Rose. P. (2003). The Technical Comparison of Forensic Voice Samples. Expert Evidence 99, Selby H. & Freckelton I. (eds). Thompson Lawbook Co., Sydney.

Out of print. pdf available from http://expert-evidence.forensic-voice-comparison.net/

99
The technical comparison of forensic voice samples

[The next text page is 99 - 51]

Abbreviations

CC	cepstral coefficient
CS	cepstrally smoothed
FFT	fast Fourier transform
FSI	forensic speaker identification
IPA	International Phonetic Association/Alphabet
LP	linear prediction
LR	likelihood ratio
LTAS	long-term average spectrum
LTAF0	long-term average fundamental frequency
LTF0	long-term fundamental frequency
SLVT	supralaryngeal vocal tract
TFSI	technical forensic speaker identification

Glossary

(arithmetical) mean — the proper statistical term for average.

(vowel) backness — an important descriptive parameter for vowels. Refers to how far forward or back the body of the tongue is.

(vowel) height — an important descriptive parameter for vowels. Refers to how high or low the body of the tongue is.

allophone — a speech sound functioning as the realisation of a phoneme.

alveolar — type of consonant. The consonants at the beginning of *two, do, Sue, zoo, now, loo* are alveolar.

aural-spectrographic identification — highly controversial method of speaker identification using both visual examination of spectrograms and listening.

between-speaker variation — the fact that different speakers of the same language can differ in some aspects of their speech. One of the conditions that makes forensic speaker identification possible.

centisecond (or csec or cs) — unit for quantifying duration in acoustic phonetics: one-hundredth of a second.

cepstrum — a very common parameter used in automatic speaker recognition, one effect of which is to smooth the spectrum.

closed set comparison — an unusual situation in forensic speaker identification where it is known that the offender is present among the suspects.

convergence — the tendency for two participants in a conversation to become more similar in their speech behavior to signal in-group membership.

conversation analysis — the study of how conversation is structured and regulated.

decibel (or db) — unit for quantifying amplitude in acoustic phonetics.

defence fallacy — an error in logical reasoning which (1) assumes that the probability of the evidence given the hypothesis of innocence $p(E \mid H_i)$ is the same as the probability of the hypothesis of innocence given the evidence $p(H_i \mid E)$, and (2) ignores how probable the evidence is under assumption of guilt $p(E \mid H_g)$.

digitising — the process of converting an analog speech signal, eg from a cassette recorder, into a digital form that can be used by a computer for speech analysis.

diphthong — a vowel in a single syllable that involves a change in quality from one target to another.

expectation effect — a well-known perceptual phenomenon whereby one hears what, or who, one expects to hear.

F- (or formant) pattern — the ensemble of formant frequencies in a given sound or word.

false negative — in speaker recognition, deciding that two speech samples have come from different speakers when, in fact, they are from the same speaker.

false positive — in speaker recognition, deciding that two speech samples have come from the same speaker when, in fact, they are from different speakers.

fast Fourier transform — a common method of spectral analysis in acoustic phonetics.

formant — a very important acoustic parameter in forensic speaker identification. Formants reflect the size and shape of the speaker's vocal tract.

formant bandwidth — an acoustic parameter that reflects the degree to which acoustic energy is absorbed in the vocal tract during speech.

fricative — type of consonant. The consonants at the beginning of *fail, veil, thing, this, Sue, zoo, shoe* are fricatives.

fundamental frequency (or F0) — a very important acoustic parameter in forensic speaker identification. F0 is the acoustic correlate of the rate of vibration of the vocal cords.

hertz (or Hz) — unit for quantifying frequency: so many times per second: 100 Hz, for example, means a frequency of one hundred times per second.

intonation — the use of pitch to signal things like questions or statements, or the emotional attitude of the speaker.

kiloherz (or kHz) — unit for quantifying frequency: so many thousand times per second: 1 kHz, for example, means a frequency of one thousand times per second.

likelihood ratio (LR) — a number which quantifies the strength of the forensic evidence, and which is thus an absolutely crucial concept in forensic identification. The strength of the evidence is reflected in the magnitude of the LR. LR values greater than 1 give support to the prosecution hypothesis that a single speaker is involved; LRs less than 1 give support to the defense hypothesis that different speakers are involved.

The LR is the ratio of two probabilities. In forensic speaker identification these are: the probability of observing the differences between the offender and suspect speech samples assuming they have come from the same speaker; and the probability of observing the differences between the suspect and offender speech samples assuming they have come different speakers.

linear prediction — a commonly used method of digital speech analysis.

long-term — a common type of quantification in forensic speaker identification whereby a parameter, usually fundamental frequency, is measured over a long stretch of speech rather than a single speech sound or word.

manner (of articulation) — the type of obstruction in the vocal tract used in making a consonant.

millisecond (or msec or ms) — common unit for quantifying duration in acoustic phonetics: one-thousandth of a second.

monophthong — a vowel in a single syllable that does not change in quality.

naïve speaker recognition — when an untrained listener attempts to recognise a speaker, as in voice lineups etc.

open set comparison — the usual situation in forensic speaker identification where it is not known whether or not the offender is present among the suspects.

phonation type — the way the vocal cords vibrate.

phone — a technical name for speech sound.

phoneme — a unit of linguistic analysis: the name for a contrastive sound in a language.

phonemics — the study of how speech sounds function contrastively, to distinguish words in a given language. Phonemics is an important conceptual framework for the comparison of forensic speech samples.

phonetic quality — one of two very important descriptive components of a voice, the other being voice quality. Describes those aspects of a voice which have to do with the realisation of speech sounds.

phonetics — the study of all aspects of speech, but especially how speech sounds are made, their acoustic properties, and how the acoustic properties of speech sounds are perceived as speech by listeners. Phonetic expertise is an important prerequisite for forensic phonetics.

phonology — one of the main sub-areas in linguistics. Phonology studies the function and organisation of speech sounds, both within a particular language, and in languages in general.

pitch — (1) an important auditory property of speech. Pitch functions primarily to signal linguistic categories of intonation tone and stress, but overall pitch and pitch range also can be used to characterise an individual's voice; (2) another term for fundamental frequency.

pitch accent — the use of pitch to signal differences between words which is partly like tone and partly like stress. Japanese is a pitch accent language.

place (of articulation) — where in the vocal tract a consonantal sound is made.

posterior odds — in forensic speaker identification, the odds in favour of the hypothesis of common origin for two or more speech samples after the forensic phonetic evidence, in the form of the likelihood ratio, is taken into account. The posterior odds are the product of prior odds and LR.

prior odds — in forensic speaker identification, the odds in favour of the hypothesis of common origin for two or more speech samples before the forensic phonetic evidence is taken into account.

probability — a number between zero and one (or 0% and 100%) quantifying one of two things: (1) the degree of belief in a particular hypothesis, like these two speech samples have come from the same speaker; (2) the frequency of occurrence of an event, for example the number of times two samples from the same speaker have the same quality for a particular vowel. In forensic speaker identification, type (2) probabilities should be used to assess the probability of the evidence under competing prosecution and defence hypotheses. This then facilitates the evaluation, by the court, of the type (1) probability as the probability of the assumptions that the speech samples come from the same speaker.

It is logically and legally incorrect for the forensic phonetician to try to assess the type (1) probability of the hypothesis that two or more speech samples come from the same/different speakers.

prosecution fallacy — an error in logical reasoning which assumes that (1) the probability of the evidence given the hypothesis of guilt $p(E \mid H_g)$ is the same as the probability of the hypothesis of guilt given the evidence $p(H_g \mid E)$, and (2) ignores how probable the evidence is under assumption of innocence $p(E \mid H_i)$.

segmentals — a generic term for vowels and consonants.

sociolinguistics — the study of how language varies with sociological variables like age, sex, income, education etc.

source-filter theory — the received theory of how the radiated speech acoustics are related to the vocal tract that produced them. Also called "the acoustic theory of speech production".

spectrogram — a picture of the distribution of acoustic energy in speech. It normally shows how frequency varies with time. Spectrograms are often used to illustrate an acoustic feature or features of importance. To be distinguished from spectro*graph*, which is the name of the analog instrument on which spectrograms used to be made. Nowadays they are made by computer.

spectrum — the result of an acoustic analysis showing how much energy is present at what frequencies in a given amount of speech.

standard deviation — a statistical measure quantifying the spread of a variable around a mean value.

stress — prominence of one syllable in a word used to signal linguistic information, like the difference between "implant" (noun) and "implant" (verb) in English.

subglottal resonance — a frequency in speech attributable to structures below the vocal cords, eg the trachea.

technical speaker recognition — to recognise a speaker informed by theory, as in forensic speaker identification, automatic speaker recognition etc.

tone — the use of pitch to signal different words, as in tone languages like Chinese.

variance — a statistical measure quantifying the variability of a variable; the square of the standard deviation.

voice quality — one of two very important descriptive components of a voice, the other being phonetic quality. Describes those long- or short-term aspects of a voice which do not have to do with the realisation of speech sounds.

voiceprint — another name for spectrogram. Usually avoided because of its association with voiceprint identification.

voiceprint identification — highly controversial method of speaker identification exclusively using visual examination of spectrograms.

voicing/phonation — refers to activity of the vocal cords.

within-speaker variation — the fact that the same speaker can differ in some aspects of her or his speech on different occasions, or under different conditions. One of the conditions that makes forensic speaker identification difficult.

[The next text page is 99 - 1051]

99
The technical comparison of forensic voice samples

by

Philip J Rose

PhD, MA, BA Hons (1st Class), Dip IPA

[The author would like to express and record his thanks to the following people who have helped in the production of this Chapter. First, two individuals must be mentioned whose positions are now defunct because of the abysmal higher education funding support in Australia. Keith Carswell (whose irreplacability as phonetics laboratory technician continues to be confirmed by the fact that, dispensed with by the Australian National University in 1998, he has still not yet been replaced) contributed his telecommunications expertise in painstakingly setting up and running the experiments on the effects of telephone transmission. Professor Pamela Davies, former head of Australia's now disbanded National Voice Centre, made available and ran some of the endoscopy equipment used to produce the figures of the author's vocal cords. Bob Boag, co-ordinator of the Language Laboratory at the Australian National University, also generously gave his time to help in the production of the vocal cord pictures. Hugh Selby, reader in the Faculty of Law at the Australian National University, provided very helpful critique on everything the author wrote, and Susan Tarua was a most patient and professional editor. The author continues to benefit from discussions on automatic feature extraction with Dr Franz Clermont, of the Department of Computer Science at the University of New South Wales. Needless to say, the views expressed in this Chapter are not necessarily those of the above-mentioned. He thanks Taylor & Francis for permission to reproduce, as Figure 2, Figure 6.1 from his book *Forensic Speaker Identification*. Some of the data used in this Chapter is from forensic-phonetic research into telephone transmission effects funded by a small grant from the Australian National University.]

Author information

Phil Rose is associate professor in phonetics and Chinese linguistics at the Australian National University. He holds a doctorate from the University of Cambridge in Chinese phonetics, as well as degrees in linguistics and German and Russian. He is author of the major reference work *Forensic Speaker Identification* (Taylor & Francis, 2002), and has also published widely on forensic speaker identification and the phonetics of tone languages, in which he is also an acknowledged expert. He is a member of the International Association for Forensic Phonetics, and a past member of the Forensic Standards Committee of the Australian Speech Science and Technology Association and former Member of Council of the International Phonetic Association. He has done research for almost 30 years on similarities and differences between individuals in their speech. He has been undertaking forensic speaker identification case work in Chinese and Australian English for over a decade.

INTRODUCTION

[99.10] Voices are never far from the news, especially when it comes to the identification of their owners. Was that Bin Laden's voice in 2002 which promised more retribution? Usually, of course, what is important is not primarily determining the owner of the voice but the implications of the identification. If it really was Bin Laden, then perhaps he is still alive and poses an enormous threat. And, of course, governments around the world will be far more inclined to vote more money to national defence and "the war against terrorism" if the voice were found to be genuine.

The ability to identify a speaker by her or his voice clearly has significant forensic implications. The forensic applications of speaker recognition, one common term for which is "forensic speaker identification" (FSI), have existed for a long time, and are becoming increasingly common. There are two main types of FSI, depending on whether it is done by naive subjects – sometimes called "earwitness identification" – or by experts. This Chapter deals only with the latter kind, which can be referred to as "technical forensic speaker identification" (TFSI). Naive forensic speaker identification is dealt with at length in Rose (2002, Ch 5) and in Hollien (2002, Ch 3).

Probably the most common task in technical forensic speaker identification involves the comparison, by an expert, of one or more samples of an unknown voice (sometimes called the questioned sample(s)) with one or more samples of a known voice. Often the unknown voice is that of the individual alleged to have committed an offence (hereafter called the offender) and the known voice belongs to the suspect. Both prosecution and defence are then concerned with being able to say whether the two samples have come from the same person, and thus be able either to identify the suspect as the offender, or eliminate them from suspicion. Sometimes it is important to be able to attach a voice to an individual, or not, irrespective of questions of guilt.

This Chapter provides an overview of essential aspects of the expert comparison of forensic voice samples in scenarios of the above type.[1] Broeders (2001) observed that "there are probably few forensic disciplines that are characterised by such a diversity of methods and procedures as the field of forensic speaker identification by experts". Such a diversity, assuming that it is great – its actual magnitude depends, of course, on the magnitude of the diversity in the other forensic disciplines – obviously presents a problem for a Chapter of this nature. This Chapter therefore concentrates on issues that are warranted by first, the nature of voices and second, the logically correct evaluation of forensic identification evidence. The assumption is that such considerations should be a part of any expert's attempt to recognise voices under real-world forensic conditions. The remainder of this Chapter is accordingly divided into the following four sections, each dealing with an important area of TFSI:

(1) It is a common, but completely erroneous, assumption on the part of most professionals involved with FSI that the aim is to say how probable it is that the speech samples were from the same speaker. This is clearly an important misunderstanding, since it relates to the aim of FSI, and a complete section is devoted to it: see **[99.50]** ff. It will be clear from this section how it is, in fact, logically not possible to say how probable it was that it was Bin Laden's voice from the voice evidence alone.

(2) The second section (see **[99.220]** ff) describes some basic facts about voices and the technical comparison of voice samples that these facts require. This attempt to set the record straight appears particularly warranted by the extremely poor understanding of what is involved in TFSI recently demonstrated both in the media's reaction to the Bin Laden voice affair, and by certain members of the legal profession.

(3) The third section (see **[99.350]** ff) describes the different types of parameters and analyses used for comparing speech samples in TFSI and includes an illustration of the proper use of spectrograms.

(4) The fourth section (see **[99.870]** ff) addresses the effect of telephone transmission.

Some examples are given in the text of TFSI comparisons using different features. Bibliographical references are at the end of the Chapter.

Case history leaves no doubt that properly conducted TFSI can be judicially and investigatorily very effective, contributing to both conviction and elimination of suspects. The reader will be able to appreciate from this Chapter, however, that TFSI is at the same time an extremely complex task. The complexity derives both from the complex nature of voices and how they relate to their owners, and the methods and disciplines that inform the analysis. This complexity is such that it cannot possibly be exhaustively covered in a short Chapter like this. For a comprehensive treatment, not only of TFSI but also of naive forensic speaker recognition, the reader is referred to Rose (2002), for which the present Chapter provides suitable preliminary reading.

In addition to the complexity of the task, the reader will also be able to appreciate from this Chapter the following main points about technical forensic speaker identification:

- it is informed by received scientific principles and theories;
- it is multidisciplinary, requiring expert knowledge of several specialist disciplines relating to speech science, including linguistics, phonetics, acoustics, signal processing and statistics;
- it is completely outside lay competence;
- it requires both acoustic and auditory comparison of both the linguistic and non-linguistic aspects of speech samples;
- one of the main problems in evaluating differences between speech samples is ubiquitous differential variation in voices, both from the same speaker and from different speakers; and
- the logically correct way of evaluating the strength of forensic-phonetic evidence is to assess, not the probability of the hypothesis that the speech samples have come from the same voice, but the probability of observing the differences between the speech samples under competing prosecution and defence hypotheses.

1 The author has been asked by the editors to keep referencing and attribution of sources to a minimum, and in particular not to give dates and pages. He thought it important to nevertheless attribute important quotes, however. Readers are referred to Rose (2002) for full references.

[The next text page is 99 - 2051]

EXPRESSING THE OUTCOME: THE BAYESIAN APPROACH TO QUANTIFYING THE STRENGTH OF EVIDENCE

Introduction

[99.50] It is a fairly safe bet that the kind of statement that police, counsel and judiciary will expect from the forensic phonetician will be something like: "Given the high degree of similarity between the two speech samples, there can be very little doubt that these two samples are from the same speaker", or "It is highly likely, given the extensive differences between the speech samples, that they are from different people". However, over the last few years much theoretical work has been carried out, and published, on the appropriate formulation of conclusions in forensic identification, and as a result it is now realised that there are serious problems with this way of expressing the outcome of a forensic investigation. This section outlines why this is so, by describing the logically correct way of evaluating the strength of forensic identification evidence.

Evaluating the strength of evidence – Bayes' Theorem

[99.60] The statements above involve quoting the probability (there is very little doubt/ it is highly likely) of a hypothesis (the samples are from the same speaker/ are from different speakers) given the evidence (the high degree of similarity/ the extensive differences). This can be expressed more conveniently, and conventionally, with Formula 1.

Formula 1

$$p(\mathrm{H} \mid \mathrm{E})$$

In Formula 1, *p* stands for probability, "H" for hypothesis, "|" for given, and "E" for evidence. So the formula can be read as "the probability that H is true given the evidence E". The amount of support for the hypothesis that the same speaker is involved – $p(\mathrm{H} \mid \mathrm{E})$ – is clearly what the court wants to know. For several reasons it is more convenient to express this probability in terms of odds in favour of the hypothesis, and this formulation is shown in Formula 2. Odds are simply a way of comparing the probability that something is true, or will happen, with the probability that it is not true, or won't happen.

Formula 2

$$\frac{p(\mathrm{H} \mid \mathrm{E})}{p(\sim\mathrm{H} \mid \mathrm{E})}$$

Formula 2 shows the odds in favour of the hypothesis H – perhaps that the accused is guilty. This is simply a ratio of two probabilities. The probability on top is of the hypothesis (eg "the accused

is guilty") given the evidence; the probability on the bottom is of the negation of the hypothesis (~H) (eg "the accused is not guilty") given the evidence. If from the evidence it is, say, 90% likely that the accused is guilty (ie $p(H \mid E) = 90\%$), then the odds in favour of guilt given the evidence are: $p(H \mid E) / p(\sim H \mid E) = 90\% / 10\% = 9$ to 1. If it is 10% likely that the two speech samples were spoken by the same person, then the odds in favour of common origin are: $p(H \mid E) / p(\sim H \mid E) = 10\% / 90\% = 1$ to 9, or 9 to 1 against. Obviously, the greater the odds in favour of the hypothesis, the stronger the prosecution case, and the greater the odds against the hypothesis, the stronger the defence case.

It has long been established that the logically correct way of evaluating the strength of evidence in favour of a hypothesis is by using Bayes' Theorem, and this is directly applicable to the evaluation of the strength of evidence in favour of a legal hypothesis like "these two speech samples were spoken by the same speaker". Bayes' Theorem makes explicit how one's belief in a hypothesis can be updated when new evidence is adduced. Thus it is directly applicable to the adduction of scientific forensic evidence. It can be appreciated that Bayes' Theorem is an extremely important theorem in science. The so-called odds form of Bayes' Theorem is shown in words at Formula 3.

Formula 3

*Posterior Odds = Prior Odds * Likelihood Ratio*

The posterior odds, to the left of the equals sign in Formula 3, represent the odds in favour of a hypothesis in the light of new evidence. To the right of the equals sign are two terms: the "prior odds" and the "likelihood ratio". The prior odds are the odds in favour of a hypothesis *before* the evidence is taken into account, and the likelihood ratio is *a measure of the strength of the evidence*. The expression at Formula 3 says, therefore, that the odds in favour of a hypothesis *after* a piece of evidence has been taken into account – the posterior odds – are quite simply the product of the prior odds and the likelihood ratio for that evidence. This formulation now allows us to be specific about the role of the forensic phonetician: it is her or his job to calculate the likelihood ratio, and thus estimate the strength of the forensic-phonetic evidence.

Before the likelihood ratio is discussed in somewhat greater detail, here is a hypothetical forensic-phonetic example to illustrate how odds in favour of a hypothesis can be updated.

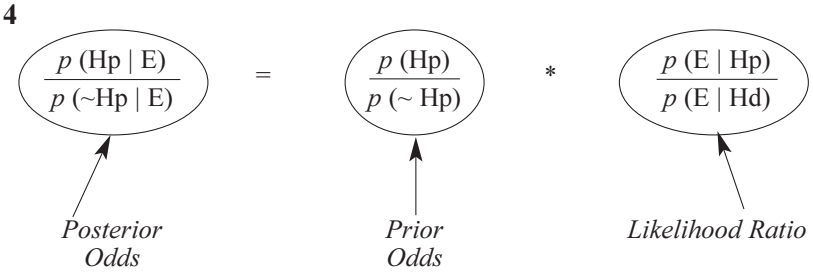
EXAMPLE 1

It is known that there are five men in a house, including the suspect. The police intercept an incriminating telephone call from the house, and want to know whether the suspect made the call. Before the forensic phonetician compares the incriminating call with known exemplars of the suspect taken from earlier calls, the odds in favour of the suspect making the call are 4-to-1 against, since there are five in the house including him. On the basis of the forensic-phonetic comparison, the forensic phonetician estimates a value for the likelihood ratio of 100. (This means that the differences between the speech samples are 100 times more likely if they have come from the same speaker than from different speakers.) Now the odds in favour of the suspect making the call can be updated by incorporating the forensic-phonetic evidence. The posterior odds in favour of the suspect making the call are: prior odds * likelihood ratio = 1/4 * 100 = 25, and move from 4-to-1 against to 25-to-1 in favour of him making the call.

The likelihood ratio

[99.70] Bayes' Theorem is shown once again at Formula 4, this time with the terms explicit, but also with them tailored to their application in forensic speaker identification. We are now dealing with the explicit prosecution hypothesis that two or more speech samples were spoken by the same speaker. Let us call this Hp (for prosecution hypothesis). The posterior odds, in favour of the prosecution hypothesis Hp to the left of the equals sign, can be seen to be the expression, already introduced, of the probability of the hypothesis given the evidence $p(\text{Hp} \mid \text{E})$ divided by the probability of the negation of the hypothesis given the evidence $p(\sim\text{Hp} \mid \text{E})$. The prior odds – the first term to the right of the equals sign – are simply the probability of the prosecution hypothesis being true $p(\text{Hp})$ to its not being true $p(\sim\text{Hp})$.

Formula 4



The likelihood ratio (LR) – the last term in Formula 4 – is the important one, for it is the likelihood ratio that the forensic-phonetic expert must provide the court with. It can be seen that the likelihood ratio is a ratio of two probabilities. The one on top refers to what the probability is of getting the evidence assuming that the prosecution hypothesis Hp is true. The one on the bottom refers to how likely you are to get the evidence assuming that the defence hypothesis – represented as Hd – is true. If you are more likely to get the evidence assuming that the prosecution hypothesis is true than if the defence hypothesis is true, then the likelihood ratio will be greater than 1 and this can be interpreted as indicating support for the prosecution hypothesis. If, on the other hand, you are more likely to get the evidence assuming the defence hypothesis is true, the likelihood ratio will be less than 1, indicating support for the defence. The more the likelihood ratio approaches 1, the more the evidence is likely under both prosecution and defence hypotheses, and thus the more useless. The more the likelihood ratio deviates from 1, the greater support for either prosecution (if it is greater than 1), or defence (if it is smaller than 1). In other words, the relative strength of the evidence is reflected in the magnitude of the likelihood ratio.

Similarity and typicality

[99.80] In order to better understand how the likelihood ratio works, it often helps to think of its numerator (the expression on the top) as a quantification of the degree of *similarity* between the offender and suspect samples, and its denominator (the expression on the bottom) as a quantification of the degree of *typicality* of the offender and suspect samples in the relevant population. Then the more similar the two samples are, the more likely they are to have come from the same speaker and the higher the ratio. But this must be balanced by their typicality: the more typical the samples, the more likely they are to have been taken at random from the population, and the lower the ratio. The value of the ratio can thus be seen to be an interplay between the two factors of similarity and typicality. Both factors are needed to evaluate forensic-phonetic evidence: it is a very common fallacy to assume that similarity is enough: that if two speech samples are similar, that indicates common origin. But the likelihood ratio of Bayes' Theorem makes it clear that similarity is only half the story: typicality is equally important.

As made clear in Rose (2002), in TFSI a completely accurate assessment of similarity and typicality (and hence an accurate LR) will usually not be possible because of currently inadequate knowledge of between- and within-speaker variation in the features used to compare speech samples. This will not preclude estimates using the information available, however. Naturally, the expert should always make it clear what limitations are involved by so doing, perhaps by saying that the LR for the evidence lies within a certain range.

Probabilities of evidence

[99.90] By far the most important point that can be made with respect to the likelihood ratio is this. *The likelihood ratio has to do with the probabilities of evidence, not hypotheses*. It can be seen that the expressions on both top and bottom are of the form $p(E \mid H)$, that is, they have to do with the probability of the evidence assuming a hypothesis. Thus the likelihood ratio has to do with the probabilities of the evidence assuming hypotheses, not the probabilities of a hypothesis assuming the evidence.

What is meant by the "probability of evidence", and why is the likelihood ratio couched in these terms? Here is another hypothetical example for explanation.

EXAMPLE 2

Assume that both suspect and offender forensic speech samples agree 100% in a particular feature (in order to make the example easy to understand, let us assume it is an easily recognised pathological feature like a "lisped s").

In order to determine how strongly the evidence of the "lisped s" in both speech samples supports the prosecution hypothesis that both offender and suspect are the same speaker, you need to know two things. The first is the probability of both speech samples having a "lisped s" (this is the evidence) *assuming that they come from the same speaker* (the prosecution hypothesis). If this is a true pathology, the same speaker will lisp in all speech, so there will be a 100% probability of both samples having a lisp if they are from the same speaker. The value for the numerator of the likelihood ratio – the probability of the evidence assuming the prosecution hypothesis, or $p(E \mid Hp)$ – will then be 1. But, of course, there is likely to be more than one speaker with a pathological lisp, so, second, you need to know the probability of observing the lisp in both samples (the evidence) assuming that different speakers are involved (the defence hypothesis).

Let us suppose one in 1000 speakers in the relevant speech community has such a lisp. The probability of observing the evidence – the agreement in the lisp – assuming that different speakers are involved is 1/1000 = 0.001. This is thus the value for the denominator of the likelihood ratio – the *probability of the evidence* assuming a defence hypothesis that the offender is simply someone else with a lisped s other than the suspect, or $p(E \mid Hd)$.

The strength of the evidence is now estimated as the ratio of the two probabilities: 1/ 0.001 = 1000. A likelihood ratio of 1000 means that you are 1000 times more likely to observe the evidence – the lisp in both samples – if the samples had come from the same person than if they had come from different people.

Two important points now need to be made with respect to the expression of the likelihood ratio outcome. First, note that the expression of the LR outcome is still couched in terms of probabilities of evidence under competing hypotheses, not in terms of probabilities of hypothesis, given evidence. Bayes' Theorem makes it perfectly explicit that it is logically not possible to quote the posterior odds in favour of a hypothesis (like "these two speech samples come from the same speaker") unless you also have access to the prior odds in favour of the hypothesis. Since the forensic phonetician seldom has this knowledge, he or she cannot logically quote a probability of the same speaker hypothesis or $p(H_{ss})$ given the forensic-phonetic evidence ($E_{f\text{-}p}$). This is the reason for the admonishment at the beginning of this section that it is logically incorrect for the expert to quote the probability of a hypothesis – like it is 90% likely that the same speaker is involved – from the evidence.

If all the prior odds *are* known, then the probability of the hypothesis *can* be estimated using Bayes' Theorem, and converting odds to probability. Using the hypothetical case of the men in the house (see Example 1), posterior odds of 25-to-1 in favour of the hypothesis translate into a probability of (25 / 25 + 1 =) ca 96% that it is the same speaker.

Second, note that, as is often pointed out, it is logically incorrect to assume that, because you are 1000 times more likely to observe the evidence assuming that the same speaker is involved, this means that it is 1000 times more likely that the same speaker is involved. This conceptual leap from $p(E \mid H)$ to $p(H \mid E)$ is part of what is known as the prosecutor's fallacy, or transposing the conditional. The two things are not interchangeable. To understand this, consider the probability of an animal having four legs if it is a cow. Since, barring road accidents and weird genetic experiments, a cow is almost certain to have four legs, we can estimate p (quadruped | cow) as 1. Now transpose the conditional – swap the cow and the quadruped – and evaluate p (cow | quadruped). Since there are many other animals – not to mention other objects – that have four legs, this will certainly not be 1, and, in fact, will be very much less than 1.

Verbal equivalents for the likelihood ratio

[99.100] Since the import of a particular value for the likelihood ratio is often difficult for interested parties to construe, some kind of verbal translation is often advocated. For example, a likelihood ratio between 100 and 1000 might be translated as offering "moderately strong evidence to support the prosecution hypothesis", or a value less than 0.0001 might be translated as offering "very strong support for the defence hypothesis". (Values for the likelihood ratio less than 1 are probably best thought of in terms of their reciprocal. Thus a value of 0.0001 can be thought of as 1/0.0001 = 10,000 more likely under the defence hypothesis than the prosecution hypothesis.) However, this usage ultimately founders on circularity, since it can always be asked what precisely is meant by "strong", to which the answer is: something corresponding to a LR between x and y.

Combining evidence

[99.110] One of the beauties of the likelihood ratio – not possible under other approaches – is that it allows the evaluation of the combined strength of separate pieces of evidence. This is an extremely useful property in forensic-phonetic comparisons, since voices can and must be compared forensically with respect to many different features. Thus it is possible to calculate an overall likelihood ratio for a forensic-phonetic comparison by combining the likelihood ratios from many different forensic-phonetic features. As long as the features are not correlated, their likelihood ratios can be combined by simply taking their product. We might find when comparing the two hypothetical speech samples with a pathological lisp, eg, that they also had very similar acoustic values for a particular vowel. A likelihood ratio – say of 200 – might be

derived for this acoustic feature, and then the overall likelihood ratio for the comparison would become 200 * 1000 = 200,000. One would now be 200,000 times more likely to observe the evidence – the similarities between the offender and suspect samples – if the samples had come from the same speaker than from different speakers.

It can be appreciated that, if likelihood ratios are derived for several independent forensic-phonetic features, the overall likelihood ratio resulting from their product stands the chance of being either much greater or much smaller than 1, and thus giving substantial support either to defence or prosecution. It is also known that it is necessary to use many features in a forensic-phonetic comparison, since it is perfectly possible for one or two to give likelihood ratios with conflicting values (ie values slightly greater or lesser than 1). The more features used, the greater the chance that the overall likelihood ratio will be forced higher or lower than 1, and thus constitute probative evidence.

Alternative hypothesis

[99.120] Note that the denominator of the likelihood ratio in Formula 4 (see **[99.70]**) is expressed, not negatively, as the negation of the prosecution hypothesis that the speech samples come from the same speaker (~Hp), but positively, as the probability of the defence hypothesis (Hd). This is because the defence hypothesis is not necessarily simply the negation of the prosecution hypothesis, but a particular alternative like "It is not my client but someone else who sounds like him". The probability of the evidence will differ, depending on the nature of the defence hypothesis. For example, a high degree of similarity between offender and suspect samples might be more likely if the alternative hypothesis was that both samples were from similar sounding, but different, speakers, rather than simply from different speakers. The important point here is that the likelihood ratio will change depending on the particular alternative hypothesis chosen by the defence.

Prior odds

[99.130] Despite the importance of the prior odds, the forensic phonetician is usually not provided with all the necessary information to estimate them for the voice evidence. Some information may be available, eg whether the incriminating phone-calls were made from the accused's house or mobile, or whether the accused has been identified by a police officer familiar with the accused's voice. However, there are also good reasons why the phonetician should insist on not being told such data and should concentrate on assessing the likelihood ratio for the voice evidence *on its own*. These reasons are presented in Rose (2002). Obviously, it will be part of the responsibility of legal counsel to ensure that the prior odds are taken into consideration at some point in the case.

Probabilistic outcome

[99.140] As pointed out above, if all the prior odds are known, a probability for the hypothesis that the same speaker is involved is possible. Even the LR, which itself is not a probability but a ratio, depends on probability estimates. Probability is a measure of uncertainty; thus it should be clear that the logically correct approach to TFSI is ultimately probabilistic and *will not yield absolute exclusion or identification of the suspect*.

Additional conditions on probability

[99.150] Strictly speaking, in addition to the actual data – in TFSI this is the observed differences between the offender and suspect speech samples, and has been referred to in this Chapter as "the evidence" – the probability of a hypothesis is conditional upon two other things. The first of these is the set of assumptions concerning statistical aspects of the data, such as how good the chosen statistical model is (eg whether the data can be considered to be normally distributed). The second is the relevant background knowledge available. Thus it might be known that the suspect is a monolingual speaker; or that he or she had lived a long time in two different places. This can be expressed in the more complete formula for the probability of the hypothesis at Formula 5, where E = evidence (ie observed differences between suspect and offender speech samples), A = assumptions, and K = relevant background knowledge.

Formula 5

$$p(\mathrm{H} \mid \mathrm{E}, \mathrm{A}, \mathrm{K})$$

It can be appreciated that the background knowledge factor can often be very important in evaluating the probability of the forensic-phonetic evidence.

Testing the LR approach

[99.160] Different jurisdictions will, of course, differ in their standards for the admissibility of evidence purporting to be scientific. However, in the light of the ruling in *Daubert v Merrell Dow Pharmaceuticals* 43 F 3d 1311; 125 L Ed (2d) 469; 509 US 579; 113 S Ct 2786 (1993), which makes criterial the fact that an approach can be, and has been, tested, it is useful to be aware of the degree to which a LR approach has been tested, both generally and on speech.

As shown above, Bayes' law predicts that same-subject data should be resolved with LRs greater than 1, and different-subject data should have LRs smaller than 1. This therefore provides a test for the method. The extent that known different-subject data are resolved with LRs smaller than 1, and known same-subject data are resolved with LRs larger than 1, reflects how well the method works. That this is, indeed, the case has already been demonstrated with three types of forensically common evidence: DNA, glass fragments and speech. Details and references are found in Rose (2002, Chs 4 and 11).

The fact that the Bayesian approach works can thus be construed both from a theoretical and a practical point of view. First, the theory can be said to work because its theoretical predictions have been empirically confirmed in several forensically important areas. Second, the theory can be said to work because it can be practically used to discriminate same-subject from different-subject data.

It is important to understand the fact that these experiments tap average behaviour, and do not say anything about the *accuracy* of a particular likelihood ratio in a particular case. This point is discussed at length in Rose (2002, Ch 11).

Summary

[99.170] The aim of a forensic-phonetic identification should be to express its outcome in terms of a Bayesian likelihood ratio. This is the logically correct way of quantifying the relative strength of forensic identification evidence, and should constitute the conceptual framework for all forensic-phonetic comparisons. This involves stating the probability of the evidence under competing prosecution and defence hypotheses. Conclusions in the form of $p(\mathrm{E} \mid \mathrm{H})$ probability

of hypothesis given evidence are logically flawed. One of the important things to check in the evaluation of forensic-phonetic speaker identification evidence must therefore be whether or not the conclusion has been stated in the appropriate way. The LR should state how much more likely it is to observe the differences or similarities between questioned and suspect voice samples assuming that they have come from the same speaker than if they have come from different speakers. Any limitations should be made explicit. The LR must be combined, but not necessarily by the forensic-phonetic expert, with the appropriate prior odds to estimate the posterior odds in favour of common provenance of offender and suspect speech samples. The nature of the defence, or alternative, hypothesis can affect the LR and the prior odds must be considered with care.

Further reading

[99.180] Detailed and very clear explanations of the application of Bayes' Theorem and the likelihood ratio to the evaluation of evidence in general can be found in Robertson and Vignaux (1995) and in their Chapter 28, "Interpreting Scientific Evidence", in this service. Rose (2002) contains chapters on both the specific application of Bayes' Theorem in forensic speaker identification and a demonstration of the method. Highly recommended shorter discussions can also be found in Hodgson (2002), which contains some demonstrations of Bayes' law as well as an assessment of its general use from the point of view of an experienced judge; Evett (1991); Broeders (1999); Champod and Evett (2000); and Champod and Meuwly (2000).

[The next text page is 99 - 3051]

SOME BASIC FACTS ABOUT VOICES AND THE FORENSIC COMPARISON OF VOICE SAMPLES

[99.220] This section introduces the most important general things to know about voices, and how voice samples should be compared forensically.

Forensic-phonetic features

[99.230] The comparison of objects is usually done with respect to certain characteristics. Two apples can be compared with respect to colour, or weight, or shape, for example. It is the same with voices. That is, one speaker's voice is understood to differ from another's not in unanalysable global terms, but in certain forensic-phonetic "features" (also called "dimensions" or "parameters"). Thus the voices of two speakers may be very similar in one particular feature or set of features like overall pitch or vowel quality, but differ in another feature like the way they say their *r* sound. Voices are compared forensically with respect to these features.

Between- and within-speaker variation

[99.240] Under ideal conditions – when enough features can be compared under well-controlled circumstances, for example – speakers can be identified reasonably easily by their voices. This probably entitles us to assume that different speakers of the same language or dialect do, indeed, have different voices, although this remains, of course, an inductive truth. We thus have to deal with variation between speakers, usually known as between-speaker (or inter-speaker) variation. As said, this variation is assumed to be in the individual voice features being compared.

Although it is a general assumption that different speakers have different voices, it is a truism of phonetics that no-one ever says the same thing in exactly the same way. This means that the voice of the same speaker will always vary as well. This is called within-speaker (or intra-speaker) variation.

Within-speaker variation occurs as the result of many different factors as well as the fact that we can never repeat exactly what we said. The most important of these factors are as follows:

(1) Non-contemporaneity: the same speaker will show greater variation in the features of her or his speech when speaking on different occasions, and the variation will often be greater the bigger the difference in time between two occasions. Thus two samples from the same speaker separated by a couple of minutes – say in consecutive phone-calls – can be expected to differ less that two same-speaker samples separated by a day, a week or a month.

(2) Emotional state and health: eg, a speaker might have different overall pitch depending on how tired the speaker is or whether he or she has a cold.

(3) Interlocutor: speakers are known to converge or diverge on features of the speech of their interlocutors, depending on how much they want to establish rapport with them.

(4) Perceived formality: it is an assumed linguistic universal that every speaker commands a range of speaking styles, depending on how formal he or she perceives the situation to be.

(5) Phonological environment: speech sounds affect each other as a matter of course. For example, the *oh* vowel in the word "road" will have different acoustic characteristics from the *oh* vowel in the word "node", because of the differential effect of the initial consonants *r* and *n*. Because of these factors it stands to reason that one of the most important considerations in forensic speech comparison is that speech samples be assessed for comparability. It would be highly questionable to compare the recorded speech of a healthy offender yelling at a bank teller during a hold-up with that of a mucosal suspect being interrogated by the police, for example.

Probably the most important consequence of within- and between-speaker variation is that there will always be differences between speech samples, even if they come from the same speaker. These differences will usually be audible, and always measurable and quantifiable. For example, a speaker might one day say the word "okay" 20 times in a conversation, and leave off the first syllable – saying "'kay" – 15 times. This would give an incidence of "'kay" of 15/20 = 75%. In another conversation the same day, he might say the word "okay" 15 times and omit the first syllable in 12 of these: an incidence of 12/15 = 80%. These two samples from this speaker would then differ in the incidence of "'kay". The feature in this case could be called "okay monosyllabicity", or some such mouthful, and its incidence quantified for different speech samples.

Technical Forensic Speaker Identification therefore involves evaluating the inevitable differences between the features in speech samples – ultimately helping the court to decide whether they are more likely to be same-speaker differences or different-speaker differences. It is the observed differences, both big and small, between the speech samples which constitute the forensic-phonetic evidence, the strength of which has to be evaluated. Of course, since in the majority of cases the same language is being spoken in both samples, many of the differences between the samples will be small, and in this case they will probably be called similarities. And if the samples come from the same speaker, it is a reasonable assumption that the differences will be small (ie they will, rather, count as similarities due to that factor too).

To illustrate further with the "'kay" example, and introduce the Bayesian approach discussed at **[99.60]**, suppose both offender and suspect speech samples are observed to contain similar incidences of "okay" said without the first syllable: perhaps the offender sample has a 75% incidence of "'kay" and the suspect sample has an 80% incidence. The difference in incidence is the evidence. Is this more likely to be a same-speaker difference or a different-speaker difference?

In order to evaluate correctly the strength of this evidence, as already explained, it is necessary to determine the probability of observing it assuming that it is a same-speaker difference, and the probability of observing it assuming it is a different-speaker difference.

Now, if it is known from the examination of very many same-speaker and different-speaker speech samples that such a difference is typical of about 10% of same-speaker speech samples, but only found in 1% of different-speaker speech samples, that means that you are (10% / 1% =) 10 times more likely to observe the agreement in "'kay" assuming the two samples came from the same speaker than if they had come from a different speaker. As explained above, this ratio

of the probabilities of the evidence under competing hypotheses as to whether the same speaker is involved or not is called a "likelihood ratio" and is the most important construct in TFSI. A value of 10 for the LR might be interpreted as providing *very limited support* for the prosecution hypothesis that the samples did, in fact, come from the same speaker.

Voice multidimensionality

[99.250] In the latter half of the 19th century in France a method of identification – eponymously called "Bertillonage" – was developed. One of its aims was to be able to identify re-offending criminals. Imagine that you wanted to determine whether someone standing in front of you was the same individual offender as described in previous prison records. Bertillonnage, properly called anthropometry, involved measuring features of the available individual's body and comparing them with existing records of the same features from allegedly the same individual. One important aspect of this approach was that the comparison was multidimensional: it was carried out with respect to very many different measurable aspects of the bodies in question. For example, initially comparisons were made of height, head length and breadth, arm span, sitting height, left middle and little finger length etc.

