MODERN SCIENTIFIC EVIDENCE

The Law and Science of Expert Testimony

Volume 3

By

DAVID L. FAIGMAN
University of California Hastings College of the Law

DAVID H. KAYE
Arizona State University College of Law

MICHAEL J. SAKS
Arizona State University College of Law

JOSEPH SANDERS
University of Houston Law Center

ST. PAUL, MINN. WEST PUBLISHING CO.
2002

A. LEGAL ISSUES

§ 31-1.0 THE JUDICIAL RESPONSE TO PROFFERED EXPERT TESTIMONY ON TALKER IDENTIFICATION

§ 31-1.1 *Pre-Daubert* Decisions

Judicial opinions on the admissibility of talker identification were widely divided before *Daubert,l* and following *Daubert* there has been only one additional case that directly considered the admissibility of "voiceprints" or "voice spectrography," though there have been several cases on the periphery of the large central issue. Thus, no consistent or coherent judicial view can be discerned, and whether *Daubert* will guide courts to increased convergence must wait for the future. The patterns and non-patterns of the courts' responses to scientific talker identification is instructive. The accompanying Table 1, Scientific Talker Identification Cases, lists the major talker identification opinions in chronological order, along with certain other information
.about the cases.

Table I Scientific Voice Identification Cases: Holdings, Legal Tests, and Citations to NAS Report

Jurisdiction	Court	Case	Cite	Date	LeJ<8l Test*	Held	NAS Report
Military	APP	WriRht	17 CMA 183	1967	Reliability	IN	
CA	APP	King	72 Cal.Rptr. 478	1968	Frye-broad	oUT	
NJ	TR	Carv	239 A.2d 680	1970	Frye-broad	OUT	
MN	SC	Trimble	192 N.W.2d 432	1971	none	IN	
FL	APP	Worley	263 So.2d 613	1972	Reliabilitv	IN	
FL	APP	Alea	265 So.2d 96	1972	none + [Reliability]	IN	
CA	APP	Hodo	106 Cal.Rptr. 547	1973	Frye-narrow	IN	
US-DC Gir.	APP	Addison	498 F.2d 741	1974	Frye-broad	OUT	
CA	APP	Law	114 Cal.Rptr. 708	1974	Frve-broad	OUT	
US-EDPA	TR	Sample	378 F.SuPP. 44	1974	McC	IN	
MA	SC	Lvkus	327 N.E.2d 671	1975	Frve-narrow	IN	
US-4th Cir.	APP	Baller	519 F.2d 463	1975	McC	IN	
OH	APP	Olderman	44 OhioApp.2d 130	1975	Reliability	IN	
US-6th Cir.	APP	Jenkins	525 F.2d 819	1975	McC + [Reliabilitv]	IN	
US-6th Cir.	APP	Franks	511 F.2d 25	1975	McC + [Reliabilitv]	IN	
CA	SC	Kelly	130 Cal.Rptr. 144	1976	'Frye-broad	OUT	
US-DC Cir.	APP	McDaniel	538 F.2d 408	1976	Frve-broad	OUT	
NY	TR	Rogers	385 N.Y.S.2d 228	1976	McC + Rei + [Frye]	IN	
PA	SC	Topa	369 A.2d 1277	1977	Frve-broad	OUT	
MI	SC	Tobey	257 N.W.2d 537	1977	FrYe-broad	OUT	

§ 31-1.0

1, Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 113 S.Ct. 2786, 125 L.Ed.2d 469 (1993).

Jurisdiction	Court	ease	Cite	Date	Legal Test.	Held	NAB Reoort
US-8DNY	TR	Williams	443 F.Supp.269	1977	Reliab + [Frve-broadl	IN	
MD	se	Reed	391 A.2d 364	1978	Frve-broad	OUT	
US-2nd Cir.	APP	Williams	583 F.2d 1194	1978	McC	IN	
NJ	TR	D'Arc	385 A.2d 278	1978	Reliab or Frye-broad	OUT	
De	APP	Brown	384 A.2d 647	1978	[Reliability + Frve]'	nei-ther	
NY	TR	Collins	405 N.Y.S.2d 365	1978	Reliab + Frve-broad	OUT	
ME	se	Williams	388 A.2d 500	1978	Reliability-Relevancv	IN	
NY	TR	Bein	453 N.Y.S.2d 343	1982	Reliab + Frye-narrow	IN	no
IN	se	Cornett	450 N.E.2d 498	1983	Frve-broad	OUT	no
OH	se	Williams	446 N .E.2d 444	1983	Reliabilitv	IN	no
AZ	se	Gortarez	686 P .2d 1224	1984	Frve-broad	OUT	ves
RI	se	Wheeler	496 A2.d 1382	1985	McC	IN	no
LA	APP	Free	493 So.2d 781	1986	Relevancy balance	OUT	slightly
NJ	se	Windmere	522 A.2d 405	1987	Frve-broad	OUT	no
eo	se	Drake	748 P.2d 1237	1988	Frve-broad	OUT	no
US-7th Cir.	APP	Smith	869 F.2d 348	1989	Reliability + [Frve]	IN	slightl y
US-DHI	TR	Maivia	728 F.Supp.1471	1990	Reliab + Frye-narrow	IN	slightly
US-6th Cir.	APP	Leon	966 F.2d 1455 (table)	1992	McC	IN	no
AK	se	Coon	974 P.2d 386	1999	Daubert	IN	no

