REVIEW ARTICLE OPEN ACCESS

Text dependent speaker Identification by using Euclidian Distance

Saurav Verma1 Harish Ojha2 Amit Nerurkar 3
1 Assistant Professor MPSTME NMIMS 2,3Assistant Professor VIT, Mumbai, India

Abstract—
In this paper we propose a unique approach to text dependent speaker identification using transformation techniques such as DFT (Discrete Fourier Transform).The speech signal spoken by a particular speaker is converted into frequency domain after detecting and removing silence in the speech signal. The concept of row mean of the transform techniques has been used for feature extraction The distribution in the transform domain is utilized to extract the feature vectors in the training and the matching phases.

Keywords---voice biometrics, silence removal, row mean method.

I. Introduction

Voiceprint Recognition System also known as a Speaker Recognition System (SRS) is the best-known commercialized forms of voice Biometrics. Automated speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices. In contrast to other biometric technologies which are mostly image based and require expensive proprietary hardware such as vendor’s fingerprint sensor or iris-scanning equipment, the speaker recognition systems are designed for use with virtually any standard telephone or on public telephone networks. The ability to work with standard telephone equipment makes it possible to support broad-based deployments of voice biometrics applications in a variety of settings. In automated speaker recognition the speech signal is processed to extract speaker-specific information. These speaker specific informations are used to generate voiceprint which cannot be replicated by any source except the original speaker. This makes speaker recognition a secure method for authenticating an individual since unlike passwords or tokens; it cannot be stolen, duplicated or forgotten[2].

“Speaker recognition” (the best known commercialized form of Voice Biometric) is the computing task of validating a user's claimed identity using characteristics extracted from their voices.

There are two major commercialized applications of speaker recognition technologies and methodologies:
1)Speaker Identification
2)Speaker Verification(or Authentication)

SIS (Speaker Identification System)

* Determines who is talking from set of known voices
* No identity claim from user(many to one mapping)
* often assumed that unknown voice must come from set of known speakers-referred to as closed set identification.
* Speaker identification is a 1: N match where the voice is compared against N templates.
* Error that can occur in speaker identification is the false identification of speaker

SVS(Speaker Verification System)

*****Also known as a Speaker Authentication System.
* Speaker Verification on the other hand is the process of accepting or rejecting the speaker claiming to be the actual one.
* Since it is assumed that imposters (those who fake as valid users) are not known to the system, this is referred to as the open set task.
* Speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" “speaker model” or "voice model").
* Errors in speaker verification can be classified into the following two categories:
(1) False rejections: a true speaker is rejected as an imposter, and
(2) False acceptances: a false speaker is accepted as a true one.
Speaker recognition systems fall into two categories:
Text-dependent

* The text is same for enrollment and verification/identification.
* In a text-dependent system, prompts can either be common across all speakers (e.g.: a common pass phrase) or unique.

Text-independent.
* The text during enrollment and verification/ identification is different.

Text-independent systems are most often used for speaker identification as they require very little if any cooperation by the speaker. In fact, the enrollment may happen without the user's knowledge, as in the case for many forensic applications [2].

II.Transform Techniques

A.Frame Blocking

In this step the continuous speech signal is blocked into frames of *N* samples, with adjacent frames being separated by *M* (*M* < *N*). The first frame consists of the first *N* samples. The second frame begins *M* samples after the first frame, and overlaps it by *N* - *M* samples and so on. This process continues until all the speech is accounted for within one or more frames. Typical values for *N* and *M* are *N* = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and *M* = 100.

B.Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n),\ 0 \le n \le N-1$, where *N* is the number of samples in each frame, then the result of windowing is the signal

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1$$

Typically the *Hamming* window is used, which has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

C .Spectrogram of speech

Spectrograms are usually created in one of two ways: approximated as a filterbank that results from a series of bandpass filters (this was the only way before the advent of modern digital signal processing), or calculated from the time signal using the short-time Fourier transform (STFT). These two methods actually form two different Time-Frequency Distributions, but are equivalent under some conditions.

The bandpass filters method usually uses analog processing to divide the input signal into frequency bands; the magnitude of each filter's output controls a transducer that records the spectrogram as an image on paper.