This is a sensible approach, given the ways in which different individuals can vary in their physical proportions. We know that it is possible for many humans to have the same height, so the chances will be quite high of picking two humans at random from a relevant population and finding that they have the same height (although the chances will differ depending on how *typical* the measurements are). Thus finding agreement in height alone would not be particular cause for thinking that the individual in front of you was the same as that described in some existing record. But what if there were agreement not only in height, but also in angle at which the ears stick out, eye-colour and missing teeth? Intuitively, the chances of finding two humans at random from the relevant population with agreement in these four features would be less. In Bayesian terms, the probability of observing the agreement in all four features together would be greater assuming that we were dealing with the same individual than were we dealing with different individuals, and we would be more inclined under these circumstances to the belief that the same individual was involved than if comparison were made on one feature alone.

In the same way as you can compare humans with respect to more than one physical feature, so it is possible, indeed mandatory, to compare voice samples with respect to many different features. It is unrealistic to envisage a forensic-phonetic comparison on a single feature only, like the omission of the first syllable in "okay". Fortunately, voices are multidimensional objects. A speaker's voice is potentially characterisable in terms of an exceedingly large number of different features – far more, in fact, than there is time to quantify them. The section below will illustrate some multidimensional comparisons, first with just two features, then with more, again using the likelihood ratio.

First we need some more features. Figure 1 is an example, from case work, of another forensic-phonetic feature, this time using an easily measurable acoustic property in the second syllable of the word "okay".

FIGURE 1 Offender and suspect speech samples compared with respect to a single acoustic feature in the word "okay"

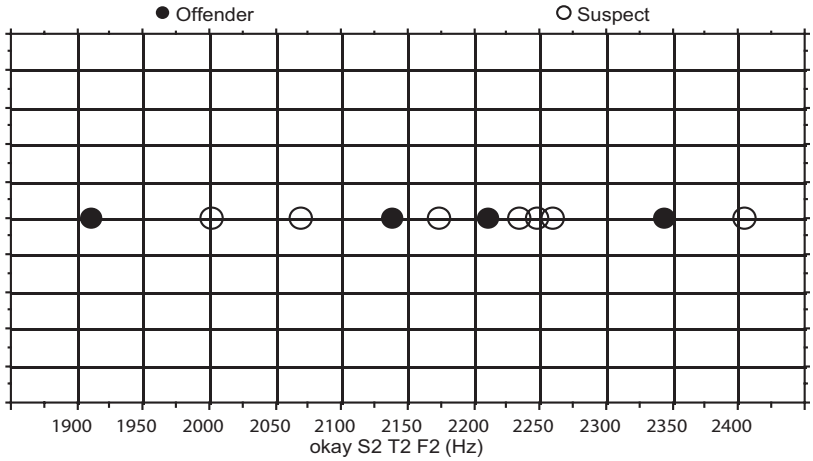


Figure 1 shows two speech samples compared with respect to this feature, called "okay S2 T2 F2" (an abbreviation for the second formant ["F2"] at the second diphthongal target ["T2"] in the second syllable ["S2"] of "okay"). (The reader need not worry about what a formant is at the moment, or a diphthongal target: they are explained below at **[99.600]**; **[99.810]**.) The feature involves frequency and is quantified in the appropriate units of Hertz (Hz). Figure 1 shows several values for this feature in both the suspect's and the offender's voice. Each dot in the offender's sample represents a single observation – one measurement from a different "okay" – in a single conversation. So the filled-in dot at the left represents a measurement of just over 1900 Hz from one of the offender's "okays".

One slight complication, but an essential and typical one, is that, whereas the filled-in offender's dots each represent a value from separate "okays" in the same conversation, the empty suspect dots do not represent values from single "okays" but mean values, each from a separate conversation in which the suspect was known to have participated. Thus each empty suspect dot represents the average of several observations from different conversations. The highest unfilled dot at just over 2400 Hz, eg, represents a mean of three measurements from the suspect's voice in a single conversation. It can be seen that there were four "okays" in the offender's speech, and means from "okays" in seven conversations from the suspect. What is being tested here, then, is the amount of strength for the hypothesis that the "okays" in the voice in the single offender sample came from the suspect. (One can usually only compare one sample of the offender's speech with known samples from the suspect, because it is not usually known whether the speaker in two offender samples is the same. It is therefore important to guard against unlawful pooling of offender/unknown samples. This topic is further discussed in Rose (2002, Ch 2), as well as the more general and crucial question – too complicated to go into here – of how many observations per feature are needed in offender and suspect samples.)

Figure 1 shows several important things typical of speech features already mentioned. First, there is clearly within-speaker variation involved: each observation within the offender sample is different, and, more importantly, each mean value from the suspect's different conversations is different. This is important because, as already mentioned, one of the greatest contributory factors to the magnitude of within-speaker variation is the fact that speech was produced on different occasions. Second, the inevitable differences between samples must be pointed out.

The values for the acoustic feature in the offender's "okays" range from just above 1900 Hz to just below 2350 Hz, and the mean values for the suspect's "okays" range from 2000 Hz to just above 2400 Hz. A final feature which is usually, but not invariably, present is that the samples heavily overlap. They would overlap even more if individual values had been shown from the suspect rather than means. (Lack of overlap would not necessarily point to provenance from different speakers: within-speaker variation can be drastic enough to result in non-overlapping samples from the same speaker.)

These between-sample differences constitute the forensic-phonetic evidence. What must be determined is the extent to which the differences support the hypothesis that they come from the same speaker. This, in turn, is determined by finding out how much more likely the differences between the samples are, assuming that they have come from the same speaker than assuming that they have not. As explained above, the ratio of these probabilities will quantify the likelihood ratio for the evidence from the "okay S2T2F2" feature and thereby its strength.

The calculation of the likelihood ratio for data like this is mathematically and statistically complicated, and need not be discussed in detail at this point, although it will be illustrated below: see **[99.810]**. The LR is actually about 10. That is, you are about 10 times more likely to observe the difference between these samples if they had come from the same rather than different speakers.

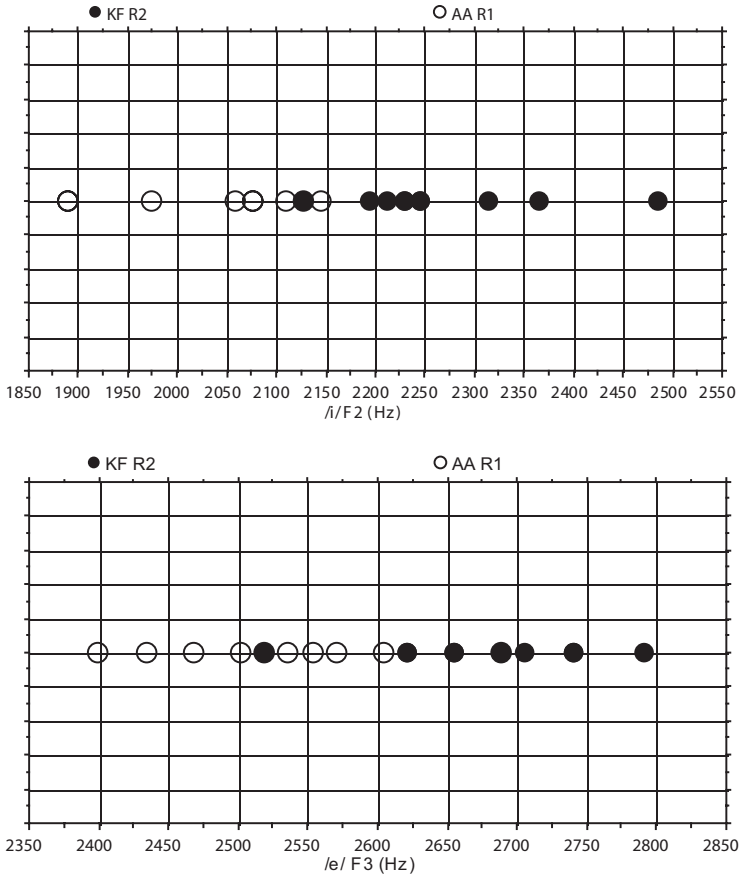
An example was given above of a likelihood ratio comparison for the single syllable "'kay" where it was shown that it was associated with a LR of 10. Let us assume that the suspect's voice and the offender's voice are now being compared with respect to both these features, ie "'kay" and "okay S2T2F2". The two LR values can be combined by taking their product to give the strength of the evidence based on both features, thus $LR_{'kay} \times LR_{okay\ S2T2F2} = 10 \times 10 = 100$. One would be now 100 times more likely to observe the evidence – the differences in the incidence of "'kay", and the differences in the acoustic feature in "okay" – assuming that the samples came from the same rather than different speakers.

The result of the comparison using two features would offer even stronger support – perhaps one could call it "moderate support" – for the prosecution hypothesis that the same speaker produced both samples. (It is important to point out that LRs for features can be combined in this way only if the two features are independent, which is a reasonable assumption in this case: see **[99.290]**.)

The example above showed data in support of the prosecution hypotheses. What does a comparison with samples from different speakers look like? Here is one, using Japanese data, that supports the defence hypothesis, again using two features. (All figures using Japanese data are quoted from Kinoshita (2001).)

Figure 2 shows data taken from a comparison between two samples taken from two different male Japanese speakers, called AA and KF. This time, both features are acoustic (both involve formant frequencies). One, labelled /i/ F2, is measured from a vowel similar to the one in the word "hid", and one, labelled /e/ F3, similar to the vowel in the word "head". Each dot represents one measurement of the feature from a vowel in a separate word, all of the words said by the speakers in a single recording session (identified as R(ecordings) 1 and 2 in Figure 2). Figure 2 shows typical examples of both between- and within-speaker variation in both features. Thus, each speaker shows a different range of values for the same acoustic feature: for /i/ F2, AA's sample has values ranging from just under 1900 Hz to 2150 Hz, and KF's sample has values ranging from about 2125 Hz to just over 2475 Hz. Note that although the two speakers' ranges are different, they still overlap for both features: this, again, is typical.

FIGURE 2 Comparison between samples from two different speakers (KF and AA) using two acoustic features (top: /i/ F2; bottom /e/ F3)



In a real forensic comparison one would not know that the two sets of data labelled AA and KF were from different speakers: they would simply constitute two speech samples being compared with respect to the acoustic features /i/ F2 and /e/ F3. Again, the estimation of the LR for the comparisons is too complicated to explain here. They turned out to be 0.24159 for the /i/ F2 feature and 0.07367 for /e/ F3. Note that both values are smaller than 1, consistent with the prediction of Bayes' Theorem, according to which same-subject data should be associated with a LR greater than 1, and different-subject data should be associated with a LR less than 1.

It is probably easiest to understand LR values smaller than 1 in terms of their reciprocal, thus the $LR_{/i/\ F2}$ of 0.24159 means that you would be (1 / 0.24159 =) 4.14 or just over four times more likely to observe this difference between the speech samples for this feature if they were from different rather than the same speakers: weak support for the defence, therefore. In the same way, the $LR_{/e/F3}$ of 0.07367 means that the difference between the samples is about (1 / 0.07367 =) 13.5 times more likely under the assumption of different speakers. Since these two features were found, perhaps surprisingly, to be negligibly correlated, they can be combined to give a LR of (4.14 x 13.57 =) 56.18. This means that the two samples, when compared with respect to these two features, differ in a way that is about 56 times more likely were they to have come from different speakers: the defence's position is now looking stronger.

The comparisons above were with two features. In Table 1 are illustrated some LR-based comparisons with more than two features. Table 1 shows some more Japanese comparisons, both between- and within-speaker, this time using not two but six acoustic features. Two different-speaker comparisons are shown, keeping one speaker (KA) constant. They are between KA and KF, and between KA and TY. Two same-speaker comparisons are also shown, with speakers KA and TN. These same-speaker comparisons are between the speech of the single speaker on one occasion, and their speech about a fortnight later.

TABLE 1 Likelihood ratios from two different-speaker and two same-speaker comparisons using the same six acoustic features

Comparison	/i/ F2	/e/ F2	/e/ F3	/o/ F3	/m/ F3	/s/ F3	Combined LR
KA-KF	***2.138***	***1.419***	***2.624***	0.032	0.051	***2.211***	**0.029**
KA-TY	***1.614***	0.031	0.406	1.8E-5	***1.066***	0.372	**1.44E-7**
KA-KA	***0.679***	2.036	1.669	6.037	5.613	2.685	**219.925**
TN-TN	3.749	***0.943***	2.933	1.021	***0.274***	2.955	**8.572**

The comparison adduces four acoustic features in addition to /i/ F2 and /e/ F3 illustrated in Figure 2. These features are arranged across the top. The values for the LR for comparison with each feature are given in the column under the feature, and in the right-most column is the LR from the combined six features. This is the product of the six LRs for the individual features. Some of the figures are very small, so so-called engineering, or scientific notation, is used (in 1.8E-5, eg, E-5 means shift the decimal point five places to the left, thus: 0.000018).

From the values for the combined LR in the right-most column, it can be seen that the two different-speaker pairs KA-KF and KA-TY are evaluated with LRs smaller than 1, and the two same-speaker pairs are evaluated with LRs bigger than 1. Thus the results of all four comparisons agree with reality.

The data in Table 1 have been chosen to illustrate three very important points associated with the use of LR-based comparisons. The first is that not all same-speaker forensic comparisons using a single feature will automatically yield a LR greater than 1; nor will all different-speaker comparisons, with a single feature, yield a LR smaller than 1. To explain this once again using the anthropometrical example, it is possible for two humans to be measured with the same, or similar, atypical height, and the comparison based on this feature thus to be evaluated with a LR appropriate for the same individual. In the same way, it is perfectly possible, eg, for two different speakers to have extremely similar, yet relatively atypical, values for a particular parameter, and consequently for samples of their speech to be evaluated with a LR appropriate for same speakers.

The same reasoning applies, mutatis mutandis, to same-subject comparisons. Intervertebral discs absorb water during sleep, and lose it during the day. Therefore if you measure my height when I get up in the morning, it will be greater than at the end of the day, after the discs have also been compressed by their usual pounding. So it is possible for a comparison based on my height at these two times to be associated with a LR that is more typical for different individuals. In the same way, there may be variation for a particular voice feature within the same individual that will be big enough to yield a LR less than 1, that is, a value more likely were different speakers involved.

That the LRs for the individual features can sometimes go against the reality is clearly shown in the data in Table 1. Table 1 shows that in each of the four comparisons, there are features for which the LR is unexpected, that is, it is bigger than 1 for different-speaker comparisons, and

smaller than 1 for same-speaker comparisons. These cases are shown in bold italics. Thus in the different-speaker comparison between KA and KF, four out of the six features were evaluated with a LR greater than 1, giving a LR of (2.138 * 1.419 * 2.624 * 2.211 =) 17.6. This means that the ratio between the similarity of the samples and their typicality was more indicative of same-speaker than different-speaker provenance: you would, eg, be twice as likely to observe the difference in the /i/ F2 feature between KA's and KF's samples had they come from the same speaker. The magnitude of the LR for the remaining two features, however, was enough to reverse this 15 into an overall value smaller than 1: (17.6 * 0.032 * 0.051 =) 0.029. Likewise, in the same-speaker comparison for TN, there are two unexpected LRs less than 1, but the magnitude of the LRs for the remaining features is great enough to result in an expected combined LR greater than 1.

The second point to be made from the data in Table 1 is that the strength of a particular feature as reflected in its LR is not invariant, but depends, as it should, on the particular comparison. If you were comparing two humans with respect to height and they both had very similar heights, the likelihood ratio associated with the difference in height would not be as big as if you were comparing two humans who differed considerably in height. This difference between the two humans' heights could even be smaller than the difference in height measured for the same subject at different times during the day (due to the height of the intervertebral discs). The likelihood ratio is also a function of typicality, so even the same magnitude of difference between samples in a particular feature will be evaluated differently depending on how typical the difference is of the relevant group of speakers to whom they were being compared. It can be seen from the LRs in Table 1 that there is no single strongest feature – the one that has the biggest effect on the LR – although for this group the /o/ F3 feature comes close. This shows that forensic-phonetic features do not, in a particular comparison, have inherent strength (although they may do in the average sense).

The third important point illustrated in the Table 1 data has to do with the combined LR values at the right of the Table, and follows from the previous two. It can be appreciated that, although they conform to reality (the LRs for the same-speaker pairs are bigger than 1; those for different-speaker pairs are smaller), they are still all very different. This reflects the different strength of the evidence in each comparison. For the comparison between KA and KF, the LR of 0.029 indicates that you are about (1 / 0.029 =) 36 times more likely to observe the difference between the six features if they had come from different rather than same speakers: really rather moderate support for a defence hypothesis. For the comparison between KA and TY, on the other hand, the LR of 1.44 E-7 means that the differences between the two samples in these features are (1 / 0.000000144 =) ca seven million times more likely under the assumption that they came from different speakers: very strong support indeed for the defence. The LRs for the two same-speaker comparisons also differ, although by less than the different-speaker examples. You would be ca 220 times more likely to get the differences between the two non-contemporaneous samples from KA if they were from the same speaker, but only 9 times more likely to get the differences from TN.

All three points illustrated and discussed above are the consequence of the differential between- and within-speaker variation in voices, whereby variation encompasses both differences in typicality and similarity. Recall that the LR is a measure of the ratio of similarity to typicality. Some voices are more similar to each other than others, some voices are more typical than others. Some voices show more than typical variation, some less than typical variation.

It is possible, theoretically, to carry the forensic comparison to as many features as desired. As LRs from new features are added – as samples are compared with respect to more and more features – the possibility increases that the resulting overall LR will be shifted increasingly lower than 1 for cases where the samples come from different speakers, and increasingly above 1 for same-speaker samples, even though the LR for some features may go against the general trend. In this way, the strength of evidence, either for or against the prosecution hypothesis, will increase.

This is another important point about the likelihood ratio approach. Two speech samples compared with respect to three features cannot be expected, in the long run, to give as strong evidence as the same speech samples compared with respect to 10 features. Two obvious questions then arise: how many features should be compared forensically, and which ones?

How many features?

[99.260] If it is assumed that all voices are different – that they all inhabit different parts of a hyperspace defined by their component features – it follows that they must be ultimately absolutely discriminable. Under these conditions a likelihood ratio type approach would not be necessary to discriminate one voice from another. This assumes, however, that we have access to all the features of a voice, which is not even remotely possible: samples are limited and will not contain all features; time is limited and will not permit extraction of all features that *are* present.

During this process it must always be borne in mind that it is quite possible that the evidence supporting same-speaker or different-speaker provenance will not be strong. In other words, no matter how many features are compared, the combined LR does not shift much either above or below unity. The reasons for this derive from the nature of voices and the use to which they are put. There is a limit to which human vocal tracts differ, and thus between-speaker variation in acoustics is limited. The presence of within-speaker variation in acoustics, and the fact that its magnitude is often increased by the lack of control over forensic speech samples, means that the ratio of between- to within-speaker variation in acoustic voice features is generally not very big. This imposes quasi limits on the magnitude of LRs for individual features. One important function of speech is to communicate. This, too, militates against endless, uncontrolled variation in linguistic features. By definition, too, most voices can be reasonably expected to have typical values for most features – again constraining LR to be similar to 1. Thus, it is possible for the voices of two different humans to have similar and typical features, and the LR comparison of them not to yield strong evidence that they are different. This is the individuality postulate in another guise, whereby no two objects (here: speakers' voices) are identical, although they may be indistinguishable.

Representativeness of samples

[99.270] It can be intuitively understood that a single random selection of an *individual* item from a group of such items – a marble from a bag of marbles perhaps – will not necessarily result in the choice of a typical item. You may have chosen the biggest, or smallest, marble, and thus have an inaccurate idea of what the typical marble size is. When comparing voice samples, it is important to know that the samples being compared are in some sense typical of the voices from which they come. Unrepresentative samples can lead to voice samples from different speakers being evaluated as more similar than they really are; or voice samples from the same speaker being evaluated as more different than they really are; both outcomes have judically undesirable consequences. It is important, therefore, for voice samples to be representative. This idea operates on two levels:

(1) There is a need for many samples of both the offender's and the suspect's voice to be taken on many different occasions, so that one can assume that the speech samples are representative of the voice(s) involved.

(2) There is a need for many tokens of each individual feature being compared, so that one can assume that the value for the feature is representative.

In many formulae for the calculation of LRs, the effect of the number of tokens in a feature, and number of voice samples, is built in. The result is that, the fewer the items available for comparison, the nearer the LR moves to unity, and the weaker the evidence.

There is thus safety in numbers, but this brings certain problems. In particular, unless there are good grounds for assuming the samples of the offender's voice are all from the same speaker, features from the offender's samples cannot be legitimately pooled to gain a more representative picture of the offender's voice. This is because the identity of the offender's voice is (of course!) not known. This means that very often comparisons have to be made between a composite suspect voice made up from a lot of suspect voice samples, and each of the offender samples in turn. As a result, there will inevitably be different LRs for each comparison. This is as it should be, because it cannot be expected that each speech sample will contain identical amounts of the same individual-identifying features.

These important questions – in particular, how much data are needed, and the architecture of observation data in forensic speech samples – are dealt with at length in Rose (2002, Ch 2).

The choice of forensic-phonetic features

[99.280] Given the fact that a voice can be characterised in terms of an effectively limitless number of features, how are the specific features chosen that are to be the basis of comparison between two or more voice samples?

The first obvious requirement is the *potential* of the feature to yield LRs which will substantially deviate from unity, and thus afford evidence which will give the strongest support to either the prosecution or defence hypothesis. (This is also an important practical question, since time is usually limited and cannot be wasted on comparing samples with respect to features that are not likely to result in large LRs.)

The examples above have shown that there is no way to tell a priori which features will have this property in a particular comparison. However, those features are preferable which are known to show small variation within a speaker and large variation between speakers. The ratio of within- to between-speaker variation for a feature is known as its "F-ratio", so it is a sensible heuristic to examine features that are known to be associated with large F-ratios. There are several acoustic features that are conventionally chosen on this basis. Very frequently, the necessary prior auditory analysis of the data will suggest additional features. Both samples might have an idiosyncratic way of pronouncing a particular sound or set of sounds, or one sample might have a recognisable speech defect not present in the other.

The choice of features will also depend on the language of the samples, since the individual-identifying potential of some speech features is language-dependent (the ratio of within- to between-speaker variation for a sound may not necessarily be the same in different languages; a sound that has good potential in one language might not even occur in another). Another important consideration – discussed below at **[99.290]** – is to avoid features that are likely to be strongly correlated.

Generally speaking, then, it constitutes part of forensic-phonetic expertise to know which features are likely to yield probative LRs, given the real-world conditions of the investigation.

Independent and dependent evidence

[99.290] As explained above, forensic-phonetic comparison has to involve many more than one piece of forensic-phonetic evidence. The combination of the LRs of different pieces of evidence is straightforward if they are independent, and involves, as illustrated above, taking the

product of the individual LRs. Here is an example of independent evidence adapted from Robertson and Vignaux (1995).

EXAMPLE 3

Suppose that the accent of the offender in an armed hold-up can be identified from the video surveillance recording as Northern Irish, and the suspect also has a Northern Irish accent. This is the first piece of evidence E_1. Suppose also that a wad of banknotes is found under the suspect's mattress – E_2. Robertson and Vignaux (1995) state that two pieces of evidence can be said to be independent if the truth or falsity of one would not affect the assessment of the probability of the other, and here this might well be the case, since there is no reasonable connection between one's accent and propensity to hide money under the bed. One could proceed to calculate a combined LR for these two independent pieces of evidence by taking their product: $p(E_1 \mid \text{suspect is guilty}) / p(E_1 \mid \text{suspect is innocent}) * p(E_2 \mid \text{suspect is guilty}) / p(E_2 \mid \text{suspect is innocent})$.

Suppose now that the forensic expert cites as two pieces of forensic-phonetic evidence: E_1 that both suspect and offender have a Northern Irish accent; and E_2 that both offender and suspect have a rising intonation (ie pitch) in statements. These are clearly dependent, since a rising intonation in statements is a characteristic of Northern Irish English. In fact, since in this case the rising intonation in statements is almost totally predictable from the type of accent, there are not two pieces of evidence here at all, but one.

If different items of evidence are not independent, combination can become very much more complicated. For example, the combined LR for two items of evidence E1 and E2, where E2 is dependent on E1, is the product of two LRs. The first LR is for the first item of evidence E1, namely $p(E1 \mid H) / p(E1 \mid \sim H)$. (The "| ~H" part of the denominator means "assuming that hypothesis H is not true".) The second LR is for the second piece of evidence E2 taking into account *both* the first piece of evidence E1 *and* the assumption H, namely $p(E2 \mid E1 \text{ and } H) / p(E2 \mid E1 \text{ and } \sim H)$.

It is not a straightforward matter, when estimating a LR, to assess and take into account the degree of interdependence of forensic-phonetic, or, indeed, any forensic data, and the problem is still being actively researched. Because of this, it is probably a good idea to choose features that are minimally correlated in the first place.

Gut feelings

[99.300] It is a natural, and probably automatic, human response, upon hearing voices, to make snap, unreflected judgments as to whether they are from the same person or not. It is only poorly understood how this occurs, but two forensically important facts are clear. It is well known that humans' judgments are very much better if the voice is familiar to them than if they have not been exposed to it before; and it is clear that the judgment can be influenced by expectation: you hear who you expect to hear. Prolonged listening can sometimes result in a change of mind.

Humans appear programmed to act in this way. Since most forensic identification experts are also human, they cannot avoid gut reactions of this kind. The question then arises of what to do with such reactions.

One possibility is to ignore them, and concentrate exclusively on calculating LRs for features. In favour of this is the fact that, although most experts will spend a long time listening to the samples in question, the voices thereby becoming familiar to them, it is not known to what extent

their initial reaction might have been influenced by expectation effects in the first place. Another possible approach is to attempt to incorporate one's overall feeling in the analysis by citing a LR for the gut reaction. That is, attempt to answer the question: "What is the probability of me feeling that these samples come from the same voice assuming that they do, and what is the probability of the feeling assuming that I am wrong?" In order to evaluate this, knowledge of past performance in similar tasks is necessary.

Summary

[99.310] From the preceding discussion, it should be clear that any TFSI report can reasonably be expected to make at least the following things explicit. It should be clear, first, what features were used to compare the speech samples, and their choice should ideally be justified on the basis of expected F-ratios. It should also be made clear to what extent the features can be considered to be independent.

The report should include an estimate for the LR associated for each feature used to compare the samples. This will be the ratio of the probability of observing the difference in the feature assuming that the samples have come from the same speaker to the probability of observing the difference assuming that they have come from different speakers. It should contain, and make clear, how the probabilities were calculated. (A distinction is sometimes drawn in this regard between so-called "hard" and "soft" probabilities. Hard probabilities are based on available data, or information calculated for the purpose; soft probabilities are based on the expert's feeling depending on their prior, unquantified, experience.)

Finally, the report should include an overall LR for the combined evidence, together with an estimate of the range of possible LRs (this is especially necessary where soft probabilities are concerned). There must also be a caveat that the overall LR still needs to be evaluated in the light of the prior odds before a statement can be made of the probability of the same speaker being involved.

[The next text page is 99 - 4051]

FORENSIC-PHONETIC FEATURES AND TYPOLOGY OF FORENSIC SPEAKER IDENTIFICATION ANALYSIS

[99.350] In the previous sections, it was explained that the forensic-phonetic expert has to determine a likelihood ratio that quantifies the probabilities of the forensic-phonetic evidence under the competing assumptions of prosecution and defence hypotheses. The forensic-phonetic evidence, in turn, consists of the differences or similarities in selected features of the speech samples. This section describes and exemplifies the features that are used in the forensic comparison of speech samples, and makes explicit the kind of knowledge that informs the identification, extraction and forensic evaluation of such features.

Differences between types of forensic-phonetic features can also be used to characterise some of the different analytical approaches to forensic speaker identification that will be encountered, and these different approaches are also described.

Forensic-phonetic features

[99.360] As explained above, the forensic comparison of voice samples works in terms of forensic-phonetic features (also called "parameters" or "dimensions"). It is one of the important properties of voices that there are many possible features, many of which are effectively independent. However, each of these features falls into one of four main types.

The primary distinction between forensic-phonetic features is whether they are acoustic or auditory, and there is a cross-cutting distinction as to whether they are linguistic or non-linguistic. It is useful to draw this second distinction because it is often erroneously assumed that speaker identity is encoded only in non-linguistic features. Thus it is possible to characterise a particular feature used in the forensic comparison of voice samples as acoustic-linguistic, acoustic non-linguistic, auditory-linguistic or auditory-non-linguistic. Some of these features, especially the acoustic ones, can be further classified. One difference is whether the feature is long-term or short-term; and a very important difference within acoustic features is between traditional and automatic features.

Acoustic versus auditory features

[99.370] Acoustic and auditory features are best characterised in terms of the analysis by which they are extracted. The raw material for forensic-phonetic analysis will usually be the magnetic patterns of tape recordings of speech, or their digitised version on a CD. There are basically two ways of analysing this raw material – acoustically or auditorily – and both of them are indispensable in TFSI. The dual nature of the analysis gives rise to the abovementioned primary classification of forensic-phonetic features as either "acoustic" or "auditory".

Auditory features are aspects of the recordings that can be heard and objectively described by an observer trained in describing how voices and their speech sounds actually *sound*. It needs to be emphasised that this is not an holistic, undifferentiated auditory response – eg, "the voices in the two samples sound as if they are the same" – but an analytical one, based on responses to individual features. One typical auditory feature might be the vowel in the first syllable of the word "hello": the suspect speech sample might have an *uh* vowel (like the vowel in "hut"), compared to an *eh* vowel (like that in "head") in the offender sample. (The reader might like at this point to see how well they have understood the section on voices and the forensic comparison of voice samples at **[99.220]** ff by asking themselves: Should this be the case (ie should the suspect and offender samples differ in this way), what is the correct question to ask within the Bayesian framework in order to evaluate the strength of the evidence?)[1] Another auditory feature might be overall pitch: both samples might sound to have lower than normal overall pitch, for example.

Acoustic features are identifiable parts or wholes in the patterns of acoustic energy radiated from a speaker, extracted, measured and compared by computer.

1 What is the probability of observing these two different vowels in "hello" assuming (a) the samples were spoken by the same speaker and (b) assuming they were spoken by different speakers?

The need for both acoustic and auditory comparison

[99.380] The distinction between acoustic and auditory features is a very important one in forensic phonetics, since there are good reasons for assuming that voice samples must be forensically compared with respect to both types. Supporting arguments can be found in Rose (2002, Ch 3), but the main one is simply that each approach on its own is known to be potentially associated with significant shortcomings. An auditory approach on its own is inadequate because it is possible for two samples to sound similar even though there are significant differences in the acoustics: see Nolan (1990). The converse also is found, namely that two samples can have very similar acoustics and yet be easily distinguishable in a single auditory feature: see Nolan and Oh (1996).

Trivially, too, a prior auditory analysis is necessary in order to decide whether the forensic samples are comparable in the first place, and, if they are, what is to be compared. Less obviously, there is also the matter of identifying boundaries between speech samples. In a recording of a telephone conversation, eg, one obviously needs to know when one speaker has ended and another has started, or to what extent two or more speakers have overlapped. It is still problematic to reliably make these decisions automatically by computer, and they are better made using the combined auditory native-speaker skill and expertise of the investigator.

Linguistic features

[99.390] Although they might contain coughs, laughs or screams, most forensic-phonetic samples are examples of speech. Speech is the most common medium in which language is realised (writing is another). Language is an incredibly complex code which links the meanings a speaker wants to convey with the sounds they produce, so that a listener can then decode the sounds they hear to understand the meaning that the speaker wanted to convey. The code of language is highly structured and aspects of this structure – in particular, how linguistic units are organised and realised – are called linguistic features. It is important to realise that, although there must, of course, be agreement in most of the language code to enable speakers to communicate, speakers of the same language can and do differ in linguistic features.

There are many different types of linguistic features, but they can be broadly categorised, in terms of the level of linguistic structure they participate in, as:

- phonological (having to do with speech sounds);

- morphological (having to do with how words are structured in terms of the smallest meaningful units in language called morphemes); and
- syntactic (having to do with the ways words are strung together into longer units like phrases or sentences).

Phonological features

[99.400] Most of the time, forensic-linguistic features will be phonological. For example, two samples might be compared with respect to the way a particular speech sound sounds to have been produced: in both samples the *r* sound may sound as if it has been pronounced in an idiosyncratic way. Samples might also be compared with respect to the acoustic structure of certain sounds, especially vowels, or whether a particular contrast in sounds is made, eg between *wh* and *w* in "what" and "watt"; "whine" and "wine". (Most speakers from Scotland and Northern Ireland will distinguish these words naturally, for example.)

There are well-established categories for the ways in which speakers can differ phonologically in their speech sounds. These are called "systemic", "phonotactic", "incidential" and "realisational", and are described in detail in Rose (2002, Ch 7).

Morphological features

[99.410] An example of a morphological feature might be if both samples clearly distinguished between "you" and "youse" (or "y'all"), that is, encoded a number distinction in the second person pronoun. Another example would be if the word "youths" were pronounced in one sample with a *th* sounding like the *th* in "thing" as opposed to the *th* in "this" in the other. This is a morphological distinction because it has to do with the differential realisation of a morpheme (smallest meaningful unit), namely {*youth*} (the curly brackets indicate that "youth" is the name of the morpheme).

Syntactic features

[99.420] An example of a syntactic difference between samples might be if one sample had "these clothes need washing" and the other had "this table needs cleaned". In the first example, the "washing" is a gerund, functioning as a noun object to the verb "need" (analogous to the noun "shave" in "you need a shave"). The second example, however, has a very different syntactic structure, and comes from the full phrase "needs to be cleaned" from which the "to be" has been optionally deleted. Another example of a syntactic difference might be if one sample had "she was pretty pissed-off but," and another had "but he was clearly pleased about it". The first has the concessional adverb "but" (meaning "nevertheless") at the end of the sentence, the second has it at the beginning.

Non-linguistic features

[99.425] As examples of non-linguistic features might be cited apparent habitual use of nasalised or breathy voice, a fast speaking rate, or a lower than average pitch. These are – in English at least – not linguistic features, because they do not have to do with the organisation or realisation of specific speech sounds.

The rider in the previous sentence is an important one, for it is important to understand that many features are generally not *inherently* linguistic or non-linguistic: it depends on the language. What may be a non-linguistic feature in one language may be a linguistic feature in another. A nice example of this is so-called "creaky voice". If you are an English speaker, say a vowel like *aah* starting on a high pitch and slowly lower your voice trying to get as low as possible.

The chances are that you will become creaky when you reach a certain low pitch: your voice will sound as if it consists of individual pulses. Some English speakers either habitually adopt creaky voice, or use it to signal short-term non-linguistic information, eg "I'm bored". Both of these are examples of creak functioning as a non-linguistic feature (they are actually examples of the feature being used extralinguistically and paralinguistically – this is a further, more detailed, taxonomy of feature use).

In some languages, however, creaky voice is simply a part of the inventory of speech sounds. Just the same as vowels and consonants, it can signal a different word if you use it. Creaky voice in this sense is obviously a linguistic feature. (Because it can signal the difference between words, it is said to be "contrastive".) Standard (Northern) Vietnamese is a good example of such a language with contrastive creak. Creak is also used as a characteristic of a particular accent – working-class Norwich speech, eg – and as such it could also be considered linguistic.

To complete what is a complicated but not atypical picture, it is not true to say that creak is exclusively non-linguistic in Standard English. It can be used linguistically as part of the realisation of a falling intonational pitch – that is, a linguistic unit – to indicate finality and/or that the speaker has finished what he or she wanted to say and is ready for the interlocutor to take over.

Because speakers can differ in linguistic features, and because the functional taxonomy of these features can be extremely complex and language-dependent – witness creak – it stands to reason that a thorough knowledge of the discipline that describes and analyses language, namely linguistics, is necessary to identify and describe them and to evaluate their forensic significance.

Moreover, offenders do not all speak English, and it cannot be assumed that what holds for English holds automatically for other languages. Therefore a thorough knowledge of the language of the samples is also crucial for forensic evaluation. One would not want to attach any significance to the fact that creak occurred frequently in two samples of Standard Vietnamese, for example. Given the fact that creak is a contrastive sound of the language, one would be just as likely to observe it assuming that the samples were from the same speaker as from different speakers, and the associated LR would be unity, and useless. Compare this with a hypothetical high incidence of creak in two English speech samples. Under these circumstances a high incidence of creak in both samples might be associated with a usefully high LR. Even here, however, linguistic knowledge is indispensable to properly evaluate a LR, since it would be necessary to distinguish between linguistic creak (at the end of a final falling intonation), and extralinguistic (and perhaps also paralinguistic) creak.

Types of forensic speaker identification analysis

[99.430] The consensus among TFSI practitioners is that voices can and must be compared forensically with both linguistic and non-linguistic features, and both acoustic and auditory features. These features are naturally the result of analysis, and below are briefly described important aspects of the different kinds of analysis used.

Auditory analysis

[99.440] It is possible to respond auditorily to many different types of information in a voice. Auditory forensic-phonetic analysis is the application of a suite of specialised conventional techniques for describing and analysing both linguistic and non-linguistic aspects of this information. Since, as described above, these features can reside in any aspect of linguistic structure from phonology to syntax, auditory analysis must presume familiarity with and ability in linguistic analysis. Phonological analysis, however, will be the most important part.

Comparison between phonological features actually involves two different types of analysis called "phonetic" and "phonemic", and both of these depend on a prior phonetic transcription. What is involved in phonetic transcription is described below. Before embarking on this, however, it is useful first to present, in a brief but essential digression, some of the anatomical features of the vocal tract, since many of the terms in phonetic and phonemic description are derived from them.

The vocal tract

[99.450] The vocal tract contains very many structures that speakers deliberately manipulate to produce speech sounds. However, only the two basic ones need mention here: the vocal cords and the supralaryngeal vocal tract (SLVT). They are basic in the sense that they constitute two independently functioning and controlled, but precisely aligned, modules in speech production.

The reader can get a good idea of the independent contribution of these two modules to speech simply by doing the following. First take a deep breath. Say an *ee* vowel (as in "heat") on a sustained, comfortable pitch, and then change to an *aah* vowel (as in "heart") while maintaining the pitch. Still maintaining the pitch, change to an *oo* vowel (as in "hoot"), and back to *ee*. What you have done is to change the contribution of the supralaryngeal vocal tract to produce different vowels, while keeping constant the contribution of the vocal cords, to produce the same pitch.

Now try changing the contribution of the vocal cords while keeping the contribution of the supralaryngeal vocal tract constant. Take a deep breath, and keep saying an *ee* vowel while you change pitch by going up and down. The pitch can be changed on any other vowel too, of course.

Vocal cords

[99.460] The vocal cords are two small lips of elastic tissue which are located in the larynx, or voice box. They stretch from front to back across the top of the wind-pipe, or trachea. There are two extremely important ways in which the cords can be manipulated. The gap between the cords can be opened (in which case the cords are said to have been abducted) or closed (adducted cords); and the cords can be tensed or relaxed. A picture of the author's vocal cords, showing them in both adducted and abducted position, is given at Figure 3 at **[99.1020]**.

Voicing

[99.470] Upon adduction, with the correct tension, the cords will vibrate when suitable pressure below the cords is supplied by the bellows action of the lungs. This vocal cord vibration is called "voicing" or "phonation". Sounds thus produced are called "voiced". Vowels are usually voiced, as are consonants like *m, n, l* and *r*.

Voicing can be easily experienced. Take a deep breath and switch, without pausing, between long *z* and *s* sounds (as at the beginning of "zoo" and "Sue"), thus: *zzzzssssszzzzssssszzzzssss* etc, and simultaneously palpate the larynx. The vibrations associated with voicing can be felt directly. By doing this while making alternating *s* and *z* it can be appreciated that *z* is voiced; *s*, on the other hand, lacks the vibrations and is consequently a "voiceless" speech sound. One of the important aspects of voicing can be seen in the fact that its absence or presence can signal the difference between speech sounds.

Pitch

[99.480] By changing the tension in the cords when they are vibrating, the frequency of their vibration can be controlled. Increased tension results in higher frequencies, relaxation in lower. This is the mechanism underlying the changing pitch of speech, since high vocal cord vibration frequencies are perceived as high pitch and low frequencies as low.

Pitch is put to very many uses in speech. For example, in English it can indicate the difference between a statement, eg "He's going", and a question, eg "He's going?" In tone languages, pitch is part of the inherent sound of a word, and differences in pitch signal difference in word identity. Say *ma* on a high level pitch. That means "mother" in Standard Chinese. Now say *ma* with a low, then rising pitch. That means "horse".

Phonation types

[99.490] The vocal cords can not only vibrate, they can be made to vibrate in different ways. This gives rise to different sounds and constitutes the third dimension of vocal cord activity in addition to the two (voicing, pitch) already described. Differing modes of vocal cord vibration are called "phonation types", the most common of which is called modal phonation, or modal voice. Creak is another phonation type. Depending on the language, different phonation types are used to signal different types of information (linguistic, paralinguisitc, extralinguistic), as has already been explained with creak.

Supralaryngeal vocal tract

[99.500] The author's supralaryngeal vocal tract – that part of the tract above the larynx – is shown in an x-ray in Figure 4 at **[99.1120]**. The SLVT consists of three main cavities – the nasal cavity, the oral cavity (or mouth), and the pharynx (throat) – and several structures. Two of the supralaryngeal structures important for speech are the tongue and lips. These can be easily identified in Figure 4. By moving these structures, the shape of the SLVT cavities can be changed, and this gives rise to different sounds. In Figure 4, note how the lips are close together and slightly pursed, and the tongue appears bunched up towards the top and back of the mouth. These are typical supralaryngeal articulations involved in producing an *oo* vowel as in British English "who". Another important structure is the soft palate, or velum. This is a muscular flap that can be held up or down, thus allowing access to the nasal cavity. When the velum is down, air can pass through the nasal cavity and produce nasal consonants, like *m* or *n*, or nasalised vowels as in French "un bon vin blanc". The velum is raised in Figure 4, so the vowel is not nasalised. The top of the larynx, the structure composed of several cartilages where the vocal cords are housed, is also visible in Figure 4.

