Language and Linguistics Compass 2/4 (2008): 671–711, 10.1111/j.1749-818x.2008.00066.x

Forensic Phonetics

Michael Jessen*
Bundeskriminalamt BKA

Abstract

An overview of forensic phonetics is presented, focusing on speaker identification as its core task. Speaker profiling/speaker classification is applied when the offender has been recorded, but no suspect has been found. Auditory speaker identification by victims and witnesses becomes relevant when no speech recording of the offender is available. It can take the form of familiar-speaker identification or unfamiliar-speaker identification, and in the latter case a voice line-up/voice parade can be carried out. When recordings of both the offender and a suspect are available, a voice comparison is done by an expert in forensic speech analysis. Current issues and domains in voice comparison analysis include the Bayesian approach to forensic reasoning and the Likelihood Ratio, the use of formant frequency measurements, non-analytic perception and Exemplar Theory, forensic automatic speaker identification, and the interaction between different methods.

1 Introduction

Forensic phonetics is the application of the knowledge, theories and methods of general phonetics to practical tasks that arise out of a context of police work or the presentation of evidence in court, as well as the development of new, specifically forensic-phonetic, knowledge, theories and methods. The most central aspect of forensic phonetics is speaker identification – also referred to as speaker recognition (in this paper these two terms will be assumed to be synonyms, although they may have specific meanings in particular legal systems or scientific contexts). Another aspect is the task of analyzing the linguistic content of passages of speech in which intelligibility is strongly reduced for technical or behavioral reasons. This task is often referred to as disputed utterance analysis/examination. Some tasks in forensic phonetics require strong interdisciplinary ties to speech technology/engineering or general acoustics. That is the case with audio enhancement, where various forms of filtering techniques are used with the goal of improving the intelligibility of a low-quality recording. It is also the case with audio authentication, where doubts have to be addressed that a recording might have been manipulated by deleting, inserting or changing passages from the original recording. More recently, speech technology has also become important in automatic speaker identification. Since, in a forensic

laboratory, phoneticians are often the only staff members who have the equipment and experience to analyze audio signals, they are sometimes asked to analyze non-speech acoustic events, such as gun shots, sounds from the cockpit of an aircraft, or bird song. This requires an even stronger interdisciplinary awareness than audio enhancement and authentication, unless it is a task that can be solved by everyday experience with common sounds. Forensic-phonetic tasks other than speaker identification will not be addressed in this paper. More information on audio enhancement, authentication and the analysis of non-speech events is provided in Hollien (1990), Broeders (2001, 2004) and Bijhold et al. (2007). Instructive examples of disputed utterance analysis are presented in French (1990) and French and Harrison (2006).

The term 'forensic phonetics' has received official status at least since the foundation of the 'International Association for Forensic Phonetics' (IAFP) in 1991 (see their website at http://www.iafpa.net). Many of the contributions from its annual meetings as well as some conference reports of these meetings are published in the *International Journal of Speech, Language and the Law* (http://www.equinoxjournals.com/ojs/index.php/IJSLL). As the former title of this journal (*Forensic Linguistics*) indicates, this journal is a forum for both forensic phoneticians and forensic linguists, who are organized in the 'International Association of Forensic Linguists' (cf. Shuy 2007). In 2004, the IAFP added 'acoustics' to its name and is now called 'International Association for Forensic Phonetics and Acoustics' (IAFPA). This change was implemented in order to explicitly honor the contributions by specialists in acoustic signal analysis and speech technology to the field. An alternative to the term 'Forensic Phonetics and Acoustics' frequently encountered is 'Forensic Speech and Audio Analysis'. Although in practice the coverage of these two terms is very much the same, the latter term does not contain the word phonetics and is therefore more neutral with respect to the expected education of its practitioners and is preferred by many governmental organizations that offer services in that field. Although it is the view of the author that a department of forensic speech and audio analysis should have at least some phoneticians or linguists among its staff, this is not the attitude in every country, where instead forensic speaker identification and related tasks might be regarded as technological domains that can be handled by engineers or by police officers who are trained in the use of certain hardware and software. Less is known about forensic speech analysis that is performed by private individuals or companies, but at least for the United Kingdom and Germany it is the case that most of the private experts are phoneticians. The first academic program directly relevant to the subject matter has been established under the name 'MSc Programme in Forensic Speech Science' at the University of York, UK, and has started in Fall 2007 (http://www.york.ac.uk/depts/lang/postgrad/forensic.htm).

This paper will focus on forensic speaker identification. Some monographs and textbooks on this topic are available, most noticeably Nolan (1983), Hollien (2002), Rose (2002), as well as (for the German-speaking reader)

Künzel (1987). A more general scope, including other aspects of forensic phonetics, is presented in Baldwin and French (1990) and Hollien (1990). Recommendable previous review articles (in English) on forensic phonetics include Nolan (1994, 1997, 2001), French (1994), Künzel (1995, 2004), Meuwly (2000), Broeders (2001, 2004, 2006), Gfroerer (2003), and French and Harrison (2006).

2 Types of Forensic Speaker Identification

Speaker identification, which has been characterized as the most central aspect of forensic phonetics (and acoustics), can be further subdivided into voice comparison, voice profiling, and the analysis of speaker identification by victims and witnesses.

2.1. VOICE COMPARISON

In voice comparisons a speech recording exists of an unknown speaker who can be associated with a crime.[1] For example, this can be a kidnapper making ransom demands over the telephone, a drug dealer arranging illegal transactions over a (tapped) telephone line, or a stalker. Furthermore, somebody is suspected to be identical with the unknown person and a speech recording of that suspect is available or can be made. Depending on the legal system, tapped telephone conversations or recorded police interviews can be used as evidence if the suspect is uncooperative, or if additional or more varied speech material is needed. When recorded material from both the unknown speaker and the suspect (or more from each category) is available, a voice comparison (or a set of different voice comparisons) can be carried out. The recordings are compared with respect to a wide variety of speech features and with one or more different methods. After the analysis is completed, a conclusion is reached that is relevant to the question whether the speech samples that are compared originate from the same or from different individuals. Conventions and conceptual frameworks as to how these conclusions are expressed can differ internationally (see French and Harrison 2007 for a position statement expressing the current procedures in the UK).[2]

Voice comparisons can be requested as part of a police investigation or privately without going to court, but most commonly, voice comparisons result in written, scientifically motivated reports that are used as evidence in court and that (depending on the law system, country or local regulations) have to be explained and defended orally in court by the expert responsible for the report. Voice comparisons will be addressed in more detail in Section 3.

2.2. VOICE PROFILING AND SPEAKER CLASSIFICATION

In some situations a recording of an unknown speaker has been made, but no suspect exists so far and hence there is nothing to compare the unknown

sample with. This is often the case in the early stages of a police investigation. In a situation like this, the forensic speech expert is requested to deliver a voice profile. Voice profiles contain information that can help police officers or people from the general public, who are laypersons with respect to linguistic/phonetic analysis, to narrow down the range of possible suspects or even to find a suspect. Depending on how long, qualitatively advanced and informative the sample from the unknown speaker is, a voice profile can contain more or less precise information about the region where the speaker has been brought up, the age, sex, level of education and social background, native language (in case of foreign-accented speech), and medical conditions that affect speech. The speech sample might also contain further aspects that are 'unusual' from a layperson's perspective. For example, a speaker might have a very high-pitched voice or might speak very rapidly. There is no point in using very technical linguistic or phonetic terminology or reporting observations (or even acoustic measurements!) which non-specialists cannot follow. In some cases it can be decided that the unknown sample should be presented directly to the public in the form of TV programs, radio, or the Internet. A more detailed overview of the range of information provided in voice profiling is found in Jessen (2007b).

It is useful to keep the terms *voice profiling* and *speaker classification* separate. Voice profiling is defined in the practical terms just stated, i.e. providing information about a speaker that is important for finding a suspect. Speaker classification is defined in more theoretical terms. It is the process of inferring from the linguistic and phonetic patterns of a speech sample the 'class' or 'category' to which a speaker belongs, including age, sex, social group, and region (see Müller 2007 for the full range of that term). The two terms voice profiling and speaker classification overlap to a large extent but not entirely. Voice profiling is primarily a speaker classification task, but it also involves the identification of 'unusual' speaker characteristics like high pitch, which might help to find a suspect but which might be of no help in assigning the speaker to a certain category. And speaker classification is performed not just for the purpose of voice profiling but also for the purpose of voice comparisons or the construction of voice line-ups (to be addressed in Section 2.3.2).

Speaker classification is very demanding on the range of knowledge required of the forensic expert. It is near to impossible to be a specialist in all the dialects of the target language, know all about its sociolinguistics, about speech pathological conditions, second language phonetics, and so forth. The important thing is to act responsibly. If a gap of knowledge in an aspect of speaker classification cannot be filled by reading up on the research literature, by consulting databases, or by carrying out a study that is tailored to the needs of the case, it is time (at the latest) to seek the assistance of another specialist. If speaker classification is performed for the purpose of voice comparison and its presentation in court, the demands

on the scientific motivation of the classification are probably stronger, whereas for the purpose of voice profiling in an ongoing police investigation, all that counts is the usefulness of the information, and often the speed at which it can be delivered.

Voice profiling is not the same as psychological profiling, and in fact the IAFPA recommends in its Code of Practice (www.iafpa.net/code.htm) that its members should not attempt to do psychological profiles.[3] Psychological profiling has become known by the work of the FBI Behavioral Science Unit and is presented in publications such as Douglas and Olshaker (1995). Based on the available (mainly crime scene) evidence, the profiler makes statements about the offender's psychological type or pathology, as well as sex, age, ethnic origin, body size, socioeconomic status, and might go as far as to infer the type of clothing, type of car, kind of job and job-related behavior, type and general location of residence, potential presence of speech defects, etc. (The term *psychological profiling* is only partially appropriate, since some of these classifications are biological or sociological rather than psychological/behavioral.) One gets the impression that many of these statements are based on the skill and experience of the profiler – including her or his ability to holistically capture the uniqueness of the case and to empathize with the criminal – but that only a subset of these statements can be scientifically backed up with background statistics that document correlations between concrete crime scene evidence and the type of offender classifications mentioned above. This is neither the place, nor is it the intention to criticise these profiling methods, and they can provide valuable information that result in catching the criminal. Furthermore, profiling methods are not the same in every country and differ, for example, between the USA and Germany, where less emphasis is given to the psychological aspect than to other aspects. Although voice profiling and psychological profiling can overlap in the offender classifications that are provided (e.g. sex and age) and in the urgency of the situation, there are differences. For example, voice profiling is limited to recorded speech evidence (forensic speech experts rarely visit crime scenes) and is performed by trained linguists or phoneticians who know which inferences can be drawn from speech evidence that are backed up scientifically. If these inferences include psychological information such as the presence of certain psychiatric disorders, the appropriate research literature should be consulted (work like Darby 1981) and cooperation with a qualified psychologist, psychiatrist or neurologist should be sought.

Voice profiling was prominent in the 'Yorkshire Ripper' case (Ellis 1994; French et al. 2006). Between 1975 and 1980, 13 women were murdered in Leeds, Bradford, Huddersfield and Manchester. Between 1978 and 1979, several letters and eventually also a tape were sent to newspapers and the police, in which responsibility for the murders was claimed and further strikes were announced. The police contacted Stanley Ellis, a phonetician and dialectologist, to analyse the tape and to provide a conclusion about

the location of the caller. Based on many years of fieldwork research, and by consulting various dialect atlases and other published sources, Ellis was able to narrow down the possibilities to Sunderland (a town close to Newcastle) and surroundings as the most likely location where the caller on the tape has been brought up. He was able to further narrow down the location by visiting Sunderland, playing the tape to the local population and making speech recordings at various locations in the area. Ellis finally came to the conclusion that the caller had been brought up in the Southwick and Castletown areas close to Sunderland, which is what he reported to the police. As it turned out later, the man did in fact come from a place west of Sunderland that was only a about mile away from the location identified by Ellis. Given such an accurate profile, and given the fact that the voice on the tape (which also had other distinctive aspects beyond the accent) was presented in the media during that time, there was a strong possibility that the caller could have been found. However – in what turned out to be a grave mistake – the police, in searching for suspects, focused on only those individuals who had a matching accent and did not have an alibi. When, after some time, no suspect was found and the murders continued, Ellis and his colleague Jack Windsor Lewis, who analysed the letters, became suspicious that the tape and the letters might be a hoax. They communicated their suspicions to the police, pointing out that the voice is very distinctive and that perhaps the caller had already been interviewed by the police, but was dismissed because he had an alibi. The police ignored these and other indications that the man who spoke on the tape (and wrote the letters) was not the murderer, but a hoaxer. It took another 2 years until the real murderer was found in 1981, and eventually sentenced to life imprisonment. The identity of the hoaxer remained unknown for many years until, in 2005, he was found through DNA analysis on the hoax letters. This initiated a series of further forensic-phonetic and -linguistic analyses, which are documented in French et al. (2006). The hoaxer was finally convicted for the offence of attempting to pervert the course of justice and sentenced to 8 years imprisonment. More details on the Yorkshire Ripper case (from the time before the hoaxer was found) are published in Bilton (2003).

