Forensic speaker identification based on spectral moments

R. Rodman, D. McAllister,* D. Bitzer,* L. Cepeda* and P. Abbitt*[†]

**Voice I/O Group: Multimedia Laboratory*
Department of Computer Science
North Carolina State University
rodman@csc.ncsu.edu

[†]Department of Statistics
North Carolina State University

ABSTRACT A new method for doing text-independent speaker identification geared to forensic situations is presented. By analysing 'isolexemic' sequences, the method addresses the issues of very short criminal exemplars and the need for open-set identification. An algorithm is given that computes an average spectral shape of the speech to be analysed for each glottal pulse period. Each such spectrum is converted to a probability density function and the first moment (i.e. the mean) and the second moment about the mean (i.e. the variance) are computed. Sequences of moment values are used as the basis for extracting variables that discriminate among speakers. Ten variables are presented all of which have sufficiently high inter- to intraspeaker variation to be effective discriminators. A case study comprising a ten-speaker database, and ten unknown speakers, is presented. A discriminant analysis is performed and the statistical measurements that result suggest that the method is potentially effective. The report represents work in progress.

KEYWORDS speaker identification, spectral moments, isolexemic sequences, glottal pulse period

PREFACE

Although it is unusual for a scholarly work to contain a preface, the controversial nature of our research requires two *caveats,* which are herein presented.

First, the case study described in our article to support our methodology was performed on sanitized data, that is, data not subjected to the degrading effect of telephone transmission or a recording medium such as a tape recorder. We acknowledge, in agreement with Künzel (1997), that studies based strictly on formant frequency values are undermined by telephone transmission. Our answer to this is that our methodology is based on *averages of entire spectral shapes of the vocal tract*. These spectra are derived by a pitch synchronous Fourier analysis that treats the vocal tract as a filter that is driven by the glottal pulse treated as an impulse function. We believe that the averaging of such spectral shapes will mitigate the degrading effect of the transmittal medium. The purpose of this study,

 1350-1771

however, is to show that the method, being novel, is promising when used on 'clean' data.

We also acknowledge, and discuss below in the 'Background' section, the fact that historically spectral parameters have not proved successful as a basis for accurate speaker identification. Our method, though certainly based on spectral parameters, considers averages of entire, pitch independent spectra as represented by spectral moments, which are then plotted in curves that appear to reflect individual speaking characteristics. The other novel part of our approach is comparing 'like-with-like'. We base speaker identification on the comparison of manually extracted 'isolexemic' sequences. This, we believe, permits accurate speaker identification to be made on very short exemplars. Our methods are novel and so far unproven on standardized testing databases (though we are in the process of remedying this lacuna). The purpose of this article is to publicize our new methodology to the forensic speech community both in the hopes of stimulating research in this area, and of engendering useful exchanges between ourselves and other researchers from which both parties may benefit.

INTRODUCTION

Speaker identification is the process of determining who spoke a recorded utterance. This process may be accomplished by humans alone, who compare a spoken exemplar with the voices of individuals. It may be accomplished by computers alone, which are programmed to identify similarities in speech patterns. It may alternatively be accomplished through a combination of humans and computers working in concert, the situation described in this article.

Whatever the case, the focus of the process is on a speech exemplar – a recorded threat, an intercepted message, a conspiracy recorded surreptitiously – together with the speech of a set of suspects, among whom may or may not be the speaker of the exemplar. The speech characteristics of the exemplar are compared with the speech characteristics of the suspects in an attempt to make the identification.

More technically and precisely, given a set of speakers $S = \{S_1 \ldots S_N\}$, a set of collected utterances $U = \{U_1 \ldots U_N\}$ made by those speakers, and a single utterance u_X made by an unknown speaker: *closed-set* speaker identification determines a value for X in [1 … N]; *open-set* speaker identification determines a value for X in [0, 1 … N], where X = 0 means 'the unknown speaker $S_X \notin S$'. 'Text independent' means that u_X is not necessarily contained in any of the U_i.

During the process, acoustic feature sets $\{F_1 \ldots F_N\}$ are extracted from the utterances $\{U_1 \ldots U_N\}$. In the same manner, a feature set F_X is extracted from u_X. A matching algorithm determines which, if any, of $\{F_1 \ldots F_N\}$ suf-

ficiently resembles F_x. The identification is based on the resemblance and may be given with a probability-of-error coefficient.

Forensic speaker identification is aimed specifically at an application area in which criminal intent occurs. This may involve espionage, blackmail, threats and warnings, suspected terrorist communications, etc. Civil matters, too, may hinge on identifying an unknown speaker, as in cases of harassing phone calls that are recorded. Often a law enforcement agency has a recording of an utterance associated with a crime such as a bomb threat or a leaked company secret. This is u_x. If there are suspects (the set S), utterances are elicited from them (the set U), and an analysis is carried out to determine the likelihood that one of the suspects was the speaker of u_x, *or that none of them was*. Another common scenario is for agents to have a wiretap of an unknown person who is a suspect in a crime, and a set of suspects to test the recording against.

Forensic speaker identification distinguishes itself in five ways. First, and of primary importance, it must be *open-set* identification. That is, the possibility that none of the suspects is the speaker of the criminal exemplar must be entertained. Second, it must be capable of dealing with very short utterances, possibly under five seconds in length. Third, it must be able to function when the exemplar has a poor signal-to-noise ratio. This may be the result of wireless communication, of communication over low-quality phone lines, or of data from a 'wire' worn by an agent or informant, among others. Fourth, it must be text independent. That is, identification must be made without requiring suspects to repeat the criminal exemplar. This is because the criminal exemplar may be too short for statistically significant comparisons. As well, it is generally true that suspects will find ways of repeating the words so as to be acoustically dissimilar from the original. Moreover, it may be of questionable legality as to whether a suspect can be forced to utter particular words. Fifth, the time constraints are more relaxed. An immediate response is generally not required so there is time for extensive analysis, and most important in our case, time for human intervention. The research described below represents work in progress.

BACKGROUND

The history of electronically assisted speaker identification began with Kersta (1962), and can be traced through these references: Baldwin and French (1990), Bolt (1969), Falcone and de Sario (1994), French (1994), Hollien (1990), Klevans and Rodman (1997), Koenig (1986), Künzel (1994), Markel and Davis (1978), O'Shaughnessy (1986), Reynolds and Rose (1995), Stevens *et al.* (1968) and Tosi (1979).