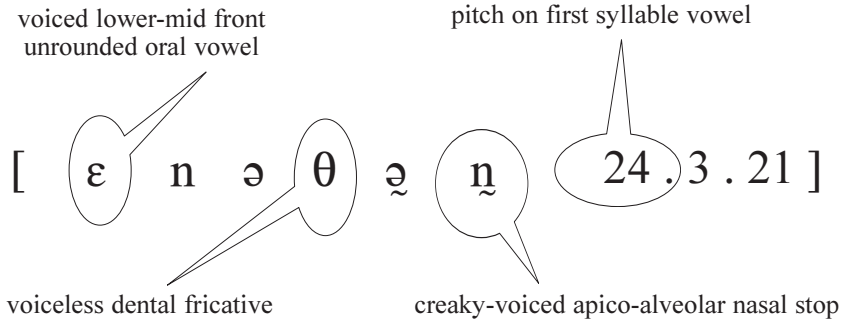
Supralaryngeal structures are also manipulated in the production of consonants. In an *s* or a *z* sound, eg, the front and sides of the tongue are raised so that a narrow constriction is formed in the region of the alveolar ridge (just behind the front teeth). When air flows through this constriction, and meets the barrier of the teeth, the hissing turbulence typical of *s/z* is generated. Consonants produced with continuous turbulence are called "fricatives"; thus *s* and *z* are both fricatives. Since for *s* and *z* the important constriction is made opposite the alveolar ridge, they are called "alveolar" fricatives. For consonants, supralaryngeal articulation is typically specified in so-called "place-manner" terms. "Place" is where the SLVT constriction is made, and "manner" refers to the type of constriction. Thus *s* and *z* both have alveolar place and fricative manner.

Apart from voicing, then, the different speech sounds *s* and *z* are produced in the same way, that is, with the same supralaryngeal articulation. All of this draws the reader's attention to the important notion of "componentiality" of speech sound structure – that speech sounds are not unanalysable wholes, but are composed of smaller units, which belong to various categories. Thus the sound *s* is composed of voicelessness, alveolarity and fricativity, each of which is a specification within the categories of voicing, place and manner. In modern phonology, components like "alveolar" are analysed in still greater detail, but this is unlikely to be of relevance for forensic comparison.

Phonetic transcription

[99.510] In phonetic transcription, a transcriber listens carefully and repeatedly to speech, and writes down what he or she hears using a conventional set of symbols, usually those sanctioned by the International Phonetic Association (IPA). This may sound rather simple, but actually phonetic transcription is a skill that involves some very complex assumptions. It is best to first give an example of phonetic transcription, and then rehearse the important aspects of phonetic transcription from a forensic point of view.

To help convey an idea of the important aspects of the nature of phonetic transcription, immediately below is an actual phonetic transcription: [ɛnəθə̰n̰ 24.3.21]. This represents a single example of the Australian English word "anything" as said by a speaker on a particular occasion (an intercepted phone call).



It can be seen that the transcription is enclosed in square brackets. These brackets are important: they are used to mark the status of the symbols within them as a *phonetic* representation: anyone who has learnt the convention (here: IPA) can understand exactly what the symbols mean, such that they could reproduce the utterance, as it was originally said, to the satisfaction of the person who said it, or another native speaker, or the original transcriber. Thus the conventional nature of phonetic transcription guarantees its potential replicability. This, in turn, enables verifiability: one transcriber's transcription of forensically relevant material can be checked by another transcriber. This is obviously a forensically important property of phonetic transcription.

Phonetic transcription embodies a long- and well-established way of objectively talking about speech sounds. To illustrate this, the meaning of some of the symbols in the transcription is given in balloons. Thus the theta [θ] symbol stands for a voiceless dental fricative. This means a consonantal sound made:

(a) with the vocal cords not vibrating (as explained above, this is what is meant by "voiceless");

(b) between the front of the tongue and the upper teeth (← "dental"); and

(c) with a constriction narrow enough to make the air flowing through it turbulent (← "fricative").

The "creaky-voiced apico-alveolar nasal stop" symbol [n̰] in the transcription indicates a sound made:

(a) with the vocal cords vibrating in a particular mode called "creaky voice" (this is indicated with the subscript squiggle);

(b) with the tip of the tongue (← "apico-"), in the region of the alveolar ridge behind the front teeth (← "alveolar");

(c) with the flow of air through the mouth totally blocked off (← "stop"), but exiting through the nasal cavity by virtue of the fact that the soft palate at the back of the mouth is open, thus allowing access to the nasal cavity (← "nasal").

The voiced lower-mid front unrounded oral vowel symbol [ɛ] in the transcription stands for a vowel made with:

(a) the vocal cords vibrating,

which sounds as if the speaker had:

(b) the tongue body positioned in the front of the mouth (← "front"), at

(c) a slightly lower than middle height (← "lower-mid"), with

(d) unrounded lips (← "unrounded") and

(e) their soft palate up (← "oral").

The numerals following the symbols in the transcription indicate pitch on a scale of one (lowest in speaker's pitch range) to five (highest in speaker's pitch range). Thus the [24] indicates that the vowel [ɛ] was said on a pitch rising from fairly low in the speaker's pitch range to fairly high.

It needs to be emphasised that speech sounds have structure, and complex structure at that. As pointed out above, they are not unanalyseable wholes, but consist of several components. Thus three of the components of the speech sound [θ] are its voicelessness, its fricativity and its dentality. If any one of these is changed, a different speech sound results. Changing the voiceless component into voiced by making the vocal cords vibrate makes a voiced dental fricative [ð] – the sound that occurs at the beginning of "this", "these", "those", the middle of "either" and the end of "mouth" when used as a verb. Changing the dental component into alveolar by moving the tongue tip backwards slightly in the mouth results in a voiceless alveolar fricative [s] – the sound that begins the words "see" and "saw", and ends the word "ice". Changing the fricative component into a stop by pressing the top of the tongue against the top teeth and thereby blocking the exit of air makes a voiceless dental stop [t̪] – the sound in the middle of the word "width" [wɪt̪θ]. A single phonetic symbol thus abbreviates a bunch of concurrent speech-sound components.

These components represent for the most part articulatory specifications: they refer to the behaviour of the speech organs. However, it is normally not the case that the professional transcriber actually thinks all the time analytically in terms of articulatory components when he or she is actually transcribing. In training, to be sure, a transcriber will have learnt to identify, distinguish and produce a very large number of speech sounds from languages all over the world by first learning to control and permute the individual features. He or she will also have learnt the appropriate symbols for these speech sounds.

However, as the learner becomes more skilled and able to identify and produce an increasingly large number of individual speech sounds, her or his ability shifts from analytic to holistic, rather in the fashion of learning to tie one's shoelaces. It is then that the transcriber can be "simply" said to identify the sound and transcribe it with the appropriate symbol. Thus it can be said that, although these phonetic symbols abbreviate inferred articulatory reality, they are most of the time used to represent the transcriber's immediate, unreflected, perceptual response to a segment or sequence of segments.

Unlike shoelace-tying, however, it is important to realise that the componential-analytical method is always available to the transcriber, and can be called up when the transcriber is not sure of some aspects of a sound. When a transcriber encounters a speech sound he or she recognises as not having heard before, eg, the transcriber can describe some of its components. He or she might say, eg, that "that is a voiceless coronal fricative of some kind" ("coronal" = made with either the tip or blade of the tongue).

Another important aspect of phonetic transcription is that it involves learning to operate in a perceptual mode that is maximally free from the phonetic categories of one's native language. It does not mean writing down the sounds you hear in some kind of English, as is often found in foreign language phrase-books, for example.

When we are born, we are programmed to hear the differences between an enormous number of speech sounds. However, as will be described below, not all phonetic differences are equally important for all languages, and part of the learning of our native language involves learning to ignore those phonetic differences that are not important for our language. The phonetic difference between the initial consonants in the Standard Thai words for "black" ([**t**am]) and "to pound" ([**d**am]) is easy to hear for native Thai speakers, and babies, but not for phonetically naive speakers of English, for whom both words sound like "dumb". The phonetic difference between the vowels in the words "head" and "had" is easy to hear for native speakers of English, but anything but for native speakers of German and French. Part of acquiring expertise in phonetic transcription, then, lies in learning to hear, and produce, these phonetic differences again.

One aspect of phonetic transcription that is forensically relevant is that it is possible to control the degree of phonetic detail that is responded to and transcriptionally encoded. Phonetic transcriptions can vary along a continuum between narrow, in which case one tries to describe speech sounds in their minutest detail, and broad, in which case one only transcribes selected relevant aspects of the sounds.

It usually makes little sense, for forensic-phonetic transcription, to try to capture every single phonetic detail of the differences between speech samples. This would take far too much time and quite often the real-world nature of the material imposes limits on ability to be sure about phonetic detail. Thus it may not be possible, eg, to decide whether a vowel over the phone is a lax [ɪ] or a half-close [e], or whether the creak extends over one or two or three segments. But it may be possible to say that, in one sample, creak is present somewhere in a word, but in the other is not present at all. The transcriber has to decide how narrow a transcription is warranted given the circumstances.

Phonetic transcription is a skill that can be taught, and some are better at it than others. In the author's experience from teaching it and testing it at university level, it takes about three months, with weekly about two hours of practical tuition, two hours theory, and several hours private practice, before the best two or three of a group of 30 or so students can agree in about 95% in their transcriptions for a set of data – say 20 medium-sized, two-to-three-syllable, words – from different languages that they have never heard before. This is actually agreement of a very high order indeed, given the fact that any single word will contain a very large number of bits of information to get right. As a combination of speech sounds from one of the world's languages, eg, the "anything" in the transcription example above contains, at a conservative estimate of six

features per segment, upwards of (6^6 =) 46,000 things to get right. Multiply this by 20 for a typical end-of-semester transcription test, and you get just under a million bits of information. It is important to mention all this, because the epithet "subjective" is sometimes found applied to phonetic transcription by people who have not been trained in it. The fact that transcribers can agree in most of these features gives the lie to this.

All this should not be taken to mean that one becomes expert in phonetic transcription after a few weeks' transcription practice. To be sure, external accreditation for phonetic transcription exists in the shape of the IPA Diploma in Phonetic Transcription and Phonetics of English administered by the University of London. But expertise in phonetic transcription is perhaps better construed in terms of the level of confidence one develops in knowing what one is hearing, and it takes many years' continuous exposure to as many of the different speech sounds of the world as possible to develop this.

In sum, a phonetic transcription is a conventional way of describing speech in any language. It permits the description of speech in considerable, largely objective detail, and can be checked.

Speech sounds and orthography

[99.520] Speech is apparently sometimes found described in forensic reports as a sequence of letters. It should be clear from the above that this conception is totally erroneous, and use of the term "letter" instead of "speech sound" in forensic reports should always be a warning sign. Because speech sounds embody an enormous amount of structure, they are very different from a conventional spelling system. Moreover, the relationship between speech sounds and spelling – in languages that have an orthography! – is never, ever one to one. The IPA phonetic symbol [g] means a voiced velar stop; the English spelling symbol g is a holistic unit and nothing more, corresponding to both a [g] and a [dʒ] in "gage", and to nothing at all in "gnat". In English, the high front unrounded oral vowel [i] is found variously spelled as *ee* ("heed"), *ie* ("believe"), *eo* ("people"), *ei* ("seize"), *ea* ("meat") and *ey* ("key"). Phonetic transcription is also obviously a far cry from attempts to use orthographical conventions of one's native language to represent non-native sounds (eg "comment alley voo" for French "Comment allez-vous?" [kɔmɔ̃talevu]).

Phonemic analysis

[99.530] It may come as a surprise that forensic speech samples cannot be described, from an auditory-linguistic point of view at least, with respect to their phonetic properties alone. In other words, it is not meaningful just to say that two samples differ (or are similar) in having different (or the same) speech sounds. In order for it to make sense, forensic-phonological comparison requires in addition the analytic conceptual framework of phonemics.

Phonemics is a conceptualisation of the sound structure of language that makes forensic comparison between the sounds of different samples possible. It is a fact of the sound structure of language that speech sounds exist on two different analytical levels. These are the "phonetic" level, which has to do with the sounds' actual articulatory and acoustic properties, and the "phonemic" level, which has to do with their functional organisation.

Phonemic analysis is again too complicated to explain in detail here: its application to forensic-phonetic comparison is treated in Rose (2002, Ch 7). However, it is important to give at least some idea of what it involves, since the reader is unlikely to be aware of the distinction between phonetic and phonemic analysis, or between their first-order units: speech sounds and phonemes. The following discussion will help give a flavour of the conceptual framework involved.

Phonemic structure

[99.540] In a given language, speech sounds are considered as realisations of abstract contrasting units called "phonemes" appropriate to that language. For example, English has a speech sound [n] – as in "gnat", "know", "dinner" or "tin" – called an alveolar nasal. English also has a velar nasal sound [ŋ] which occurs, eg, in "singer" and "long". [ŋ] is often spelled as "ng", and often called "eng" (as in "Ghengis"), or "engma". The terms "alveolar" and "velar" were introduced above as specific locations at which a consonant is made: the alveolar ridge is just behind the front teeth, and the velum is the soft palate, at the back of the oral cavity.

In English the speech sounds [n] and [ŋ] are said to "contrast", because the difference between them can signal the difference between words. The words "thin" and "thing", eg, or "sinner" and "singer", contrast solely in that one has an alveolar nasal and the other a velar. Because the sounds [n] and [ŋ] contrast, they are said to "realise different phonemes".

Phonemes are conventionally represented in oblique slashes, and speech sounds functioning as realisations of phonemes are called "allophones", so for English we say that the alveolar nasal phoneme /n/ is realised by the alveolar nasal allophone [n], and the velar nasal phoneme /ŋ/ is realised by the velar nasal allophone [ŋ].

This is usually formalised as:

$$/n/ \rightarrow [n]$$

and

$$/ŋ/ \rightarrow [ŋ]$$

where the arrow represents the relationship of realisation.

A single phoneme in a given language can be realised by one or more phonetically similar allophones, which is one of the reasons why the dual existence of speech sounds on both levels is not redundant. The phonetic nature of the allophone may be determined by the systematic context in which the phoneme occurs, in which case the allophones are said to be in "complementary distribution": one allophone of a phoneme occurs in one environment, and another allophone in another, mutually exclusive, environment. In most varieties of English, eg, vowel phonemes are audibly longer when followed by certain consonants. Say the words "beat" and "bead", paying attention to the length of the vowels, and you will easily appreciate that the vowel in "bead" is longer. From a phonetic point of view, then, the vowels in these words are different. One (in "beat") is long: [iː] (the two dots represent length); the other (in "bead") is longer still – represented as [iːˑ]. But from a *phonemic* point of view they both represent the same phoneme /i/. It is actually the difference between the two post-vocalic consonantal phonemes, here /t/ and /d/, that conditions the phonetic length difference in the vocalic allophones. This situation can be formalised thus as:

$$/i/ \rightarrow [iːˑ] / \_/d/$$

and

$$/i/ \rightarrow [iː] / \_/t/$$

(In words, "the (English!) phoneme /i/ is realised as [iːˑ] when it occurs before /d/, and [iː] as when it occurs before /t/.") The oblique slash / and the underline indicate "occurs in the environment of", so "/ _/t/" means "occurs in the environment before the phoneme /t/".

This is a systematic effect, in English, in two ways. First, it applies to all vowel phonemes, not just /i/. Second, it is conditioned not just by the difference between /t/ and /d/, but the same difference between analogous pairs /p/ and /b/, and between /k/ and /g/ (and, in fact, all pairs of phonemes that differ in the way that these pairs differ, namely, in voicing: /p/, /t/ and /k/ are

voiceless phonemes; /b/, /d/ and /g/ are voiced.). To appreciate this, try saying "bag" and "back" (/bæg/ /bæk/) or "loop" and "lube" (/lup/ /lub/) and note the differences in phonetic vowel length: [bæːˑg] [bæːk]; [luːp] [luːˑb]. The vowels are longer before voiced /g/ ("bag") than before voiceless /k/ ("back"); and before voiced /b/ ("lube") than before voiceless /p/ ("loop").

It is also possible for a given phoneme to be realised by one or more phonetically similar allophones in the same environment, in which case the allophones are said to be in "free variation". For example, when I say the English word "bat", I can choose to make my vocal cords vibrate at the same time as I make the initial consonant spelled with the *b*; or I can wait until the vowel starts before making my vocal cords start to vibrate. Thus two phonetically different sounds – two sounds which sound different – can be produced in this way: the first is transcribed as [b], and the second as [p].

Although [p] and [b] are phonetically different sounds, they are not contrastive in English at the beginning of a word because it does not matter whether I say [bæt] with a [b] or [pæt] with a [p]: it is still a "bat". In English, [b] and [p] are thus said to be allophones in free variation word-initially of the same phoneme /b/.

At this stage the reader is almost certain to be confused by the use of the symbol [p]. Isn't this the sound at the beginning of "pat", in which case how can it realise the same phoneme as in "bat"? The answer is: no, [p] is not the same sound as at the beginning of "pat". That sound – the "pat" sound – is a different sound again, and transcribed phonetically as [pʰ]. The reader must take on faith the fact that English words like "bat" and "boy" – words beginning with the phoneme /b/ – can begin with two phonetically different sounds [p] and [b], which are different from [pʰ].

One can at least make a virtue out of this confusion by pointing out that it demonstrates nicely the indispensability of having the objective method of talking about speech sounds that is part of the discipline of phonetics and linguistics. If I were trying to communicate this state of affairs with initial [p] and [b] in English to another phonetician or linguist who did not know about English and had never heard it, it would be sufficient for me to say that "in English, coincident voice-onset and lead voice-onset stops are in free variation word-initially". The listener would then be able to give acceptably varying tokens of words like "bat" and "boy" (assuming they knew how to say the rest of the word) without even having heard them.

Returning now to the argument, phonetic differences are thus not automatically contrastive. This applies either within a language (like the conditioned difference in English vowel length, or the free-variation difference in English vocal cord vibration just demonstrated for [p] and [b]), or between languages. In Italian, eg, unlike English, the phonetic difference between a velar and an alveolar nasal is not contrastive. This is because the velar nasal always occurs before a velar consonant like /g/ or /k/ and the alveolar nasal never does. The two sounds [n] and [ŋ] cannot therefore contrast by definition. In order to contrast, two sounds must be able to occur in the same environment, and these never do because one ([ŋ]) is always followed by a velar and one ([n]) never is. In Italian [n] and [ŋ] are, in fact, complementarily distributed allophones of the same /n/ phoneme. This is formalised thus:

$$/n/ \rightarrow \begin{cases} [ŋ] \,/\, \_ \text{ velar consonant} \\ [n] \,/\, \text{elsewhere.} \end{cases}$$

Since the speech sounds of language are structured in the ways described above, the only way to make sense of – ie, to compare and evaluate – the inevitable phonetic variation between forensic speech samples is by assuming the constructs of phonemic analysis: that the sounds themselves are realisations of abstract contrasting units called phonemes.

To understand this, imagine that one voice sample contained the word "fight" pronounced as in Standard British English ([faɪt]), but the other had the word "light" pronounced like *loyt* ([lɔɪt]). Without assuming phonemic structure it is not possible to say that the two samples differ in anything, because we have no indication that the diphthong in "fight" is comparable with that in "light": they are just two phonetically different diphthongs. With phonemes, however, the comparison is possible, because we know that both diphthongs are allophones of the same phoneme. It would then be possible to say that the two samples differed in the realisation of the phoneme /aɪ/, with one having /aɪ/ → [aɪ], and one having /aɪ/ → [ɔɪ]. Phonemic structure thus guarantees phonological comparability between samples.

A final point is that since languages, dialects and speakers can differ with respect to their phonemic structure, it is essential that the forensic analyst be completely familiar with the phonemic structure of the language(s) in question.

Example of forensic comparison with linguistic-auditory features

[99.550] The following is a typical example, adapted from case work, of the comparison of two forensic speech samples using linguistic-auditory features. The language involved is Mandarin Chinese, and the data are from intercepted telephone calls. Some comments on Mandarin and its use in the People's Republic of China are necessary before presenting the data.

EXAMPLE 4

The People's Republic of China has an official Standard Language, called *Putonghua*. (One translation of *Putonghua* is Mandarin, but since the term "Mandarin" can refer to quite a few other varieties of Chinese, eg an ensemble of homogeneous dialects that spreads over about two-thirds of China, it is better to retain the term *Putonghua*). As a Standard Language, *Putonghua* is used in broadcasting and education, and aspects of its structure are normatively encoded in various ways, like grammars, textbooks and dictionaries. One such aspect of its structure is its sounds: the sounds of *Putonghua* are based on the sounds of a particular historically prestigious dialect – that of Peking.

Of relevance for the current example is that Peking dialect has two contrasting sets of consonantal phonemes: a so-called *retroflex* set, and a so-called *dental* set. In dental phonemes, the front of the tongue touches the front teeth; retroflex phonemes can be thought of as being made by curling the front of the tongue up and backwards in the mouth. (Sean Connery is often stereotyped as having a retroflex allophone for his /s/.) As already explained, "contrast" is a technical phonological term meaning that the difference between the members of the two sets has the potential to signal a difference in word meaning. An example of two contrasting sets of phonemes in English would be /p, t and k/ (one set) versus /b, d, and g/ (the other set). These two English sets contrast because the substitution of the phoneme /p/ in the word "pig" by a /b/ would result in a different word, namely "big", and the same applies mutatis mutandis for /t/ and /d/, and /k/ and /g/: compare "to" and "do", and "cot" and "got".

example 4 - *continued*

TABLE 2 Retroflex and dental phonemes in Peking dialect

Phonemes	/ ʂ /	/ s /	/ tʂʰ /	/ tsʰ /	/ tʂ /	/ ts /
Approximate English equivalent	*shoes*	*Sue's*	*choose*	-	*Jews*	-
Retroflexion	retroflex	dental	retroflex	dental	retroflex	dental
Continuance	fricative		affricate			
Aspiration	N/A		aspirated		unaspirated	
Word	ʂaŋ *on*	saŋ *mulberry tree*	tʂʰa *to differ*	tsʰa *to rub*	tʂaɪ *stockade*	tsaɪ *at*

Table 2 gives some Peking dialect words with retroflex and dental initial consonantal phonemes. The first row, labelled "phonemes", lists the six different phonemes involved. The next row gives the nearest English phonetic equivalents to four of the six sounds in a minimal quadruplet. The next row, "retroflexion", identifies the Peking dialect phoneme above it as retroflex or dental – it can be seen that the retroflex sounds are transcribed with an *s* with a little subscript curl (this was supposed to represent iconically the tongue tip curling backwards in the mouth). The next row classifies the six phonemes into two groups – fricatives (/ ʂ & s /) versus affricates (/ tʂʰ tsʰ tʂ & ts /) – according to a feature called "continuance".

The next row classifies the affricate phonemes according to whether they are aspirated or not. Aspiration refers to the amount of time between releasing the consonant and starting vocal cord vibration for the following vowel. Sounds with a long time between releasing the consonant and starting vocal cord vibration (a long time in articulatory phonetics is about 10 hundredths of a second) are called aspirated (and commonly transcribed with a superscript ʰ); sounds with a short time are called unaspirated. A similar contrast occurs in English between "choose" and "Jews", or between "char" and "jar".

These features (retroflexion, continuance, aspiration) can be used to separately classify all six Peking dialect phonemes. Thus Peking dialect /ʂ/ is properly described as a retroflex fricative; /tsʰ/ is an aspirated dental affricate.

The next two rows give six words, each beginning with one of the six different phonemes. The words' phonemes are given above, and their meaning below. The words have been chosen to show a minimal contrast in retroflexion; thus the only difference (apart from tone) between pairs of words like ʂaŋ "on" versus saŋ "mulberry tree", or tʂʰa "to differ" versus tsʰa "to rub" is that one has a retroflex initial consonant and one has a dental.

Table 2 represents part of the unconscious knowledge a speaker of Peking dialect has of their language (apart from their English equivalent, of course!). They know, eg, that the word "on" begins with a retroflex fricative, and that the word "mulberry tree" begins with a dental. This is just the same as a speaker of English "knows" that "cot" begins with a /k/ and "got" with a /g/.

example 4 - *continued*

Since the phonological structure of *Putonghua* is based on Peking dialect, *Putonghua* also has two sets of retroflexes and dentals. Now, not all Chinese dialects – rather few, in fact – have such a contrast between retroflexes and dentals. Most, in fact, only have a single, dental, set of phonemes. It can be imagined that this represents one difficulty that a speaker of one of these dialects – eg, Shanghai – would have in learning proper *Putonghua*. They would have to learn, in the same way as when one learns a foreign language, which words have the retroflex phonemes and which do not. The reality, however, is that speakers seldom bother about such a degree of precision in normal conversation. They generally do not bother to observe it, and let the sound patterns of their particular dialect show through. Thus a speaker of Shanghai dialect, which does not have the two sets, is liable to pronounce the words in Table 2, when they are speaking *Putonghua*, in the way shown in Table 3. As can be seen, no difference is made between *Putonghua* ʂaŋ "on" and saŋ "mulberry tree", both being said with an /s/.

TABLE 3 Typical pronunciation, by a Shanghai speaker, of Putonghua words with contrasting retroflexes and dentals

Putonghua word	ʂaŋ	saŋ	tʂ^{h}a	ts^{h}a	tʂaɪ	tsaɪ
	on	*mulberry tree*	*to differ*	*to rub*	*stockade*	*at*
Shanghai speaker speaking *Putonghua*	saŋ	saŋ	ts^{h}a	ts^{h}a	tsaɪ	tsaɪ

The forensic data can now be addressed. Table 4A and B shows several utterances, mostly words, from intercepted phone calls. The examples in B are from a single conversation involving an unknown offender. Those in A are from several conversations involving the same, known suspect.

The suspect is known to have been born and to have grown up in Peking, and to speak typical Peking dialect. He is also known to speak near-fluent Cantonese, from having lived for a long period in Hong Kong. The utterances are transcribed phonemically, and also, to their right, is given a gloss in italics, and below the gloss in brackets the utterances are given in the standard romanisation called *Pinyin*.

Putonghua is a tone language, which means the pitch on a word is just as important a part of its structure as the vowels and consonants. Since the pitch is not relevant here, it has not been transcribed phonetically: it can be assumed that the offender and suspect samples do not differ in tonal pitch. It can also be assumed that both offender and suspect samples have the same voice quality, and that on the basis of this same voice quality, they would sound, to lay ears, as if they had come from the same speaker. (The important term "voice quality" is being used here in its precise sense of contrast with "phonetic quality". Both these are very important terms in the comparison of voice samples, as explained in Rose (2002, Ch 10).)

The prosecution hypothesis is that both the suspect and offender utterances come from the same speaker.

example 4 - *continued*

TABLE 4 Example of auditory-linguistic comparison of forensic voice samples in Putonghua

A. Suspect's samples

	[utterance]	*gloss* (Pinyin)		[utterance]	*gloss* (Pinyin)
1	**tʂ**aʊ	*to look for* (zhǎo)	5	**ʂ**wɔ	*speak* (shuō)
2	**ʂ**ʐ̩wu	*fifteen* (shíwǔ)	6	**tʂ^{h}**ʐ̩	*eat* (chī)
3	**ʂ**ʐ̩fɐn**t**ʂʊŋ	*ten minutes* (shífēnzhōng)	7	**tʂ**ʐ̩taʊ	*know* (zhīdao)
4	aᴵ**ʂ**ʐ̩lɪɔʊ	*twenty six* (èrshíliòu)	8	ni **ɕɛn** tɕ^{h}y ba	*better you go first* (nǐ xiān qù ba)

B. Offender sample

	[utterance]	*gloss* (Pinyin)		[utterance]	*gloss* (Pinyin)
1	lu**s**aŋ	*on the way* (lùshang)	5	**s**wɔ	*speak* (shuō)
2	**s**zxɔʊ	*time* (shíhou)	6	**l**ali	*where?* (nǎlǐ)
3	**s**zfɐn**t**ʃʊŋ	*ten minutes* (shífēnzhōng)	7	**l**a	*in that case* (nà)
4	**tʃ**ɔjɔʊ	*approximately* (zuǒyòu)	8	tɐŋ iɕa **ɕɛn** a	*wait a bit first* (děng yíxià xiān a)

The suspect's examples 1-7, in Table 4A, show a strict and consistent observance of the contrast between retroflexes and dentals: all words that are supposed to have retroflexes, like "speak" and "ten" and "eat" have them. This is what would be expected, given that we know that the suspect is a native speaker of Peking dialect.

The offender's examples 1-5 in Table 4B, on the other hand, do not show this contrast: they have dentals in words which in Peking dialect, and proper *Putonghua*, should have retroflexes. This can best be seen in the words for "speak" and "ten (minutes)" which occur in both offender and suspect samples.

The offender's sample also differs from the suspect's samples in three other ways – two phonological and one syntactic:

(1) The offender's sample has a lateral (an l sound) in the words "where" and "in that case", which should in *Putonghua* or Peking dialect have an n: nali and na.

example 4 - *continued*

(2) The offender's sample has a palato-alveolar affricate [tʃ] in the word "approximately" which should, in *Putonghua* and Peking dialect, have a dental: **ts**wɔjɔʊ.

(3) The offender's sample has a postverbal time adverb ɕɛn "first" in the utterance "wait a bit first", compared to the suspect's preverbal position in "better you go first". The preverbal position is again typical of *Putonghua* and Peking dialect.

These three features, as well as the failure to make a retroflex/dental contrast, are typical of a speaker of Cantonese speaking *Putonghua*. Cantonese is one of the many Chinese dialects which does not distinguish a retroflex from a dental set of phonemes; does not contrast /n/ and /l/; has palatal allophones of dental affricate and fricative phonemes before rounded vowels (/ts/ → [tʃ]; /s/ → [ʃ] / _ + rounded vowel), and postposes the time adverb "first".

The proper forensic evaluation of the auditory-linguistic differences in Table 4A and B involves estimating their LR. Generally, this is the ratio of the probability of observing these auditory-linguistic differences between the samples assuming that they were spoken by the same speaker, to the probability of observing the differences assuming that they were spoken by different speakers. Specifically, the additional knowledge can also be incorporated concerning the linguistic competence of the suspect as a native speaker of Peking dialect who is near fluent in Cantonese; and the fact that the offender's speech is typical of the *Putonghua* of a native Cantonese speaker. The questions now become:

- What is the probability of observing the differences between the samples, assuming that both samples have come from the suspect (who is a native speaker of Peking dialect near fluent in Cantonese)?
- What is the probability of observing the differences between the samples, assuming that both samples have come from different speakers, the suspect being a native speaker of Peking dialect near fluent in Cantonese, and the offender being a native speaker of Cantonese speaking Mandarin?

Consider first the prosecution hypothesis: what is the probability of observing the evidence assuming that the same speaker was involved? It is important to realise that it is not automatically zero, because it is possible to imagine cases where a single speaker might sometimes observe the difference between retroflex and non-retroflex words, and sometimes not. For example, if a speaker of a retroflex dialect was speaking *Putonghua* to a speaker of a non-retroflex dialect who was not making the distinction, it *might* just be possible for the former to signal solidarity or intimacy with their interlocutor by not making the distinction.

This would be an example of what is termed "convergence" in speech behaviour, and is a well-documented phenomenon (although not for this particular feature). (A speaker can also "diverge" in her or his speech behaviour from an interlocutor to emphasise, for whatever reason, the difference between them.) In the present case, convergence can be ruled out, because the other speaker in the offender sample actually observes the retroflex distinction (and also clearly distinguishes /l/ from /n/ etc.)

Another instance of the same speaker differing in her or his retroflexion behaviour might be if the speaker was of a non-retroflex dialect, but the particular situation required observance of the distinction – eg if the speaker was desiring to impress with her or his knowledge of *Putonghua*. This possibility can be ruled out in this case, since it is known that the suspect is a native speaker of Peking dialect which makes the retroflex distinction.

example 4 - *continued*

Finally, since it is known that the suspect has excellent command of Cantonese, it might just be possible that his Cantonese competence is showing through in his native Peking Dialect. Not enough is known about the phonological behaviour of bilinguals or near bilinguals to accurately assess the probability of this. Since it is common for native accent phonology to show through in a second language, but not vice versa; and since – more importantly – there is no incidence of such behaviour in any of the suspect's speech samples, this must also be accorded a low probability.

We are therefore left with a rather low probability of the evidence assuming the samples come from the same speaker.

Turning now to the refined alternative hypothesis: all the linguistic traits in the offender's speech are very typical of the Mandarin of Cantonese speakers, such that one would expect that they could be observed in a large proportion of Cantonese speakers speaking Mandarin. Thus the probability of observing them assuming that the offender is a native Cantonese speaker is very high.

It should be clear from the above that the linguistic differences between the samples would be far more probable under the alternative, defence hypothesis, than under the prosecution hypothesis. In cases like this, it is difficult to quantify the ratio because little quantified data are available. One might estimate a soft probability of perhaps 80% to 90% for the denominator, on the assumption that most native speakers of Cantonese could be expected to show the observed features when they try to speak Mandarin. But, as already pointed out, although it will clearly be very small, it is difficult to give even an approximate value for the numerator: it might be 1%, or 0.1% or less. In cases like this – and this will happen frequently with auditory linguistic analysis – it would be probably better to say that, at the worst the evidence would be enough to constitute support for reasonable doubt that different speakers were involved (and thus to counter the similarity in voice quality between the samples); and at best it would offer strong support that different speakers were involved. As usual, the most important thing is that the bases for the assumptions are made explicit.

Auditory analysis of non-linguistic phonological features

[99.560] In the same way as the IPA is available for the cross-linguistic description and comparison of speech sounds, an analytical framework exists, developed by John Laver and described in Laver (1980), for the auditory description of voices.

According to Laver's framework, the researcher can describe a voice with respect to a constellation of auditory features. For example, a voice might be described as having a "narrow pitch range" (self-explanatory), and a "nasal velopharyngeal setting" (the speaker sounds as if he or she is talking with soft-palate down as a default position, making the speaker sound as if he or she were talking through the nose all the time).

Although most forensic phoneticians can be expected to pay attention to and make use of general auditory non-linguistic phonological features, like overall pitch range or overall pitch height, not many have been explicitly trained in the Laver system to the extent that they can describe a voice in detail with respect to all the relevant features.

Since the quantitative nature of between- and within-speaker variation in auditory non-linguistic features is generally not well known, their evidentiary strength as forensic-speaker identification features will often not be very great.

Acoustic analysis

[99.570] Acoustic analysis is the means by which acoustic features are derived from properties of the speech wave, and are quantified with the aid of a computer. Sound is rapid fluctuations in air pressure, and when someone speaks, these pressure fluctuations radiate outwards from the mouth, throat and nostrils. The acoustic energy in the pressure fluctuations can be transduced by a microphone at a distance from the speaker and converted into other types of energy for acoustic analysis by computer.

The knowledge and assumptions motivating the use of acoustic analysis for TFSI comparison are as follows. A speaker radiates acoustic energy when he or she speaks, and this acoustic energy is relatively easily quantifiable in terms of acoustic features. It is known that the acoustic output of a speaker's vocal tract is a unique function of her or his vocal tract anatomy, thus the acoustic energy radiated by a speaker carries the imprint of the vocal tract that produced it. Since speakers are assumed to differ in their vocal tract anatomy, and the way they use it to produce speech, acoustic features can be extracted from the radiated acoustics which will reflect this, and offer means of discriminating between acoustics produced by the same vocal tract and acoustics produced by different vocal tracts.

These assumptions are correct, but there are non-trivial problems associated with making use of them forensically. First, human vocal tracts are not invariant, but highly deformable, and speakers make routine linguistic, paralinguistic and extralinguistic use of this plasticity. The organic state of an individual's vocal tract can also differ, depending eg on the individual's health. All this means that a speaker's acoustic output, although a function of the speaker's vocal tract, will show a range of values, potentially overlapping with other speakers' ranges, according to how much their (ie, the individual speaker's and other speakers') vocal tract varies. Thus, although the radiated acoustics contain information on the size and shape and condition of the vocal tract that produced them, this information is also convolved (ie, mixed up in a complex way) with acoustically encoded information on the particular speech sound being produced; the particular emotion that is being conveyed; and perhaps also the particular habitual articulatory setting the speaker selects. (This is actually a very crude, oversimplifying, characterisation of the information content in a voice: the reader is referred to Rose (2002, Ch 10) for a full discussion of what information is conveyed, acoustically and otherwise, in a voice.) All the above factors contribute to within-speaker variation in acoustics.

Second, there is not limitless variation in the dimensions of the human vocal tract. This means that the ranges of acoustics output by vocal tracts are also not big, and thus not necessarily easy to discriminate. Moreover, one is not concerned forensically with the total range of acoustic differences potentially output by the human vocal tract, but the subset thereof that will present the greatest difficulty. This is because the acoustics output by two vocal tracts of very different dimensions, like a male and a female, or a male and a child, are going to be so different that the voices associated with them are highly unlikely to be confused in the first place. In TFSI, rather, one is obviously going to have to be able to discriminate between similar-sounding speakers (speakers that sound similar to the lay ear). And one of the reasons that speakers sound similar is that some of their acoustics are similar because of the similarly dimensioned vocal tracts that produced them. Thus in TFSI one needs to be able to discriminate within a very much smaller range of acoustics.

Traditional versus automatic approaches

[99.580] There are two very different approaches to the acoustic comparison of forensic speech samples. These can be referred to as "traditional" and "automatic". In a traditional acoustic analysis, the acoustic speech signal is treated analytically as the output of a vocal tract that is executing all the complex gestures that are required to make speech sounds.

Thus in a traditional acoustic analysis, features are extracted that relate in a relatively straightforward way to aspects of speech production, like what the speaker is doing with her or his tongue, or vocal cords, to produce a particular speech sound. The traditional approach derives originally from phoneticians' and speech engineers' attempts to find out what the acoustic correlates of speech sounds were: what made a *oo* vowel sound like a *oo* vowel; what made an *l* sound different from an *r* sound; what made the rising pitch of a question different from the falling pitch of a statement etc.

The theory which informs the traditional acoustic analysis of speech relates the radiated speech acoustics to the configuration and dimensions of the vocal tract that produced them and is called the "acoustic theory of speech production", or "source-filter theory". It is a received scientific theory in the strongest, falsifiable sense of that word, and all forensic phoneticians should be expected to know it in detail. This section invokes some aspects of source-filter theory, but for a detailed account of its application to vowels, Rose (2002, Ch 8) needs to be consulted.

In an automatic approach, the acoustic speech signal is treated purely statistically, as a time-varying signal, with no special attention to any particular linguistically meaningful subpart or deliberate production thereof. The underlying theory is taken from signal detection. The origins of the automatic approach lie in attempts to get computers to perform speech and speaker recognition automatically.

Both traditional and automatic approaches differ considerably with respect to the two forensically significant characteristics of "interpretability" and "strength", and it is important for those involved with evaluating expert testimony to understand the difference. The following sections exemplify and clarify the difference using two features that the reader is likely to encounter in forensic speaker identification reports – "formants" (traditional) and "cepstrum" (automatic). A second traditional feature which is very commonly used in the forensic comparison of speech samples – fundamental frequency – is then briefly introduced.

Formants

[99.590] One of the traditional acoustic parameters, or sets of parameters, that are commonly used in the forensic comparison of speech samples are formant frequencies, in particular the formant frequencies of vowels. During the production of a vowel, air is being expelled at a fairly constant rate through the vocal tract. The movement of air is initiated by the lungs, and its direction is outwards. Because vowels are usually voiced, this pulmonic egressive airstream is interrupted at the larynx by the vibratory action of the vocal cords. The result is that during the production of a vowel a sequence of high-velocity jets of air is injected into the supralaryngeal vocal tract. The effect of these high-velocity jets is to cause the air already present in the supralaryngeal vocal tract to vibrate, and it vibrates at some frequencies with greater ease than at others.

The frequencies at which the air in the supralaryngeal vocal tract vibrates with the greatest ease are called its "resonant frequencies", or "formant frequencies". The lowest two or three formant frequencies are the primary determinants of vowel quality. They are what makes a vowel sound like an *ee*, an *aah* or an *oo*. Formants, and their frequencies, are often referred to as F1 ("eff one"), F2, F3 etc, or the first, second, third formant ... , with F1 indicating the formant with the lowest frequency. An ensemble of formant frequencies for a given sound is often called its "F-pattern".

The formant frequencies/F-pattern are determined by the size and shape of the speaker's supralaryngeal vocal tract. The size of the tract roughly reflects the speaker's stature: taller speakers will eg generally have longer tracts than shorter speakers; females will have shorter tracts than males. The shape of the supralaryngeal vocal tract depends on the vowel being produced: different shapes result in different sounding vowels. To make these different supralaryngeal vocal tract shapes, a speaker will position her or his tongue body according to

coordinates running from front to back and high to low, and will also control the size of the mouth opening, making it smaller by rounding the lips, and more open by spreading the lips.

