DNA CAN'T TALK - SOME FACTS ABOUT FORENSIC SPEAKER IDENTIFICATION[1]

Phil Rose (Australian National University)

Summary

In this paper I describe some important ways in which real Forensic Speaker Identification differs from its portrayal in popular TV and film: (1) the kind of evidence used; (2) the role of computers; and (3) the logically correct expression of the outcome. These give an idea of the kind of work involved in real Forensic Speaker Identification. I hope that some of this is news to at least some of you!

Introduction

Forensic Speaker Identification (FSI) typically involves the comparison of one or more samples of an unknown voice with one or more samples of a known voice. Often the unknown voice is that of the individual alleged to have committed an offence, and the known voice belongs to the suspect. Obviously, both prosecution and defence are then concerned with being able to say on the basis of the evidence whether the two samples have come from the same person, and thus be able either to identify the suspect as the offender, or eliminate them from suspicion.

The way FSI is done in real life departs considerably from its depiction in film and popular TV crime shows like *Law and Order* and *Crime Scene Investigation*. On the TV, the speech samples are typically compared with a computer. The button is pushed and the computer promptly tells the investigators that there is a match (or not, as the case may be). The match implies that the offender sample was spoken by the suspect, who is convicted. Below I explain how this differs from reality.

It's a match!

Because of the nature of human voices, speech samples from different speakers will differ, but so will samples from the same speaker, even if they are repeats of the same thing said right after another. We never say the same thing in exactly the same way twice. Moreover, human speech is produced by vocal tracts of similar dimensions, and speech has a function - that of communication. There is thus a limit on the amount of between-speaker variation that can exist. Therefore there can never be an exact 'match' between speech samples. Rather, FSI involves an attempt to classify the inevitable differences between the offender and suspect speech samples as *same-speaker differences* or *different-speaker differences*. Again, because of the nature of speech and the real-world context, this classification is probabilistic, not absolute.

But, probably counter to your expectations, the probabilities that the FSI expert must try to estimate are *not* the probabilities of *hypotheses given the evidence* like "... given the high degree of similarity between these two speech samples, there can be very little doubt that they are from the same speaker”, or "... it is highly likely, given the extensive differences between the speech samples, that they are from different people." Although that is certainly what is expected of the expert from the courts and police, it will probably come a surprise to the reader that it is logically not usually

[1] Paper given and distributed at the *2002 Media Science Forum – The Science of Terrorism*, University of Technology, Sydney, 29th October

possible for the identification expert to quote the probability of a hypothesis given the evidence. The sections below explain why.

Similarity versus typicality: the Likelihood Ratio

The FSI expert is concerned first and foremost with estimating a crucial measure called the *Likelihood Ratio* (LR). The LR quantifies the strength of the scientific evidence. The bigger the LR is than one, the more the evidence supports the prosecution hypothesis, and the smaller it is than one, the greater the strength of support for the defence.

In FSI, it helps to think of the LR as the ratio of the degree of *similarity* between the offender and suspect samples to the degree of *typicality* of the offender and suspect samples in the relevant population. Then the more similar the two samples are, the more likely they are to have come from the same speaker and the higher the ratio. But this must be balanced by their typicality. The more typical the samples, the more likely they are to have been taken at random from different speakers in the population, and the lower the ratio. The value of the LR can thus be seen to be an interplay between the two factors of similarity and typicality. *Both factors are needed to evaluate forensic-phonetic evidence*: it is a very common fallacy, and one we also see on the TV, to assume that similarity is enough: that if two speech samples are similar, that indicates common origin; or, if they are different, that indicates different speakers. But the likelihood ratio makes it clear that similarity is only half the story: typicality is equally important.

A typical, correct, forensic conclusion incorporating a LR of 100 would be "I have compared the suspect and offender speech samples and concluded that you would be one hundred times more likely to observe the difference between them had they come from the same than different speakers." Note there is no inference from the LR, which has to do with the *probability of the evidence*, to the *probability of the hypothesis*.

