Forensic Semi-Automatic Voice Comparison – an Explanation using Chinese Speech Sounds

Phil Rose

君子於其所不知、蓋闕如也。
“A superior man, in regard to what he does not know, shows a cautious reserve.”
(Confucius, Analects)

“Things are going to slide in all directions. Won’t be nothin’ you can measure anymore.”
(Lenard Cohen, The Future)

philjohn.rose@gmail.com https://philjohnrose.net

ABSTRACT

This chapter explains how forensic semi-automatic voice comparison is done using speech acoustics from Cantonese, Shanghai and North-Western mandarin.

1. INTRODUCTION

Offenders often speak when committing crimes: to threaten, instruct, coerce, extort. Fortunately, their voices are also often recorded at the same time. Here is part of a recording of a financial institution in the process of being defrauded, over the phone, of $150 million:

clerk: “… and we’re going to pay Hong Kong dollars 118,678,543 spot 29 to HSBC erh … Hong Kong?”.

fraudster: “Correct”.

Here is part of a hold-up of a service station:

robber: “Where’s yer fuckin’ till, cunt, I’ll blow yer head off!

attendant: “It’s here!!!”

robber: “I’ll blow yer fuckin’ head off! Push the wrong button and yer dead”.

Recordings of speech like this are often compared with a suspect’s voice in order to help decide whether the suspect said the incriminating speech. This chapter uses examples from Chinese to explain what is involved so that readers can better understand the scientific literature and perhaps do their own testing and experiments.

Your voice carries an enormous amount of information, and not just on what you intended to say. Imprinted on your voice is also information you convey willy-nilly: for example your sex and state of health (Nolan 1983). Some of this information is often of forensic use. Of the forensic trinity of fingerprints, DNA and speech, speech constitutes a special kind of evidence. For speech is the only evidence type of the three that can be directly incriminating. Crime scene DNA matching the suspect can help establish the suspect’s presence and therefore be relevant at a source level, but it is not directly indicative of their guilt. However it is often the case that the speech is an inalienable part of the crime, as in the examples above, or in the recording of a confession. If the offender’s voice can be related to the voice of a suspect, that often constitutes evidence relevant at an offence level (Lucy 2005).

The scientific evaluation of information from speech in forensics is called Forensic Speech Science, and by far the commonest Forensic Speech Science task involves comparison of the speech of an unknown voice with speech
from a known voice. The unknown voice is usually from the offender; the known, that of a suspect/accused/defendant. Since this is the commonest scenario, I will use the terms suspect voice and offender voice below. The interested parties want to know if the suspect said the incriminating speech, so that they can be identified as the offender, or exonerated.

How this is done in the real world bears little resemblance to its portrayal in current TV or cinema. There are several approaches, which differ mainly depending on what kind of information is used and how it is processed. The approach described in this chapter uses information based on the acoustics of speech sounds, much of which is extracted and processed automatically.

Voices are inherently complex, mostly because of the nature of the variation involved. There are differences between different speakers of course – although there are many different speakers who sound alike – but, unlike with DNA and fingerprints, a single speaker’s voice can also vary considerably, both on one occasion and on different occasions. Neither do speakers all vary the same way: some show little variation from occasion to occasion; some a lot; and some vary in their variation, showing a lot of variation on one occasion and little on another. The interaction between this variation and the inherent lack of control over forensic speech samples compounds the difficulty of the task; and finally it is made even more complicated because of the necessary interaction with the Law, which often has problems coming to terms with science (Black, Ayala & Saffran-Brinks 1994) and its language of mathematics and probability.

Given this complexity, with arguments and demonstrations to do with probability and speech acoustics, it is sensible to first give a summary of the chapter’s structure. You might think that, in a chapter involving speech acoustics, the speech acoustics would take centre stage. But you would be wrong. By far the most important thing is to understand what is logically involved in arriving at an answer to the typical question: did the suspect in fact say the incriminating speech? The key term here is Bayes’ Theorem, and the first part of the chapter is taken over with what I hope is a relatively painless explanation of its central role in forensic inference. The speech acoustics then follow in the second part, where I use examples from Chinese – both Cantonese and Mandarin – to demonstrate how one implements the ideas in the first part.

2. EVIDENCE, AND HOW TO EVALUATE IT LOGICALLY

I want you to imagine the following scenario … Actually, you do not have to imagine it. The past two years with covid have transformed this originally fictitious example into frightening reality for all of us.

There is a very nasty life-threatening disease going around. It is often at first asymptomatic and so, although you feel pretty well at the moment, you think should have a test to check, just to make sure that you are not infected. So you have the test, and, to your horror, you test positive! You see your doctor, who does not decrease your worries by saying that he has read that the test is pretty accurate – 96% in fact.

If you are a normal human, even a fairly optimistic one, you will now probably believe in the face of this information – especially that “96% accuracy” the doctor mentioned – that you have caught the disease, and are now in for some long and nasty treatment.

But have you really got the disease? You know that tests are fallible and can give false positives – so you know it is possible the test can say you are infected when you are not. So, is your positive result a true result or is it a false positive? You intuit that it will have something to do with that figure of “96% accurate”, but how do you rationally decide whether you are likely to have the disease, given that you don’t know for sure? Below is how you work out the probability that you are infected, given your positive test result.

There are two main things to take on board. First, let’s be clear again about what it is you want to know. This is: how likely are you to be infected, given that you have tested positive for the disease? Secondly, the information you currently have – the positive test result and its 96% effectiveness – are not enough to answer this question. You need to know more in order to estimate whether or not you are infected. You need to know two things, in fact: you need to know how rare the disease is; and you need to know more about the test.

First, the disease: let’s assume it is very rare – say about one person in 1000 in the country has it. To make things easier to calculate, let us assume a very small population for the country you live in of 100,000, so that of these
100,000 people can be expected to be infected, and \((100,000 - 100 =) 99,900\) can be expected not to have the disease. This is shown in table 2.1 below.

### Table 2.1. Incidence of infected people in the population.

<table>
<thead>
<tr>
<th></th>
<th>no of people</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>100,000</td>
</tr>
<tr>
<td>infected</td>
<td>100</td>
</tr>
<tr>
<td>not infected</td>
<td>99,900</td>
</tr>
</tbody>
</table>

Now let’s look at the test. As far as the test is concerned, we need to know more than its “96% accuracy” to know how good it actually is. By reporting that the test is 96% accurate what the doctor presumably means is that of 100 people who actually have the disease, 96 will test positive (and the disease will go undetected in the remaining four unfortunates). These 96 positives will be true positives, because all 100 people tested had the disease.

We know that no test is perfect, but in addition to knowing how many infected people it correctly identifies, we also need to know how many uninfected people it incorrectly diagnoses as having the disease when they don’t. Let’s assume the test is pretty good, with a 2% false positive rate – out of 100 people who do not have the disease, only two test positive. And these positive results will be, of course, false positives.

These two figures characterise its strength. Imagine a test which got the presence of the disease correct 50% of the time (out of 100 infected people 50 were correctly diagnosed as positive), but also incorrectly diagnosed it 50% of the time (out of 100 healthy people 50 were diagnosed as infected). You will agree that is a pretty useless test, because with that test you are just as likely to get a positive result if you are healthy as if you are infected. Its positive test does not tell you anything. Compare that to the test we are considering: you are quite a lot more likely to get a positive result if you are infected than if you are healthy. How much more likely? \(96/2 = 48\) times more likely, in fact. This value of 48 represents the strength of the test as far as positive results are concerned. With this test a positive result is 48 times more likely if you have the disease than if you do not. This is summarised in table 2.2.

### Table 2.2. Strength of test.

<table>
<thead>
<tr>
<th>people taking test</th>
<th>number who test positive</th>
<th>strength of test</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 infected people</td>
<td>96</td>
<td>96/2 = 48</td>
</tr>
<tr>
<td>100 healthy people</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Now let us combine the figures characterising the test’s strength with the incidence of the disease in the population. Given the disease’s incidence, and the fact that the test correctly diagnoses 96 out of 100 truly infected people, we would expect 96 out of the 100 people in the population who have the disease to be correctly diagnosed by the test as positive. But there are also 99,900 uninfected people to be considered. Of these the test will incorrectly say \(2%\) of \(99,900 =\) 1998 have the disease when they don’t. This is shown in table 2.3.

### Table 2.3. Positive tests in the population.

<table>
<thead>
<tr>
<th>number of people</th>
<th>number testing positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>infected</td>
<td>100</td>
</tr>
<tr>
<td>not infected</td>
<td>99,900</td>
</tr>
<tr>
<td>total</td>
<td>100,000</td>
</tr>
</tbody>
</table>

From table 2.3 we can see that the number of positive results given by the test is \((96 + 1998 =)\) 2094. Of these 2094 positives, 96 are true: the person is really infected. But 1998 positive results are not true: the person is not infected.

So which is your positive result? Is it a true positive or not? You don’t know for sure, but think of it like this: you have a bag containing five balls; one is red, 4 are black. Close your eyes, put your hand in and pull out a ball. How likely is it to be red? There is a \((1/5 =)\) one in five chance of it being red. Now imagine you have a lot more balls in the bag: 2094 in fact, to represent the number of positive results. Of these balls 96 are red, to represent the true positive results, and 1998 are black to represent the false positives. Close your eyes, pull out a ball: what
It is useful to dwell a little on what we mean here by probability. Probability is a wonderful thing, in that it enables us to be definite about things that are inherently uncertain. It is also rather a mysterious thing, in that it does not have a single definition. There are actually several ways that probability can be construed, but in this context, it is best to understand it as an expression of degree of rational belief. It takes a value between 0 (or 0%), which stands for "certain that it is not the case", and 1 (or 100%), which stands for "certain that it is the case". So the % chance of infection just obtained can be taken to mean I should rationally believe my chance of being infected is very low. Had the probability been, say, 80%, I would have to assume it was pretty likely I was infected. And a 50% probability would mean that I cannot say one way or the other. So, as far as your positive test result is concerned, the % probability means there is a rather low chance that you are actually infected.

The normal response to all this is: "However can this be so, given that you tested positive and the test was 96% accurate?!" The answer is that, as we have seen, the correct evaluation of the positive test result depends not only on the test’s 96% accuracy (which you will remember represents its true positive rate), BUT ALSO on the test’s false positive rate, AND how rare the disease was in the first place (recall we set this as a very low 1 in 1000).

