Chapter 3
Aural/Acoustic vs. Automatic Methods in Forensic Phonetic Case Work

Anders Eriksson

Abstract In this chapter focus will be on speech analysis in a forensic context. Both so called aural/acoustic approaches and automatic methods will be considered and their application in a forensic context described. Forensic casework introduces many challenges not found in the laboratory settings where the applied methods were originally developed. Forensic phonetic case work may involve many different types of tasks like voice comparison, speaker profiling, phonetic transcription and enhancement of poor quality recordings but in the present chapter only voice comparison will be described in any detail. The purpose of forensic speech science is to produce evidence that can be used in court. This introduces other types of challenges like the choice of presentation format. Pros and cons applying the traditional approach using verbal scales vs. using a likelihood ratio approach will be considered and put in the context of a current debate within forensics in general about the presentation of evidence.

3.1 Introduction

Until recently, most forensic case work involving speech has been carried out using what we may call aural/acoustic methods. During the past decade, however, automatic methods have been successively gaining ground. We have no statistics showing the precise proportion of case work done using automatic methods vs. aural/acoustic methods. We may say, however, that the majority of case work is still done using aural/acoustic methods only. A combination of both types is used by the police crime labs in Germany and France and by independent analysts in other places. In this chapter we will present pros and cons of both methods based, among other things, on our experience from a substantial amount of case work for the Swedish police. The focus will be on application and usefulness in actual forensic speech science case work rather than the scientific research that lies behind it.
This chapter will primarily be concerned with forensic phonetic case work involving what is often referred to as speaker identification, although we will see that appropriateness of this term may be questioned. But forensic speech science case work may also involve other tasks that should not be forgotten. One such task is transcription of recordings. This may sound as an easy task that anyone can do without any specific forensic or linguistic phonetic training. But this is very far from the truth. In forensic transcription it is often crucial to know exactly what is said. It is also absolutely necessary not to allow any guessing. If the transcriber cannot say with absolute certainty what words, expressions, figures are used this must be clearly stated. Untrained transcribers have a strong tendency to write down not what is actually said, including mispronunciations, speech errors, hesitations, repetitions etc., but what they “know” based on their linguistic knowledge as native speakers that the speaker “must have said” although it may in some cases be absolutely impossible to say with certainty if that was actually what was said. There are many reasons why transcription may be difficult and require a lot of training and experience to get it right. Poor recoding quality is one such factor. Other factors are overlapping speech, slur, holding the telephone away from the mouth, background noise, poor transmission of telephone calls etc. Let me cite just two examples where accurate transcription played a crucial role. The first case involved an illegal arms deal between criminal gangs. A conversation during the arms pick-up was recorded from the bugged mobile phone used by one of the criminals. But the telephone call had nothing to do with the deal itself. It just so happened that the speaker tried to speak to someone over the phone while at the same time instructing someone else who was physically present about where to get the automatic rifles. While doing so he moved the mobile phone away from the mouth making the reception over the phone extremely poor. It was nevertheless possible to extract enough of the content to say with certainty that what went on was connected with an arms deal. Another case involved suspected illegal trading on the stock market. In his case the sound quality of the bugged mobile phone calls was fairly good but here the speakers frequently used code words which made it absolutely essential to transcribe the exact words used even though they in many cases made no logical sense to the uninformed listener. I mention these cases just to make it clear that forensic speech science can be a lot more than just determining whether speech sample A and speech sample B come from the same speaker.

I also want to say a few words about cleaning poor quality sound files. As we all know from films and TV crime series, this is as easy as a, b, c. A few clicks on the buttons on the computer screen and the previously unintelligible speech in the sound file comes out as clear as crystal. In real life it is not quite like that. What most people, including many police officers, do not know is that cleaning in almost all cases involves the removal of potentially useful information. This may make the recording less annoying to listen to, but may at the same time remove information that in uncontrollable ways distorts acoustically based analyses. In certain cases, where the goal is to better hear and understand what is said, acoustic cleaning may be helpful. Cleaning is most successful (i.e. removes potentially important information only minimally) if the noise is strictly periodic and has a fairly narrow bandwidth. One such case involved a robbery of a taxi driver. During the robbery, which was recorded over the surveillance system installed in the taxi, there was a very annoying, narrow-band beep repeated with a regular frequency. Such sounds are fairly easy to remove without removing much else. Doing so made the conversation considerably easier to transcribe. This example illustrates that cleaning may indeed be useful but also that rather special conditions apply when this is the best solution. In most normal cases cleaning attempts make little or no difference or may even be harmful.

3.2 A Brief History of Forensic Speaker Recognition

The rest of the chapter will be about speaker recognition or comparison as we prefer to see it. But let me begin by giving a brief historic background of speaker recognition in a forensic context in order to give you a better understanding of where we are now and why we are just were we happen to be. For obvious reasons speaker recognition in the old days could only be aural. It was what we today would refer to as earwitness evidence. By that we mean, technically speaking, that the evidence is based solely on memory of the voice by the witness. The first case on record where earwitness testimony in a modern sense played a crucial role in a trial dates back to 1660. The case involved a certain William Hulet who was accused of regicide in the execution of King Charles I. A witness testified that he had heard the executioner, whose face was obscured, and that he knew that it was Hulet "by his speech". The jury found Hulet guilty of high treason and he was sentenced to death. Earwitness testimony should properly be regarded as a field of forensic voice comparison and there is a rich research literature on the topic. But since the present volume is devoted to the analysis of recorded material we will have to leave earwitness evidence for some other occasion.

The advent of recording machines opened new avenues for forensic speech science. A particular milestone in that development was the invention of new analysis techniques most notably the spectrograph. The basic ideas were published in the thirties [56] and the first commercially available machine (the Sonagraph) was manufactured in the early forties. Its usefulness for phonetic analysis is undisputable, but the possibility of performing speaker identification based on visual inspection of the spectrograms, a technique that came to be known by the name of voiceprinting, was grossly overestimated. Possibly the illusion was created by the superficial similarities between the patterns in a spectrogram and fingerprints. Anyhow for many years to come, voiceprinting, that is speaker recognition based on visual voice comparison of spectrograms, played a significant role as an investigative tool although it was not generally accepted as evidence in court. It also has to be said that it was not recognized as a reliable method by the leading phoneticians of the time and the accuracy claimed by its supporters could
never be verified in scientific studies. If we look at this development from a theory of science perspective we may perhaps call it a paradigm shift away from aural-only identification towards acoustically based methods. Today voiceprints are only used by some private investigators and interestingly also as an investigative tool by the FBI but its role in forensic speech science and case work is otherwise negligible. A comparable development took place in the Soviet Union and there is some evidence that automatic (or at least semiautomatic) methods were also tried. Very little about this is known today but there is an interesting account of such attempts at semiautomatic speaker recognition in the novel The First Circle by Solzhenitsyn.

3.3 Current Methods

As was mentioned in the introduction, the predominantly used approach in forensic phonetic case work today and since two or three decades back is the aural/acoustic approach. This approach is used by most forensic phonetic analysts in Europe and also elsewhere. Among the members of the only professional organization of forensic speech scientists, the International Association for Forensic Phonetics and Acoustics (IAFPA), the majority of members work within this framework. But as was also mentioned in the introduction, automatic methods are gaining ground. Both approaches will be briefly outlined in the following two sections. More details will follow in subsequent sections.

3.3.1 Aural/Acoustic Analysis Methods

Aural/acoustic methods mean a combination of linguistic judgements made by the examiner listening to the speech material combined with the use of acoustic measurements traditionally used by phoneticians to describe speech. In the former category we find such factors as speech errors, pathological speech problems, idiosyncrasies, dialect, foreign accent etc. In the latter we typically find measurement of formants, fundamental frequency mean and standard deviation, speaking rate etc.

We may schematically divide an aural/acoustic analysis process into the following steps:

1. Careful listening to the speech material
   This is to provide a first assessment of general factors like the duration of useful material and sound quality but also to identify traits that seem to be worth exploring in more depth.

2. Transcription of the material
   I am not aware to what extent transcription is routinely used by others but in our case work in Gothenburg we have found it very useful. First of all the entire material becomes searchable in a convenient fashion. Secondly, the transcription may be used in order to find and extract units in an automatic way for various types of analysis, for example formant analysis. The way to do this can be realized in many different but equivalent ways, but the way we do it is by using scripts in the acoustic analysis package Praat.

3. Linguistic analysis
   Valuable information may be found by examining the speech samples from a linguistic point of view analysing dialect, sociolect, accent, grammar, etc. This is the part that is the most difficult to imagine how it might be replaced by automatic methods.

4. Acoustic analysis
   The acoustic analysis always incorporates an analysis of the most common parameters like formants, fundamental frequency, intonation patterns etc., but many more parameters may be used. This will be explained in some detail in Sect. 3.4.

5. Summarizing the findings in a report
   This part is, in my view, the Achilles heel of the aural/acoustic approach as it is most commonly applied. If we were to follow the guidelines suggested by (p. 10, [42]) the analysis should be performed in two stages—feature extraction and feature comparison. Steps 1–4 may represent the feature extraction part. From the brief outline above it may be seen that these steps are fairly straightforward. But the last step, feature comparison, is more problematic. Nolan suggests that each compared speaker should be represented by a point in a multi-dimensional space using the values from the analysis of each parameter as co-ordinate values. Based on these representations some distance metric should be applied in order to estimate the inter speaker distances. It is just that this procedure is hardly ever followed and if we were to do so, there are no commonly agreed standards on how to do it.

An alternative way of summarizing the results would be by applying the Likelihood Ratio framework, but this is at present fairly uncommon. One reason is that if we are to use all or many of the parameters described in Sect. 3.4 there is simply not enough data available on the distribution of parameter values in the relevant population. Some forensic speech analysts therefore suggest that the Likelihood Ratio approach is not a realistic approach at the present moment. An alternative approach has therefore been suggested in a UK position statement [19]. How this approach differs will be discussed in Sect. 3.8.3.

The aural/acoustic approach is largely experience based and analyses may therefore differ between individual analysts. This does not necessarily mean that the conclusions are inaccurate, but it is not easy to see how accuracy could be measured and compared. This is in my view the principle weakness of the approach.

Most references in Sect. 3.4, where strengths and weaknesses of various parameters used in forensic case work will be discussed, refer to studies or case work carried out in the aural/acoustic tradition reflecting the predominance so far of this approach.
For the interested reader who wants an in-depth account of the aural/acoustic approach there are numerous text books that may be consulted. The following may be mentioned as a guide: [3, 23, 42, 49].

3.3.2 Automatic Methods

Strictly speaking the word ‘automatic’ as in ‘Automatic Speaker Recognition’ (ASR) only means that we leave the task to be carried out to a computer (or some other type of machine) rather than do it manually. Analysis of all the various parameters described in Sect. 3.4 could therefore in principle be carried out using fully automatic or semiautomatic methods in order to save time if for nothing else.

As things stand today, however, most automatic speaker recognition systems should more properly be called automatic voice recognition systems or voice comparison systems as we prefer to call them since ‘recognition’ makes reference to individuals rather than speech samples which is what the systems work with. The use of the word ‘speaker’ is misleading in the same way. What is recognized is not, strictly speaking, a speaker interpreted as an individual. That would require knowledge of all kinds of other evidence (e.g. fingerprints, DNA, eyewitness reports and so on) which is not available to the forensic speech analyst. The reason why ‘speaker’ has come to be used is that there is also ‘speech recognition’ which is something completely different.

Most existing ASR systems do not compare speech but voices. This is an important distinction to keep in mind. Various techniques used in ASR are described in other chapters of the book so it suffices just to remind the reader that what is analyzed in the most widely used ASR systems is the voice, or more precisely what we may refer to as the timbre of the voice. The Mel Frequency Cepstral Coefficients (MFCCs) that are used in those systems are well suited to describe timbre and moreover modeled on studies of the human perceptual system. So, if it is voice (in this particular sense) comparison we are after, this approach is just right. But as we have seen, speech is much more than voice. A given speaker may be described in many other ways like speaking style, dialect, speaking rate, patterns of intonation, word stress etc., but also more speaker specific traits like speech errors, idiosyncratic choice or words, mispronunciations etc. There are ongoing attempts to incorporate some such factors in ASR but to our knowledge no such system is used today in actual forensic case work. In our work we make a clear distinction between voice and speech. Our understanding of ‘voice’ is inspired by the Modulation Theory of Speech which will be described later. It differs a bit from how voice is usually seen but timbre is an important factor.

Recognizing an individual, in our case a speaker, would require a method by which it is possible to single out a single individual in an open set. This is a particularly important aspect to keep in mind in connection with forensic speech science. In some areas we may perhaps approach open set recognition. With respect to DNA and fingerprints it would, at least in principle, be possible to construct a complete population database. Given such a database and sufficiently reliable analysis procedures, it would perhaps be possible to pick out a single individual matching a given sample.

Would something similar be possible for speech even in principle? Well, in theory it would be possible to construct a population database for speech, but it has to be remembered then that such a database, in contrast to a DNA or fingerprint database, would soon be outdated. Whereas DNA or fingerprints do not change over time for a given individual (except in the case of hand injury destroying the fingerprints) human voice and speech changes over time. Voice changes are for example caused by ageing, disease, smoking habits, and injury. Also, voice characteristics may change even in a short time perspective for example due to a cold or the emotional state of the speaker, under the influence of drugs, or environmental conditions such as background noise.

So, the bottom line here is that we should perhaps reserve the expression ‘automatic speaker recognition’ for the Science Fiction domain and focus on what we can do, namely speaker comparison, at present by and large limited to voice comparison. Automatic Voice Comparison (AVC) would be a more appropriate name for it.

A known problem for automatic systems is so called mismatch conditions, for example differences in sound quality between speech samples due to transmission channel differences. If the questioned speech samples are recorded over the telephone and the reference database consists of direct recorded speech then the performance of the system is worse than if the reference data are also telephone speech. Now simple cases of variation like that can to some extent be taken care of, but the in forensic phonetic case work variation in channel quality is much wider. What if the questioned sample is recorded with a poor quality microphone three meters away from a shouting bank robber in a room with severe reverberation, a case which is typical rather than uncommon? In real life case work there are countless examples of mismatch conditions of this type. The very severe types of mismatch conditions that we see in actual case work are a serious problem for automatic as well as aural/ acoustic analysis.