2.3. SPEAKER IDENTIFICATION BY VICTIMS AND WITNESSES

In all the scenarios mentioned so far, audio recordings of the unknown speaker (the offender) were available. There are situations, however, where no recordings were made, and all that is available as evidence is the perception of the offender's voice by a witness, who might at the same time be the victim of the crime. These can be situations of rape or other attacks on the victim. Two kinds of situation should be distinguished – one in which the offender is familiar to the witness and one in which he is not. Generally, speaker identification has been shown to be much more accurate in the former than the latter kind of situation.

2.3.1 Familiar-speaker identification

In the first kind of situation, the offender is known to the witness from the time prior to the crime, and, based on this knowledge, the witness is able to identify the offender, often by name. This voice identification can be treated in court as a regular witness statement – with all the merits and limitations that witness statements entail (Sporer et al. 1996). In some of these cases, however, the reliability of such a voice identification statement can be questioned. Factors that cast doubts on the reliability of earwitness statements can be divided into those involving the channel, the speaker (offender), or the listener (witness).

Channel limitations are in place if, for example, the distance between speaker and listener is large, so that the amplitude of the speech is relatively low compared to environmental sound and other air pressure variations. Other instances of channel limitations occur if there are loud sound sources that mask part of the speech, or if there is an obstacle between speaker and listener that filters out some of the resonances important for speech and voice recognition (including the case that the offender wore a mask, a motorcycle helmet or the like). Voice identification has also been shown to be somewhat less reliable in telephone speech, where the speech signal is bandpass filtered, than in direct contact between speaker and listener (see http://www.mml.cam.ac.uk/ling/research/voice_similarity.html for a current project on that topic). Difficulties arise particularly in situations of channel *mismatch*, where, for example, the known voice was only heard under good channel conditions whereas the offender's voice was only heard under poor channel conditions (this point is even more relevant with the identification of voices that were not previously known). The mismatch problem also extends to different speech styles (e.g. shouted vs. non-shouted speech), as will be addressed now.

The speaker factor involves cases where the offender spoke in unusual ways or for only a short period of time. The most extreme kind of unusual speaking can be voice disguise, such as the use of falsetto voice, which is chosen deliberately for the purpose (though not necessarily the effect) of making voice identification impossible. But there can be other more natural kinds of speech production that can make voice identification difficult. One such kind of difficult speech production mode is shouting (others include emotionally marked or stressed speech). Blatchford and Foulkes (2006) report a case where two women were killed in a gang-related drive-by shooting. Another person was shot at, but was able to escape. The offenders followed him while shouting 'get him'. Based on that utterance, the victim was able to identify and name the offender, whom he knew because they had served time together in a young offenders' institution. Blatchford and Foulkes carried out an experiment where utterances of different length were shouted – one of them was the two-syllable utterance 'get him' from the case, the other one was of the length of a sentence containing 12 syllables. When presented to lay listeners, to whom the

producers of the shouted utterances were familiar in a close social network, they found that correct identification dropped from 81% to 52% from the longer to the shorter utterance, showing that shouting and short duration interact. Speaker identification in shouted speech was found to be possible, in principle, but strongly depended on both the speaker and the listener. This brings us to another speaker factor: not every voice is equally distinctive; some speakers are identified better than others. Rose and Duncan (1995) and Foulkes and Barron (2000), who both show this speaker effect in familiar-speaker identification, highlight different possible explanations. According to Rose and Duncan (1995), speakers might be hard to identify when they exhibit large amounts of intraspeaker variation (cf. Section 3.1.3), some of which intersects with the auditory space of other speakers. Foulkes and Barron (2000) were able to correlate low identification difficulty with rare phonetic features, most strikingly from the domain of dialectal variation and fundamental frequency.

Turning to the listener factor, one aspect that came out of the study by Blatchford and Foulkes (2006), but was also found in Foulkes and Barron (2000) and other studies, is that not every listener is equally well able to identify voices. They conclude from this finding that earwitnesses should be tested for their ability to identify voices. Perceptual abilities may also be located on less cognitive and more peripheral levels. This is the case when the listener exhibits hearing problems that are relevant for voice identification, which could be determined by audiological testing. Some other listener factors, such as vocal memory, are more relevant for the identification of unfamiliar voices, to which we turn now (and of course, other factors discussed here in connection with familiar-speaker identification are also relevant for unfamiliar-speaker identification). More detailed overviews of factors that influence the performance of auditory speaker identification by lay listeners include Hammersley and Read (1996), Hollien (2002) and Rose (2002).

2.3.2 Unfamiliar-speaker identification and voice line-ups

The second kind of situation relevant to speaker identification by victims and witnesses occurs if the offender was not known to the witness from before the crime. If in such a situation a suspect can be found, a procedure can be applied that is analogous to a visual line-up (also called eyewitness parade), where the suspect is presented, along with a set of other persons who are unrelated to the crime (foils). Only here the line-up is not visual, but auditory, and the presentations of suspect and foils are not submitted simultaneously but rather sequentially. An auditory line-up of different voices is called a 'voice line-up' or 'voice parade'. In reality it turns out that voice line-ups are much more difficult to plan and carry out than visual line-ups. Ormerod (2001), in a paper directed to lawyers, points out that while selecting foils for visual line-ups can be done by experienced police officers, selecting foils for voice line-ups – and in fact carrying out the

entire line-up – should be done by experts (phoneticians and linguists). Guidelines as to how a voice line-up should be planned and carried out have been proposed (e.g. Hammersley and Read 1996; Ormerod 2001). The most specific guideline has been published by Broeders and van Amelsvoort (1999), and this is the one that has been accepted by many forensic laboratories in Europe, which are organized in the European Network of Forensic Science Institutes (ENFSI). Only a few aspects from this and other guidelines can be addressed here.

The first thing to keep in mind is that the memory for voice identities decays rapidly. This means that the witness who has heard the voice of the offender for the first time must be interviewed as soon as possible after the criminal act, in order to elicit any voice characteristics of the offender that the victim can remember. The results from this interview can help find a suspect and select foil speakers with similar voice characteristics. Again because of the decay in vocal memory, the line-up must be carried out as soon as possible.

The first systematic research into the forensically relevant topic of vocal memory was motivated by a well-known case in criminal history: the kidnapping and murder of the baby of the aviator Charles Lindbergh in 1932 (see Fisher 2000). Lindbergh witnessed the voice of the kidnapper at a graveyard in the Bronx were the transfer of ransom money was arranged. He went to that location together with Dr. Condon, a friend of Lindbergh, who had arranged the meeting. The kidnapper uttered: 'Here doctor, over here, over here' (Fisher 2000, p. 81). Two and a half years later, the voice of the suspect Bruno Richard Hauptmann was presented to Lindbergh. After that presentation Lindbergh said: 'That is the voice I heard that night' (p. 249). Hauptmann was eventually convicted of murder and executed. There is still some debate in the literature whether Hauptmann was guilty or innocent (Fisher 1999 for the debates on this issue), but it is clear that there was more substantial incriminating evidence against Hauptmann than the voice identification statement.[4]

The psychologist McGehee (1937) reported an experiment where a text was read by a person and presented to listeners who were on the other side of the room, separated by a screen. The same text was read after different time intervals by the previous speaker and four other speakers (in further variations of the experiment that number of speakers was varied). The listeners had to check on an answer sheet which of the five voices they thought they had heard before. McGehee found that after 1 day between first and second presentation, 83% of the listeners were able to recognize the correct voice. That recognition level was essentially sustained until it dropped to 69% between the first and the second week and eventually ended at 13% after 5 months. According to these results, Lindbergh's earwitness testimony must be treated with caution. Similar results were obtained in later studies (for overview Hammersley and Read 1996 and other sources mentioned above). What became more apparent later was that

vocal memory does not decay according to a simple rule, but that a number of interacting factors are relevant. One important finding was that if the exposure to the unknown voice is short, vocal memory decays more rapidly than with longer exposure. This duration factor would cast further doubts on the ability of Lindbergh to have remembered the voice of the kidnapper better than chance, since his exposure to that voice lasted only a few words. Another unfavorable factor was that Lindbergh heard the voice from a long (200 feet) distance (Fisher 2000, p. 250). It should be added that if Lindbergh's earwitness statement was perhaps not much better than chance level, it was not his fault but the fault of those who, without scientific backup, assumed that voice identification under these poor conditions is possible and allowed it as evidence in court. It has subsequently been found that a high confidence rating in naïve voice identification does not imply that its accuracy must be high as well (Perfect et al. 2002). Finally, in addition to putting too much faith in voice memory, the suspect's voice in the Lindbergh's case was presented alone, which is likely to predispose a witness to say 'yes', as opposed to a line-up, which tests the witness's memory.

Knowing that a voice line-up has to be carried out as soon after the crime as possible, how is it planned and carried out? Much of the effort goes into selecting suitable foils. Usually about five or more foils in addition to the suspect should be included in a line-up. The general idea is that the foils have the same speaker classification as the suspect (cf. Section 2.2). It would be unacceptable, for example, if the suspect spoke with a lower-class accent whereas all the foils sounded more upper-class like. The suspect must also not differ from the foils in terms of other unusual speech characteristics. If suitable foils are found, recordings are made of all speakers. If the suspect is unwilling to provide a voice sample, a recording from a police interview can be used. Then the foils have to be recorded in ways that are technically and stylistically similar to that police interview recording, or authentic police interviews with other speakers are used. Whether the selection of the foils and the recording under comparative conditions was successful has to be tested on listeners who have no knowledge of the case. Various testing formats are possible. If it turns out that one of the speakers stands out from the rest in some way, foil selection or recording has to be done again, until the line-up is unbiased and every voice has an equal chance of being selected (it is therefore advised to begin with a larger number of foils than eventually used, so that less well matched ones can be eliminated from the line-up that is eventually used with the witness). After all this preparation has been performed, the actual line-up can begin. Every voice is played in sequence by someone who does not know who the suspect is. After each voice, the witness is asked if that is the voice from the crime. It is possible to play a voice several times but not to go back to previously played voices. The point is not to promote the selection of the voice that best matches the offender's voice with the set of speakers of the line-up (closed-set identification), but to have the witness decide for each voice

whether that was the voice heard during the crime (open-set verification). To implement that principle more rigorously, it is also possible to run a line-up session where the suspect is not included. The entire line-up should be recorded on video or observed by means of a two-way mirror. One reason for such documentation or scrutiny is that the witness might react non-verbally in ways that are different for the suspect than the foils (perhaps with a strong emotional reaction to the suspect voice). Another reason is to confirm that the assistant who presented the voice samples did not in any way cue the witness toward or away from the suspect voice. If the methodology of the line-up is challenged in court and the criticism is valid, the line-up cannot be carried out a second time because it might be possible then that the witness recalls the voice from the first line-up, not the one from the day of the crime.[5]

Given the substantial effort it takes to plan and carry out a voice line-up – and for various legal reasons – careful consideration should be given to whether a voice line-up is realistic at all. If, for example, much time has passed since the crime, if exposure to the voice was extremely short, or if any of the adverse conditions apply that were mentioned above in connection with the identification of familiar voices, the application of a voice line-up might be ill-advised, both from the prosecution and the defense perspective.

3 Voice Comparison

3.1 INTRODUCTION

3.1.1. Experts

A voice comparison can be performed as soon as recorded speech from both the person who is associated with the crime – referred to as the unknown or the anonymous speaker – and from a suspect is available. Voice comparisons are carried out by experts in forensic speech analysis. These are often phoneticians, who after their academic training in linguistics, psychology, speech and hearing sciences, etc. (or, of course, phonetics, in those countries where phonetics is a separate academic discipline), have obtained additional training in forensic science and perhaps in technological domains such as speech enhancement. They usually have practical experience in working with authentic forensic speech material, as well as in the writing of expert witness reports and their presentation in court (cf. the chapter 'the professionals' in Hollien 2002). Sometimes the experts are engineers or computer scientists who use automatic speaker identification. Due to the involvement of experts, voice comparison activities are often referred to as 'speaker identification by expert' or 'technical speaker identification' as opposed to speaker identification by victims or witnesses or 'naïve speaker identification' (see Section 2.3). Of course, suitable expert involvement is required not only in voice comparisons, but also in voice

profiling and in the elicitation and evaluation of evidence from speaker identification by victims and witnesses.

3.1.2. The Bayesian approach

In a voice comparison, two aspects have to be kept in mind. The first aspect is how similar or how different the two voices are with respect to the phonetic dimensions on which they are compared. This is the *similarity* aspect in speaker identification. The more similar the voices are the more likely it is – everything else being equal – that they originate from the same speaker. It is a common misunderstanding, both in speaker identification and in other branches of forensic science, that as soon as a strong degree of similarity has been established, a match has occurred and the two probes (speech samples) that are compared have the same source. As pointed out by Rose (2006a), how many times have we heard the phrase 'it's a match' in 'CSI' and other TV shows that feature forensic detection? This notion of match vs. no match implies a categorical decision between identity and non-identity, whereas in current forensic science, the practice advocated is to express conclusions with probabilities (verbal or quantitative) rather than absolute decisions (with the exception of a few countries). In order to prove that a match has occurred in the categorical sense of this notion, it would have to be shown that the matching pattern does not occur a second time between different probes. It has been assumed for a long time that this is the type of situation that occurs in fingerprint analysis. However, more recently problems have emerged with this assumption, so that even fingerprint comparisons should be (and often are) expressed in probabilistic terms. In DNA analysis probabilistic thinking was dominant all along, but it is also a much younger discipline.