To make this clearer – and to see how well you have understood the example! – I invite you now to redo the calculation, but this time assume that the disease is much, much, much more common. Instead of 1 person in 1000, assume now that 1 person in 25 in the population has the disease. The strength of the test remains the same: 96% true positives, 2% false positives. Again you test positive – now what is the probability that you are infected? It might help to work through the following separate questions:

1. If 1 person in 25 in the population is infected, how many in our hypothetical population of 100,000 are infected?
2. How many healthy people does that leave in our hypothetical population?
3. If, out of 100 infected people who are tested, 96 return a positive test, how many positive tests are there from those in the population who are truly infected?
4. If, out of 100 healthy people who are tested, 2 return a positive test, how many positive tests are there from those in the population who are not infected?
5. How many positive tests are there in all?
6. What proportion of those tests are true results?
   [That is the probability that your test is a true result, and not a false positive.]
7. Should you be worried?

You will find the answers in the footnote.

Things do not end there, of course. You still have to make the decision as to whether this 5% is a small enough chance to ignore the result. But we will leave the problem at this point, in order to concentrate on the lessons from the demonstration.

Bayes’ Theorem

Now let us introduce a little bit of terminology and formalism. You wanted to estimate how probable it is that you have the disease, given the positive test result. The thing you want to find out – whether you have the disease or not – is the Hypothesis. Call it H for hypothesis that you are infected. The information that is relevant to this hypothesis – the positive test result – is the Evidence in support of the hypothesis. Call it (E). The probability of the Hypothesis, given the Evidence is then represented as p(H | E), where p stands for probability, and the vertical stroke " vertical bar" stands for "given", or “conditional upon”. This is a so-called conditional probability: the probability that something is true, given some other information.

1 Out of our hypothetical population of 100,000, 4000 people now have the disease, and 96,000 are healthy. Of the 4000 infected people, 3840 test positive; and of the 96,000 healthy people, 1920 also test positive. That makes 3840 + 1920 = 5760 positive tests, of which 3840 are true positives. So the probability that you have the disease, given that you test positive, is 3840/5760, or 67%. Yes, you should be worried: 67% is considerably more likely than not that you have the disease.
The rather long scenario above showed an empirical solution to estimating the probability of a hypothesis, given the evidence \( p(H \mid E) \). There is, however, an analytical solution to the problem. And, no, it is not some recently derived mathematical formulation or hi-tech solution (although its successful implementation is undoubtedly due to increase in computer processing power). Bayes’ Theorem has been around for about three hundred years. It was worked out by the reverend Bayes, a protestant minister who lived in a quiet little town in the south east of England in the 18th century. Bayes’ Theorem gives you the formula for estimating a conditional probability, and belongs to those wonderful mathematical objects which are at once both profound and simple (Ellenberg 2015: 14, Silver 2013: 243). It says that if you want to know the probability of a hypothesis, given the evidence adduced in its support, you need to know two things: the strength of the evidence; and the probability of the hypothesis before the evidence is adduced.

**Likelihood Ratio**

Let us look first at the strength of the evidence. In the thought experiment above, the evidence was the positive test, and its strength was the ratio of the true positive rate to the false positive rate. Both these terms – the rate of true positives and the rate of false positives – are actually conditional probabilities. Think about it - the test’s true positive rate of 96% (out of 100 infected people tested, 96 correctly show a positive test) is actually the probability of getting a positive test, given that the person being tested is actually infected: \( p(E \mid H) \). The test’s false positive rate of 2% (out of 100 healthy people tested, 2 incorrectly return a positive test) is the probability of getting a positive test, given that the person is not infected: \( p(E \mid \sim H) \). (“\( \sim \)” stands for the Hypothesis is not true). Their ratio \( \frac{p(E \mid H)}{p(E \mid \sim H)} = \frac{96}{2} = 48 \) quantifies the strength of the test. This ratio is called the **Likelihood Ratio**, and it is arguably the most important number in forensic science. Because it is so important it is shown again in table 2.4. Note once again that the likelihood ratio is the ratio of the probabilities of the *same* evidence under competing hypotheses. It has nothing to do with the probabilities of the hypothesis.

<table>
<thead>
<tr>
<th>Table 2.4. The Likelihood Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In words</strong></td>
</tr>
<tr>
<td>Likelihood Ratio =</td>
</tr>
<tr>
<td>probability of evidence E, given</td>
</tr>
<tr>
<td>hypothesis H is true</td>
</tr>
<tr>
<td>probability of evidence E, given</td>
</tr>
<tr>
<td>hypothesis H is not true</td>
</tr>
</tbody>
</table>

The likelihood ratio is also probably the most unfortunately named number in forensic science. Take a conditional probability, for example \( x = p(H \mid E) \). (Recall in the above example we obtained \( x = p(H \mid E) = 5\% \)). Now, mathematically, given \( x = p(H \mid E) \), the term \( y = p(E \mid H) \), with the \( H \) and the \( E \) reversed, is called the *likelihood of* \( p(H \mid E) \). \( p(E \mid \sim H) \) is also a likelihood, and so their ratio is called, sensibly enough, a likelihood ratio.

The trouble is that the word *likelihood* is much more common in its non-mathematical sense of *probability* (e.g. what is the likelihood of rain today?), and since humans worry much more about the probability of a hypothesis, or event, it is very common to find the likelihood ratio understood as representing the probability of a hypothesis rather than having to do with the probability of evidence. But the probability of a hypothesis given the evidence \( [p(H \mid E)] \) is not the same as the probability of the evidence given the hypothesis \( [p(E \mid H)] \). Suppose I tell you that, although you can’t see it, behind that door there is a cow. What is the probability that it has four legs? Barr ing genetic mutations, road accidents and such like, most cows have four legs. So the probability that it has four legs, given that it is a cow [ \( p(\text{four legs} \mid \text{cow}) \) ], is very high. Now swap the terms around: \( p(\text{cow} \mid \text{four legs}) \) – what is the probability that it is a cow, given that it has four legs? A lot of things have four legs that are not cows, so this probability will be much lower. This is, unfortunately, a very common error in legal cases, where it is called the prosecutor’s fallacy or transposing the conditional. For example, assume that the probability of the incriminating speech evidence, given that the suspect said it, is 90% [ \( p(E \mid H) \) ] = 90%. It is common to find this interpreted to imply that it was 90% likely the suspect said the incriminating speech \( [p(H \mid E) = 90\%] \). These two statements are not interchangeable: the probability that it is a cow, given that it has four legs – \( p(\text{cow} \mid \text{quadruped}) \) – is not the same as \( p(\text{quadruped} \mid \text{cow}) \).

**Prior probability**

Now the probability of the hypothesis before the evidence is adduced. This is called the *prior probability* and is represented as \( p(H) \). In our thought experiment, this refers to the probability that you are infected before you are tested, and its best estimate was the probability of infection in the community at large. We set that in the first example at 1 in 1000, so the probability that you are infected before you know the result of your test \( p(H) = \frac{1}{1000} = 0.001 \). In the second example, with the disease being much more common, we made one person in 25 have it, and so we had to assume that the probability you are infected before we know the test result is \( \frac{1}{25} = 0.04 \).
Odds Form of Bayes’ Theorem
Now, finally, comes the formula. For important reasons which have to do with correspondence to the legal system, and which will be explained later, we use a formula which implements what is known as the Odds Form of Bayes’ Theorem. Here it is: posterior odds = prior odds * likelihood ratio. It is shown in table 2.5.

Table 2.5. Formula for Odds Form of Bayes’ Theorem

\[
\frac{p(H|E)}{p(\neg H|E)} = \frac{p(H)}{p(\neg H)} \times \frac{p(E|H)}{p(E|\neg H)}
\]

\(p = \text{probability}, \ | = \text{given}\)
\(E_s = \text{speech evidence}\)
\(H_{SS}/H_{DS} = \text{same/different speaker hypothesis}\)

Instead of using probabilities for the prior and posterior, odds are used. So instead of expressing the prior as a probability, we express it as odds. Odds simply compare the probability of something happening to the probability of it not happening. Odds are easily calculated from probability and vice versa. A probability of 1 in 5, for example, corresponds to odds of 1 to 4, and odds of 10 to 1 correspond to a probability of (10 / 10+1 = ) 0.91 or 91%. The prior odds in this case are therefore the ratio of the probability \(p(H)\) that the hypothesis is true – you are infected – to the probability \(p(\neg H)\) the hypothesis is not true – before you are tested! We know that the incidence in the population is 1 in 1000, so the prior odds are 1 to 999 against you being infected.

The likelihood ratio is the ratio of two likelihoods: the probability of getting the evidence – the positive test – assuming you are infected \(p(E|H)\) and the probability of getting the positive test assuming you are healthy \(p(E|\neg H)\). We were told that out of 100 infected people, 96 test positive, so \(p(E|H) = 0.96\). And we were also told that out of 100 healthy people, 2 test positive, so \(p(E|\neg H) = 0.02\). So the likelihood ratio for the positive test is \(0.96/0.02 = 48\).

Table 2.6 plugs the values for the prior odds and the likelihood ratio in the formula for the odds form of Bayes’ Theorem. It shows that the posterior odds in favour of you being infected, given the positive test result are 0.96 to 19.98. This corresponds to a posterior probability of [0.96 / (19.98 + 0.96) = ] 0.045 or about 5%. And this, using Bayes’ Theorem, is the same answer that we obtained above by working through the numbers empirically, but laboriously.

Table 2.6. Odds form of Bayes’ Theorem with low priors for disease incidence.

\[
\begin{array}{ccc}
0.96 & 1 & 0.96 \\
19.98 & 999 & 0.02 \\
\end{array}
\]

\(Posterior = \frac{Prior \times Likelihood}{Odds} \)

It may be useful to plug into the formula the values for the second example, where we made the disease much more common, with an incidence of 1 in 25 in the population. These are shown in table 2.7.

Table 2.7. Odds form of Bayes’ Theorem with higher priors for disease incidence.

\[
\begin{array}{ccc}
0.96 & 1 & 0.96 \\
0.5 & 25 & 0.02 \\
\end{array}
\]

\(Posterior = \frac{Prior \times Likelihood}{Odds} \)

This time, with the same strength test but a much more prevalent disease, the posterior probability is [0.96 / (0.5 + 0.96)] = 0.657 … or about 66%.

Further Reading
I hope I have now covered enough ground to help establish concepts like prior odds and likelihood ratio. If you wish to read more about their central role in forensic inference, you could not do better than Robertson and Vignaux’ (1995) Interpreting Evidence. A far more technical account, but still with examples very similar to the
**Visualising the Likelihood Ratio**

We have seen above that the strength of evidence in support of a hypothesis is quantified by the likelihood ratio: the ratio of the probability of the evidence under the assumption the hypothesis is true, and under the assumption it is not. To help reinforce the crucial idea of the probability of evidence, here is another made-up example, this time with speech data, from which you are to estimate a likelihood ratio. Imagine the offender speech sample contains a well-defined speech defect, a lisp perhaps, which is also found, permanently, in the suspect’s voice. The evidence is that both offender and suspect samples contain a lisp. You know from research that one in 100 people have a lisp like this. What is the likelihood ratio/strength of the lisp evidence? Again it might be helpful to ask the following separate questions:

- What is the probability of observing a lisp in the offender sample supposing the latter has come from the suspect? \[ p(\text{lisp} \mid \text{suspect is offender}) \].
- What is the probability of observing the offender lisp assuming it has come from someone other than the suspect? \[ p(\text{lisp} \mid \text{suspect is someone other than offender}) \].
- What is the ratio of these probabilities?
- How would you express the likelihood ratio in words?