In our own work at the University of Gothenburg, we use an ASR system for voice comparison. The system we use is the French ALIZE SpkDet packages which are developed and released as open source software under a so-called LGPL license [4–5]. Voice comparison results are then combined with traditional aural/ acoustic analysis and the combined results expressed using the verbal 9-point scale described in a later section. A step forward would be to combine the results in a Bayesian Likelihood Ratio framework, but at present it is not clear how this may be implemented.

As far as we are aware, the most commonly used or tested system by police labs is the Spanish Batvox system by Agnito. It has been tested by the police in Sweden and several other countries. We have no information, however, to what extent it is being used in actual case work. Batvox has consistently performed well in evalu-

3.4 A Look at Some of the Most Important Parameters Used in Forensic Speech Science

In the following no sharp distinction will be made between analyses performed within the aural/acoustic paradigm and those using automatic methods. Although it is true that most automatic methods so far have been based on voice analysis, most notably using a Mel Frequency Cepstral Coefficient (MFCC) approach, there is no reason why automatic or semi-automatic methods should not be possible to extend to other types of analyses. And the aural/acoustic approach may, of course, also include voice analysis. Several studies referred to below have been performed within a Likelihood Ration framework. This becomes important when presenting the evidence in court. Here the Likelihood Ratio based approach offers a standardised way of presenting the results and all types of analyses may in principle be presented in identical ways. The largely experience based aural/acoustic approach does not easily lend itself to any similar degree of standardization. This issue will be discussed in the section on how to present the results in court together with an account of new trends in forensic science in a legal context in general and what some claim is the beginning of a paradigm shift in the forensic speech science domain [38, 39].

One problem with many of the commonly used acoustic measures is that they are not particularly robust with respect to such factors as speaking style, recording quality, channel bandwidth etc. In the following a brief description will be given of the problems encountered in the context of speaker comparison when using the different parameters and measures but also something about how the problems may be dealt with.

The following account will allow a rather prominent place for a description of the use of two traditional parameters—fundamental frequency and formants. As will become apparent, intra-speaker variation and many other factors make these two parameters rather dull instruments for speaker comparison. Analysis based primarily on formants and fundamental frequency is therefore slowly being replaced by more sophisticated approaches but since formants and fundamental frequency have played such a prominent role in the historical development of forensic speech analysis and are still widely used in forensic case work, they merit a reasonably prominent place in a summary of forensic speech analysis. Describing them in some detail will also serve as an introduction that paves the way for a description of more sophisticated approaches.

3.4.1 Fundamental Frequency (F0)

Fundamental frequency is the frequency of vibration of the vocal cords in phonated speech. Its usefulness for speaker comparison in forensic case work has been an issue for a long time. As descriptors of individual differences in fundamental frequency, long-term distribution measures such as arithmetic mean and standard deviation are often suggested [49]. To some extent means and standard deviations depend on the duration of the speech sample but there is no general agreement on what minimum duration is required to yield representative results. Horii [24] suggests that recordings should exceed 14 s. In other studies, ranges from 60 s [42] up to 2 min [3] have been suggested as a minimum. Rose [49], reports that F0 measurements "... for seven Chinese dialect speakers stabilised very much earlier than after 60 seconds", as suggested by [42], implying that duration requirements may be language specific. Braun [7] points out that the measures are also dependent on physiological factors such as age, smoking habits, dequeue, intoxication and psychological factors such as the emotional state of the speaker, sleepiness or vocal effort, but maintains that 15–20 s should be sufficient "if the communicative behaviour may be considered "normal"". She also mentions the influence of ambient noise on speaking style. In spite of all these sources of variation, fundamental frequency has nevertheless been shown to be a useful forensic phonetic parameter in several investigations [2, 3, 22, 29, 49]. Voice disguise is another factor that may affect fundamental frequency. A rather common type of disguise is to raise or lower the fundamental frequency. In a study by [31] where 100 subjects were asked to read a text using a voice disguise of their own choosing, a very common choice was to either raise or lower fundamental frequency. There was a tendency for speakers with lower than average fundamental frequency to lower it even further and for those with a normally high level to raise it. For obvious reasons such manipulations will render fundamental frequency comparisons rather meaningless in forensic case work.

Positive skewing of the F0 distribution for a speaker is typical (Jassem et al. 1973). This means that the values will not be symmetrically distributed around the mean. Using the traditional measures for describing fundamental frequency level, such as mean or median values, may therefore yield misleading results. Positive skewing occurs primarily because there is much more room for fundamental fre-
quency variation upwards that downwards. If fundamental frequency is measured in semitones skewness is reduced, but not eliminated.

In order to make the F0 measure insensitive to some of the above mentioned types of variation alternative ways of representing fundamental frequency have been searched for. We use what we call the baseline as a representation of fundamental frequency since it better represents a speaker specific fundamental frequency level. A further advantage is that this measure is considerably more robust than the mean or median. In a study by [35] it was shown to be almost completely insensitive to variation in channel quality and emotional expression. The concept of baseline will be explained in Sect. 3.6.

All the above factors concern intra-speaker variation, but in forensic case work the usefulness of a parameter also depends on inter-speaker variation. Even if we are able to determine the characteristic frequency for a given speaker quite precisely, the predictive power of fundamental frequency may still be low if the fundamental frequency of the speaker falls in a region where were most other speakers cluster. The F-ratio has been proposed as way of expressing the intra- vs. inter-speaker variation [5, 25, 28]. The likelihood ratio, which will be explained later, provides a framework for combining data on intra- and inter-speaker variation. Given that we have access to sufficiently large and appropriate databases this would give us a fair estimate of the usefulness of fundamental frequency data in a given case.

3.4.2 Formants

Formants are resonance frequencies in the vocal tract. They are determined by the shapes and volumes of the different cavities of the tract. From this description alone one may draw the conclusion that the formant frequencies will change continuously as we vary jaw opening, tongue position and lip shape during speech. There are, however, certain regularities in the formant frequencies. During normal speaking conditions they are fairly invariant for a given speaker, speech rate and vowel. Also, the vowel schwa, which is produced with the speech organs in a relaxed neutral position, may be correlated with vocal tract length. The schwa vowel may also be used as a measure of the timbre of the voice under neutral conditions. Traunmüller [58] in his Modulation Theory of Speech has suggested that this quality may be seen as a carrier wave which, when we speak, is modulated to convey the linguistic message but also to express various paralinguistic qualities. This will be further explained in Sect. 3.6.

Even though formant frequencies vary with the articulation of speech sounds, they also depend on the size and proportions of the speech organs of an individual

---

2 A typical male speaker reading a sentence in a neutral tone may have a mean of 110 Hz, a minimum F0 of 85 and a maximum value of 160 Hz. In a liveller reading the minimum stays about the same, whereas the maximum may exceed 200 Hz.