Speaker identification can be categorized into three major approaches. The first is to use long-term averages of acoustic features. Some features that have been used are inverse filter spectral coefficients, pitch, and

cepstral coefficients (Doddington 1985). The purpose is to smooth across factors influencing acoustic features, such as choice of words, leaving behind speaker-specific information. The disadvantage of this class of methods is that the process discards useful speaker-discriminating data, and can require lengthy speech utterances for stable statistics.

The second approach is the use of neural networks to discriminate speakers. Various types of neural nets have been applied (Rudasi and Zahorian 1991, Bennani and Gallinari 1991, Oglesby and Mason 1990). A major drawback to the neural net methods is the excessive amount of data needed to 'train' the speaker models, and the fact that when a new speaker enters the database the entire neural net must be retrained.

The third approach – the segmentation method – compares speakers based on similar utterances or at least using similar phonetic sequences. Then the comparison measures differences that originate with the speakers rather than the utterances. To date, attempts to do a 'like phonetic' comparison have been carried out using speech recognition front-ends. As noted in Reynolds and Rose (1995), 'It was found in both studies [Matsui and Furui 1991, Kao *et al*. 1992] that the front-end speech recognizer provided little or no improvement in speaker recognition performance compared to no front-end segmentation.'

The Gaussian mixture model (GMM) of speakers described in Reynolds and Rose (1995) is an implicit segmentation approach in which like sounds are (probabilistically) compared with like. The acoustic features are of the mel-cepstral variety (with some other preprocessing of the speech signal). Their best results in a closed-set test using five second exemplars was correct identification in 94.5% ±1.8 of cases using a population of 16 speakers (Reynolds and Rose 1995: 80). Open-set testing was not attempted.

Probabilistic models such as Hidden Markov Models (HMMs) have also been used for text-independent speaker recognition. These methods suffer in two ways. One is that they require long exemplars for effective modelling. Second, the HMMs model temporal sequencing of sounds, which 'for text-independent tasks … contains little speaker-dependent information' (Reynolds and Rose 1995: 73).

A different kind of implicit segmentation was pursued in Klevans and Rodman (1997) using a two-level cascading segregating method. Accuracies in the high 90s were achieved in closed-set tests over populations (taken from the TIMIT database) ranging in size from 25 to 65 from similar dialect regions. However, no open-set results were attempted.

In fact, we believe the third approach – comparing like utterance fragments with like – has much merit, and that the difficulties lie in the speech recognition process of explicit segmentation, and the various clustering and probabilistic techniques that underlie implicit segmentation. In forensic applications, it is entirely feasible to do a manual segmentation

that guarantees that *lexically* similar partial utterances are compared. This is discussed in the following section.

SEMI-AUTOMATIC SPEAKER IDENTIFICATION

Semi-automatic speaker identification permits human intervention at one or more stages of computer processing. For example, the computer may be used to produce spectrograms (or any of a large number of similar displays) that are interpreted by human analysts who make final decisions (Hollien 1990).

One of the lessons that has emerged from nearly half a century of computer science is that the best results are often achieved by a collaboration of humans and computers. Machine translation is an example. Humans translate better, but slower; machines translate poorly, but faster. Together they translate both better and faster, as witnessed by the rise in popularity of so-called CAT (Computer-aided Translation) software packages. (The EAMT – European Association for Machine Translation – is a source of copious material on this subject, for example, the *Fifth EAMT Workshop* held in Ljubljana, Slovenia in May of 2000.)

The history of computer science also teaches us that while computers can achieve many of the same intellectual goals as humans, they do not always do so by imitating human behaviour. Rather, they have their own distinctly computational style. For example, computers play excellent chess but they choose moves in a decidedly non-human way.

Our speaker identification method uses computers and humans to extract *isolexemic* sound sequences, which are then heavily analysed by computers alone to extract personal voice traits. The method is appropriate for forensic applications, where analysts may have days or even weeks to collect and process data for speaker identification.

Isolexemic sequences may consist of a single phone (sound); several phones such as the rime (vowel plus closing consonant(s)) of a syllable (e.g. the *ill* of *pill* or *mill)*; a whole syllable; a word; sounds that span syllables or words; etc. What is vital is that the sequence be 'iso' in the sense that it comes from the same word or words of the language as pronounced by the speakers being compared. A concrete example illustrates the concept. The two pronunciations of the vowel in the word *pie,* as uttered by a northern American and a southern American, are isolexemic because they are drawn from the same English word. That vowel, however, will be pronounced in a distinctly different manner by the two individuals, assuming they speak a typical dialect of the area. By comparing isolexemic sequences, the bulk of the acoustic differences will be ascribable to the speakers. Speech recognizers are *not* effective at identifying isolexemic sequences that are phonetically wide apart, nor are any of the implicit segmentation techniques. Only humans, with deep knowledge of the language, know that *pie* is the same word regardless of the fact that the

vowels are phonetically different, and despite the fact that the same phonetic difference, in other circumstances, may function phonemically to distinguish between different words. The same word need not be involved. We can compare the 'enny' of *penny* with the same sound in *Jenny* knowing that differences – some people pronounce it 'inny' – will be individual, not linguistic. Moreover, the human analyst, using a speech editor such as Sound Forge™, is able to isolate the 'enny' at a point in the vowel where coarticulatory effects from the *j* and the *p* are minimal.

In determining what sound sequences are 'iso', the analyst need not be concerned with prosodics (pitch or intonation in particular) because, as we shall see, the analysis of the spectra is glottal pulse or pitch synchronous, the effect of which is to minimize the influence of the absolute pitch of the exemplars under analysis. In fact, one of the breakthroughs in the research reported here is an accurate means of determining glottal pulse length so that the pitch synchronicity can hold throughout the analysis of hundreds of spectra (Rodman *et al*. 2000).

Isolexemic comparisons cut much more rapidly to the quick than any other way of comparing the speech of multiple speakers. Even three seconds of speech may contain a dozen syllables, and two dozen phonetic units, all of which could hypothetically be used to discriminate among speakers.

