

Establishing the Discriminative Power of Biometric Data with Application to Speaker and Language Individuality

Narish Trilok

Abstract

Establishing the discriminative power of biometric data quantitatively, for the establishing the individual modality voice. The number of classes of classification is an unspecified, large number, hence establishing that biometric modality is capable of distinguishing every person and every language is difficult. The proposed methodology is statistically inferable. A many class problem is transformed into a dichotomy by using "distance" between two different classes and two samples of the same class, to establish thorough distinction of classes and thereby to validate distinct individuality. This model shall remain statistically inferable even when it does not observe all the classes. The process shall be illustrated in establishing individuality of voice and language with a given set of feature measurements.

Introduction

The task considered is of establishing the individuality of voice of each individual in a population since voice is an inherent variable for each individual. This task of establishing individuality is the same as showing the distinctiveness of classes with a very small error rate in discrimination. Individuality in handwriting has been shown in [1]. This paper proposes to give a validation of methodology used in [1] for individuality in voice and also generalize results to other domains. The same model to establish individuality in fingerprints has been recently shown [2].

Motivation

Voice recognition is the field of computer science that deals with designing computer systems that can recognize spoken words. Note that voice recognition implies only that the computer can take dictation, not that it *understands* what is being said. Comprehending human languages falls under a different field of computer science called *natural language processing*. Although Handwriting, fingerprints, face etc have been recognized as distinct per individual and used for verification purposes, voice of the speaker is not used. This paper proposes to discriminate users with Biometric with application to Speaker Individuality. The latest Voice recognition systems are based on the *Polychotomy* principle with a distinct disadvantage of being statistically non-inferential and thereby requiring many more observable instances of the same class in the training data. This paper proposes to show the individuality of languages by *dichotomy*. The advantage being that this method is statistically inferential.

Problem statement

Taking two audio inputs from speakers and use it for the Biometric discrimination of the Speaker Individuality and thereby determining whether they belong to the same person.

Present Approach (Polychotomy)

When a multiple class problem with a small number of classes and where one can observe enough instances of each class is considered, the instances can be clustered into one class and can be inferred to the entire population. This would show that individuality of classes statistically. This is a valid setup and easy but is limited to classes, which have substantial number of instances that are available. Without knowing the geometrical distribution of the unseen classes (populations), statistical inference cannot be drawn; true error of the entire from the error estimate of the sample population.

Proposed Approach (Dichotomy)

When a many class problem, where the number of classes is too large to be observed, is considered, the classification technique as mentioned in the previous paragraph cannot be applied to establish individuality as the number of classes is too large or unspecified. Many pattern identification problems especially in forensic sciences for establishing individuality fall under this category of many class problems.

Inferential statistics is the measure of reliability of individuality about the entire population based on data obtained from a sample drawn out of that population. The *Identification Model* is claimed to be not statistically inferable for a many class problem. In the many class problem, a population is all the biometric data samples of each person and is a very large or unspecified number. Samples from every single individual must be observed so that a conclusion could be drawn and is impossible. To draw statistical inference, the knowledge of the geometry of the unseen classes is a basic requirement. Since there are unseen classes, the error estimate of a sample population cannot infer the true error estimate of the entire population.

The alternative approach to be taken is that of transforming the many class problem into a dichotomy by taking the “distance” two samples of the same class and those of two different classes. This model allows inferential classification although there is no requirement for all the classes to be observed. In this model, two patterns are categorized into only one of the two classes; they either belong to the same class or are from two different classes.

Given two biometric data samples, the distance between the two samples is computed first. This distance value is used as data to be classified as positive or negative. Positive value of distance is intra-variation, within a person or identity and negative value is inter-variation, between different people or non-identity.

Pattern recognition techniques typically require that the features be homogeneous. The proposed model however, has a tremendous additional advantage. It allows the use of

multiple heterogeneous features. The *dichotomy* model overcomes the non-homogeneity of features as multiple type features are integrated into feature distinct scalar values.

Feature Extraction

The Effect of Human Anatomy

Human ear is a non-linear receptor and as such amplifies arbitrary features of acoustic data and dampens others. In addition, the hearing cells and nervous system can adapt to different situations and reduce the impact of noise and other random phenomena. By taking into consideration the effect of anatomy on the way we perceive sounds one can in theory eliminate all the irrelevant properties of a speech signal and use the remainder as a basis for recognition. The most important notion is that our auditory system works primarily on the frequency domain. The bandwidth we can hear is about 20 to 20000 Hz, but both the amplitude and frequency scales are non-linear. From a technical perspective there are various empirical frequency scales, which are applicable and utilize the knowledge of critical bandwidths. The critical bandwidth around a center frequency is the range at which all frequencies sound the same. Despite the demand for a spectral representation, some waveform analyzing methods in time domain have proved useful when digitally processing speech signals. Although the ear may ignore these additional features, they are very powerful support for computational analysis. They are used in determining the beginning and the end of an utterance and in classifying phonemes, for example.

Speech Recognition and Computers

The average rate at which humans can produce different phonemes is about 10 per second [9]. If we set a binary number for each phoneme and converted the acoustic signal into data stream, the bit rate would be less than 100 bits/sec. In reality, digitized speech uses high sample rates (>8kHz) often with 16-bit resolution, so interpreting the complex audio signal as a highly simplified set of discrete symbols is a demanding task for computers.

The inherent differences between the biological system in our heads and the logical but uncreative machine extend far beyond the physical methods of receiving acoustic signals. The computer does not understand anything – it only traverses a given algorithm. The extracted features should be such that there is no room for doubt or interpretation within fixed boundaries and the amount of data should remain reasonably low. In real-time systems the available computing power becomes the crucial problem so the computational load must not grow excessively.

