Proceedings of The Institute of Acoustics

PHILOSOPHY OF SPEAKER IDENTIFICATION

S. R. Lewis

SRDB Home Office

1. Introduction

The most common speaker identification situation confronting the Police is one of receiving a tape recording made in connection with some criminal activity, for example blackmail, kidnapping or fraud, for which they have one or more suspects and they wish to determine if the voice of a suspect is one and the same voice as that on the recording. More precisely they would like to be able to infer from two examples of speech whether they were produced by the same individual speaker or not. The immediate question this raises is common to most identification evidence: Given evidence X, what is the probability it was produced by Y? A direct scientific answer can only be given in specific cases (1). On the basis of Bayesian inference, if the a priori odds, O_o, that an hypothesis such as "Y produced evidence X" is true on the basis of evidence, I, then the posterior odds, O_1, given additional evidence X is given by

$$O_1 = \frac{P(X|Y,I)}{P(X|\bar{Y},I)} O_o$$

Where $P(X|Y,I)$ is the conditional probability of obtaining evidence X given the other evidence, I, and hypothesis Y is true, and $P(X|\bar{Y},I)$ is the same conditional probability but assuming the alternative hypothesis, $\bar{Y}$, eg "Y did not produce evidence X", is true.

Normally evidence does not consist only of scientific data. In general, science can only provide additional information to assist with an answer that must be ultimately arrived at inductively. Speaker identification is the epitome of this problem. If X is a recording made in connection with a crime and Y is a suspect on the basis of information, I, can science contribute a meaningful quantity for the above factor relating the posterior odds to the a priori odds?

Machines are being developed that recognise words in speech or objects in images, ie they have identification ability. It is unlikely that such pattern recognition machines will be developed that can compete with human performance at most identification tasks. However, this may not be the case for all, and in particular those where we have most doubt about human performance, eg identification of a person from his handwriting or voice. Should a machine be produced that for a sample of speakers could be shown to perform the speaker identification task better than human listeners, would that contribute to deciding if two utterances were made by the same speaker in general? Clearly to make such a comparison it is necessary to define quantitatively what is meant by better performance for both man and machine.

PHILOSOPHY OF SPEAKER IDENTIFICATION

2. Concept of Identity

Identity is a human concept which arises directly from our experience of the physical world. Objects, specific or generic, are defined from our grouping or organising the memory of past experiences according to common patterns contained within them. It is these personal definitions of objects which we refer to as the "identities" of these objects. We can know nothing external to ourselves except via our sensory data. Therefore identities must be established inductively.

A fundamental purpose for this organisation of memory is to enable us to relate present experiences to past experiences. Another is to attempt to predict or infer knowledge associated with the personally unknown or unexperienced, whether from the past or the future. These two purposes must be clearly distinguished if we are to prevent confusion in our understanding of what is meant by identification evidence. Relating present to past experience is done by deductive reasoning. Those deductions must be based on the identities established inductively. They must assume the identities established in the past are correct until a contradiction with the present is recognised. Consequently identities and therefore deduced relationships based upon them can and do change with added experience. Predicting the future or inferring knowledge of the personally unexperienced past, such as in the forensic situation, may be done using both deductive and inductive reasoning.

It is essential to be clear throughout any analysis of the identification process what has been based on inductive reasoning and what has been based on deductive reasoning. An attempt to clarify these will be constantly made. The contribution of science is in the realms of deductive reasoning. All we can say is that usually its results are a major influence on the conclusions of inductive reasoning, and often determine many of the constraints by which they are reached. The identification process is so closely associated with inductive reasoning that its part must be included in any description of identification evidence. For most identification tasks the identity involved is based on non-statistical features that are essentially fixed. The inductive process involved in comparing two such identities at the same time is not usually consciously or critically considered. However, it forms the basis for observing scientific information such as the statistical probability of an identity occuring in a population. In the case of speaker identification no such features are known. Repetitions of the same utterance by the same speaker vary from utterance to utterance. There is within-speaker variation as well as between-speaker variation. For speaker identification the inductive process of determining and comparing the identities of voices has itself uncertainty associated with it.

3. Inductive and Statistical Probabilities

It is recognised by many statisticians that we operate two different concepts of probability, statistical and inductive (2). The basis for many inductive probabilities is Laplace's Principle of Indifference which involves "reducing all events of the same kind to a certain number of cases equally possible,

that is, such that we are equally undecided as to their existence". Statistical probability is the observed relative frequency with which an event occurs, in the limit of the sample in which it is observed approaching that of the full population. In relation to identification evidence both types of probability are used, a fact which is not always recognised. It is the operation of inductive probability that causes the greatest difficulty, probably because it is outside the realms of science and quantitive analysis.