The manual intervention converts a *text-independent* analysis to the more effective *text-dependent* analysis without the artifice of making suspects repeat incriminating messages, which does not work if the talker is uncooperative in any case, for he may disguise his voice (Hollien 1990: 233). (The disguise may take many forms: an alteration of the rhythm by altering vowel lengths and stress patterns, switching dialects for multidialectal persons, or faking an 'accent'.)

For example, suppose the criminal exemplar is 'There's a bomb in Olympic Park and it's set to go off in ten minutes.' Suspects are interviewed and recorded (text independent), possibly at great length over several sessions, until they have eventually uttered sufficient isolexemic parts from the original exemplar. For example, the suspect may say 'we met to go to the ball game' in the course of the interview, permitting the isolexemic '[s]<u>et to go</u>' and '[m]<u>et to go</u>' to be compared (text dependent). A clever interrogator may be able to elicit key phrases more quickly by asking pointed questions such as 'What took place in Sydney, Australia last summer?', which might elicit the word *Olympics* among others. Or indeed, the interrogator could ask for words directly, one or two at a time, by asking the suspect to say things like 'Let's take a break <u>in ten minutes</u>.'

The criminal exemplar and all of the recorded interviews are digitized (see below) and loaded into a computer. The extraction of the isolexemic sequences is accomplished by a human operator using a sound editor such as Sound Forge™. This activity is what makes the procedure semi-automatic.

FEATURE EXTRACTION

All the speech to be processed is digitized at 22.050 kHz, 16 bit quantization, and stored in .wav files. This format is suitable for input to any sound editor, which is used to extract the isolexemes to be analysed. Once data are collected and the isolexemes are isolated, both from the criminal exemplar and the utterances of suspects (in effect, the training speech), the process of feature extraction can begin.

Feature extraction takes place in two stages. The first is the creation of 'tracks', essentially an abbreviated trace of successive spectra. The second is the measurement of various properties of the tracks, which serve as the features for the identification of speakers.

Creating 'tracks'

We discuss the processing of voiced sounds, that is, those in which the vocal cords are vibrating throughout. The processing of voiceless sounds is grossly similar but differs in details not pertinent to this article. (The interested reader may consult Fu *et al*. 1999.) Our method requires the computation of an average spectrum for each glottal pulse (GP) – opening and closing of the vocal cords – in the speech signal of the current isolexeme. We developed an algorithm for the accurate computation of the glottal pulse period (GPP) of a succession of GPs. The method, and the mathematical proofs that underlie it, and a comparison with other methods, are published as Rodman *et al*. (2000).

By using a precise, pitch synchronous Fourier analysis, we produce spectral shapes that reflect the shape of the vocal tract, and are essentially unaffected by pitch. In effect, we treat the vocal tract as a filter that is driven by the glottal pulse, which is treated as an impulse function. The resulting spectra are highly determined by vocal tract shapes and glottal pulse *shapes* (not spacing). These shapes are speaker dependent and this provides the basis for speaker identification.

We use spectral moments as representative values of these spectral shapes. We use them as opposed, say, to formant frequencies, because they contain information across the entire range of frequencies up to 4 kHz for voiced sounds, and 11 kHz for voiceless sounds (not discussed in this article). The higher formants, and the distribution of higher frequencies in general, have given us better results than in our experiments with pure formant frequencies and even with moments of higher orders (Koster 1995).

Knowing the GPP permits us to apply the following steps to compute a sequence of *spectral moments*.

(Assume the current GPP contains N samples.)

1. Compute the discrete Fourier transform (DFT) using a window width of N, thus transforming the signal from the time domain to the frequency domain.
2. Take the absolute value of the result (so the result is a real number).
3. Shift over 1 sample.
4. Repeat steps 1–3 N times.
5. Average the N transforms and scale by taking the cube root to reduce the influence of the first formant, drop the DC term, and interpolate it with a cubic spline to produce a continuous spectrum.
6. Convert the spectrum to a probability density function by dividing it by its mass, then calculate the first moment m_1 (mean) and the second central moment about the mean m_2 (variance) of that function in the range of 0 to 4000 Hz and put them in two lists L_1 and L_2. Let *S(f)* be the spectrum. The following formulae are used, appropriately modified for the discrete signal:

$$mass = \int_0^{4000} S(f)df$$

$$P(f) = S(f) \div mass$$

$$mean = m_1 = \int_0^{4000} f * P(f)df$$

$$variance = m_2 = \int_0^{4000} (f-m_1)^2 * P(f)df$$

7. Repeat Steps 1 through 6 until less than 3N samples remain.
8. Scale each moment: m_1 by 10^{-3} and m_2 by 10^{-6}.

Several comments about the algorithm are in order. The shifting and averaging in Steps 1–3 are effective in removing noise resulting from small fluctuations in the spectra, but preserving idiosyncratic features of the vocal tract shape. Although the window spans the length of two glottal pulses as it slides across, there is one spectral shape computed per glottal pulse. The overlapping windows improve the sensitivity of the method. The process is computationally intense but it yields track points that are reliable and consistent in distinguishing talkers. The procedure also removes the pitch as a parameter affecting the shape of the transform, as noted above.

In Step 5 the cube root is taken – at one time we took the logarithm – because the first formant of voiced speech contains a disproportionate amount of the spectral energy. The effect of taking the cube root 'levels' the peaks in the spectrum and renders the spectrum more sensitive to speaker differences. The means and variances of Step 6 are chosen as 'figures of merit' for the individual spectra. Although representing a single spectrum over a 4 kHz bandwidth with two numbers appears to give up information, it has the advantage of allowing us to track *every* spectrum in the isolexeme and to measure the changes that occur. This dynamism leads to features that we believe to be highly individuating because they capture the shape, position and movement of the speaker's articulators, which are unique to each speaker. (This is argued in more detail in Klevans and Rodman 1997.) Also in Step 6, the division by the spectral mass removes the effect of loudness, so that two exemplars, identical except for intensity, will produce identical measurements. Finally, the scaling in Step 8 is performed so that we are looking at numbers in [0, 3] for both means and variances. This is done as a matter of convenience. It makes the resulting data more readable and presentable.

