International Journal of Research in Computer Science
eISSN 2249-8265 Volume 3 Issue 5 (2013) pp. 13-17
www.ijorcs.org, A Unit of White Globe Publications
doi: 10.7815/ijorcs. 35.2013.070

VOICE RECOGNITION SYSTEM USING TEMPLATE MATCHING

Luqman Gbadamosi
Computer Science Department, Lagos State Polytechnic, Lagos, Nigeria
Email: luqmangbadamosi@yahoo.com

Abstract: *It is easy for human to recognize familiar voice but using computer programs to identify a voice when compared with others is a herculean task. This is due to the problem that is encountered when developing the algorithm to recognize human voice. It is impossible to say a word the same way in two different occasions. Human speech analysis by computer gives different interpretation based on varying speed of speech delivery. This research paper gives detail description of the process behind implementation of an effective voice recognition algorithm. The algorithm utilize discrete Fourier transform to compare the frequency spectra of two voice samples because it remained unchanged as speech is slightly varied. Chebyshev inequality is then used to determine whether the two voices came from the same person. The algorithm is implemented and tested using MATLAB.*

Keywords: *chebyshev's inequality, discrete fourier transform, frequency spectra, voice recognition.*

I. INTRODUCTION

Voice Recognition or Voice Authentication is an automated method of identification of the person who is speaking by the characteristics of their voice biometrics. Voice is one of many forms of biometrics used to identify an individual and verify their identity. Naturally human can recognize a familiar voice but getting computer to do the same is more difficult task. This is due to the fact that it is impossible to say a word exactly the same way on two different occasions. Advancement in computing capabilities has led to a more effective way of recognizing human voice using feature extraction. Voice recognition system is one of the best and highly effective biometrics technique which could be used for telephone banking and forensic investigation by law enforcement agency. [9][10]

A. What is Human Voice?

The voice is made up of sound made by human being using vocal folds for talking, singing, laughing, crying, screaming etc. The human voice is specifically that part of human sound production in which the vocal folds are the primary sound source. The mechanism for generating the human voice can be subdivided into three; the lungs, the vocal folds within the larynx, and the articulators. [11]

Figure 1: The spectrogram of human voice reveals its reach harmonic content.

B. What is Voice Recognition?

Voice Recognition (sometimes referred to as Speaker Recognition) is the identification of the person who is speaking by extracting the feature of their voices when a questioned voice print is compared against a known voice print. This technology involves sounds, words or phrases spoken by humans are converted into electrical signals, and these signals are transformed into coding patterns to which meaning has been assigned. There are two major applications of voice recognition technologies and methodologies. The first is voice verification or authentication which is used to verify the speaker claims to be of a certain identity and the voice is used to verify this claim. The second is voice identification which is the task of determining an unknown speaker's identity. In a better perspective, voice verification is one to one matching where one speaker's voice is matched to one template or voice print, whereas voice identification is one to many matching where the speaker's voice is compared against many voice templates.

Speaker recognition system has two phases: Enrollment and Verification. During enrollment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print or template. In the verification phase, a speech sample or "utterance" is compared against a previously created

voice print. For identification systems, the utterance is compared against multiple voice prints in order to determine the best match while verification systems compare an utterance against a single voice print. Voice Recognition Systems can also be categorized into two: text independent and text dependent. [9]

Text-Dependent: This means text must be the same for the enrollment and verification. The use of shared-secret passwords and PINs or knowledge-based information can be employed in order to create a multi-factor authentication scenario.

Text Independent: Text-Independent systems are most often used for speaker identification as they require very little cooperation by the speaker. In this case the text used during enrollment is different from the text during verification. In fact, the enrollment may happen without the user's knowledge, as in the case for many forensic applications. [9]

C. Voice Recognition Techniques

The most common approaches to voice recognition can be divided into two classes: Template Matching and Feature Analysis.