The size of the vocabulary the system should recognize can vary from a few tens to thousands. With many possible words, the recognition process becomes demanding and time-consuming. Another issue is the type of speech: is it continuous or discrete, conversational or dictation. Furthermore, the type of use of the system determines whether it is speaker dependent or can recognize speech without training. The final important aspect is the environment and

equipment that are available, which dictates e.g. the available signal-to-noise ratio.

Time Domain Analysis

A digitized audio signal is a set of discrete values on regular intervals and if the sample rate is high and the values are connected with a line one can easily see the original waveform. In this report all sound samples are sampled at 44kHz and 16-bit resolution, which provide seemingly smooth graphs and more than enough data. It is good to examine the waveform of speech partly because of finding useful features and partly so as to understand the basis of subsequent processing.

Speech as a Quasi-stationary Signal

Producing a word is in effect a purely mechanical process. By changing the shape of the vocal tract we can modify the pulsative sound that originates in the vocal cords and ultimately emanates around us. At a given time frame though, say 20ms, the passage way is close to stationary and propagating sound waves remain unchanged. This is true especially for vowels and voiced consonants. By dividing the apparently stochastic acoustic data into frames it is now possible to easily calculate some useful average features on each frame. The y-axis associated with time domain tells the signal level from -1.0 to 1.0 throughout this report.

Cepstral Analysis

Cepstral Parameters are normally used instead of spectral or log-spectral domain parameters since they are

Compact – The same information can be represented with fewer parameters. High-order cepstra can be discarded since they represent high-frequency variations in log-spectrum.

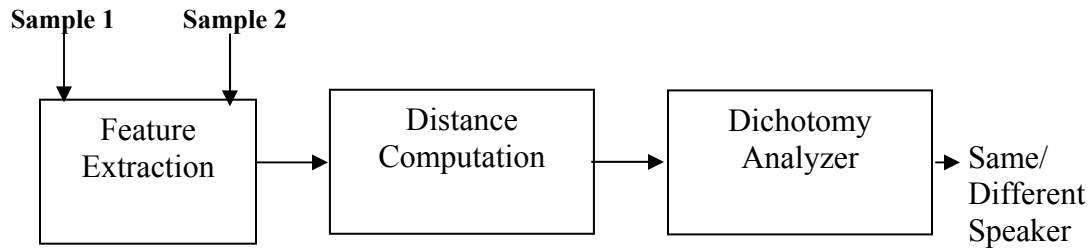
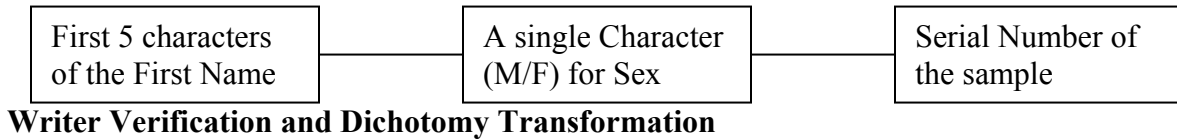
Uncorrelated – The cepstral coefficients are approximately uncorrelated. In fact, for Speech signals, DCT (Discrete Cosine Transform is an approximation that makes it uncorrelated)

Gain Independent – Only the Zeroth Cepstral value (a function of power) is dependent on energy (power) of the signal

Calculating the cepstral coefficients for a speech signal consists of preemphasis, Frame blocking and windowing followed usually by Fourier transformation, mel-scaling and cosine transformation for each time frame

Experimental Database

The data used in this in this experiment is that of audio files (*.wav). The data was collected from 10 different individuals. 10 voice samples of each individual was collected and stored using the 5 characters of the first name of the person, followed by Sex and then followed by the numeral corresponding to the number of the sample. Since there were 10 subjects, the convention was designed to provide unique names to the samples within the collected data.



On-going work

Currently MFCC feature extraction of the voice samples has been performed. Work is being undertaken in distance measurement between vectors. A dichotomy analyzer would be implemented using an Artificial Neural Network, and the results would be obtained. The hypothesis would be proved after establishing the various required statistical tests.

References:

[1] S. N. Srihari, S. -H. Cha, H. Arora and S. Lee, "Individuality of handwriting," *Journal of forensic sciences*, vol. 47, no. 4, pp. 856-872, 2002.

[2] S. Pankanti, S. Prabhakar, and A. K. Jain, "On the individuality of finger prints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1010-1025, 2002.

[3] S. -H. Cha, "Establishing the Discriminative power of a Biometric, with Application to Handwriting Individuality", 2002.

[4] Jonathan Law, Zhong-hua wang, Charles Tappert, "Corpus Collection Framework using VoiceXML", *MASPLAS*, 2002.

[5] Constantinides, G., "A framework for evaluating Multilingual systems." *Surprise 96'*, 1996

[6] N. A. Weiss, *Introductory Statistics*. Addison-Wesley, 5th ed., 1999.

[7] O. J. Dunn and V. A. Clark, *Applied Statistics: Analysis of Variance and Regression*. John Wiley & Sons, 2nd ed., 1987.

[8] T. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[9] L. R. Rabiner, R. W. Schafer; “Digital Processing of Speech Signals”;
Prentice-Hall; New Jersey; 1978

[10] M. Brookes, “Voicebox: Speech Processing Toolbox for Matlab”, Imperial
College, London, <http://www.ee.ic.ac.uk>.