The result of applying the algorithm is a sequence of points in two-dimensional m_1-m_2 space that can be interpolated to give a *track*. These are the values from the lists L_1 and L_2. The tracks are smoothed by a three-stage cascading filter: median-5, average-3, median-3. That is, the first pass replaces each value (except endpoints) with the median of itself and the four surrounding values. The second pass takes that median-5 output and replaces each point by the average of itself with the two surrounding values. That output is subjected to the median-3 filter to give the final, smoothed track. The smoothing takes place because the means and variances of the spectra make small jumps when the speech under analysis is in a (more or less) steady state as in the pronunciation of vowels. This is true especially for monophthongal vowels such as the 'e' in *bed*, but even in diphthongs such as the 'ow' in *cow*, there are steady states that span several glottal pulse periods. The smoothing removes much of the irrelevant effect of the jumps. (See also Fu *et al.* 1999, Rodman *et al.* 1999.)

A visual impression of intra- and interspeaker variation may be seen in Figure 1. The first two tracks in the figure are a single speaker saying *owie* on two different occasions. The third and fourth tracks in the figure are two different speakers saying *owie*. Figures 2 and 3 are similar data for the utterances *ayo* and *eya*.

Our research shows that the *inter*speaker variation of tracks of isolexemic sequences will be measurably larger than the *intra*speaker variation, and therefore that an unknown speaker can be identified through these tracks.

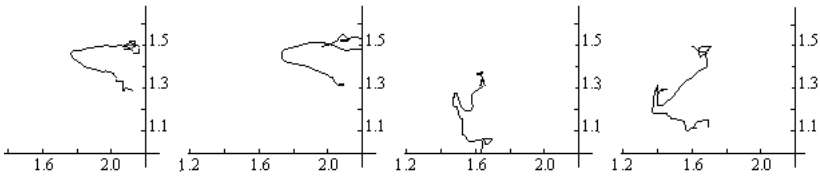

Figure 1 The first two plots are the same speaker saying 'owie'; the third and fourth plots are different speakers saying 'owie'

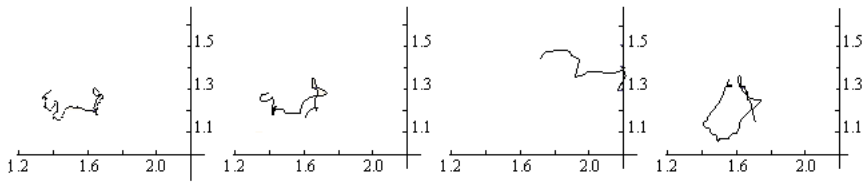

Figure 2 The first two plots are the same speaker saying 'ayo'; the third and fourth plots are different speakers saying 'ayo'

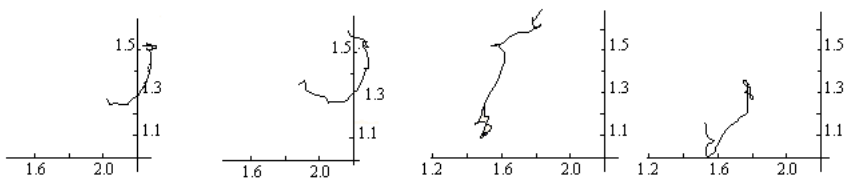

Figure 3 The first two plots are the same speaker saying 'eya'; the third and fourth plots are different speakers saying 'eya'

Extracting features from tracks

To compare tracks, several factors must be considered: the region of moment space occupied by the track; the shape of the track; the centre of gravity of the track; and the orientation of the track. Each of these characteristics displays larger interspeaker than intraspeaker variation when reduced to statistical variables. One way to extract variables is to surround the track by a *minimal enclosing rectangle* (MER), which is the rectangle of minimal area containing the entire track. The MER is computed by

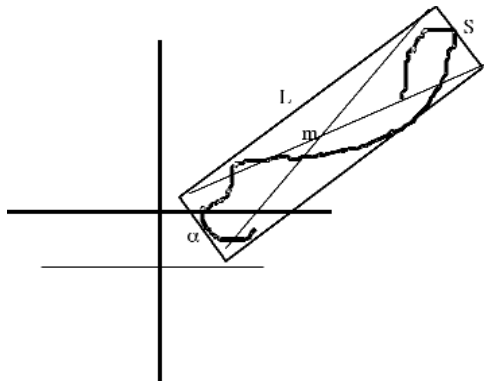

Figure 4 A minimal enclosing rectangle

rotating the track about an endpoint one degree at a time and computing the area of a bounding rectangle whose sides are parallel to the axes each time, and then taking the minimum. The minimum is necessarily found within 90 degrees of rotation. This is illustrated in Figure 4.

From the MER of the curve in its original orientation, we extract four of the ten variables to be used to characterize the tracks, viz. the x-value of the midpoint, the y-value of the midpoint, the length of the long side (L), and the angle of orientation (α). (The length of the short side was not an effective discriminator for this study.) Four more variables are the minimal x-coordinate, the minimal y-coordinate, the maximal x-coordinate, and the maximal y-coordinate of the track. They are derived by surrounding the track in its original orientation with a minimal rectangle parallel to the axes and taking the four corner points. These eight parameters measure the track's location and orientation in moment space.

The final two variables attempt to reflect the shape of the track. Note that the spacing and number of track points in an utterance depend on the fundamental pitch. The higher the frequency the fewer the number of samples in the period and hence the greater the number of track points that will be computed over a given time period. To obviate this remaining manifestation of pitch and hence, the number of track points, as a factor affecting the measurement of the shape of a curve, we reparameterize the curve based on the distance between track points. We normalize the process so that the curve always lies in the same interval thus removing track length as a factor. (Other variables take it into account.)

More particularly, we parameterize the tracks in m_1-m_2 space into two integrable curves by plotting the m_1-value of a point *p* (the ordinate) versus the distance in m_1-m_2 space to point $p+1$ (abscissa), providing the distance exceeds a certain threshold. If it does not, the point $p+1$ is thrown out and

the next point taken, and so on until the threshold is exceeded. The abscissa is then normalized to [0, 1] and the points interpolated into a smooth curve by a cubic spline. This is known as a *normalized arc length* parameterization. A second curve is produced via the same process using the m_2-value of the point *p*. The two quadrature-based variables are calculated by integrating each curve over the interval [0, 1].

The ten variables are most likely not completely independent. With a data set of this size, it is nearly impossible to estimate the correlations meaningfully. The first eight variables were selected through exploratory analysis to characterize the MER. The last two variables are related to the shape of the track as opposed to its location and orientation in m_1-m_2 space and are therefore likely to have a high degree of independence from the other eight.