It is important to understand that, because you are one hundred times more likely to observe the difference assuming same speakers does not mean that it is one hundred times more likely that it is the same speaker. Thinking like this is called the 'prosecutors' fallacy', or 'transposing the conditional'. That the probability of the evidence given the hypothesis is not the same as the probability of the hypothesis given the evidence can be seen from working out the following. What is the probability of its having four legs if it's a cow? Barring genetic and road accidents, nearly 100%. Is that the same as the probability of it being a cow if it has four legs?

The whole picture: *Bayes' Theorem*

The court needs to reach a decision, in the face of uncertainty, concerning the guilt of the defendant. It needs to determine the probability of a hypothesis, for example that the defendant is guilty, or that two speech samples were spoken by the same speaker. How do you estimate probabilities like these? Such estimation is possible, at least where quantitative scientific evidence is concerned, for example evidence based on DNA or speech samples.

It has long been established - since the eighteenth century in fact - that the logically correct way of evaluating the strength of evidence in favour of a hypothesis, and the probability that the hypothesis itself is true, is by using Bayes' Theorem, and this is directly applicable to the evaluation of the strength of evidence in favour of a legal hypothesis like 'these two speech samples were spoken by the same speaker'. Bayes' Theorem makes explicit how one's belief in a hypothesis, like the accused's guilt, can be updated when new evidence is adduced. The so-called odds form of Bayes' Theorem is shown in words at 1.1.

*Odds in favour of hypothesis = Prior Odds * Likelihood Ratio* 1.1

The odds in favour of the hypothesis, which is one way of expressing how probable the hypothesis is, are to the left of the equals sign in 1.1. They represent the odds in favour of a hypothesis, e.g. it's the same speaker, in the light of the scientific evidence. To the right of the equals sign are two terms. The first term is called the *prior odds* and the second is called the *likelihood ratio*. The prior odds are the odds in favour of a hypothesis *before* the evidence is taken into account, and the likelihood ratio, as already explained, is a measure of the strength of the forensic evidence. The expression at 1.1 says, therefore, that the odds in favour of a hypothesis *after* a piece of evidence has been taken into account are quite simply the product of the prior odds and the likelihood ratio for that evidence.

Imagine that it is known that there are five men in a house, including the suspect. The police intercept an incriminating telephone call from the house, and want to know whether the suspect made the call. Before the forensic phonetician compares the incriminating call with known exemplars of the suspect taken from earlier calls, the prior odds in favour of the suspect making the call are actually 1 to 4 *against*, since there are four other people in the house as well as him.

On the basis of the forensic-phonetic comparison, the forensic phonetician estimates a value for the likelihood ratio of 100. This means that the differences between the speech samples are 100 times more likely if they have come from the same speaker than from different speakers. Now the odds in favour of the suspect making the call can be updated by incorporating the forensic-phonetic evidence. The posterior odds in favour of the suspect making the call are: prior odds * likelihood ratio = 1/4 * 100 = 25, and move from 1-to-4 *against* to 25-to-1 *in favour of* him making the call. Odds of 25-to-1 in favour of the hypothesis correspond to a probability of 25/(25+1) = 96% that the suspect made the incriminating call.

Now, at last, we reach the main point. *Under normal circumstances, the forensic speaker identification expert is not privy to the prior odds*. (In fact, in order to preserve maximum objectivity, there are good reasons for the expert to insist on being kept totally in the dark about them - this is the current practice at the ANU, for example.) But Bayes' Theorem makes it clear that you cannot estimate the probability of the hypothesis unless you know the prior odds. (You can easily see how the odds in favour of the hypothesis are dependent on the prior odds by changing the number of men in the house to two in the example above. Now odds in favour of the same speaker being involved move from prior odds of evens (i.e. one-to-one) to 1/1 * 100 = *one hundred times* more likely that it was the suspect, a probability of 99%.

Therefore, if the expert does not know the prior odds, it will not usually be logically possible for them to say anything about the probability of the suspect having made the call or not. What the expert *will* be able to do - and this is their proper role - is *to* *calculate the likelihood ratio, and thus estimate the strength of the forensic phonetic evidence*.