This brings us to the second aspect that needs to be kept in mind in voice comparisons: *typicality*. The phonetic characteristics which are analyzed in the speech of the unknown speaker and the suspect can be relatively typical patterns in the entire population of speakers or they can be relatively rare. Evidence for the identity of two speakers is stronger – everything else being equal – when typicality is low than when it is high. An illustration of a situation with low and with high typicality will be provided once we get to the analysis of fundamental frequency in Section 3.2.2.

The conceptual and mathematical framework behind this balancing of similarity and typicality is known as the *Bayesian approach*.[6] A core concept in this approach is the Likelihood Ratio (LR). It can be expressed as follows (Rose 2002, 58):

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_D)}$$

The expression in the numerator refers to the probability of obtaining the given speech evidence E if the prosecution hypothesis is correct (the

two probes having the same origin). The denominator expresses the probability of obtaining the same evidence if the defense hypothesis is correct (the two probes having a different origin). If the LR is larger than one there is relatively more evidence that the probes have the same source and with values smaller than one there is relatively more evidence that the source of the two probes was different.[7] In its numerator the LR captures the similarity aspect of a comparison: if similarity is high, the likelihood that the source of two probes is the same is relatively high as well; if similarity is low, the likelihood that the source of two probes is the same is also relatively low. In its denominator the LR captures the typicality aspect: if typicality is high, the likelihood that some other entity (here: speaker) is the source of the unknown probe is relatively high as well, and if typicality is low the likelihood that some other than the suspected entity is involved is also relatively low.

The Bayesian approach has gained wide acceptance in the forensic sciences since the 1990s and has played an important role in the development of DNA analysis and the presentation of the DNA results in court (see Robertson and Vignaux 1995 for an introduction). In forensic speech analysis, the Bayesian approach has been used first and most intensively in automatic speaker identification (Meuwly 2000 for overview). But soon after, it was also discussed in relation to the application of acoustic- and auditory-phonetic methods (see Nolan 2001; Rose 2002). These and other methods will be explained in Section 3.2.

In order to be able to quantify the typicality aspect in an LR it is necessary to create or have access to *population statistics* about the speech properties that are used in a voice comparison. These do exist in automatic speaker identification (see Section 3.2.4), but only rarely so with phonetic properties (see Section 3.2.2). Even in areas where no such population statistics exist, and therefore no quantification is possible, the Bayesian approach should be used as a conceptual framework that provides the logical backbone of voice comparison analysis. The Bayesian approach in forensic phonetics is most comprehensively explained in Rose (2002, 2005).

3.1.3. Limited quality/quantity and the mismatch problem

Similarity, typicality and the importance of population statistics are necessary ingredients of any form of voice comparison activities, including those that are performed for commercial reasons or for pure research purposes. These are contexts where the quality and quantity of the speech material usually leaves nothing to be desired. In forensic casework, however, there are often restrictions on the quality and quantity of the material that have an influence on the methods that can be used and the conclusions that can be drawn. One frequent problem with forensic recordings is that intraspeaker variation is higher than in the controlled settings of studio or laboratory recordings. Intraspeaker variation (also referred to as within-speaker variation)

refers to the fact that the same speaker can speak differently (with respect to the parameters that are analyzed in a voice comparison) on different occasions or in different contexts. Intraspeaker variation occurs, for example, if the pitch of the voice is higher when a person speaks loudly than when the same person speaks with normal loudness. In Bayesian terms, if intraspeaker variation is high, similarity, and hence the numerator of the LR, is low despite the fact that the same speaker is involved. This could lead to false rejection. This form of behavioral intraspeaker variation is a forensic problem that DNA analysis and fingerprinting do not have to contend with. It is one of the complications that make forensic speaker identification (or any behavioral analysis) a very difficult task indeed.

Intraspeaker variation comes in many forms. Some of them are necessary ingredients of the linguistic system. For example, the pitch of the voice and its main acoustic correlate fundamental frequency, is not constant for a speaker, but varies according to the intonation patterns or tonal patterns (for tone languages), and the formant frequencies of a vowel are influenced by the neighboring sounds due to coarticulation. This means that in a voice comparison, either speech portions have to be selected that are linguistically equivalent or enough material has to be processed so that it can be safely assumed that the entire composition of linguistic influences in two speech samples is approximately the same. More difficult to handle, and often more substantial in size, are sources of intraspeaker variation whose status can be classified as paralinguistic or stylistic. The influence of vocal effort (loud vs. soft speaking) has already been mentioned. Other such sources of intraspeaker variation include stress and emotion, reading vs. speaking spontaneously, being drunk vs. being sober, having a cold vs. being healthy. Often there is a mismatch between the speech sample of the anonymous speaker and the suspect with respect to these paralinguistic or stylistic influences. Sometimes the anonymous person speaks loudly and in an emotionally agitated manner whereas the suspect speaks more softly and sounds as though he is bored. These types of mismatch problems can be addressed either by selecting portions in the recordings where the mismatching influence is minimal or by applying knowledge about the phonetic influences of these factors in an effort to compensate (technically or conceptually) for the mismatching influence.

The most extreme forms of intraspeaker variation are those that are introduced deliberately by the speaker in the form of *voice disguise*. A huge mismatch in phonetic expression can arise if the anonymous person speaks with voice disguise, in order to conceal his identity, whereas the suspect speaks without voice disguise. Even the opposite situation is possible, but then the voice disguise is usually more subtle (which does not make the analysis any easier). When disregarding the regrettable circumstances under which it occurs in a forensic context, voice disguise can be an interesting topic in its own right, since it highlights the flexibility of the speech production apparatus and the inventiveness of language users in an effort

to conceal their identity. Many forms of voice disguise focus on pitch or voice quality as determined by the vocal folds. That was the case in a well-known German case in the late 1980s and early 1990s, where a criminal who blackmailed major German department stores – enforced by bomb detonations where fortunately nobody got hurt – used the alias *Onkel Dagobert* (after the German equivalent of the Disney character 'Uncle Scrooge') and conducted ransom negotiations in a high-pitched falsetto voice. Other voice disguise strategies include the change of speech tempo or the change of supralaryngeal characteristics through nose pinching, putting objects in the mouth, etc. Voice disguise can also focus on more purely linguistic aspects, such as imitating a foreign accent or talking in a foreign language altogether. Usually, voice comparisons can be conducted despite the presence of voice disguise because the speaker is rarely able to focus on more than one or two voice disguise strategies. Furthermore, because of physical, emotional or cognitive exhaustion the speaker often cannot maintain voice disguise for all of the conversation. Voice disguise can also be reduced or terminated by the offender if it turns out that it reduces intelligibility to an extent that his communicative goals cannot be achieved.

The possible limitations and problems with forensic speech material that have been discussed so far occur with respect to behavioral aspects (speech production). But there are also technical limitations and problems. One technical limitation that occurs in almost every forensic case is that at least one of the samples that have to be compared is from recorded telephone conversations. In telephone communication – whether landline or mobile phone – the speech signal is bandpass limited from about 200 Hz to about 3500 Hz. This means that any speaker specifics that occur outside this range cannot be used in forensic phonetics. Fundamental frequency, which in male adult speakers ranges from about 80 to 170 Hz under normal circumstances (hence below 200 Hz), is not a problem because fundamental frequency can be calculated from the distance between the harmonics of the fundamental, which occur abundantly within the telephone passband. Problematic are vowel formants above the third formant or fricative spectra for sounds like [s]. While speaker specifics involving these kinds of information are interesting from a research point of view, they are usually not applicable forensically. (This could change if telephone companies on a broad basis decided to extend the frequency range of telephone communication.) The frequency range that a forensic phonetician has to work with can be further limited if, due to poor transmissions or recoding equipment, the upper frequency boundary is further lowered or the lower boundary further increased beyond what is expected in telephony. Aside from frequency limitations, other types of difficulties occur with speech that is distorted (e.g. because the recording levels were set too high or because of poor mobile phone connections), because it contains strong echo or because it is mixed with background noise or overlapping voices. Still another problem is mismatches in the recording and playback speed

of analog recordings or mismatching sampling frequencies in analog-digital and digital-analog conversion.

Another type of technical problem is reduced quantity. It is a common misunderstanding in the general public that one can identify a voice from a very short utterance, perhaps just a single vowel. Those with a background in linguistics or phonetics can imagine why short duration is a problem. The shorter the speech recording, the less representative the speech patterns in the recording are in relation to the whole range of the speech patterns of the speaker. And, if it is too short, it might not contain the type of sounds that are rich in speaker specifics. Also, dialectal analysis can be severely limited if lexical items with 'diagnostic value' are not present in the material, such as words that contain /r/ in the syllable coda. There is no fixed limit of duration below which a voice comparison cannot be carried out, but at least something like eight seconds of speech from the anonymous speaker and at least about double that time for the suspect is recommended. This does not exclude the possibility that even a two-second sample contains very distinct information.

The limitations and problems with forensic speech material that were summarized here have two consequences. First, since these types of limitations and problems are to be expected, research and development in forensic phonetics should focus on those phonetic parameters that are maximally robust against them, or they should find a way to deal with them (Nolan 1983, 11–14; Rose 2002, 51f). For example, since mobile phone communication is to be expected in forensic casework, research is necessary on how this technology influences voice comparison methods, such as formant frequency measurements. This research has shown that the second and third formant are largely unaffected, whereas the first formant is raised relative to high-fidelity speech (Byrne and Foulkes 2004). Second, for each individual forensic case, a careful assessment of the limitations and problems of the material at hand has to be made. Based on that assessment, it has to be decided, first if the case can be processed at all, and if it can, what phonetic parameters can be analyzed and how strong a conclusion of identity or non-identity can be.

3.2. DIFFERENT METHODS IN VOICE COMPARISONS

Depending on the expert or the forensic lab, different methods can be used in voice comparisons. One way of classifying these different methods is shown in Table 1.

The analytical approach, which is the approach most commonly applied in phonetics, is to divide entire speech events into constituent parts, such as sounds, sound features or prosodic properties, and then to investigate some of these components separately or with respect to the way in which they influence each other. In an auditory-perceptual approach the components of speech can be transcribed phonetically by means of the International

Table 1. Classification of different methods in forensic voice comparisons.

	Analytical	**Holistic**
Auditory-perceptual	Categorical-phonetic transcription and description	Holistic voice perception
Acoustic	Acoustic phonetics	Automatic speaker identification

Phonetic Alphabet (see International Phonetic Association 1999). An acoustic-phonetic approach (see Stevens 1998) builds upon an auditory-perceptual sound categorization and then investigates the acoustic manifestations of the perceptual categories. Acoustic phonetic analysis usually reveals that in acoustic reality, sound distinctions and sound separations in time are more gradient and less categorical than in perception. Within a forensic context acoustic-phonetic analysis has the advantage that very accurate quantitative values can be provided, which would be impossible with auditory-perceptual analysis. However, it might not always be the case that additional accuracy actually increases the performance of speaker identification.

The sound properties that a perceptual or acoustic analysis should focus on are those that are known to be subject to large differences between speakers, i.e. have large 'interspeaker variation' (also referred to as between-speaker variation). This should include those properties whose interspeaker variation has some anatomic-physiological foundation. That would be the case with voice pitch (auditory-perceptual) or fundamental frequency (acoustic), which depends on the length of the vocal folds, which in turn differs between speakers. However, even though some properties have at least some anatomic-physiological foundation, they are not exhaustively determined by those factors and can change due to linguistic and paralinguistic influences. On the one hand, these additional linguistic and paralinguistic factors can introduce further speaker-specific information, but on the other hand they can introduce undesired intraspeaker variation (cf. Braun 1995 for an overview of intraspeaker factors that influence fundamental frequency).

Proceeding further in the table, in a holistic approach speech is not 'taken apart' mentally or technologically but kept as a whole. Applying a holistic approach in the auditory-perceptual domain means to listen to speech without trying to analyze it componentially, while focusing mentally on the question whether two speech recordings have been produced by the same speaker or by different speakers. On the one hand, this approach resembles naïve speaker identification, on the other hand, there are reasons, to be mentioned below, why expert listeners have additional advantages when they listen to voice holistically as opposed to naïve listeners. The final possibility in the table occurs when speech is approached holistically with acoustic methods. This is the combination that is used in automatic

speaker identification.[8] Instead of using the term 'holistic' it is also possible to talk about 'global', 'nonanalytic', 'Gestalt', or 'overall' approaches.

The success rate of holistic methods in speaker identification is relatively straightforward to test because they can be applied more quickly than analytical speaker identification and, hence, more test comparisons can be run. In fact, it has been shown that the speaker identification performance of holistic methods is clearly good enough to be of practical use (Alexander et al. 2005). However, holistic methods do not fully satisfy the scientific interests and demands of the phonetic expert or the court, since there is no convincing answer to the question of what exactly it is that makes two particular voices so similar vs. different, or that makes a particular voice so common vs. rare compared to other voices in the population.

At present there is no comprehensive study on which of the four methods shown in Table 1 is the most accurate with respect to the task of distinguishing same-speaker pairs from different-speakers pairs in a forensic context. Given this situation, it would be clearly ill-advised to rely entirely on one method, and probably two are not enough either. It would also look bad from the point of view of the scientific basis of forensic speaker identification if it relied entirely on holistic methods. Currently, most forensic experiences exist with the three-way combination of methods that excludes automatic speaker identification. It is that combination that has been used for about 20 years at the German Bundeskriminalamt (BKA) and elsewhere in Germany, and has been used by most labs or individuals in the UK, Sweden, Finland, the Netherlands and Austria. Automatic speaker identification, which has a strong acceptance in France and Spain, deserves serious attention and should be tested more intensively with forensically realistic material – a task that is currently being explored at the BKA. The following subsections provide more concrete information on the four methods shown in Table 1.

3.2.1. Categorical-phonetic transcription and description

One important domain in which categorical-phonetic transcription and description is applied in forensic casework can be characterized by the term *linguistic phonetics* (cf. Ladefoged 1971).[9] The way it is used here, linguistic phonetics concerns the segmental and suprasegmental aspects of speech that distinguish different languages or different language varieties within the same language. Linguistic-phonetic transcription and description is important in speaker classification, specifically in the identification of dialects/regional accents, sociolects (understood here as any native-language variation that is correlated with social structure), as well as foreign accents. Language identification also belongs in this domain, if that task has not been already completed in the preparatory stages of the case by other parties. However, in many regions around the world, languages and dialects, or even different languages, might be hard to tell apart, so that the forensic

phonetician (or linguist) gets involved, eventually, if matters are difficult. As mentioned in Section 2.2, speaker classification is performed not just for the purpose of voice profiling, but also for the purpose of voice comparisons. The main importance of linguistic-phonetic speaker classification within voice comparisons is twofold. First, it can be very powerful evidence *against* the identity of two speakers if they differ with respect to dialect, sociolect or foreign accent. Second, even though speaker classification usually does not, by itself, lead to identification (because by definition it provides information about entire classes of speakers), it can substantially narrow down the range of possible speakers, especially with certain *combinations* of speaker classifications. For example, the author has encountered a case involving drug dealing where both the anonymous speaker and the suspect spoke German with a combination of a Russian foreign accent and a Swabian regional accent. Such kinds of combinations can strongly reduce the set of possible speakers.

Linguistic-phonetic speaker classification is not necessarily limited to the use of the auditory-perceptual method, although this method predominates in that domain. Ladefoged and Maddieson (1996) provide many convincing examples of the use of acoustic-phonetic methods in linguistic phonetics, and there are many examples in the literature where acoustic phonetics is used in dialectology, sociolinguistics and second-language phonetics (see Foulkes and Docherty 2006, for examples of acoustic methods in sociophonetics). It should be kept in mind that differences between varieties and even languages are not necessarily limited to categorical aspects, but can also involve gradient aspects and fine phonetic detail (cf. Keating 1990 and much subsequent work in 'Laboratory Phonology').

Another domain within voice comparison (but outside of speaker classification) where categorical-phonetic transcription and description is used by some forensic phoneticians is *voice quality*. Speakers can differ, for example, in whether they speak with a neutral (i.e. modal) voice quality or whether they have striking deviations from such a neutral voice quality, such as breathy voice, creaky voice, rough (harsh) voice, or pressed (tense) voice (or any combination thereof). Speakers can also differ with respect to the degree of a particular voice quality like light vs. strong breathy voice. So far only voice qualities that are produced in the larynx have been mentioned. But voice quality can also be understood more broadly to include supralaryngeal voice qualities such as nasal voice (cf. footnote 1). A theoretically motivated and practicable framework for the description and classification of voice quality, which covers both laryngeal and supralaryngeal aspects, is Laver (1980, 1994). Another important auditory-perceptual approach that is limited to laryngeal voice qualities is the GRBAS scale (G = total grade of hoarseness, R = rough, B = breathy, A = asthenic (i.e. weak), S = strained (i.e. similar to pressed voice)) (Hirano 1981). The GRBAS scale has been developed and tested in phoniatrics (a medical discipline dealing with voice disorders and speech disorders more generally), but has also proven useful for forensic

purposes. Both the Laver and the GRBAS system (in its form adapted for German phoniatrics by Nawka and Anders 1996) are used at the BKA (Köster and Köster 2004). It has been shown recently that training and forensic experience in the use of voice quality ratings can improve accuracy and inter-rater agreement (i.e. relative lack of differences between different listeners) in the perception of voice quality (Köster et al. 2007).

Nolan (2005) makes some important critical remarks about the limitations of the use of voice quality in forensic phonetics. For example, the bandpass filtering of telephone communication (cutting off frequencies below about 200 and above about 3500 Hz) eliminates some of the information that is important for the acoustics and ultimately the perception of some voice qualities. An example is breathy voice, which is cued in part by information from the first harmonic, which usually lies outside the telephone passband in male speech. Further problems occur when the recording is noisy, which can also interfere with the perception of breathiness. Problems like these are one reason why, in the domain of forensic voice quality analysis, the auditory-perceptual approach is used more frequently than the acoustic-phonetic approach. But, as pointed out by Nolan, if crucial acoustic components are missing (e.g. due to telephone transmission) or disturbing non-speech acoustic information is added, the perceptual voice quality evaluation process has to rely on 'a rather elaborate process of perceptually reconstructing what a sample would have sounded like had it not been passed through the telephone' (Nolan 2005: 396). It is possible that this reconstructing process, due to its interpretative nature, is to a certain extent idiosyncratic. This could exacerbate an effect that is inherent in voice quality perception, even with good-quality signals; namely, that inter-rater agreement can be relatively low (Kreiman et al. 1992). Nolan furthermore mentions that voice quality can be partially determined by dialectal differences and by stylistic variation. These influences will have to be factored out perceptually or intellectually in the voice comparison analysis, where the focus is on speaker differences. These problems have to be acknowledged and addressed when performing forensic voice quality analysis.

Phonetic transcription and description can also be applied in other domains that can be investigated in voice comparisons. One such domain is the use of *filled pauses*, i.e. utterances that are written in English like 'uh' and 'um' (cf. Clark and Fox Tree 2002). These are utterance types that speakers who use them are usually not aware that they do, and even if they are, still do not have much control over them. This can be relevant in voice disguise, where a criminal might concentrate on one type of disguise (such as falsetto voice), but be unable to control other aspects of his speech, such as dialect or filled pauses. Filled pauses can differ with respect to presence or absence of a nasal consonant, presence or absence of glottalization, nasalization, the exact vowel quality, etc., which can all be transcribed phonetically. They can also differ in other respects, such as their duration

or frequency of occurrence in a given time interval, which is better captured with measurements in the acoustic signal.

3.2.2. Acoustic phonetics

Some phonetic domains are better captured acoustically than auditorily. The most classic one of these in speaker identification is the average *fundamental frequency* (*f0*) of the voice. Fundamental frequency is the acoustic correlate of the vibration frequency of the vocal folds in voice production. Perceptually, voices differing in average *f0* can be described on a scale between 'high-pitched' and 'low-pitched'. However, such an auditory-perceptual scale would be less accurate than a scale based on acoustic *f0* values. Furthermore, it cannot be ruled out that high vs. low voices are confused with 'light' vs. 'dark' voices which reflect speaker difference in terms of vocal tract length rather than vocal fold vibration (cf. Jessen 2007b about this problem). No such confusion is possible acoustically, which is one reason why, in addition to being more accurate, *f0* analysis is also more objective than the perceptual estimation of speaker-specific pitch. Average *f0* is one of the few areas in which population statistics are available. The first statistics of this kind in forensic phonetics were presented by Künzel (1987; also shown in Künzel 1995).[10] They were based on the speech of 100 German-speaking men and 50 women in a reading task. More recently, another study has been conducted on German-speaking men, in which the results from Künzel (1987) were replicated (Jessen et al. 2005). Furthermore, data from other speech styles were presented, specifically on the difference between read and spontaneous speech and the difference between speech with neutral vocal loudness and loud speech. Loud speech was induced by a Lombard experiment, where loud white noise was presented over headphones (Jessen et al. 2005). In that study data on the standard deviation of *f0* values over the course of a recording were also presented.[11] This parameter is a correlate of the difference between speakers on the axis between very melodic and very monotonous ways of speaking. Figure 1 presents some of the results on average *f0*.

Figure 1 presents the distribution of the *f0*-mean in spontaneous speech at normal vocal loudness. From that graph the principle of typicality in the Bayesian approach (see Section 3.1.2) can be illustrated (cf. Rose 2002, 306–310). Assume two cases A and B. In case A the anonymous speaker has an average *f0* of 114 Hz and the suspect an average *f0* of 117 Hz whereas in case B the anonymous speaker has an average *f0* of 160 Hz and the suspect an average *f0* of 163 Hz. In both cases the similarity is the same (a 3 Hz difference, which is well within the expected intraspeaker variation). Hence, the numerator of the likelihood ratio (LR) would be the same and high; let us assume it is 1 in both cases. Now, the typicality would differ in case A and B, being higher in A than B. In case A, there is a likelihood of 0.31 that someone else than the suspect is identical with the anonymous speaker (see in Figure 1 that 31% of the speakers have average *f0* values

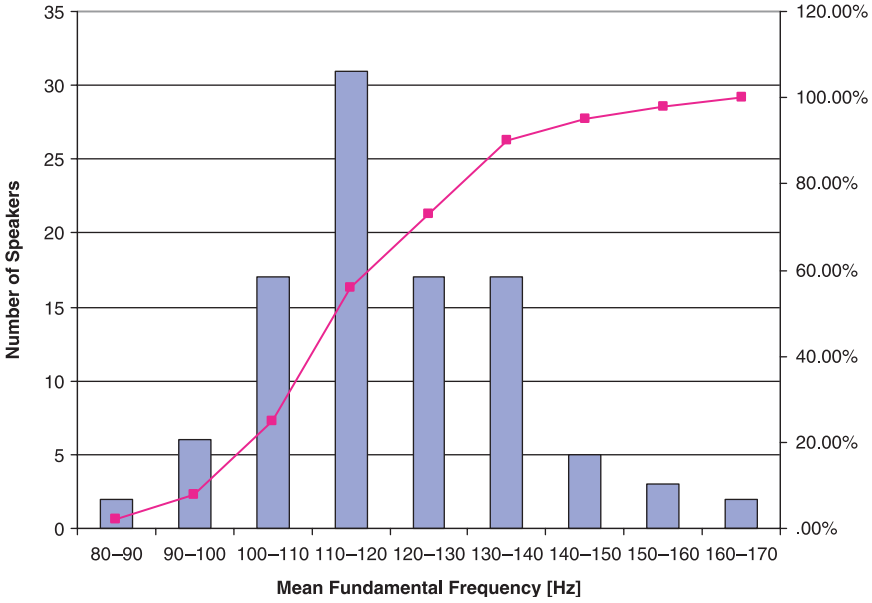

Fig. 1. Histogram (blue bars, *y*-axis left) and cumulative percentages (purple line graph, *y*-axis right) of mean fundamental frequency (intervals on *x*-axis) among 100 German-speaking men using spontaneous speech at normal vocal loudness.

in the 110–120 Hz interval, which is where the values in case A are located). With 1 divided by 0.31, the LR for case A would be 3.2. This still supports the identity between the anonymous speaker and the suspect (because LR > 1), but the LR is not very high and, therefore, the strength of the evidence is relatively weak. In case B, the likelihood that someone else than the suspect is identical with the anonymous speaker is 0.02, because according to Figure 1 only two of 100 speakers have an average *f0* that is between 160 and 170 Hz. With 1 divided by 0.02, we arrive at an LR of 50 which is much stronger evidence for the identity between the suspect and the anonymous speaker than in case A.

Figure 2 shows that the distributions of average *f0* values in the population differ quite substantially with different levels of vocal loudness. If for example the anonymous person spoke loudly whereas the suspect spoke with neutral loudness, the data cannot be directly compared (normally, although not always, there are at least some passages where loudness levels are equivalent). This illustrates that *f0* can be subject to a large amount of intraspeaker variation (see Braun 1995 for an overview). This is indeed the major problem with the forensic use of *f0*, which shows that *f0* has to be used in conjunction with several other phonetic parameters in order to arrive at a conclusion. Recently, population statistics on average *f0* based on 100 men speaking Southern British English have been presented by Hudson et al. (2007).