Figures 5A–C illustrate the discriminatory power of these variables. Figures 5A and 5B represent two different utterances of *ayo* by speaker JT. The first plot in each figure is the track in moment space. The second and third plots are the normalized arc length parameterizations for m_1 and m_2. (The actual variable used will be the quadrature of these curves.) The similarity in shape of corresponding plots for the same-speaker utterances is evident. Figure 5C is the set of plots for the utterance of *ayo* by speaker BB. The different curve shapes in Figure 5C indicate that a different person spoke.

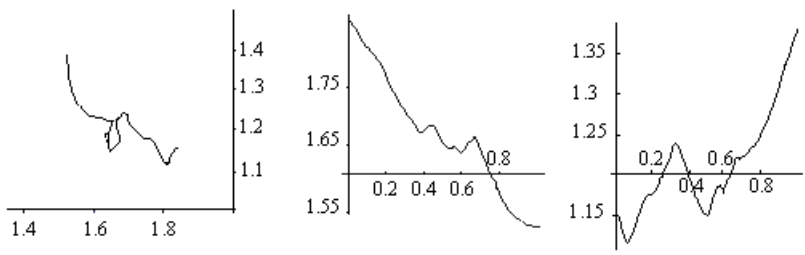

Figure 5A JT speaking *ayo*-1: the m_1-m_2 track, the normalized arc length parameterization of m_1, the normalized arc length parameterization of m_2

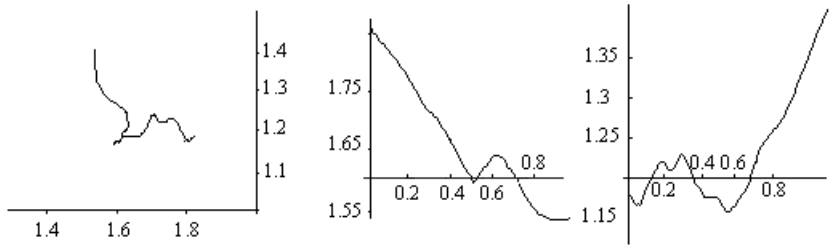

Figure 5B JT speaking *ayo*-2: the m_1-m_2 track, the normalized arc length parameterization of m_1 the normalized arc length parameterization of m_2

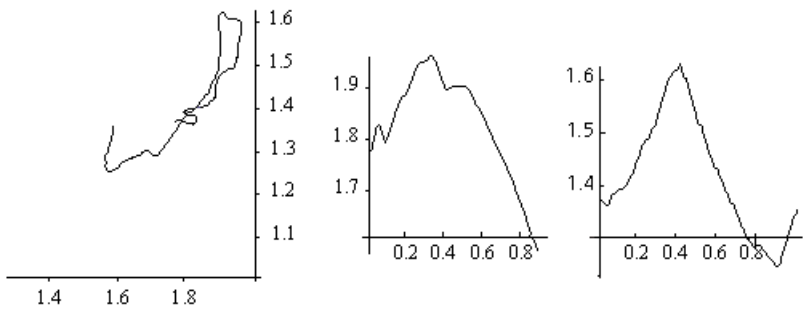

Figure 5C BB speaking *ayo*: the m_1-m_2 track, the normalized arc length parameterization of m_1, the normalized arc length parameterization of m_2

A CASE STUDY

The experiment

From an imagined extortion threat containing 'Now we see about the payola', we identified three potential isolexemes: *owie, eya,* and *ayo*, as might be isolated from the underlined parts of the exemplar. Single utterances of *owie, eya,* and *ayo* were extracted from the speech of ten unknown speakers. The set consisted of eight males of whom five were native speakers of American English, and three were near accent-less fluent English speakers whose native language was Venezuelan Spanish. The two females were both native speakers of American English. This is the testing database. We then asked the ten speakers – BB, BS, DM, DS, JT, KB, LC, NM, RR, VN – to utter *owie, eya,* and *ayo* four times to simulate

the results of interrogations in which those sounds were extracted from the elicited dialogue. This is the training database. All the speech samples were processed to create tracks as described in the 'Creating "tracks"' subsection above.

The objective of the experiment is to see if the 10 features described in the previous subsection are useful in discriminating among individuals. The approach of using several variables to distinguish between groups or classes is referred to as discriminant analysis. (See, for example, Mardia *et al.* 1997.) As mentioned in the 'Background' section above, many authors have employed methods such as neural networks and hidden Markov models to discriminate between individuals. (See Klevans and Rodman (1997) for a general discussion.) A disadvantage of these methods is that they require a large amount of training data. We present a fairly simple discriminant analysis, which is easily implemented and can be used with a small amount of training data.

Determining effective discriminators

The set of variables described in the 'Extracting features from tracks' subsection above seemed to capture important features of the *ayo*, *eya* and *owie* tracks. We therefore used an analysis of variance (ANOVA) to confirm that these variables are effective in discriminating between individuals. ANOVA is a method of comparing means between groups (see, for example, Snedecor and Cochran 1989). In this case, a group is a set of replicates from an individual. If the mean of a feature varies across individuals, then this variable may be useful for discriminating between at least some of the individuals. In an ANOVA, the F-statistic is the ratio of the interspeaker variation to the intraspeaker variation. If this ratio is large (much larger than one), then we conclude that there is a significant difference in feature means between individuals.

Table 1 contains the F-statistics for each of the ten variables described in the 'Extracting features from tracks' subsection. In this analysis, each variable is considered separately, so the F-statistic is a measure of a variable's effectiveness in distinguishing individuals when used alone. For these data an F-statistic larger than 2.2 can be considered large, meaning the variable will be a good discriminator. Note that a large F-value does not imply that we can separate *all* individuals well using the single feature; however, it will be useful in separating the individuals into at least two groups. All of the variables discussed in the 'Extracting features from tracks' subsection have a large F-statistic. (Indeed, we used the F-statistic to eliminate as ineffective such potential variables as the length of the short side of the MER.)

