Establishing the Uniqueness of the Human Voice for Security Applications

Naresh P. Trilok, Sung-Hyuk Cha, and Charles C. Tappert

Abstract

Establishing the individuality (uniqueness) of a person's voice is a necessary precursor to the successful development and deployment of security-related voice applications, such as speaker verification and speaker identification systems. Due to the large world population, however, the task of establishing the uniqueness of a person's voice is difficult, and one that has not yet been demonstrated. Our approach is to transform a many-class problem into a dichotomous one by using, as pattern classifier features, the distances between measurements extracted from utterances by the same and by different speakers. The resulting statistically inferable model allows us to experiment with a relatively small sample of the population and then to generalize the results to the entire population. This dichotomous model has previously been used to establish the individuality of handwriting and of fingerprints.

Keywords: biometrics, identity authentication, security, speaker verification, speaker identification, security informatics.

Introduction

Many security applications have adopted voice to verify or to identify a person. These speaker verification and identification systems are based on the hypothesis that the human voice is unique. Because it is important to validate the individuality of voice in order to use the voice biometric for security related systems, we consider here the task of establishing the individuality (uniqueness) of a person's voice. Earlier studies have used a statistically inferable dichotomy model to show the individuality of handwriting [3, 12] and of fingerprints [8]. Here, we propose using the same model to demonstrate the individuality of a person's voice. Statistical inference infers a conclusion about the entire population of interest from a sample of the population. That is, if the error rate of a random sample of the population is the same as that of the entire population, the procedure is said to be statistically inferential. This study summarizes and extends a dissertation study [13].

In addition to the transfer of meaning through linguistic components, human speech conveys information about the speaker, such as gender and identity. Here, we are interested in determining whether a person's voice has sufficient information to uniquely identify him/her. There has been considerable work on automatic speaker verification and speaker identification systems [2, 4, 6, 7, 10], yet it is not a solved problem and research in the area continues. Furthermore, even although these systems rely on the assumption that a person's voice is unique, such uniqueness has not yet been established.

There are various practical applications of automatic speaker verification and speaker identification, many of which have relevance to security informatics and national security. The human voice can serve as the means for security checks, and it is not something easily lost or forgotten like an ID card. For example, the human voice can be used to verify identity for access to physical facilities by storing a speaker model in a small circuit chip to be used as an access tag instead of a pin code. Another speaker identification application might be to monitor people by their voices. For instance, it could be used for information retrieval by speaker indexing of recorded debates or news, and then retrieving the speech of only certain speakers. It can also be used to monitor criminal activity in common places by associating individuals with their voices.

The approach taken here is to transform the many-class problem into a dichotomy one by taking the distances between feature vectors of samples of the same class and those of samples of different classes [3, 12]. This model allows inferential classification without the requirement to observe all the classes. In this model, two input patterns are classified either as belonging to the same class or to two different classes. Given two biometric data samples, the distance between the features of the two samples is computed, and this distance is used as input to be classified as positive or negative, positive for intra-class (same person) and negative for inter-class (different people) distances, as shown in Fig. 1 for two features as an illustrative example.