Template Matching: Template matching is the simplest technique and has the highest accuracy when used properly, but it also suffers from the most limitations. As with any approach to voice recognition, the first step is for the user to speak a word or phrase into a microphone. The electrical signal from the microphone is digitized by an "analog-to-digital (A/D) converter", and is stored in memory. To determine the "meaning" of this voice input, the computer attempts to match the input with a digitized voice sample, or template that has a known meaning. This technique is a close analogy to the traditional command inputs from a keyboard. The program contains the input template, and attempts to match this template with the actual input using a simple conditional statement. This type of system is known as "speaker dependent." and recognition accuracy can be about 98 percent.

Feature Analysis: A more general form of voice recognition is available through feature analysis and this technique usually leads to "speaker-independent" voice recognition. Instead of trying to find an exact or near-exact match between the actual voice input and a previously stored voice template, this method first processes the voice input using "Fourier transforms" or "linear predictive coding (LPC)", then attempts to find characteristic similarities between the expected inputs and the actual digitized voice input. These similarities will be present for a wide range of speakers, and so the system need not be trained by each new user. The types of speech differences that the speaker-independent method can deal with, but which pattern matching would fail to handle, include accents, and varying speed of delivery, pitch, volume, and inflection. Speaker-independent speech recognition has proven to be very difficult, with some of the greatest hurdles being the variety of accents and inflections used by speakers of different nationalities. Recognition accuracy for speaker-independent systems is somewhat less than for speaker-dependent systems, usually between 90 and 95 percent. [12]

I have implemented template matching technique. This approach has been intensively studied and is also the back bone of most voice recognition products in the market.

II. IMPLEMENTATION

A. Design Description

The voice recognition system using template matching technique require the user to first create a template for matching comparison by first recording 10 samples of the speaker's voice by calling a phrase which is going to be the known voice. Thereafter, the questioned speaker's voice can now be recorded which would now be further analyzed using Discrete Fourier Transform.

Discrete Fourier Transform: Voice recognition in time domain would be extremely be impractical based on the difficulties explained above. Instead an analysis in frequency spectra in a voice which remain predominately unchanged as speech is slightly varied turn out to be a more viable option. The conversion of all the recording into frequency domain is done using discrete Fourier transform greatly simplified the process of comparing two recordings. [3][6]

Finding the Norm: Due to the nature of human speech all the data pertaining to frequency above 600Hz is safely discarded. Therefore, once a recording is converted into frequency domain, it could then be simply regarded as a vector in 600-dimensional Euclidean space. At this point, a comparison between two vectors could easily be carried out by normalizing the vectors (giving them length 1) then computing the norm of the difference between the two (of course, the difference between two vectors in R600 is performed by subtracting component wise). Unfortunately, exactly which norm to use is not immediately clear? After carefully comparing and contrasting the use of the Taxicab, Euclidean, and Maximum norms.[13]

It became clear that the Euclidean norm most accurately measured the closeness between different frequency spectra. Once the norm function was chosen, all that remained was to decide exactly how small the norm of the difference of two vectors had to be in order to determine that both recordings originated from the same person.

Chebyshev's Inequality: Chebyshev's inequality says that at least 1-1/K2 of data from a sample must fall within K standard deviations from the mean, where K is any positive real number greater than one. To illustrate the inequality, we will look at it for a few values of K:

- For $K = 2$ we have $1 – 1/K^2 = 1 - 1/4 = 3/4 = 75\%$. So Chebyshev's inequality says that at least 75% of the data values of any distribution must be within two standard deviations of the mean.
- For $K = 3$ we have $1 – 1/K^2 = 1 - 1/9 = 8/9 = 89\%$. So Chebyshev's inequality says that at least 89% of the data values of any distribution must be within three standard deviations of the mean.
- For $K = 4$ we have $1 – 1/K^2 = 1 - 1/16 = 15/16 = 93.75\%$. So Chebyshev's inequality says that at least 93.75% of the data values of any distribution must be within four standard deviations of the mean.[13]

Template Matching: The above analysis has revealed that Chebyshe v's Inequality states that in particular, at least 3/4 of all measurements from the same population fall within two standard deviations of the mean. Hence, in response to the problem posed at the end of the previous paragraph, the following solution can be formulated: By requiring that the norm of the difference fall within 2 standard deviations of the normal average voice, I have ensured that at least 75% of the time, the algorithm would recognize a voice correctly.