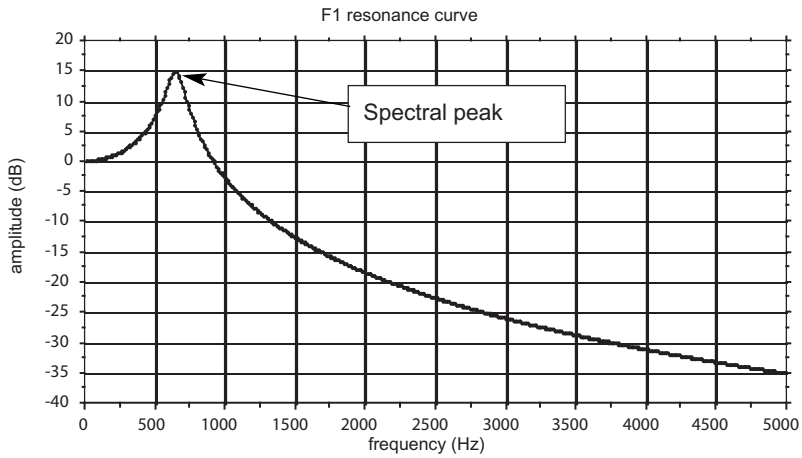
Thus for an *ee* vowel, as in the British and American English word *lea*, the speaker will put the body of the tongue high and forward in the mouth, thus making the oral cavity relatively short, with a small cross-sectional area, and the pharynx relatively long, with a large cross-sectional area. The speaker will also have the lips fairly spread. For an *or* vowel, as in the British English word *law*, the speaker will locate the body of the tongue low in the mouth and at the back, so that it is the pharynx that is relatively narrow and short and the oral cavity relatively wide and long. In addition, the lips will be rounded and somewhat protruded. A speaker of General American English will also have the tongue body low and back for this vowel, but the lips are likely to be unrounded, not rounded. For a *oo* vowel, as in *woo*, a speaker will put the tongue body in a high back position, creating oral and pharyngeal cavities of about equal size, and round the lips. Such an articulation was pointed out in the vocal tract x-ray of Figure 4 at **[99.1120]**.

The positioning of the tongue and lips provides the set of conventional auditory-phonetic descriptive terms for vowels which refer to the height and backness of the tongue body and whether the lips are rounded or not. Describing a vowel, eg like *ee,* as "high front and unrounded" means that it sounds as if it has been produced with the tongue body high in the front of the mouth, and with the lips unrounded. With the same assumptions, *oo* is described as "high back rounded" etc. Height, backness and rounding do not, by a long way, exhaust the set of parameters that languages use in the production of vowels; other important parameters will be found in Rose (2002, Ch 6).

Formant resonance curves

[99.600] Figure 5 shows what a vowel formant looks like. It shows the first formant, or F1, at effectively a single instant in the diphthong in the second syllable of a single token of the word "hello", as spoken by a young male from Adelaide. The word "hello" consists of two syllables: *he-* and *-lo*. In Australian English, the vowel in the second syllable *-lo* is a diphthong. A diphthong is a movement, within a single syllable, from an initial vowel target (T1) to a second vowel target (T2). Figure 5 shows the first formant at T1, ie first diphthongal target. Like all vowels, this target is specified in terms of presumed height and backness of the tongue body and rounding of the lips. This particular target sounds lowish, central and unrounded, somewhat like the vowel in Australian or British English "hut".

FIGURE 5 Spectrum of first formant resonance for the first target of the second-syllable diphthong in a single token of Australian "hello"



Representations of the type shown in Figure 5 are called "spectra". The concept of a "spectrum" is one of the most important things in acoustic phonetics, and, indeed, in hard science in general. A spectrum shows distribution of energy. The horizontal axis in Figure 5 shows frequency, which is quantified in terms of so-many repeats per second, or Hertz (Hz). The frequency axis runs from 0 Hz at the left to 5000 Hz at the right. The vertical axis shows amplitude, or amount of energy. It is quantified in decibels (dB).

A spectrum can show either real or potential distribution of energy; this one shows potential distribution. That is, Figure 5 shows exactly how much energy would be present at what frequencies for this formant in this sound. Its main feature is the single spectral peak located a little above 500 Hz. The actual frequency location of the spectral peak is called the "formant centre frequency", or just "formant frequency", and this is the important quantity in question. The formant (centre) frequency is often abbreviated as F_n, where *n* is the number of the formant. Here, the actual value was 651 Hz, so in this example F_1 = 651 Hz. The amplitude of the spectrum at this frequency of 651 Hz can be visually estimated at about 15 dB. The amplitude of the spectrum is also referred to as "spectral amplitude", "spectral height", or just "amplitude" or "height".

What the curve in Figure 5 actually represents is quite complicated. It was mentioned above that the air in the supralaryngeal vocal tract vibrates in response to the energy input from the pulses of air generated by vocal cord activity on the outgoing airstream, and that the supralaryngeal air vibrates at some frequencies with greater ease than at others. For "greater ease" now substitute "at greater amplitudes", and it can be appreciated that Figure 5 shows the amplitudes at which the air in the supralaryngeal vocal tract vibrates for the frequency range from 0 Hz to 5000 Hz in this particular sound. (Another term for response is "resonance", and so Figure 5 shows the spectrum of a resonance curve.)

The representation in Figure 5 is abstract, however: it shows the amplitudes at which the air *would* vibrate, if energy *were* present at that frequency. It says, eg, that if energy were present at 1000 Hz, the response of the air in the vocal tract would be to vibrate at about -3 dB for that frequency. (The reason for this is part of source-filter theory and is explained in Rose (2002, Ch 8).) The formant centre frequency can now be understood as the frequency at which the air in the vocal tract would show its maximum response, or would resonate with maximum amplitude.

It is also possible to be more specific, for this particular sound, about the part of the vocal tract involved. For males, the first formant is typically associated with the pharynx in sounds like this, so Figure 5 shows how the air *in the pharynx* would respond at different frequencies.

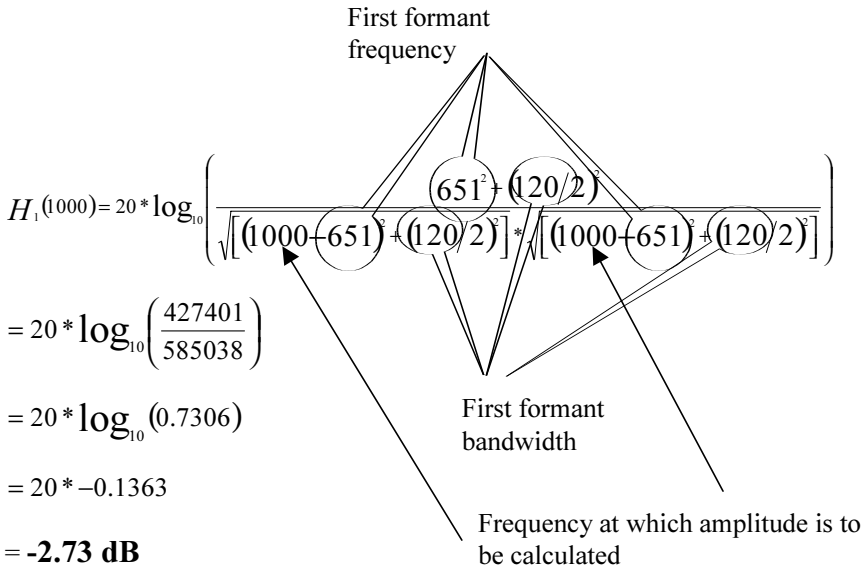
The other important quantity of a resonance curve, apart from the formant centre frequency, is the "formant bandwidth". The formant bandwidth quantifies the fatness of the formant peak from side to side. It is actually the width, in frequency, of the formant 3 dB down from its peak. Looking at the curve in Figure 5, it can be appreciated that at 3 dB down from peak – so at an amplitude of about (15 dB - 3 dB =) 12 dB – the body of the formant seems very roughly to be about 200 Hz across. The bandwidth actually reflects the amount of energy loss: the greater the loss the wider the bandwidth. For this particular formant in this particular sound, the bandwidth is considerably larger than it would normally be. This is because of greater than normal energy losses connected with the *h* sound at the beginning of "hello", and the fact that these *h*-related losses will be manifested primarily in the response of the air in the pharynx.

Formula 6

$$H_n(f) = 20 * \log_{10}\left(\frac{F_n^{\,2} + (B_n/2)^2}{\sqrt{\left[(f - F_n)^2 + (B_n/2)^2\right]} * \sqrt{\left[(f + F_n)^2 + (B_n/2)^2\right]}}\right)$$

The spectrum of the resonance curve in Figure 5 is plotted with a formula that uses only two figures as variables: the formant centre frequency and the formant bandwidth: see Formula 6. The $H_n(f)$ to the left of the equals sign stands for "the amplitude, or height, H of the spectrum of the resonance curve for the nth formant as a function of frequency f". The part of the formula to the right of the equals sign specifies how this amplitude is calculated from just the two variables of formant centre frequency and formant bandwidth. The former is represented by F_n and the latter by B_n. To show how it works, some actual values have been inserted in Formula 7. These values show how the amplitude of the resonance curve is found corresponding to a frequency of 1000 Hz. The reader should first try to estimate this value visually from Figure 5, where it can be seen that at a frequency of 1000 Hz, the amplitude is between -5 dB and 0 dB – say about -3 dB.

Formula 7



In the set of values at Formula 7, $H_n(f)$ on the left is now replaced by H_1(1000 Hz), since it is desired to determine the spectral amplitude, or height, of the first formant resonance curve when the frequency is 1000 Hz. The centre frequency for the first formant is taken as 651 Hz and its bandwidth as 120 Hz, thus $F_n = F_1 = 651$ Hz and $B_n = B_1 = 120$ Hz. The frequency value at which the amplitude is being calculated, namely 1000 Hz, also has to be substituted in two positions on the right-hand side of the equation. It can be seen from the working at Formula 7 that the amplitude of the spectrum for this resonance curve corresponding to a frequency of 1000 Hz is actually -2.73 dB, nicely agreeing with its visual estimate.

In order to plot the whole resonance curve for this formant, the amplitudes corresponding to the complete range of frequencies from 0 Hz to 5000 Hz would have to be calculated with the formula. Thus the calculation would have to be repeated with values of f = 0 Hz, f = 1 Hz, f = 2 Hz etc. If lesser resolution were required, f values can be taken at greater intervals, eg, f = 0 Hz, f = 10 Hz, f = 20 Hz etc.

Thus it can be seen that the combination of the low frequency location of the formant centre frequency, at 651 Hz, with its bandwidth of 120 Hz results in a resonance curve that drops off considerably at frequencies higher than the centre frequency, so that at 5000 Hz, eg, the energy is some (15 dB - -35 dB =) 50 dB down from peak.

FIGURE 6 Spectrum of F2 resonance curve for the first target of the second-syllable diphthong in a single token of Australian "hello"

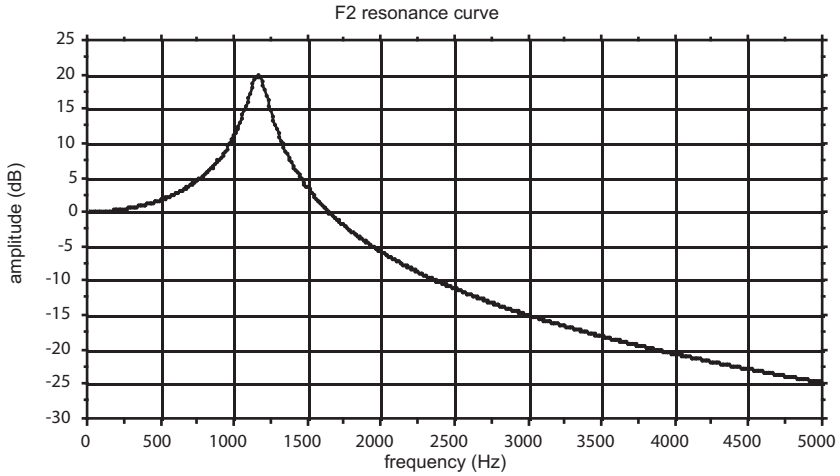
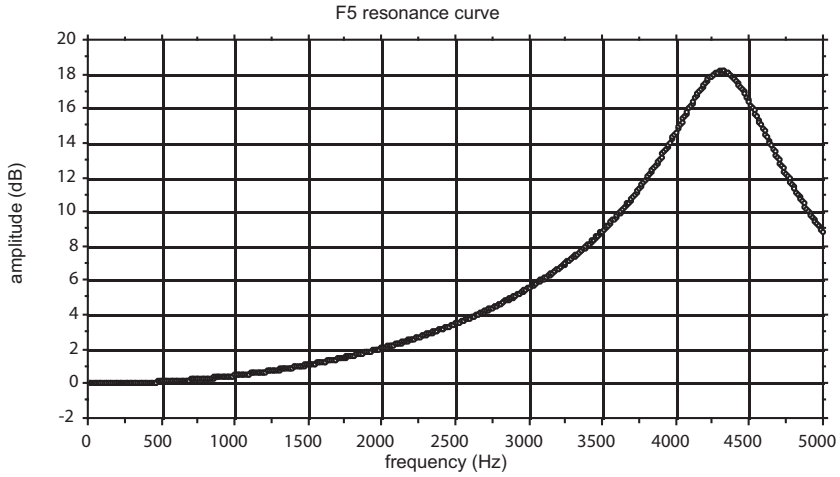


Figure 6 shows the resonance curve for the second formant in this sound. The frequency scale is the same as in Figure 5, but the amplitude scale is slightly different to accommodate the slightly higher peak amplitude value (ca 20 dB) of the second formant. The centre frequency of F2 can be seen to be somewhat higher than that of F1 in Figure 5. The reader should try to roughly estimate by eye this centre frequency, say to within 100 or 200 Hz. (It is about 1200 Hz.) The second formant in sounds like this is typically an oral cavity response in males, so this curve represents the way the air in the oral cavity would vibrate.

Finally, in order to show what a formant resonance curve looks like with a wider bandwidth, the resonance curve for F5 in this sound is shown in Figure 7. Once again the amplitude scale has been changed to accommodate the resonance curve peak amplitude. F_5 here is 4320 Hz, and B_5 (fifth formant bandwidth) is 536 Hz, and it can be seen that the formant is, indeed, fatter at 3 dB down from peak: at an amplitude of about (18 dB - 3 dB =) 15 dB it appears to be about 500 Hz wide.

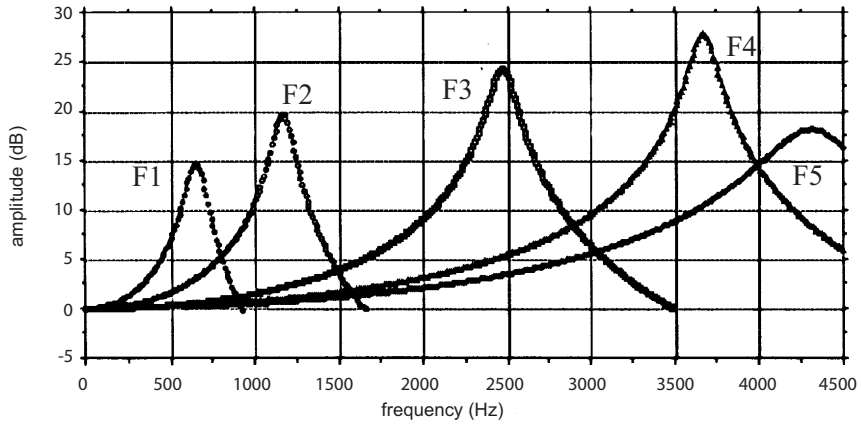
FIGURE 7 Spectrum of F5 resonance curve for the first target of the second-syllable diphthong in a single token of Australian "hello"



Combining the formant resonance curves

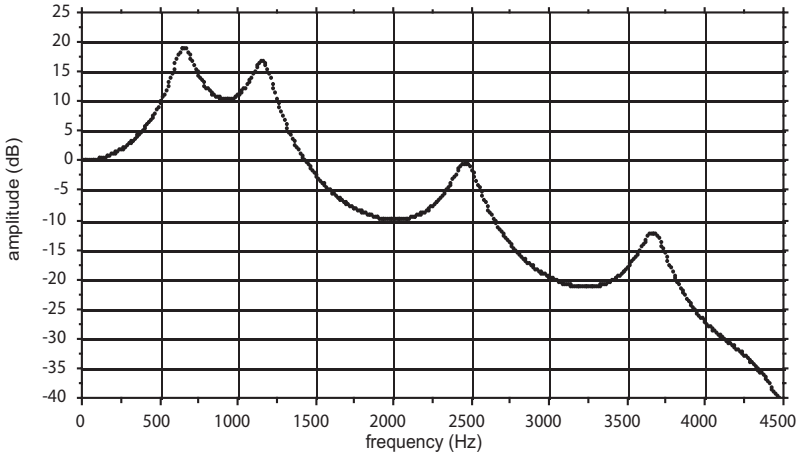
[99.610] The air in the vocal tract vibrates at all the frequencies specified by all the formant resonance curves effectively simultaneously. In order to represent this, the individual resonance curves for all the formants can be plotted in superposition, as shown in Figure 8. (Due to software memory limitations, the values of the curves were not plotted below 0 dB, and the limit of the upper frequency range was lowered to 4500 Hz.) The reader should try to identify the curves for F1, F2 and F5, which were individually presented above, by their formant centre frequencies and the amplitude at these frequencies.

FIGURE 8 Superposed resonance curves for F1 to F5 for the first target of the second-syllable diphthong in a single token of Australian "hello"



The individual formant resonant curves have to be combined mathematically into a single curve in order to show precisely how the air in all the vocal tract would vibrate for the sound. The curves are combined by adding, for each separate frequency, the amplitude values of each formant curve, as derived by using Formula 6: see **[99.600]**. The spectrum of the resulting combined resonance curve for all five formants is shown in Figure 9.

FIGURE 9 Individual resonant curves for F1 to F5 in the first target of the second-syllable diphthong in a single token of Australian "hello" combined to form an overall resonance spectrum



In order to show how the values are added, and the overall resonance curve in Figure 9 derived, some examples are given in Table 5. Three values have been selected on the frequency axis which are nearest to the centre frequencies of the first three formants, namely 650 Hz (F1), 1160 Hz (F2) and 2460 Hz (F3). These are shown along the top of Table 5. In the columns are shown the amplitude values for each of the five formants at these frequencies. Thus at 1160 Hz, the first formant curve had an amplitude of -6.74 dB. This can be seen in Figure 5: at just over 1000 Hz, the amplitude appears to be about -5 dB. At 1160 Hz the second formant is at its peak – Figure 6 shows that it is about 20 dB at this frequency value. Table 5 shows that at 1160 Hz the F2 curve has, indeed, an amplitude of 19.75 dB. Adding up the amplitude values for all five formants at 1160 Hz gives a value of 16.72 dB, and it can be seen in Figure 9 that at 1160 Hz (the second spectral peak, or F2 peak, in the overall envelope) the amplitude indeed lies at about 17 dB. At the third spectral peak, or F3 (at 2460 Hz), the overall amplitude can be seen in Figure 9 to be just under 0 dB, and the actual value from Table 5 can be seen to be -0.35 dB.

TABLE 5 Summing amplitude values of individual resonance curves to obtain the overall resonance spectrum at given frequencies

at freq =	650 Hz	1160 Hz	2460 Hz
F1 amp is ..	14.76	-6.74	-22.4
F2 amp is ..	3.20	19.75	-10.77
F3 amp is ..	0.62	2.16	24.31
F4 amp is ..	0.28	0.91	5.17
F5 amp is ..	0.20	0.64	3.34
combined amp (dB)	**19.06**	**16.72**	**-0.35**

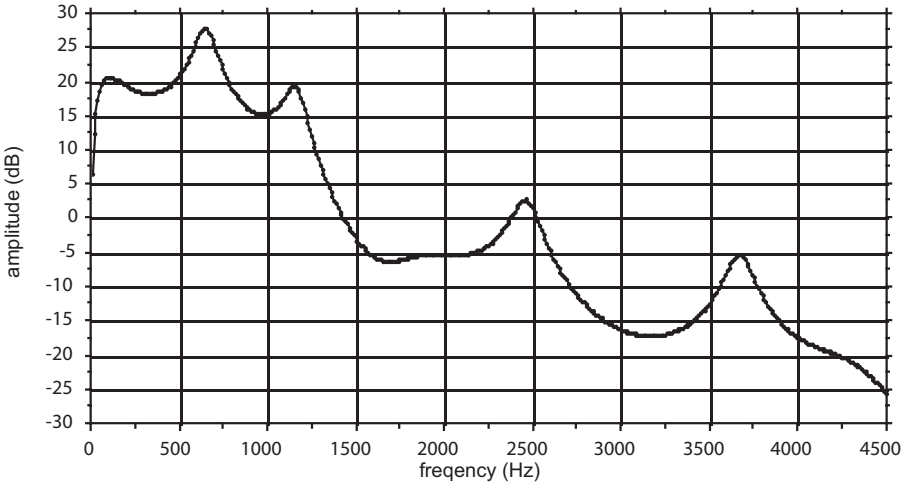
There are several other factors besides the supralaryngeal resonances which have an effect on the overall spectral shape of the sound as it would be radiated from the speaker, and these also have to be taken into account before the spectral shape corresponds to that which might be picked up by a microphone at a distance from a speaker and used in forensic comparison.

The actual energy for the sound comes from the vibrating vocal cords, and this needs to be taken into account. Moreover, the sound is usually recorded at a distance from the speaker's lips, and radiation from the head has an effect on the spectral profile. The effect on the overall spectrum of the formants higher than F5 is another factor that needs to be considered.

Finally, in this particular word ("hello") there is often an additional specific factor to be incorporated. Resonances are not all supralaryngeal: with certain settings of the vocal cords, the air in the trachea, below the vocal cords, is also set into vibration. This gives rise to so-called "sub-glottal resonances", as well as "sub-glottal zeros" (which are frequencies at which energy is absorbed). One such sound typically associated with sub-glottal resonances and zeros is the *h* at the beginning of "hello", and the effect can last through the whole word, certainly as far as the first diphthongal target. In this particular token of "hello", there was a clear sub-glottal resonance at about 1850 Hz, and a zero about 200 Hz below it. The resonance curves associated with these sub-glottal effects – a curve with a centre frequency at about 1850 Hz, and an inverse curve subtracting energy at 1650 Hz, both with wide bandwidths of about 500 Hz – also have to be incorporated. The overall effect is to introduce a very low amplitude spectral peak between the F2 and F3 peaks.

The effect of introducing all these factors is shown in Figure 10. It can be seen that the formant peaks are still clear, and their centre frequencies are the same, but in comparison to Figure 9, there is now some more energy present in the 500 Hz range below F1. This represents part of the energy from vocal cord vibration. The profile between the F2 and F3 peaks is also slightly different in Figure 10, which now shows a very slight dip in the spectrum just above 1500 Hz and a flattening out of the profile above it. These reflect the addition of the sub-glottal zero and resonance associated with the *h*.

FIGURE 10 Radiated spectral envelope for the first target of the second-syllable diphthong in a single token of Australian "hello"

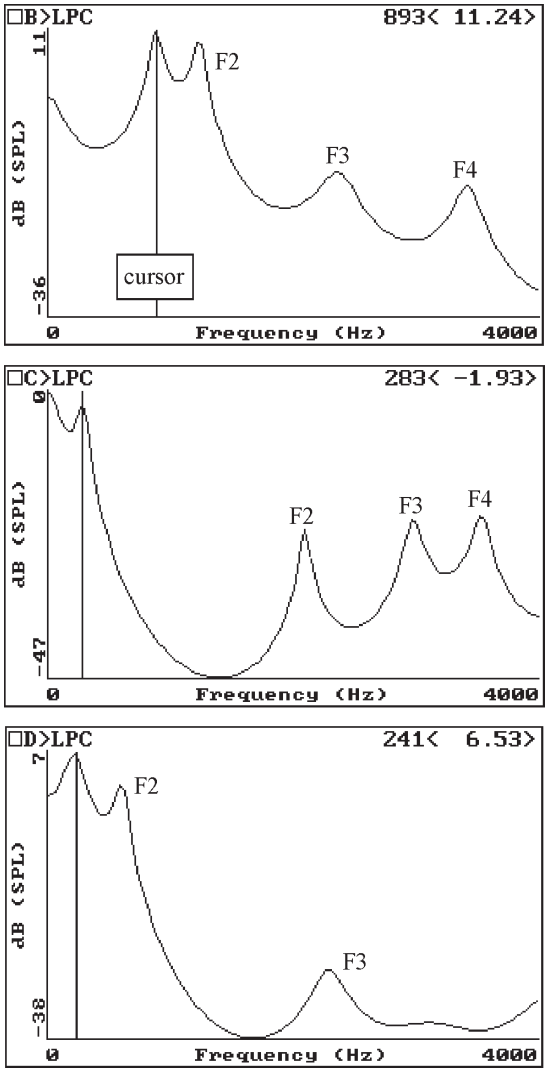


Formant structure in different vowels

[99.620] The previous section has shown how the radiated spectrum for a particular vowel sound, with its formant centre frequencies, is constituted from a set of formant resonance curves in combination with other factors like the energy input from the vocal cord vibration. Spectra for different vowels are now examined to demonstrate how differences in vowel quality are determined by the frequencies of the lowest two or three formants.

Figure 11 shows the spectra of three Cantonese vowels spoken by a young male when reading out a story. The vowels are a low central unrounded vowel [a] (similar to the vowel in the Australian English word "hard"); a high back rounded [u] (similar to the vowel in "woo"); and a high front unrounded [i] (similar to the vowel in "yee"). The latter two vowels occurred in the word "gusih" ("story") [k**u**s**i**] (pronounced somewhat like *goosee*), and the first occurred in the word "pa" ("to fear") [p^{h}**a**] (pronounced somewhat like *pa*). Cantonese is a tone language, but the pitch of the tones is not relevant for these examples and has not been shown in the phonetic transcription.

FIGURE 11 Linear prediction spectra for three Cantonese vowel tokens spoken by a male (CM1)



Top = [a], middle = [i], bottom = [u]. The vertical cursor has been placed through the first formant centre frequency (F1) in each vowel.

The panels of Figure 11 show the distribution of the acoustic energy at a single instant in each of the three vowels. This instant was near the middle of the vowel in each case. The top panel shows the spectrum for the vowel [a]. Below it is [i] and at the bottom is [u].

The axes are the same as in the spectra already presented. Frequency, in Hz, runs horizontally from left to right, from 0 to 4000 Hz. In theory, the frequency range is infinite, but in practice, especially forensic-phonetic practice, there is seldom anything of use above values between about 3000 Hz and 3500 Hz. This is because frequencies above these values will be severely attenuated by the telephone transmission. Amplitude, in dB, is shown vertically.

The wavy line in each panel, already familiar from previous examples, shows the spectral amplitude. Although the spectra look very similar to the radiated spectrum demonstrated above for the sound in "hello", these spectra have been produced by a different method called "linear prediction" (LP). Linear prediction is a complicated, but very common, computerised signal-processing technique, the aim of which is to estimate the vocal tract resonant frequencies from a short portion of acoustic wave-form. It is often used forensically to estimate formant frequencies.

The spectra in Figure 11 estimate the vocal tract resonant frequencies for these vowels at the particular instant they were sampled. In each spectrum, the vertical cursor has been placed so that it goes through the spectral peak with the lowest frequency, ie, F1. The actual frequency, and the amplitude of the peak, can be read off in the top right hand corner of each panel. Thus the lowest spectral peak in [a] – F1 – is at 893 Hz and has an amplitude of 11.24 dB. Although it is a little hard because there are no frequency intervals indicated, the reader should again try to visually estimate the frequency of F1 in the other vowels, before reading them in the top right-hand corner.

Looking now at the spectrum for [a] in the top panel of Figure 11, it can be seen that the lower half of the spectral frequency range – from 0 Hz to 2000 Hz – contains two high narrow spectral peaks close together (the lowest has the cursor through it). These peaks in spectral energy are the peaks of the first two individual formant resonance curves, F1 and F2: the second formant has been marked as F2. The upper half of the spectral frequency range – from 2000 Hz to 4000 Hz – contains two more formants – F3 and F4 – but they do not have such a great amplitude, have greater bandwidths, and are more widely spaced in frequency. The overall spectral profile of the [a] is clearly similar to that of the radiated spectrum in the first diphthongal target in "hello" (see Figure 10). This is because both vowels are very similar in quality.

The reader should now identify the formants of the vowels in the other two panels, and especially note the position of F1 and F2 in the frequency range from 0 to about 2000 Hz.

It can be seen that the three vowels [a], [i] and [u] differ in the location of their first three formants. It is, in fact, the centre frequency of the first two, or sometimes three, formants that are the primary determinants of perceived vowel quality. An [i] sounds like an [i], and not an [u] or an [a], because it has a low F1 and a relatively high F2; a [u] sounds like an [u] because it has a low F1 and F2; and an [a] sounds like an [a] because its F1 and F2 are close together in a mid-frequency location. (In talking about vowel quality, it is conventional to assume a reference range, for males, of about 0 Hz to 2000 or 3000 Hz, such that values towards 2000 Hz or 3000 Hz count as high in the range, and values towards 0 Hz count as low.)

It was pointed out above that formant frequencies are determined articulatorily, by the shape the speaker makes her or his supralaryngeal vocal tract assume. The supralaryngeal shape determines the relative positions of the formant frequencies in different vowels. However, it was pointed out that the formant frequencies are also determined by the size of the supralaryngeal vocal tract. Since different speakers can, of course, have different-sized vocal tracts, this is obviously important for TFSI. How vocal tract size is reflected in the acoustics is now addressed.

F-pattern and vocal tract length

[99.630] Formant frequencies are a function of the overall length of a speaker's supralaryngeal vocal tract. This can be most easily demonstrated by means of a formula that specifies the formant frequencies for a vowel like that in the British or Australian English word "herd". This is the least complex of vowels because, unlike [i] and [u], eg, it does not involve differential oral and pharyngeal cavity shaping. Instead it is made with a supralaryngeal vocal tract that approximates a tube of uniform cross-sectional area. Consequently its F-pattern is the least complex and can be considered to be effectively the same as that of a tube of uniform cross-sectional area closed at one end (the larynx end). This vowel, which is transcribed [ə], is special enough to have its own name – "schwa".

Formula 8

$$F_n = (2n-1)\cdot\frac{C}{4l}$$

Formula 8 gives the resonant frequencies (F-pattern) of a tube of uniform cross-sectional area closed at one end. This formula says that the frequency of any given formant with such a tube is determined by two things: the speed of sound (*C*) and the length of the tube (*l*). *C* can be conventionally taken as 35,000 cm/sec. Since the length of the tube *l* is equivalent to the length of the supralaryngeal vocal tract, let us assume (again conventionally, because it gives results easy to remember) an individual with a tract length of 17.5 cm. Thus a schwa said by a speaker with a 17.5 cm long vocal tract will have a first formant frequency F_1 of:

([2*1] - 1) * [35,000 / [4*17.5]]
= (1 * [35,000 / 70])
= (1* 500)
= **500 Hz**.

In the same way, F_2 will be:

([2*2] - 1) * [35,000 / [4*17.5]]
= 3 * [35,000 / 70]
= **1500 Hz**;

F_3 will be:

([2*3] - 1) * [35,000/ [4 * 17.5]]
= 5 * [35,000 / 70]
= **2500 Hz**;

and F_4 will be:

([2*4] - 1) * [35,000/ [4 * 17.5]]
= 7 * [35,000 / 70]
= **3500 Hz**.

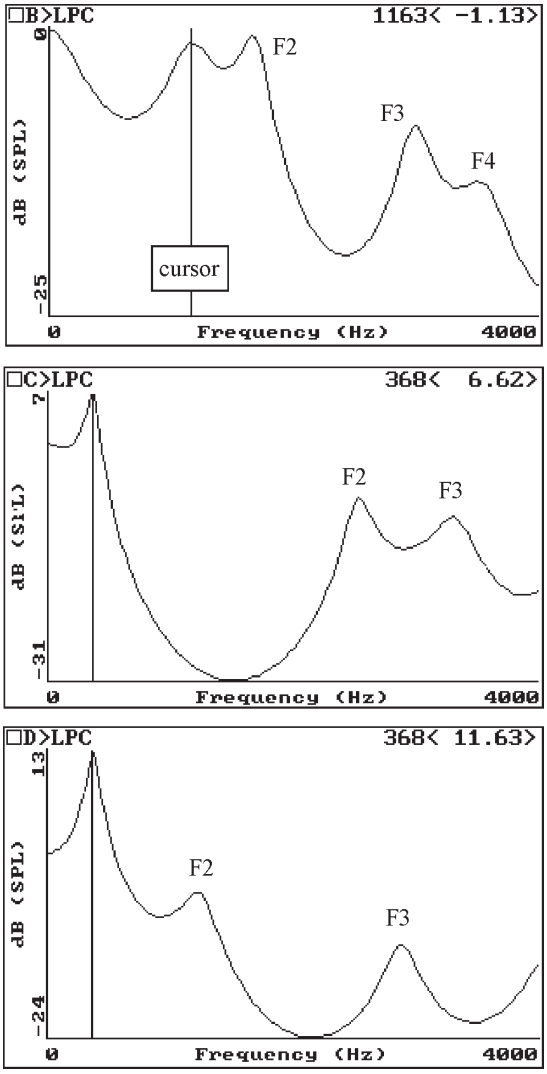
It can be seen from Formula 8 that, since the speed of sound can normally be taken as a constant, what determines the actual F-pattern frequencies for schwa is the length of the supralaryngeal vocal tract. Furthermore, Formula 8 shows that the frequencies vary directly with the length: schwas produced with longer tracts will have lower F-pattern frequencies; schwas produced with shorter tracts will have higher F-pattern frequencies.

Although the details are more complicated for other vowels, involving as they do deviations away from a uniform supralaryngeal vocal tract tube, and more complicated formulae, the general idea still applies. A longer vocal tract will be associated with overall lower vowel formants; a shorter one with higher. Since Caucasian male vocal tracts are on average about 17 cm compared to about 14 cm for females, they are about 20% longer than females, and their F-pattern will be on average 20% lower. The reader can try inserting in Formula 8 a value of 14.0 cms for vocal tract length *l* to derive the F-pattern values for a typical female schwa. The formant frequencies will come out at F_1 = 625 Hz; F_2 = 1875 Hz, F_3 = 3125 Hz and so on for formants higher than F_3. (These are ca 20% higher than the 17cm long vocal tract values of 515 Hz (F_1), 1544 Hz (F_2), 2574 (F_3).)

The principle applies to vocal tracts of differing dimensions within the same sex too, of course, and thus gives a nice illustration of the relation of speech acoustics to anatomy.

To illustrate these points, Figure 12 shows analogous data to those in Figure 11, this time from a young female Cantonese speaker. It can be seen from Figure 12 that the general spectral profiles of the three vowels are similar to those of the male in Figure 11. Thus F1 is low and F2 relatively high for [i]; F1 and F2 are relatively low for [u$_{\downarrow}$]; and F1 and F2 are close and mid for [a]. The formant frequencies for the female's vowels, however, can be seen to be generally higher than for the male. Thus, eg, her F1 in [i] is 368 Hz compared to his at 283 Hz, and her F1 in [a] is 1163 Hz compared to his at 893 Hz. Her F4 in [i] is above the 4000 Hz upper limit of the display. These differences are ascribable, presumably, to the fact that her vocal tract was shorter than the male's. (Her F2 in [u$_{\downarrow}$] is also relatively higher than for the male. However, this reflects a slight difference in articulation as well as the presumed difference in overall length, with the female vowel sounding fronter – this is what is indicated by the $_{\downarrow}$ diacritic.)

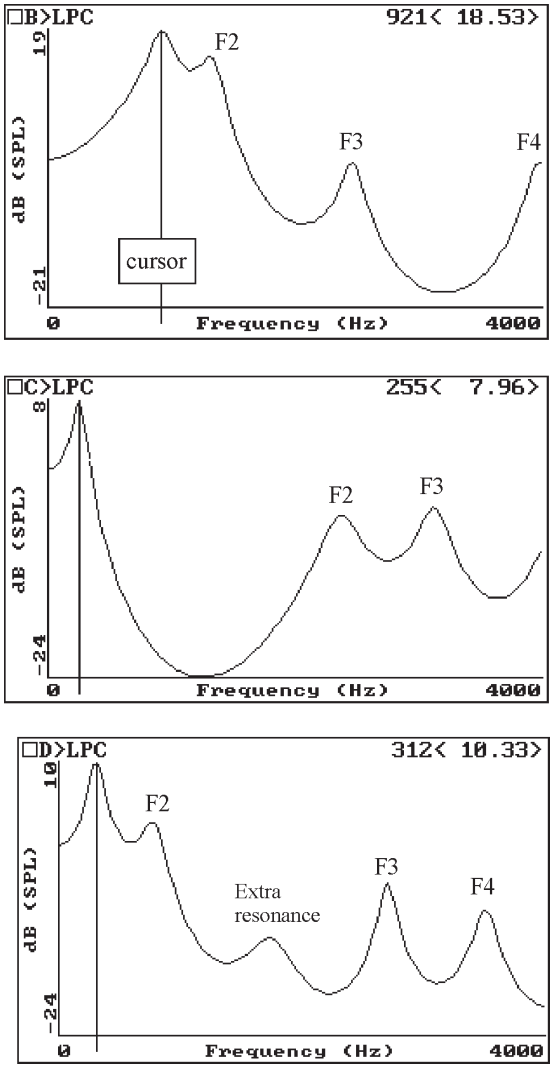
FIGURE 12 Linear prediction spectra for three Cantonese vowel tokens spoken by a female (CF1)



Top = [a], middle = [i], bottom = [u$_{\text{ɻ}}$]. The vertical cursor has been placed through the first formant frequency in each vowel.

In order to show some more between-speaker variation in F-pattern, Figure 13 shows vowel spectra from yet another young Cantonese speaker, this time another male (CM3). It can be seen that similar spectral profiles characterise his three vowels as with the previous two speakers, although some differences are also apparent. CM3's fourth formant is considerably higher than CM1's in both [a] and [i], eg, and his [u] profile is very different in the upper frequency regions. It has a very strong F3 and F4, together with an extra resonance between F2 and F3.

FIGURE 13 Linear prediction spectra for three Cantonese vowel tokens spoken by a male (CM3)



Top = [a], middle = [i], bottom = [u]. The vertical cursor has been placed through the first formant centre frequency (F1) in each vowel.

Acoustic vowel charts

[99.640] At this point it is convenient to introduce a useful means of graphical representation of formant frequencies that will be commonly encountered, eg to show in TFSI reports the differences/similarities in the F-pattern of offender and suspect samples. This is called an "acoustic vowel chart". In order to make an acoustic vowel chart, the actual values for the formant frequencies are needed. These are given in Table 6. They were extracted by the computer using linear prediction.

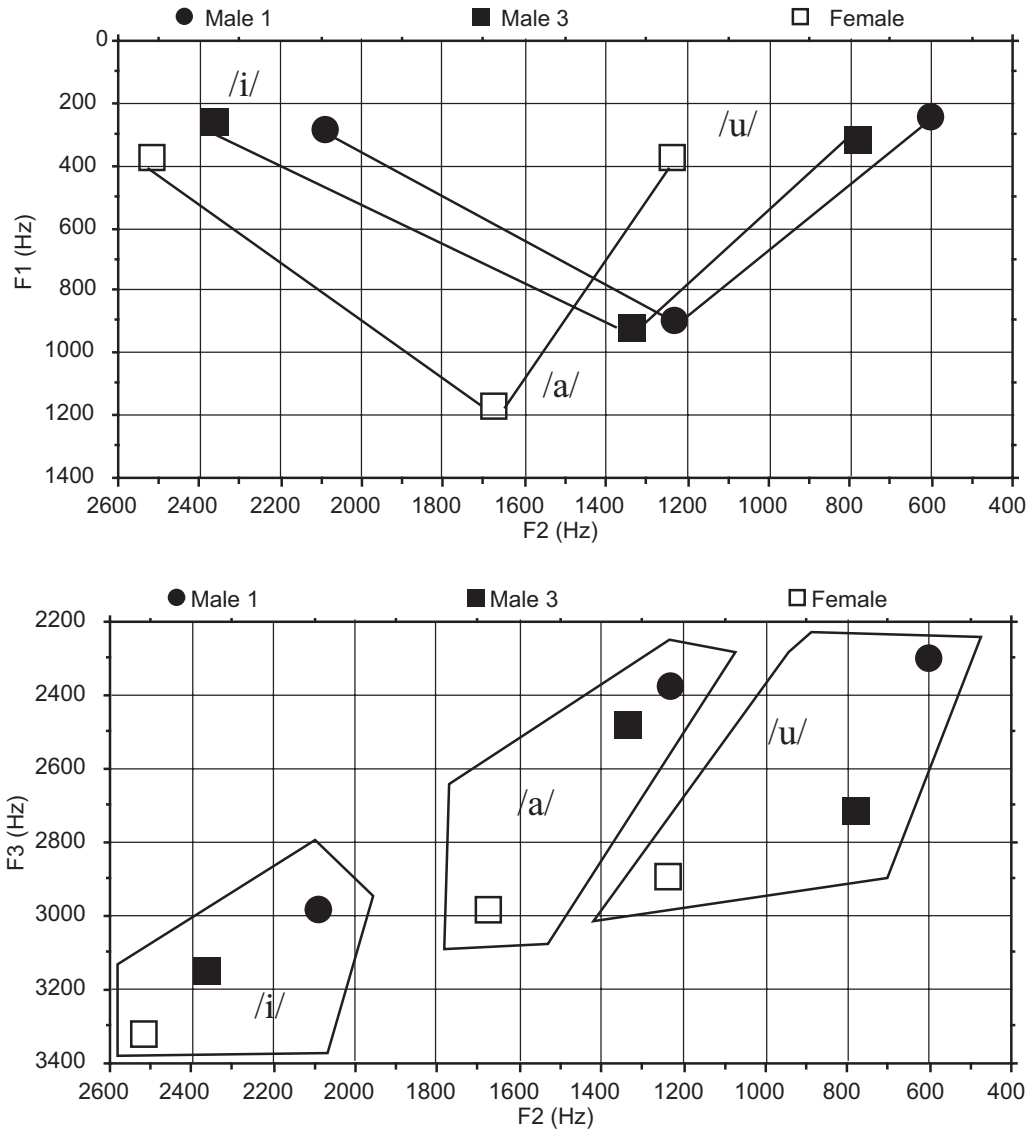
TABLE 6 Formant frequency values for the three Cantonese speakers' /a/ /i/ and /u/ vowels in Figures 11, 12 and 13

	F1	F2	F3	F4
		/a/		
CM1	893	1236	2371	3439
CM3	921	1338	2475	3982
CF1	1163	1678	2976	3560
		/i/		
CM1	283	2091	2978	3541
CM3	255	2365	3143	4143
CF1	368	2520	3316	4312
		/u/		
CM1	241	605	2296	(3145?)
CM3	312	782	2713	3526
CF1	368	1241	2887	4133

CM = Cantonese male, CF = Cantonese female.