Let us now move from the fictitious to the real, in a demonstration of likelihood ratios and Bayes’ Theorem with some Chinese speech acoustics. The aim will be to see how well we can tell the sex of a speaker from their speech acoustics. The acoustic data is taken from a small experiment carried out at China’s Southwestern University of Political Science and Law under the supervision of Dr Zhang Cuiling, herself a well-known practicing forensic speech scientist. A class of 33 of Dr Zhang’s students was asked to record themselves saying various vowels, like ee ah and uu. They then had to measure their vowels with respect to several acoustic features. One of these features is called fundamental frequency, or F0. F0 reflects the rate of vibration of the vocal folds and is measured in units of Hz, or repeats per second. So an F0 value of, say, 100 Hz means that the speaker’s vocal folds were vibrating one hundred times per second at the time the F0 was measured.

To demonstrate this, Figure 2.1 shows a picture of my vocal folds when I was saying an ee vowel (the proper phonetic symbol for this is [i], but it will not hurt just to write ee). The picture was taken using a laryngoscope inserted through the nose and down into the throat. The view is from above and slightly to the left. The vocal folds are the two small white strips of tissue in the middle of the picture stretching from front to back, and, although you can’t see it, they are actually vibrating. The rate at which they are vibrating can be measured from the speech acoustics of the ee vowel, which are shown in figure 2.2. Fig. 2.2 shows the wave-form from the middle of the ee vowel. The x axis is duration, quantified in seconds. It can be seen that the wave-form is complex, but four repeats are evident. Each of these four repeats corresponds to a single vibration of the vocal folds, so if we take any point in the wave-form and measure how long it takes to repeat we can get the F0. One such duration is shown in red. It can be seen that at over this duration the wave-form took 0.0068 seconds to repeat. If the wave-form took 0.0068 seconds to repeat, then it is repeating at a rate of about \( 1/0.0068 = 146 \) times per second, or a fundamental frequency of about 146 Hz.

---

2 If the lisp is always present in the suspect’s speech, the probability of observing a lisp assuming it has come from the suspect is 100%.
3 If the incidence of the lisp in the general population is 1 in 100, the probability of observing a lisp assuming it has come from someone else is \( (1/100 = ) 1\% \).
4 \( 100\% / 1\% = 100 \). That is the likelihood ratio for the lisp evidence.
5 “You would be 100 times more likely to get the lisp if it came from the suspect than if it came from someone else.”
6 The experiment on sex and F0 was conducted during my stay at the School of Criminal Investigation. This was made possible by grants from the Chongqing Municipality Attracting Overseas Expertise Scheme 巴渝海外引智计划 for which I am most grateful.
Figure 2.1 Laryngoscopic view of author's larynx showing the vocal folds during an ee vowel. I am facing towards the bottom.

Figure 2.2. Four periods from the wave-form in the middle of the ee vowel in figure 2.1. Red text shows the duration of one of the pulses. The fundamental frequency at this point is about 147 Hz.

Now, one of the things that makes forensic voice comparison possible is the fact that the acoustics radiated from someone when they speak are determined by their anatomy, so that speakers with different dimensioned vocal tracts will output different acoustics for the same sound. (Of course, what you are wearing on your head – balaclava, Covid mask … - might also play a role, but the primary determinant is still your vocal tract.) Now, females tend to have shorter, less massive vocal folds than males, and because of this their vocal folds will vibrate at a higher frequency, other things being equal. Your brain hears F0 as pitch – a higher F0 from a speaker will be heard as a higher pitch than a lower F0. So F0 is the main acoustic feature corresponding to tone in Chinese. Figure 2.3 shows mean F0 values for the three long tones of Shanghai spoken by two males and two females (the data are from Rose 1993). Shanghai has three long tones, with high falling, mid dipping and low rising pitch (Zhu 2004). It can be seen that, although there is a little overlap, the two male speakers have overall lower tonal F0 than the females.
Fig. 2.3. Mean F0 values for the three long tones of Shanghai showing correlation between tonal F0 and sex. Left two panels = male speakers, right two panels = female speakers. Red = high falling tone, black = mid dipping tone, brown = low rising tone. X-axis = duration (csec.), y-axis = F0 (Hz).

We can now return to the experiment to estimate sex from F0, where the class of 33 Chinese students were asked to extract the F0 from their recorded ee vowels. The left panel of fig. 2.4 shows the distribution of the F0 measurements of the class separated according to sex. It can be seen that the female values are located mostly between 200 and 300 Hz, with the commonest values around about 250 Hz. The male values are lower: most lie between about 90 and 170 Hz, with the commonest values around about 130 Hz. These values correspond roughly to the Shanghai values in fig. 2.3. Because, as can be seen, the mean values of the distributions are quite close to the modal values it is sensible to assume that these values are distributed normally, and the data modelled by a normal, or Gaussian distribution, are shown in the right panel of fig. 2.4. Assuming normality makes it much easier to do calculations.

It is easy to see, given such distributions, that you would be vastly more likely to get an F0 value of, say, 200 Hz if it came from a female than from a male, and a value of 180 Hz would be more likely if the speaker was male. Note that such expressions, e.g. $p(200 \text{ Hz} | \text{female})$, are likelihoods. So distributions like this make it easy to visualise likelihood ratios.

Now suppose we are given an F0 value from the class of 200 Hz, and are asked to say whether it is from a female student. The hypothesis $H_f$ is therefore that the 200 Hz F0 value came from a female student, and the evidence $E_{F0}$ is the F0 value of 200 Hz. We have seen above that, from the odds form of Bayes’ theorem, the odds in favour of a hypothesis, given the evidence, is the prior odds * likelihood ratio. What is the likelihood ratio in this case? Recall it is the probability of the evidence assuming the hypothesis $H_f$ is true – that the F0 value came from a female – relative to the probability of the evidence assuming that the hypothesis $H_f$ is not true (in this case, of course, that means that it came from a male). A visualisation of this likelihood ratio – for the evidence $E_{F0} = 200 \text{ Hz}$ – is given in fig. 2.5. The vertical line at 200 Hz represents the F0 to be evaluated. It can be seen that it crosses the female distribution where the probability density is 0.003, so the probability density of getting the value of 200 Hz assuming it has come from a female, is 0.003; but it crosses the male distribution where the probability density is 0.00024, so the probability
density of getting the value of 200 Hz assuming it has come from a male is 0.00024. The likelihood ratio is the ratio of these values: $0.003 / 0.00024 = 12.5$. In words, you would be twelve and a half times more likely to get the value of 200 Hz if it had come from a female than from a male.

![Figure 2.5. Visualisation of likelihood ratio with evidence = 200 Hz. Red = female F0 distribution. X-axis = F0, y-axis = probability density.](image)

To estimate the probability that our value of 200 Hz is from a female, we still have to incorporate the prior odds. The class contained 20 female students and 13 males. So the odds in favour of any F0 value coming from a female would be 20 to 13, or in favour of a female. Now it is easy to estimate the posterior odds in favour of the value of 200 Hz coming from a female: it is the prior odds * likelihood ratio: $20/13 \times 12.5/1 = \text{ca.}19.2$. The posterior odds, given the value of 200 Hz, are about 19 to 1 in favour of it being a female. This corresponds to a posterior probability of $(19 / 19+1) = 95\%$. Therefore your rational belief, given a value of 200 Hz, would strongly favour that it came from a female student.

A final point is that likelihood ratios in forensic voice comparison are usually expressed as base 10 logarithms. So a likelihood ratio of 10 would become a log_{10}LR of 1, a likelihood ratio of 100 would be log_{10}LR 2, a likelihood ratio of 0.1 would have a log_{10}LR value of $-1$ and a likelihood ratio of 0.01 would be log_{10}LR $-2$. This has several advantages. It becomes easier to appreciate the equivalence of likelihood ratios in support of opposite hypotheses. It is not quite so obvious, for example, that a LR of 0.1 is the same strength as a LR of 10, only in support of the alternative hypothesis; whereas it is perhaps easier to understand this with the equivalent log_{10} LRs of -1 and 1. The use of logarithms also has the effect of neatly representing useless strength of evidence – where you are just as likely to get it under both hypotheses – as log_{10} 0. Finally, use of logarithms also suggests the additive property of evidence, which we invoke when we talk about its weight (Good 1991: 89). Turing, for example, suggested a base of 10, calling the unit a *ban*, with a *deciban* being the smallest change in weight of evidence we can conceptually process (Good 1991: 90; Jaynes 2003: 92).

### 3. EVALUATING SPEECH EVIDENCE IN FORENSIC VOICE COMPARISON

The nature of voices and the essentially uncontrolled circumstances in which forensic voice samples are obtained means that the question *did the suspect in fact say the incriminating speech* cannot be answered definitively. There is uncertainty involved. (A trial is, after all, about making decisions in the face of uncertainty). This is no different from the *are you infected?* and *what sex is this F0?* questions we have evaluated above: their answer is a probability.
So it is usually asked: how probable is it that the suspect said the incriminating speech? This is a very reasonable way of putting it, since philosophers and statisticians agree that the best way of quantifying uncertainty is by using probability (Lindley 1991). Implied, of course, will also be the crucial role of evidence. That is, the question is really: how probable is it, given the voice evidence, that the questioned and known samples have been said by the same person? This is what the interested parties want to know. We can represent it as $p(H_s \mid E_p)$ – the probability that suspect and offender are the same speaker, given the speech evidence. You will recognise this as a conditional probability. The answer, invoking the odds form of Bayes’ Theorem, is: the prior odds that the suspect said the incriminating speech times the likelihood ratio for the speech evidence. This is shown in table 3.1.

Table 3.1. Odds form of Bayes’ Theorem tailored for forensic voice comparison.