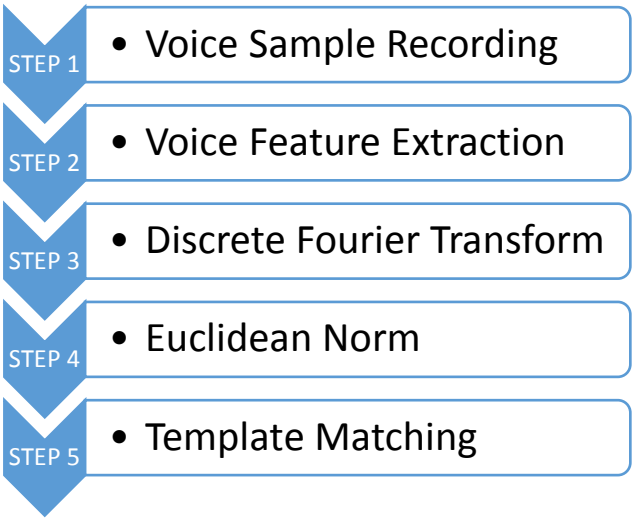

Figure 2: Detail Design Description

III. RESULTS

The performance rating of the voice recognition technique adopted would recognize the speaker's voice 75% of the time of enrollment.

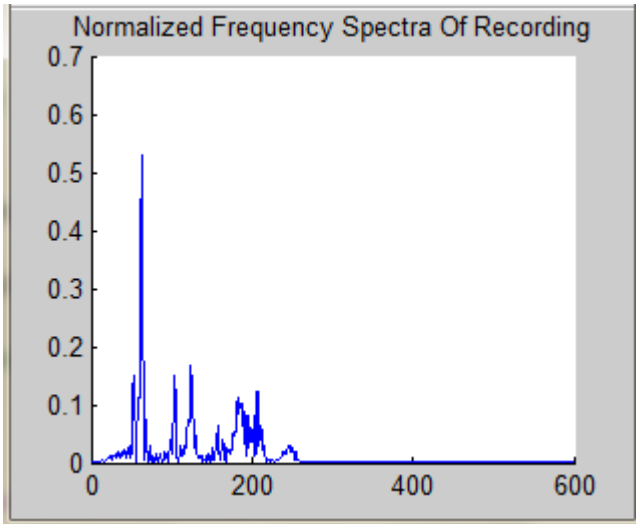

Figure 3: Graph showing normalized frequency spectra of recorded questioned voice sample

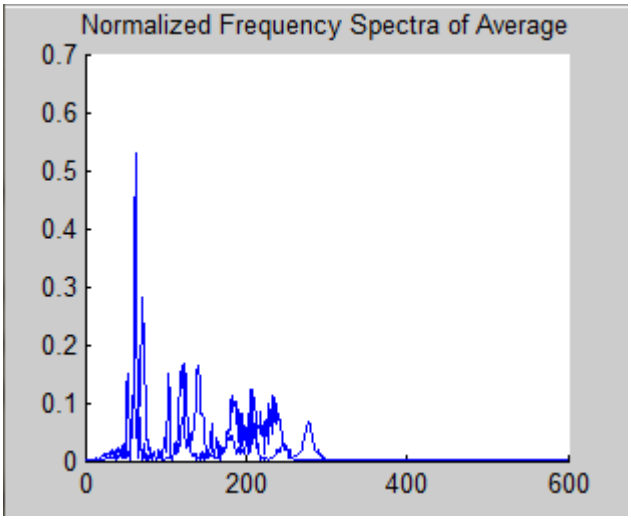

Figure 4: Graph showing normalized frequency spectra of average template voice sample.