---

speaker. Given that we were able to separate the contribution of the “neutral” vocal tract size and proportions from the modulations caused by articulation, formant frequencies should in principle be possible to use as characteristics of a given speaker. There are, however, many normally occurring conditions which may alter the formant frequencies making such a separation very difficult. Speaking rate is one such factor. Imaizumi and Kiritani (1989) have shown that the two highest of the formants normally considered (F3 and F4) “vary significantly with rate”.

Formant frequencies have also been shown to change as a function of vocal effort [61]. In loud or shouted speech, which by the way is the speaking style we find in most recordings of bank robberies, the first formant (F1) has been shown to increase [14]. Reference recordings in bank robbery cases are, however, often taken from interrogations with the suspect speaking in a relaxed tone of voice, which makes acoustic comparison problematic. To some extent we may predict the change in formant frequencies as a function of vocal effort but far too little is known about how vocal effort affects formant frequencies in general to make any more precise predictions.

Channel bandwidth is another factor known to affect formant frequencies. This is the case for telephone recorded speech samples. The effect has been examined in several studies involving aural voice recognition. The first studies of this kind used telephone quality simulations like pass-band filters [53]. In other studies (e.g. Rathborn et al. 1981) landline phone recorded signal samples were used. Studies of the effect of telephone sound quality on aural voice recognition have, however, in general shown no significant effect of telephone quality on recognition accuracy. In acoustic analyses, however, sound quality effects have been observed. A marked effect on the first formant was observed in a study by [34] for example. In a study with the somewhat alarmist title “Beware of the telephone effect”, [32] reports a strong effect on primarily the first formant (F1) in telephone mediated speech samples. In many cases the questioned speech samples are bugged telephone calls and the reference samples direct recorded police interrogations. In such cases, KüNZEL argues in rather strong terms, formant analyses should not be used. KüNZEL’s views are not universally endorsed but we may say that there is general agreement that formant analysis should be treated with caution where a mixture of direct and telephone recorded samples are involved.

Today most bugged telephone calls in connection with crimes are made using mobile phones. Reports from forensic investigators in Sweden, the UK and Germany indicate that a substantial and increasing number of cases involve mobile phone recorded speech [43]. A detailed technical overview of the various ways in which mobile phone transmission affects sound quality may be found in [21]. Mobile phone transmission introduces a number of other sound quality degrading effects in addition to bandwidth limitations and these effects have been shown to affect formant analysis. Byrne and Foulkes [8] showed that mobile phone transmission had a significant effect on the first formant (“29 per cent higher than in the direct condition”) and in the study by [21] the three first formants were all affected and under certain transmission conditions rather dramatically so.
Some types of disguise may also affect formant frequencies. Papcn [44] looked
at imitation performed by both professional and amateur mimics and found that
mimics succeeded to approach the formant values of the targets, at least to some
degree. Similar results have been obtained in other studies [15, 17]. It has to be said,
however, that imitations of a given target are rare in actual case work but changing
ones voice in similar ways may also have the effect of changing the formant values
away from those of the speaker’s normal voice.

To sum up we may say that although formant positions depend, among other
things, on the physical properties of the vocal tract of a speaker and are thus to some
extent typical of a given speaker, there are so many other factors that may influence
formant positions that one may seriously question the value of formant analysis for
forensic speaker comparison.

3.4.3 Rhythm and Other Patterns in the Time Domain

Various measures conventionally connected with speech rhythm have been suggest-
ed also for speaker comparison but also other measures that reflect motor patterns
connected with speech articulation have been suggested. Not enough is yet known
so it is difficult to predict how successful these attempts will be for speaker com-
parison but what little we already know is reason for a certain degree of optimism.
If we begin at a rather fundamental level we may observe that highly automatic motor
patterns have been shown to vary from individual to individual but tend to be stable
and invariant once they have been firmly established. This has for example been
demonstrated to be the case for typing [57] and gait [6, 65]. If such motor patterns
are sufficiently stable and inter-individual differences sufficiently large it should be
possible to use them in person recognition. This has also been demonstrated for
articulation (see for example the two references cited above). Speech articulation is another
example of such highly automatic motor activities and if articulatory patterns are
also fairly characteristic of a given speaker then they may be useful for speaker
comparison. A study of impersonation by [17] showed, for example, that although
the professional impersonator was quite successful at approaching the target values
in terms of mean fundamental frequency, over all timing, global speaking rate,
and to some extent also formant frequencies, his timing at the segmental level was
always closer to his own personal timing patterns than to those of the target voice.

Three renditions of the same texts were compared—the target version, the imitation,
and the impersonator’s rendition in his own natural speaking style. The three
versions were transcribed at the segmental level marking the segment onsets, produc-
ing three timing sequences (target, imitation, and natural) for each utterance. The
timing deviations from the target, segment by segment, for the imitations and the
natural renditions were plotted for whole utterances, selected phrases and words.
In all cases there were minimal differences between the two versions produced by
the impersonator but rather substantial deviations from the targets. In other words,
although the impersonator was quite successful at approaching the targets for the other
tested parameters he seemed unable to change his own personal timing patterns in
the direction of the target to any appreciable degree. Another example along the
same lines are the studies of formant dynamics by McDougall [36, 37] suggesting
that individual timing differences in selected articulatory movements may be used
for speaker comparison. Several attempts have been made to describe the rhythmic
character of languages [20, 48]. Various duration based measures used for language
classification like the Pairwise Variability Index (PVI) and %V (percent of the total
duration that consists of vowels) have been tested in experiments aiming at speaker
comparison [10, 64, 66]. In a study by [11], 10 German speakers were recorded
reading a short text 5 times with speech rates varying between normal and “as fast
as possible”. Test variables were %V and PVI (mentioned above) and AC (standard
deviation of consonantal interval duration). All test variables remained stable within
a given speaker across speaking conditions as did inter-speaker differences. In an
experiment involving disguise [12], one native speaker of English read 29 sentences
in his own normal voice and using disguise in the form of dialect imitation. In
the disguise condition the speaker also increased fundamental frequency mean and
standard deviation and voice breathiness. In spite of these manipulations, a number
of tested temporal measures “revealed no differences between the normal and
disguised conditions”. It is too early to say how useful these measures will turn out to
be in real life case work, but the results so far are very promising.

3.4.4 Speaking Rate

Speaking rate is the number of speech units produced per minute or per second. The
most common measure is words per minute (wpm). Syllables per second (syll/s) is
a more fine grained measure and should be preferred. In addition there is the ques-
tion of how to handle pauses. When we talk about speech rate pauses are included.
In most cases it makes sense, however, to separate the rate when actually speaking
from pauses. In this case we speak of articulation rate. Typical articulation rates for
normal speech are 5–7 syllables per second. Syllable rate is measured syllable onset
to syllable onset or vowel onset to vowel onset. In reasonably good quality speech
syllable onset may be detected automatically, but as speech quality deteriorates
so does the precision in syllable onset detection. In poor quality recordings there
is often no other alternative than manual annotation. We may say that speech rate
reflects speaking style and that articulation rate reflects motor timing much in the
sense described in the previous section. Based on what we know about the stability
of motor timing we would predict that articulation rate might also be reasonably
speaker specific showing moderate intra-speaker variability making it an interest-
ing parameter to investigate for forensic purposes. And this assumption has indeed
been confirmed in at least a handful of studies. Künzel [30] studied various aspects
of speaking tempo, varying the speaking style from read to spontaneous speech.
Articulation rate turned out to be “almost entirely unaffected” by speaking style and
“remarkably constant within speakers”. The discrimination power for articulation
rate (taking both intra- and inter-speaker variation into account) also turned out to be quite good. Speech rate which also reflects speaking style showed, as might be expected, a higher degree of intra-speaker variation and a considerably lower discrimination power.