Ears and computers

Undoubtedly the picture of FSI in popular consciousness involves a computer programmed to implement, without human intervention, powerful signal-processing statistical techniques on speech acoustics to recognise a voice. This picture differs in two main ways from reality. Computers are certainly an indispensable part of FSI. But they are used interactively, not automatically, to analyse speech acoustics, and this is done only after careful listening. Also, a substantial amount of FSI evidence is not based on computers at all, but the investigator's auditory analysis. Below are examples of a typical auditory and acoustic analysis.

Auditory comparison On listening carefully to offender and suspect speech samples, it is noted that they both have a certain percentage (higher than 10%) of *okay*s said with just one syllable, thus: *'kay*. Both samples are also observed to have a certain incidence (greater than 20%) of creaky voice, (a way of vibrating your vocal cords that Bob Hawke is often stereotyped with). It can be estimated from results of current research[2] that about 30% of same-speaker pairs agree in having this incidence of *'kay*, and 20% agree in having this amount of creak. It can also be estimated that about 10% of different-speaker pairs agree in *'kay*, and 2.2% of different-speaker pairs agree in creak. Therefore the LRs associated with these two auditory features are about 30% / 10% = 3, and 20% / 2.2% = 9 respectively. Since these features are independent, their combined LR is their product: 3 * 9 = 27. This means that the likelihood ratio for the comparison of the two samples with respect to these two auditory features is 27. You are 27 times more likely to observe this agreement in *'kay* and creak assuming that the samples have come from the same rather than different speakers.

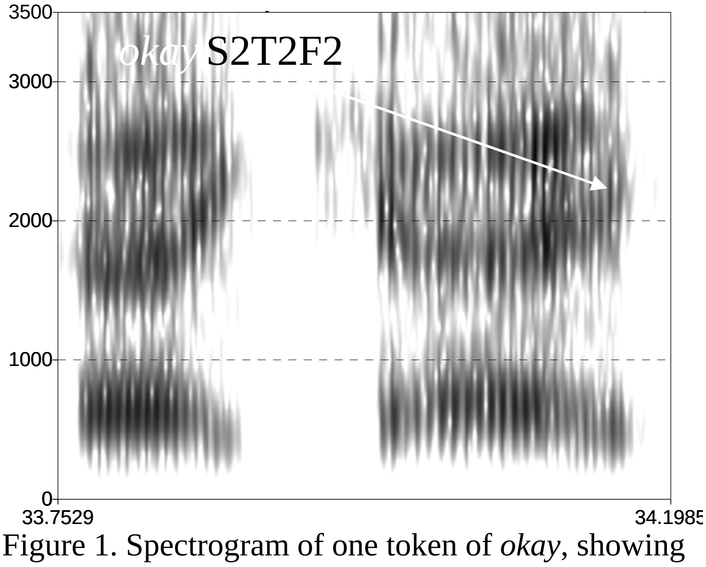

Figure 1. Spectrogram of one token of *okay*, showing the feature *okay*$_{\text{S2T2F2}}$.

Acoustic comparison Figure 1 is a spectrogram of a single *okay* taken from an intercepted phone-call. A spectrogram is a very useful picture of the distribution of acoustic energy in speech. It has two main dimensions: duration (shown horizontally) and frequency (shown vertically). Spectrograms are sometimes called voiceprints, but the term is highly misleading in its implied analogy to fingerprints. Spectrograms, or voiceprints, are nothing whatsoever like fingerprints, no matter what might be claimed in *Law and Order*.

[2] Postgraduate forensic-phonetic research conducted in the ANU *School of Language Studies* by J.Elliott.

A spectrogram shows how much acoustic energy is present at what frequencies, and how this changes over time. The amount of energy is shown by the darkness of the trace - the greater the amount of energy the darker the trace. The spectrogram in figure 1 shows two regions of acoustic energy, one corresponding to the first syllable of *okay*, and the second to the *-ay* part of the second syllable. The low energy white gap between the two relates to the *k* sound.

Very clear in figure 1 are three thick black bands that appear to be running predominantly horizontally. These band-like areas of high acoustic energy are called formants. Formants relate to the frequencies of vibration of the air in the mouth and the throat during speech. Their frequencies change depending on what shape the mouth is, and what shape the mouth is depends on the sound being produced.

It can be seen that the frequencies of the formants do not remain static, but change over time, and this reflects the movements of the vocal organs as they produced the sounds in this particular token of *okay*. It is actually the frequency value of the second formant at the second diphthongal target in *-kay* (called "*okay*$_{\mathrm{S2\ F2\ T2}}$") that is the forensic acoustic feature under consideration. This is shown in the figure.

Fairly sophisticated digital algorithms exist to estimate and track formant frequencies, and according to one algorithm, the value of okay$_{\mathrm{S2\ F2\ T2}}$ for this token is 2210 Hz. If sufficient comparable *okay* tokens are available from both suspect and offender speech samples, they can be measured for this feature, their average values calculated, and the difference between the two samples' average values compared against a reference background of other speakers to determine a LR. (The difference between the mean values of the two samples is a measure of their similarity; the comparison with the other speakers' mean values is a measure of the samples' typicality. Figure 2 helps to make this easier to understand. It represents the reference population as a probability density - this is the slightly bumpy line that goes up and down. This shows what the probability is of observing values of the feature in the population at large. The values for *okay*$_{\mathrm{S2F2T2}}$ run along the horizontal axis from 1550 Hz to 2250 Hz.

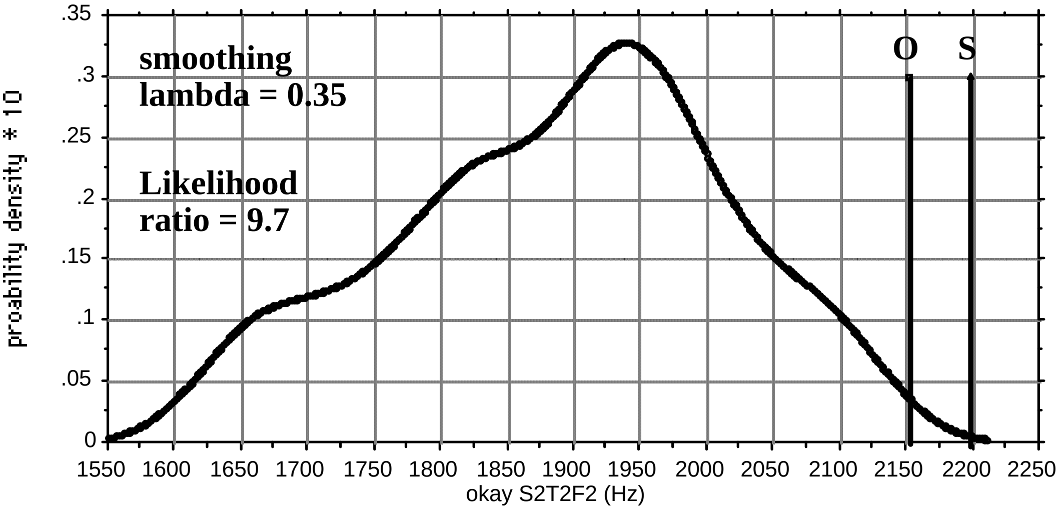

Figure 2. Illustration of estimation of typicality of offender (O) and suspect (S) samples for the forensic-phonetic acoustic feature *okay*$_{\mathrm{S2T2F2}}$.

Imagine that mean values of 2151 Hz are derived for the offender’s *okay*s and 2199 Hz for the suspect’s, and these values are shown in figure 2 by vertical lines marked O and S. It can be seen that they lie at the extreme upper end of the skirt of the probability density curve, and are therefore not very typical of the population. It can be calculated using the rather complicated-looking formula in figure 3 that the probability of getting two values between 2150 Hz and 2200 Hz at random from two different speakers in the population is about 0.008, or less than one percent. The probability of the same two measurements assuming they came from the same speaker is a bit bigger - 0.0776 - so the value for the LR for these data is about 0.0776 / 0.008 = 9.7, which means that the similarity between the offender and suspect samples is 9.7 times greater than their typicality. This means that one would be about ten more likely to observe the difference between the offender and suspect values for *okay*$_{\mathrm{S2T2F2}}$ assuming that they came from the same speaker. This would constitute weak support for the prosecution.