A. Performance Evaluation Index

The indexes well accepted to determine the recognition rate of voice recognition system is endpoint detection algorithm using Zero crossing rates (ZCR) and Variable Frame Rates (VFR). This techniques involves using a clean enrollment of speech signal. The signal is recorded for 2seconds and the testing speech is polluted by additive noise at different noise decibel levels. The performance of the four endpoint algorithm has been plotted in the figure below. Three varieties of additive noise, babble noise, and F-16 noise have been used to test. Table (1-3) shows the accuracy rates. The additive noise has been taken at different levels of 20dB, 15dB, 10dB, 5dB and 0dB SNR.[15]

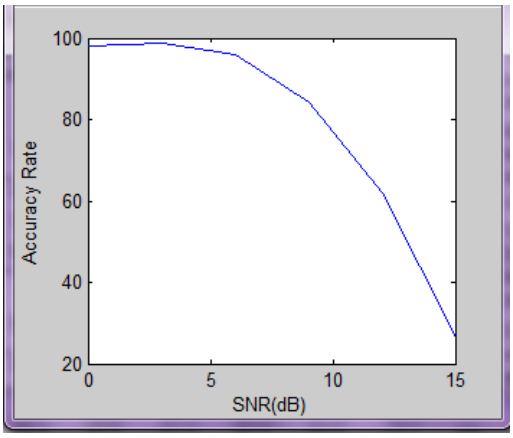

Figure 5: Factory Noise

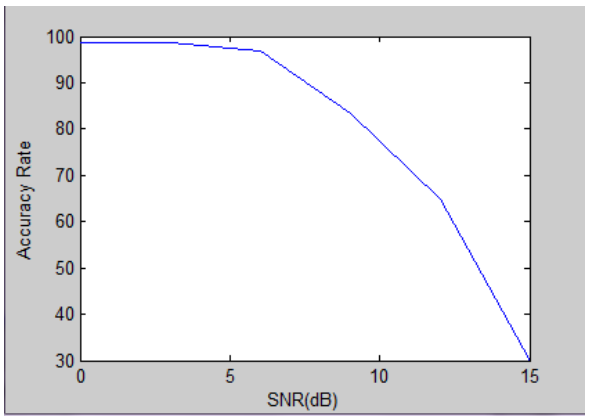

Figure 6: Babble Noise

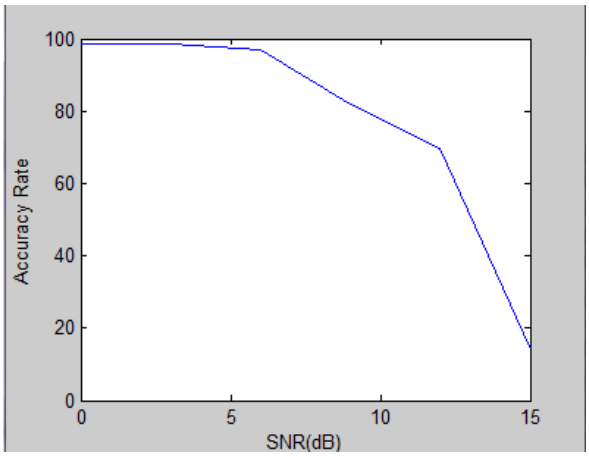

Figure 7: F-6 Noise

Table 1: Endpoint Detection (Babble Noise)

	Clean	20dB	15dB	10dB	5dB	0dB
VFR	98.0	98.6	96.0	84.3	62.0	26.6

Table 2: Endpoint Detection (Babble Noise)

	Clean	20dB	15dB	10dB	5dB	0dB
VFR	98.7	98.6	97.0	83.3	65.0	30.0

Table 3: Endpoint Detection (F-16 Noise)

	Clean	20dB	15dB	10dB	5dB	0dB
VFR	98.7	98.6	97.0	82.0	69.6	14.0

The experimental results above was derived from speech data collected from speaker using the different voice recognition algorithm. clean speech was achieved when the effect background noise and channel distortion are minimized.

The experimental results using comparative analysis of different algorithm for voice recognition at different noise levels has revealed that inaccurate endpoint detection can cause misclassification rather than other possible errors. The accuracy of endpoint detection is much higher for the algorithm which integrate both time domain and frequency domain. This has actually proven beyond any reasonable doubt that voice recognition system using template matching still remain the best algorithm for recognizing an unknown voice.