Table 1 F-statistics for each variable

VARIABLE NAME	AYO	EYA	OWIE
Midpoint-x	14.93	34.09	19.56
Midpoint-y	23.25	39.52	33.78
Long side	9.30	13.27	4.61
Alpha	7.66	2.31	4.61
Maximum-x	18.31	6.02	12.61
Minimum-x	13.51	16.11	20.83
Maximum-y	11.95	57.17	18.90
Minimum-y	19.89	18.85	49.11
Mean Quadrature	34.14	32.13	30.77
Variance Quadrature	65.44	48.75	67.49

Measures of similarity

Having determined that all ten features are useful for all three sounds, the discriminant analysis will be based on these 30 variables. Let $\bar{y}_i$ be the 30-dimensional vector of sample means for speaker *i*. In our training database, this mean is based on four repetitions for each speaker. It will be easy to discriminate between individuals if the $\bar{y}_i$'s are 'far apart' in 30-dimensional space. One way to measure the distance between means is to use Euclidean distance. However, this metric is not appropriate in this situation because it does not account for differing variances and covariances. For example, a change in one unit of the angle of orientation variable α is not equivalent to a change of one unit of a quadrature-based variable. Also, with a one-unit change in maximum-y, we might expect a change in minimum-y or the long side variables. Mahalanobis distance is a metric that accounts for variances and covariances between variables (see, for example, Mardia *et al*. 1979).

Let $\sum$ be the 30x30 dimensional covariance matrix. We will partition $\sum$ into nine 10x10 matrices, six of which are distinct. The matrix has the form

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AE} & \Sigma_{AO} \\ \Sigma_{AE} & \Sigma_{EE} & \Sigma_{EO} \\ \Sigma_{AO} & \Sigma_{EO} & \Sigma_{OO} \end{pmatrix}$$

For example, the submatrix $\sum_{AA}$ represents the covariance matrix of the ten variables associated with the *ayo* sound. The submatrix $\sum_{AE}$ represents the covariance matrix of the ten *ayo* variables and the ten *eya* variables. We make two assumptions about the structure of this matrix. First, we assume

that the diagonal submatrices are constant for all individuals, so that $\sum_{AA}$ $\sum_{EE}$ and $\sum_{OO}$ can be estimated by pooling the corresponding sample covariance matrices across individuals. This is a fairly strong assumption, but with the size of the training data set, we cannot reliably estimate these matrices separately for each individual without making even more stringent distributional assumptions.

Secondly, we assume that $\sum$ has a block diagonal structure. That is, the matrices $\sum_{AE}$, $\sum_{AO}$ and $\sum_{EO}$ are assumed to be matrices of zeros. This is also a strong assumption, but again, the size of the training data set does not allow for reliable estimation of these submatrices. Let $\hat{\sum}$ be the estimate of $\sum$ using the zero matrices and estimated matrices described above. The squared Mahalanobis distance between individuals *i* and *j* is

$$M_{ij} = \left(\bar{y}_i - \bar{y}_j\right)^T \hat{\Sigma}^{-1}\left(\bar{y}_i - \bar{y}_j\right)$$

Table 2 contains squared Mahalanobis distances for the ten individual means in the training database. The lower triangle of the table is blank because these cells are redundant. Relatively small distances indicate that the individuals are similar with the respect to the variables used in the analysis. For example, the most similar individuals are KB and NM (the two female speakers) while the most dissimilar are DM and RR (both native speakers of American English).

Table 2 Squared Mahalanobis distances between individual means

Individual	**BB**	**BS**	**DM**	**DS**	**JT**	**KB**	**LC**	**NM**	**RR**	**VN**
BB	0	178	589	165	346	194	107	341	277	132
BS		0	322	228	192	148	224	202	271	145
DM			0	637	243	258	670	134	708	642
DS				0	533	215	137	259	387	112
JT					0	216	406	179	447	467
KB						0	285	98	475	248
LC							0	427	169	119
NM								0	586	374
RR									0	297
VN										0

Classifying exemplars

For features extracted from a set of three utterances (*ayo, eya, owie*) from a speaker, we can calculate the squared Mahalanobis distance from the exemplar to each individual mean by

$$M_{Xi} = (y_X - \bar{y}_i)^T \hat{\Sigma}^{-1} (y_X - \bar{y}_i)$$

For the closed-set problem, we identify S_x by choosing the individual mean to which y_x is closest. We first tested this identification rule on each exemplar in the training set. The rule correctly identified the speaker for all training exemplars. We would expect to have a low error rate in this case, since each exemplar was also used in estimating $\bar{y}_i$ and $\hat{\Sigma}$.

The rule was also applied to unknowns 1–7 in the testing database. These exemplars came from speakers in the training set. (Unknowns 8–10 were 'ringers' introduced for the open-set test. They consisted of one male and two female native speakers of American English, replacing one female and one male native speaker of American English, and one male speaker whose native language was Venezuelan Spanish.) Table 3 contains the squared Mahalanobis distances from each y_x to each individual mean. Each speaker was identified correctly. For each unknown(1–7), the minimum distance is less than 100, except for Unknown 6. The asterisk marks the minimum distance for unknowns 8–10. The minimum distances are lower, in general, than the interspeaker distances given in Table 2. This confirms that this set of variables is useful for discriminating between individuals. Also, the distances from each speaker in the test set seem to follow approximately the same trends as in Table 2. For example, in the training data, DM was the most dissimilar to BB. For Unknown 1 (BB), the largest distance is to DM.

In many cases, it will be desirable to report not only the individual selected by the rule, but also to provide an estimate of the reliability of the procedure. The reliability may be determined empirically. We may use the observed error rate for the closed-set classification rule when applied to the training data and test speakers 1–7. Cross-validation can also be used to estimate error rates. However, due to the size of the study, neither method will provide a reasonable estimate of reliability of the procedure. Another method of estimating reliability would be to make distributional assumptions, e.g. multivariate normality. Any such assumptions would be difficult to verify with such a small data set. Developing a framework for estimating the reliability of such a procedure with a small data set is planned for future work.

For the open-set problem, the rule must be modified to allow us to conclude that the unknown speaker is not in the training set ($X=0$) One way of modifying this rule would be to establish a distance threshold. If none of the distances M_{xi} fall below this threshold, then we conclude $X=0$. As in the closed-set problem, estimates of reliability are desirable. In general, error rates will depend on the choice of the threshold.