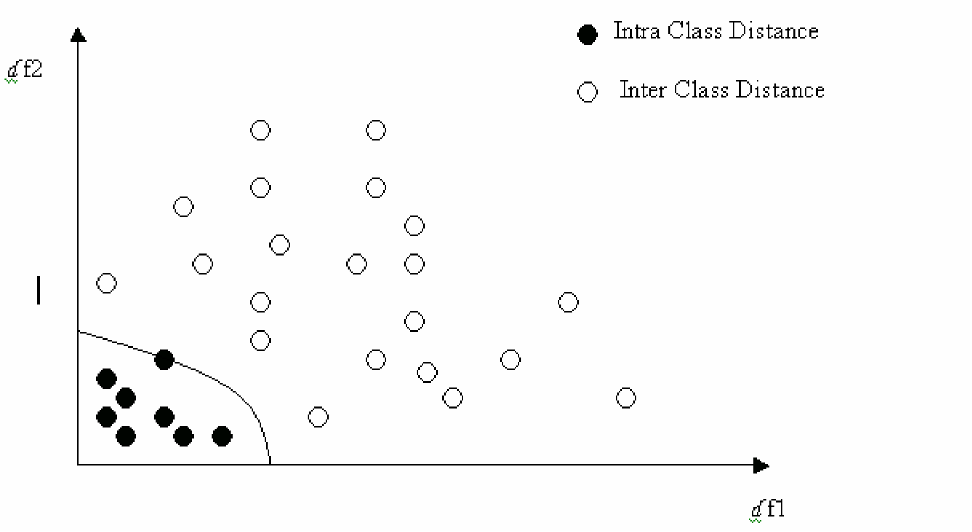
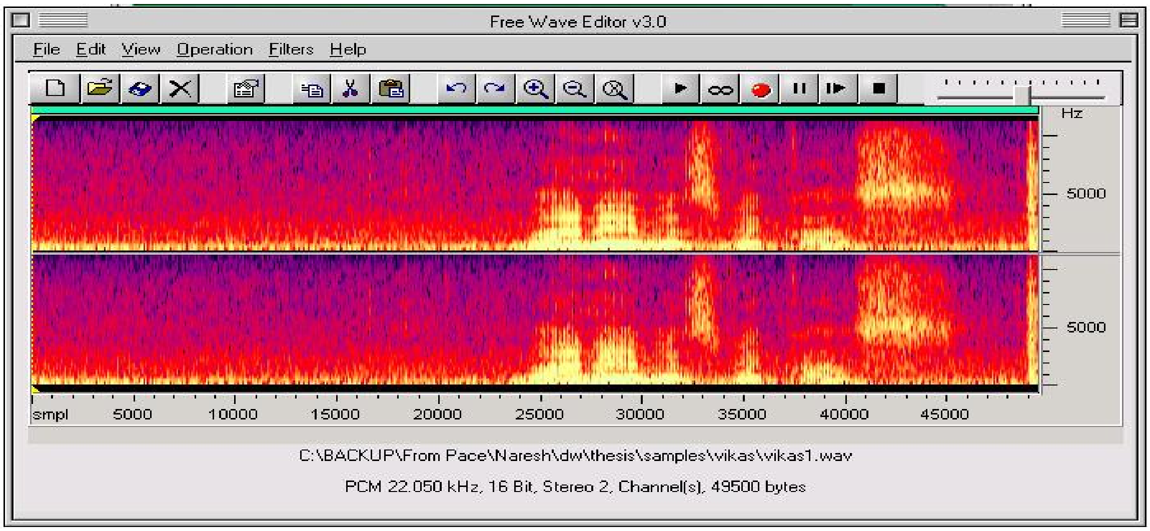


Figure 1. Dichotomy model.

Methodology

This study involved speech sample collection, segmentation, feature extraction, and classification. Speech samples were collected from ten subjects, each subject repeating the utterance "My name is (person's name)" ten times, for a total of 100 speech samples. The speech samples were recorded over a standard microphone (*Cyber Acoustics OEM AC-200 Stereo Speech Headset and Microphone*) attached to a PC (*Dell Dimension*^{TM} *2400, with a Pentium IV Processor*) running the *Windows XP* Operating System. We collected the speech samples using *Microsoft Sound Recorder* that comes as the part of the XP Operating System. The speech samples and speaker information were stored in a database for later processing.

The speech samples were segmented to isolate the “My name is” portion of the speech utterance in common to all the samples. We used *Free Wave Editor* [14], a freeware application downloaded from the Internet, to manually perform the segmentation. This application software allows the user to view the speech signal in both the time and frequency domains, and we used primarily the spectrographic view (Fig. 2).