Creating a spectrogram using the STFT is usually a digital process. Digitally sampled data, in the time domain, is broken up into chunks, which usually overlap, and Fourier transformed to calculate the magnitude of the frequency spectrum for each chunk. Each chunk then corresponds to a vertical line in the image; a measurement of magnitude versus frequency for a specific moment in time. The spectrums or time plots are then "laid side by side" to form the image or a three-dimensional surface.

The spectrogram of a signal s(t) can be estimated by computing the squared magnitude of the STFT of the signal s(t), as follows :

$$\text{Spectrogram}(t,\omega)=|\text{STFT}(t, \omega)|^2$$

D.Euclidian Distance

Compute the Euclidean distance for one dimension. The distance between two points in one dimension is simply the absolute value of the difference between their coordinates. Mathematically, this is shown as |p1 - q1| where p1 is the first coordinate of the first point and q1 is the first coordinate of the second point. We use the absolute value of this difference since distance is normally considered to have only a non-negative value.

Take two points P and Q in two dimensional Euclidean spaces. We will describe P with the coordinates (p1,p2) and Q with the coordinates (q1,q2). Now construct a line segment with the endpoints of P and Q. This line segment will form the hypotenuse of a right triangle. Extending the results obtained in Step 1, we note that the lengths of the legs of this triangle are given by |p1 - q1| and |p2 - q2|. The distance between the two points will then be given as the length of the hypotenuse

$$d\ (p\ ,q)= \sqrt{}\ \sum\ (q - p)^{2}$$

III. Proposed Method for feature extraction :

After application of hamming window, the spectrum of wave file is plotted and saved as image. This image is divided vertically in 16 equal parts.

The signal energy is calculated for each section as follows.

* Signal Energy: Let xi(n); n = 1…..N the audio samples of the i th frame, of length N. Then, for each frame i the energy is calculated by taking average of summation of square of signal. This simple feature can be used for detecting silent periods in audio signals, but also for discriminating between audio classes[1].

$$E=\log \sum_{n=1}^{n} S_n{}^{2}$$

Then the normalized signal energy is calculated. The mean of above feature extracted is returned.Thus for each wave file , column vector of 16 elements is returned & store in database .We have tested 120 samples i.e 30 speakers with same sentence in four different sessions.

Speaker identification is done by calculating euclidian distance of stored files with test file .The speaker in database having minimum euclidian distance is identified.

The test file given as input is the sentence spoken by one the speaker in database .

IV.Results

In this paper, we have review the topic and proposed the new solution for the detection of speaker using voice recogination system by euclidian distance method.

References

[1] A method for silence removal and segmentation of speech signals, implemented in Matlab-; Theodoros Giannakopoulos ;Computational Intelligence Laboratory (CIL);Insititute of Informatics and Telecommunications (IIT)

[2] Voiceprint Recognition Systems for Remote Authentication-A Survey;Zia Saquib, Nirmala Salam, Rekha Nair, Nipun Pandey;International Journal of Hybrid Information TechnologInternational Technology Vol. 4, No. 2, April, 2011

[3] Channel Compensation for Speaker Recognition Systems -Katrina Lee Neville, B Eng. School of Electrical and Computer Engineering Science, Engineering and Technology Portfolio; RMIT University; November 2006

[4] Fast Euclidean distance transformation in two scans using a 3x3 neighborhood, Frank Y. Shih* and Yi-Ta Wu Computer Vision Laboratory, College of Computing Sciences, New Jersey Institute of Technology, Newark, NJ 07102, USA Received 10 September 2002; accepted 18 September 2003

Authors :

Saurav Verma received B.Tech (Electronics & Communication) and M.Tech (Electronics and Telecommunication) degrees from Punjab Technical University and Mukesh Patel School of technology, NMIMS, Mumbai in 2011 and 2013, respectively. Area of interest is MEMs design, Wireless communication and digital communication. He is currently working as Assistant Professor in the department of Information Technology Engineering at MPSTME, NMIMS, Mumbai and also an IEEE Member.

Harish Ojha received B.E (Biomedical) and M.Tech (Electronics) degrees from Mumbai and Mukesh Patel School of technology, NMIMS, Mumbai in 2011 and 2013, respectively. Area of interest is Biomedical and digital communication. He is currently working as Assistant Professor in the department of Biomedical Engineering at Vidyalankar Institute of technology, Mumbai.

Amit Nerurkar completed ME in Computer Engineering from Vidyalankar Institute of Technology in February 2012, completed BE in Computer Engineering from Mumbai University. Research domain focused on Distributed systems, network security, Databases.