<table>
<thead>
<tr>
<th>$p (H_{SS} \mid E_{SS})$</th>
<th>$p (H_{DS} \mid E_{SP})$</th>
<th>$p (H_{SS})$</th>
<th>$p (H_{DS})$</th>
<th>$p (E_{SS} \mid H_{SS})$</th>
<th>$p (E_{SS} \mid H_{DS})$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Posterior</strong></td>
<td><strong>Prior</strong></td>
<td><strong>Odds</strong></td>
<td><strong>Likelihood Ratio</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p (H_{DS} \mid E_{SP})$</td>
<td>$p (H_{SS})$</td>
<td>$p (H_{DS})$</td>
<td>$p (E_{Sp} \mid H_{DS})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where $p = probability, \mid = given, E_{SP} = speech evidence, H_{SS}/H_{DS} = same/-different-speaker hypothesis

As we have seen, Bayes’ Theorem is of paramount importance when one wants to know the probability of a hypothesis given the evidence, and thus it is crucial in forensic identification. It has in fact been called “...the fundamental formula of forensic science interpretation” (Evett 1998: 200).

Making this a probabilistic question means that one is actually considering two competing and mutually exclusive claims: the suspect said the incriminating speech; and someone other than the suspect said the incriminating speech. In forensic voice comparison the first is often called the prosecution hypothesis, and the second, alternative, claim is often called the defence hypothesis.

Now comes an extremely important point. Note that in the odds form of Bayes’ Theorem the posterior odds are neatly factored into two separate terms: the prior odds and the likelihood ratio. The forensic expert, however, is usually not privy to information that informs the prior odds – they don’t know all the other relevant information in the case – so they cannot logically estimate a posterior. Therefore they cannot say how likely it is that the same speaker was involved, given the evidence. This is a point unfortunately still not well enough understood outside forensic statistics (Rose & Morrison 2009). It has been the author’s experience, for example, to often be asked by counsel to say whether he thought the suspect said the incriminating speech, when he has tried very hard to explain that he cannot logically do that because he does not know the prior odds!

The force of that word logically needs emphasising here. Bayes’ Theorem is a theorem, like Pythagoras’ Theorem. As long as you accept its premises – which it would be irrational not to accept, because they are based solidly on probability theory – it does not have the property of being wrong. It is inescapable: if you want to know the probability that the suspect said the incriminating speech, given the evidence, you must know both the strength of the evidence AND the probability that they said the incriminating speech BEFORE the evidence is adduced. If you do not know this, you can ONLY estimate the strength of the evidence.

It is also an important point in some jurisdictions that the question of guilt – whether the suspect said the incriminating speech – is a question the trier-of-fact must decide, not the expert. And from this point of view, too, Bayes’ Theorem is legally correct, since it allows the expert to concentrate exclusively on the strength of the evidence. The odds form of Bayes’ Theorem allows the expert mathematically to confine themselves to estimating the strength of evidence in support of the hypothesis as quantified by the likelihood ratio (Aitken & Taroni 2004). Since no identification or recognition is involved with the estimation of a likelihood ratio, the term forensic voice comparison is now being preferred to forensic voice recognition or identification, which imply a posterior.

The sections above have explained the logically and legally correct way of comparing forensic speech samples: it is a question of estimating the strength of the speech evidence with a likelihood ratio, and this is done by estimating the probability of the evidence assuming the suspect said the incriminating speech, and comparing that with the probability of the evidence assuming they did not.
Now it is time to examine how speech acoustics can be used to estimate likelihood ratios and this will be done in the following section, with data from Chinese.

**Evaluating Evidence from Speech Sounds**

In this section we look at how voice recordings can be compared forensically using the acoustics of speech sounds. We start with the Cantonese vowel *ee* in words like *sì* 'market', *yih* 'two', *yìhga* '而家' 'now', *sì* 'time' etc. Details of the experiment described below may be found in Wang & Rose (2012).

Fig. 3.1 shows what a Cantonese *ee* vowel looks like acoustically (the word was *yih* 'two', spoken by a young male). This three-dimensional object shows the amount of acoustic energy in the vowel at different frequencies during its production. You can see that the vowel lasted about 35 hundredths of a second (from csec. 12 to csec. 46), during which time the acoustics are dominated by several bands of high amplitude energy running through the vowel. For example, the band with the highest amount of energy is located at a frequency of about 250 Hz. These ridges are called formants, and the frequency with the highest amount of energy is called a formant centre-frequency, or just formant frequency.

![Figure 3.1. Three-dimensional plot of the acoustics of a Cantonese ee vowel.](image)

When visualising speech acoustics, it is usual, and much less cumbersome, to compress the vertical, amplitude, dimension into a two-dimensional representation called a spectrogram. A spectrogram of the vowel in fig. 3.1 is shown in fig. 3.2, where you can see that the duration now runs along the horizontal axis, with the frequency being shown along the vertical axis. The amount of energy present is shown by the darkness of the trace, and the formants can be easily seen as the dark bands of energy running through the vowel. The formant centre frequencies have been automatically extracted and superimposed with a dotted red line.

Formants are numbered from the lowest upwards. It is easy to see that the centre frequency of the first formant (F1) is a little above 250 Hz. Now see if you can say *from the spectrogram* approximately what the F2 frequency is — see the footnote for the answer. The ensemble of formant frequencies is often called an *F-pattern*.

---

7 Many thanks to Dr Franz Clermont for his Matlab code to produce this waterfall plot.
8 F2 is at about 2000 Hz, or 2 kilohertz (kHz).
Formant frequencies indicate the frequency at which air in the speaker’s vocal tract was vibrating with maximum amplitude when the speaker said the sound. The lowest two or three formant frequencies are what you hear as vowel quality: what makes you hear a vowel as ee or ah or ooh.

Formant frequencies are a function of two things. Firstly, they relate to the speech sound the speaker was making. In this case it was an ee vowel. If it had been an ah vowel the F-pattern would have been different. Secondly, formant frequencies relate to the overall dimensions of a speaker’s vocal tract, especially its length. This is best illustrated with a picture of an actual vocal tract saying an ee vowel. The left panel of fig. 3.3. shows an x-ray of my vocal tract while saying ee. You can see that my tongue is high in my mouth, and towards the front (that is why phonetically ee is called a high front vowel).

The high front position of the tongue in ee means that the shape of the vocal tract consists mostly of two tube-like portions: a narrow oral cavity is formed by the approximation of the top of the tongue to the hard palate; and a wide pharyngeal cavity is constituted by the pharynx behind the tongue. The right panel of fig. 3.3, in a model of the vocal tract in a Russian ee vowel (Barlow et al. 2001) shows these two tubes more clearly, as well as the flaring of the vocal tract at the lips, and the short tube of the epilarynx at the bottom.

Figure 3.2. Spectrogram of the ee vowel in fig. 3.1, with extracted formant centre-frequencies superimposed in red.  
X-axis = duration (csec.), y-axis = frequency (Hz).
Figure 3.3. Supralaryngeal vocal tract configuration for an ee vowel. Left = x-ray, Right = 3D model of vocal tract.

Now, formant frequencies are a function of the properties of the vocal tract that produced them. These properties are mostly dimensional, and mostly to do with the length of various cavities acting as acoustic resonators, although the compliance of the vocal tract walls is also a factor. So for example in an ee vowel the F1 frequency can be shown to be a function of both the volume of the pharyngeal cavity and the length and cross-sectional area of the oral constriction; and its F2 frequency is a function of the length of the pharyngeal cavity (Stevens 1998: 262, Fant 1960: 121). For example, assuming values of 50 cms$^3$ for the pharyngeal volume, with 4 cms and 0.4 cms$^2$ respectively for the length and cross-sectional area of the oral cavity, gives an F1 value of 250 Hz. A pharyngeal cavity in ee about 9 cms long will result in an F2 of about 2000 Hz.

Not all formants reflect resonances of the supralaryngeal cavities shown in fig. 3.3. If you look again at the 3D representation in fig. 3.1 you will see a fairly prominent ridge of energy at about 1000 Hz. This energy is also visible – stronger at first, then getting weaker – in the spectrogram in fig. 3.2. This is a subglottal formant, reflecting the resonance in the trachea below the vocal folds (Stevens 1998: 300). It is not always present, although it is more commonly found in female voices.

As we have already seen with the vocal folds, the greater the length, the lower the frequency. This is a further example of the correlation between speech production and radiated acoustics which makes forensic voice comparison a possibility.

**Between- and within-speaker variation**

I wrote *possibility* in the previous paragraph for two reasons. Although speakers with differently dimensioned vocal tracts will output different acoustics, it is not the case that every speaker has differently dimensioned vocal tracts. There is not infinite variation in human vocal tract dimensions, and, to the extent that speakers have very similar sized vocal tracts, they will output similar acoustics. Between-speaker variation in the acoustics of speech sounds as a function of difference in vocal tract dimensions can therefore be expected to be limited, and it is additionally constrained by the fact that formants have to realize specific speech sounds. If you want to say an ee vowel there is a limited range of configurations you can put your vocal tract into: otherwise it would be producing another vowel!

But there is worse to come. It is a phonetic truism that an individual never produces the same sound in exactly the same way (Perkel & Klatt 1986: ii), and this gives rise to within-speaker variation in speech acoustics. There is a tendency for within-speaker variation to vary directly with time, so that the same individual will differ more across the space of a month than a day. So although different speakers may differ in their acoustic output, an individual will definitely do so. And yet worse – as already mentioned, individuals differ in their variability. Some may differ a lot, some not so much, and some may differ a lot on one occasion and not so much on another.

It can be appreciated that, if individuals differ to the same extent as different speakers – if within-speaker variation is the same as between-speaker variation – it will be difficult to distinguish same-speaker recordings from different-speaker recordings.

The final piece of bad news is that different speech sounds can differ in the amount of between- and within-speaker variance in their acoustics. Thus, in order to be useful for forensic speaker comparison, a speech sound has to typically show large between-speaker variation, or small within-speaker variation, or, hopefully, both. The next stage in the forensic voice comparison process is therefore to find out how well acoustics from ee vowels can be used to distinguish recordings from the same speaker from recordings from different speakers.

**Likelihood ratio based forensic voice comparison**

The likelihood ratio provides a marvelous way of doing that. A positive log$_{10}$ likelihood ratio (i.e. greater than 0) implies that one is more likely to get the difference in acoustics if the recordings came from the same speaker, whereas a negative log$_{10}$ likelihood ratio implies that one is more likely to get the difference in acoustics if the recordings had come from different speakers. In addition, the more the likelihood ratio deviates from 0, the greater its strength. So we have in the likelihood ratio a potential discriminant function for discriminating between same-speaker recordings and different-speaker recordings. In addition the magnitude of the likelihood ratio also quantifies how strong the discrimination is. So how do we use all this on the ee vowel acoustics?
The idea is simple. We need to have recordings from different speakers and recordings from the same speaker. These constitute the known data, the ground truth. We then need to estimate likelihood ratios for comparisons using same-speaker recordings, and likelihood ratios for comparisons using different-speaker recordings. To the extent that comparisons of same-speaker recordings yield $\log_{10}$ likelihood ratios greater than 0, and to the extent that comparisons of different-speaker recordings yield $\log_{10}$ likelihood ratios less than 0, the feature used to compare the recordings can be considered potentially valid. How this is done is described below.