IV. CONCLUSION

The above research work implementation is an effort to understand how voice recognition is used as one of the best forms of biometric to recognize the identity of human being. It briefly describe all the stages from voice recording, voice feature extraction, discrete Fourier transform to template matching which generate a good percentage of matching score. Various standard technique are used at the intermediate stage of the processing.

Low percentage verification rate arise due to the difficulty of developing algorithm to recognize human voice as different data are obtained for voice samples recorded on different occasions. New technique and highly effective algorithm have been discovered which gives better results.

Also a major challenge is the inability of the technique to recognize a different word phrase aside from the one stored in the database during enrollment. The technique adopted only recognize human voice 70% of the time. It is highly recommended that future research work should focus on achieving up 95% recognition rate should recognize different word phrase.

V. REFERENCES

[1] Kinnunen, Tomi; Li, Haizhou. "An overview of text-independent speaker recognition: From features to super vectors". *Speech Communication* 52 (1): 12–40. doi:10.1016/j.specom.2009.08.009

[2] Homayoon Beigi, "Speaker Recognition, Biometrics / Book 1, Jucheng Yang (ed.), Intech Open Access Publisher, 2011, pp. 3-28, ISBN 978-953-307-618-8. doi: 10.1007/978-0-387-77592-0

[3] Duhamel, P. and M. Vetterli, "Fast Fourier Transforms: A Tutorial Review and a State of the Art," Signal Processing, Vol. 19, April 1990, pp. 259-299. doi: 10.1016/0165-1684(90)90158-U

[4] Oppenheim, A. V. and R. W. Schafer, “Discrete-Time Signal Processing”, Prentice-Hall, 1989, p. 611.

[5] Oppenheim, A. V. and R. W. Schafer, Discrete-Time Signal Processing, Prentice-Hall, 1989, p. 619.

[6] Rader, C. M., "Discrete Fourier Transforms when the Number of Data Samples Is Prime," Proceedings of the IEEE, Vol. 56, June 1968 (Current Version: June 2005), pp. 1107-1108. doi: 10.1109/PROC.1968.6477

[7] Oppenheim, A. V. and R.W. Schafer. Discrete-Time Signal Processing, Englewood Cliffs, NJ: Prentice-Hall, 1989, pp. 311-312.

[8] ITU-T Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," General Aspects of Digital Transmission Systems; Terminal Equipments, International Telecommunication Union (ITU), 1993.

[9] Beigi, Homayoon (2011). “ Fundamentals of Speaker Recognition.”. [Online]. Available: http:// www.wikipedia.org/wiki/speaker_recognition.

[10] Course project (Fall 2009) “Voice Recognition Using MATLAB”. California State University Northridge during the semester. [Online]. Available: http://www.cnx.org/content/m33347/1.3/module_export?format=zip

[11] “Article on Human Voice” [Online]. Available: http://www.wikipedia.org/wiki/Human voice.

[12] “Techniques of Voice Recognition System” [Online]. Available:http://www.hitl.washington.edu/scllw/EVE/I.D.2.d.VoiceRecognition.htm

[13] “Probability Tutorials on Chebyshevs-Inequality” [Online]. Available: http://www.statistics.about.com/od/probHelpandTutorials/a/Chebyshevs-Inequality.htm.

[14] Sangram Bana, Dr. Davinder Kaur, “Fingerprint Recognition System using Image Segmentation”. International Journal of Advanced Engineering Sciences and technologies Vol No. 5, Issue No. 1, 012 – 023

[15] Kapil Sharma, H.P Sinha & R.K Aggarwal “Comparative study of speech Recognition System using various feature extraction techniques”. International Journal of Information Technology and Knowledge Management Vol 3, No2, pp. 695-698

How to cite

Luqman Gbadamosi, " Voice Recognition System using Template Matching ". *International Journal of Research in Computer Science*, 3 (5): pp. 13-17, September 2013. doi: 10.7815/ijorcs. 35.2013.070