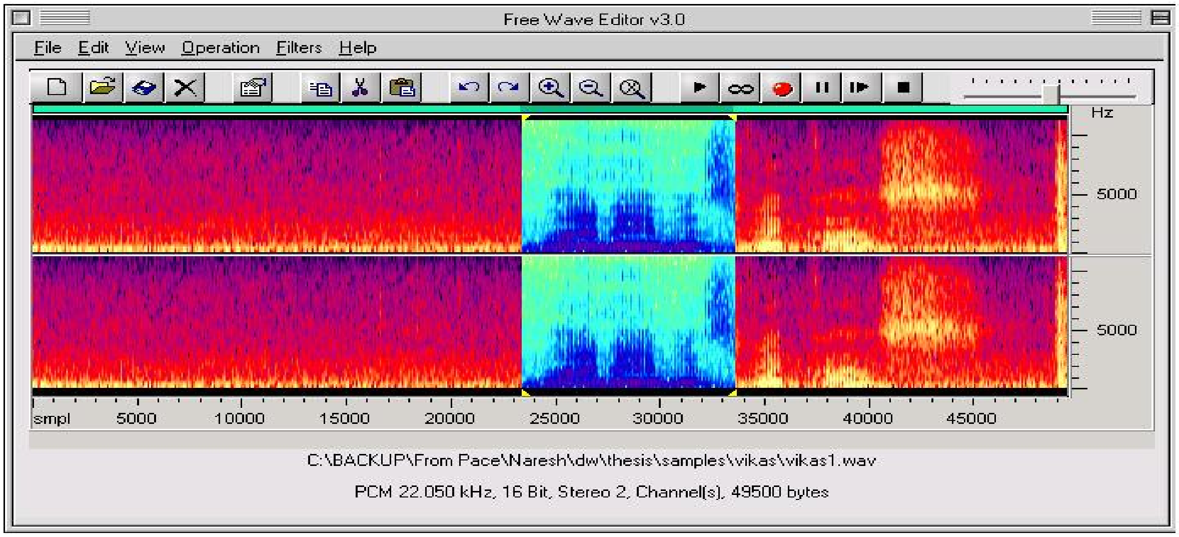


C:\BACKUP\From Pace\Naresh\dw\thesis\samples\vikas\vikas1.wav

PCM 22.050 kHz, 16 Bit, Stereo 2, Channel(s), 49500 bytes

Figure 2. Speech sample spectrogram in Free Wave Editor.

The software allows the user to segment an utterance by left mouse clicking at the start of the desired phrase to get a dotted yellow line and right clicking at the end of the phrase to get a shaded blue area between the lines (Fig. 3). The selected portion can be played, and these lines can be adjusted to get the required speech segment before saving it as a separate *wav* file.



C:\BACKUP\From Pace\Naresh\dw\thesis\samples\vikas\vikas1.wav

PCM 22.050 kHz, 16 Bit, Stereo 2, Channel(s), 49500 bytes

Figure 3. Example of the segmented portion of the utterance.

The duration of the “My name is” speech utterances is about one second. Fig. 4 shows the ten segmented speech samples from three different subjects. They show the similarity of the within-speaker samples, the dissimilarity of the between-speaker samples, and the variability in utterance length.