3.4.5 Voice Quality

Voice quality may also be a useful parameter in speaker comparison. Until now most research on voice quality has been carried out in connection with studies of speech pathology. It may well be the case, however, that some of the measures used in this field may also be useful for speaker comparison, but this possibility has been tested to a rather minor extent. Two of most commonly used acoustic voice quality descriptors in speech pathology are jitter and shimmer. Jitter is the period-to-period variation in fundamental frequency and shimmer the corresponding variation in amplitude. Some studies do exist that have tested the usefulness of jitter and shimmer for speaker comparison. Wagner [63] used measures of jitter to successfully distinguish between a group of speakers with pathological voices and speakers with normal voices. Farrús and Hernando [18] used both jitter and shimmer as additional factors in an automatic speaker verification system. Jitter and shimmer used on their own did not produce “good enough” se results but adding them to a system based on prosodic and spectral parameters improved the results of the system.

Other possible, but so far untested, voice descriptors include glottal pulse shape and glottal source spectrum. Voice creak should also be mentioned. Voice creak is in general terms a perceptual category which may have several different causes. When the transglottal pressure decreases below a certain level, typically towards the end of a phrase, this often results in an abrupt halving of the fundamental frequency. It may also result in what is called diplonaphia which is a regular pulse-to-pulse variation in amplitude with every second pulse being considerably weaker. The degree of creak varies between speakers but if the inter speaker variation is great enough to be forensically useful has not been tested. Most other descriptors are primarily auditory based and therefore suffer from inevitable subjectivity. Some of those are leaky voice (glottal folds do not close completely letting a stream of air pass through), harsh voice (tense vocal folds), whispery voice (incomplete voicing of normally voiced segments), hoarse voice (irregular closing) and falsetto (increased fundamental frequency to a point resulting in sharply falling glottal spectrum). None of this has been tested with a view to forensic application as far as I am aware.

3.4.6 Linguistic Factors

Last but not least, the importance of linguistic factors must be pointed out. Such factors are dialect, accent, idiosyncratic expressions, unusual pronunciation errors etc. In actual forensic case work it is not uncommon for such factors to carry a rather decisive weight.

3.5 Disguise

Recordings of disguised voices are not very common in forensic case work but they do occur. As mentioned above, changing fundamental frequency is one rather common type of disguise but there are others. In the study by [31] mentioned above changing the fundamental frequency was one of the chosen types of disguise. Künzel further reports that the most common other types of disguise found in cases analysed at the crime lab of the German federal police (BKA) were falsetto, persistent creaky voice, whispering, faking a foreign accent and pinching one’s nose. Although all these types of disguise are of a fairly simple nature they may nevertheless make speaker comparison considerably more difficult. A more detailed account of various types of disguise and their effect on speaker identification may be found in [16].

Harsh voice is sometimes used as a form of disguise. It is quite easy to do and may severely diminish the chances for successful speaker comparison. Automatic methods in particular suffer greatly. We had a kidnapping case where the perpetrator used harsh voice for his ransom calls delivered over a mobile phone. The reference was two mobile phone calls the suspect made to a female acquaintance. The automatic analysis failed to produce any conclusive results and although there was quite a bit of similarity in terms of certain expressions and speaking style that matched the known background of the suspect and made us fairly convinced that the speaker in the ransom calls was the same as the speaker in the reference calls, it was not possible to draw any conclusions that would have a chance of standing up in court. This is an example of how even a very simple form of disguise may effectively destroy the possibilities of successful forensic speaker comparison.

Speaker comparison of disguised speech may be performed in four steps, (1) Identifying the type of disguise, (2) Applying known models of the acoustic consequences of the disguise in question, (3) Eliminating as far as possible the distortions caused by the disguise from the speech sample and (4) Using the now ‘cleaned’ speech sample for speaker comparison. There is ongoing research addressing the problem along these lines [45–46] but it is still too early to say how far it will take us.

3.6 The Modulation Theory of Speech

In our work we use two concepts inspired by the Modulation Theory of Speech [58]—a definition of ‘voice’ inspired by the notion of carrier in the modulation theory and a measure of fundamental frequency we call the ‘baseline’. A brief de-
cription of the modulation theory and why it is relevant in an account of forensic speaker comparison is therefore in its place.

As you may notice you do not find any mention of "voice source" in the Table 3.1. That is because voice source does not play any role as an independent parameter in this theory. The partly corresponding parameter is instead the carrier which is a fundamental concept. The concept of carrier is inspired by, not identical to, the concept of carrier wave in signal theory. In the modulation theory all types of information in speech communication is transmitted as modulations of the carrier, hence the name "Modulation Theory". What then is the carrier? In the 1994 paper where Traunmüller first presented his theory the carrier is described in the following way: "we should think of the carrier as having the properties characteristic of a 'neutral' vowel, approximately [O]".

The relevance of the concept of carrier in the present context is that it seems as a very reasonable way of thinking about what 'voice' should mean; a schwa vowel produced with a relaxed vocal effort and fundamental frequency. The nice thing about it is that one way of representing the voice of a speaker acoustically would then be an MFCC representation of the carrier. In our case work we make a distinction between voice and speech and our definition of voice is the carrier. We do not use the word carrier in our reports, however, since the term is not well known but refer to it as 'voice'. Now, an MFCC analysis is not limited to the analysis of neutrally produced schwa vowels but include a great variation of speech sounds; that is sounds represented by modulations of the carrier in the modulation theory sense. How to handle this in voice comparison while maintaining the idea of carrier comparison will be explained in 3.6.1.

Another idea inspired by the modulation theory is the idea of a neutral, speaker specific fundamental frequency level. In [58] this is described as "a stable point about 1.5 SD below the mean value of F0". This level has been empirically determined by studying speaker behaviour. For example this level has been shown to behave like an eigenvector for fundamental frequency when speakers vary the liveliness in their speech. We refer to this level as the fundamental frequency baseline and it may be thought of as the frequency a speaker returns to after fundamental frequency excursions used for prosodic or other purposes. In a study by [60] this level was estimated as 1,43 standard deviations below the mean. This can be translated as approximately 7% up from the lowest frequencies used by the speaker and going much lower will normally result in creak which speakers tend to avoid. There is, in principle, no corresponding upper limit, however, resulting in a distribution bias towards higher frequencies.

### 3.6.1 More About Voice vs. Carrier

In the modulation theory it is assumed that a complete separation may be made between the carrier and the information (both linguistic and paralinguistic) that is transmitted by modulating the carrier. As was mentioned, a neutrally produced schwa vowel would be a close approximation of the carrier. That said it should be clear that a neutrally produced /æ/ comes less close to the carrier. From this we may draw the conclusion that the MFCC's used in the automatic voice comparison systems are "contaminated" by the message, in particular the phonology of the language (the set of contrastive speech sounds used by the language in question). Does this mean that for automatic voice comparison to produce reliable result one must make sure to use a reference database which closely matches the particular language, dialect, speech style etc. used in the known and questioned samples to be analysed? This is basically an empirical question that has to be examined in experiments comparing different combinations of test material and reference databases.