Inductive probability is essentially a statement of our opinion of how likely it is that an event will occur or has occurred. It depends heavily on the state of our knowledge, but even when this is the same, for a group of people, it will in general differ from individual to individual. For example, most persons presented with a die would assign a probability of 1/6 to any particular side falling uppermost at the next throw. This is an example of an inductive probability based on Laplace's Principle. We may observe the statistical probability by throwing the die a large number of times. It is unlikely to be precisely 1/6. Some gamblers knowing more about the manufacture of a particular die may work with significantly different values of inductive probability for each possible outcome. However, a wise gambler who does not have access to that information would remain at a disadvantage only as long as it took to observe a representative number of outcomes. Wise gamblers know the superiority of statistical probability over inductive probability. Inductive probability can differ significantly from statistical probability when the latter has not been observed, but once observed most reasonable persons would equate their inductive probability to the statistical result. That however would depend on factors such as how representative he felt the sample had been of the population as a whole.

A more important example of inductive probability is of course identification evidence provided by a witness in association with a crime. The identification is deduced from identities established inductively, and the reliability placed on that evidence is the inductive probability assigned to it by all individuals including the witness, after considering it is association with any other information they personally know and decide is relevant.

4. Speaker Identification by Induction

Speaker identification evidence used by the police consists of either the opinion of a phonetician, a person familiar with the suspect's voice, or a witness. It is important to appreciate that there are in fact four inductive probabilities and four identities involved. The latter are: Vc, the voice associated directly with the crime; Vs, the voice of the suspect; Ic, the criminal with the voice Vc; and Is, the suspect with the voice Vs. The four probabilities are the conditional probability of Is given Vs, P (Is/Vs); the probability that Vs and Vc are the same, P(Vs=Vc); the conditional probability that Is is Ic given P(Vs=Vc), P(Is=Ic/P(Vs=Vc)); and the probability that the person giving his opinion will in fact be correct. To clarify the meaning to be associated with these, comparison will be made to fingerprint and blood group identification evidence which is better understood.

Here the P(Is|Vs) will normally be equal to 1 provided the recording of the suspect is directly witnessed. None the less it should be remembered that that is not absolutely certain always to be the case. In relation to fingerprints the equivalent term is the probability the fingerprint believed taken from the suspect was indeed that print. It would take a most unlikely administrative error for that not to occur. However in the case of speaker identification the voice of the suspect may have been observed remotely and that direct relationship may not be assumed with such confidence. This relationship is necessarily considered inductively.

P(Vs=Vc) presents a far more difficult decision, and corresponds to the strength of conviction of the person giving an opinion on whether two utterances could have been made by the same speaker. This is analagous to the decision about whether two fingerprints match or two blood samples have the same blood group. The decision is made by human observation ie by induction. For blood and fingerprint evidence it only becomes of great concern when the sample associated directly with the crime is difficult to observe. This problem also occurs in association with speaker identification, e.g. a poor quality recording or a very short utterance. However even when we believe we have high quality recordings of long representative utterances it still does not follow that we would assign it the value 1 or 0. The corresponding decision for all identification evidence has always been essentially an inductive one. Even with the advent of successful automatic fingerprint searching the final selection and decision on whether there is a match is performed manually.

For the speaker identification evidence in present use the associated P(Is=Ic|P(Vs=Vc)) is inductive. In the case of fingerprint evidence and blood group evidence it is statistical. If the result of the conditional information is 0 (determined inductively), ie there is no match, the equivalent probability (determined statistically) will also be 0. The same is true for speaker identification although both the conditional information and the P(Is=Ic|P(Vs=Vc)) are determined inductively. This is simply stating that, if the accepted view is that the voices being compared are considered definitely dissimilar, then we would say they were certainly made by different speakers. When there is a positive match of blood groups, then dependent on the circumstances, an appropriate statistical probability is assigned. In the case of fingerprints a value of 1 is generally accepted, reflecting the fact that no two persons have ever been observed with identical fingerprints. To be precise it could be a value extremely near to 1. In the case of speaker identification the need to distinguish between P(Is=Ic|P(Vs=Vc)) and P(Vs=Vc) needs emphasis. Often it is assumed that if two voices are judged the same then that necessarily implies that they may only be associated with one speaker ie P(Is=Ic|P(Vs=Vc)=1)=1. The statistical value of P(Is=Ic|P(Vs=Vc)) is in principle the relative frequency of distinguishable voices in the speaking population. As with fingerprints or blood we could observe this for a sample of the population if the inductive probability P(Vs=Vc)=1 or 0 and we always knew which. Unfortunately that is not the case. The problem is that we have a degree of uncertainty in the inductive process that forms our basis for observing such a statistic.