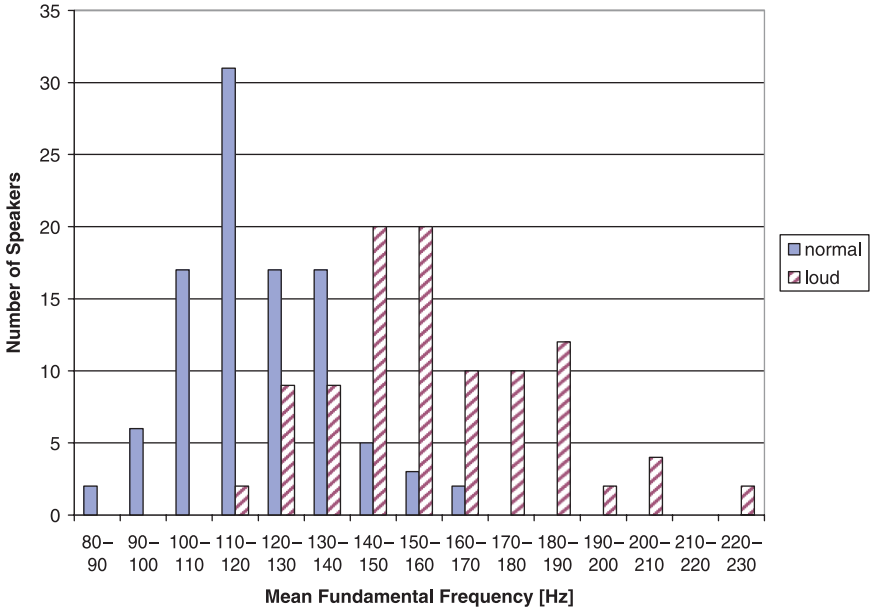

Fig. 2. Histogram of mean fundamental frequency (intervals on *x*-axis) among 100 German-speaking men using spontaneous speech at normal vocal loudness (filled blue bars) and increased vocal loudness (hatched red bars), induced by a Lombard experiment.

Other population statistics have been presented on *articulation rate* in German by Jessen (2007a). Articulation rate (expressed in syllables per second) is a quantification of speaker differences with respect to their speech tempo. Strictly speaking, articulation rate is a hybrid between auditory and acoustic processing: the syllables are counted on an auditory-perceptual basis, whereas the time it takes to complete a chunk of syllables is measured acoustically. Previously it had been shown that among the speech tempo parameters 'syllable rate' (also referred to as 'speech rate'), where pausing behavior is taken into account, and 'articulation rate', where only fluent speech is considered and all pausing ignored, the interspeaker variation is relatively greater and the intraspeaker variation smaller with articulation rate than syllable rate (Goldman Eisler 1968; Künzel 1997). This makes articulation rate the more promising measure for forensic purposes.

Another very promising domain of acoustic-phonetic speaker identification is *formant frequencies*. Although to date no similar population statistics exist for formants as they do for fundamental frequency and articulation rate, several research projects are currently working on it, the most advanced of them at Cambridge University (see www.ling.cam.ac.uk/dyvis). Formant frequencies are important correlates of distinctions between different consonants and vowels but they also carry much speaker information (cf. Stevens 1998). One of the anatomic foundations of formant frequencies is their

dependence on vocal tract length (i.e. the distance from the larynx to the lips): longer vocal tracts lead to lower vowel formants than shorter vocal tracts. This is why women tend to have higher formant frequencies than men, but there are also individual differences within the sexes (cf. Rose 2002). Not only the overall vocal tract length, but also proportions within the vocal tract determine formant frequencies. Notice that the pharynx is longer in relation to the rest of the vocal tract among male than female adults (Fitch and Giedd 1999). This probably also holds to a smaller extent for individual differences within these populations.

Several approaches to forensic formant frequency measurements are possible. The most classical one is to measure the center frequencies of different vowel categories. Rose (2002, 2006a,b) and Rose et al. (2003) have shown how these measurements can be used in the Bayesian evaluation of phonetic evidence. As Rose points out, the difficulties, yet also the potential of this approach, lie in the fact that the formant frequencies of different vowels and different formants (especially the second and the third formant; others usually lie outside the telephone passband or are compromised by it) are less than fully correlated. One expected reason for this lack of high correlation is the circumstance – just mentioned – that individual patterns are not limited to differences in overall vocal tract length but also involve differences in the various sections of the vocal tract. This limited correlation requires sophisticated statistical modeling but offers the opportunity that by combining results from different formants and vowels high likelihood ratios can be obtained.

Another way of approaching formant analysis is by focusing on the formant dynamics. Recent work on this subject has been conducted by McDougall (2004, 2006) and McDougall and Nolan (2007). As pointed out by McDougall (2006), it is plausible to assume that while the formant targets (i.e. those regions addressed by measurement of formant center frequencies) are largely determined by the phonological requirements of a language, speakers have some freedom in how they move articulatorily from one target to the next. Therefore, some traces of speaker-specific motor strategies should emerge that add to the speaker-specifics that are given by the static differences in vocal tract shape. These predictions were borne out, and indeed formant dynamics have high speaker-differentiating power. Due to the fact that formant dynamics are also influenced by segmental and prosodic contexts, most strongly in spontaneous speech, more research as to how these sources of intraspeaker variation can be treated is necessary.

A third method of measuring formant frequencies is by using Long Term Formant Distributions (LTF), as proposed by Nolan and Grigoras (2005). Instead of selecting specific vowel targets, LTF analysis captures the information from all portions of the speech recordings where formant structure is visible and reliable with the application of Linear Predictive Coding (LPC)-based formant tracking. When signal quality is sufficient, the formant information that is included is usually from all or most of the

vowels and glides, otherwise from only a subset of these. The advantages of this method are its relatively time-efficient application and its applicability even with languages that are not spoken by the expert (because identification and segmentation of phonological vowel categories is not required, only a good knowledge of general acoustic phonetics). A disadvantage is that because of the pooling of all vowel information, the results are more difficult to interpret than with segmented vowels. By pooling different vowels, probably some opportunity for better speaker-discrimination power is lost which would exist if results from different vowels were evaluated separately and then combined in a Bayesian analysis (in the sense proposed by Rose, discussed above). A more recent version of LTF, implemented by Grigoras, allows for the automatic identification of the vowel categories of a language and the listing of results separately for each vowel. However, for any given case, which might involve poor-quality material, the reliability of this algorithm would have to be checked for at least a subportion of the material.

Aside from the range of acoustic-phonetic parameters in speaker identification that have been mentioned so far, and for which at least some forensic experiences are available, many more acoustic-phonetic parameters are conceivable which could carry speaker-specific information, and which could prove useful in forensic phonetics. The need for more research in this area is clearly indicated. Although practically every acoustic-phonetic study reveals some speaker differences, the important point for forensic applicability is that the speaker specifics must be robust enough against technical limitations, short duration, different speaking styles, natural speech phenomena like stress and emotion, and so forth (cf. Section 3.1.3). Today nobody knows how many of the acoustic properties that show significant speaker differences in carefully controlled experiments with good recording conditions and equipment would pass this test of forensic realism (Nolan 1983, 11–14; Rose 2002, 51f) and could become valuable innovations in forensic phonetics. One promising area is duration measurements of sounds, sound features or prosodic units. Since it is known that articulation rate or other speech tempo measures can differ between speakers, duration measures other than those pertaining to speech tempo should be normalized against tempo. Allen et al. (2003) addressed the question whether known speaker differences in Voice Onset Time (i.e. the time from the release of a stop consonant to the voicing onset of the following vowel) are merely a side effect of speaker differences in speech tempo. They normalized for the tempo factor and found speaker differences even after normalization. Similar effects were shown by Pfitzinger (2002), who found speaker differences in the duration of different segments, even after normalizing for speech tempo differences.

3.2.3. Holistic voice perception

As discussed earlier, lay listeners have a certain ability to recognize voices. Since lay listeners are not trained in the systematic analysis of language

and speech, their voice perception is essentially holistic. Experts in language and speech are able to perceive voices analytically, but there is no reason why they should not make use of their ability for holistic perception as well. After all, there is the possibility that the principle of Gestalt theory, 'the whole is greater than the sum of its parts', applies to voice identification. Some studies have shown that phoneticians are somewhat better in holistic speaker identification than lay listeners (see Hollien 2002, 37–39, for discussion and further literature).

In one study, this advantage has been demonstrated for forensic phoneticians in particular (Schiller and Köster 1998). In that study, speech from six German native speakers was split into 4- to 8-second samples. Seventeen German-speaking phonetically untrained listeners and 10 German-speaking forensic phoneticians had to respond in a listening test whether the voice they had been familiarized with 5 minutes earlier was the same as the voice they heard in each of the 108 speech samples that were presented. The untrained listeners had a hit rate (answer 'identical' where the voices were in fact identical) of 92% and a false alarm rate (answer 'identical' where the voices were not identical) of 2%. The forensic phoneticians had a hit rate of 98% and a false alarm rate of 1%. These results show that the forensic phoneticians were both better in identifying identical speakers and in rejecting foils. So far no systematic research has addressed the question of what it is that gives general phoneticians or even forensic phoneticians the advantage over lay listeners with respect to holistic perceptual speaker identification.

A plausible explanation of holistic voice perception can be found in *Exemplar Theory* (Johnson 1997). Exemplar Theory and other theories of non-analytic cognition (cf. Pisoni 1997) stand in contrast to theories in which the perception process is viewed as analytical activity that is analogous to the analytical activities of a scientist. For example, in classical analytical theories of speech perception, speaker and speech recognition have been understood as a sequential process where the listener first estimates the speaker characteristics and then uses this information to perform speaker normalization that is subsequently required in speech recognition. In Exemplar Theory, on the other hand, speaker and speech recognition are performed simultaneously rather than sequentially. In analytical theories only the pure speaker characteristics and the information necessary for speech perception (e.g. feature detection), lexical access and other important aspects of speech recognition are stored in memory, whereas in Exemplar Theory the entire speech event is stored in the form of 'exemplars' that contain all the phonetic details. According to Johnson (1997), an exemplar is an association between auditory properties – which are the output from the peripheral auditory system – and category labels. These category labels can tag any type of information necessary for communication, including the linguistic entity of the exemplar, the sex of the speaker or the identity of the speaker. All speech samples from the same speaker, for example, would have the same category label for that speaker. This information, which is stored in memory,

can be used later when the task is to identify that speaker on another occasion. As is apparent from this short summary, Exemplar Theory provides a straightforward explanation, not only for holistic speaker identification, but also for the holistic perception of speaker classification properties. For example, research has shown that lay listeners are, to a certain degree, able to classify age (e.g. Braun 1996) or dialect (Clopper and Pisoni 2005).

Exemplar Theory has many advantages; for example, it is able to explain interactions between speech and speaker recognition (see Nygaard 2005) that analytical theories have more difficulties with. On the other hand, Exemplar Theory is confronted with the 'head-filling-up problem' (Johnson 1997). Although psychological research has shown that memory capacity is larger than previously assumed, there must be limits that would be reached much earlier with holistic processing than analytical processing, where only the essentials of speaker identification/classification and speech recognition are stored. Problems might, for example, occur when very many labels of a category type have to be distinguished, as when very many voices have to be distinguished (Henning Reetz, personal communication). Notice also that the ability of listeners to identify speakers holistically (i.e. without consciously using or being able to use abilities of language and speech analysis) is not a proof for Exemplar Theory per se. It is still possible that the cognitive processing that is behind this ability is analytical, only that, in contrast to an expert analysis, it is taking place unconsciously. One argument for this position is that it could possibly explain the advantage that phoneticians have over non-phoneticians in holistic voice perception, which was mentioned above. An explanation of the type 'the conscious meets the unconscious' would be that in their unconscious analytical processing while performing holistic voice perception, experts are aided by their conscious abilities for voice analysis, which they have because of their training.

On the other hand, a 'verbal overshadowing' effect in voice recognition has been found, meaning that when, in an experimental line-up paradigm, subjects are asked to provide a description of the target voice prior to the recognition task, their recognition rate is lower than in a control task, where no voice description is requested (Perfect et al. 2002). This is a more general effect that has also been found in face recognition and is interpreted as an indication of the importance of holistic as opposed to feature-based processing. The practical implications for forensic speaker identification have to be worked out in more detail. The implications are probably stronger for speaker identification by victims and witnesses than for speaker identification by experts because some of the explanations offered for verbal overshadowing do not apply to voice comparisons. For example, the explanation that the verbal description task reduces the availability of the original memory trace, which is mentioned by Perfect et al. (2002), is not relevant in auditory-holistic (or any other) expert examination of audio recordings, where the material can be played as

often as necessary and at any point during the analysis, so that memory is not an issue.

Nolan (2005) reminds us that holistic voice perception, like any other method in forensic speaker identification, needs to respect the principles of the Bayesian approach. He points out that in holistic voice perception in particular, there is a danger that too much attention is given to the similarity aspect and too little to the typicality aspect. He warns that the expert might become 'overwhelmed by similarity' (p. 401), not noticing that many other speakers would evoke the same gestalt perception. Since holistic voice perception can draw upon any kind of voice characteristics in the broadest sense possible – including dialect and dialectally determined voice quality – it is possible, according to Nolan, that some similarity perceptions would be shared by a broad range of dialectally homogeneous speakers.