**Hong Kong MTR database**

The forensic potential of the *ee* vowel formants can be investigated by using the Hong Kong Mass Transit Railway database. This is a small (ca. 50 speakers), quasi map-task database, collected in 2012 as part of a go-to-whoa Cantonese forensic voice comparison experiment run as a one-semester postgraduate course at the Hong Kong University of Science and Technology. Its purpose was two-fold: to be a reference sample for two real-world cases for which preliminary reports had been commissioned by a Hong Kong-based mandating authority; and to be resource for controlled but natural speech for testing likelihood ratio-based approaches to forensic voice comparison in Cantonese. In addition to its obvious educational aim – learning how to properly evaluate the strength of evidence supporting competing hypothesis should surely be part of every student’s education – the course had two other objectives. The continuing poor understanding of likelihood ratios, especially on the part of the legal profession, had suggested that the LR framework actually might be much more of a hindrance to the court than a help. I wanted to see, therefore, whether it was indeed too difficult to be taught and learnt.

In order to collect the Cantonese data, subjects were given a map of the Hong Kong Mass Transit Railway and asked different questions about it. Among these were ten questions designed to elicit the answer 第二個站大圍hoi the second station, which contained an *ee* vowel in the word 二字 *yih* *two*. Subjects were instructed to answer in a whole sentence, so a typical exchange between experimenter and subject was:

**Q**: Tin Hauh haih Wànnjáí *jìh* hauh daihyatgo dihnghaih daihyihgo jahhm a?

天侯係灣仔之後第一個定係第二個站呢

Is Tin Hau the first or second station after Wan Chai?

**A**: Tin Hauh haih Wànnjáí *jìh* hauh daihyihgo jahhm.

天侯係灣仔之後第二個站

Tin Hau is the first station after Wan Chai.

Provision was made in the task for seven such answers containing the *ee* vowel to be elicited in one recording session. In reality we sometimes got less due to incorrect answers, and sometimes extra *yih* tokens could be used if the subjects counted the number of stations out loud. In all, however, we obtained between 5 and 10 *ee* replicates per speaker per recording.

If one wants to mirror a forensically realistic situation one of the essential components of a forensic voice comparison database is that it include non-contemporaneous recordings (Enzinger 2012). This is because suspect and offender recordings are perforce non-contemporaneous, and within-speaker variance is greater for non-contemporaneous recordings. For this experiment speakers were recorded on two occasions separated by about a month. Following a protocol for collection of forensic speech data (Morrison et al. 2012), participants communicated by phone while high quality recordings were made from their lapel microphones.

26 young male subjects from the database were used. The data for testing thus consisted of 26 pairs of non-contemporaneous same-speaker recordings and 325 pairs of different-speaker recordings. The first three formants were extracted with Praat’s Burg option. It was found that a setting of four formants below four kHz usually gave acceptable results, in the conventional acoustic-phonetic sense of the extracted formant centre-frequencies.

---

9 I thank the Hong Kong University of Science and Technology for making it possible to run the Mass Transit Railway forensic voice comparison experiment from which many of the Cantonese examples were taken. This was done as part of their postgraduate Humanities course *Topics in Chinese Phonetics: Forensic Voice Comparison in Cantonese* which I ran in 2012. Most of the thanks though must go to the course participants who did all the hard work collecting, measuring and processing the data.
appearing to track through the middle of the spectrographic formants – as can be seen in figure 3.2 – and this was adopted as default. When this was not the case, the settings were changed to get a better visual resolution. This usually involved increasing the number of formants to extract to five. An example of the formants for a speaker’s seven yih tokens in one of their recordings is shown in fig.3.4(A), where they are plotted as functions of raw duration. It can be seen that the tokens range from about 6 to 10 centiseconds, and the formant frequency values cluster fairly tightly. Nevertheless, the within-speaker variation is still obvious, even for a single session recording!

The formant frequency trajectories were then parametrised using polynomial coefficients. This has been shown (Morrison 2008) to be superior to sampling formant frequencies at given durational points, even for monophthongs where little change in F-pattern trajectories can be expected (Rose 2015). As the formant trajectories in yih do not change very much it is not expected that high-order polynomials are necessary, and in fact the optimum parametrization was obtained with only the intercept and slope (i.e. first-order polynomials).

One consideration in calculating polynomial coefficients is what kind of duration base to use. It was found that, as previously, e.g. Morrison (2009a, p. 2392, 2395), generally better results are obtained when polynomials are calculated using equalised as opposed to raw duration. Panel B of fig. 3.4 shows the corresponding equalized duration first-order polynomial trajectories for the three formants, with the mean polynomial overlaid in a thick black line. You can see that the duration is now equalised and expressed in percent. Panel C of fig. 3.4 compares the mean polynomial F-pattern of the speaker’s two non-contemporaneous recordings. It can be seen that, plotted as a function of equalized duration, the second recording F2 is very slightly lower, but otherwise this speaker’s formants have not changed much at all between the first recording and the recording about a month later. The probability of getting this degree of similarity assuming both recordings had come from the same speaker is in fact very much bigger than the probability of getting the agreement assuming they had come from different speakers.

**Figure 3.4:** Stages in yih F-pattern acoustic data extraction and comparison for a single speaker. A, B = Raw and polynomially-smoothed data from one recording. C = comparison between mean polynomially-smoothed F-pattern in two recordings separated by about a month. Thick black lines in B = mean values. Y-axis = frequency (Hz). X-axes: (A) = duration (csec.), (B,C) = equalized duration (%).
Each speaker’s mean $ee$ F-pattern acoustics from their first recording were then compared with their second recording (as in figure 3.4 panel C), and the difference between each speaker’s first and second mean recordings expressed with a likelihood ratio. In other words, it was estimated how likely it was to get the difference between the mean values assuming they had come from the same speaker and how likely it was to get difference between the mean values assuming they had come from different speakers, and the ratio of these likelihoods used as a measure of the difference. Differences that are quantified taking both similarity and typicality into account, as done in this case with a likelihood ratio, are technically called scores, and so the measures of differences in this case are called likelihood ratio scores. Because there were 26 same-speaker comparisons, there were 26 known same-speaker likelihood ratio scores. The same thing was done for the different-speaker comparisons: each speaker’s mean $ee$ F-pattern acoustics from their first recording was compared with the mean $ee$ acoustics of the other speakers’ first recordings to get 325 known different-speaker likelihood ratio scores.

The likelihood ratio scores were calculated with the multivariate kernel-density likelihood ratio (MVKD) formula developed at the Joseph Bell Centre for Forensic Statistics and Legal Reasoning (Aitken & Lucy 2004). This work-horse has been used in many previous studies as well as real-world case-work (Rose 2013a). A Matlab implementation is available (Morrison MVKD), and a version is also available as part of the $R$ package Comparison (Lucy 2013).

The MVKD formula works by treating the difference between the means of the two samples to be tested as the evidence, and estimating a likelihood ratio for it. Its main strength lies in its ability to handle many variables at once, which is why it is called multivariate. It also takes any correlation between variables into account. This is vital when calculating likelihood ratios from multivariate data: if two variables are independent, then the likelihood ratio from their combination is the product of their individual likelihood ratios. But if the two variables are correlated, their combined likelihood ratio will be less than the product of their individual likelihood ratios (Rose 2010). Another noteworthy feature is that takes the number of replicates in each sample into account, as it makes sense that the more replicates from which to derive a mean value, the more representative the mean value will be and the stronger the evidence. The scores can either be treated as normally distributed, or modelled with a kernel density (Aitken & Taroni 2004: 330-341). The MVKD formula is discussed in greater detail in Rose (2013b: 92-95).

The MVKD calculation of the likelihood ratio score for each same-speaker and different-speaker comparison was carried out with cross-validation, using a leave-one-out approach, whereby all data for the particular pair being tested are removed from the reference sample used to estimate the typicality of the comparison. Failure to do this with a small number of speakers will over-estimate the performance.

**Validation**

It surely stands to reason that the forensic expert should be able to demonstrate to the trier-of-fact that their system does what they claim it to do:

> “Without actual empirical evidence of the ability of a forensic feature-comparison method to produce conclusions at a level of accuracy appropriate to its intended use under circumstances reasonably related to this use, an examiner’s conclusion that two samples are likely to have come from the same source is completely meaningless.” Lander (2017: 1-2)

The importance of testing your approach, or validation, was articulated already in a 1993 US Supreme Court ruling on the admissibility requirements for scientific evidence, where it was stated that a judge must take into account the validity and reliability of a method, especially whether the method has been tested and found to have an acceptable error rate. This ruling is usually known as Daubert v Merrell Dow Pharmaceuticals or Daubert for short. More recently, the Daubert requirements have been seen as a major force in the paradigm shift in forensic evidence evaluation (Saks & Kohler 2005, Morrison 2009b). The need to test the validity of forensic analyses, and to demonstrate the commitment of an approach to testing, was reiterated in the USA President’s Council of Advisors on Science and Technology report (Holdren et al. 2007). Forensic voice comparison has arguably been at the forefront of system validation (e.g. González-Rodriguez et al. 2007).

So, how well does our likelihood ratio score based system do in distinguishing same-speaker recordings from different-speaker recordings? The probability distributions of the 26 likelihood ratio scores from the known same-speaker comparisons and likelihood ratio scores from the 325 known different-speaker comparisons are shown in fig. 3.5. Recall that different-speaker comparisons are expected to have scores less than $\log_{10} 0$, and same-speaker
comparisons are expected to have scores greater than \( \log_{10} 0 \), and it can be seen that, as might be hoped, most of the scores from the same-speaker comparisons (red) are greater than 0 and most of the different-speaker scores (blue) are less than 0. Nevertheless there are clearly some same-speaker comparisons which were evaluated counterfactually, with scores less than 0, and there are also some incorrect different-speaker scores – indeed, some different-speaker scores are peeping out from above the highest same-speaker score!

**Figure. 3.5.** Kernel-density probability distributions of multivariate likelihood ratio scores for same-speaker (red) and different-speaker comparisons using yih F-pattern. X-axis = \( \log_{10} \) multivariate likelihood ratio score, y-axis = probability density.

