.4%
D	93.0%	85.5%
E	97.7%	85.7%

Table 7. The accuracies of using the same words during the training and test phases.

system / tester	Our	MFCC
A	82.5%	70.8%
B	85.0%	73.3%
C	80.0%	74.6%
D	89.5%	71.1%
E	87.5%	71.8%

Comparing Tables 6 and 7, it can be seen that, because the acoustic models used to describe the speakers are not established in MFCC, the system using MFCC alone does not effectively identity the distribution of voice features, resulting in the fact that the variability of accuracies is larger, no matter whether or not we use the same words in the training and test phases. The recognition rate is relatively low, since it is difficult to resist vocal talk, the sound volume, and tone changes.

VI. Conclusion and Future studies

The purpose of this study is to establish a phonetic model of the speaker, and to provide a set of practical system processes. We summarize the operation flow of the whole identification system and implement the speaker identification system with python. The reason for choosing python is that it provides many basic math-related extensions, which are extensible, user-friendly, and easy to implement on a variety of platforms. From the test results of the system, we can find that the recognition rate of the system is higher when the training voice content of the system covers the test voices. When the test voice is not included in the training phase, different words will produce different vocal patterns, and will affect the overall system identification rate. Three items extracted from our system.

1) The use of python makes a system to be integrated and expanded quite conveniently, because the data processing format of various processing steps have been unified. Thus, new development or function modifications are convenient and easier to do.

2) Because the system is developed for voice recognition, during training phase, a large number of voice training is required, causing long training time.

3) The number of vocabularies is very large, the general daily conversation, singing, speech or having a cold, etc., will greatly affect the accuracy of system identification. Therefore, it is hard to produce very high recognition rate.

Voice recognition has gradually entered human lives. It is helpful in convening people's everyday lives, such as the applications on personal computer by using voice control, data access by in putting human voice and playing games by using speech. Applications in our daily lives include life services, the automatic ticketing system, requesting human services via talk, and voice shopping services. We believe in the near future, these can bring great changes to the human society.

References

[1] C. Zhan, W. Li and P. Ogunbona, "Face Recognition from Single Sample based on Human Face Perception," *International Conference Image and Vision Computing New Zealand*, pp. 56-61, 2009.

[2] http://www.apple.com/tw/ios/siri/

[3] https://cloud.google.com/speech/

[4] D.A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," *IEEE International Conference on Acoustics, Speech,*

and Signal Processing, Vol. 4, pp. 4072-4075, 2002.
[5] http://htk.eng.cam.ac.uk/
[6] http://kaldi-asr.org/doc/about.html
[7] https://en.wikipedia.org/wiki/Speaker_recognition
[8] T. Stafylakis, M.J. Alam and P. Kenny, "Text-Dependent Speaker Recognition with Random Digit Strings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, Issue 7, pp. 1194-1203, 2016.
[9] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, Issue 1, pp. 72-83, 1995.
[10] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, Vol. 67, Issue 12, pp. 1586-1604, 1979.
[11] N. Cristianini and J. Shawe-Taylor, "Support Vector Machines," *in Cambridge University Press*, 2000.
[12] B.H. Juang and T. Chen, "The Past, Present, and Future of Speech Processing," *IEEE Signal Processing Magazine*, Vol. 15, Issue 3, pp. 23-48, 1998.
[13] N.E. Huang et al, "On Instantaneous Frequency," *in World Scientific Publishing Company*, pp. 177-229, 2009.
[14] R. Vergin, D. O'Shaughnessy and A. Farhat, "Generalized Mel Frequency Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 7, Issue 5, pp. 525-532, 1999.
[15] X. Peng, W. Xu and B. Wang, "Speaker Clustering via Novel Seudo-Divergence of Gaussian Mixture Models," *International Conference on Natural Language Processing and Knowledge Engineering*, pp. 111-114, 2005.
[16] A. Goel and A. Gupta, "Design of Satellite Payload Filter Emulator Using Hamming Window," *International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, pp. 202-205, 2014.
[17] K.R. Rao and P. Yip, "Discrete Cosine Transform: Algorithms, Advantages, Applications," *in Academic Press*, 1990.
[18] https://www.python.org/
[19] https://www.scipy.org/
[20] https://www.hdfgroup.org/
[21] J.P. Openshaw, Z.P. Sun and J.S. Mason, "A Comparison of Composite Features under Degraded Speech in Speaker Recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 371-374, 1993.
[22] http://www.oxfordlearnersdictionaries.com/us/about/pronunciation_english