3.2.4. Automatic speaker identification

Automatic speaker identification (more frequently found under the name automatic speaker recognition) is a method in forensic speaker identification that is essentially no more than 10 years old. Automatic speaker identification as a discipline outside the forensic arena (to be referred to as general automatic speaker identification) is much older, dating back at least to the 1960s. General automatic speaker identification is a speech-technological discipline that has been (and still is) developed for purposes such as access control to sensitive information or to high-security buildings, devices, etc., or for other biometric applications. General automatic speaker identification is, in several ways, a simpler task than forensic automatic speaker identification. For example, the recordings of the speakers who want to be given future access, as well as all the required reference voices, can be made in qualitatively and quantitatively optimal ways, and there is no mismatch problem of the sort mentioned for forensic data in Section 3.1.3. General automatic speaker identification is also conceptually different from automatic speaker identification in forensics. Whereas concepts like closed-set identification or the setting of thresholds for acceptable false identification or false rejection scenarios are common in general automatic speaker identification they are uncommon in forensics, except for some specialized applications. On the other hand, the Bayesian approach, which is fundamental to forensics (see Section 3.1.2), is no absolute requirement in general automatic speaker identification.

It is probably due to a convergence of three factors – increased efficiency in the methods of general automatic speaker identification, better computer technology, and the advance of Bayesian thinking in forensics – that serious efforts have been made during the past few years to bring automatic speaker identification into the field of forensic speech and audio analysis.

A system for forensic automatic speaker identification essentially consists of three components, which correspond to three sequential stages of the

automatic speaker identification process. These stages are parameter extraction, parameter modeling and the calculation of distances. In the first stage, *parameter extraction*, acoustic parameters are automatically extracted from the speech signal. Different proposals exist as to which acoustic parameters are most efficient for the task. The parameter that has been used most frequently is called Mel Frequency Cepstral Coefficients (MFCC). MFCC feature vectors are generated by the following steps. First, the Power Spectrum is obtained. This is usually done by application of the Discrete Fourier Transform, which is a common method in acoustic phonetics. As a means of smoothing the spectral shape and of making the outcome more realistic psycho-acoustically, the spectrum is then passed through a filterbank based on the non-linear Mel scale. The logarithms of the filter coefficients are transferred to the cepstrum by application of the Discrete Cosine Transform. The resulting vectors are now called cepstral coefficients (Bimbot et al. 2004 for more details and further literature).

About 20 MFCCs are extracted within windows of about 20-ms duration every 10 ms (i.e. with 50% overlap of the windows). The distribution of the MFCCs over the entire course of the recording of a speaker is determined. Notice that with this method no segmentation of the speech stream into different linguistic categories, such as consonants, vowels or syllables is performed. This is one of the reasons why automatic speaker identification was classified as a holistic method in Table 1.

An important aspect worth knowing about MFCCs is that they allow for a good source-filter separation. In phonetics, the source-filter model refers to the distinction between the voice source at the level of the larynx and the contribution of the supralaryngeal structures that act as a filter of the source energy. It has been shown that it is beneficial for the automatic identification results if MFCC components that capture the contribution from the source – in particular the fundamental frequency – are ignored and if only the MFCC components that capture the filter are used. Since formants are an important filter characteristic, this choice seems to imply that formants are more important for speaker identification than fundamental frequency. Notice, however, that the extraction of MFCCs is not the same as automatic formant tracking (as mentioned in Section 2.2.2). In formant tracking, algorithms are used that attempt to find formants. Since formants are objects that are well studied in phonetics – both with respect to how formants are generated by the structures and configurations of speech production and with respect to the consequences that formants have for speech perception – the phonetician has the knowledge to determine where automatic formant tracking fails, which can often happen in low-quality speech signals (in that case the phonetician can manually correct the mistakes). In contrast, MFCC extraction never fails because there is no sense of correct vs. incorrect tracking of a predefined structure in the acoustic signal. However, the values that result from MFCC analysis are not interpretable in the same sense that formants are. Very little is known

about how specifics of MFCC information relate to specifics of speech production and perception. Perhaps this is something phoneticians should turn their attention to, but so far such attempts are rare, such as Mokhtari et al. (2007) (cf. Rose 2002 for similar discussion of the differences between formant and cepstral analysis). This lack of phonetic/linguistic interpretability of cepstral patterns is the second reason why automatic speaker identification was classified as a holistic method in Table 1.[12]

Since the patterns of MFCC extraction cannot be interpreted phonetically, it is consequently not clear which aspects of MFCCs contain speaker information and which contain information about the channel, about background noise and other irrelevant and disturbing acoustic influences. Indeed, separating voice information from irrelevant information is an important and difficult task in automatic speaker identification, which is particularly challenging in forensics as opposed to other applications, where the quality of the material is usually better. It is often performed in several steps. One of them is an audio enhancement stage that can be performed before MFCC extraction takes place, another one is called cepstral mean subtraction, which means that each of the local MFCC results is subtracted from the MFCC average of the entire recording, working with the assumption that the channel information and noise problems remain constant throughout the recording.

MFCCs, since they holistically capture all aspects of speech information, also contain information about paralinguistic and stylistic aspects, such as reading vs. speaking spontaneously or speaking with or without a particular emotion. If the anonymous speaker, the suspect and all the speakers in the reference population use the same paralinguistic or stylistic setting there is no problem, but often there is a mismatch with respect to these behavioral aspects. In order to address this problem, automatic speaker identification uses 'compensation for mismatched conditions' (Botti et al. 2004). It is implemented at the third stage (calculation of similarities), which will be explained further below. Compensation for mismatched conditions can also be used for technical mismatches or frequently recurring combinations of technical and behavioral mismatches.

Another possible type of mismatch is different languages. Not uncommonly, practitioners of forensic automatic speaker identification use a reference population of speakers that speak a different language than the anonymous speaker and the suspect (or use a reference population where a variety of languages are spoken). This is possible because experience has shown that language differences have only a very small effect on the MFCC patterns. This language independence is usually regarded as a major advantage of forensic automatic speaker identification as opposed to other methods. However, recent research in automatic speaker identification has demonstrated that language *does* matter. Specifically, in the speech of bilingual speakers, recognition rate was lower when training and test material were in different languages than when the language was matched (Przybocki et al. 2007).

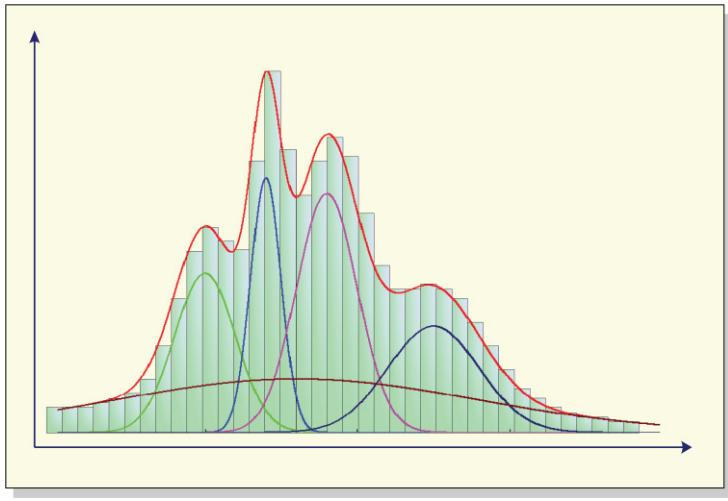

Fig. 3. Abstract representation of a Gaussian Mixture Model.

Turning now to the second stage of automatic speaker identification – *parameter modeling* – this stage has the purpose of modeling the distribution of the raw MFCC results. Again several methods have been proposed, but the one that has been used most frequently results in what is called Gaussian Mixture Models (GMM). Figure 3 provides an illustration of a GMM.

Figure 3 demonstrates how the distribution is modeled with a set of Gaussian functions, in this case five (see the five Gauss curves in different colors). The resulting GMM is shown as the red-line envelope of the distribution. The *x*-axis abstractly represents different values for an MFCC feature vector and the *y*-axis shows how frequently a given interval of feature vector values is found in a speech sample. This is a one-dimensional display, whereas in reality, multidimensional GMM functions with the data from about 20 feature vectors are calculated. A Gaussian function is also referred to as a 'component'. The number of components is one of the variables that can be selected by the user and can be varied in the optimization of automatic speaker identification procedures. The result of GMM modeling is a speaker model (i.e. it is a model of the speech patterns of a particular speaker).

The third and final stage is the calculation of distances. Here we will talk about the *calculation of similarities*, which is the converse perspective and easier to understand. This process is illustrated in Figure 4.

In Figure 4, two distributions are shown. These must not be confused with a GMM; they are second-order distributions (these second-order distributions can be modeled in a number of ways, one of which is called kernel density function). The distribution to the right (in blue) is a second-order model of the speaker that is the suspect in a voice comparison. It

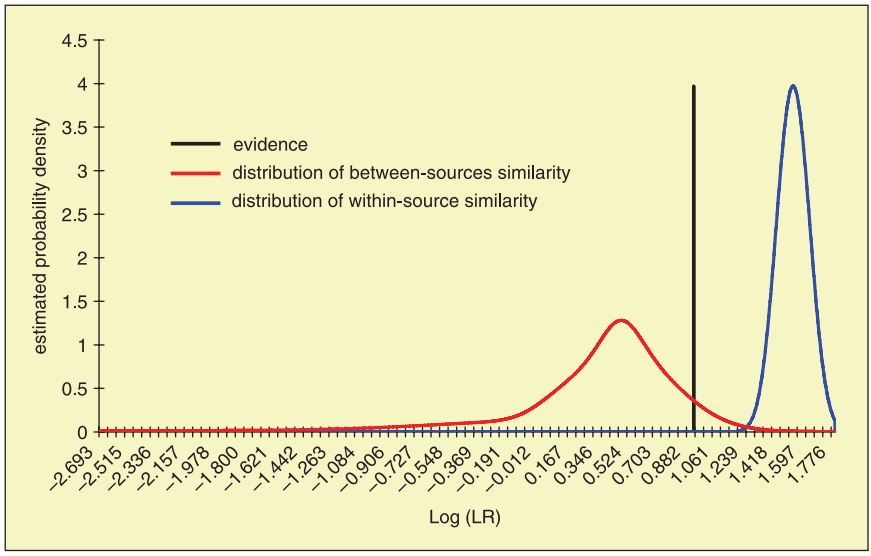

Fig. 4. Calculation of likelihood ratios based on similarity scores in forensic automatic speaker identification: Example of non-identity. Depending on the method, the similarity scores themselves (shown on the *x*-axis) are expressed in terms likelihood ratios, but these must not be confused with the likelihood ratios that result from the comparison of the evidence with the within-speaker and the between-speaker distribution of similarity scores, which is explained in the text.

is derived by calculating several similarities between the GMM of the suspect speaker and different speech samples from the same speaker. Ideally, the speaker should speak in different conditions. The *x*-axis provides an index of the different degrees of similarities between speech samples and GMM. Since the samples in the blue distribution are all from the same speaker, it is expected that the degrees of similarity are high, i.e. the blue distribution is located towards the higher end of the similarity scale on the *x*-axis. The distribution to the left (in red) results from the calculation of the similarities between the speech sample of the anonymous speaker and the GMMs from each of the speakers in what is called a reference population. The reference population is a collection of speakers, usually 30 or more, that is necessary in order to quantify in a Bayesian way whether some speaker other than the suspect could have been the anonymous speaker. Since it can safely be assumed that the anonymous speaker is not identical with any of the speakers from the reference population, it is expected that the similarity scores are lower than in the case of same-speaker comparisons and, hence, are located more towards the lower values of the similarity scale. In summary, the blue distribution captures within-speaker similarity, i.e. the similarities that result when the suspect is compared with himself, whereas the red distribution captures between-speaker similarity. The final ingredient that is necessary at this stage of the automatic speaker identification

procedure is called the Evidence (E). It is represented as a black bar. The Evidence is the similarity value that results when the speech sample of the anonymous speaker is compared with the GMM of the suspect. The Evidence is evaluated according to the LR formula shown in Section 3.1.2. The numerator of the LR is the likelihood of obtaining the degree-of-similarity value of the Evidence (see where the black bar is located on the *x*-axis) assuming that the Evidence is part of the within-speaker distribution. In Figure 4 this is the point where the black bar intersects the blue distribution, which is at an extremely low likelihood value (see *y*-axis). The denominator of the LR is the likelihood of obtaining the degree-of-similarity value of the Evidence assuming that the Evidence is part of the between-speaker distribution. This is the point where the black bar intersects the red distribution, which is at a likelihood value clearly higher than in the comparison with the within-speaker distribution. When the numerator is divided by the denominator the LR will obtain a value smaller than one. This LR smaller than one, as illustrated in Figure 4, is evidence that the suspect is not identical with the anonymous speaker. In the case of an identity result, the black bar would be shifted to the right and would occur more within the within-speaker than the between-speaker distribution.

A more detailed overview of the automatic speaker identification procedure described here, particularly of the third stage, is provided by Drygajlo (2007). Reviews of current developments are also found in Gonzalez-Rodriguez et al. (2006) and Müller (2007). Further references are listed in Bijhold et al. (2007). In this rapidly developing field, the current state of the art can be updated by consulting the outcome of the 'Speaker Odyssey' conferences that take place about every 2 years. This conference has its strongest foothold in general automatic speaker identification, but also has regular sessions on the forensic applications of automatic speaker identification.