In its most basic form – many variations are possible – an acoustic vowel chart is two-dimensional, and involves plotting the frequency of the second formant on the horizontal axis against the frequency of the first formant on the vertical axis. This is shown in the top panel of Figure 14, which shows the three vowels of the two male and one female Cantonese speakers plotted with respect to their first two formants.

FIGURE 14 Acoustic vowel charts for the three Cantonese speakers' vowels in Figures 11, 12 and 13



Top = F1 vs F2; bottom = F3 vs F2.

It can be seen in the top panel that for each speaker the three vowels form a triangle with the high front unrounded vowel /i/ at the top left, the high back rounded vowel /u/ at the top right, and the low central vowel /a/ at the bottom. A plot of F1 vs F2 demonstrates a long-known simple working relationship between perceived vowel quality and acoustics. Perceived vowel height is inversely proportional to F1, so that high vowels like [i] and [u] have low F1, and low vowels like [a] have high F1. The reader should verify this in Figures 11 to 13. F2 relates both to perceived vowel backness and lip rounding: the further back a vowel sounds, the lower the F2, and the rounder the vowel sounds, the lower the F2. Thus [i], a front, unrounded vowel, has a high F2, whereas [u], a back rounded vowel, has a low F2; [a], which sounds neither front nor back, and is unrounded, has an intermediate F2. Again, the reader should verify this.

The female's configuration in the top panel of Figure 14 is displaced down and to the left of the males', which is as expected, given the relationship between tract length and F-pattern previously explained. With the exception of F1 in /i/, CM3's plot is also down and to the left of CM1's, which suggests that CM3 has an overall shorter resting vocal tract than CM1. (It is possible for CM3 to have an overall shorter vocal tract than CM1 and yet have a lower F1 in /i/ than CM1, because F1 in /i/ is dependant on factors in addition to the length of the oral and pharyngeal cavities like the magnitude of the ratio between their cross-sectional areas. It is likely that CM3's pharyngeal volume was greater than CM1's in the production of this vowel.) Also, there is no guarantee that just because a speaker says one vowel, like /i/, with a vocal tract of a certain length, he or she will maintain a comparable length for other vowels.

Acoustic plots are not restricted to F1 and F2. The bottom panel of Figure 14 shows F3 plotted against F2 for the three speakers. Greater separation can be seen between the two male speakers than in the F1 F2 plot, because there are generally bigger between-speaker differences in F2 and F3. The relative position of the vowels can also be shown in three-dimensional space, using three formants.

It must be stressed that formant frequencies actually reflect a myriad of other factors as well as tract size and vowel target. The F-pattern of the same vowel said by the same speaker can, eg, vary as a function of whether the vowel is in a stressed syllable or not; how fast the speaker is speaking; what pitch the vowel is said on; the sounds surrounding the vowel; and, of course, if the vowel is said on two separate occasions. Ideally, then, all these factors need to be controlled as well as possible to ensure maximum comparability between samples and accuracy of LR. Since it is never possible to control all of these factors under the real-world conditions of forensic speaker identification, it is part of forensic-phonetic expertise to know how to control for them as well as possible.

Formants as traditional parameters

[99.650] The examples above have shown how a vowel's formant frequencies are determined by the twin factors of the articulatory shape required by the vowel and the size of the individual speaker's vocal tract (reflected in overall length). Thus the speech acoustics contain information on both what the sound is and the dimensions of the tract that produced it. The information on these separate factors is, however, convolved in the speech signal and not packaged in separate channels.

Vowel formants are coherent constructs from the point of view of phonetics. They actually represent something that can be directly related to both production and perception, and therefore constitute traditional acoustic forensic-phonetic parameters. Their relationship to production lies in the fact, already demonstrated, that formants are the resonant frequencies of the vocal tract during the production of a particular sound by that particular vocal tract. Their relationship to perception lies in the fact that they are the acoustical correlates of perceived vowel quality: change F1, or F2, or both, and the vowel changes.

Because they reflect the dimensions of the vocal tract that produced them, and because they can be directly related to production and perception, vowel formants are commonly used in the forensic comparison of speech samples. Not every formant of every vowel is equally suitable. Some vowel formants show a much greater ratio of between- to within-speaker variation (a so-called "F-ratio") and thus have a greater potential as a forensic-phonetic parameter. F2 and F3 in front and mid vowels [i] [e] and [ɛ], eg, have typically high F-ratios, and will be compared if possible. F3 in low central and non-high back vowels like [a], [ɑ] and [ɔ] is also good if its amplitude is not too low.

Some vowel formants will be typically too high or too low in frequency and thus be compromised by the telephone bandpass (this is demonstrated at **[99.870]-[99.880]**). This will apply especially to the F1 of high vowels like [i] and [u]. The higher formants (F4 and above) of all vowels, although they are often associated with high F-ratios, are usually attenuated out of existence by telephone transmission and will not be available for comparison. Some formants, eg F3 in [u], are typically of too low amplitude to be reliably measured.

It constitutes part of forensic-phonetic expertise to know which are the best formants to compare under the circumstances of the individual case. Whichever set of formants is chosen, however, the question remains the same: what is the probability of observing the difference between the F-pattern of offender and suspect speech samples assuming they have come from the same speaker, and assuming they have come from different speakers?

Automatic features

[99.660] The sections above have described a little about what formants are, and what it is about them that makes them typical *traditional* forensic-acoustic parameters. This section illustrates comparison using parameters that are typical of the other, *automatic* approach. The automatic approach does not distinguish sub-parts of the speech acoustics with respect to their phonetic import, but treats the speech wave simply as a time-varying signal to be processed statistically. The algorithmic mainstay of the automatic approach is a rather beautiful signal-processing technique called the "cepstrum" (pronounced with a hard c: *kep*-). The cepstrum permits spectral detail to be smoothed to any desired degree, and one of its results, illustrated here, is a "cepstrally smoothed linear prediction spectrum" (referred to below as a "cepstrally smoothed spectum"). As its name implies, this spectrum is the result of a cepstral signal-processing operation on a linear prediction spectrum of the type illustrated in the previous section with Cantonese vowels.

To illustrate the cepstrally smoothed (CS) spectrum, we return to the same sound as analysed at **[99.600]-[99.610]**, namely the first diphthongal target in the Adelaide speaker's "hello". The top part of Figure 15 shows a CS spectrum for this sound. Below it is reproduced, for comparison, the radiated spectrum from Figure 10 which was derived above from the individual formants' resonance curves plus the effects of vocal cord vibration, radiation, higher formant contribution and sub-glottal resonance/zero.

The radiated spectrum for this vowel clearly shows the two peaks of F1 and F2 where they are expected to be, given how it sounds: F1 is at about 600 Hz, and F2 at about 1200 Hz. The peaks of F3 and F4, and the blunt F5 peak, can also be seen higher up the frequency range. Compared to the pointy radiated spectrum, the CS spectrum can be seen to have a much smoother profile. In fact, F1 and F2 are so close in frequency that they have been cepstrally smoothed into a broad-bandwidth single hump and are no longer separate.

The power of the cepstrum lies in its smoothing, which can be adjusted to be just enough to preserve the differences between spectra that are important for distinguishing individuals, or speech sounds, but just enough also to eliminate the differences that will tend to hamper such differentiation.

The CS spectrum is built up, like the radiated and linear prediction vowel spectra illustrated above, from more simple spectral components. However, unlike the radiated/LP vowel spectra, these components are not the resonant curves of individual formants. Rather, the CS spectrum is constituted from a set of much more simple individual spectral components, each determined by a separate "cepstral coefficient" (CC). The number of individual spectra used (equivalently, the number of cepstral coefficients) depends on several factors, one of which is how well one wants to approximate the details of an actual spectrum. A typical number of CCs is between 12 and 14. The CS spectrum in Figure 15 was derived from 14 CCs.

FIGURE 15 Comparison of cepstrally-smoothed spectrum (top) and radiated spectrum (bottom) for the first target of the second-syllable diphthong in a single token of Australian "hello"

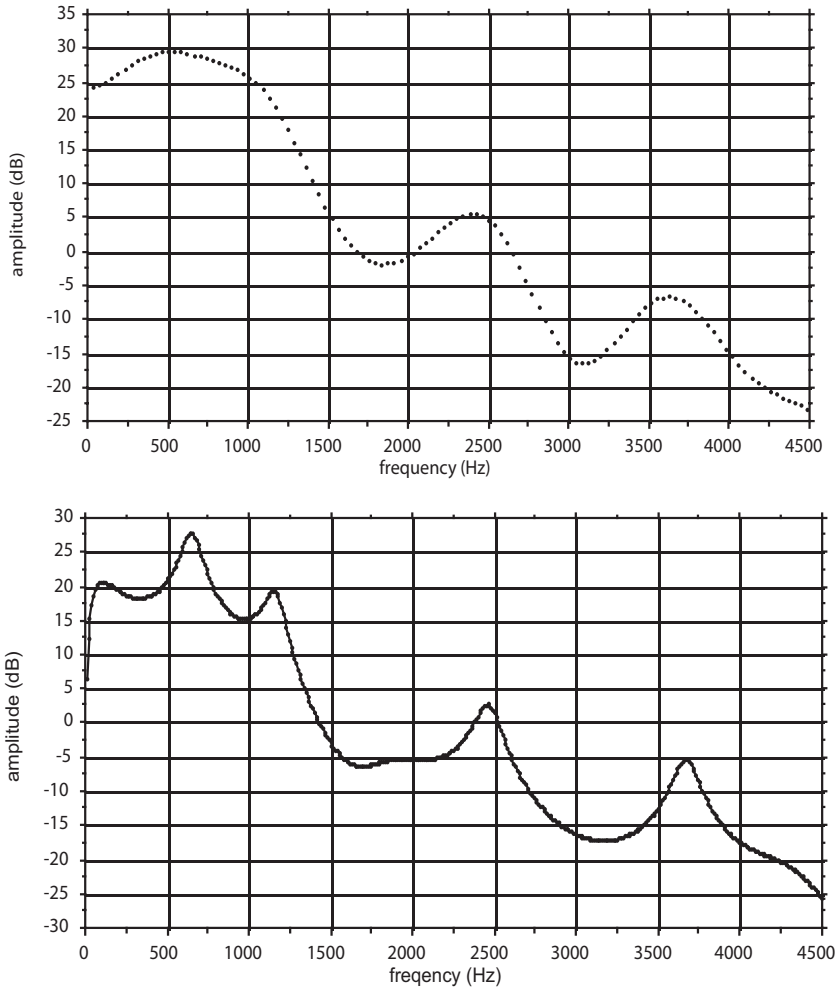
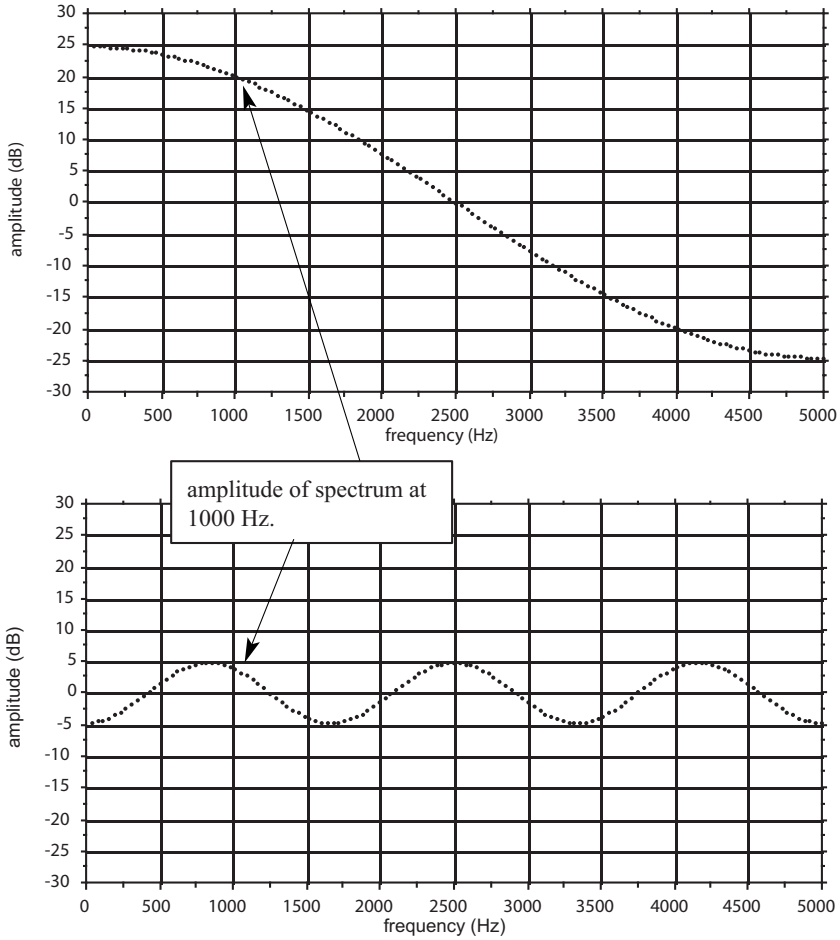


Figure 16 shows the spectra for two of the individual component CCs for this sound: the first and the sixth CCs. As can be seen, the spectral amplitudes are very simple sinusoidal curves, symmetrical around 0 dB, and very unlike the formant resonance curve components described above (in Figures 5 to 7, for example). This is because the curves are based on the trigonometical cosine function. The spectrum of the sixth CC, eg, represents three repeats of a cosine curve, the spectrum of the first CC represents a half of a cosine curve. The maximum deviation of a curve away from 0 dB is determined by the magnitude of the cepstral coefficient (it is actually 10 times the value of the CC); whether the initial amplitude value, at 0 Hz, is positive or negative is determined by the coefficient's sign (positive or negative). Thus it can be seen that the maximum deviation of the sixth CC from zero dB is just under 5 dB, and its initial value, at 0 Hz, is negative. Thus the sixth CC will be negative and just under one-tenth of 5: its actual value is -0.477. The reader might like to visually estimate the first cepstral coefficient from the top panel of Figure 16. (The amplitude of the first CC curve at 0 Hz is just under 25 dB, so its CC will be positive and just under 25/10. The actual value is 2.4699.)

FIGURE 16 Spectra of first (top) and sixth (bottom) cepstral coefficient for the first target of the second-syllable diphthong in a single token of Australian "hello"



The idea, then, is to analyse the complicated shape of a radiated speech spectrum in terms of a set of cosinusoidal components, each defined by their cepstral coefficients, which when added together reconstitute the complicated spectral shape to any desired degree.

Formula 9

$$H_k(f) = C_k \cdot \cos\left[\left(\pi/_R \cdot f\right) \cdot k\right] \cdot 10$$

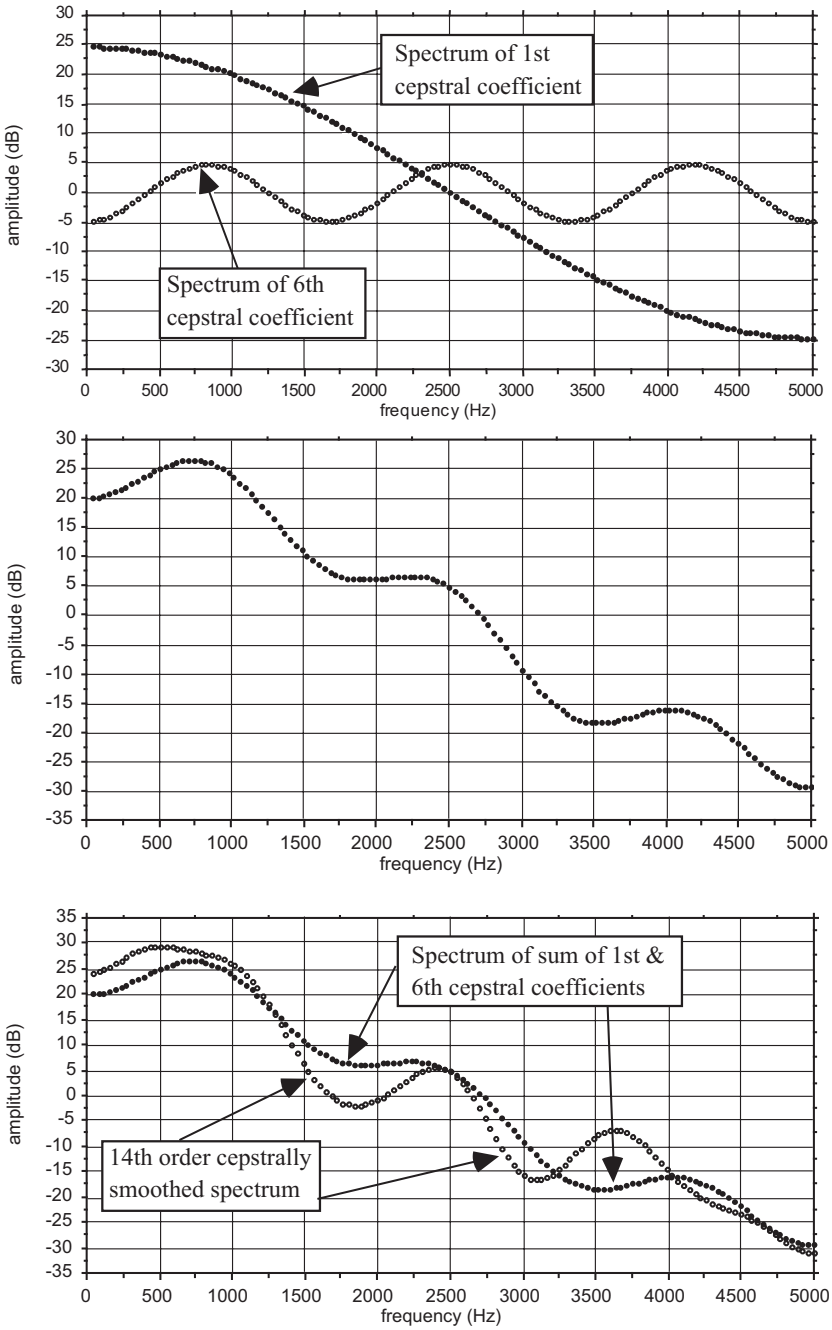
The formula which allows the amplitude in a CS spectrum to be calculated at a given frequency is shown as Formula 9. Note the "cos": this is the part that determines the cosinusoidal nature of the curve. In the formula, the $H_k(f)$ to the left of the equals sign stands for the "height, or amplitude, of the spectrum of the kth cepstral coefficient at the frequency f". This might be, eg, the amplitude of the spectrum of the first cepstral coefficient corresponding to a frequency of 1000 Hz. From the top panel of Figure 16 it can be seen that this value is about 20 dB, and to show how the formula works, this value will be calculated using it.

The C_k immediately to the right of the equals sign stands for the value of the kth cepstral coefficient. In this demonstration, the value for the first cepstral coefficient is 2.4699, so $C_k = C_1 = 2.4699$. The R represents the frequency range over which the spectrum is to be calculated: from Figure 16 this can be seen to be from 0 to 5000 Hz, so $R = 5000$ Hz. The f represents the frequency at which the amplitude is to be calculated (here $f = 1000$ Hz), and the k is the number of the cepstral coefficient, which is 1, since we are using the first CC. These values are shown substituted into Formula 10, where it can be seen how the amplitude at 1000 Hz in the spectrum of the first cepstral coefficient ("H_1 (1000)") actually comes out as 19.98 (dB).

Formula 10

$$\begin{aligned} H_1(1000) &= 2.4699 \cdot \cos\left[\left(3.1416/_{5000} \cdot 1000\right) \cdot 1\right] \cdot 10 \\ &= 2.4699 \cdot \cos[0.6283] \cdot 10 \\ &= 2.4699 \cdot 0.809 \cdot 10 \\ &= 19.98 \end{aligned}$$

FIGURE 17 Combining cepstral coefficient spectra to approximate a spectrum



Top: superposed spectra for first and sixth cepstral coefficients. Middle: summed spectrum from 1st and 6th CCs. Bottom: spectrum of summed 1st and 6th CC spectra compared with 14th order cepstrally-smoothed spectrum. All examples are from the first target of the second-syllable diphthong in a single token of Australian "hello".

Figure 17 shows the result of adding together the spectra of the first and sixth CCs shown in Figure 16, and how well the result approximates the final CS spectrum based on all 14 CCs. The top panel shows the two cepstral coefficients' spectra superposed, and the middle panel shows the curve resulting from the sum of the two CCs' spectra. The 6th CC curve seems to be sitting on top of the 1st CC curve. It has just been demonstrated that at 1000 Hz, the amplitude of the 1st CC spectrum is about 20 dB. The amplitude of the 6th CC spectrum at 1000 Hz can be estimated visually from the top panel of Figure 17 to be about 4 dB. Adding these two amplitude values gives about 24 dB for the height of the combined 1st and 6th CC spectral curve at 1000 Hz, and it can be seen that the amplitude of the combined curve in the middle panel has, indeed, a value of about 24 dB at 1000 Hz.

In the bottom panel of Figure 17, the combined 1st and 6th CC spectrum is compared with the full 14th order CS spectrum that was shown in the top panel of Figure 15. It can be appreciated that the more spectra that are combined, the better the approximation to the final 14th order spectrum. When the spectra from the first and sixth CCs are combined, as in Figure 17, the approximation is not bad up to about 3000 Hz, and above 4000 Hz, but between 3000 Hz and 4000 Hz the agreement is poor. Adding more spectra from the remaining 12 CCs improves the fit, and when the spectra from all 14 CCs are combined, of course, the curves are identical and the fit is perfect.

Table 7 lists the amplitudes for all 14 CC curves at 1000 Hz, each calculated in the same way as illustrated above from the 14 individual CCs, which are also given. In Table 7, the leftmost column gives the number of the cepstral coefficient. The middle column gives the actual value for that coefficient, and the rightmost column gives the amplitude for the spectrum of that coefficient at 1000 Hz. Thus the value of the first cepstral coefficient is 2.4699, and this, as has been shown, gives an amplitude at 1000 Hz of 19.98 dB. Likewise, the amplitude of the sixth CC spectrum at 1000 Hz – estimated visually at about 4 dB – can be seen to be 3.86 dB. The sum of the amplitude values for all 14 CC spectra can be seen to be 25.38 dB, and it can be seen in the bottom panel of Figure 17 that this is, indeed, the amplitude of the 14th order curve at 1000 Hz.

TABLE 7 Values for calculation of the 14th order cepstral amplitude at 1000 Hz for the first diphthongal target in the second syllable of a token of Australian "hello"

k	C_k	$H1000_k$ (dB)
1	2.4699	***19.98***
2	2.84E-01	0.88
3	4.19E-01	-1.30
4	-1.81E-01	1.47
5	3.17E-02	-0.32
6	-4.77E-01	***3.86***
7	-2.62E-01	0.81
8	2.81E-01	0.87
9	9.04E-02	0.73
10	-1.74E-01	-1.74
11	4.76E-02	0.38
12	-3.52E-02	-0.11
13	-3.78E-02	0.12
14	3.27E-02	-0.26
	$\Sigma H_k 1000$ =	**25.38**

k = number of cepstral coefficient, C_k = value for the kth cepstral coefficient, $H1000_k$ = amplitude at 1000Hz in the spectrum of the kth cepstral coefficient. ΣH_k 1000 = Sum of amplitudes for all 14 cepstral coefficients at 1000 Hz.

In order to calculate the whole of this 14th order CS spectrum, the calculation would have to be repeated for each frequency value from 0 Hz to the upper bound of the frequency range (here, 5000 Hz). Thus it can be seen that each of the spectra of the individual cepstral coefficients contributes to all parts of the final spectrum: the amplitude at any one frequency point is a contribution from each of the 14 cepstral coefficients, although some CCs contribute at some frequencies more than others. With the cepstrum therefore it is strictly speaking not possible to attribute a particular spectral feature – say, a second formant peak – to the spectrum of a particular cepstral coefficient.

The most important thing to understand from these calculations is that the cosinusoidal components which are added together to constitute the final smoothed spectral shape do not correspond to anything real in the process of speech production. They are simply abstract mathematical constructs that, when combined, give the best approximation to the shape. It is in this sense that they are automatic parameters. This is unlike the traditional features of formant centre frequencies which, as explained above, *can* be related to aspects of speech production and perception.

Between- and within-speaker cepstral differences

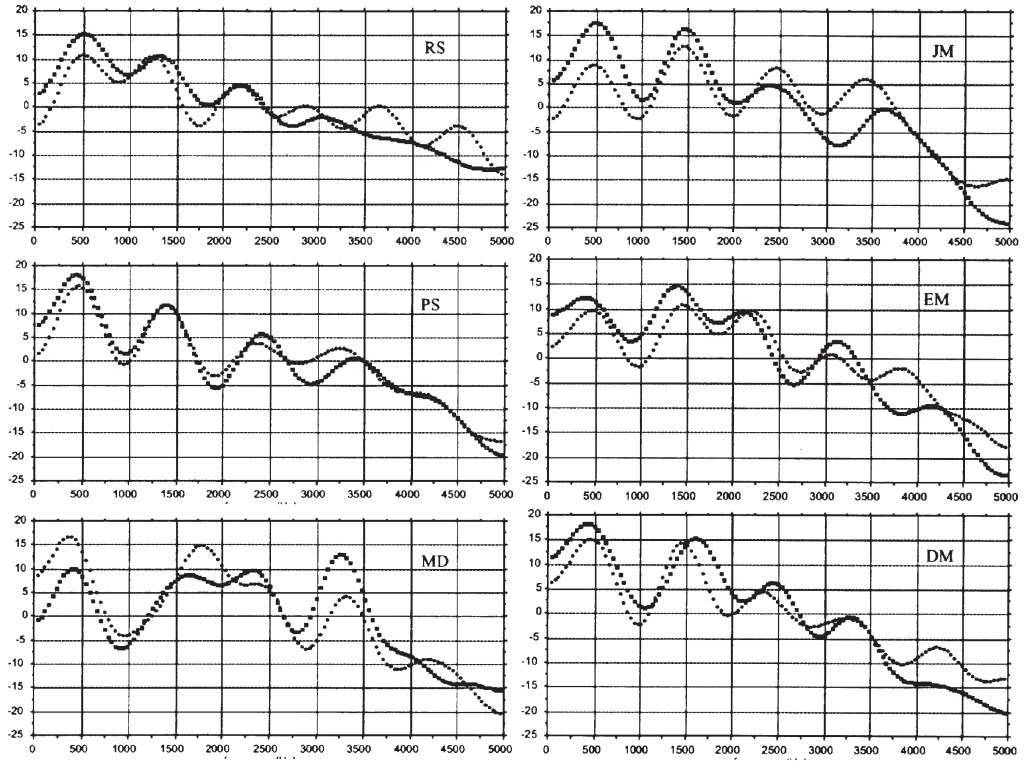
[99.670] Figure 18 shows some typical between- and within-speaker differences in cepstrally-smoothed spectra. The sound illustrated is once again taken from the second syllable of Australian "hello", but this time it is the *second* diphthongal target (the end of the *o*). This is the point in Australian "hello" where there is typically the highest ratio of between- to within-speaker variation. Six male speakers' data are shown, each speaker in a different panel. Within each panel are shown two spectra, with each spectrum representing the average value from a set of "hellos" for the same speaker. Both sets within a panel were recorded about a year apart, so what is shown in each panel is clearly non-contemporaneous within-speaker variation for this particular sound.

Figure 18 also shows some typical between-speaker differences in the CS spectrum of the second diphthongal target. MD's two spectra, eg, are very different from PS's. (MD has considerable energy present at 2000 Hz, whereas PS has a trough; MD also has a more prominent spectral peak between 3000 Hz and 4000 Hz.) There are also regions where the different speakers' spectra appear rather similar: between 0 Hz and 1000 Hz, eg, all speakers show a peak centred around 500 Hz (this is actually the cepstrally smoothed F1) which drops to a trough at about 1000 Hz.

There is generally considerable within-speaker variation above about 3500 Hz. RS's two spectra differ in absence or presence of two spectral peaks above 3500 Hz, for example. The amount of between-speaker similarity below about 1000 Hz, together with the within-speaker differences above about 3500 Hz, means that the strongest evidence from the CS spectra – the largest LR deviations away from unity – will, at least for this sound, probably come from the spectral profile between 1000 Hz and 3500 Hz.

The spectra in Figure 18 can also be used to point out that speakers differ with respect to the amount and nature of within-speaker variation they show. PS does not show much difference between his two recordings separated by a year, whereas JM shows a considerable difference. (In fact, the amount of difference for JM is more typical of a different-speaker pair.) Thus there is between-speaker variation in within-speaker variation. These characteristics – differential between-speaker differences and differential within-speaker differences – are all typical for speech acoustics, and constitute a major part of the problem that technical speaker identification has to solve.

FIGURE 18 Mean cepstrally smoothed spectra from the second target of the second-syllable diphthong in a set of Australian "hellos", illustrating between- and within-speaker variation for six male speakers



Power versus interpretability: Cepstrum versus formants

[99.680] Differences between forensic speech samples can be calculated using either formants or the cepstral coefficients, or both, and the strength of the evidence assessed by determining the likelihood ratio for the differences. At the present, formants constitute the more common approach, and automatic parameters are only likely to be used by experts who have a background in speech engineering and/or automatic speaker recognition.

It was pointed out above that, as traditional versus automatic features, the formants and the cepstrum represent two rather different types of forensic features. Both have been sufficiently exemplified for the reader to appreciate the difference between them, and it is now possible to list their pros and cons.

Several arguments can be adduced in support of both traditional and automatic approaches. It is generally acknowledged that both the identification of formants, and the extraction of their centre frequencies, is often difficult to accomplish automatically. This is especially true for the higher formants, and for formants in certain sounds that are known to be useful in speaker identification, like nasal consonants. Use of the cepstrum avoids this problem entirely, since it is not based at all on identifying and extracting individual formants.

Moreover, with the cepstrum, the whole of the relevant part of the spectral shape can be quantified and compared, rather than only one or two frequencies of its spectral peaks. Thus there is theoretically more information in a cepstral than a formant comparison, and there is therefore a greater chance of picking up important similarities or differences between samples with the cepstrum. More information per se is not necessarily invariably a good thing. It is a truism of automatic speech and speaker recognition that too much information will hamper performance, but as has already been pointed out and illustrated, one of the beauties of the cepstrum is that the degree to which it smooths can be controlled, by selecting only the set of CCs which give maximum resolution. For example, discrimination performance often improves if the first, as well as the higher order (eg, CC11 and above) cepstral coefficients are ignored.

Of course, it is possible in principle to quantify the difference between samples in *overall* linear prediction spectral shape too. However, this cannot be done unless all the formant frequencies and bandwidths can be accurately estimated, and, as already pointed out, this is not an easy thing to do automatically, especially with bandwidths. Thus forensic comparison between overall LP spectra is not a practical proposition.

A further consideration is that cepstral coefficients by their nature are minimally correlated. This means that the calculation of a combined LR from the LRs of the individual CCs is less problematic than with formants, many of which are known to be correlated to different degrees.

Probably the most important forensic advantage of cepstrally based comparisons is that the *average* strength of evidence – as reflected in the *average* magnitude of likelihood ratios – can be expected to be considerably greater than with formants. In other words, the cepstrum is on average forensically more powerful. For example, the author showed in a recent paper involving the comparison of 240 same-speaker pairs and 28,320 different-speaker pairs that the average LR was about 50 for a formant-based approach compared to 900 with the cepstrum: see Rose et al (2002). This will almost certainly have had to do with the fact that the cepstral LR was the result of combining more features (ie, more spectra) – up to 10 per speech segment perhaps – whereas a formant-based LR is usually simply based on a LR from a few features (ie, formant centre frequencies).

Of course, it must always be remembered that averages will not predict specific instances, and it is specific instances with which the court is concerned. (This important point is discussed in Rose (2002, p 322ff).) It may be that a specific comparison with one or two formants performs effectively as well as one based on 10 CCs. To illustrate this, Table 8 shows the results of comparing "hellos" with respect to two of their sounds using cepstral and formant methods.

TABLE 8 Likelihood ratios from comparisons of different-speaker "hello" pairs using cepstrum and formants

Different speaker pair	Mode of comparison	LR T2	LR /l/	Combined LR
DM 2.3-PS 2.1	Cepstrum (CC2 - CC10)	0.569	1.41E-8	**8.0E-9**
	Formant (F2)	*2.008*	0.140	**0.28**
DM 2.1-RS 2.2	Cepstrum (CC2 - CC10)	1.1E-20	*19.2*	**2.0E-19**
	Formant (F2)	3.4E-14	0.015	**5.3E-16**

T2 = second diphthongal target.

In Table 8 two different-speaker comparisons are shown, one between sets of "hellos" from speakers DM and PS, and one between sets of "hellos" from speakers DM and RS (the figures after the initials index the recording session – two different sets of "hellos" from DM are represented).

Each pair of speakers is compared with respect to features of two different sound segments in the word "hello": the /l/, and the second diphthongal target. The formant comparison used only two pieces of information – the F2 frequency in /l/ and the F2 frequency in the second diphthongal target. The cepstral comparison, on the other hand, was based on the 2nd to the 10th CC inclusive, and therefore used 18 pieces of information. The LRs for both segments are shown in Table 8, and the combined LR, which is simply the product of both individual segments' LRs, is given in the right-most column.

The rightmost column of Table 8 shows, first, that the combined LR from the two segments is in all cases lower than unity, which is correct, given that these are different-speaker pairs. Note in passing that both types of comparison have LRs for individual segments which run counter to reality in showing values greater than unity. In the first pair, it is the formant in the diphthong (LR = 2.008); in the second pair it is the /l/ cepstrum (LR = 19.2). The same effects were also demonstrated in Table 1 at **[99.250]** above; this example shows that the cepstrum, despite its power, is not immune from them.

The second thing to see from Table 8 is that the LRs for the formant-based comparisons are nearer to unity than those of the cepstral-based comparison, which means that the evidence from the latter is stronger.

Third, there is a big difference between the cepstrally-based and formant-based LRs for the first pair of speakers, but the difference is nowhere near as big for the second. In the first pair – DM 2.3 versus PS 2.1 – the cepstrally based LR of 8.0E-9 says that one would be *very many millions of times* more likely to observe the difference between the cepstrally smoothed spectra of the two samples had they come from different rather than same speakers. The formant-based LR of 0.28 says that one would be just over (1/0.28 =) *three and a half times* more likely to observe the difference between the F2 values of the two samples had they come from different speakers. Whereas this is an enormous difference, the difference for the second pair – DM 2.1 and RS 2.2 – is neither so big, nor so important: both formant- and cepstrally-based analyses give equally astronomically large figures in support of the defence hypothesis.

All the above advantages for the cepstrum do not appear to leave too much going for the traditional formant approach, although one thing that can be said immediately in defence of formants is that the formants that are most often used – F2 and F3 in mid and high front vowels, eg – are neither difficult to identify or extract. However, all spectral shape features, like the cepstrum, are very sensitive to the properties of the channel(s) through which they have been passed. Thus they will reflect properties of the transmission channel, and these will be added to the properties of the original signal. Thus the same signal over two different channels will show a different cepstrum. It is possible – indeed, it is mandatory in automatic speaker recognition – to use sophisticated channel-normalisation techniques which are able to control very well for the differential influence of channel transmission, and the componential nature of the cepstrum lends itself nicely to this. However, their use in forensic speaker identification is still being developed. In contrast, the frequencies of some formants are not adversely affected by channel transmission, which is a clear advantage.

But there is one extremely important positive characteristic for the traditional approach which is almost completely lacking in the cepstrum. Because they are related to aspects of the production and perception of speech sounds, formants (and traditional parameters in general) have high interpretability. The interpretability works in two directions. The formants can be interpreted in terms of inferred articulatory reality, and the speech percept in turn can be interpreted in terms of the formants. Thus, eg, if a difference between two speech samples in the quality of a particular vowel is perceived, the expert will know exactly what particular aspect of the vowel acoustics – what particular formant, eg – to measure in order to quantify this difference acoustically. This cannot be possible with the cepstrum: since no single CC is responsible for any particular part of a spectrum, individual parts of the spectrum cannot be attributed to a particular CC.

The result of this differential interpretability is that it will be very much more difficult to explain in detail to the court how one has come to a conclusion about the probability of observing a cepstrally quantified difference between speech samples. Any such explanations will tend to have a "black-box" quality, whereas it is much more easy to justify how one has reached a conclusion with traditional parameters.

Clearly, the ideal forensic approach would be to use both traditional and automatic parameters, using perhaps the cepstrum for sounds for which individual formants are difficult to identify and extract. Formants might then be compared for sounds in which they are easily identifiable and extractable, and where a clear relationship exists between the sound's percept and the formant.

Automatic parameters versus automated approaches

[99.690] It is important to understand that the selective use of automatic parameters does not imply that a fully automated approach is possible in TFSI. There are currently serious attempts to make TFSI fully automated, but the consensus among practitioners is that a fully automated approach will never be possible. Reasons for this are discussed in Rose (2002, Ch 5).

Fundamental frequency

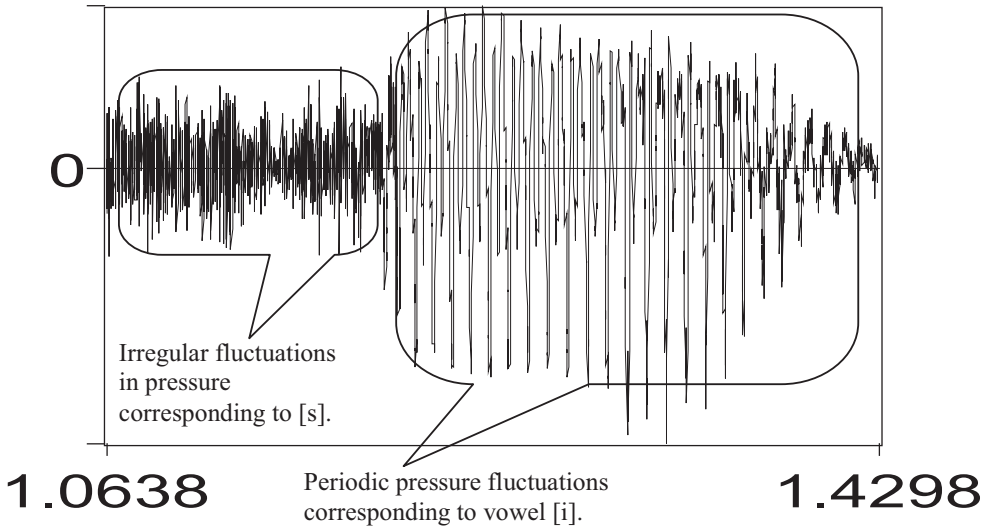
[99.700] It was explained above that a central component in the production of speech sounds is the vibration of the vocal cords. The presence or absence of vocal cord vibration – signalling whether a sound is "voiced" or "voiceless" – can be the only thing distinguishing between two different speech sounds. An example of this is the sounds [z] and [s]. Both have the same articulation – both are alveolar fricatives – but the former is voiced, with vibrating vocal cords, and the latter voiceless, without vibration.

The rate of vibration of the cords determines pitch, the perceptual dimension within which many different types of linguistic categories like tone, intonation and stress are signalled. The acoustic correlate of rate of vocal cord vibration, or "fundamental frequency", is one of the most commonly used traditional acoustic features in forensic comparison of voice samples, and therefore it is important for the reader to understand its main aspects. This section explains what fundamental frequency is, so that its most common forensic use as a long-term feature can then be described at **[99.720]-[99.780]**.

Fundamental frequency is clearly a traditional acoustic parameter because it can be said to be a meaningful property of the speech wave. It reflects fairly directly things that speakers do with their cords to signal linguistic information, and it is also the acoustic cue that listeners use to make perceptual sense of the linguistically vital dimension of pitch. Its popularity in forensic speech comparison also rests on the fact that it was shown to be a fairly powerful parameter in early automatic speaker identification; thus under well-controlled conditions it shows greater between- than within-speaker variation. It is also robust in telephone transmission; relatively easy to measure; and one can expect to find enough of it in forensic speech samples (most languages will have more voiced speech sounds than voiceless, both in terms of phonemic inventory and incidence in normal speech).

Fundamental frequency (abbreviated as F0 and commonly referred to as "eff oh" or "eff sub-zero") is the basic rate of repetition of the quasi-periodic part of the speech wave, and corresponds to – is the principal acoustical correlate of – the rate of vibration of the speaker's vocal cords. This is illustrated in Figure 19.

FIGURE 19 Acoustic wave form of part of the Cantonese word "gusih" ("story")



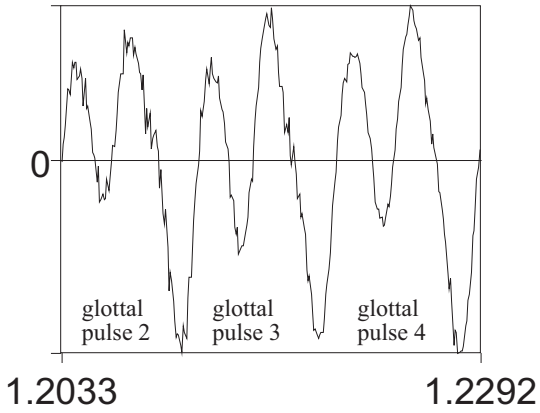
Horizontal axis is duration in seconds; vertical axis is pressure (voltage) in arbitrary units.

Figure 19 shows the acoustic wave-form from part of a Cantonese utterance (the second syllable of the word "gusih" ("story") (the syllable sounds like the word "see"). The horizontal dimension is time, in seconds; the vertical dimension is the electrical analog of pressure (ie, voltage) in arbitrary units. It can be seen that the speech wave for this word lasts for (1.4298 - 1.0638 =) 0.366 seconds, or about a third of a second, and consists of rapid positive (above zero) and negative (below zero) fluctuations in pressure.

The pressure fluctuations in roughly the first half of the wave – these correspond to the *s* – are irregular, but those in the second part – which corresponds to the *ee* vowel – seem to be periodic, ie, recur at regular intervals. These regularly occurring pulses are often called "glottal pulses", because they are the result of vocal cord activity (the glottis is the space between the cords). This regular wave-form for the vowel is what is known as "quasi-periodic", and it contains about 26 approximately evenly-spaced glottal pulses.