If the MFCC's represented the carrier and only the carrier such tests would not be needed, at least not as a function of language. The carrier only depends on the physiological properties of the speaker (size and proportions of the speech organs). It seems reasonable to suggest, however, that these properties are independent of language. If we had a perfectly working automatic carrier comparison system, then if the known and questioned samples were spoken in German it would not matter if the reference database was Spanish or German. With the MFCC approach it does matter but precisely how much has not been studied to the extent that it would be needed to give a precise answer. Informal test using German speech samples tested against a Spanish database indicate, however, that matching the language is not as crucial as one might think (Künzel, p.c.). Dialect within a given language should influence the results even less. If we assume that there are no important anatomical differences between speakers of different Spanish dialects in Spain, and that the inventory of meaningful speech sounds is about the same in the dialects then it should not make much of a difference if we use one dialect or another for the reference database. This hypothesis has in fact also been tested. In a study presented by Moreno et al. at the IAFPA annual meeting in 2006, it was shown that the differences between three tested dialects of Spanish were not great enough to make

<table>
<thead>
<tr>
<th>Quality</th>
<th>Information</th>
<th>Phenomena</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conventional, social</td>
<td>Dialect, accent, speech style, ...</td>
<td>Words, speech sounds, prosodic patterns, ...</td>
</tr>
<tr>
<td><strong>Expressive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychosocial within speaker variation</td>
<td>Emotion, attitude, ...</td>
<td>Phonation type, register, vocal effort, speech rate, ...</td>
</tr>
<tr>
<td><strong>Organic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anatomical between speaker variation</td>
<td>Age, sex, pathology, ...</td>
<td>Larynx size, vocal tract length, ...</td>
</tr>
<tr>
<td><strong>Perspectival</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial, transmittal</td>
<td>Place, orientation, channel, ...</td>
<td>Distance, channel distortion, ...</td>
</tr>
</tbody>
</table>
any appreciable difference for the recognition results. In the study, using the MFCC based Batvox system from Agnitio, the “suspects” were speakers of Andalusian and the reference databases that were used were Andalusian, Castilian and Galician. The test results were by and large the same regardless of reference dialect. The conclusion we may draw from this, albeit with some caution given that this is just one study, is that although the MFCC:s represent not just the carrier but also the phonology of a given dialect, the differences between voices obtained in the automatic voice comparison come close to the assumed differences in carriers between the speakers. It therefore makes some sense to think or the inter-speaker differences obtained in automatic voice comparison as carrier differences while keeping in mind that the MFCC:s themselves are not the carrier.

3.7 A Summary So Far

In a book on phonetically based speaker recognition by [42] he suggested that the following criteria should apply to parameters used in speaker recognition.

1. High between speaker variability.
2. Low within speaker variability.
3. Resistance to attempted disguise or mimicry.
4. Availability.
5. Robustness in transmission.

As we have seen in the survey above, the commonly used parameters do not in general comply fully with these criteria. Criteria 4 and 6, availability and measurability, is usually not a problem but all the parameters reviewed above have shortcomings with respect to the other criteria. A way to express the interdependence of criteria 1 and 2 is to calculate the Likelihood Ratio (See Appendix B). This requires access to suitable databases, but representative databases are seldom available. This is a severe shortcoming. There are, however, attempts at the level of basic research that show how this may work. Using a small database of elicited speech data from 13 male speakers of Standard Japanese, [25] used formant data from the database recordings in a recognition test comprising of 90 same-speaker pairs and 180 different-speaker pairs using the Likelihood Ratio approach. Five of the 180 different-speaker pairs were wrongly classified as the same speaker and 9 of the 90 same-speaker pairs were classified as different speakers. Others have performed similar experiments. Alderman [5] used formant values for F1, F2 and F3 extracted from recordings of eleven male speakers of Australian English. Vowels were elicited in a /h s d/ context to be comparable to data in the reference database (the so called Bernard data). Same-speaker pairs were correctly identified in 72.2% of the cases and different-speaker pairs correctly identified in 99.6% of the cases. Descriptions of other experiments along similar lines may be found in [52] and [50].

In Sect. 3.4 it was seen that using single measures to describe a parameter, say the mean value to describe fundamental frequency, does not lead to very precise speaker discrimination. What we may learn from other scientific fields, for example genetics, is that combining several measures representing different aspects of the same parameter may substantially increase the predictive power. The effect of using such a combination of measures is nicely illustrated in [26] and [27]. In the first of the two studies [26], she tested the discriminative power of fundamental frequency long term mean and Standard Deviation. The reference data were from a corpus of recorded spontaneous speech by 90 male native Japanese speakers. Test data were casual speech by 12 native Japanese speakers recorded for the experiment. The discriminative power of the tested measures turned out to be low. Kinoshita sums up the results the following way: “These LRs are practically unity; meaning that the evidence produced using this method has no informative value with respect to the given speaker’s identity”. In the second study [27] not only long term F0 mean and SD but also kurtosis, skew, modal F0 and the probability density of modal F0 were included. All measures were found to be useful to some degree for speaker discrimination, but the real advantage came from combining them all. What we may learn from these two experiments is that the use of a single descriptor of a given parameter may suggest that the parameter is a weak or even useless parameter in speaker comparison while combining many aspects relevant for the description of the given parameter may make a radical change.

Robustness in transmission is another factor to be considered according to [42]. One should perhaps generalize that to “robustness” in general. Transmission in the technical sense of channel quality, like GSM transmission, may indeed be a problem, but there are many more threats to robustness. Some of them, like speech liveliness, vocal effort and disguise have been mentioned above in connection with describing the analysis of different parameters. We may also mention background noise and reverbation which is a common problem. Combining many aspects of a given parameter or using a Likelihood Ratio approach does not solve these problems. If the robustness of the used factors is poor then the combined results will also be poor and an LR approach does not change that. It is therefore important to find ways of representing the various components that are as robust as possible. One such measure, mentioned above, is the base line as a measure of fundamental frequency suggested by [35] which is far more robust against some of the more common sources of error like speaking style and recording quality than the mean or median commonly used. A very promising approach from the point of view of robustness is also the use of measures in the time domain described above. These measures should be particularly robust against channel quality degradation. Measuring durations is often possible even in recordings of such poor quality that normal measurement or even F0 measurement is more or less meaningless. We need to see more studies where timing measures are used, however, before any more far reaching conclusions may be drawn.
3.8 Presenting Evidence in Court

In this section I will discuss how results may be presented in court. As will be seen, there is no single agreed upon way of doing this and we seem to be in the middle of a transition period, some say moving towards a paradigm shift [54], where several competing ways of presenting the results co-exist. One might expect that the only existing professional organization of forensic speech scientists, the IAFPS mentioned earlier, should have guidelines instructing members how to present analysis results. But this is at present not the case. This situation is not unique but rather a mirror image of a lively international debate over the presentation of forensic evidence in general.