Plotting the cumulative distribution of scores allows us to better visualize the values involved and such a plot is shown in fig. 3.6. Scores from different-speaker comparisons increase towards the left and are shown in green; same-speaker scores, in red, towards the right. This particular way of plotting – it is called a type II Tippett plot (Drygajlo et al. 2015, p. 20-22) – means that the horizontal axis has two interpretations, depending on whether one is reading-off scores from same- or different-speaker comparisons. For different-speaker comparisons the scale indicates scores greater than the given value. Fig. 3.5 thus shows that about 85% of different-speaker comparisons had \( \log_{10} \) scores less than 0. For same-speaker comparisons the scale quantifies scores smaller than the given value. Thus it can be seen that about 82% of same-speaker comparisons had \( \log_{10} \) scores greater than 0. Recall that different-speaker comparisons are expected to have scores less than 0, and same-speaker comparisons are expected to have scores greater than 0, so the likelihood ratio score based on F2 and F3 trajectories in yih appears to be able to distinguish same-speaker from different-speaker comparisons reasonably well. It will be correct in about 84% of cases overall. Equivalently, if you assume that scores greater than 0 indicate a same-speaker comparison and scores less than 0 indicate a different-speaker comparison you will be wrong about 16% of the time.
Figure 3.5. Type II Tippett plot with scores derived from comparisons based on polynomial coefficients of yih F2 and F3. X-axis = Log_{10} LR score, y-axis = cumulative proportion of scores. red = same-speaker comparisons, green = different-speaker comparisons.

But all is not as good as it seems. With the performance of likelihood ratio-based detection systems it is not just discriminatory potential that is important. The strength of the evidence – quantified by the magnitude of the ratio – counts just as much. On this count, the different-speaker comparisons perform fairly well: of the cases which were evaluated counterfactually – with scores greater than 0 – the worst cases (about 3% of comparisons) had scores between about log_{10} 1 and log_{10} 2. This is saying that you will find a few different-speaker comparisons incorrectly evaluated as being between 10 and 100 times more likely if they were from the same rather than different speakers. But look at the magnitude of incorrectly evaluated same-speaker comparisons. It is hair-raising: they almost go as high as log_{10} -6! In fact, over 10% of same-speaker comparisons are going to have counterfactual scores suggesting that they were between 100 and 1,000,000 times more likely if they were different speaker comparisons! You get this result mostly because of the nature of within- and between-speaker variation observed in speech sound acoustics. Generally, between-speaker variation is greater than within-speaker variation. But the differences between some speakers are the same, or even less than you might find between the same speaker on different occasions; and some individuals can differ on different occasions more than is sometimes found between different speakers.

We need a measure of how good, or bad, the performance with ee F-pattern in fig. 3.5 is. The performance of a likelihood ratio-based detection system like this is currently assessed by the *log likelihood ratio cost* $C_{llr}$. $C_{llr}$ heavily penalizes large counterfactual values of the type evident in fig. 3.5. It can be interpreted in several ways (Brümmer and du Preez, 2006, p. 273). Its information-theoretic construal (Brümmer and du Preez, 2006, p. 264-266) relates to the average amount of information the system provides to its end-user. $C_{llr}$ values below 1 – the smaller the $C_{llr}$, the better – indicate that the system has the capability of reducing the user’s uncertainty in the hypothesis. Fig. 3.5 shows the $C_{llr}$ for this system is 1.2, indicating that it is actually not giving us any useful information at all.

Now from the point of the legal system, obviously, the thing to avoid is convicting the innocent. So even the smaller values for incorrect different-speaker comparisons are cause for concern. But one would also not want the same-speaker comparisons to be susceptible of such egregious errors. The solution to this is what is known as calibration. There is no space to go into this in detail – Morrison (2013) provides an excellent tutorial – but briefly in calibration, likelihood ratio scores are transformed into true likelihood ratios by various means, one of which is logistic regression.

Fig. 3.6 shows the results of a logistic-regression calibration on the data in fig. 3.5. The true multivariate likelihood ratios from the calibrated scores are shown in solid lines, and the scores in dotted lines for comparisons. It can be seen that the calibration has effectively reduced the magnitude of counterfactual scores for both same-
speaker and different-speaker comparisons. The reduction is dramatic for the same-speaker comparisons, the worst of which is now only about associated with a counterfactual likelihood ratio LR about 10, rather than 1000000. The different-speaker comparisons also show a reduction in magnitude of counterfactual likelihood ratios: the worst case is now less than 10. The improvement comes at the cost of magnitude of overall strength of evidence, however: the strongest evidence that one can expect for same-speaker comparisons has reduced from slightly less than 100 to slightly less than 10, although the calibration has not changed the average strength of evidence expected: about 50% of same-speaker comparisons can be expected to have a likelihood ratio of about 4 or less. There is a considerable drop in expected likelihood ratio values for different speaker comparisons: 50% of different-speaker comparisons can be expected to have a likelihood ratio of about 5 or greater, whereas for the uncalibrated data 50% of different-speaker comparisons can be expected to have a likelihood ratio of about 30000 or greater. The Cllr value for the calibrated data is 0.63, which shows that now the system is reasonably valid and provides some degree of information. It is of interest to note that similar results were obtained with a likelihood ratio-based discrimination with ee and ü vowel F-pattern acoustics in Mandarin, using 64 speakers (Morrison et al. 2008).

Figure 3.6. Type II Tippett plot showing effect of calibration on MVLR scores derived from comparisons based on polynomial coefficients of yih F2 and F3. X-axis = Log₁₀ LR score, y-axis cumulative proportion of scores. Dotted lines = uncalibrated scores, solid line = calibrated likelihood ratios. red = same-speaker comparisons, green = different-speaker comparisons.

Whether this performance is adequate for the mandating authority is a question for the mandating authority. But there are a few indicators that can help them. If the hypothesis to be tested is that the suspect said the incriminating speech, we can use Bayes’ Theorem to look at the range of posterior probabilities the system would furnish for same-speaker comparisons. It is not great. Given this system, the greatest strength of evidence in support of a same-speaker hypothesis is a little below 10, and the average strength of evidence to be expected is about 3 or 4. Consider the most advantageous prior odds for the prosecution of 1 to 1, that is, there is only one other person the offender could be. With the maximum likelihood ratio for this system of 10, Bayes’ Theorem tells us that with these priors and no other evidence the maximum posterior probability will be 91%¹⁰. Does 91% sure it is the suspect leave room for reasonable doubt? It depends on the seriousness of the crime. But with less advantageous priors, and/or weaker evidence strength the posterior will surely fall well below that required for a reasonable doubt. With priors of 1 to 9 (there are 10 people, including the suspect, who could have said the incriminating speech), and with even a maximum likelihood ratio of 10, with no other evidence the posterior probability is now

¹⁰ With prior odds of 1:1 and a likelihood ratio of 10:1, the posterior odds are 1/1 * 10/1 = 10/1, which is equivalent to a probability of 10/11, or about 91%.
53\%^{11}. We probably would need more stronger evidence than that provided by just yih F-pattern acoustics! With this in mind, the next sections look at evidence from other speech sounds.

**Strength of evidence from the Cantonese diphthong ei**

The previous section has shown how forensic semi-automatic voice comparison can be performed using the second and third formants of the Cantonese monophthong ee. The important idea was to see what strength of evidence the formants would provide. As a monophthong, the formant trajectories in ee do not change much and so they could be modelled by a low order polynomial involving just the intercept and the slope. This means that just two parameters were involved in estimating a likelihood ratio. Varieties of Chinese are typologically fairly rich in diphthongs, however – where there tongue glides from one position to another during the vocalic portion of the syllable, and there may also be changes in lip rounding. Because the articulators are changing during the vocalic portion of the syllable, so will the formants. Fig. 3.7 illustrates this with the Cantonese word sei four, which contains the diphthong [ei] (I will write it as ei). Formant centre-frequencies are superimposed in red.

In ei the tongue shifts forwards and upwards in the front part of the mouth. Since the first formant is correlated inversely with vowel height, its frequency falls, in this case from about 500 to 300 Hz. The second formant is correlated with the position of the tongue in a front-back dimension and therefore shifts a little upwards, in this case from about 1500 to 2200 Hz. (F2 is also correlated with lip rounding, but since that does not change during ei, changes in F2 just reflect the slight change in the front-back position of the tongue.) The third formant also shifts upwards a little, from about 2400 to 2600 Hz. The F-pattern trajectories are clearly more complex than in the monophthong ee, and the polynomial needed to model them will also be more complex, thus offering the possibility of more parameters qua polynomial coefficients upon which to estimate a likelihood ratio. This section describes what strengths of evidence are obtained with a Cantonese diphthong ei. The details are in Pang & Rose (2012).

![Figure 3.7. Spectrogram of sei four with superimposed formant centre-frequencies showing the dynamic F-pattern typical of diphthongs.](image)

To do this, the responses of 18 speakers from the HKMTR database were used replying to a question expecting the answer 四 sei four. For example:

Q: Taaigu töhngmâih Töhnglōhwaän jügān yâuh géídōgo jaahm a?

大古同埋銅邁灣之間有幾多個站阿

How many stations are there between Prince Edward and Causeway Bay?

---

11 With prior odds of 1:9 and a likelihood ratio of 10:1, the posterior odds are \( \frac{1}{9} \times \frac{10}{1} = \frac{10}{9} \), which is equivalent to a probability of \( \frac{10}{19} \), or about 53\%. 
A: Taaign tòhngmāih Tōhnglōhwaān jīgaān … yāuh seigo jaahm.

Between Prince Edward and Causeway Bay there are … four stations.

The F-pattern trajectories were modelled with cubic polynomials. The non-contemporaneous within-speaker variation was typical: some speakers did not differ much; some differed in F2; some in F3 and some in both. Fig. 3.8 shows some examples of the cubically modelled F-pattern in sets of nine replicates from each recording session. The speaker on the left shows little variation over one month in his F-pattern. The speaker on the right shows some differences in both F2 and F3.

**Figure 3.8:** Examples of non-contemporaneous F-pattern differences (modeled with cubic polynomials) in Cantonese /ei/ for two speakers.

The results of the LR-based analysis of the 16 speakers’ ei F-pattern are shown in the Tippett plot in fig. 3.9. This shows the likelihood scores and ratios that were derived from the cubic polynomial coefficients of all three formants. The plot for the uncalibrated MVLR scores has been included as it gives a clearer idea of the ‘shifting and scaling’ effects of calibration when compared with the calibrated likelihood ratios. As can be seen, the $C_{\text{LR}}$ is 0.48, and so a little better than for the F-pattern in ee. This indicates that there is a little more forensically useful speaker-specific information in the F-pattern trajectories in the diphthong ei than the vowel ee.
Figure 3.9. Tippett plot for 16 speakers’ F-pattern trajectories in et. Dotted lines indicate uncalibrated scores. X-axis = Log_{10} LR score, y-axis cumulative proportion of scores. Dashed lines = uncalibrated scores, solid lines = calibrated likelihood ratios. red = same-speaker comparisons, green = different-speaker comparisons.

The Cantonese diphthong [ɔy] in the word 水 séui water was also tested using 15 speakers from the HKMTR database (Li & Rose 2012). It returned a C_{llr} of 0.55, again indicating usefulness for forensic voice comparison.

Triphthongs
What about triphthongs? Their F-pattern trajectories are even more dynamic than diphthongs, as they realise glides between three vocalic targets, so one might expect some forensic voice comparison potential from them too. The Cantonese triphthong [iau] in the word 油 yàuh oil was likelihood ratio-based tested using 15 speakers from the HKMTR database (Chen & Rose 2012) and returned a C_{llr} of 0.6. The Mandarin triphthong [iau] in the word 一 yāo one was tested by Zhang et al. (2011) using 20 speakers and returned a much more impressive C_{llr} of 0.35. Whether this difference in C_{llr} is due to differences in phonetic structure, accompanying tone, or speaker number is not clear. Perhaps the most important thing is that the triphthongs are both giving useful information. Zhang et al. (2011) and Huang et al. (2012) also found that when combined with the results from a likelihood ratio-based system using a completely automatic approach the information from the triphthong resulted in a substantial improvement in the combined system performance.

Cepstral spectrum in nǐh
Monophthongs, diphthongs and triphthongs are not the only segments that are of potential use in forensic voice comparison. Cantonese, like many other varieties of Chinese, also has syllabic nasal consonants with durations as long as vowels. Although conservative Cantonese has both bilabial and velar syllabic nasals, e.g. in /m/ 無 nǐh not and /ŋ/ 五 nīh five, many speakers now have just a bilabial nasal /n/. Nasal cavity resonances are prima facie promising as forensic voice comparison features because they can be expected to show a high ratio of between- to within-speaker variance which is probably the most important factor in speaker recognition (Nolan 1983). The complicated internal structure of the nasal cavities and sinuses gives potential for between-speaker variation, and their relative rigidity would contribute to low within-speaker variation.

Because of its potential, one of the segments deliberately elicited in the HKMTR study was the syllabic nasal [n], as in the word nǐh ‘five’. For details see Rose (2017: 487-490) and Yim & Rose (2012).

Tokens were elicited with the following type of exchange:

Q: Taaiág tóhngmáiëh Waänjái jīgaän yáuh géidōgo jaahm a?
大古同埋灣仔之間有幾多個站阿
How many stations are there between Prince Edward and Wanchai?
Quantifying the acoustics of nasals, however, is considerably more complicated than with oral vowels. With oral vowels the supralaryngeal tract can be treated as a uniform acoustic tube with a single source at the glottis, and the identification and interpretation of F-pattern, in conjunction with the auditory percept, is reasonably straightforward for forensic purposes. With nasals or liquids, however, extra supralaryngeal cavities are created which have the effect of introducing additional resonances. The extra cavities also introduce antiresonances, where energy is absorbed, and these can shift, attenuate or even obliterate formant energy. These effects make it difficult to identify comparable formants either between or within speech samples, and so other methods must be used to quantify nasal acoustics.

Figure 3.10 is a typical example of the complexity of a syllabic nasal. Its left panel shows a spectrogram of a bilabial syllabic nasal /m/ in the phrase 有五個 yáuh míhgo there are five … . It is immediately clear that this involves not just additional resonances but also a complex sequence of supralaryngeal labial, dorsal and velic articulations extending over a relatively long duration of ca. 25 csec., between supralaryngeal closure at ca. csec. 15 and supralaryngeal release at ca. csec. 35. The speaker first articulates a bilabial nasal [m] with the tongue position inherited from the preceding vowel in yáuh. This portion extends from bilabial closure at ca. csec. 15 to ca. csec. 20. The speaker then raises their tongue dorsum in anticipation of the following velar stop /k/ in go thus producing a co-articulated [m̥]. This lasts until ca. csec. 27, when transients (most visible between ca. 3.5 kHz and 4.5 kHz) show the release of the bilabial closure resulting in a velar nasal [n̥]. A weak transient at ca. csec. 35 (visible around ca. 3.5 kHz) indicates the release of the dorso-velar occlusion, but it is clear that at this time the velic port is still lowered, as the low frequency acoustic nasalization can be seen to continue well into the following /o/ vowel, so there cannot have been much pressure build-up behind the occlusion. The portion from the bilabial release, although a velar nasal, can probably therefore best be regarded as the hold phase of the velar stop /k/ in go.

The spectral details of nasals do not appear with great clarity in spectrograms, and so the right panel of fig. 3.10 shows FFT and LPC spectra corresponding to the portion with bilabial closure. The lowest resonance, at about 250 Hz is the low quasi-Helmholtz resonance expected in nasals and nasalized sounds. This can also be seen on the spectrogram extending further on both sides representing nasalisation of the surrounding vowels. The next lowest pole, at ca. 800 Hz, is probably associated with the nasal cavity tuned by the open velopharyngeal port, since this is expected to occur between 750 Hz and 1 kHz (Stevens, 1998, p. 489) and has fairly high amplitude. There are four additional resonances up to 3 kHz which presumably reflect nasal and oral cavity, and possibly also sinus resonances. Depending on how much of the oral cavity is available from the nasal’s place of articulation, the lowest antiresonance will be present from above 1 kHz in bilabials to ca. 3 kHz in velars. Antiresonances do not
usually show up well in spectrograms, but there is certainly a drop in amplitude in the expected region, above the 2nd resonance.

When comparing nasals with respect to their acoustic properties, then, their complexity usually makes it impractical to quantify formants. Instead their spectrum can be quantified. This can be done in several ways, but usually it involves calculating a set of cepstral coefficients which define the spectral shape over a given frequency range. Further explanation of how cepstral coefficients define a vocalic spectrum and their use in likelihood ratio-based testing may be found in Rose (2013b). In the case of the Cantonese bilabial nasal seven coefficients were used to capture the shape of its spectrum between 350 Hz and 3.5 kHz, which corresponds to a typical range of frequencies usually passed in a telephone. Figure 3.11 shows the cepstral spectrum of the same token of ōn shown in figure 3.10. The portion in red indicates the spectrum captured by the seven cepstral coefficients used.

![Figure 3.11. Illustration of cepstral bandlimiting in Cantonese syllabic nasal [m]. Dotted line = cepstrally-smoothed 12th order LPC spectrum to 5 kHz. Solid line = derived cepstrally-smoothed sub-band spectrum over nominal telephone pass-band of 350 Hz – 3.5 kHz. X axis = frequency (Hz), y axis = log magnitude.](image)

How did this parameterisation of the syllabic nasal perform? Fig. 3.12 shows the Tippett plot for 34 same- and 561 different-speaker comparisons using the seven cepstral coefficients from 34 of the HKMTR speakers’ bandlimited cepstral spectrum in ōn five. As with the diphthong ei, the uncalibrated scores lie way to the left of the calibrated data. This is typical when many parameters are involved in the comparison, as here. As can be seen, the strength of evidence from the calibrated likelihood ratios is about the same as with yi, a C_{LR} of 0.65. Testing shows therefore that syllabic nasals are forensically useful speech sounds, but, contrary to expectations, not especially privileged.

---

12 Thanks to Dr. Frantz Clermont for both the idea and the Matlab code for band-limited cepstrum.
Strength of evidence from tonal F0.

In the first part of this chapter it was demonstrated how Bayes’ Theorem operates, by attempting to estimate a class member’s sex from their fundamental frequency – the acoustical correlate of the rate of vibration of the vocal folds. It was explained that this works because fundamental frequency (F0) is a function of the length and mass of a speaker’s vocal folds. And since males tend to have longer, more massive folds than females, their F0 values tend to be lower.

The principle also applies irrespective of sex: differences between speakers in the length and mass of their folds will be reflected in their F0 values, and F0 is actually a very popular feature in forensic speaker identification. Varieties of Chinese are of course tonal, so an obvious question, given that F0 is the main acoustical correlate of tone, is: what sort of strength of evidence does tonal fundamental frequency have? Conservative Cantonese varieties have six tonemes (Rose 2000) and the strength of evidence for three of them was investigated using the HKMTR database: the low level, low to high rise and low falling tones. Details are in Wang and Rose (2012), Li and Rose (2012) and Chen and Rose (2012), but basically the tonal F0 in 二 yiḥ two (low level tone), 水 séui water (low to high rise tone) and 油 yù oil (low falling tone) was modelled with low-order polynomials in the same way as the vowels above, and then tested to see how well same-speaker recordings could be distinguished from different-speaker recordings using likelihood ratios. The results were uniformly poor, with $C_{llr}$s of 0.77 (low level tone), 0.95 (low falling tone) and 0.86 (low to high rising tone). In order to show just what a bad Tippett plot looks like, fig. 3.13 shows the low falling tone. The calibrated plots for the same- and different-speaker comparisons are almost touching. This magnitude of $C_{llr}$, although less than 1, can hardly be expected to make the trier-of-fact change their belief at all when presented with new evidence.
Figure 3.13. Tippet plot for 15 Cantonese speakers’ low falling tonal F0. X-axis = Log_{10} LR score, y-axis cumulative proportion of scores. Dashed lines = uncalibrated scores, solid lines = calibrated likelihood ratios. red = same-speaker comparisons, green = different-speaker comparisons.

Short-term fundamental frequency
Instead of just using the tonal F0 trajectory in a single monosyllable, as above, a profile can be built-up of a speaker’s F0 values over a stretch of speech. It is called long- or short-term F0 depending on the amount of speech used. Fig. 3.14. shows the distribution of F0 measurements of a Cantonese male speaker over about 2.5 minutes of his speech. It can be seen that his F0 values lie between about 50 and 200 Hz, but most of them are found between about 80 and 150 Hz. The F0 values below about 80 Hz represent creaky phonation type, where the rate of vocal fold vibration is low because the vocal folds are dampened by the false cords. The F0 distribution is slightly positively skewed with slightly more higher values than lower. The distribution has been quantified according to six parameters: mean, mode, standard deviation, skew, kurtosis, and maximum probability density.

Figure 3.14. Distribution of long-term F0 from a male speaker of Cantonese. X-axis = F0 (Hz), y-axis = probability density.

The same caveat applies with F0 distributions as with formants. Because there is not a wide range in vocal fold length and mass, it is perfectly possible to find different speakers with very similar values and between-speaker variation is limited. In addition, within-speaker variation in F0 is not advantageous. As well as the many linguistic uses of F0, which encodes tone, intonation and stress, many non-linguistic factors, like state of health, are also
known to affect it. This multiplicity of factors has an adverse effect on its between-to-within-speaker variance ratio by increasing the latter. Since the inherent strength of forensic speaker recognition features relies primarily on their ratio of within-to-between-speaker variance, one would not expect particularly good strength of evidence from global F0 properties. So it is of interest to see how well F0 performs with the HKMTR Cantonese data.

F0 distributions over a stretch of voiced speech more typical of forensically realistic samples were obtained with a task in the HKMTR elicitation where speakers had to describe how to get from one station to another, where the journey was particularly complicated, or sometimes not possible. Subjects’ responses to these how to get from a to b questions typically lasted for several seconds – especially when they realised they had got lost and had to backtrack! The F0 distributions were then quantified with the six shape parameters listed in fig. 3.14, viz: mean, mode, standard deviation, skew, kurtosis and maximum probability density. These parameters were used as input to the MVKD formula.

As a salutary reminder that it is always possible to find different speakers with similar acoustic values, figure 3.15 illustrates data from two speakers whose non-contemporaneous short-term F0 distributions were sufficiently similar for the difference between their distributional parameters to be counter-factually evaluated (with a log_{10}LR of 0.77). The upper row shows data from the speakers’ first, and the bottom from their second recording session. The different colours represent the distributions from the different journeys and the box contains amount of net voiced speech in each individual journey described (so for example Alan’s first recording contained the description of five journeys, the shortest of which contained 9.4 and the longest 19.59 seconds net voiced speech. The grey shading represents distribution from all journeys combined.

Fig. 3.16 shows a Tippett plot for comparisons using parameters from 29 HKMTR speakers’ short-term F0 distributions. The C_{LR} of 0.55 is on a par with the other values demonstrated from the vocalic segments, and indicates a certain degree of usefulness.
The deleterious effect of mismatch

A very similar experiment was carried out using short-term F0 parameters from 90 male N.E. Mandarin speakers carrying out three different tasks (Rose & Zhang 2018). One task, involving a co-operative exercise over the phone, gave a similar strength of evidence to the Cantonese data ($C_{llr} = 0.53$). However, other tasks yielded considerably worse strengths of evidence, with $C_{llr}$s of 0.79, 0.75, showing the sensitivity of F0 to different genres of speech interaction. Worse still, the performance degraded further when speech from different tasks was used. Simulating a typical forensic scenario where the offender’s conversational offender speech is compared with the suspect’s police interview speech gave a $C_{llr}$ of 0.87. This shows that F0 is sensitive to mismatch in recording conditions.

Combining evidence with likelihood ratios

The validity testing of the various individual speech sounds described above is typical in not showing particularly good strength of evidence. Normally, one does not expect a single segment to be capable of delivering a forensically useful likelihood ratio, although Bayes’ Theorem tells us that ultimately that will depend on the prior odds. However, if likelihood ratios from many segments can be combined, the magnitude of the overall likelihood typically shifts away from 0, and with it the strength of the evidence increases also.

The likelihood-ratio framework makes it very simple to combine evidence of either the same or different types. As long as the features are not correlated, their likelihood ratios can be combined, as in the example above, by simply taking their product. But this is rarely the case with speech sounds. However, methods have been developed within the automatic speaker recognition community for handling this problem. These are usually known by the term fusion. One well-known method of fusion is done with a very similar process to calibration, and in fact fusing a set of likelihood ratio scores simultaneously calibrates them (Brümmer Focal, Brümmer & Swart 2014).

In order to demonstrate how combining likelihood ratios from different speech sounds increases the strength of evidence, the likelihood ratio scores from the yih F-pattern, the short-term F0 and the bandlimited mh cepstral spectrum for 23 speakers of the HKMTR database were used. Fig. 3.17 shows, with thin lines, the individual Tippet plots for these three speech sounds. The Tippett plot for the fused data is shown with thicker red lines. It can be seen that the fused data plot pretty much encloses the plots for the individual data, and a glance at the $C_{llr}$

---

13 Dr Zhang is responsible for collecting the data used in the experiment on mismatch in short-term fundamental frequency, which was supported by the National Social Science Foundation of China Key Program (Grant No. 16AYY015), Southwest University of Political Science and Law Research Funding (2015-XZRCXM003), and Chongqing Social Enterprise and People’s Livelihood Guarantee Scientific and Technological Innovation Special Research and Development Key Project (cstc2017shms-zdyfX0060).
values in the legend shows that fusing the scores from the individual speech sounds has resulted in an increase in the strength of evidence. The $C_{llr}$ values for the individual scores are 0.64, 0.63 and 0.46, compared to 0.26 for the combined evidence. One can also see that the average likelihood ratio for same-speaker comparisons from the fused data has increased from about $\log_{10} 0.6$ to about $\log_{10} 1.5$, or from about 4 to 30. This magnitude of likelihood ratio is beginning to have some heft. Another demonstration of two speech sounds being fused to yield an increase in strength of evidence may be found in Zhang & Ding (2020) where the cepstral spectra of a vowel and the consonant $s$ in Mandarin were used. Another variation on the theme of speech sound acoustics is reported in Wang & Zhang (2015), who extracted F0 and values for the first four formants automatically, not for particular speech sounds but for the whole of a speaker’s speech. They obtained very good results from the fusion of the features.

Figure 3.17. Tippett plots illustrating fusion of three individual speech sounds. X-axis = $\log_{10}$ LR s, y-axis cumulative proportion of likelihood ratios.

Summary and Envoi

This chapter has attempted to explain, using examples from Chinese, what is involved in forensic semi-automatic voice comparison.

In the first part of the chapter I gave examples of how Bayes’ Theorem works, because that is the logical framework in which the comparison must proceed. The essence of Bayes’ Theorem is to show how you should rationally change your prior belief in something when confronted by evidence (Jaynes 2003: 92). Your new belief is your old belief updated with the strength of the evidence.

Bayes’ Theorem states informally that if you want to know the probability that a hypothesis is true, given the evidence adduced in support of that hypothesis, then you have to know not only the strength of the evidence – the likelihood ratio – but also the probability that the hypothesis is true before the evidence is considered – in the form of the prior odds. Thus forensic voice comparison involves estimating the strength of the evidence in support of a hypothesis that a questioned voice has come from a known speaker: in other words, a likelihood ratio. It is not, and cannot be, about estimating the probability that the questioned voice comes from the known speaker. That would logically require the expert to know the priors, which they usually don’t; and it might also involve the expert in deciding the ultimate issue of guilt, which is usually up to the trier of fact to decide. Like Confucius’ 君子 (‘superior man’) in the quote at the beginning of the chapter, therefore, one needs to exercise restraint with what one cannot know.

The second part of the chapter gave some examples of how one goes about estimating a likelihood ratio from speech sounds. Various Chinese speech sounds were examined to see the kind of strength of evidence they can furnish: monophthongs, diphthongs, triphthongs, syllabic nasals and fundamental frequency. It was shown that, with the exception of tonal F0, they can all be expected to furnish information that will help the trier-of-fact. In particular, it was shown that evidence from speech sounds can get much stronger when they are fused. Although this chapter has used Chinese speech sounds, it should be noted that there is actually nothing special about them
when it comes to forensic voice comparison. Similar strength-of-evidence results have been obtained from speech sounds in other languages.

One important message from all of this was that it is not just the known and questioned recordings that are involved. If you want to know the probability of getting the questioned recording assuming it has come from someone else, you have to know about them. In other words you have to have an idea of the characteristics of the relevant population, and that requires a lot of hard work recording and analyzing different speakers other than the known and questioned voices.

But the leitmotiv of the chapter is really that speech sounds can be quantified in such a way as to allow the most important task of validating the forensic voice comparison system in a forensically realistic manner. Quantification is central to this. For the poet/singer Leonard Cohen, hell in the quote at the beginning of the chapter is a place where you can no longer quantify anything. A good way of saying that numeracy is the crucial underpinning to forensic voice comparison and forensic science in general, of course.

It would be remiss of me to leave it there, with the implication that this is all hunky-dory and the accepted way under the law. The likelihood ratio in forensic speaker recognition has been around for over 20 years now; in DNA, for much longer. The likelihood ratio framework was recently endorsed as best practice in forensic automatic and semi-automatic speaker recognition by the Board of the European Network of Forensic Science Institutes, representing 58 laboratories in 33 countries (Drygajlo et al. 2015). Despite this, the unfortunate fact is that the law is akin to the ‘foreign country’ in the famous opening line of Hartley’s The Go-Between: ‘The past is a foreign country. They do things differently there’. (A doubly appropriate metaphor, as the law, like Hartley’s ‘past’ is of course inherently conservative.) It has to be acknowledged that, with very few exceptions, the legal profession still has enormous trouble understanding things like likelihood ratios and the necessity for experts to demonstrate that their approach works. Having worked through this rather dense chapter you will probably feel some sympathy for that! There are no indications that things will change any time soon. This is a pity, given that the law justifiably prides itself on rationality, and these approaches offer rational solutions to the typical problems of validity and reliability of expert evidence which constantly trouble the legal profession.

It is a similar story in China. Likelihood ratio-based experimentation there has taken-off under the impetus from researchers like Zhang Cuiling and her associates, and as is clear from the examples used in this chapter, Dr. Zhang has been a prolific contributor to Chinese forensic voice comparison. Nevertheless, an outdated and discredited forensic speaker recognition method known as voiceprinting remains in use in real-world cases. Moreover, Chinese courts usually want to hear from their experts not just the non-logical probability of the hypothesis given the evidence, but a simple binary judgment of whether or not the suspect said the offending speech.

But, take consolation in the sheer interest of forensic voice comparison! Forensic voice comparison lies at the intersection of several disciplines, both scientific and humanistic, each one fascinating on their own: linguistics, phonetics, phonology, speech engineering, acoustic phonetics, probability. Voices, too, like language, are simply inherently interesting. Experimenting with them empirically, exploring the forensic potential of speech sounds, trying to improve the performance of a forensic voice comparison system: all are rewarding and absorbing undertakings of potential real-world value. I hope this chapter will encourage you to make a start.

Further reading
For further reading on the use of likelihood ratios in forensic voice comparison, the chapters in the legal reference series Expert Evidence by Rose (2003) and Morrison et al. (2018) are good places to start. Zhang & Rose (2008) is an introductory account in Chinese. Morrison et al. (2018) is also an excellent introduction to forensic voice comparison with automatic systems. These systems, which also use likelihood ratios, boast impressive, constantly improving performance and anyone interested in forensic voice comparison should be familiar with them. You can also read about automatic forensic speaker recognition in Gonzalez-Rodriguez et al. (2006). Finally, you can read about a successfully prosecuted case using likelihood ratios from speech sounds in Rose (2013a), and an account of how evidence in a simulated real-world case was evaluated with an automatic system in Enzinger & Morrison (2016).
4. REFERENCES


Brümmer, N., Focal Toolkit http://www.dsp.sun.ac.za/nbrummer/focal