Figure 20 zooms-in on a short portion at the beginning of the quasi-periodic vowel wave-form in Figure 19 (actually, glottal pulses two to four have been magnified). It can be seen that their structure is quite complicated, involving subsidiary pressure fluctuations within each glottal pulse. Within each pulse, the pressure goes up and down twice; and there are also smaller ups and downs visible – very roughly about 26 per pulse.

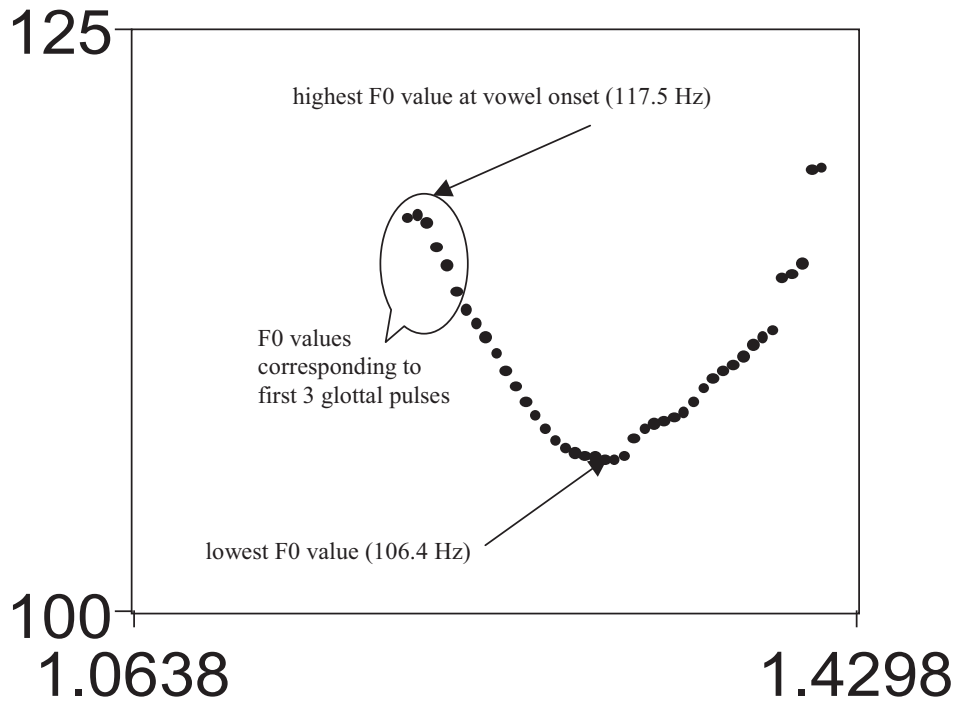
FIGURE 20 Magnified portion of the vowel wave-form from Figure 19



Fundamental frequency is quantified in Hertz (Hz). 1 Hz is a rate of 1 per second. The three glottal pulses in Figure 20 repeat in the space of (1.2292 - 1.2033 =) 0.0259 seconds. This means that in one second they would repeat on average (3/0.0259 =) 116 times. Their fundamental frequency is therefore 116 Hz. Since F0 is the acoustic consequence of the rate of vibration of the vocal cords, it can be said that the Cantonese speaker's cords were vibrating at the average rate of 116 times per second when they gave rise to these three pulses.

Fundamental frequency can be relatively easily extracted automatically by computer, provided the signal is not too noisy, and the speaker's phonation type is fairly normal (automatic F0 extraction does not perform very well with creaky voice, or harsh voice, for example). Figure 21 shows the fundamental frequency on the same Cantonese syllable as shown in Figure 19 automatically extracted by a well-known and widely used software package. The horizontal dimension still shows time in seconds, and is coterminous with the axis of Figure 19. The vertical dimension is fundamental frequency, in Hz, and spans 25 Hz, from 100 Hz to 125 Hz.

FIGURE 21 Fundamental frequency of the [i] vowel in Figure 19 automatically extracted by computer



Horizontal axis is duration in seconds and coterminous with that of Figure 19. Vertical axis is fundamental frequency in Hz.

Figure 21 shows that the F0 on the vowel [i] has a V shape, first falling from a value of about 118 Hz to a lowest value of about 106 Hz in mid-syllable duration, and then rising again. There are some F0 discontinuities at the end of the syllable, caused by a change in the mode of vocal cord vibration which typically occurs at the offset of phonation.

The first six dots of the F0 trace in Figure 21 correspond to the F0 values of the three magnified pulses in Figure 20. (The F0 was estimated by computer at the approximate rate of twice every glottal pulse, so there are six estimates per three pulses.) It can be appreciated visually that these dots indicate F0 values of between 113 Hz and 117 Hz, and in fact the computer program returns the value of 115.9 Hz for the mean F0 for these first three pulses. This is as near as makes no difference to the average value of 116 Hz measured directly from the expanded wave-form in Figure 20 above, so the automatic extraction is working very well here.

As far as the average F0 for the whole vowel is concerned, the computer returns a value of 111.5 Hz. This value compares excellently with 111.3 Hz estimated directly from the wave-form in Figure 19, in the following way. The first of the 26 glottal pulses of the whole vowel occurred at about sec 1.1961 in Figure 19, and therefore the 26 pulses lasted for about (1.4298 - 1.1961 =) 0.2337 seconds. The average duration of the 26 glottal pulses is therefore 0.2337/26 seconds, or 0.008988 seconds, which means an average F0 for the whole vowel of (1/0.008988 =) 111.3 Hz.

The aspect of an individual's vocal tract that is assumed to be imprinted on her or his acoustical output of F0 is the size of the individual's vocal cords. Like formants, then, F0 reflects the anatomy of the vocal tract that produced it. However, the relationship between F0 and vocal

cords is rather more complicated than that between F-pattern frequency and vocal tract length discussed above at **[99.630]**. It is a useful simplification to say that, other things being equal, the rate of vibration of the vocal cords, and hence the F0 they produce, is determined by their length and mass: larger, more massive, cords vibrate at lower frequencies than smaller, less massive ones. As a simple fact of different anatomical endowment, different speakers may have different-sized vocal cords, and this may translate into different F0 values. For example, just as females are generally more gracile than males, so do they generally have smaller and shorter vocal cords, and this will be reflected in overall higher average F0 values for females than males. There are, of course, size differences within sex too, and these can also be assumed to be potentially reflected in F0. Average F0 values are quoted to lie within a range of 180 Hz to 300 Hz for females, and 90 to 140 Hz for males.

Just as with formant frequencies, however, a particular set of F0 values in a speech sample will simultaneously reflect very many more factors than just the vibratory length and mass of the vocal cords of the speaker that produced them. They will, first of all, reflect linguistic factors like the speech sound being made: a rising intonational pitch versus a falling intonational pitch; a high-pitched tone versus a low-pitched tone in a tone language; a high-pitched syllable versus a low-pitched syllable in a pitch accent language; a stressed syllable versus an unstressed syllable in a stress accent language. F0 is also affected by segmental qualities. Other things being equal, high vowels have higher F0 than low vowels; voiced consonants induce lower F0 values on a following vowel than voiceless consonants. Some different languages appear to have different default values for F0 height and range, which makes F0 comparison of samples problematic if they are in different languages.

F0 will also reflect paralinguistic features like temporary emotional state: anger, eg, is often signalled by higher F0. This is relevant forensically, because paralinguistic factors will obviously need to be controlled in order for a comparison between two voice samples involving F0 to be valid. It would be wrong, eg, to use F0 to compare a voice sample that sounds angry with one that does not.

F0 can also signal extralinguistic features: a speaker might choose to speak habitually at a lower pitch than warranted by the speaker's cord size, perhaps in an attempt to convey gravitas.

And, of course, F0 will reflect differing states of health of the cords (or the brain that controls them), showing influence from laryngitis, smoking or intoxication.

Most importantly, F0 can reflect varying ambient circumstances, like the amount of background noise. In order to make themselves heard above background noise, speakers will often speak louder, especially on the telephone. The increase in subglottal pressure that speaking louder necessitates will increase F0, since F0 is also a function of subglottal pressure. Again, such factors need to be controlled for forensically.

Thus it can be seen that the rate of vibration of the vocal cords at a particular instant is determined by very many linguistic, paralinguistic and extralinguistic factors, and far from directly reflecting some kind of inherent anatomical feature, there are very many degrees of freedom involved in the production of F0, and in principle all of them need to be controlled for optimum forensic comparison. The realities of the forensic situation make this impossible, however, and it is part of TFSI expertise to know what is safely and usefully comparable. One usual expedient is simply not to entertain comparison of samples with respect to short-term F0 measurements, like values on the same tone in tone languages, or the lowest value in falling intonations (often assumed to be a speaker-dependent F0 feature), but to restrict the forensic use of F0 to so-called long-term comparisons. These are explained in the next section.

Long-term features

[99.710] A distinction can be drawn between long-term and short-term features. Short-term features are usually part of the acoustic reflexes of a particular speech sound like the F2 of an individual vowel, or the cepstrum of a nasal consonant. Since speech sounds are ephemeral, short-term features are usually calculated over a very short space of time – effectively an instant of time of the order of a few milliseconds. The different kinds of spectra presented so far have been short-term: the F-pattern for three single different Cantonese vowels, eg, or the cepstra for the first and second diphthongal targets in the *-lo* of an Australian "hello".

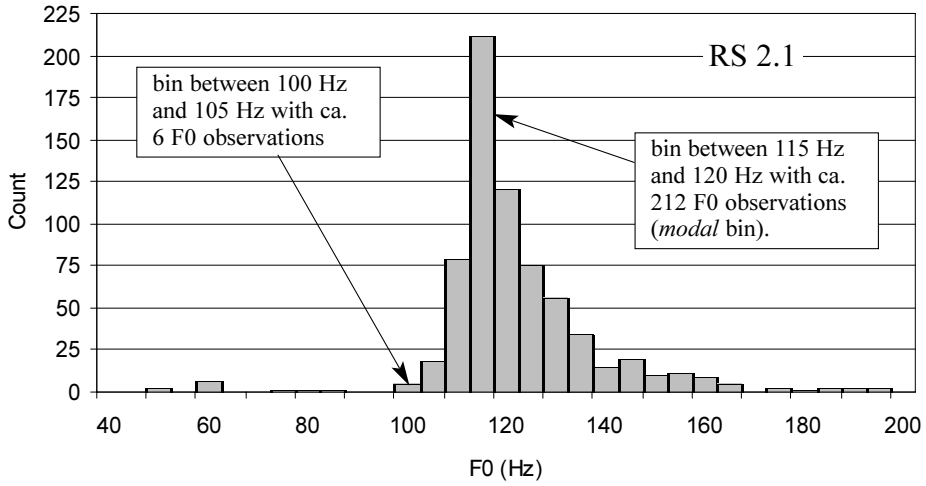
Long-term features, on the other hand, represent average values calculated over a long period of speech – perhaps in the order of tens of seconds or for as long as the speaker speaks in a telephone conversation. It has been stressed above that short-term speech acoustics always contain the acoustic reflex of the speech sound being made at that instant. Since it is often problematic to exert adequate control under these circumstances, the idea with long-term features is that, by taking average values from a long stretch of speech, acoustic features that are due to individual sounds will cancel each other out, and thus leave the acoustic features which characterise the speaker.

Long-term fundamental frequency

[99.720] One feature that readily lends itself to long-term comparison is fundamental frequency (although the approach can be used for other features like F2 and F3, or the cepstrum). Aspects of long-term fundamental frequency (LTF0) are commonly used in the forensic comparison of voice samples, and are discussed and exemplified in detail in Rose (2002, Ch 8). In LTF0, a speaker's F0 is measured (another term is "sampled") at very short intervals – perhaps of the order of every hundredth of a second – throughout her or his speech. A LTF0 "distribution" is then built-up from these individual F0 measurements which has certain statistical properties. Speech samples can then be compared with respect to these properties.

Figure 22 shows a LTF0 distribution for a male speaker of Australian English. This speaker's F0 was measured by computer from the recording of an approximately 30 second passage he read out in the laboratory. F0 was sampled every 20 milliseconds, and there were 689 F0 measurements (or "observations") made. This means that the distribution was built up from about (689 * 0.02 =) 14 seconds of voiced speech, which means that the rest of the 30 seconds – approximately half – was taken up by voiceless speech sounds and pauses. The horizontal axis shows fundamental frequency in Hz, and the vertical axis, labelled "count", shows the number of F0 observations made.

FIGURE 22 Histogram of long-term fundamental frequency distribution of a male speaker of Australian English from a 30 second read-out passage



The F0 observations have been grouped into bins of 5 Hz width, starting from 50 Hz, and the number of observations in each 5 Hz bin counted and plotted. (The choice of bin width and where to start are up to the analyst: the bin width determines the amount of smoothing; the starting point will affect details of the profile.) So, eg, it can be seen that were very few F0 observations – about six – in the interval between 100 Hz and 105 Hz, and that the majority of F0 observations – just under 212 in all – were in the interval between 115 Hz and 120 Hz. These figures represent about (6/689 =) less than 1% and (212/689 =) about 31% of all the F0 observations in the distribution. This method of plotting a distribution is called a "histogram".

The profile of the LTF0 distribution in Figure 22 is typical for many, perhaps most, languages. It is unimodal – it has one main peak – and positively skewed – the F0 observations tend to cluster more in the lower frequency bins than in the higher. There is a small number of isolated observations in the bins between 50 Hz and 90 Hz. These probably represent phonation that was creaky, since creak has a typically low fundamental frequency. A few of the observations above 180 Hz represent F0 values extracted in error, where visual examination of the wave-form showed the computer to have extracted an F0 value in the absence of true periodicity. (This, again, is typical for automatic F0 extraction, and makes it mandatory to always check the results, and perhaps manually remove any incorrect values that may have skewed them.) In this case there were so few incorrect values that any effect on overall results can be discounted.

Long-term average F0

[99.730] The dimension of a LTF0 distribution that is most commonly compared forensically is its mean, when it is called the "long-term average (or mean) F0" (LTAF0). This is simply the sum of the F0 measurements divided by the number of observations. The LTAF0 was found to be a strong parameter in early automatic speaker identification experiments. The LTAF0 for the data in Figure 22 was 124 Hz, which is typical for unemotionally read-out male Australian English speech.

Sometimes it makes sense to omit the putative low F0 creaked values from the sum, since if there are a lot of them, it can pull the mean down. (Samples thus treated might then be compared separately with respect to LTAF0, and the amount of low frequency F0 values.)

Long-term modal F0

[99.740] Since the LTF0 distribution is positively skewed, its mean value will be higher than the F0 value that occurred the most often, which is its "modal F0". For this distribution, the modal F0 was 118 Hz. Although this depends on the start point chosen, the modal F0 is usually in the bin with the highest count (the modal bin), and this is the case here: the bin with the highest count is in the interval between 115 Hz and 120 Hz. During this passage, then, the speaker's vocal cords were, on average, vibrating at a rate of 124 times a second, but for a large portion of the time – about 30% – they were vibrating at a rate between 115 Hz and 120 Hz. For skewed distributions typical of LTF0 it often makes more sense – better likelihood ratio values will be obtained on average – to compare the samples' LT modal F0 values than LT mean values.

Long-term standard deviation F0

[99.750] The LTAF0 value is a measure of central tendency – it indicates where the centre of a distribution lies. Another dimension that is sometimes used to compare samples is a measure of their spread – whether most of the speaker's F0 values are concentrated in a relatively narrow F0 region, or whether they spread out.

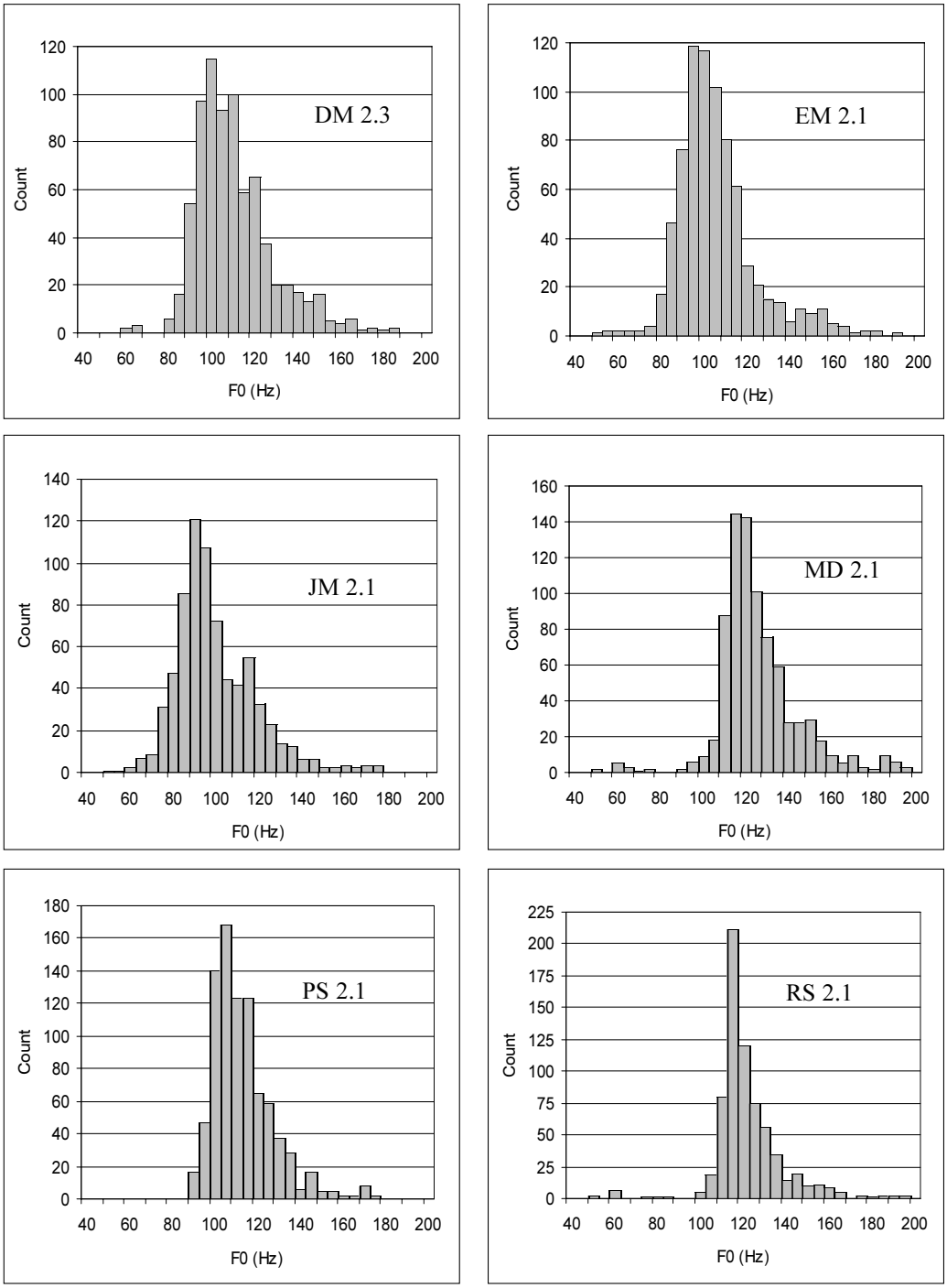
The appropriate statistical measure of this is called the standard deviation, which quantifies the average deviation of each F0 observation away from the mean. When the distribution is not skewed, a large proportion – 94% – of all its observations will lie between +/- 2 standard deviations around the mean. The LTSDF0 for these data is 16 Hz, so, if it was not skewed, about 94% of its F0 observations would lie in the range between (124 - 32 =) 92 Hz and (124 + 32 =) 156 Hz. Since this F0 distribution is skewed, this percentage will not be completely correct, but since its skewing is only moderate, it can be seen that a higher percentage of F0 observations do, in fact, fall in the interval between 92 Hz and 156 Hz.

The LTF0 distribution can be further characterised, and quantified, with respect to its amount of skew, and how flat or peaked it is (called its "kurtosis"). Standard deviation, skew and kurtosis are sometimes referred to as the second, third and fourth moments around the mean respectively.

Between-speaker comparison in long-term F0 distribution

[99.760] Figure 23 shows the LTF0 distributions for another five male Australians, obtained under identical circumstances to those in Figure 22. (Speaker RS's distribution, already shown in Figure 22, has been repeated to facilitate comparison.)

FIGURE 23 Histograms of long-term F0 distributions of six Australian males from a 30-second reading passage



It can be seen in Figure 23 that all distributions are positively skewed and unimodal. Summary statistics (giving mean, standard deviation and modal long-term F0) are given in Table 9.

TABLE 9 Summary statistics on between-speaker variation in the long-term F0 distribution of six similar-sounding Australian males from a ca 30-second read-out passage

Speaker	No of observations	LTAF0 (Hz)	LTSDF0 (Hz)	LTF0 Mode (Hz)
DM 2.3	754	112.3	18.0	99.2
EM 2.1	762	107.5	18.4	98.7
JM 2.1	732	101.4	18.4	94.4
RS 2.1	689	123.8	16.1	117.6
PS 2.1	854	114.6	14.3	107.0
MD 2.1	811	127.6	19.2	122.1

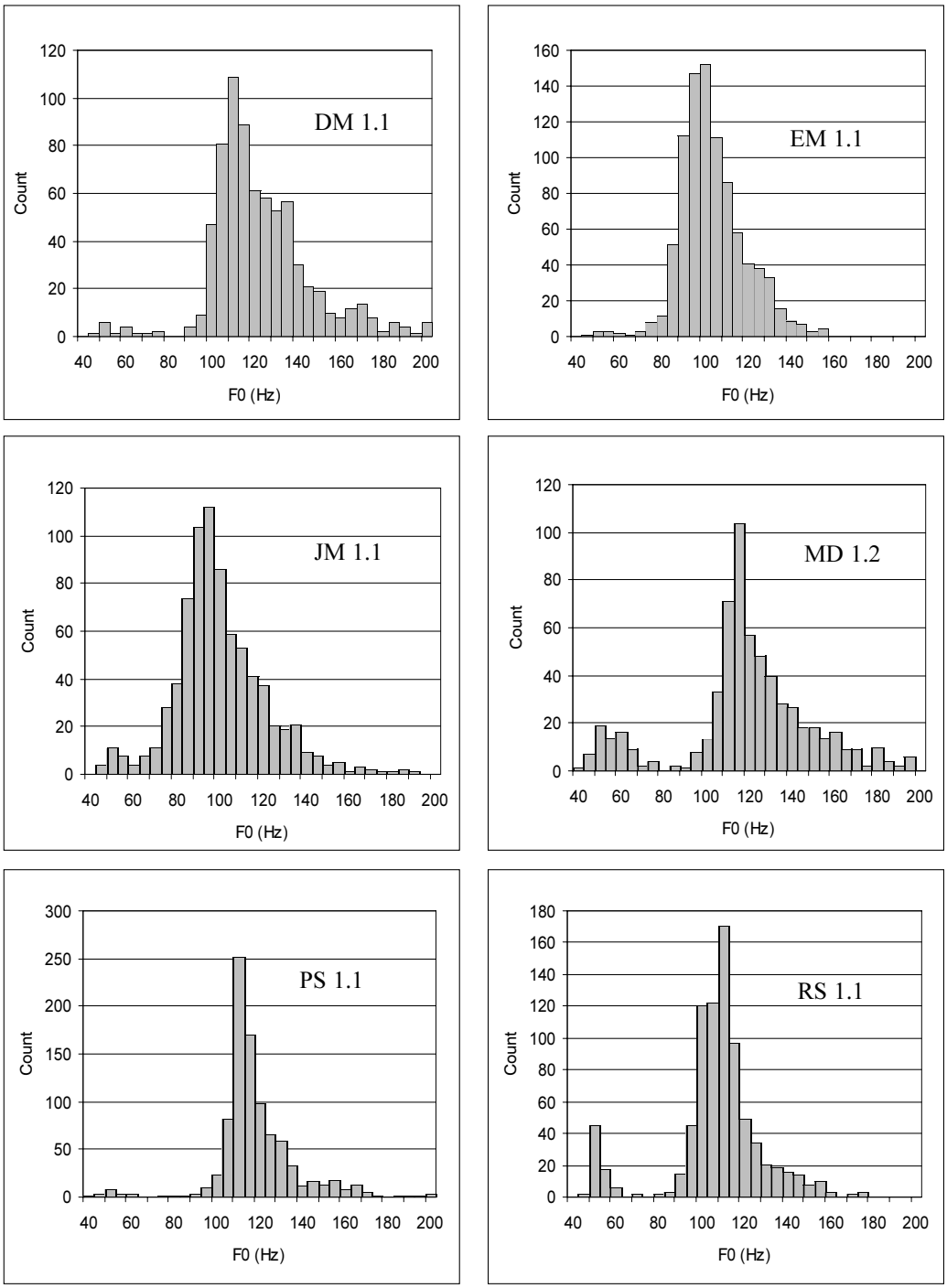
Some clear between-speaker differences in LTAF0 and LT modal F0 can be seen: JM's and EM's distributions can be seen to be slightly lower than the others', and MD's slightly higher. The lower and higher mean and modal LTF0 for these speakers can be seen in Table 9. There are also differences in the amount of low F0 values (ie, below about 90 Hz) present (as already mentioned, these probably correspond to creaky phonation). RS, EM and MD have quite a few, whereas DM has only a few and PS has none. There do not seem to be any big between-speaker differences in LTSDF0, however. (RS's distribution looks somewhat narrower than the others', but this is not reflected in his LTSDF0, probably because of the amount of his low frequency observations.)

Despite the above-mentioned differences, the long-term F0 distribution for least one pair – DM and EM – is clearly very similar in many respects. This is one of those occasions, mentioned above, where two different speakers can show considerable agreement. One would probably be more likely to observe the difference between most aspects of these two distributions if they had come from the same speaker rather than different speakers.

Within-speaker variation in long-term fundamental frequency

[99.770] The reader will recall that all forensic speaker identification features also show within-speaker variation. In order to demonstrate this for LTF0, Figure 24 shows long-term F0 distributions for the same six males when they read very nearly the same passage at least one year *earlier*, and Table 10 gives the numerical data.

FIGURE 24 Histograms of long-term F0 distributions of the same six Australian males, and from very nearly the same reading passage as in Figure 23



Comparison with Figure 23 shows that whereas each speaker's profile has remained fairly similar across the space of a year or more, it is by no means invariant. Within-speaker variation is observable in the amount of low frequency (ie, creak) F0 observations: all speakers except EM have more in their first reading (Figure 24) than in their second (Figure 23), for example. Some speakers have also appeared to have shifted slightly: DM, MD and PS have shifted up; RS has shifted down. These shifts mean that DM and EM were not so similar in the first reading as in the second.

TABLE 10 Summary statistics on between-speaker variation in the long-term F0 distribution of six similar-sounding Australian males from very nearly the same passage as in Table 9 read out at least one year earlier

Speaker	No of observations	LTAF0 (Hz)	LTSDF0 (Hz)	LTF0 Mode (Hz)
DM 1.1	725	124.8	23.0	112.9
EM 1.2	900	105.4	15.24	95.6
JM 1.1	775	101.9	21.11	95.7
RS 1.1	826	108.9	21.9	110.3
PS 1.1	908	119.6	18.4	112.9
MD 1.2	613	120.9	30.5	116.3

All the above within-speaker variation is typical for LTF0. Nevertheless, the data presented here still conform to one of the essential criteria for forensic-phonetic features, namely, that they still show greater between-speaker variation than within-speaker variation – at least for long-term mean and modal F0 values. For the long-term mean, the between-speaker variation is about twice the within-speaker variation, and for the long-term mode, the between-speaker variation is about four times bigger.

These ratios are actually not very big, and their magnitude can perhaps be better appreciated by comparison of average differences. The average differences for within-speaker LTF0 mean and modal F0 for these data are about 7 Hz and 6 Hz respectively, compared to about 11 Hz and 12 Hz respectively for the average between-speaker difference.

In contrast to the long-term mean and mode, however, the long-term standard deviation for these data actually shows more variation within-speaker than between-speaker, and therefore would be no use for discriminating between same-speaker and different-speaker pairs from these data.

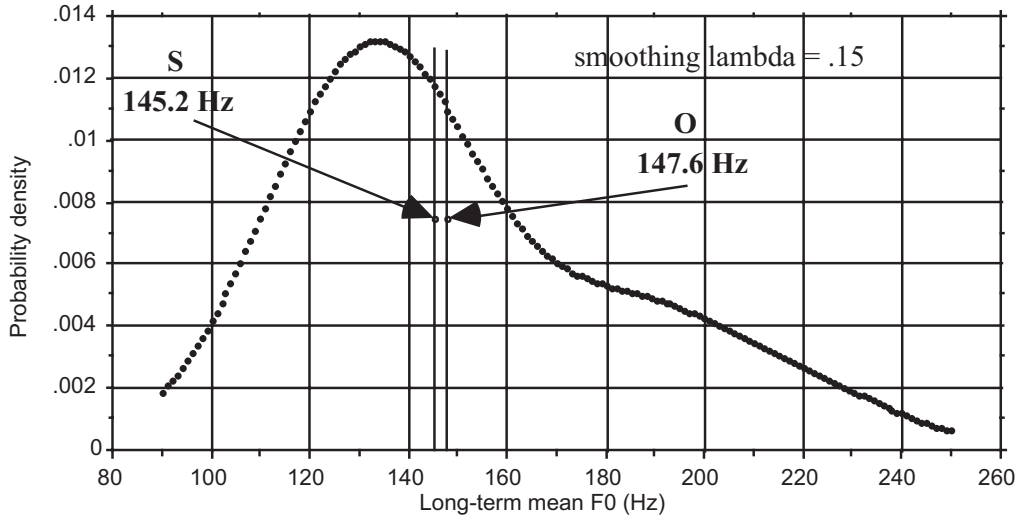
It needs to be stressed that the data used to illustrate these LTF0 features were obtained under very similar, well-controlled circumstances atypical of real-world forensic comparison. Therefore it is to be expected that the within-speaker variation in LTF0 features will be greater than that shown. This factor, on the other hand, will be offset at least to a certain extent by the fact that the six speakers used were selected for their typicality and homogeneity, and therefore can be expected to show rather small between-speaker variation. Other things being equal, then, one is probably justified in expecting, on average, useable but small likelihood ratios from LTF0 features.

Since LTF0 is sensitive to paralinguistic and extralinguistic factors, it cannot be used indiscriminately, and samples to be compared should be basically a priori comparable with respect to these factors.

Example of likelihood ratio comparison using LTF0

[99.780] Figure 25 shows a comparison, from case work, between suspect and offender voice samples in long-term F0. The language is Cantonese. The suspect's mean long-term fundamental frequency, estimated from 14 separate phone calls, was 145.2 Hz. The vertical line linked to "S" shows this value in Figure 25. The offender sample, measured in one incriminating phone call, was 147.5 Hz. This is marked with "O" in Figure 25. The curvy line represents an estimate of the distribution of male Cantonese speakers' long-term F0 measurements. It was constructed, using a method called "kernel density estimation", from the means of 17 different Cantonese males speaking over the phone under comparable circumstances.

FIGURE 25 Mean Suspect and Offender LTF0 samples compared against a reference distribution of LTF0 in Cantonese



Kernel density estimation is a method of modelling distributions which deviate from normality – in being skewed or multimodal, for example. Such distributions are quite often found in forensic-phonetic features. It can be seen that the reference distribution in this case was somewhat positively skewed, which warranted the use of the method.

The actual formula used to estimate the LR value for these data from a kernel density is shown at Formula 11. This formula is taken from Aitken (1995, p 188). The formula highlights the fact that the estimate is based on principles derived from probability theory and logic as well as speech acoustics.

Formula 11

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

where

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

and

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

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

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

σ^2 = within-speaker variance

λ = smoothing factor for kernel density estimate

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

k = number of kernel functions

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

The 2.3 Hz difference in LTF0 between the suspect and offender values shown in Figure 25 is extremely small relative to an expected range for a given speaker. A speaker's F0 range (or "compass") is sometimes estimated at twice the standard deviation above and below the speaker's average F0 value. For data of this kind, the expected range is about 100 Hz, so the value of 2.3 Hz represents only about 2% of compass (for the 17 reference males, and also for the suspect and offender, the standard deviation was about 26 Hz).

Although both these mean long-term F0 values are very similar because the difference between them is very small, it can be seen that they are also very typical: they occur close to the centre of gravity of the distribution of the population. Thus it is to be expected, since the LR is the ratio of similarity to typicality, that the LR for a difference of this size but of this typicality is not going to be very big. Using Formula 11 to estimate the LR, it is found to be about 2.3. One would only be about 2.3 times more likely to get this kind of difference – even though it is so small – assuming the samples came from the same rather than different speakers.

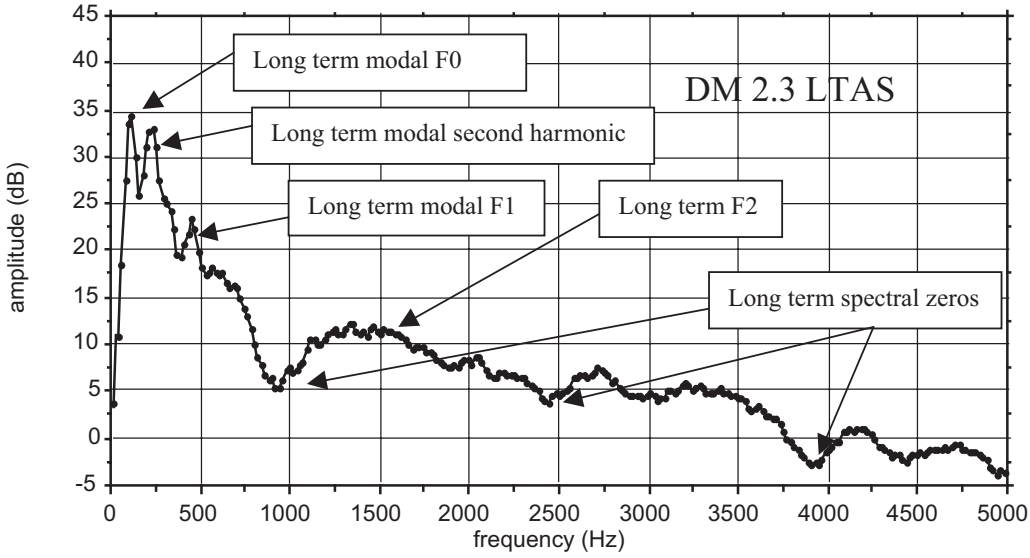
This is a good illustration of the principle that similarity (or difference) on its own is not enough to evaluate speech samples. The effect of typicality can also be appreciated from the fact that, had the samples been located at the extreme upper end of the distribution – say at 230 Hz and 232.3 Hz – the LR would have been greater: about 13 times more likely assuming same than different speakers. It is also instructive to see how different the offender sample would have to be from a suspect value of 145.2 Hz in order for the difference to be evaluated as more probable assuming different speakers. For these data, distances greater than about 25 Hz in either direction will be evaluated with a LR less than 1.

Long-term average spectrum

[99.790] Another long-term feature that will sometimes be encountered in forensic speech comparison is the long-term average spectrum (LTAS). This is calculated by averaging a succession of short-term spectra over a long period of speech. The long-term average spectrum thus shows the average distribution of acoustic energy in the speaker's voice as the speaker is speaking on a particular occasion, like during a telephone call.

Figure 26 shows the long-term average spectrum of about 30 seconds of read-out speech from a young Australian male. Its axes should by now be familiar: frequency (increasing along the bottom from 0 to 5000 Hz) and amplitude (increasing vertically, in dB). This spectrum is the average of some thousands of short-term spectra. (These are yet again another type of spectrum called a "fast fourier transform" or FFT.) The LTAS profile rises abruptly to a peak at about 100 Hz, from which the amplitude falls, first fairly abruptly to about 1000 Hz, and after that more gently to 5000 Hz.

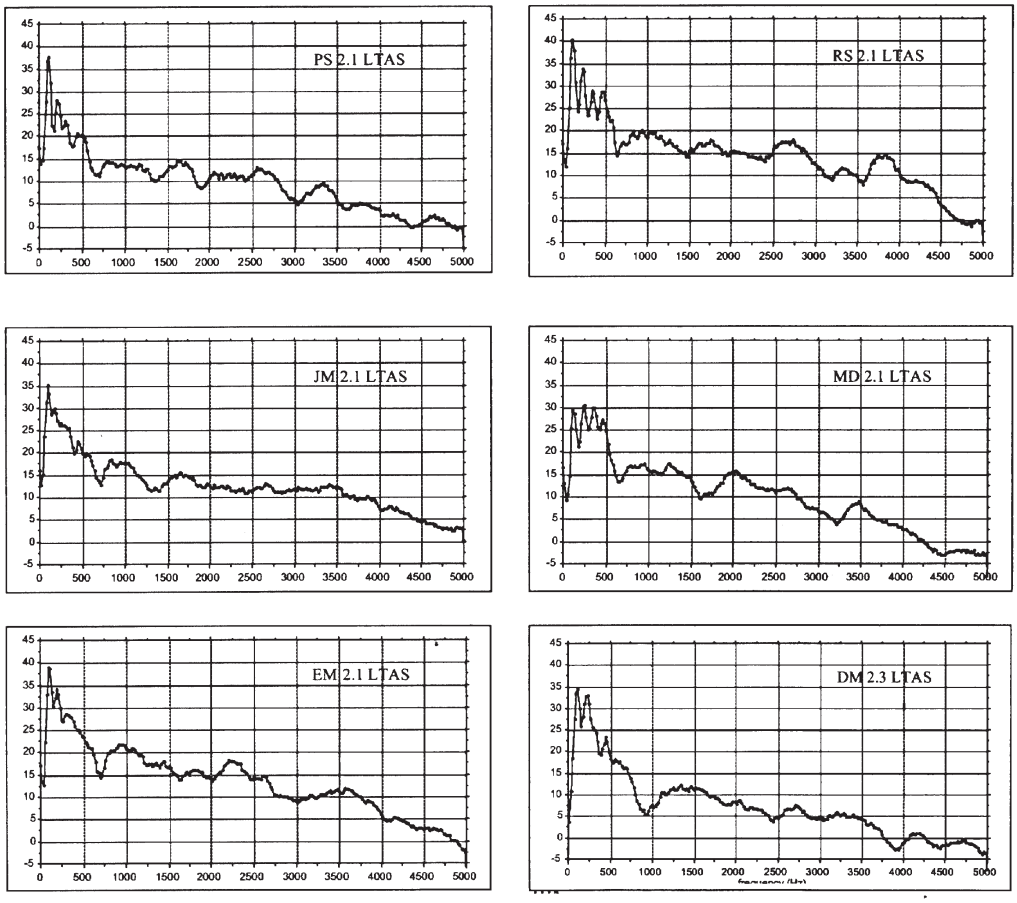
FIGURE 26 Long-term average spectrum for a monologue from an Australian male speaker



Some fine detail in the spectral profile can be identified: the lowest frequency peak, at about 100 Hz, is probably the long-term modal fundamental frequency. (This is the same passage as that used for DM in the LTF0 illustration above, so this value is actually 99 Hz.) The next spectral peak, just below 250 Hz, is probably the long-term modal second harmonic (LT H2) since it is about double the F0. (Harmonics are energy present at whole number multiples of the F0. Harmonics higher than the second have not been separately resolved.) The next highest spectral peak, just below 500 Hz, might reflect the long-term modal first formant. The broad band peak at about 1500 Hz probably represents the long-term second formant.

Of particular note are the spectral dips in the LTAS. This particular spectrum has three approximately equidistant dips: a conspicuous one just below 1000 Hz; another, less conspicuous one about 1500 Hz above it at just below 2500 Hz; and another just below 4000 Hz. These dips probably represent absorption of acoustic energy, or "zeros", at particular frequencies, and carry potentially speaker-dependent information.

FIGURE 27 Long-term average spectra for six Australian males from a 30-seconds reading passage



In order to show some between-speaker differences in LTAS, Figure 27 presents long-term average spectra from six similar-sounding Australian males. These include DM, already illustrated in Figure 26. They can be seen to differ in several respects, including presence and location of zeros; relative amplitude of the modal F0; and resolution of lower modal harmonics. The long-term average spectrum is still subject to within-speaker variation, of course.

The LTAS, or cepstrally smoothed transforms of it, contains much speaker-dependent information and has been shown to be a very powerful parameter in automatic speaker recognition, where sophisticated channel-normalising techniques are applied to counteract the effect of transmission over different – especially telephone – channels, a factor to which the LTAS is particularly sensitive. The conditions under which it can be used forensically are probably rather limited, however, and are discussed in Rose (2002, Ch 8).

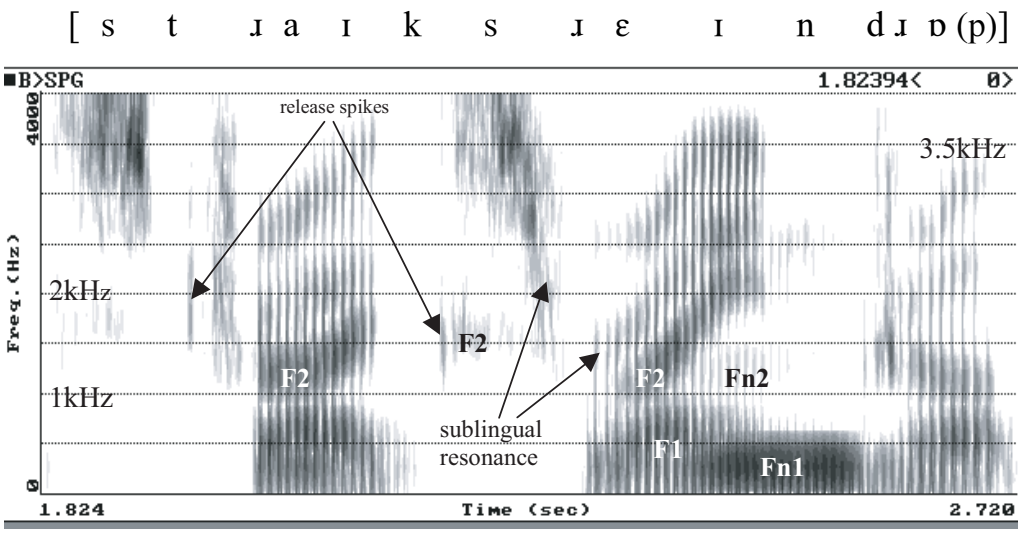
Spectrograms

[99.800] A spectrogram is a very common way of displaying speech acoustics, and may often be encountered in forensic reports. Spectrograms, their interpretation, and how they relate to the theory of speech production and other representations of speech acoustics are discussed in detail in Rose (2002, Ch 8).