PHILOSOPHY OF SPEAKER IDENTIFICATION

The final probability relates to the question of quantifying human performance. At present that remains essentially a matter of opinion. There have been a few studies of human performance on small numbers of voices with highly variable results. For example an early study (3) made in response to the acceptance of aural speaker identification evidence by a court in 1935, a panel of listeners heard 5 unknown speakers read a paragraph of text. The listeners were able to identify the speakers in 83% of trials one day later. However this fell to 69% after a two week delay. Several other studies have measured the performance of listeners under different circumstances and observed error rates of between 2% and 15% (4,5,6,7). The sample of voices used, because of our limited capacity to retain such data, is necessarily too small to be considered representative of speakers in general. This is considered further below. We simply do not know at present how close inductive probability is to statistical probability for human speaker identification. Reliance is necessarily placed on the specialised experience and listening skills of phoneticians or the familiarity of a suspect's acquaintances without quantitative assessment. That we do have some capability to identify speakers from their voice alone is indisputable, but how much is unquantified.

Some years ago speech spectograms were introduced in the USA as speaker identification evidence. The features used for comparison were not adequately defined by their proponants to be subjected to objective analysis. They may therefore be viewed as a technique to convert the process from an aural comparison to a visual comparison. The process remained inductive. The technique is not accepted as a reliable method of speaker identification by many speech scientists (8).

5. Speaker Identification by Machine

Speaker verification machines that confirm a person's identity from his voice, have been successfully implemented. One system for example has operated for more than 5 years with an imposter acceptance rate of less than 2% and user rejection rate of 0.5% (9). However, it does not correspond to the speaker identification problem where the voice should in principle be compared with all possible speakers' voices. In a speaker identification or verification machine a precise definition of the identity of a voice is expressed as a set of features measured from the speech waveform. Many alternative features sets have been studied in association with machine recognition of speakers (10). They are not based on any understanding of how humans perform the task, as little is known in that respect. Most are based on conventional signal processing techniques and models of the speech production process developed in association with vocoders, speech compression, analysis and synthesis. They define multidimensional statistical distributions for each speaker in feature space. These are estimated or represented by training samples taken from a library of speakers (11). Using the techniques of statistical pattern recognition a speaker identification machine attempts to decide on an objective basis which speaker in its library was most likely to have produced a given test pattern representing an unknown voice (12). It may decide the sample was not produced by any of the speakers and offer no decision. A speaker verification system makes the comparison with the library data for

only one speaker and decides whether the test sample could have been made by that speaker.

Many speaker identification systems have been developed which work with recognition rates in excess of 97%, but with libraries of fewer than 30 speakers and high quality recordings, ie unrepresentative of the forensic situation. However, some have been evaluated with larger library sizes and "telephone quality" data. With data recorded under "telephone booth" conditions from 172 speakers speaking 5 isolated digits one system produced recognition rates of 84% with single words and 94% with pairs of words (11). Another study used data from 50 Polish and 50 American speakers (13). Little difference was observed due to language and average recognition rates of 95% for wide bandwidth data (80Hz to 12.5 KHz) and 76% for telephone bandwidth data (315 to 3150 Hz) were obtained. Telephone bandwidth data from 25 subjects recorded under three conditions: normal, stress and disguise, resulted in 88%, 68% and 32% recognition rates respectively. This system, using features based on long time averaged spectra, was therefore very susceptible to disguise. Parameters based on auto correlation coefficients, contours and statistics of pitch, signal energy, cepstrum coefficients and LPC coefficients are other examples of features used in speaker identification machines.

Figure 1 represents schematically the conditional probability distributions for obtaining pattern X given it was produced by speaker Y, where Y=A, B...H, mapped along one dimension. Within this small population we could say the voices of speakers A,F,G and H have unique identities as defined by the particular feature set used. However there is uncertainty associated with distinguishing between the remaining speakers. The performance of a speaker identification system will clearly be very sensitive to the library set chosen. In order to be representative of the population at large a library of speakers would have to contain all generic groupings affecting the voice such as age, dialect, sex, social class etc with the same relative frequency as the full population. This selection would necessarily be subjective. The library would have to be impractically large before it was likely to be regarded as representative of all speakers. Speaker identification by induction is also performed with a library. It corresponds to all voices experienced and remembered by the person giving their opinion. How representative that is of all speakers would also appear to be questionable.