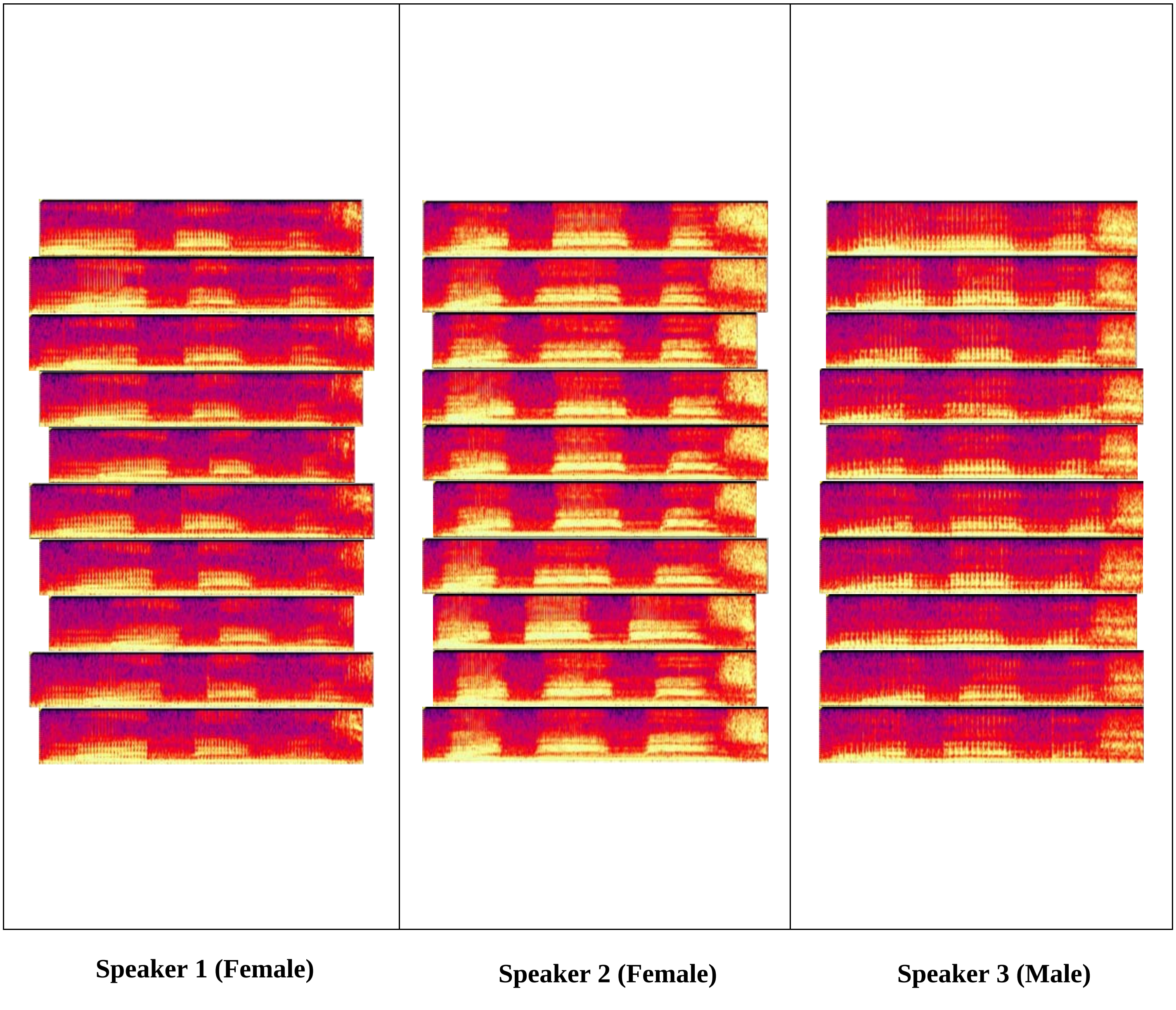


Figure 4. Speech samples from three speakers.

We used the Speech Processing Toolbox written in Matlab for *wav* files [1] to perform standard speech signal processing [9, 11] on the segmented utterances. The 13 lowest Mel-frequency Cepstral coefficients (MFCC) were computed from 40 Mel-spaced filters: 13 spaced linearly with 133.33 Hz between center frequencies, and 27 spaced logarithmically by a frequency factor of 1.07 between adjacent filters. The spectral analysis frame was a 30 ms Hamming window with 10 ms overlap between adjacent windows. The number of time windows per utterance varied since they were of fixed size and the lengths of the voice samples varied. We computed features from the 13 Cepstral coefficients over specified time intervals.

We extracted three different feature sets from the spectral data, and these feature sets normalized for the varying lengths of the speech utterances. The first consisted of taking the means and variances of each of the 13 frequency bands over the entire utterance, for a total of 26 features per utterance. For the second, we used the same 13 frequency bands but divided each utterance into its 7 speech sounds (Fig 5), averaging the energy in the 13 frequency bands within each of the 7 sounds, for a total of 91 features. Finally, for the third feature set, we omitted the first Cepstral component since it represents the energy of the signal and is not speaker specific, and used the remaining 12 frequency bands in the 7 sounds for a total of 84 features.

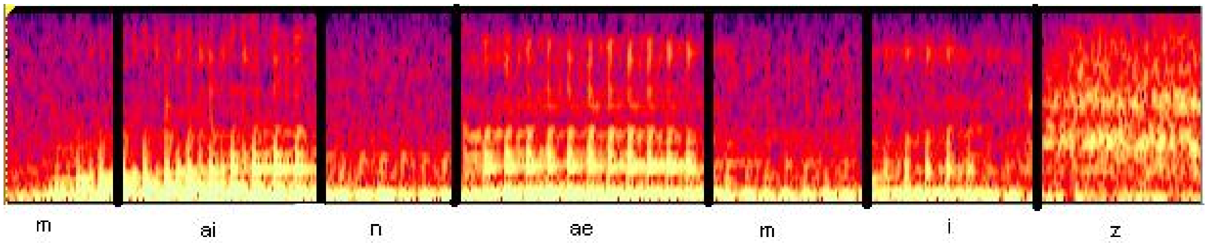


Figure 5. Example of the "My name is" utterance divided into seven sound units.

The classifier described above was implemented with a neural network (Fig. 6), which was trained using the back-propagation algorithm.