A spectrogram is a very useful picture of the distribution of acoustic energy in speech. Figure 28 shows an example of a spectrogram – specifically a wide-band spectrogram – of part of the utterance "When the sunlight strikes raindrops in the air" corresponding to "strikes raindro". The speaker is an Australian male. The spectrogram was generated by computer, using a conventional software package. Because the utterance was recorded in a sound-proofed studio, the spectrogram is nice and clear.

The spectrogram has two main dimensions: time and frequency. Time is shown horizontally and progresses from left to right. (Some spectrograms of Arabic can be found going right to left, highlighting the conventional nature of the orientation.) The time in seconds of the start and end of the spectrogram is shown in the bottom left and right corners. It can be seen that the two words were spoken in a little more than about (2.72 - 1.824 =) a ninth of a second. Precise duration measurements can be made on such computer-generated wide-band spectrograms using the cursor.

FIGURE 28 Annotated wide band spectrogram of "strikes raindro(ps)" spoken by an Australian male



Panel at top shows aligned broad phonetic transcription

The second spectrographic dimension, frequency, is shown vertically. In the spectrogram in Figure 28, the frequency range goes from 0 Hz at the bottom to 4000 Hz at the top, and there are horizontal grid lines at 500 Hz intervals (0 Hz, 500 Hz, 1000 Hz etc). Because of their size, it is often more convenient to express these frequencies in thousands of Hz, or kiloherz (kHz). The range of frequencies of interest present in a voice is quite large. It goes from about 50 Hz to about 7 kHz. In forensics, this range is usually considerably smaller, due to the influence of the telephone which, amongst other unfortunate things, limits the range to about 3000 Hz (from about 350 Hz to about 3.5 kHz).

Not only is the frequency range wide in speech, but there are always many different frequencies – eg, fundamental frequency and formant frequencies – involved at the same time, and these also change rapidly over time. A spectrogram shows how much energy is present at what frequencies, and how this changes over time. The amount of energy is shown by the darkness of the trace – the greater the amount of energy the darker the trace. One term for amount of energy, as explained earlier, is amplitude – thus amplitude constitutes an effective third dimension to the spectrogram in addition to time and frequency.

The spectrogram in Figure 28 shows several different-looking regions of acoustic energy, each corresponding approximately to the speech sounds in "strikes raindro(ps)". To help the reader relate the acoustics to the speech sounds involved, these are shown above the spectrogram, in broad phonetic transcription, approximately aligned with the acoustics.

The acoustics corresponding to the two voiceless alveolar fricative /s/ sounds in "strikes" can be easily seen as energy located mostly above 3.0 kHz. This /s/ energy would extend far above the 4.0 kHz upper limit of the display.

The three parts consisting of closely spaced vertical striations correspond to the two diphthongs – /aɪ/ in "strikes", /ɛɪ/ in "rain" – and the short monophthong /ɒ/ in "drops". The rhotic /r/ in "strikes" and "rain" is also partially encoded at the beginning of these striated sections. The vertical striations reflect the acoustic energy from the vibration of the vocal cords that arises when the cords are coming together. In the part corresponding to the /(r)ɛɪ/ in "rain", 20 striations can be counted. These occur in about 0.161 seconds, which means that the speaker's vocal cords were vibrating at an average rate of (1/[0.161 /20] =)124 Hz during this sequence of sounds (ie, his average F0 was 124 Hz).

The gaps immediately after the first /s/ and before the second relate to the voiceless stops /t/ and /k/ in "strikes". The short isolated vertical striations after these gaps – the first between about 1.5 kHz and 2.5 kHz, and the second between about 1.2 kHz and 1.75 kHz – show the energy from the release of the stops.

The alveolar nasal /n/ in "rain" is partly reflected in most of the large gap after the /ɛɪ/.

Also very clear in the voiced portions of Figure 28 (ie, those portions with striations) are several thick black bands. In the /(r)aɪ/ portion of "strikes", eg, there is a black band starting at about 1.2 kHz, curving up in frequency and ending at about 1.5 kHz.

These band-like areas of high acoustic energy are the formants. The one just referred in "strikes" is its second formant. Two more formants, both rising in frequency, are visible above the F2 which do not have quite as much amplitude (are not as dark).

In previous sections, it has been explained how formants, or vocal tract resonances, relate to the frequencies of vibration of the air in the mouth and the throat (and the nose if the soft palate is down), and that the frequency of the lowest two or three formants are responsible for cueing the phonetic quality of speech sounds, especially vowels. The frequencies of the vibration of the air in the mouth and throat change depending on what shape the mouth is, and what shape the mouth is depends primarily on the sound being produced.

Formants have been described in previous sections with respect to their centre frequency at a particular instant of time, but not with respect to how that frequency changes over time. The spectrogram shows these time-varying changes nicely. It can be seen that the frequencies of the formants do not remain static, but change over time, and this reflects the movements of the tongue, lips and soft palate in the speaker's supralaryngeal vocal tract as he said "strikes raindro(ps)". Some particularly large changes in formant frequency are observable (that was why these two words were chosen). The reader should try to see how, eg, F2 in /rɛɪ/ shoots up from about 1 kHz to about 2 kHz. The large changes are due in part to the difference between the acoustic reflexes of /r/, which is associated with a set of very low resonances, and the offglide

of /ɛɪ/ which, as a high front unrounded vowel, consequently has high F2 and F3, and as a vowel has high F4.

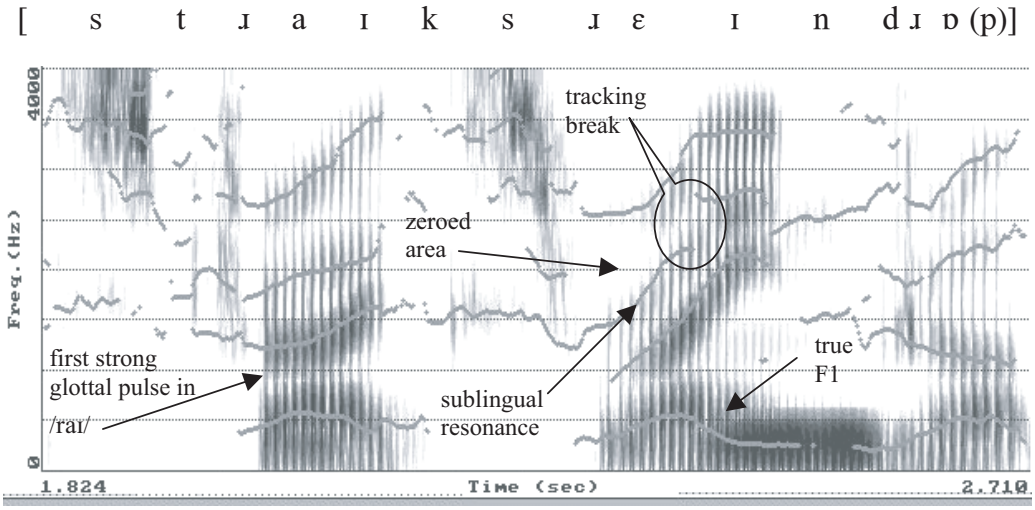
The low frequency area of the highest amplitude extending into the gap for the /n/ in "rain" is the first nasal formant (Fn1). This reflects the vibration of the air both in the mouth and throat, and the nasal cavities, since the soft palate is down for /n/. Fn1 can be seen to have started before the end of the /ɛɪ/ region, which shows that the speaker had lowered his soft palate during the production of the /ɛɪ/ in anticipation of the following /n/. The second nasal formant is also already visible in the /ɛɪ/ at about 1.25 kHz, although it is not as clear as the first.

This overlap in articulatory gestures is typical of speech and is the reason why the correspondence between the speech sounds and the different-looking stretches of acoustic energy was described above as approximate. It is approximate because speech sounds are not encoded sequentially, but overlappingly, in the acoustics, so that any stretch of acoustics contains information on more than one sound. There are, to be sure, clear spectral discontinuitites in the speech acoustics, eg the abrupt drop in overall amplitude at the end of the /ɛɪ/ which reflects the point at which the tongue tip makes contact with the alveolar ridge for the /n/. However, it is not possible to use such discontinuities to segment the acoustics in such a way that one stretch of acoustics corresponds exclusively to one speech sound. Another clear instance of this multiple encoding is the stretch of acoustics identified as corresponding mostly to the second /s/. Its F2 is faint, but can still be made out, at 1.5 kHz. However, its next highest resonance descends rapidly from above 3 kHz to about 2.5 kHz as part of the realisation of the following /r/. The decrease in the frequency of this resonance, which is admittedly not very clear in the spectrogram, reflects the formation of the sublingual cavity which is one of the essential parts of the English /r/.

Fairly sophisticated digital algorithms exist to estimate and track formant centre frequencies, and the result of applying one to the "strikes raindro(ps)" and superimposing it on the spectrogram in Figure 28 is shown in Figure 29, where the formant centre frequencies have been estimated by the method of linear prediction, and their trajectories tracked over time.

Although these algorithms are automatically performed by computer, the analyst still always has to choose appropriate values for various analysis settings before the programs can be implemented to generate the spectrogram, and do the formant estimation and tracking. Thus, just as with any "automatic" approach, some interaction with the analyst is always necessary. These settings are too complex to explain in this Chapter but many of them can have considerable effect on the outcome, and therefore it is also essential to know them if an analysis has to be replicated.[1] Thus they must always be clear in a forensic report. Also, of course, once decided on, they cannot be changed in the course of an investigation to accommodate to different circumstances.

FIGURE 29 Wide-band spectrogram of "strikes raindro(ps)" with formant centre frequencies estimated by linear prediction and their trajectories tracked



It can be seen from Figure 29 that the programs have estimated and tracked the centre frequencies of the formants well, in that the trajectories appear to go through the middle of most of the formant bands (the estimate in the F2 of "rain" /(r)ɛɪ/ seems slightly too high, however). Even the weak F2 in the second /s/ has been picked up. Given that this utterance contains sounds, like the nasal and the rhotic, which have exceedingly complex acoustic structure, this is a good result.

There have been some minor errors in tracking. A relatively easy one for the reader to see occurs between the first formant in /rɛɪ/ and the F_n1 associated with the following nasal. The true F1 can be seen to continue at about 500 Hz before falling in frequency "into" the nasal resonance. Continuities between the F-pattern of the second /s/ and the following voiced portion are also not correct,[2] although the tracking break noted is probably real. These errors highlight the point that there is always an element of interpretation necessary in the analysis of computer-generated spectrograms and extracted resonances, and this is informed by a knowledge of the acoustic theory of speech production in conjunction with the perceived sound and its phonological representation.

The computer program can be interrogated for the centre frequency value (or bandwidth) of any formant at any point in time. Thus at the first strong glottal pulse in /raɪ/ in "strike" the LP estimated values of the formant centre frequencies were 443 Hz (F1), 1205 Hz (F2), 1785 Hz (F3), and 2640 Hz (F4). The reader should check these values visually.

Formant centre frequencies can also be, and sometimes have to be, estimated by eye. It has been demonstrated that, with good recordings, experienced spectrographers are as accurate as computers in the estimation of F1 and F2, where overall absolute errors are of the order of +/- 60 Hz, but are worse for F3, where the computer still performs at +/- 60 Hz, but the humans degrade to +/- 110 Hz. However, as might be expected, humans outperform computers in specific circumstances where additional top-down interpretation can be applied. This is, eg, when two formants are close, as in F1 and F2 for low back vowels; where formants have large bandwidths or where nasalisation is concerned. (The linear prediction model does not incorporate spectral zeros, and has to model a spectrum containing zeros, as in all examples of nasalisation or nasal sounds, with resonances alone.)

It can also be assumed that humans will usually outperform computers when working with noisy or degraded speech, since the accuracy of linear prediction decreases considerably under noise. Since degraded recordings are the norm in forensics, it is vital to visually assess the performance of automated extraction.

Spectrograms are a very useful tool in the forensic comparison of speech samples. They can be used as an easy visual reference for the extraction of acoustic features. For example, one might want to compare two speech samples with respect to the lowest frequency of the sublingual cavity resonance in prevocalic /r/ in stressed syllables. This point can be easily visually located using a spectrogram. In Figure 29, eg, the reader can see where the lowest value in the sub-lingual cavity resonance in /rɛɪ/ is located. The actual value can be then extracted automatically by locating the cursor at this spot, clicking, and storing the result for further statistical processing. Spectrograms can also be useful for defining a particular acoustic feature used in comparing samples, or for demonstrating the absence or presence of a particular acoustic feature in a report. They can be useful in checking visually for the acoustic reflex of a particular auditory feature, in order to quantify it. Or they can be used for simply eyeballing the acoustics for any anomaly that might not have been audible. However, spectrograms have no unique properties as wholes: they cannot constitute features in themselves with which to compare forensic speech samples. This point is taken further in the discussion below on voiceprints: see **[99.820]**.

1 The most important settings are the filter order, preemphasis and frame length for the linear prediction and tracking; and the bandwidth for the spectrogram. Another important datum is the sampling frequency - how many times per second the analog signal was measured as it was digitised (went into the computer). The settings used for Figures 28 and 29 were: 20 order filter; 0.90 preemphasis; 25 ms frame length; 234 Hz spectrogram bandwidth; and 16 kHz sampling frequency.

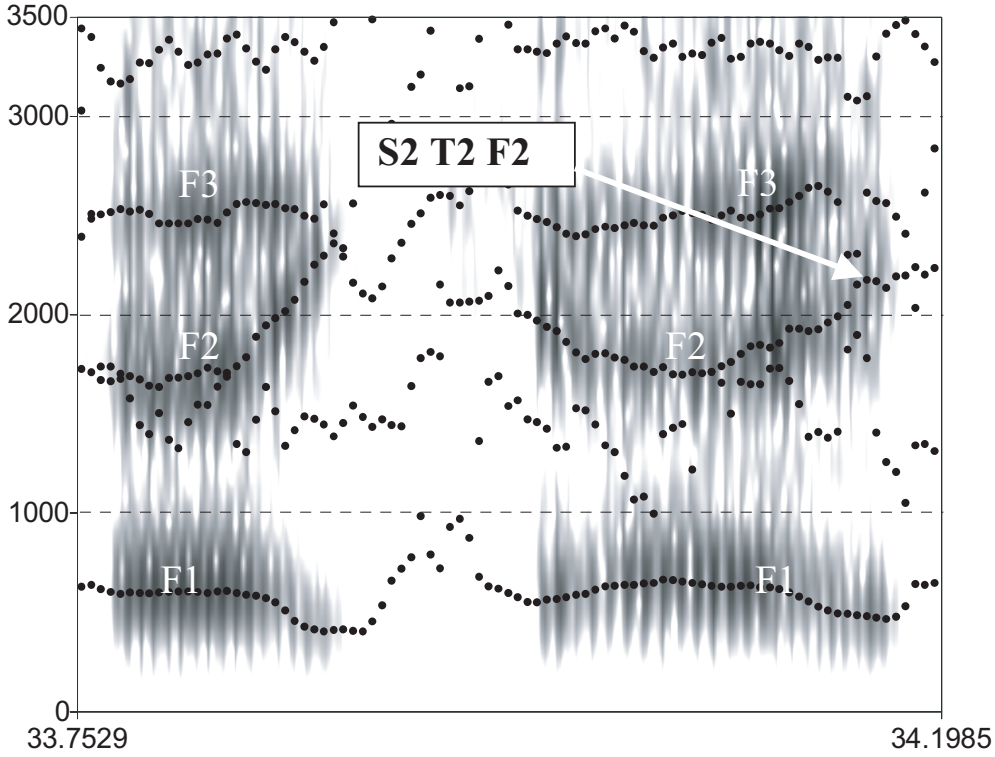
2 F2 in the second /s/ has been tracked continuous with the following third (sublingual) resonance, whereas in theory it should connect with F2, and the sublingual resonance on /rɛɪ/ should be continuous with the rapidly descending resonance on the /s/ (which has not been well tracked as a single resonance).

The tracking break in the third resonance in /rɛɪ/ may not be an error, but reflect a real complexity in the relationship between /r/ and its acoustics. The articulation of /r/ characteristically involves a sublingual cavity, which is responsible for an extra resonance. This extra resonance can be clearly seen in /rɛɪ/ increasing from a frequency of about 1.5 kHz. The sublingual cavity also contributes a zero at about 2 kHz which attenuates the true F3. The clearly visible absence of energy centred at 2 kHz at the beginning of the /rɛɪ/ portion is due to this zero. As the articulators move from the /r/ to the first portion of the /ɛɪ/ diphthong, the extra resonance and the zero cancel each other out and the true F3, no longer attenuated by the zero, emerges. The observed tracking discontinuity might therefore reflect the end of the acoustic effects of the sub-lingual cavity and the beginning of the true F3.

Example of likelihood ratio comparison using formants as linguistic-acoustic features

[99.810] As an example of a likelihood ratio comparison with a linguistic-acoustic feature is taken the acoustic feature in "okay", called *okay* $_{S2\ T2\ F2}$, mentioned in at **[99.250]** above. Figure 30 shows what the feature looks like in a wide-band spectrogram of the type just illustrated and explained, and also demonstrates the use of a spectrogram to show a particular acoustic feature.

FIGURE 30 Spectrogram of offender "okay" showing feature "okay $_{S2\ T2\ F2}$"



Horizontal axis = time in seconds; vertical axis = frequency in Hz.

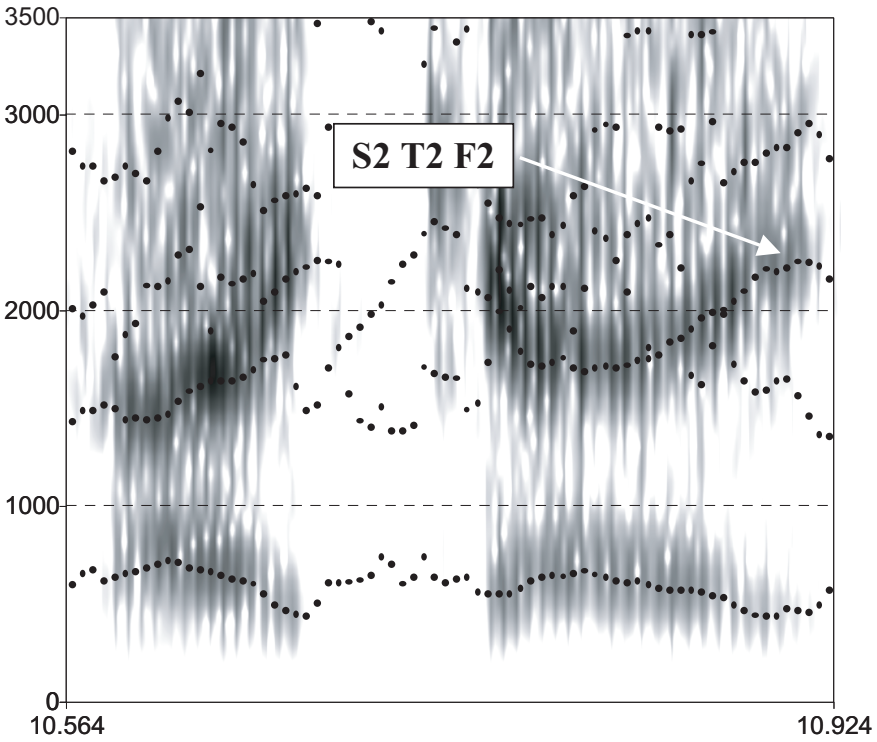
Figure 30 shows a wide-band spectrogram of "okay" from the offender with formant centre frequencies estimated by linear prediction and tracked. The "okay" was excerpted from an intercepted telephone conversation. It can be seen that this particular token of "okay" was said in a little less than (34.199 - 33.753 =) 44.5 hundredths of a second. The spectrogram's frequency range goes from 0 Hz at the bottom to 3500 Hz at the top, and there are grid lines at kiloherz intervals.

The spectrogram in Figure 30 shows two major regions of acoustic energy with vertical glottal pulse striations, one corresponding to the first syllable of "okay", and the second to the -*ay* part of the second syllable. The low energy white gap between the two relates to the *k* sound. The formants are very clear as several thick black bands that appear to be running predominantly horizontally.

It can be seen that the frequencies of the formants do not remain static, but change over time, and this reflects the movements of the vocal organs as they produce the sounds in "okay". For example, at the end of the *o* part, F2 shoots up in frequency from about 1600 Hz to about 2300 Hz. This reflects the movement of the tongue body to start the production of the *k*. Or in the *-ay* part, F2 dips in frequency to about 2700 Hz and then rises again to about 2200 Hz. This reflects the tongue-body movement away from the *k* to produce the *-ay* diphthong. In *-kay*, the tongue body first moves towards an initial target (T1) where it is fairly low in the front of the mouth, and then to a second target (T2) where it is high in the front of the mouth. It is actually the frequency value of the second formant at the second diphthongal target in *-kay* ("S2 T2 F2") that is the forensic feature under consideration.

At the point marked S2 T2 F2 on Figure 30, the computer-extracted formant centre frequency is 2210 Hz. So *okay* $_{\text{S2 T2 F2}}$ for this token is 2210 Hz, and this is one of the four offender values that was plotted in Figure 1 at **[99.250]**.

FIGURE 31 Spectrogram of suspect "okay" showing feature "okay $_{\text{S2 T2 F2}}$"



Horizontal axis = time in seconds; vertical axis = frequency in Hz.

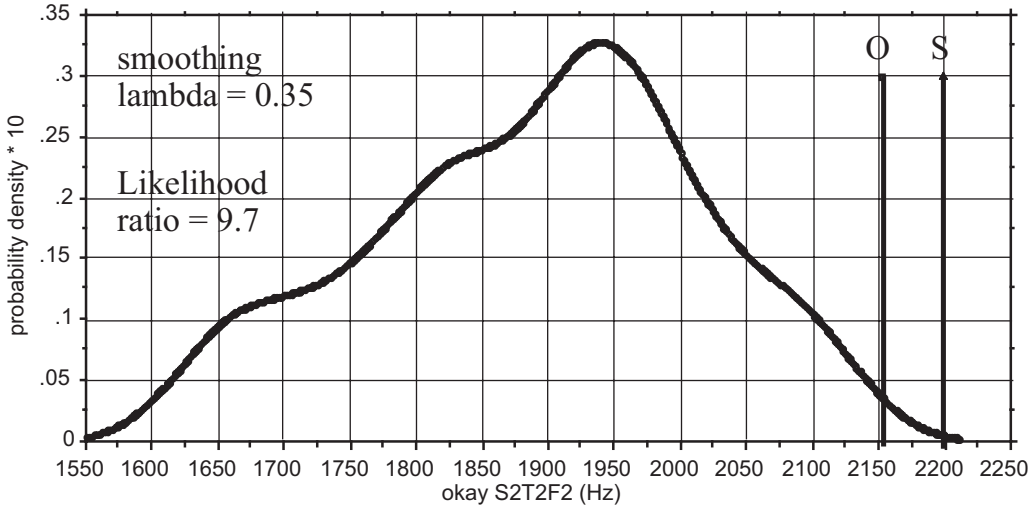
Figure 31 shows a spectrogram with superimposed formant centre frequencies for another "okay" token, this time one produced by the suspect, in an intercepted phone call. The *okay* $_{\text{S2 T2 F2}}$ value for this token is 2259 Hz, and the mean value for the conversation in which this token occurred is also plotted in Figure 1 at **[99.250]**.

It can now be shown how acoustic features like *okay* $_{\text{S2 T2 F2}}$ are used to estimate a likelihood ratio. The evidence has already been shown in Figure 1: recall it consisted of four values for *okay* $_{\text{S2 T2 F2}}$ taken from four "okays" in the offender's speech in one conversation, and seven

mean values for *okay* $_{S2\ T2\ F2}$ taken from seven different conversations involving the suspect. First, a mean value is calculated for the values of *okay* $_{S2\ T2\ F2}$ in the four offender "okays". This is ([2344 + **2210** + 2139 + 1910] /4 =) 2151 Hz. Next, the mean of the seven suspect means for *okay* $_{S2\ T2\ F2}$, is found, which is ([2404 + 2248 + 2235 + **2259** + 2002 + 2069 + 2174] / 7 =) 2199 Hz. (Note the value of 2259 Hz from the token in Figure 31.) It is the 48 Hz difference between the offender mean of 2151 Hz and the suspect grand mean of 2199 Hz that has to be evaluated using a likelihood ratio.

It was explained above at **[99.80]** how the LR is the ratio of similarity to typicality. It is estimated therefore by comparing the difference between the two samples (this is a measure of the similarity) against the background of the reference population (this quantifies how typical the values are). Figure 32 helps to make this easier to understand. Figure 32 represents the reference population as a probability density – this is the slightly bumpy line which goes up and down. This is an estimate of what the probability is of observing values of the feature in the population at large, and it is constructed using a kernel density method as described above for the LTF0 comparison: see **[99.780]**. The values for *okay* $_{S2\ T2\ F2}$ run along the horizontal axis from 1550 Hz to 2250 Hz.

FIGURE 32 Illustration of estimation of typicality of offender (O) and suspect (S) samples for the forensic-phonetic acoustic feature "okay $_{S2\ T2\ F2}$"



For estimating probabilities of continuous variables of the *okay* $_{S2\ T2\ F2}$ kind, a probability density works in terms of *areas under the curve*. For example, if we want to estimate the probability of getting a value lower than 1900 Hz at random from the population, we calculate the area under the curve from 1900 Hz to 1550 Hz. Figure 32 has been plotted against a background of squares to enable this to be estimated by eye. There are about 41.5 squares under the curve, and there are about 20.5 squares below the 1900 Hz value. This means a probability of (20.5 / 40.5 =) 0.51, or 51%, of observing a value lower than 1900 Hz. This means that, if you chose someone at random from this population, and obtained a mean value for *okay* $_{S2\ T2\ F2}$ from a sample of their "okays", you would expect that about half the time – 50 times in 100 choices of individual, for example – you would obtain a value lower than 1900 Hz.

The location of the mean values of 2151 Hz for the offender and 2199 Hz for the suspect samples is shown in Figure 32 by vertical lines marked O for offender and S for suspect. It can be seen that they lie at the extreme upper end of the skirt of the probability density curve, and are therefore not very typical of the population (the reader might like to calculate at this point just how atypical they are). The area under the curve between 2151 Hz and 2199 Hz is almost the same as the area from 2150 Hz to 2200 Hz and by visual examination this is a bit less than half a square – say one third. Therefore the probability of getting two values between 2150 Hz and 2200 Hz at random from the population is (0.33/41.5 =) 0.008, or less than 1%. Since the denominator of the LR is $p(E \mid DS)$ – the probability of observing the evidence assuming different speakers are involved – this is the estimate of the denominator of the LR.

It would be nice to be able to demonstrate the calculation of the similarity between the offender and suspect samples – the numerator of the LR – with the same ease, but this is unfortunately too complicated for this Chapter and has to be done with the kernel density formula at **[99.780]**. The LR for these data, calculated using Formula 11, is about 9.7, which means that the similarity between the offender and suspect samples is 9.7 times greater than their typicality. This means that one would be about 10 times more likely to observe the difference between the offender and suspect values for *okay* $_{S2\ T2\ F2}$ assuming that they came from the same speaker. This would constitute weak support for the prosecution.

Although the calculation of only one feature has been illustrated, it is important to point out that there are several other features, both auditory and acoustic, in "okay" that have forensic-phonetic potential. Since "okay" occurs with high frequency in conversations, it is clearly a useful word in TFSI.

Voiceprint/aural-spectrographic identification and fingerprints

[99.820] In the context of spectrograms it is important to mention one further approach to forensic speaker identification: aural-spectrographic voice identification. This was originally termed "voiceprint identification", and was first based on visual comparison of spectrograms only. (Spectrograms, as has just been shown, are very useful graphic representations of the acoustics of speech.) Later, a combination of visual examination of spectrograms and auditory response was introduced, with the examiner making a decision after both looking at the spectrograms and listening to the speech samples. Because of the inclusion of the listening part, and also because of problems associated with the term "voiceprint", which naturally invites misleading comparison with "fingerprint", the name of the method was changed by some practitioners to "aural-spectrographic" method.

The idea that everyone has a unique voice that can be fingerprinted and used to identify them forensically is widespread. It surfaces from time to time in movies and television crime series, and in 2002 hit the headlines in conjunction with the Bin Laden voice affair. For example, the CNN announced on the web in December 2002:

> The CIA, FBI and National Security Agency have computers that use special programs to identify voice prints. The idea is that every voice has a unique pattern like a fingerprint.

On 25 November 2002 and 30 September 2002, the *Canberra Times* ran articles that claimed:

> Embedded in the voice template is an encoding of the physical characteristics of the vocal tract that make a voice unique.
>
> Voice prints have much the same legal status now as identity proof as fingerprints, Iris scans and facial scans, but have the advantage of not requiring any special equipment – just a microphone hooked to a computer with the appropriate software.

Given such egregious nonsense, it is really no wonder that forensic speaker identification is not well understood. First of all, yes, voices may very well be unique, but voice samples do not carry any invariant feature, or set of invariant features, that will enable them to be uniquely identified.

They can be identified very well automatically, under ideal circumstances, but this is by using a statistical conjunction of *variable* features. Another, more complicated reason why the statements are misleading is this. Source-filter theory tells us that the radiated acoustics are uniquely determined by the anatomy of the vocal tract that produced them on the occasion of speaking. So it is, indeed, correct to say that the physical characteristics are imprinted in the signal. However, the directionality of the function is one way only: from vocal tract to acoustics. Source-filter theory tells us that a unique vocal tract shape is not recoverable from the acoustics (although it is possible, using vocal tract modelling analysis, to estimate a *possible* anatomy from the acoustics).

Second, it cannot be emphasised enough that, in spite of the name "voiceprint", there is no analog in forensic speech identification to the fingerprint. Because voices are nothing like fingerprints, fingerprints and spectrograms (or voiceprints) are highly dissimilar. Both, to be sure, are *representations* of something purporting to contain information on the individual who produced it. Both are ultimately functions of the individual's anatomy; and both are always degraded by real-world forensic circumstances. Like fingerprints, voices also have the forensically useful property that they can be divorced from their owners and left behind. However, whereas a fingerprint (however partial) is a direct record of an anatomical feature – an individual's friction ridges – a spectrogram is only an indirect record of a person's vocal tract anatomy because an additional level of transformation, namely the speech acoustics, intervenes.

Discounting the effect of real-world circumstances on how faithfully the spectrogram represents the acoustics actually radiated from a speaker on a given occasion, and how much of the actual friction ridges of the individual's finger a particular fingerprint shows, it can be appreciated that the thing represented in both cases differs enormously in variability. The friction ridges are effectively invariable for a given individual, whereas a person's speech acoustics are inherently variable, this variability arising from the interaction between the individual's communicative intent and the vocal tract he or she has to use to implement it.

Moreover, the nature of variation in the structures that are represented by fingerprints and spectrograms is very different in both, whether this is taken to be the anatomical structures of fingerprints and vocal tracts, or the anatomical structures of fingerprints and speech acoustics. The friction ridges on an individual's finger are not normally subject to variation. Considerable within-subject variation in vocal tract anatomy/speech acoustics exists simply by virtue of speaking, quite apart from organic differences from different states of health. Between-subject variation in friction ridges is partially genetically conditioned but arises mostly as the result of random in utero pressure differences. Its patterns are not functionally constrained, and thus are free to vary widely. Between-subject variation in the anatomy of the vocal tract as it produces speech, or the acoustic patterns of speech, is largely not random, because they are constrained by the function of speech communication.

Aural-spectrographic voice identification evidence can still be found in investigative, corroborative and substantive use today, especially in the United States of America. In spite of its fairly widespread use, aural-spectrographic recognition in forensics continues to be the centre of considerable controversy on both scientific and legal fronts. It differs from the kind of auditory and acoustic forensic analysis described above in lacking a theoretical base and appearing to rely on naive linguistic and visual pattern-matching abilities. It is the aural-spectrographic method that Ormerod (2002), in his critique of TFSI, mistakenly assumes to be synonymous with the kind of forensic speaker identification described in this Chapter.

Although the term "voiceprint" can sometimes be found used in a reasonably bona fide sense in automatic speaker identification (as a set of statistical features, or template, that can be used to identify or verify voices), the use of the terms "voiceprint", or "aural-spectrographic" identification in FSI reports nevertheless should be treated as warning signs. The reader is referred to several in-depth critiques of the method: see Rose (2002, Ch 5); Hollien (2002, Ch 6); Gruber and Poza (1995).

Summary

[99.830] Speech samples are compared forensically by means of different types of features. It is necessary to compare samples with respect to both auditory – especially auditory-*linguistic* – and acoustic features. Thus the omission of either in a report requires justification.

Features are extracted using received analytical frameworks from the speech and information sciences (eg, source-filter theory, phonemics, signal detection theory). This permits the detailed linguistic characterisation of speech samples and quantification of their acoustics. The speech samples' features are compared using statistical approaches from Bayesian inference. The use of any method outside of speech or information science, eg aural-spectrographic/voiceprint identification, can be legitimately criticised for being non-scientific.

The acoustic output of an individual's vocal tract is relatively easily extracted and quantified by computer, but extraction always requires interaction with the expert, and any report must detail this. Acoustic representations of different degrees and types of abstraction and power (eg, linear-prediction derived formants, cepstrally-smoothed spectra, long-term and short-term features) can be extracted from the signal. Automatic features require much less interaction and yield on average higher likelihood ratios than traditional features, but are still infrequently met with and lack interpretability. The measurement of traditional features requires considerable theoretically-informed interpretation by the expert.

Although the acoustic output of an individual's vocal tract is uniquely determined by its size and shape, these factors reflect the interaction of the tract's dimensions and organic state with the choices the individual makes in conveying linguistic, paralinguistic and extralinguistic information. Acoustics are never invariant for the same speaker, and always reflect how the speaker is speaking on that particular occasion. Spectrograms are not like fingerprints.

[The next text page is 99 - 5101]

EFFECT OF TELEPHONE TRANSMISSION

[99.870] A frequently asked question concerns the effect of telephone transmission in TFSI. This is a very sensible question, given that the majority of speech samples in TFSI are taken from telephone conversations.

The distortion involved in telephone transmission is manifested in two ways:

(1) Telephone transmission acts as a band-pass filter that passes energy within a particular frequency band. This band nominally extends from about 300 Hz to 3500 Hz, but can vary greatly. Energy above and below these cut-off frequencies will be attenuated, to the extent that any information will be severely compromised and rendered useless for comparison.

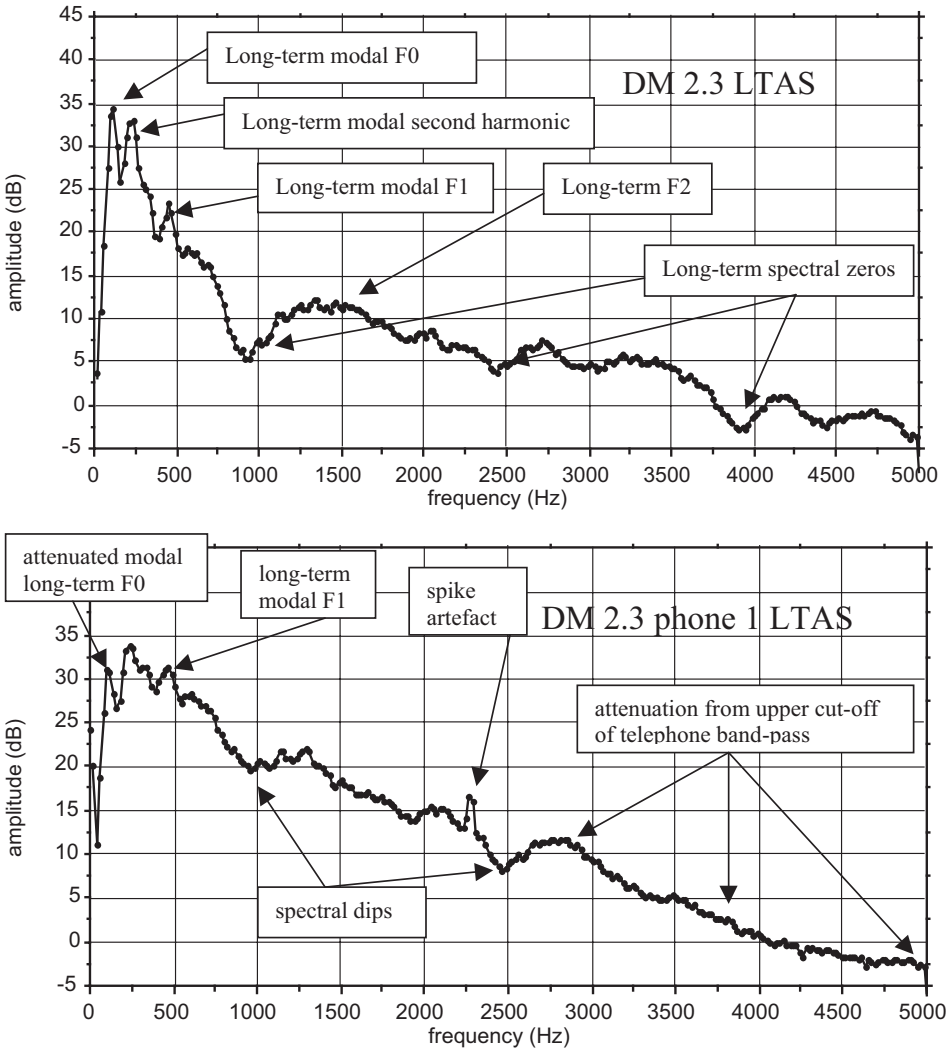
(2) There will also be *spectral distortion* within the nominal 3.5 kHz range of the band itself. This is partly due to the effect of the band-pass filter, which has the potential to shift frequencies slightly above the lower cut-off upwards, and frequencies slightly below the upper cut-off downwards. Spectral distortion will also be manifested in all sorts of artefactual peaks and dips in the band spectrum introduced by many other uncontrollable real-world factors like the particular hand-set used; the acoustic properties of the immediate physical environment (whether the call is being made in a phone booth, an acoustically reflective room, while walking along the street); the technology (eg mobile, digital landline, analog); and the particular routing of the call. In addition, the telephone transmission can be expected to have a narrower dynamic range, and to not be as spectrally differentiated (with peaks and troughs not so prominent).

Figure 33 shows some of these typical telephone transmission distortions using a long-term average spectrum. It shows what the signal in Figure 26 looks like when sent over a telephone line involving both analog and digital transmission (Figure 26 has been reproduced for ease of comparison). This involved an experiment where a recording of the utterance was played from a loudspeaker into the handset of a conventional analog telephone to simulate someone speaking into a telephone. Both the loudspeaker and the handset were contained in a specially treated box to control for reflectivity. The signal was transmitted over a digital landline, recorded from the phone line at the receiving end, and a long-term average spectrum made of the recording. Thus the signal shows the effects of the loudspeaker, handset, box and the landline transmission, as well as the response of the tape-recorder.

Comparison with the long-term average spectrum of the original signal shows both distortion and preservation of the original spectral features. Probably the most conspicuous difference lies in the spectral drop-off starting just below 3000 Hz. This is due to the high frequency attenuation of the telephone channel mentioned above, although in this case its effect cuts in considerably below the nominal 3500 Hz. There will be no usable information above about 2.8 kHz in this signal. The low-frequency attenuation from the band-pass effect can be seen in the relatively reduced amplitude of the long-term modal F0 peak, which is now lower in amplitude than that of the modal second harmonic.

The dynamic ranges of the two spectra up to about 3000 Hz are fairly comparable. The original spectrum has a range of about 30 dB (from about 35 dB to about 5 dB), whereas the phone spectrum does, indeed, have a slightly narrower range of about 22 dB (from about 34 dB to about 12 dB). The spectral profile in the phone spectrum is less differentiated, with the troughs appearing filled-in.

FIGURE 33 Effect of telephone transmission on long-term average spectrum of speech



Top = LTAS of original signal (reproduced from Figure 26). Bottom: LTAS of the same signal passed through an analog telephone line.

Another obvious distortion is the telephone-induced spike at just over 2250 Hz. This is not a problem, since the source-filter theory of speech production makes clear the artefactual nature of such a feature at such a frequency location: it cannot have been produced by a vocal tract, and therefore could be discounted as a bona fide difference between the two spectra.

Despite these distortions, it can be seen that the telephone has appeared to preserve the frequency locations of the two lowest spectral zeros, and also the frequency locations of the lowest three peaks (long-term modal F0, modal H2 and modal F1 peak). However, there are still small discrepancies – in the order of between 0 and 30 Hz – between the two spectra in the actual frequencies of these events. $Z(ero)_1$ occurs at 918 Hz for the original signal, and at 957 Hz in the phone; and Z_2 occurs at 2441 Hz in the original signal and at 2461 Hz in the phone. The modal F0 peak occurs at 118 Hz in the original signal and at 107 Hz in the phone spectrum; the modal H2 peak occurs at the same frequency in both; and the modal F1 occurs at 450 Hz in the original signal, and at 468 Hz in the phone spectrum.

It is assumed that telephone transmission does not distort fundamental frequency, although it can certainly affect its computerised extraction.

Spectrographic demonstration of effects of telephone transmission

[99.880] Some of the effects of telephone transmission are best demonstrated spectrographically. Figure 34 shows a spectrogram of the same "strikes raindrops" utterance as shown in Figures 28 and 29 when sent over a telephone line (the conditions were the same as described above). The formants have been automatically extracted and tracked. Comparison with the spectrogram of the original utterance and its superimposed formants in Figure 34 shows the following differences due to the effects of the loudspeaker, handset, box and the landline transmission.