Table 3 Squared Mahalanobis distances between unknowns and individual means

Unknown	BB	BS	DM	DS	JT	KB	LC	NM	RR	VN
1-BB	**72**	157	595	303	324	228	141	357	212	171
2-BS	222	**37**	398	273	295	177	262	250	327	145
3-DM	539	298	**46**	603	261	246	601	129	685	551
4-DS	263	327	643	**54**	621	237	212	422	542	248
5-JT	326	176	213	489	**68**	211	406	183	406	435
6-KB	187	140	394	277	260	**122**	319	190	437	241
7-NM	278	191	153	335	239	65	400	**27**	596	325
8-RB	198	213	276	271	199	126*	184	**298**	280	294
9-SG	119	91	294	188	241	66*	179	163	**347**	150
10-TM	221	140*	452	364	215	234	173	308	154	**222**

We investigated empirical choices of thresholds for this experiment. For the test data set, if we choose a distance threshold, we will misclassify at least one of the ten unknowns. For example, if we choose a distance threshold of 100, Unknown 6 will be incorrectly assigned to S_0 and Unknown 9 will be incorrectly classified as KB. In this testing situation, we can pick a distance threshold that minimizes the number of misclassification errors. However, this will not be possible in a practical situation. A framework for choosing thresholds for the open-set problem is planned for future work.

SUMMARY AND FUTURE DIRECTIONS

The results we obtained are encouraging because of the sparseness of data. The known speakers had about 8–12 seconds of speech data per speaker. The unknowns had one-quarter of that amount, 2–3 seconds. In an actual forensic situation there is an excellent likelihood of having many times the amount of data for the criminal exemplar (unknown speaker), and as much data as needed for suspects (known speakers).

The identification process is cumulative in nature. As additional data become available, there is more information for individuating speakers, and the error probabilities diminish. In practice the only limitation is the amount of data in the criminal exemplar (the testing data). Often, authorities are able to collect as large an amount of training data as needed. Each new sound sequence that undergoes analysis makes its small contribution to the overall discrimination. In even as short an utterance as 'There's a bomb in Olympic Park and it's set to go off in ten minutes' there are easily a dozen or more sequences that may be extracted for analysis. Thus we are sanguine about the ability of this method to work in practice.

When the case study is regarded as closed-set speaker identification, the system performed without error. While it is unreal to expect zero error rates in general, the results forecast a relatively low error rate in cases of this kind. Many practical scenarios require only closed-set identification. For example, in a corporate espionage case, where a particular phone line is tapped, there are a limited number of persons who have access to that phone line. Similar cases are described in Klevans and Rodman (1997).

The more difficult and more general open-set identification yielded error rates between 10 and 20 per cent depending on how thresholds are set. Our current research is strongly concerned with reducing this error rate.

Future research: short term

Our research in this area is expanding in three directions. The first is to use a larger quantity of data for identification. Simplistically, this might have ten repetitions of ten vowel transitional segments similar to *owie* for the training database. It is expected that the F-values of the variables would rise, meaning that the ratio of interspeaker variation to intraspeaker variation will climb. At one time we used only three utterances per sound per speaker in the training base and when we went to four utterances the F-values increased significantly, which validates our expectation. (Naturally this implies lengthier interrogating sessions in a forensic application, but when a serious crime is involved, the extra effort may be justified.)

The second direction is to use more phonetically varied data. The vowel transitions of this study were chosen primarily to determine if the methodology was promising. They do not span the entire moment space encompassed by the totality of speech sounds. There are speech sounds such as voiced fricatives that produce tracks that extend beyond the union of the MERs for the above utterances. Moreover, we are also able to process voiceless sounds to produce moment tracks, but using a different processing method that analyses the speech signal at frequencies up to 11 kHz (Fu *et al.* 1999). We are also able to process liquid [l], [r] and nasal sounds [m], [n], [nʲ], [ŋ]. We hypothesize that the use of other transitions, for example, vowel-fricative-vowel as in *lesson*, will increase the discriminatory power of the method because it 'views' a different aspect of the speaker's vocal tract. An interesting, open, minor question is whether particular types of sequences (e.g. vowel-nasal-vowel, diphthong alone, etc.) will be more effective discriminators than others.

We are currently moving from producing our own data to using standardized databases such as those available from the Linguistic Data Consortium. While this makes the data extraction process more difficult and time-consuming, it has the advantage of providing test data of the kind encountered in actual scenarios, particularly if one of the many telephone-based databases are used.

The third direction is to find more and better discriminating variables. Eight of the ten variables are basically 'range statistics', a class of statistics well known for their lack of robustness and extreme sensitivity to outliers, and as noted above, are not entirely, mutually independent. Both more and varied data would obviate these shortcomings, but what is truly needed is a more precise measurement of curve shape, since the shape appears to be highly correlated to the speaker.

We are experimenting with methods to characterize the shape of a curve. The visual appearance of the shape of tracks for a given speaker for a given utterance, and the differences between the shapes of the tracks among speakers for the same utterance, suggest that curve shape should be used for speaker identification.

Curvature scale space (Mokhtarian and Mackworth 1986, Mokhtarian 1995, Sonka *et al.* 1999) is a method that has been proposed to measure the similarity of 2D curves for the purpose of retrieving curves of similar shape from a database of planar curves.

The method tries to quantify shape by smoothing the curve (the scaling process) and watching where the curvature changes sign. When the scaling process produces no more curvature changes, the resulting behaviour history of the changes throughout the smoothing process is used to do curve matching (Mokhtarian 1995). We are currently exploiting this methodology to extract variables that are linked to the shape of the moment tracks in m_1-m_2 space. These variables should provide discriminating power highly independent of the variables currently in use, and hence would improve the effectiveness of the identification process.

Other methods for exploiting shape differences are also being considered. Matching shapes, while visually somewhat straightforward, is a difficult problem to quantify algorithmically and methods for its solution have only recently begun to appear in the literature.

Future research: long term

Our long-term future research is also pointed in three slightly different directions. They are (1) noisy data, (2) channel impacted data, and (3) disguised voice data. All three of these data-distorting situations may compromise the integrity of a speaker identification system based on 'clean' data. A system for practical use in a forensic setting would need methods for accommodating to messy data. This is a vast and complex topic, and most of the work needed would necessarily follow the development of the speaker identification system as used under less unfavourable circumstances.