4. Conclusion

This review has attempted to give an impression of the wide range of topics, tasks and scientific challenges that characterize the field of forensic phonetics today. While other areas of forensic phonetics have been mentioned in the introduction, the main focus of this presentation has been on speaker identification, which is at the core of the scientific responsibilities of forensic phonetics. Within speaker identification, most of the current practical developments and theoretical issues concern voice comparison. The question of this task is simple: Given two different recordings, are the two recordings spoken by the same person or by different persons? As simple as the question is, it requires substantial experience and scientific backup to provide appropriate answers in forensic casework.

What has become clear through the advance of the Bayesian approach in the forensic sciences is that voice comparison is not just a matter of

comparing the similarities and differences between the voice patterns of different recordings, but also of knowing how common these voice patterns are in an estimate of the relevant population of speakers. If the similarity aspect and the typicality aspect can be expressed quantitatively, a likelihood ratio can be calculated. But even if knowledge about these aspects is more qualitative (such as in auditory phonetics) or more implicit (such as in holistic voice perception) the concept of a likelihood ratio and the Bayesian approach should still be a guiding principle in forensic speaker identification (Rose 2002).

Another point that has become clear over the years is that speaker identification should not be confined to a single method, but that a variety of different methods should be used. Four such methods were described here (Table 1). With several methods, conclusions of identity or non-identity become more stable and the risks of false identifications or false rejections are reduced. This is mainly because each of the methods covers a different aspect of the entire range of speaker specific information. Automatic speaker identification, for example, scans the speech signal more comprehensively than analytical methods, so that hardly any potential speaker specific component will be left out, yet on the other hand, through the expert knowledge that is required in analytical methods, those pieces of information can be identified that are of particular importance for the speaker identification process and which might get lost in the wealth of information collected with automatic methods. For example, presently automatic systems would not usually be able to detect the type of dialectal information that leads to the conclusion that two speakers are probably not identical.

Forensic phonetics is still a young discipline. Although solid experience from forensic laboratories and private practitioners has been accumulated and crucial research has been conducted, there are still many areas that can benefit from more research. One important research domain is acoustic phonetics. Although it is known that formant structure carries important speaker specific information, it is still not known how maximal use can be made of the formant evidence. Combining information from different formants and vowels, or capturing formant dynamics have been discussed as possibilities. Another important area is the exploration of the perceptual abilities of humans (and perhaps non-humans as well) to identify voices. Exemplar Theory is one of the relevant theoretical accounts in this quest. Still another question is to what extent different methods capture similar speaker information (and thereby contain redundancy, which can add confidence to the results) or whether they capture different information that can be treated as independent in a probabilistic sense (which would justify multiplying likelihood ratios, leading to very strong results). And what if there are conflicts between the different methods? Such situations can occur if, for example, auditory methods point towards identity whereas acoustic-phonetic evidence speaks against identity (Nolan 1990 for such a case), or if automatic speaker identification reports non-identity whereas all other methods support identity. Another

important issue is the relation between formant analysis and automatic speaker identification. We have seen that both concentrate on vocal tract characteristics. More research should address the question of whether these two methods lead to similar or different results on the identity or non-identity of speaker pairs and what the reasons for these similarities or differences might be (see Rose et al. 2003 for an important study in that direction).

One final point should be emphasized. Although it might seem that the issue of speaker characteristics is a genuinely forensic task, it is not. What is genuinely forensic are sub-issues, such as how formants are influenced by mobile phone technology, how to handle voice disguise, or how to compensate for mismatched conditions. But speaker characteristics (often referred to as talker characteristics) are also an issue of general phonetics. For example, in many phonetic studies the speaker is one of the factors that lead to significant effects. These types of research results can be very inspiring for forensic phonetics. Forensic phoneticians should not bury themselves completely in casework, but should keep themselves informed about speaker characteristics that emerge from general phonetics. Likewise, general phoneticians should be aware that speaker characteristics are not merely a forensic issue, but something that is of importance for their own field. General phoneticians can make an important contribution to the forensic field, for example by running experiments with many speakers, by investigating natural speech conditions, or simply by presenting experimental results in a way that speaker variation can be seen, rather than pooling across speakers.

Acknowledgments

Thanks to Paul Foulkes and another reviewer for their valuable comments which helped me to improve this paper. Among many other points, thank you for guiding my attention to the paper by Perfect et al. (2002). I also appreciate the comments of Franz Broß and Timo Becker about the section on automatic speaker identification and I thank Franz Broß for letting me use the illustrations in Figures 3 and 4.

Short Biography

Michael Jessen received his MA degree in linguistics from Universität Bielefeld, Germany, and his PhD in linguistics from Cornell University with a study on German phonetics and phonology. He worked as lecturer for linguistics and phonetics in the Department of Computational Linguistics and Phonetics at Universität Stuttgart, Germany, for 8 years. During that time he did phonetic fieldwork on Xhosa as a postdoctoral fellow at Stellenbosch University, South Africa. In 2001, he joined the Speaker Identification and Audio Analysis Department of the federal forensics laboratory of Bundeskriminalamt, Germany. He is involved both in practical

casework and in research and development of methods in forensic speaker identification.

Notes

⋆ Correspondence address: Michael Jessen, Bundeskriminalamt BKA, Wiesbaden 65173, Germany. E-mail: michael.jessen@bka.bund.de.

[1] The use of the expression *voice* in *voice comparison*, *voice profiling* and related terms is meant as a *pars pro toto* for all aspects of speech production. This includes voice in its most literal sense – the vibration patterns of the vocal folds – in a broader sense of including supralaryngeal structures like in nasal voice, and in the broadest sense of including linguistic characteristics such as the dialect of a speaker. Laypersons when claiming 'I recognize that voice' are often not aware of the full range of phonetic and linguistic features that carry speaker-specific information. Alternative terms to voice comparison, voice profiling, etc., are speaker comparison, speaker profiling, etc.

[2] A website has been set up containing this position statement and related information (www.forensic-speech-science.info).

[3] The IAFPA also recommends against making assessments of the sincerity of speakers. This is because it has not been proven that verbal lie detection is possible (cf. Hollien 1990, chapter 13; Eriksson and Lacerda 2007).

[4] In the opinion of the New Jersey's Court of Errors and Appeals, which affirmed Hauptmann's murder conviction, the three most substantive and independent pieces of evidence against Hauptmann were the possession and use of the ransom money, which was hidden in the garage of Hauptmann's residence, the statement of several handwriting experts that Hauptmann was the writer of the ransom notes, and the point that the wood used in the construction of the ladder that was used to climb into the nursery of the Lindbergh residence in order to kidnap the baby could be connected to Hauptmann's attic (Fisher 2000, esp. p. 385). Specifically, evidence from wood patterning and tool marks strongly suggested that a plank that was missing in the attic floor was identical to one of the rails in the ladder (Fisher 1999, 125, for further details).

[5] The guidelines to voice line-ups summarized in this paragraph derive from the paper by Broeders and van Amelsvoort (1999), which was mentioned above. Not all aspects of these guidelines are universally accepted and there are good arguments for some alternatives in the details of voice line-up construction and implementation. For example, there are different views on the procedure of requiring a decision after each voice is heard (cf. Nolan 2003 for further discussion and illustration).

[6] Named after Thomas Bayes, an English Presbyterian minister and mathematician who lived in the eighteenth century and worked in the town of Tunbridge Wells. A recommendable popular-science account of the Bayesian approach in the judicial system is provided by Kaplan and Kaplan (2006).

[7] The modification 'relatively' is used here because the final probability for or against identity also depends on the 'prior odds', which are facts relevant to the case (e.g. witness statements, alibi status), but unrelated to speech analysis. Notice that the final outcome (after prior odds have been taken into account) is reported as the probability of identity or non-identity given the evidence, whereas the LR expresses the probability of the evidence given identity (numerator) or non-identity (denominator). The present discussion is simplified insofar as it focuses on the LR. For more details see Rose (2002).

[8] Another approach, which can be characterized as 'visually holistic', is the 'voice print' method. Essentially, the voice print method is the comparison of spectrograms taken from the anonymous speaker and the suspect in speech portions that are preferably text-identical. The comparison is primarily based on 'eye balling' the differences and similarities on two spectrograms. Since the voice print method is offered in compact courses for people with no previous education in fields such as phonetics and linguistics, it is clear that the comparison of spectrograms is not grounded on a systematic acoustic-phonetic analysis by somebody appropriately trained to do so. The voice print method has a long history, which is presented by Nolan (1983), Hollien (2002) and Rose (2002). Due to several problems with this method, it is rejected by most

practitioners and most courts. There is a very clear consensus in the forensic-phonetic community that the voice print method must not be used. Caution should be taken, however, that by rejecting voice printing, one does not reject spectrographic analysis or even acoustic phonetics as a whole. This is an unfortunate baby-out-with-the-bath-water response that has occurred in some US states (Foulkes, personal communication).

[9] The conjunction 'transcription and description' is used here because the use of phonetic transcription is not obligatory in a voice comparison report. Crucial pronunciation aspects can also be described by the classification matrix of the International Phonetic Alphabet (e.g. talking about a postalveolar voiceless fricative instead of using its symbol) or transcribed or described in ways that might be more directly comprehensible by the court. Furthermore, voice qualities and some other prosodic phenomena are more conveniently described than transcribed.

[10] See also Künzel (2000) where aspects from that population study are presented in which subjects spoke with different kinds of voice disguise.

[11] It was shown in Jessen et al. (2005) that mean *f0* and *f0* standard deviation are positively correlated. By expressing standard deviation as coefficient of variation (standard deviation divided by mean), the correlation essentially disappeared. Since different forensic-phonetic parameters should be as uncorrelated as possible in order to obtain high LR (Rose 2002, 52), it is preferable to express *f0* variability as coefficient of variation.

[12] In a relatively new approach to automatic speaker identification, 'higher-level features' are used in addition to 'acoustic features' such as MFCCs. One of the first and comprehensive examples of this approach is the Super-SID project at MIT Lincoln Lab (http://www.clsp.jhu.edu/ws2002/groups/supersid/SuperSID_Closing_Talk_files/frame.htm). A recent overview of higher-level features is provided by Shriberg (2007). Higher-level features are features that are phonetically and linguistically interpretable, such as segment phonotactics, segment durations, intonation properties, etc. Many of these higher-level features are derived with the help of automatic speaker recognition, which leads to a segmentation of the speech stream into segments, syllables and other linguistic units. Due to linguistic segmentation and the use of phonetically interpretable features, this is an analytical rather than holistic approach. However, since this higher-level features approach is still fully automatic it can – unlike the use of MFCCs – make objective mistakes in linguistic/phonetic classification. Especially with forensic material and within a context of forensic decision-making, complete reliance on the analytical abilities of an automatic system is dangerous. At least some systematic supervision of such a system by the analytical abilities of an expert is recommended. This is not to say that developers of higher-level feature recognition generally recommend their use in forensic speaker identification. Presently, higher-level feature analysis is more a matter of general than forensic automatic speaker identification.

References

Alexander, Anil, Damien Dessimoz, Filippo Botti, and Andrzej Drygajlo. 2005. Aural and automatic forensic speaker recognition in mismatched conditions. International Journal of Speech, Language and the Law 12.214–34.

Allen, J. Sean, Joanne L. Miller, and David DeSteno. 2003. Individual talker differences in voice-onset-time. Journal of the Acoustical Society of America 113.544–52.

Baldwin, John, and Peter French. 1990. Forensic phonetics. London, UK: Pinter.

Bijhold, Jurrien, Arnout Ruifrok, Michael Jessen, Zeno Geradts, Sabine Ehrhardt, and Ivo Alberink. 2007. Forensic audio and visual evidence 2004–2007: A review. 15th INTERPOL Forensic Science Symposium, October 2007, Lyon, France, http://www.forensic.to/webhome/enfsidiwg/Interpol-review-2007-audio-visual-evidence-paper.pdf

Bilton, Michael. 2003. Wicked beyond belief. The hunt for the Yorkshire ripper. London, UK: Harper Collins.

Bimbot, Frédéric, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A. Reynolds. 2004. A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing 4.430–51.

Blatchford, Helen, and Paul Foulkes. 2006. Identification of voices in shouting. International Journal of Speech, Language and the Law 13.241–54.

Botti, Filippo, Anil Alexander, and Andrzej Drygajlo. 2004. On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition. Forensic Science International 146S.101–6.

Braun, Angelika. 1996. Age estimation by different listener groups. Forensic Linguistics 3.65–73.

——. 1995. Fundamental frequency – How speaker-specific is it? Studies in forensic phonetics, ed. by Angelika Braun and Jens-Peter Köster, 9–23. Trier, Germany: Wissenschaftlicher Verlag Trier.

Broeders, A.P.A. 2001. Forensic speech and audio analysis, forensic linguistics 1998 to 2001: A review. 13th INTERPOL Forensic Science Symposium, 16–19 October 2001, Lyon, France, http://www.interpol.int/Public/Forensic/IFSS/meeting13/Reviews/ForensicLinguistics.pdf

——. 2004. Forensic speech and audio analysis, forensic linguistics. A review: 2001 to 2004. 14th INTERPOL Forensic Science Symposium, 19–22 October 2004, Lyon, France, 171–88, http://www.interpol.int/Public/Forensic/IFSS/meeting14/ReviewPapers.pdf

——. 2006. Phonetics, Forensic. Encyclopedia of language and linguistics (2nd edn), ed. by Keith Brown, 460–3. Amsterdam, The Netherlands: Elsevier.