Note: The legal tests are abbreviated as follows: Frye with the relevant fields defined broadly (Frye-broad), or narrowly (Frye oarrow), reliability (reIiab) or relevancy (relev), or McCormick weighting (McC). The Court levels are: court of last resort (SC) imntermediate court of appeals (CA), trial (TR) Brackets indicate a test a court stated it was applying but where there is no indication in the opinion that the court actually applied that test.

First, we can see from the Table that the extent of agreement in recent years is no greater than in the earliest days of scientific talker identification. Of the first ten courts to consider the technique, six admitted it and four excluded it. The most recent ten to consider it were similarly divided, six for admission and four for exclusion.

Second, we can see that the legal test of admissibility applied by the courts is highly correlated with the holding.[2] Of those courts that applied the classical broad *Frye*[3] test-that is, an understanding of the relevant scientific community as consisting of a range of applicable fields[4] and not merely the one or two narrowly concerned with performing the particular application that constituted the technique at issue-not one admitted expert testimony of talker identification.[5] Of courts that employed a narrow *Frye* test-narrowing

2. .Notice that we merely say "correlated." We venture no guess as to whether the rule dictated the conclusion or vice-versa.

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

4. Concerning scientific talker identification, that could mean acoustical engineering, anatomy, electrical engineering, linguistics, phonetics, physics, physiology, psychology, physiology, and statistics-beca~se the technique of voice spectrography made assumptions about or borrowed principles from each of these fields.

5. People v. King, 266 Cal.App.2d 437, 72 Cal.Rptr. 478 (Cal.Ct.App.1968); State v. Cary, 99 N.J.Super. 323, 239 A2d 680 (N.J. Super. Ct. Law Div. 1968); United States v. Addison, 498 F.2d 741, (D.C.Cir.1974); People v. Law, 40 Cal.App.3d 69, 114 Cal.Rptr. 708 (Cal.Ct.App. 1974); People v. Kelly, 17 Cal.3d 24, 130 Cal. Rptr. 144, 549 P.2d 1240 (Cal. 1976); United States v. McDaniel, 538 F.2d 408 (D.C.Gir. 1976); Com. v. Topa, 471 Pa. 223, 369 A.2d 1277 (Pa.1977); People v. Tobey, 401 Mich. 141, 257 N.W.2d 537 <Mich. 1977); Reed v. State, 283 Md. 374, 391 A2d 364 (Md.1978);

the relevant scientific field to those that performed the test at issue-not one excluded the testimony.6 These two versions of the *Frye* test, and their predictably opposite conclusions, illustrate one of the important criticisms of *Frye,* namely, that defining the relevant scientific fields broadly or narrowly largely dictates the conclusion that will be reached.

Of courts that employed a "relevancy" or "reliability" test-frequently equated, at least in the past, with the test embodied in the Federal Rules of Evidence-eleven admitted7 talker identification expert testimony and three excluded it.[8] The one case that was decided after and under *Daubert* admitted voice identification expert testimony.[9] The courts varied considerably in what they required for the expertise to be found sufficiently "reliable." Most were satisfied that as long as there was something to be said on behalf of talker identification, that was enough to let it in. One court noted only that the witness was a credentialed expert and cited other jurisdictions that had admitted such testimony.[10] Using a similarly minimal threshold, however, another 'court excluded the evidence, concluding that its almost presumptive reliability was outweighed by its risk of being given excessive weight by factfinders.[11] Yet another court gave the scientific evidence on the proffered expertise a close and thoughtful examination, much like what the *Daubert* gloss on the Federal Rules would seem to require. That court concluded that talker identification expert testimony was inadmissible. [12]

People v. Collins, 94 M~c.2d 704, 405 N.¥.S.2d 365 (Sup. Ct. 1978); D'Arc v. D'Arc, 157 N.J. Super. 553, 385 A.2d 278 (N.J. Super. Ct. Ch. Div. 1978); Cornett v. State, 450 N.E.2d 498 (Ind.1983); State v. Gortarez, 141 Ariz. 254,686 P.2d 1224 (Ariz.1984); Windmere, Inc. v. International Ins. Co., 105 N.J. 373, 522 A.2d 405 (N.J.Super.Ct.App.Div.1987); People v. Drake, 748 P.2d 1237 (Colo.1988).

6. Hodo v. Super. Ct., Riverside County, 30 Cal.App.3d 778, 106 Cal.Rptr. 547 (Cal.Ct.App. 1973); