SPEAKER RECOGNITION SYSTEM WITH PITCH DETECTION ALGORITHM

Anjana parua
C.U.Shah College of engg. & technology
India
aparua@yahoo.com

ABSTRACT

A simple password is not enough to differentiate between an authorized and a fraudent person. It can be guessed or it can be stolen. So, as technology advances, the security and authentication terms are becoming more intricate and strict, and that has introduced the term 'biometric' in our technical dictionary. In this attempt a speaker recognition system has been developed using pitch detection algorithm. The use of the pitch detection algorithm has been done to differentiate between male and female voice. The categorized voice sample then subjected to extract the feature using cepstrum coefficient. Dynamic time warping method is used for the purpose of pattern matching. To evaluate the system each sample utterance of the user is compared with the rest of the utterance in the database. For each comparison, the distance measured is calculated. A lower distance indicates a higher similarity. A lower distance indicates a higher similarity.

KEYWORDS Pitch, Pitch detection algorithm, Biometrics, Speech production, Speaker recognition, Cepstrum coefficient, Auto correlation, Dynamic Time Warping (DTW).

1. Introduction

Biometrics is automated methods of recognizing individual based on their physical and behavioral characteristics. Some common examples are fingerprint, face, iris, hand geometry, voice and dynamic signature. The biometric methods of identification are preferred to traditional methods involving passwords and personal identification number (PIN) mainly for 2 reasons:

1. The person to be identified has to be physically present at the point of identification.
2. Identification based on biometric technique obviates the need to remember a password or carry a token.

The biometric system consists of three basic elements. Enrollment or the process of collecting biometric samples fro an individual, known as enrollee, and the subsequent generation of template. Second is the templates or the data representing the enrollee's biometric and the last is the matching or the process of comparing a live biometric sample against one or many templates in the system's database. In fig 1 below the block diagram of a typical biometric system is shown.

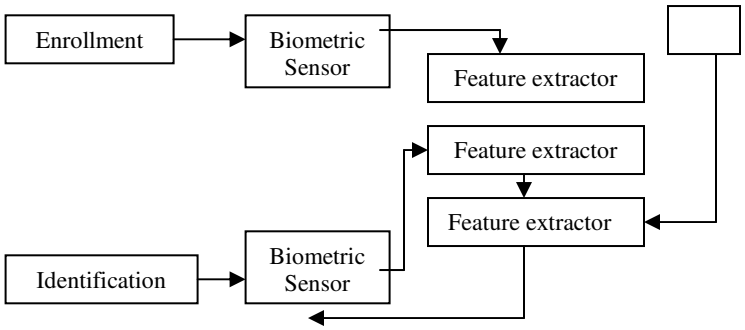

Fig 1. A biometric generic system

2. Biometric technologies

Biometric authentication system facilitates controlled access to applications, network and personal computers and physical facilities. The biometric authentication system cannot fall prey to hackers; it can't be shared, lost or guessed. Therefore it is an efficient way to replace the traditional password- based authentication system. Basically there are two types of biometric systems. One is contact biometric type and contact less biometric type.

Contact biometric technologies: because of the inherent need to make direct contact with the electronic device (scanner) in order to gain logical or physical access, reasons that this technology be known as contact biometric. The famous example of which being fingerprint scanning, hand/finger geometry, dynamic signature reading, keystroke dynamics etc.

Contact less biometric technologies: A contact less biometric can come in the form of either a passive (the biometric device continuously monitors for the correct activation frequency) or an active (the user initiates activation at will). In either event, authentication of the user voluntarily agrees to present the biometric sample. It does not require undesirable contact in order to extract the required data sample of the biological characteristics. The most probable example of this type of system being facial recognition, voice recognition, iris scans, retinal scans etc.

3. Advantages and disadvantages of biometric technologies.

Biometric technologies can be applied to application requiring logical access control, such as personal computers, networks, financial accounts etc. The biometric authentication system can be linked to the business processes of the company to increase the accountability of the financial systems, vendors and supplier transactions; the results can be extremely beneficial.

The global reach of the internet has made the services and products of the company available 24X 7, provided the customer has a user name and password to login. In case the customer forgets his/her password, then with a biometric authentication system implanted, consumers can opt to register their biometric trait.

Physical access control applications for biometric authentication systems include entry into a building, room or a safe. These systems may also be used to start motorized vehicles or linked to computer-based applications used to monitor time and attendance of the employees as they enter and leave company facilities.

The major disadvantages being people with disabilities may have problems with contact biometrics. Some times for the regular user dues to the change in the environment and other such factors, the system rejects to accept the user.

4. Introduction to speech signal, its production and classification.

Speech has evolved as a primary form of communication between humans. Nevertheless, there often occur conditions under which we measure and then transform the speech signal to another form in order to enhance our ability to communicate. At the physiological level of communication, the brain creates electric signals that move along the motor nerves; these electric signals activate muscles in the vocal tract and vocal cords. The vocal tract and vocal cord movement results in pressure changes within the vocal tract, and , in particular, at the lips, initiating a sound wave that propagates through space. The sound wave propagates through space as a chain reaction among the air particles, resulting in the pressure change at the ear canal and thus vibrating the eardrum. The vibration at the ear drum induces electric signals that move along the sensory nerves to the brain. Finally at the linguistic level of the listener, the brain performs speech recognition and understanding.

The speech organs can be primarily divided into three main groups; the lungs, larynx and vocal tract.

Lungs: one purpose of the lungs is the inhalation and exhalation of air. When we inhale, we enlarge the chest cavity by expanding the rib cage surrounding the lungs and lowering the diaphragm that sits at the bottom of the lungs and separates the lungs from the abdomen; this action lowers the air pressure in the lungs, thus causing air to rush in through the vocal tract and down the trachea into the lungs. When we exhale, we reduce the volume of the chest cavity by contracting the muscles in the rib cage, thus increasing the lung air pressure. This increase in pressure then causes air to flow through the trachea into the larynx.

In short the lungs act as a power supply unit to the speech production system.

Larynx: The larynx is the complicated system of the cartilages, muscles and ligaments whose primary purpose, in the context of the speech production, is to control the vocal cords or vocal folds. The vocal folds are two masses of flesh, ligaments and muscles, which stretch between the front and back of the larynx, as illustrated in the fig.3. The folds are about 15mm long in men and 13mm long in women.

The glottis is the slit like orifice between the two folds. The folds are fixed at the front of the larynx where they are attached to the stationary thyroid cartilage. The folds are free to move at the back and sides of the larynx. The size of the glottis is controlled in the part by the *Arytenoid Cartilages,* and in part by muscles within the fold. Another important property of the vocal folds, in addition to the size of the glottis, is their tension. The tension is controlled primarily by the muscle within the folds, as well as the cartilage around the folds.

Vocal tract: The vocal tract is comprised of the oral cavity from the larynx to the lips and nasal passage that is couples to the oral cavity by the way of the velum. One purpose of the vocal tract is to spectrally “color” the source, which is

important for making perceptually distinct speech sounds. A second purpose is to generate new sources for sound production. The vocal tract modifies the spectral contents of the speech signal as it passes through it.

5. Speaker recognition system categories.

Generally, two applications of the speaker recognition can be distinguished: if the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called speaker verification. On the other hand, speaker identification is task of determining an unknown speaker's identity. In a sense speaker verification is a 1:1 match where one speaker's voice is matched to one template whereas speaker identification is a 1: N match where the voice is matched to N templates.

Speaker recognition systems employ 2 styles of spoken input as text-dependent and text-independent. If the text is same for enrollment and test, this is called text-dependent recognition. Text-independent systems are often used for speaker recognition as they require very little if any cooperation by the speaker. In this case the text during enrollment and test is different.

Here the system is text-dependent. The whole operation can be explained as , firstly the person need to enroll himself/herself, which requires the basic information e.g. name, employee no, date of joining etc and most important a sample of his/her voice, which should be done by saying a fixed set of the texts. The voice sample goes through the pitch detection algorithm. By pitch detection algorithm either the voice is classified as female voice or male voice, depending upon that the basic information is added to female or male database. Now when the same speaker next time enter the system he/she should say the previous fixed set of text, which will again go through the pitch detection algorithm. Then the process of feature extraction and pattern matching will follow. But the pattern matching takes place from the databases defined from the pitch detection. Lastly if the claimant's pattern is matched against any of the previously defined pattern, then he/she will be accepted, otherwise will be rejected.