3.8.1 Presenting Evidence in the Aural/Acoustic Tradition

The presently most common way of presenting the results of a forensic phonetic investigation is to use a verbal scale. The following nine-point scale is the one recommended by the Swedish police [62] and used by us in most case work.

+4 The results support the hypothesis with near certainty
+3 The results strongly support the hypothesis
+2 The results support the hypothesis
+1 The results are inconclusive
0 The results contradict the hypothesis to some degree
−2 The results contradict the hypothesis
−3 The results strongly contradict the hypothesis
−4 The results contradict the hypothesis with near certainty

The “hypothesis” is usually that the analysed speech samples are from the same speaker. Verbal assessments of the same or a similar type are used by the Finnish, French and German police (p.c.), and most likely in many other places. It goes without saying that such a scale invites subjectivity, in particular when weighing and combining the results from analyses of several parameters. This does not necessarily mean that the analyses are unreliable. It is likely that several qualified and experienced analysts working with the same material would come to the same conclusions in most cases. When we have performed analyses in co-operation with others there has been little disagreement about the conclusions.

3.8.2 The Likelihood Ratio Framework: A Possible Paradigm Shift

There has, however, been criticism of the relative subjectivity of the present practice and a move towards a Likelihood Ratio based approach [38–39, 51]. A broader criticism of forensic science in general has been raised both in the US [41] and the UK [33]. Forensic speech science is not covered in the two reports but most of what is said is obviously relevant also with respect to speech science. The points of criticism include: Disparities in the Forensic Science Community, Lack of mandatory Standardization, Certification, and Accreditation, Interpretation, and Measures of performance (pp. 5–8, [41]) and further that “The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine its validity” (p. 9). Not surprisingly, DNA evidence is seen as a model.

With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source (p. 7, [41]).

We may see the trend towards an increased use of the Likelihood Ratio approach in forensic speech science as a reaction to the subjectively based reporting described above and there is an obvious attempt to raise the status of forensic speech evidence to a DNA like level if possible. This is all very well, but we must at the same time recognize the substantial differences between the two. A more detailed comparison with DNA is misleading for several reasons. A DNA sequence is discrete (after removing measurement errors) whereas speech variables are for the most part scale variables. Speech may also be expressed by a large number of different (and at least partly independent) variables like voice quality, fundamental frequency, formant structure, timing etc. DNA is stable and invariant over the life time of an individual, speech is not. Even if we consider only one parameter like F0, we have seen that it varies not just over the life time of an individual but sometimes even on a minute by minute basis. Also, whereas the Likelihood Ratio approach may be applied to all parameters that may be described using scale variables, there are other aspects of speech for which there are no obvious descriptions of this type. We may think of speaking style, dialect, foreign accent, speech errors etc. In principle we may think of estimates of the number of speakers of a given dialect etc., but that too introduces a number of problems; dialectality may also be seen as a gradual variable for example. In the Daubert ruling (See Appendix A) the requirement for expert testimony is that it should be based on “scientific knowledge” and “evidentiary reliability”, but the Supreme Court also describes the Daubert standard as “flexible”. We may perhaps interpret this as a licence to use DNA-like methods as far as this is possible, but may also allow ourselves to supplement this type of evidence with scientifically based judgement that must not be readily expressible in quantitative terms but has been shown to possess some degree of “evidentiary reliability”. While we should indeed strive towards more stringent descriptions, we must also be aware of what is and is not realistic. These last observations are more or less what motivated the UK position statement described in the next section.

An advantage with the verbal scale quoted above is that it is easy to understand and also that it is commonly used, the disadvantage is that it is subjective. The Likelihood Ratio is the exact opposite, it is to a very minor extent subjective, but it is far from obvious how the results should be presented in a way that may be easily understood by non-specialists. Does a Likelihood Ratio of 2374 really mean anything to a jury or a judge? Strictly following the Bayesian framework the number
means that whatever your prior evidence led you to believe it will now have to be
revised by saying that it is now 2374 times more likely that the combined evidence
weighs against the suspect. And if the judge or the jury are sitting there with Bayes’
formula in front of them eagerly waiting to fill in these numbers, then fine. But this is
a highly unlikely scenario. They are more likely not even to have heard about
Bayes. Concerns like these have been raised also by statisticians. For an instructive
example based on a real life case see [13].
Attempts have been made to translate LRs to verbal expressions ([19]; Gonzalez-
Rodriguez et al. 2001).

<table>
<thead>
<tr>
<th>LR Range</th>
<th>Support Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–10</td>
<td>Limited support</td>
</tr>
<tr>
<td>10–100</td>
<td>Moderate support</td>
</tr>
<tr>
<td>100–1,000</td>
<td>Strong support</td>
</tr>
<tr>
<td>Over 1,000</td>
<td>Very strong support</td>
</tr>
</tbody>
</table>

The suggestion makes some sense, especially since we may see the verbal scale
only as a suggested interpretation of what the reported LRs mean. There is, howev-
er, at present no general agreement on what is the best way to present LRs in court.

### 3.8.3 The UK Position Statement

The UK Position Statement [19] grew out of awareness that the traditional way
of presenting the evidence invites an interpretation known as the prosecutor’s fal-
lacy. If the results of a voice comparison is presented as “it is highly likely that the
samples come from the same speaker” but without considering the possibility that
other speech samples, not analysed, could have been equally similar, the presenta-
tion is unfairly biased against the defendant. As we have seen the likelihood ratio
approach takes care of this problem so why not just switch to this approach then?
The main reason mentioned by the proponents of the statement is the present lack
of suitable reference databases which makes numerical assessments impossible for
most parameters. The proposed solution is what is basically a verbal version of
the likelihood approach. The key concepts here are consistency and distinctiveness
(which roughly parallels similarity and typicality). Consistency is “a decision ... con-
cerning whether the known and questioned samples are compatible, or consist-
tent, with having been produced by the same speaker”. The outcome can be: Consis-
tent, not consistent or no decision. Distinctiveness is an evaluation of to what extent
the same features are shared by other people in the population. Distinctiveness can
be: Exceptionally distinctive, highly distinctive, distinctive, moderately distinctive
or not distinctive.

This approach is currently accepted by most analysts in the UK, but I have per-
sonal information that they too are considering moving towards a likelihood ratio
approach. Likelihood ratios are also discussed by police labs in Finland, Germany
and France (p.c.) but no decisions have been taken so far one way or the other.

### 3.9 Summary and Conclusions

It should hopefully be clear from what has been said in this chapter that speaker
comparison for forensic purposes is quite a lot more complex than what many
people may think. The complexity arises primarily from the fact that those traits in
speech which may be used for speaker comparison are also used for speech commu-
nication and are therefore not constant. An example that was mentioned is formant
frequencies. It is true that they are partly a function of the physical size and
shape of the vocal tract of an individual, they are also constantly changing due to the
changes in tongue shape and position, jaw opening, and lip opening used to produce
speech. This in stark contrast to DNA or fingerprints for example which remain
constant over the life time of an individual.