ACKNOWLEDGEMENT

The authors wish to acknowledge the editors for helpful assistance in improving the presentation of the foregoing work.

REFERENCES

Baldwin, J. R. and French, P. (1990) *Forensic Phonetics*, London: Pinter Publishers.

Bennani, Y. and Gallinari, P. (1991) 'On the Use of TDNN-Extracted Features Information in Talker Identification', *ICASSP* (International Conference on Acoustics, Speech and Signal Processing), 385–8.

Bolt, R. H., Cooper. F. S., David, E. E., Denes, P. B., Pickett, J. M., and Stevens, K. N. (1969) 'Identification of a speaker by speech spectrograms', *Science,* 166: 338–43.

Doddington, G. (1985) 'Speaker Recognition – Identifying People by Their Voices', in *Proceedings of the IEEE* (Institute of Electronics and Electronic Engineers), 73(11): 1651–63.

Falcone, M. and de Sario, N. (1994) 'A PC speaker identification system for forensic use: IDEM', in *Procedings of the ESCA* (European Speech Communication Association) *Workshop on Automatic Speaker Recognition, Identification and Verification,* Martigny, Switzerland, 169–72.

French, P. (1994) 'An overview of forensic phonetics with particular reference to speaker identification', *Forensic Linguistics,* 1(2):169–81.

Fu, H., Rodman, R., McAllister, D., Bitzer, D. and Xu, B. (1999) 'Classification of Voiceless Fricatives through Spectral Moments', in *Proceedings of the 5th International Conference on Information Systems Analysis and Synthesis* (ISAS'99), Skokie, Ill.: International Institute of Informatics and Systemics, 307–11.

Hollien, H. (1990) *The Acoustics of Crime: The New Science of Forensic Phonetics,* New York: Plenum Press.

Kao, Y., Rajasekaran, P. and Baras, J. (1992) 'Free-text speaker identification over long distance telephone channel using hypothesized phonetic segmentation, *ICASSP* (International Conference on Acoustics, Speech and Signal Processing), II.177–II.180.

Kersta, L. G. (1962) 'Voiceprint identification', *Nature,* 5(196): 1253–7.

Klevans, R. L. and Rodman, R. D. (1997) *Voice Recognition,* Norwood, Mass.: Artech House Publishers.

Koenig, B. E. (1986) 'Spectrographic voice identification: a forensic survey', *Journal of the Acoustical Society of America,* 79: 2088–90.

Koster, B. E. (1995) *Automatic Lip-Sync: Direct Translation of Speech-Sound to Mouth-Animation,* PhD dissertation, Department of Computer Science, North Carolina State University.

Künzel, H. J. (1994) 'Current Approaches to Forensic Speaker Recognition', in *Proceedings of the ESCA* (European Speech Communication Association) *Workshop on Automatic Speaker Recognition, Identification and Verification,* Martigny, Switzerland, 135–41.

Künzel, H. J. (1997) 'Some general phonetic and forensic aspects of speaking tempo', *Forensic Linguistics,* 4(1): 48–83.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*, London: Academic Press.

Markel, J. D. and Davis, S. B. (1978) 'Text-independent speaker identification from a

large linguistically unconstrained time-spaced data base', *ICASSP* (International Conference on Acoustics, Speech and Signal Processing), 287–9.

Matsui, T. and Furui, S. (1991) 'A text-independent speaker recognition method robust against utterance variations', *ICASSP* (International Conference on Acoustics, Speech and Signal Processing), 377–80.

Mokhtarian, F. (1995) 'Silhouette-based isolated object recognition through curvature scale space', *IEEE* (Institute of Electronics and Electronic Engineers) *Transactions on Pattern Analysis and Machine Intelligence*, 17 (5): 539–44.

Mokhtarian, F. and Mackworth, A. (1986) 'Scale-based description and recognition of planar curves and two-dimensional shapes', *IEEE* (Institute of Electronics and Electronic Engineers) *Transactions on Pattern Analysis and Machine Intelligence,* V. Pami-8 (1): 34–43.

Oglesby, J. and Mason, J. S. (1990) 'Optimization of Neural Models for Speaker Identification', *ICASSP,* 393–6.

O'Shaughnessy, D. (1986) 'Speaker Recognitiion', *IEEE* (Institute of Electronics and Electronic Engineers) *ASSP* (Acoustics, Speech and Signal processing) *Magazine,* October, 4–17.

Reynolds, D. A. and Rose, R. C. (1995) 'Robust text-independent speaker identification using Gaussian mixture speaker models', *IEEE* (Institute of Electronics and Electronic Engineers) *Transactions on Speech and Audio Processing,* 3(1): 72–83.

Rodman, R. D. (1998) 'Speaker recognition of disguised voices', in *Proceedings of the COST 250 Workshop on Speaker Recognition by Man and Machine: Directions for Forensic Applications,* Ankara, Turkey, 9–22.

Rodman, R. D. (1999) *Computer Speech Technology*, Boston, Mass.: Artech House Publishers.

Rodman, R., McAllister, D., Bitzer, D., Fu, H. and Xu, B. (1999) 'A pitch tracker for identifying voiced consonants', in *Proceedings of the 10th International Conference on Signal Processing Applications and Technology* (ICSPAT'99).

Rodman, R., McAllister, D., Bitzer, D. and Chappell, D. (2000) 'A High-Resolution Glottal Pulse Tracker', in *International Conference on Spoken Language Processing (ICSLP)*, October 16–20, Beijing, China (CD-ROM).

Rudasi, L. and Zahorian, S. A. (1991) 'Text-independent talker identification with neural networks', *ICASSP* (International Conference on Acoustics, Speech and Signal Processing), 389–92.

Snedecor, G. W. and Cochran, W. G. (1989) *Statistical Methods* (8th edn), Ames, IA: Iowa State University Press.

Sonka, M., Hlavac, V. and Boyle, R. (1999) *Image Processing, Analysis, and Machine Vision* (2nd edn), Boston, MA, PWS Publishing, ch. 6.

Stevens, K. N., Williams, C. E., Carbonelli, J. R. and Woods, B. (1968) 'Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material', *Journal of the Acoustical Society of America,* (43): 1596–1607.

Tosi, O. (1979) *Voice Identification: Theory and Legal Applications,* Baltimore, Md.: University Park Press.