6. Filtering

The very first step of the system is to filter the sample of the voice which is taken inside the PC using a microphone. The whole system has been implemented using MatLab 7.0. So the filter is designed using the MatLab code. A filter is essentially a system or network that selectively changes the wave shape, amplitude-frequency and/ or phase-frequency characteristics of a signal in a desired manner. Common filtering objective are to improve the quality of a desired signal to noise ratio, to extract information from signals or to separate two or more signals previously combined to make efficient use of an available communication channel. Here since MatLab 7.0 is used to develop a filter, so the filter is obviously a digital filter.

In digital filtering two types of filters are there: IIR (infinite impulse response) and FIR (finite impulse response). This system uses the IIR low pass digital filter for removing the high frequency or we can say to lower the background noises, which can interfere in the process of recognition. The specification being, the specific cut off frequency w_p = 400Hz, the sampling frequency Fs= 7000Hz and the final stop band edge frequency w_s = 2500Hz. The order of the filter is 2 and the bilinear method of digital filter development is used.

7. Pitch detection algorithm

Pitch is not a physical parameter, but a perceptive one. The pitch can be called as 'fundamental frequency'. The voices of women and men differ relative to larynx size, the space between the vocal folds, speaking pitch and pitch range. The higher pitch of women compared to men means that the vocal folds vibrate or come together almost twice as many times per second as those of the male. The pitch range is about 60Hz to 400 Hz. To detect the pitch, a pitch detection algorithm is required. There are several ways of detection of pitch of a suitable signal. Basically the algorithms or the methods of determining the pitch are classified as Time domain approach and Frequency domain approach. The time domain approach seems more straightforward, which consists of the looking at the input signal as fluctuating amplitude in time domain and try to find the repeating pattern in the waveform that give clue as to its periodicity. The most common maximum likelihood, adaptive filters methods. The next category is the frequency domain approach, which includes the short time Fourier transform (STFT). In this system, time domain Autocorrelation method is employed to detect the pitch.

The goal of the autocorrelation routine is to find the similarity between the signal and a shifted

version of itself. If the signal is periodic, the autocorrelation function also will be, and if the signal is harmonic the autocorrelation function will have peaks in multiples of the fundamental frequency. This technique is most efficient at mid or lower frequencies. Thus it has been used in speaker recognition system where the pitch range is limited.

8. Feature extraction

Now after the pitch detection algorithm, the voice sample now is categorized into female and male voice. The next step is to extract the feature from the voice sample which will make it unique. For the feature extraction, the Cepstrum coefficient is used and then the feature is normalized using parameter domain normalization technique, in which the cepstral coefficients are averaged over the duration of an entire utterance and the averaged values subtracted from the cepstral coefficient of each frame. This procedure is done so that the feature remains stable for a long time.

9. Pattern matching

The speaker recognition system can be classified on the basis of text-dependent and text-independent system. So as this system is text-dependent system, so dynamic time warping method of pattern matching is used. Other pattern matching algorithm being Hidden Markov's Method.

The Dynamic time warping is a technique that finds the optimal alignment between two time series, if one time series may be 'warped' non-linearly by stretching and shrinking it along its time axis. The warping between two time series can then be used to find corresponding regions between two time series or to determine the similarity between two time series. Dynamic time warping is used in speaker recognition system to determine if two waveforms represent the same spoken words.

10. Tools utilized

The whole system has been implemented on MatLab 7.0 which used windows Xp as the operating system.

The MatLab 7.0 Graphic User Interface (GUI)has been used for the ease of end user. For the database generation MS Access has been employed.

11. Conclusion

From system which is designed it is concluded that the human voice is depends on the shape of vocal tract. For pitch detection, autocorrelation method is used as it is relatively impervious to noise but it is quite expensive. Instead of Euclidean method of distance measurement between two time series, Dynamic Time warping method is employed as in the previous method if there is slight shift in two time series in the time axis it will produce quite unintuitive results. This short coming is fulfilled by DTW algorithms by ignoring the global and local time shifts between two time series.

12. References

Books:

A.V.Oppenhein and R.W.Schafer, *Digital Signal Processing,* Prentice Hall

John G. Proakais and Dimitris G. Manolakis, *Digital signal Processing:Principles Algorithms and Applications,* Prentice Hall of India Pvt. Ltd., New Delhi. Third edition.

Thomas E. Quatieri, *Discrete time Speech Signal Processing, Principles and Practice,* Prentice Hall PTR.