Given that it is necessary to work with an unrepresentive sample of all speakers and that performance depends primarily on the particular sample used, it would appear necessary to define samples accepted as representative of subsets of the population sharing generic identities. Membership of a subset may be determined inductively when we are confident that we can recognise the attributes defining the group. For example we would normally be confident that we can objectively define and determine the language, age, social class, sex and recent domicile to be associated with a speaker. However other attributes such as dialect and intonation are less clearly definable or determined objectively. Another approach may be to attempt to define samples of similar voices by machine. A pattern recognition technique known as clustering analysis (12) may be used to determine any natural groupings of pattern distributions in the feature space used by the machine. Alternatively

pattern recognition techniques of feature extraction and selection may be used to objectively search for feature sets useful for recognition of such attributes, although it will still be necessary to train the machine with patterns labelled by attributes assigned inductively. The ability to reliably recognise such generic features in voices would also be useful to the police in the investigation of a crime.

The relative performance of a speaker identification machine, phonetician, or unskilled listener may be objectively determined by observing the recognition rate for a small sample of speakers, but only if the same sample is used every time. This may indicate the appropriate procedure to follow in the provision of speaker identification evidence in the forensic situation. But unlike most other identification evidence, reliable statistical information applicable to a positive match would only appear possible if subsets of similar voices can be defined. However, definite negative matches remain possible and the hypothesis that the same speaker may have made a recorded utterance may still be tested.

PHILOSOPHY OF SPEAKER IDENTIFICATION

1. I EVETT 1982, JOURNAL FORENSIC SCIENCE SOC. 23, PAGES 35-39
What is the Probability that This Blood Came from That Person?
2. M G BULMER 1967, 2ND ED. OLIVER & BOYD EDINBURGH & LONDON
Principles of Statistics.
3. F MCGEHEE 1944, JOURNAL GEN. PSYCHOL. 31, PAGE 53
An Experimental Study in Voice Recognition
4. A M COLLINS 1973, Dept Eng. Physic, Australian National University, Canberra. Dept EP RR37
5. K STEVENS et al 1968, JASA 44, PAGE 1596
Speech Authentication and Identification: A Comparison of Spectrographic and Auditory Presentation of Speech Material
6. A E ROSENBURG 1973, IEEE TRANS. AUDIO ELECTRO ACOUST AU 21, PAGE 221
Listener Performance in Speaker Verification Tasks
7. P D BRICKER & S PRUZANSKY 1968, JASA 40, PAGE 1441
Effects of Stimulus Content and Duration on Talker Identification
8. R H BOLT et al 1970, JASA 47, PAGE 597
Speaker Identification by Speech Spectrograms: A Scientist's View of its Reliability for Legal Purposes.
9. B M HYDRICK & G R DODDINGTON 1978, PAPER NNN24 JASA 64 SUPPL. FALL
Performance Evaluation of Speaker Verification in Entry Control
10. B S ATAL 1976, PROC. IEEE 64, PAGE 456
Automatic Recognition of Speakers From Their Voices
11. P D BRICKER et al 1971, BELL SYST J50, PAGE 1427
Statistical Techniques for Talker Identification
12. P A DEVIJVER & J KITTLER 1982, PRENTICE-HALL INC LONDON
Pattern Recognition: A Statistical Approach
13. H HOLLIEN & W MAJEWSKI 1977, JASA 62, PAGE 975
Speaker Identification by Long Term Spectra Under Normal and Distorted Speech Conditions

PHILOSOPHY OF SPEAKER IDENTIFICATION

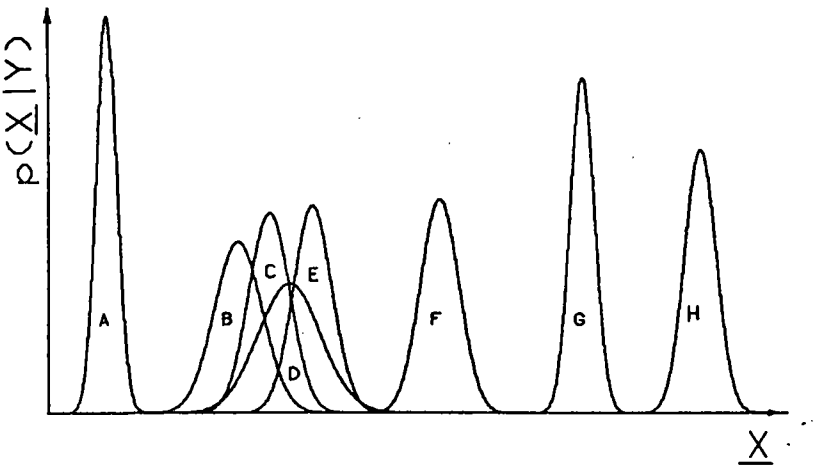

FIG 1 Hypothetical distributions for individual speakers.