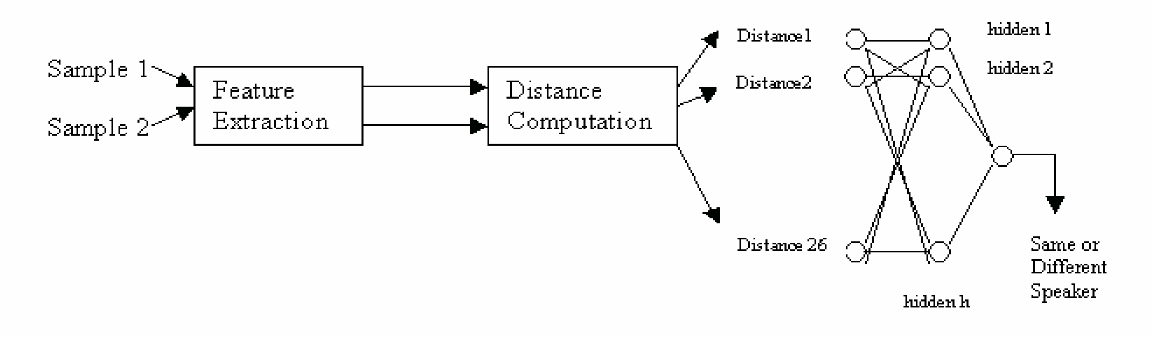


Figure 6. Neural network classifier.

Experimental Results

The absolute distances between the features of each sample and those of all the other samples of the same subject (intra-class distance) and those of all the samples of other subjects (inter-class distance) were calculated. A total of 450 intra-class distances and 4500 inter-class distances were obtained. These were divided into groups of 225 each. Of the two groups of 225 intra-class distances, one was used for training and one for testing, and of the 20 groups of 225 inter-class distances, one was used for training and one for testing. The experimental results are shown in Table 1.

Normalized Features	**Hidden Units**	**Type I Error**	**Type II Error**	**Accuracy**
26 Features	10	3%	20%	77%
91 Features	32	0%	11%	89%
84 Features	28	0%	6%	94%

Table 1. Experimental results.

Upon examination of the speech samples in error on the 84 feature experiment, we found that the recording quality of four samples was unusually poor. Removing these samples and rerunning the experiments reduced the type II error rate to 2%, increasing the accuracy to 98%, further supporting the hypothesis of the individuality of the human voice.

6.9 Conclusions

We described a methodology for establishing the individuality of the human voice and thus the discriminative power of speech biometric data. Our preliminary experimental results using a statistically inferential dichotomy model demonstrated that an individual's voice print does indeed appear to be unique to the individual.

Additional work will be required to develop a working speaker verification system using this model and procedures. For example, the manual segmentation process could be automated by using the dynamic programming, time warping algorithm [5].

References

[1] M. Brookes, "Voicebox: Speech Processing Toolbox for Matlab," Imperial College, London, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (last visited 04/2003).

[2] J.P. Campbell, "Speaker Recognition: A Tutorial," *Proc. IEEE*, vol. 85, no. 9, Sept 1997, pp. 1437-1462.

[3] S.–H. Cha and S.N. Srihari, "Writer Identification: Statistical Analysis and Dichotomizer," *Proc. SPR & SSPR*, Alicante, LNCS-Advances in Pattern Recognition, vol. 1876, pp 123-132, 2000.

[4] X. Huang, A. Acero and H.-W. Hon, *Spoken Language Processing*, Prentice, 2001.

[5] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, Prentice, 2000.

[6] J.W. Koolwaaij, "Automatic Speaker Verification in Telephony: a Probabilistic Approach," 2000, http://himalaya.lab.telin.nl/~koolwaaij/research/Pub/koolwaaij. 2000.4.shtml (last visited 12/2003)

[7] J. M.Naik, "Speaker Verification: A Tutorial," *IEEE Communications Magazine*, Jan 1990, pp. 42-48.

[8] S. Pankanti, S. Prabhakar, and A. K. Jain, "On the individuality of finger prints," *IEEE Trans. PAMI*, vol. 24, no. 8, pp. 1010-1025, 2002.

[9] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Pearson Education, 1993.

[10] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," *ICASSP*, 2002, pp. 4072-4075.

[11] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, 1997.

[12] 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.

[13] N. P. Trilok, "Assessing the discriminative power of voice," M.S. Dissertation, School of CSIS, Pace University, 2004.

[14] Wave Editor – an advanced WAV file editor and recorder, http://www. webspeakster.com/ wave_editor.htm (last visited 11/2003).