The task is now to find what, if anything is reasonably invariant in the speech of
a given individual and to find out if such traits show sufficient inter-speaker
variation to be useful for speaker comparison. What we know from the various at-
ttempts so far seems to suggest that no single trait possesses these properties but that
we must look for combinations of factors characterizing a given parameter. One
example mentioned [27] was the combination of mean, standard deviation, kur-
tosis, skew, modal F0 and the probability density of modal F0 for the fundamental
frequency parameter.

In our case work we have found it useful to make a distinction between voice
and speech. Voice similarity is not the same as speaker identity. It is quite possible
for a voice comparison to yield very high Likelihood Ratios in an automatic voice
comparisons analysis while the analyses of other factors indicate that the samples
come from different speakers. A somewhat drastic but illustrative example would be
a test of two voice samples that results in a LR of 1000. In good old television series
style you might be tempted to cry “We have a match”, but what if it turns out that
the two speakers speak different languages? In doing actual case work it is therefore
useful to keep mind that voice and speech are not the same thing.

Another important take home message is that we should abandon the idea of
identification. Reliable recognition in the sense of identification of a specific indi-
vidual from speech samples is not just difficult but virtually impossible. And this is
true, some would say, not just for speech based evidence but for all types of forensic
evidence including DNA. (For an excellent overview see [54]).

Access to relevant reference databases is presently a practical problem for most
types of analyses. Work is in progress in many places to remedy this shortage by
creating new databases, but for now and several years to come we will have to base
typicality estimates on less reliable data.

Finally some concern was mentioned over what it means to be a forensic expert
witness. With more automatic, user friendly methods available there is a risk that
should not be underestimated that unqualified analysts will rely completely on the
machines they operate without really knowing what goes on behind the screen. We
have no evidence that this is a real problem today but we should be aware of a de-
velopment in that direction as a potential threat to the scientific validity of forensic case work.

Appendix A: The Daubert Ruling

The US Supreme Court Daubert ruling in 1993 came in the aftermath of a controversy over the admissibility of scientific evidence in connection with the Daubert v. Merrill Dow Pharmaceuticals trial. Prior to the Daubert ruling most courts relied on the Frye ruling from 1923, which held that scientific evidence should be based on principles generally accepted within the relevant scientific community. We may say that the question of admissibility rested on the degree of consensus in the scientific community in a given case. Without going into detail we may observe that there are obvious pros and cons concerning the Frye ruling. On the one hand, consensus in the scientific community gives the evidence a high degree of scientific credibility. On the other hand we also know that scientific progress almost always begins with minority views.

The Daubert ruling, replacing the Frye ruling, presented a different view replacing it with a set of criteria meant as a general guide [47]:

1. Is the evidence based on a testable theory or technique;
2. Has the theory or technique been peer reviewed;
3. In the case of a particular technique does it have a known error rate and standards controlling the techniques operation;
4. Is the underlying science generally accepted?

The court cautioned, however, that the list should not be regarded as “a definitive checklist or test”.

At first sight this seems very reasonable, stressing scientific validity and the competence of the expert witness, but without requiring consensus in the scientific community. The Daubert ruling has not been without critics, however, and it is easy to see where the weak point lies. The final word concerning admissibility is shifted from the scientific community to the trial judge who in most cases is likely to possess little or no competence to evaluate scientific validity and reliability. Two Supreme Court Justices (Stevens and Rehnquist) voiced concern over such a development warning against judges taking on the role of “amateur scientists”.

This also seems to have become the case to a considerable extent. Many judges have invented their own admissibility standards that well exceed what is required in the scientific community, the result of which is that it has become very difficult for plaintiffs in cases involving health effects of toxic waste or side effects of drugs even to have their cases tried before a jury. The cases are often dismissed by a summary judgement and most often it seems in favour of the defendant (For a summary

---

4 Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).

see: [47]. The effect in criminal cases seems to have been less severe, at least until now.

If we can develop techniques and methods that comply with the four guiding criteria formulated by the Supreme Court we may, however, have a chance to bring back the power over admissibility to the scientific community, at least to some extent. It also has to be said that the question of admissibility does not exist in all countries and where it does it may differ from the rules that apply in the US, but the same kind of reasoning concerning validity and reliability reappears in the recommendations by the National Research Council and the UK Law Commission and is likely to be a topic of great significance for the future development of forensic phonetic expert testimonies.

Appendix B: The Likelihood Ratio Framework

I have referred to Likelihood Ratios (LR) on several occasions in the chapter. For some readers this is a very familiar concept, for others less so. The very brief account below should be sufficient to understand its place in the discussion for readers who are not already sufficiently familiar with the subject.

The Likelihood Ratio expresses the quotient between the strength of the evidence that the known and questioned samples have the same origin and the alternative hypothesis that they do not. Another way of expressing it is that it represents the quotient between a similarity score and a typicality score.

The technical way of expressing the LR is the following equation:

$$LR = \frac{p(E|H_{so})}{p(E|H_{do})}$$

In the equation, the numerator expresses the probability of the evidence (E) under the hypothesis that the known and questioned samples have the same origin. The denominator expresses the strength of the evidence under the alternative hypothesis, namely that the samples have different origins. If the two probabilities are the same, i.e. the LR is equal to 1, then obviously no definite conclusion may be drawn. If the quotient is greater than 1, then the evidence speaks in favour of the “same origin” hypothesis; the greater the number, the stronger the evidence. For LR's less than 1, the opposite obviously holds.

A few words also need to be said about the so called Bayesian framework. This approach is based on Bayes' theorem:

$$\frac{p(H_{so}|E)}{p(H_{do}|E)} = \frac{p(E|H_{so})}{p(E|H_{do})} \times \frac{p(H_{so})}{p(H_{do})}$$

In the equation the term to the left of the equality sign is referred to as posterior odds, the first term on the right hand side is the LR and the second term is the so
called prior odds. The posterior odds is the combined weight of the all the evidence. One way of verbally defining Prior odds would be to say that it represents the trier of fact’s belief about the two competing hypotheses prior to the “new” (hence ‘prior’) evidence the strength of which is expressed by LR. The posterior odds will as a consequence have to me modified by a factor expressed by the LR. It goes without saying that the prior odds should be based on some other type of solid evidence and not just beliefs in general. The term ‘prior’ should, however, not be interpreted as any requirement of an ordering in time of various types of evidence. It is just a way of expressing how beliefs may be modified as more and more evidence is presented. This said, it should be obvious that this is something for the judge, jury, lawyers to think about but not the expert witness presenting the evidence. There may for example be fingerprints, DNA, eyewitness reports etc. which the court needs to consider, but which should be completely disregarded by the expert when performing the forensic speech analysis. It therefore makes sense to call the approach used by the expert witness the Likelihood Ratio framework rather than the Bayesian framework. This is also the suggestion by several forensic speech scientists [40]. There is, however, one type of prior odds that the forensic speech science expert may have to consider and that has to do with the choice of a relevant reference database. There is generally complete agreement about the sex of the speaker, often also about approximate age and dialect. In such cases the speech analyst may limit the reference database accordingly.

References

47. Project on Scientific Knowledge and Public Policy (2003) Daubert—the most influential Supreme Court ruling you’ve never heard of