FIGURE 34 Spectrogram with LP estimated and tracked formants of a recording of the "strikes raindro(ps)" utterance in Figure 28 when passed over a telephone line

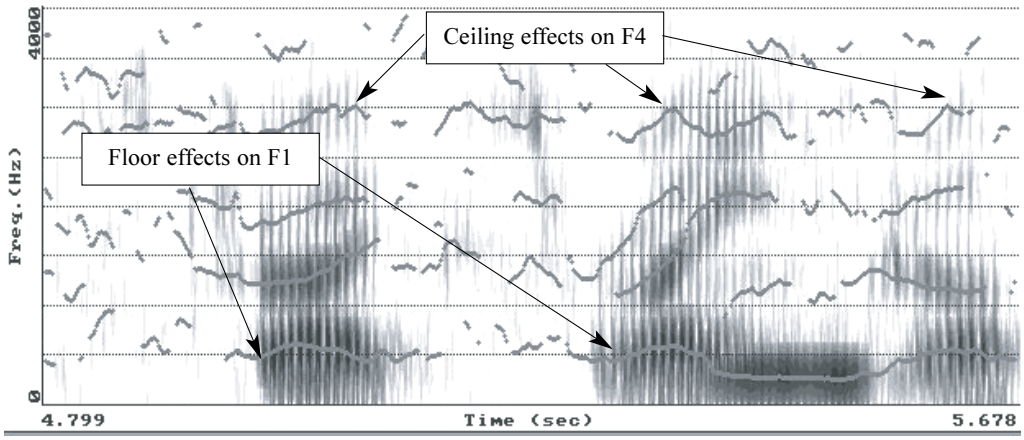


Figure 34 looks a lot more wormy than Figure 29, with spurious resonance traces extracted throughout the frequency range in areas where the amplitude was low in the original recording. This is due to the greater amount of channel noise present in the telephone recording. Generally, energy can be seen to have disappeared above about 3.5 kHz, due to the upper limit of the band-pass. In the original spectrogram, energy present above 3.5 kHz was found in the two /s/ sounds, and in the F4 of the /ɛɪ/ diphthong. This is no longer visible. The effect on the LP extraction of the centre frequencies of the /ɛɪ/ diphthong is noteworthy. Its sublingual cavity resonance, as in the original signal, is resolved correctly up to about 2.2 kHz. However, its F4 has been resolved

up to about 3.0 kHz, after which its centre frequency is shown as decreasing and becoming continuous with the true F3. A similar ceiling effect has occurred with F4 both in the /aɪ/ diphthong, which reaches 3.5 kHz in the original, but only 3.0 kHz in the telephone spectrogram, and in the /ɒ/.

An analogous ceiling (probably better called a "floor") effect is also present as a result of the lower limit of the band-pass. This can be seen by examining some of the frequencies of the lowest resonances, eg the F1 at the first glottal pulse in the first two /r/ phonemes, and the first nasal formant in the /n/. It can be seen from a comparison of the phone and original spectrogram that the onset frequencies of the first formant in /raɪ/ and /rɛɪ/ in the phone spectrogram are higher than in the original. An effect in the /n/ formant cannot be easily seen. Table 11 shows the actual values for these three points of comparison extracted by linear prediction. It can be seen that all three estimates in the phone recording are higher than in the original. The magnitude of the difference is not the same in each case, however: for the two /r/'s the difference between the original and phone recording is greater the lower the original value; for the /n/ the difference is very small, even though the original value is low. Thus the value in the phone recording for low frequencies might not be a sole function of the original value (information on the fundamental frequency and the rest of the segment's F-pattern is probably needed to attempt a prediction). Another explanation is that the differential is due to the low frequency boost of the loudspeaker involved.

TABLE 11 Lowest resonant frequencies (Hz) for the /n/ and the first two /r/'s extracted by linear prediction from the "strikes rainbow" utterance

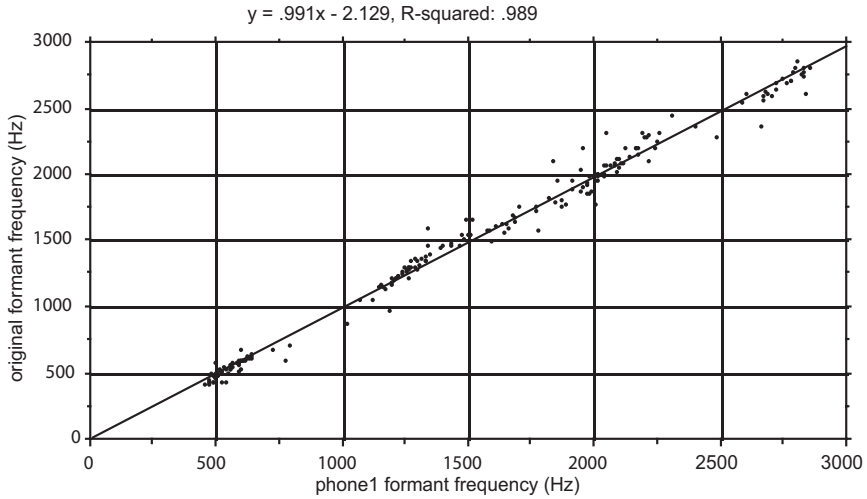
	Original	**Phone1**	$\Delta_{\text{phone-original}}$
/r/ / _ /aɪ/ (mmt. at first glottal pulse)	443	481	38
/r/ / _ /ɛɪ/ (mmt. at first glottal pulse)	327	441	114
/n/ (mean of mmts. at first 3 pulses)	252	267	15

These floor and ceiling effects, as well as the poor performance of automatic formant extraction in telephone speech, are fairly well known. They were reported in Rose and Simmons (1996), and in Künzel (2001). They mean that formant comparisons are counter-indicated on at least some segments with low F1, and perhaps all measurements above 3 kHz. (This will usually include F4, except on rhotics.) Further research has shown that it is probably a sensible heuristic to automatically exclude F1 measurements on all except low vowels, which have high F1: see **[99.640]**. Since F1 does not generally have much individual-identifying power, this is not a loss; the sacrifice of F4 is not such good news, as it often shows good F-ratios, with consequent high potential for LRs deviating considerably from 1.

It is important to emphasise that formant measurements from nearer the centre of the band-passed range, however – say between about 500 Hz and 3 kHz – do not generally seem to be adversely affected by the telephone transmission. This can be appreciated from a visual comparison of the tracked F2 and F3 trajectories in Figures 34 and 29, but is best quantified by plotting the original recording measurements against the phone recording measurements, the idea being to see how well the measurements from the phone recording agree with those from the original signal.

Figure 35 shows original formant values in the sentence "When the sunlight strikes raindrops in the air they act like a prism and form a rainbow" plotted against 197 putatively valid formant frequency estimates in the phone recording within the frequency range of 400 Hz to 3 kHz. (This includes F1 measurements above 400 Hz, most F2, F3 and sublingual cavity resonance measurements, and those F4 measurements below 3 kHz.)

FIGURE 35 Simple regression of LP estimated formant values from original signal on values in phone recording



Since the dependent variable – here the values from the original signal – is known to be the same as the independent variable – the values from the phone data – every value on the horizontal axis should correspond to the same value on the vertical axis, and the data must be modelled with the simple linear equation y = βx + α, where *x* is the phone value; *y* is the original value; β, the proportionality constant, is 1; α, the intercept, or value of *y* when *x* is 0, is 0. (So that, eg, if the value *x* for a particular formant measurement from the phone recording was 500 Hz, the original value *y* would be βx + α = (1 x 500) + 0 = 500 Hz.)

The extent to which the data fit such a model is examined by doing a simple linear regression of the original data on the phone data. From Figure 35, which shows the results of such a regression, it can be seen that the data do, in fact, come very close to this. The equation which best linearly models the data has a constant of proportionality (represented in the equation in Figure 35 as x) of very nearly 1, namely 0.991; and an intercept of about -2 Hz. The R-squared value of 0.989 indicates that about 98.9% of the variance in the original data is accounted for by the model. The square root of this value – ca 0.99 – indicates that the data are almost perfectly linearly (99%) correlated.

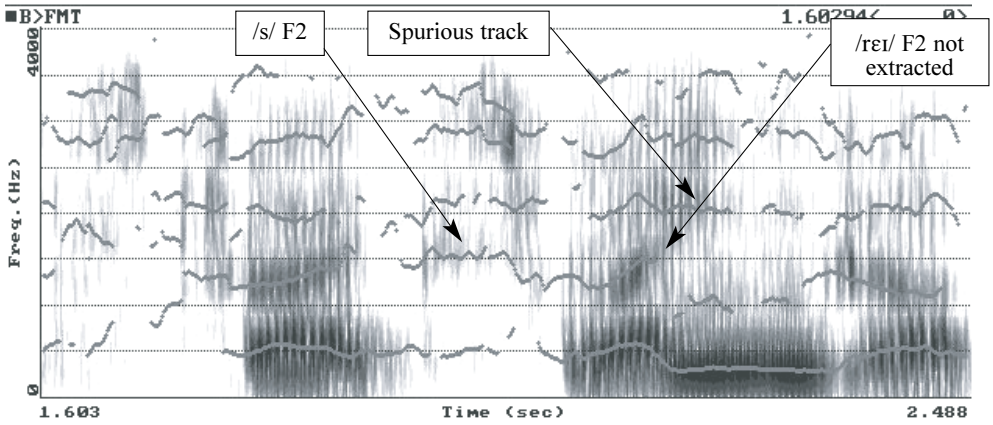
The most important question is how *accurately* the phone measurements mirror the original measurements. The discrepancy involved in each estimate of the original from the phone arises partly from measurement error and partly from the effect of the phone transmission. Normally this needs to be quantified with a statistic derived from the regression analysis called the "standard deviation of the residuals", but since here the output of the underlying model is the same as the input, it can be done by simply calculating the discrepancies between the phone and original measurements. The standard deviation of the signed differences between the phone and original data is ca 75 Hz, which means that about 70% of the time the original measurement will lie between ca +/- 75 Hz of the phone measurement, and about 90% of the time it will lie between ca +/- 150 Hz of the phone measurement.

These confidence limits are rather large. For example, there is about an 81% probability that any one phone measurement will lie within about 80 Hz above or below the original measurement, but a probability of only about 61% that any one phone measurement will be accurate to within 50 Hz. This shows the need to base formant estimates from telephone recordings on a lot of data, so that the inevitable discrepancies have a chance of cancelling each other out in the long run.

In the 197 measurements in Figure 35, eg, there is no clear tendency for the phone estimates to be higher or lower than the original, and differences do tend to cancel each other out: on average, the original measurements are overall about 16 Hz lower than the phone measurements. This is quite a small discrepancy and within the tolerances expected from human or machine measurement on telephone speech.

No phone transmission characteristics are ever the same, and because of this the actual properties of the original signal can never be recovered. Some lines are good, appearing to involve little distortion, some are extremely bad. Figure 36 shows a spectrogram and LP estimated and tracked formants of the same "strikes raindro(ps)" utterance recorded over another phone line. This recording sounded somewhat worse – more muffled – than that of the first phone recording. The same ceiling and floor effects are observable as in the previous example, but the extraction of the formant frequencies – eg, in /rɛɪ/ – is clearly not as good. The reader will be able to spot how the latter part of the trajectory of F2 in /rɛɪ/ has not been picked up at all; and also that the level trace at about 2 kHz cannot reflect a true F3 for this sound, but is actually the result of an illegal continuity from the true F3 onto the true F2.)

FIGURE 36 Spectrogram with LP estimated and tracked formants of a recording of the "strikes raindro(ps)" utterance in Figure 29 passed through a slightly worse phone line than in Figure 34



Despite all this, it is clear that some parts of the formants have been correctly extracted. Note, too, that the F2 of /s/ has, in fact, been extracted better than in the previous phone recording. The most important point here, however, is that the performance can be seen to be inferior, given a knowledge of the relationships between sound and acoustics afforded by theory. In such cases, the analyst can simply decide not to make a measurement, so no adverse effect occurs other than a decrease in the amount of data available for comparison.

Effect of telephone transmission on auditory quality

[99.890] Because it distorts frequencies in the vicinity of the upper and lower limits of the band-pass, it is to be expected that telephone transmission will also have an effect on how the speech sounds are perceived in telephone speech. Since low F1 values are shifted up by the telephone transmission, and F1 is indirectly proportional to vowel height, it is to be expected that high vowels, with low F1, will sound more open over the phone than they actually were when

said. This may also apply to vowels with originally higher F1, like mid vowels. Again, this creates no problem, since the analyst will know to discount any perceived between-sample differences in the height of high or mid vowels. The same principles apply to perceived auditory features encoded in frequencies near the upper limit of the band-pass, like some voice quality differences, and the difference between some fricatives.

Because there does not appear to be any adverse effect on the frequencies towards the centre of the band-pass, auditory features encoded in this area, like vowel backness or rounding, can be assumed not to be compromised, thus allowing auditory comparison to be made with such features.

Realistic comparison involving telephone transmission

[99.900] The section above has demonstrated some typical effects of telephone transmission by comparing the *same* signal before and after it has been passed over a telephone line. Although the use of the same signal is logically the only way to demonstrate these effects, this never, of course, happens in forensic reality, where the comparison is always between two *different* signals. Comparison can be between samples, all of which have been intercepted from telephone speech, or between samples from telephone speech and samples from direct speech. The latter case, since it is more complicated, is demonstrated in Figures 37 and 38.

Figures 37 and 38 show a comparison, for three male Australian speakers, between the acoustics of a subset of their vowels recorded in the laboratory, and of a *separate* set recorded over the phone. The data are from Rose and Simmons (1996). The three speakers are two brothers and their father; the vowels were measured from stressed words said at the end of sentences like "where is the *deed*?"

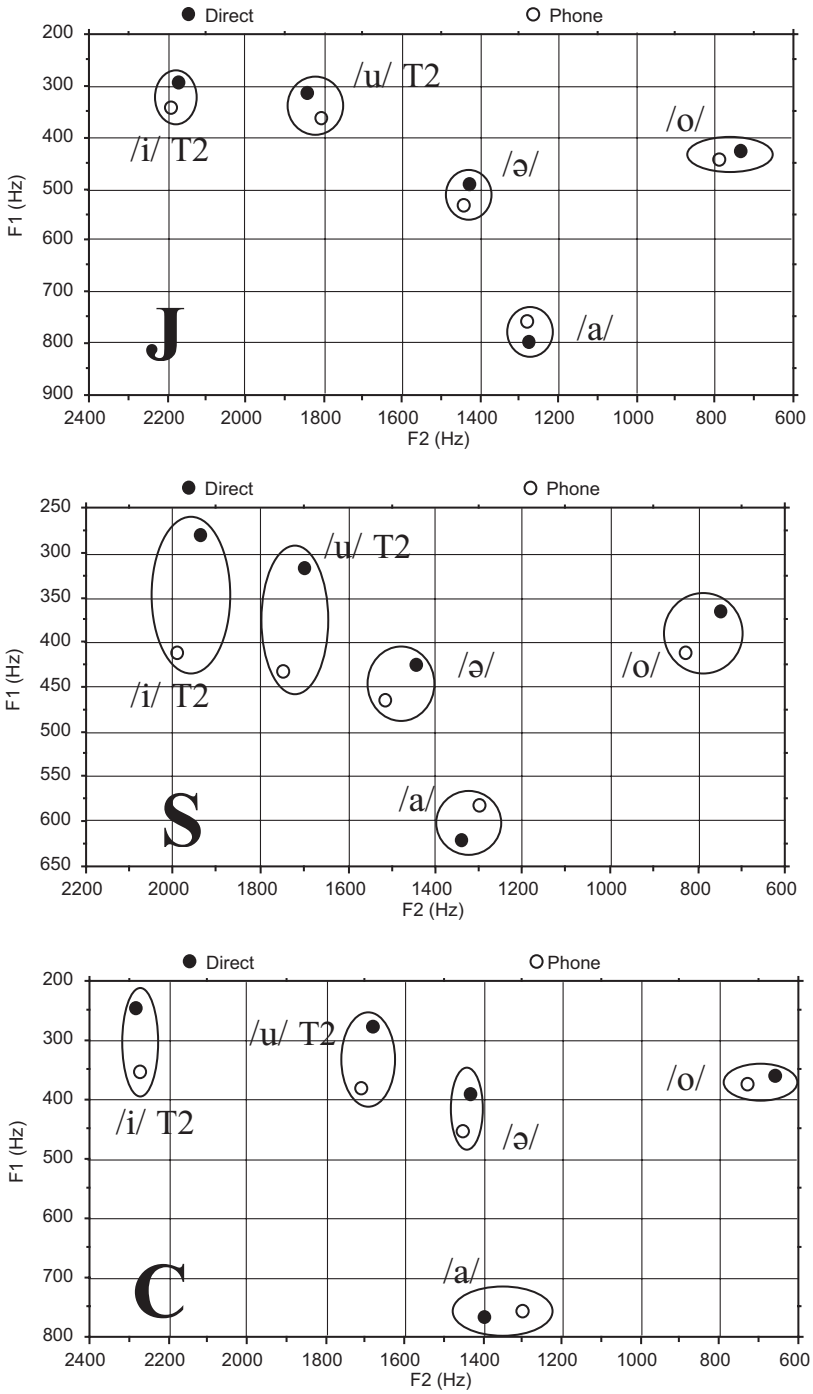
Australian vowels, as indeed vowels in most varieties of English, are very complicated. Their phonological structure is explained in Rose (2002, Ch 6). Comparison is shown for five of the speakers' long vowel phonemes: /i/, /u/, /o/, /ə/ and /a/. /i/ and /u/ in this accent – technically called "Broad Australian" – are phonetically actually diphthongs: only their second targets, labelled *T2*, are shown in Figure 37 for clarity, but their first targets, labelled *T1*, have been included in Figure 38. The vowel acoustics of the two sets of recordings, labelled "phone" and "direct", are shown using conventional vowel acoustic plots of F1 vs F2 (Figure 37), and F2 vs F3 (Figure 38). These types of plots were introduced above in the section on formants: see **[99.590]** ff.

The F1 - F2 plots in Figure 37 all nicely show the floor effect on F1 of high vowels /i/ and /u/ from the telephone transmission: the telephone high vowels are all located lower on the chart than the direct recordings. The effect seems also to apply to /o/ and /ə/ as well, however, which illustrates the need to exclude F1 measurements for these vowels. It can also be seen that the magnitude of the effect is different for different speakers, with S showing large differences, C intermediate differences, and J small differences.

Although there are clearly differences between the phone and direct recordings in F2, they are small – small enough to be not statistically significant.

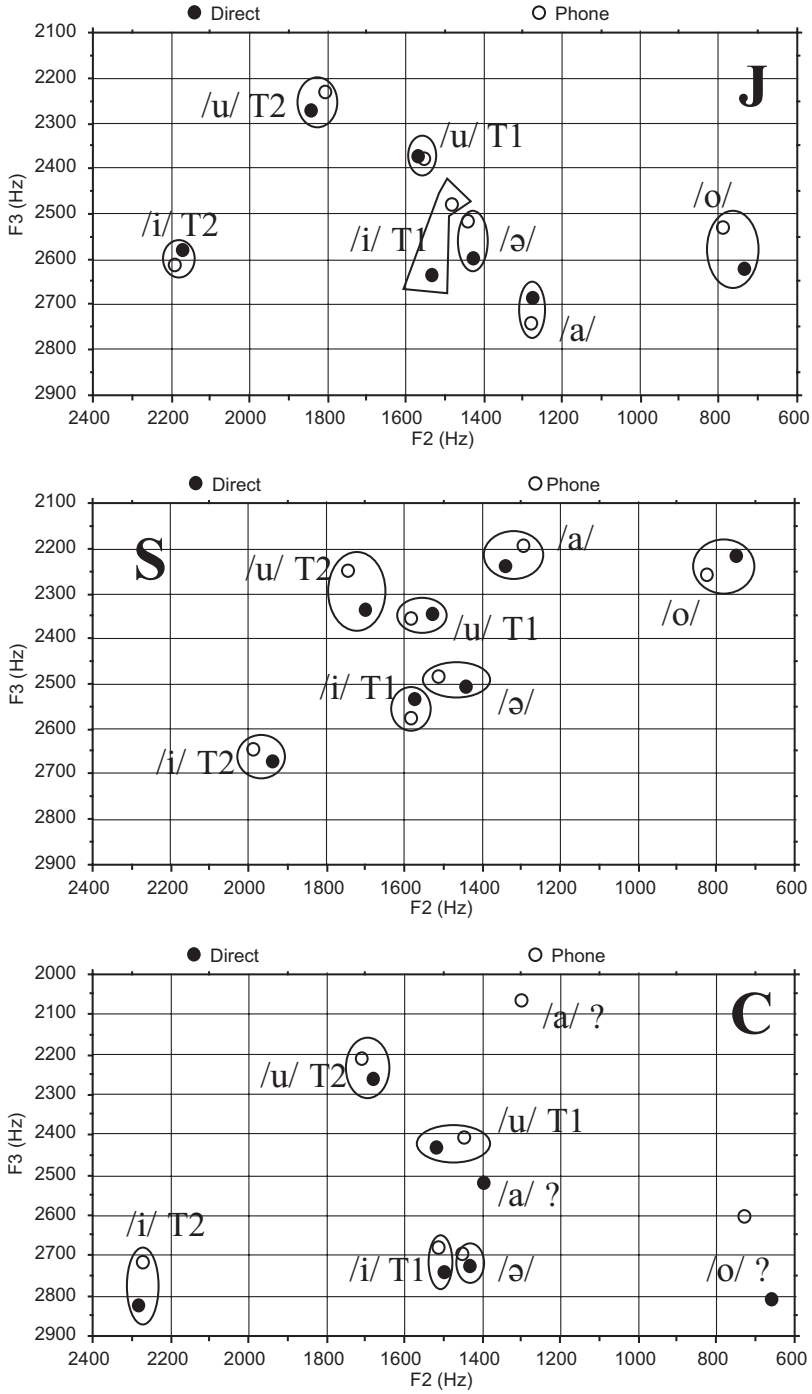
The F2 - F3 plots in Figure 38 show that the phone and direct recordings are very similar. One anomaly is /a/ in speaker C, where the phone F3 was below 2.1 kHz compared to above 2.5 kHz for the direct speech. Comparison with the plots of the other two speakers shows that these two extracted values probably reflect different formants, incorrectly extracted as F3 by the automatic extraction. (It is well known that automatic extraction very often mis-identifies formants and consequently always needs to be checked.) Whereas the phone and direct F2 and F3 measurements group well for each speaker, Figure 38 shows that they clearly differ for different speakers – compare, eg, the position of /i/ T2 for the three speakers.

FIGURE 37 F1 - F2 plot of phone and direct acoustics for five vowels of three Australian male speakers



T2 = measurements at 2nd diphthongal target in /i/ and /u/.

FIGURE 38 F2 - F3 plot of phone and direct acoustics for five vowels of three Australian male speakers.



T1/2 = measurement of 1st/2nd diphthongal targets.

Example of likelihood ratio based forensic discrimination with telephone speech

[99.910] The crucial thing is the extent to which the phone transmission interferes with correct discrimination of same-speaker samples from different-speaker samples. For these data, at least, it does not. That is, it is possible, for any pair of samples in the data, to correctly say whether it is a pair of samples from the same speaker or a pair from different speakers. This can be demonstrated with a *forensically appropriate* discrimination on the F2 and F3 values of the speakers' five vowels, using the likelihood ratio as a *discriminant function*. Theory predicts that, given appropriate conditions, the LR should be smaller than 1 for different subject data, greater than 1 for same-subject data. Thus a LR can be calculated for any pair of samples and used to predict whether they constitute a same-speaker pair or different-speaker pair. This very important point is explained with another example from Japanese vowels in Rose (2002, Ch 11) which, since it also discusses the limitations of such an approach, is essential reading.

Table 12 shows that all three same-speaker comparisons – between the direct and phone speech for each speaker – can, on the basis of the LR based on just F2 and F3 of the five vowels, be *absolutely discriminated* from all 12 different-speaker comparisons – eg, between speaker S's phone vowels and speaker J's phone vowels, or between speaker C's phone vowels and speaker J's direct vowels.

TABLE 12 Results of LR-based forensic discrimination of phone and direct speech for the F2 and F3 of three Broad Australian males

SAME-SPEAKER COMPARISONS (ie direct vs phone)							
		/a/	/o/	/ə/	/i/	/u/	Combined LR
J	F2	5.111	1.894	1.991	3.894	4.833	
	F3	1.71	0.175	***0.446***	5.404	3.923	**1028.76**
C	F2	***0.743***	2.00	2.099	1.875	2.098	
	F3	n/a	n/a	4.876	***0.815***	2.827	**137.82**
S	F2	1.374	***0.014***	***0.801***	12.49	2.412	
	F3	5.94	3.461	1.838	2.095	1.549	**57.26**
DIFFERENT-SPEAKER COMPARISONS							
				J vs C			
J direct	F2	1E-14	0.434	2.397	0.527	0.027	
C direct	F3	0.846	0.863	0.100	0.038	6.006	**3.12 E-18**
J direct	F2	***2.952***	***9.744***	***2.715***	0.027	0.134	
C phone	F3	n/a	***6.197***	0.331	***6.006***	0.072	**0.25**
J phone	F2	5E-04	0.048	9.772	0.872	0.131	
C direct	F3	0.618	n/a	3E-16	0.067	***5.205***	**1.47 E-21**
J phone	F2	***2.8***	***2.122***	***1.803***	1.237	0.797	
C phone	F3	n/a	0.057	3E-30	6.724	0.114	**1.49 E-30**

J vs S

J direct	F2	0.133	***10.3***	***3.441***	0.001	0.003	
S direct	F3	n/a	3E-04	0.969	1.503	1.932	**1.45 E-8**
J direct	F2	***3.813***	0.01	0.882	3E-09	***1.399***	
S phone	F3	n/a	1E-08	0.163	***2.028***	***3.622***	**2.4 E-18**
J phone	F2	***3.524***	***1.781***	***1.196***	8E-05	2.77	
S phone	F3	n/a	1E-06	***2.926***	***4.103***	4.083	**1.02 E-7**
J phone	F2	0.709	***2.533***	***2.021***	0.003	0.021	
S direct	F3	n/a	0.007	***1.924***	***1.917***	0.76	**4.0 E-6**

S vs C

S direct	F2	0.23	0.005	2.431	9E-04	3.073	
C direct	F3	0.397	n/a	0.051	0.537	1.431	**1.28 E-7**
S direct	F2	***1.737***	***8.259***	***3.182***	1E-05	***2.574***	
C phone	F3	2.376	5E-04	0.113	***1.656***	0.345	**1.3 E-7**
S phone	F2	0.001	6E-11	***1.161***	5E-04	***1.688***	
C direct	F3	n/a	n/a	4E-11	0.202	4.329	**2.24 E-27**
S phone	F2	***3.082***	0.14	***1.389***	2E-08	***2.133***	
C phone	F3	***3.872***	2E-09	3E-12	***1.115***	3.894	**2.31 E-27**

n = 2 per sample (/ə/); 4 (other vowels).
Figures in bold italics indicate LRs which are counter to reality.
n/a = comparison not possible because different formants were extracted.

In Table 12, the values for the LR for F2 and F3 of each of the vowels are given separately, in the five columns to the left, and the combined LR derived from their product is shown in the right-hand column. Thus, eg, when the F2 of /a/ in speaker J's phone sample was compared with the /a/ F2 in his direct sample, the difference was about five times more probable assuming they had come from the same rather than different samples (LR = 5.111). When LRs from all F2 and F3 of all five of J's vowels were combined, the difference between the phone and direct samples was just over 1000 times more likely assuming that they had come from the same speaker (LR = 1028.76). As usual, it can be seen that some LRs for some individual comparisons ran counter to reality – these are shown in italics in Table 12. It can also be seen that some different-speaker comparisons gave combined LRs that are so small that they have to be indicated with scientific notation.

A LR has to be calculated against the background of suitable multispeaker reference data, independent of the test data, which give a good estimate of the distribution in the relevant population. Often such a distribution is not available, but in this case it is: see Bernard's (1967) study on acoustics of Australian vowels, from which mean and standard deviation F2 and F3 measurements were calculated from vowels spoken by 56 Broad-speaking males. The vowels were spoken in stressed words at the end of sentences, just as with the tested data. The reference data used in the discrimination, which, of course, are totally independent of the tested data, are given in Table 13.

TABLE 13 Data for calculation of reference parameters in LR discrimination

	F2			F3		
	x	s	n	x	s	n
/i/ T2	2255	143	57	2754	150	56
/u/ T2	1623	160	43	2415	174	43
/ə/	1565	112	56	2534	150	56
/o/	870	77	56	2471	232	38
/a/	1367	102	55	2506	210	51

x = mean (Hz); s = standard deviation (Hz); n = number of speakers. Source: Bernard (1967) (conditions: vowels of *Broad* speakers before alveolar consonants in stressed syllables in read out sentences).

The discrimination just described was performed on very well-controlled data that will have undoubtedly contributed to its good performance – comparing vowels in the same words in the same prosodic environments for example. Also, no attempt was made to take into account the almost certain correlations within the data which will have had the effect of boosting the magnitude of some of the combined LRs. The comparison was nevertheless forensically realistic in being based on non-contemporaneous speech – the laboratory recordings and the phone recordings having taken place on different days – and shows that the telephone effect per se is probably not going to inevitably compromise such comparisons.

It is, rather, more probable that the problems that arise with forensic comparisons involving telephone speech do not so much have to do with distortion associated with the telephone line as with the speaker who is using it. For all sorts of reasons, people can speak differently on the phone than face-to-face. The degree of formality might be different, eg, or the amount of background noise. Since many aspects of speech co-vary with such factors, the magnitude of within-speaker variation is thereby increased. This makes it more difficult to correctly evaluate the probability of whether differences between samples are between- or within-speaker differences.

This is obviously an important consideration, since comparison is often requested and/or undertaken between recordings of telephone intercepts and the speech of a suspect under questioning by police. Such comparisons therefore need care, especially since the difference in perceived formality can give rise to linguistic differences (this is discussed in detail in Rose (2002, Ch 3)). In view of this, the expert always needs to be confident that the samples are comparable with respect to the features they want to use, and forensic comparison of phone and direct speech samples should probably require individual justification. Knowledge from the sub-discipline of socio-linguistics will usually be necessary to evaluate the comparability of the speech samples in circumstances like this. Care must also be taken to justify comparability between samples that are *both* from telephone recordings. Here, other factors may affect comparability of telephone speech samples, like differences in background noise level, or even whether the receiver thinks he or she knows from looking at her or his mobile who is calling the receiver. An assessment of comparability is required for *all* comparisons, of course, not just those involving the telephone.

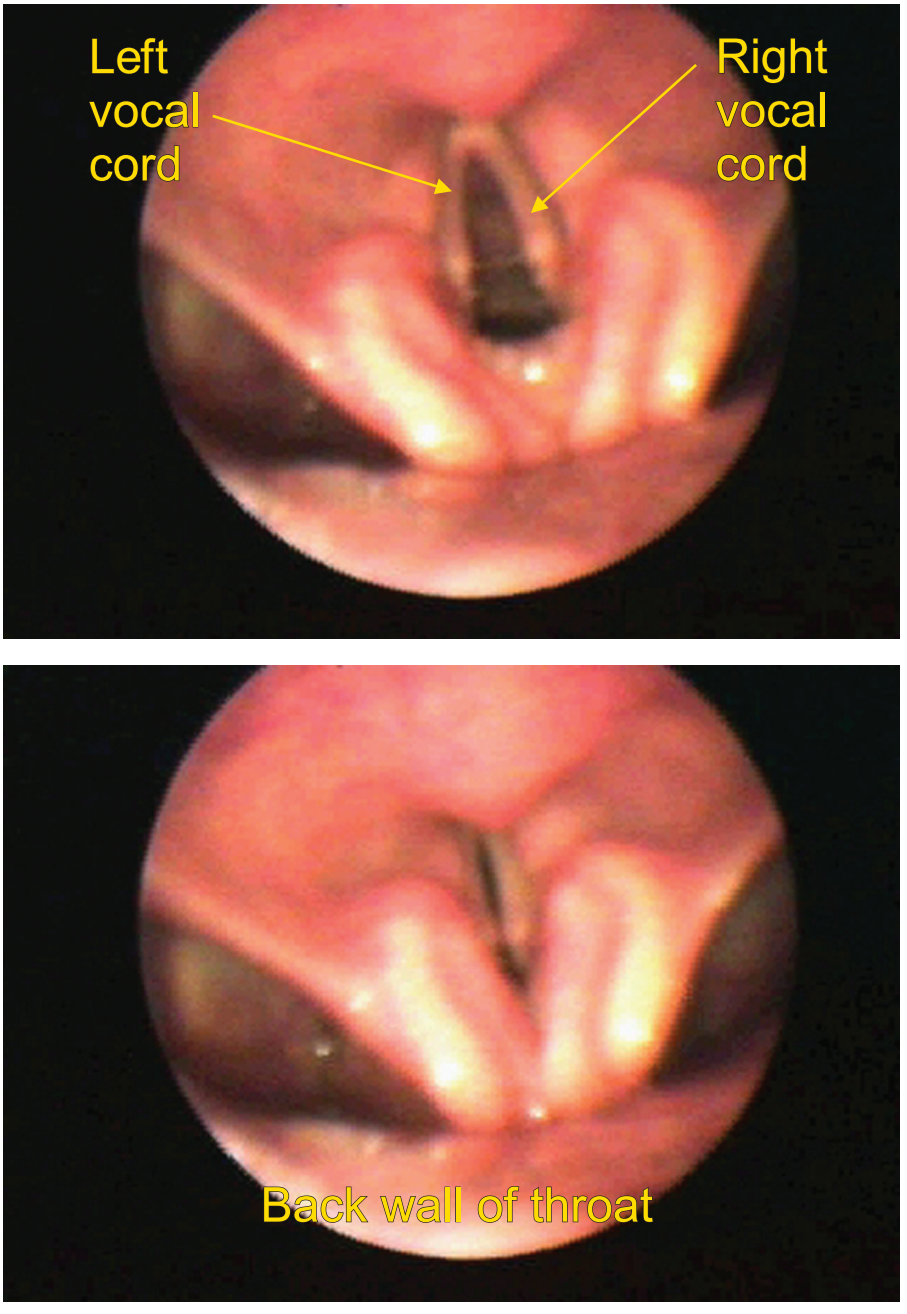
Summary

[99.920] Although telephone transmission distorts the spectral characteristics of speech acoustics, and as a result some features are irretrievably compromised, it does not necessarily do so in such a way as to totally preclude forensic comparison of the acoustics of samples over the telephone, or of their auditory qualities. It is up to the expert to decide whether the distortion has been too great. This is done both by auditory assessment – if it is difficult to actually hear what has been said, for example – and by examining how well the estimation and tracking work in the light of what their expected values are for given speech sounds. Often the degradation can be quantified by signal-to-noise ratios. Even if automatic tracking fails, which it often will, it will often be possible to visually estimate formant frequencies from spectrograms.

A more potentially deleterious effect of the telephone comes, not from the inevitable physical distortions of its transmission, but from what the *speaker* does when speaking over it, and particular care has to be taken to assess the comparability in comparisons involving telephone speech.

[The next page is FIG 99 - 1 and the next text page is 99 - 6101]

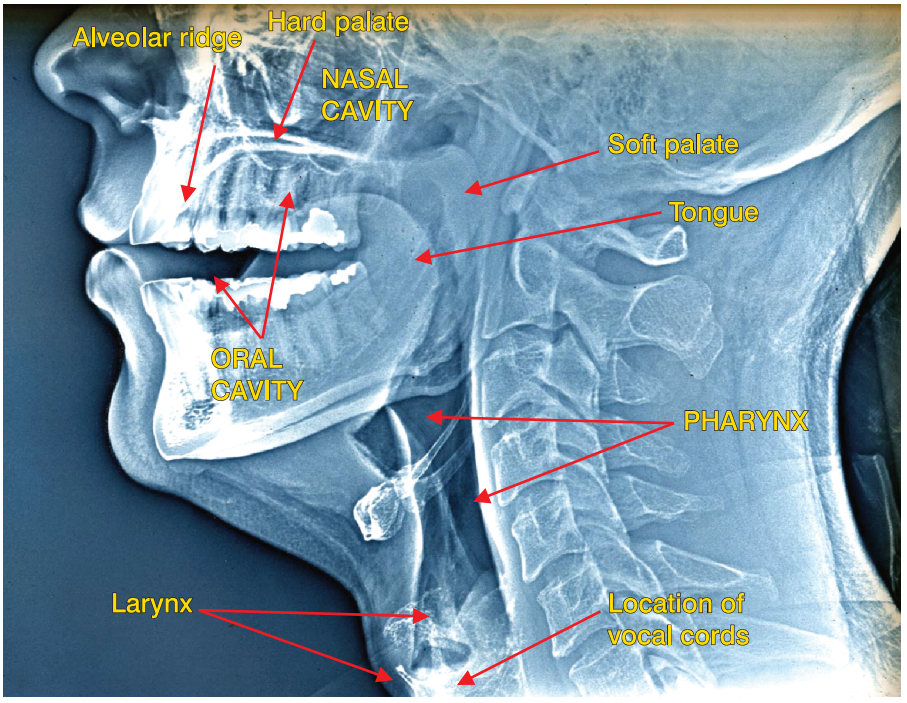
Figure 3—Endoscopic pictures of vocal cords in abducted and adducted position [99.1020]



Close-up of the author's vocal cords when seen from above at a distance of a few centimetres. The front of the author's head is facing towards the top and slightly towards the left. The top picture shows the cords in abducted, the bottom in adducted position.

Figure 4—X-ray of vocal tract for [u] vowel

[99.1120]



X-ray of the author's supralaryngeal vocal tract taken in the middle of a sustained *oo* vowel similar to that in the (British) English word "who". Note how the lips are close together and slightly pursed, and the tongue appears bunched up towards the top and back of the mouth. These are typical supralaryngeal articulations involved in producing an *oo* vowel. The soft palate can be seen to be raised, so the vowel is not nasalised. The top of the larynx is visible in front of the body of the 6th cervical vertebra. The vocal cords cannot be seen directly, but are located immediately below the small dark triangular shape in the middle of the larynx.

Select bibliography and references

Aitken CGG, *Statistics and the Evaluation of Evidence for Forensic Science* (Wiley, 1995).

Bernard JRL, "Some Measurements of Some Sounds of Australian English", Unpublished PhD thesis, Sydney University, 1967.

Broeders APA, "Some Observations on the Use of Probability Scales in Forensic Identification" (1999) 6 (No 2) *Forensic Linguistics* 228.

Broeders APA, "Forensic Speech and Audio Analysis Forensic Linguistics 1998 to 2001: A Review", Paper at the13th INTERPOL Forensic Science Symposium, 2001.

Champod C and Evett I, "Commentary on Broeders" (1999) (2000) 7 (No 2) *Forensic Linguistics* 238.

Champod C and Meuwly D, "The Inference of Identity in Forensic Speaker Recognition" (2000) 31 *Speech Communication* 193.

Elliott JR, "Auditory and F-Pattern Variation in Australian 'Okay': A Forensic Investigation" (2001) 29 (No 1) *Acoustics Australia* 37.

Elliott JR, "The Application of a Bayesian Approach to Auditory Analysis in Forensic Speaker Identification" in C Bow (ed), *Proceedings of the 9th Australian International Conference on Speech Science and Technology*. (Australian Speech Science and Technology Association, 2002).

Elliott JR, "Okay, What are the Odds?" Unpublished M.A. Thesis, Australian National University, 2002).

Evett IW, "Interpretation: A Personal Odyssey" in Aitken CGG and Stoney DA (eds), *The Use of Statistics in Forensic Science* (Ellis Horwood, 1991).

Gruber JS and Poza FT, "Voicegram Identification Evidence" (1995) *American Jurisprudence Trials* 54.

Hodgson D, "A Lawyer looks at Bayes' Theorem" (2002) 76 *Australian Law Journal* 109.

Hollien H, *The Acoustics of Crime* (Plenum, 1990).

Hollien H, *Forensic Voice Identification* (Academic Press, 2002).

Kinoshita Y, "Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants", Unpublished PhD Thesis, the Australian National University, 2001.

Künzel HJ, "Beware of the Telephone Effect: The Influence of Transmission on the Measurement of Formant Frequencies" (2001) 8 (No 1) *Forensic Linguistics* 80.

Nolan F, (1990) "The Limitations of Auditory-Phonetic Speaker Identification" in Kniffka H (ed), *Texte zur Theorie and Praxis forensischer Linguistik* (Max Niemayer Verlag).

Nolan F, "Auditory and Acoustic Analysis in Speaker Recognition" in Gibbons J (ed), *Language and the Law* (Longman, 1994).

Nolan F, "Speaker Recognition and Forensic Phonetics" in Hardcastle WJ and Laver J (eds), *The Handbook of Phonetic Sciences* (Blackwell, 1997).

Nolan F and Oh T, "Identical Twins, Different Voices" (1996) 3 *Forensic Linguistics* 39.

Ormerod D, "Sounding Out Expert Voice Identification" (2002) *Criminal Law Review* 771.

Robertson B and Vignaux GA, *Interpreting Evidence* (Wiley, 1995).

Rose P, *Forensic Speaker Identification* (Taylor & Francis, 2002).

Rose P, Osanai T and Kinoshita Y, "Strength of Forensic Speaker Identification Evidence – Multispeaker Formant And Cepstrum Based Segmental Discrimination With a Bayesian Likelihood Ratio as Threshold" in C Bow (ed), *Proceedings of the 9th Australian International Conference on Speech Science and Technology* (Australian Speech Science and Technology Association, 2002) (expanded version to appear in *Forensic Linguistics*).

Rose P and Simmons A, "F-pattern Variability in Disguise and Over the Telephone – Comparisons for Forensic Speaker Identification" in McCormack P and Russell A (eds) *Proceedings of the 6th Australian International Conference on Speech Science and Technology* (Australian Speech Science and Technology Association, 1996).

[End of Volume 4]