Broeders, A.P.A., and A.G. van Amelsvoort. 1999. Lineup construction for forensic earwitness identification: a practical approach. Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, CA, vol. 2, 1373–6.

Byrne, Catherine, and Paul Foulkes. 2004. The 'mobile phone effect' on vowel formants. International Journal of Speech, Language and the Law 11.83–102.

Clark, Herbert H., and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. Cognition 84.73–111.

Clopper, Cynthia G., and David B. Pisoni. 2005. Perception of dialect variation. The Handbook of speech perception, ed. by David B. Pisoni and Robert E. Remez, 313–337. Oxford, UK: Blackwell.

Darby, John K. (ed.). 1981. Speech evaluation in psychiatry. New York, NY: Grune and Statton.

Douglas, John, and Mark Olshaker. 1995. Mind hunter. Inside the FBI's elite serial crime unit. New York, NY: Simon and Schuster.

Drygajlo, Andrzej. 2007. Forensic automatic speaker recognition. IEEE Signal Processing Magazine. March.132–5.

Ellis, Stanley. 1994. The Yorkshire ripper enquiry: Part I. Forensic Linguistics 1.197–206.

Eriksson, Anders, and Francisco Lacerda. 2007. Charlatanry in forensic speech science: a problem to be taken seriously. International Journal of Speech, Language and the Law 14.169–93.

Fisher, Jim. 1999. The ghosts of Hopewell: Setting the record straight in the Lindbergh case. Carbondale, IL: Southern Illinois University Press.

——. 2000. The Lindbergh case. New Brunswick, NJ: Rutgers University Press (paperback revision of the first 1987 edition).

Fitch, W. Tecumseh, and Jay Giedd. 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. Journal of the Acoustical Society of America 106.1511–22.

Foulkes, Paul, and Anthony Barron. 2000. Telephone speaker recognition amongst members of a close social network. Forensic Linguistics 7.180–98.

Foulkes, Paul, and Gerard Docherty. 2006. The social life of phonetics and phonology. Journal of Phonetics 34.409–38.

French, Peter. 1990. Analytic procedures for the determination of disputed utterances. Texte zur Theorie und Praxis forensischer Linguistik, ed. by Hannes Kniffka, 201–13. Tübingen, Germany: Niemeyer.

——. 1994. An overview of forensic phonetics with particular reference to speaker identification. Forensic Linguistics 1.169–81.

French, Peter, and Philip Harrison. 2006. Investigative and evidential applications of forensic speech science. Witness testimony, psychological, investigative and evidential perspectives, ed. by Anthony Heaton-Armstrong, Eric Shepherd, Gisli Gudjonsson, and Davis Wolchover, 247–62. Oxford, UK: Oxford University Press.

——. (eds.). 2007. Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. International Journal of Speech, Language and the Law 14.137–44.

French, Peter, Philip Harrison, and Jack Windsor Lewis. 2006. R vs John Samuel Humble: the Yorkshire Ripper hoaxer trial. International Journal of Speech, Language and the Law 13.255–73.

Gfroerer, Stefan. 2003. Auditory-instrumental forensic speaker recognition. Proceedings of EUROSPEECH 2003, Geneva, Switzerland, 705–8.

Goldman Eisler, F. 1968. Psycholinguistics: Experiments in spontaneous speech. London, UK: Academic Press.

Gonzalez-Rodriguez, Joaquin, Andrzej Drygajlo, Daniel Ramos-Castro, Marta Garcia-Gomar, and Javier Ortega-Garcia. 2006. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech and Language 20.331–55.

Hammersley, Richard, and J. Don Read. 1996. Voice identification by humans and computers. Psychological issues in eyewitness identification, ed. by Siegfried Ludwig Sporer, Roy S. Malpass, and Guenter Koehnken, 117–52. Mahwah, NJ: Lawrence Erlbaum Associates.

Hirano, M. 1981. Clinical examination of voice. Berlin, Germany: Springer.

Hollien, Harry. 1990. The Acoustics of crime. The new science of forensic phonetics. New York, NY: Plenum.

——. 2002. Forensic voice identification. San Diego, CA: Academic Press.

Hudson, Toby, Gea de Jong, Kirsty McDougall, Philip Harrison, and Francis Nolan. 2007. F0 statistics for 100 young male speakers of Standard Southern British English. Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, 6–10 August 2007, 1809–1812. http://www.icphs2007.de/conference/Papers/1570/1570.pdf

International Phonetic Association. 1999. Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet. Cambridge, UK: Cambridge University Press.

Jessen, Michael. 2007a. Forensic reference data on articulation rate in German. Science and Justice 47.50–67.

——. 2007b. Speaker classification in forensic phonetics and acoustics. Speaker classification I: Fundamentals, features, and methods, ed. by Christian Müller, 180–204. Berlin, Germany: Springer.

Jessen, Michael, Olaf Köster, and Stefan Gfroerer. 2005. Influence of vocal effort on average and variability of fundamental frequency. International Journal of Speech, Language and the Law 12.174–213.

Johnson, Keith. 1997. Speech perception without speaker normalization: an exemplar model. Talker variability in speech processing, ed. by Keith Johnson and John W. Mullennix, 145–165. San Diego, CA: Academic Press.

Kaplan, Ellen, and Michael Kaplan. 2006. Chances are . . . : Adventures in probability. New York, NY: Viking Books.

Keating, Patricia A. 1990. Phonetic representations in a generative grammar. Journal of Phonetics 18.321–34.

Köster, Olaf, and Jens-Peter Köster. 2004. The auditory-perceptual evaluation of voice quality in forensic speaker recognition. The Phonetician 89.9–37.

Köster, Olaf, Michael Jessen, Freshta Khairi, and Hartwig Eckert. 2007. Auditory-perceptual identification of voice quality by expert and non-expert listeners. Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Gertmany, 6–10 August 2007, 1845–1848, http://www.icphs2007.de/conference/Papers/1152/1152.pdf

Kreiman, Jody, Bruce. R. Gerrat, Kristin Precoda, and Gerald S. Berke. 1992. Individual differences in voice quality perception. Journal of Speech and Hearing Research 35.512–20.

Künzel, Hermann J. 1987. Sprechererkennung: Grundzüge forensischer Sprachverarbeitung. Heidelberg, Germany: Kriminalistik Verlag.

——. 1995. Field procedures in forensic speaker recognition. Studies in General and English Phonetics. Essays in Honour of Professor J. D. O'Connor, ed. by Jack Windsor Lewis, 68–84. London, UK: Routledge.

——. 1997. Some general phonetic and forensic aspects of speaking tempo. Forensic Linguistics 4.48–83.

——. 2000. Effects of voice disguise on speaking fundamental frequency. Forensic Linguistics 7.149–79.

——. 2004. Tasks in forensic speech and audio analysis: a tutorial. The Phonetician 90.9–22.
Ladefoged, Peter. 1971. Preliminaries to linguistic phonetics. Chicago, IL: University of Chicago Press.
Ladefoged, Peter, and Ian Maddieson. 1996. The sound of the world's languages. Oxford, UK: Blackwell.
Laver, John. 1980. The phonetic description of voice quality. Cambridge, UK: Cambridge University Press.
——. 1994. Principles of phonetics. Cambridge, UK: Cambridge University Press.
McDougall, Kirsty. 2004. Speaker-specific formant dynamics: an experiment on Australian English /ai/. International Journal of Speech, Language and the Law 11.103–30.
——. 2006. Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. International Journal of Speech, Language and the Law 13.89–126.
McDougall, Kirsty, and Francis Nolan. 2007. Discrimination of speakers using the formant dynamics of /u:/ in British English. Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, 6–10 August 2007, 1825–1828, http://www.icphs2007.de/conference/Papers/1567/1567.pdf
McGehee, Frances. 1937. The reliability of the identification of the human voice. Journal of General Psychology 17.249–71.
Meuwly, D. 2000. Voice analysis. Encyclopedia of forensic sciences, ed. by Jay A. Siegel, Pekka J. Saukko, and Geoffrey C. Knupfer, 1413–21. San Diego, CA: Academic Press.
Mokhtari, Parham, Tatsuya Kitamura, Hironori Takemoto, and Kiyoshi Honda. 2007. Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients. Journal of Phonetics 35.20–39.
Müller, Christian (ed.). 2007. Speaker classification vol. I and II. Berlin, Germany: Springer.
Nawka, Tadeus, and Lutz Christian Anders. 1996. Die auditive Bewertung heiserer Stimmen nach dem RBH-System. Stuttgart, Germany: Thieme (2 CDs).
Nolan, Francis. 1983. The phonetic bases of speaker recognition. Cambridge, UK: Cambridge University Press.
——. 1990. The limitations of auditory-phonetic speaker identification. Texte zur Theorie und Praxis forensischer Linguistik, ed. by Hannes Kniffka, 457–79. Tübingen, Germany: Niemeyer.
——. 1994. Auditory and acoustic analysis in speaker recognition. Language and the law, ed. by John Gibbon, 326–45. London, UK: Longman.
——. 1997. Speaker recognition and forensic phonetics. The handbook of phonetic sciences, ed. by William J. Hardcastle and John Laver, 744–67. Oxford, UK: Blackwell.
——. 2001. Speaker identification evidence: its forms, limitations and roles. Proceedings of the conference 'Law and Language: Prospect and Retrospect', 12–15 December 2001, Levi (Finnish Lapland). http://www.ling.cam.ac.uk/francis/LawLang.doc
——. 2003. A recent voice parade. International Journal of Speech, Language and the Law 10.277–91.
——. 2005. Forensic speaker identification and the phonetic description of voice quality. A figure of speech. A festschrift for John Laver, ed. by W.J. Hardcastle and J. Mackenzie Beck, 385–411. Mahwah, NJ: Lawrence Erlbaum Associates.
Nolan, Francis, and Catalin Grigoras. 2005. A case for formant analysis in forensic speaker identification. International Journal of Speech, Language and the Law 12.143–73.
Nygaard, Lynne C. 2005. Perceptual integration of linguistic and nonlinguistic properties of speech. The handbook of speech perception, ed. by David B. Pisoni and Robert E. Remez, 390–413. Oxford, UK: Blackwell.
Ormerod, David. 2001. Sounds familiar? – Voice identification evidence. Criminal Law Review 595–622.
Perfect, Timothy J., Laura J. Hunt, and Christopher M. Harris. 2002. Verbal overshadowing in voice recognition. Applied Cognitive Psychology 16.973–80.
Pfitzinger, Hartmut. 2002. Intrinsic phone durations are speaker-specific. Proceedings of the International Conference on Spoken Language Processing 2002, 2.1113–6, http://www.phonetik.uni-muenchen.de/~hpt/
Pisoni, David. 1997. Some thoughts on 'normalization' in speech perception. Talker variability in speech processing, ed. by Keith Johnson and John W. Mullennix, 9–32. San Diego, CA: Academic Press.

Przybocki, Mark A., Alvin F. Martin, and Audrey N. Le. 2007. NIST speaker recognition evaluations utilizing the Mixer corpora – 2004, 2005, 2006. IEEE Transactions on Audio, Speech, and Language Processing 15.1951–59.

Robertson, Bernard, and G.A. Vignaux. 1995. Interpreting evidence. Chichester, UK: Wiley.

Rose, Phil. 2005. Forensic speaker recognition at the beginning of the twenty-first century – An overview and a demonstration. Australian Journal of Forensic Sciences 37.49–72.

——. 2006a. Technical forensic speaker recognition: evaluation, types and testing of evidence. Computer Speech and Language 20.159–91.

——. 2006b. Accounting for correlation in linguistic-acoustic likelihood ratio-based forensic speaker discrimination. Proceedings of the Speaker and Language Recognition Workshop, 2006, Puerto Rico: IEEE Odyssey.

Rose, Phil, and Sally Duncan. 1995. Naïve auditory identification and discrimination of similar voices by familiar listeners. Forensic Linguistics 2.1–17.

Rose, Phil, Takashi Osanai, and Yuko Kinoshita. 2003. Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. International Journal of Speech, Language and the Law 10.179–202.

Rose, Philip. 2002. Forensic speaker identification. London, UK: Taylor and Francis.

Schiller, Niels O., and Olaf Köster. 1998. The ability of expert witnesses to identify voices: a comparison between trained and untrained listeners. Forensic Linguistics 5.1–9.

Shriberg, Elizabeth. 2007. Higher-level features in speaker recognition. Speaker classification I: Fundamentals, features, and methods, ed. by Christian Müller, 241–59. Berlin, Germany: Springer.

Shuy, Roger W. 2007. Language in the American courtroom. Language and Linguistics Compass 1.100–14.

Sporer, Siegfried Ludwig, Roy S. Malpass, and Guenter Koehnken (eds.). 1996. Psychological issues in eyewitness identification. Mahwah, NJ: Lawrence Erlbaum Associates.

Stevens, Kenneth N. 1998. Acoustic phonetics. Cambridge, MA: MIT Press.