Figure 3. Formula for calculating LR with kernel probability density:

$$LR = \frac{K\exp\left\{-\frac{\left(\bar{x}-\bar{y}\right)^2}{2a^2\sigma^2}\right\}\sum_{i=1}^{k}\exp\left\{-\frac{(m+n)(w-z_i)^2}{2\left[\sigma^2+(m+n)s^2\lambda^2\right]}\right\}}{\sum_{i=1}^{k}\exp\left\{-\frac{m\left(\bar{x}-z_i\right)^2}{2\left(\sigma^2+ms^2\lambda^2\right)}\right\}\sum_{i=1}^{k}\exp\left\{-\frac{n\left(\bar{y}-z_i\right)^2}{2\left(\sigma^2+ns^2\lambda^2\right)}\right\}}$$

where

$$K = \frac{k\sqrt{(m+n)}\sqrt{\left(\sigma^2+ms^2\lambda^2\right)}\sqrt{\left(\sigma^2+ns^2\lambda^2\right)}}{a\sigma\sqrt{(mn)}\sqrt{\left[\sigma^2+(m+n)s^2\lambda^2\right]}}$$

and

$\bar{x}, \bar{y} =$ means of offender, suspect samples

$m, n =$ number of observations in offender, suspect samples

$s^2 =$ variance in reference population (between - speaker variance)

$\sigma^2 =$ within - speaker variance

$\lambda =$ smoothing factor for kernel density estimate

$a = \sqrt{(1/m)+(1/n)}$ $w = \left(m\bar{x}+n\bar{y}\right)/(m+n)$

$k =$ number of kernel functions

$z_i =$ value at which probability density is evaluated for the *i*th kernel

Combining LRs

Normally, forensic speech samples are compared with respect to many independent features, both acoustic and auditory. This is possible because a voice contains an almost limitless number of features. If the acoustic and auditory examples above were all from the same forensic comparison, it would be possible to combine them, again by multiplication: 9.7 * 27 = ca. 260. You would now be about 260 times more likely to observe the acoustic and auditory differences between the two samples assuming they came from same than different speakers. It can be appreciated that LRs from the different speech features, when they are multiplied together, can get very big, or very small, and sometimes as big, or small, as typical LRs from DNA analysis.

A final point of difference from its usual portrayal is that the evaluation of forensic speech samples is not instantaneous, at the click of a mouse, but takes a lot of time - usually in the order of months - to complete.

Summary

A far cry from its portrayal in TV and film, forensic speaker identification is done not only by computer but also by ear. It is not automatic but depends on human interaction with a computer; and it is evaluated within the logically correct Bayesian probability theory which does not try to say what the probability is that the same speaker is involved, but what the probability is of getting the evidence under competing defence and prosecution hypothesis. In the real world of forensic speaker identification there is no such thing as "the computer says it's a match".

Profile

Phil Rose is a forensic-phonetic consultant, associate professor, and head of the phonetics laboratory at the Australian National University in Canberra, where he directs forensic-phonetic research. For almost 30 years, he has researched similarities and differences between individuals in their speech, and has undertaken forensic speaker identification casework in Chinese and Australian English for over a decade. He holds a doctorate from the University of Cambridge in Chinese phonetics, as well as degrees in linguistics and German and Russian. He is author of the major reference work *Forensic Speaker Identification* and the chapter on the *Technical Comparison of Forensic Voice Samples* in the legal reference series *Expert Evidence*. He has also published widely on forensic speaker identification and the phonetics of tone languages, in which he is also an acknowledged expert. He is a member of the *International Association for Forensic Phonetics*, a past member of the *Forensic Standards Committee* of the *Australian Speech Science and Technology Association*, and former Member of Council of the *International Phonetic Association*.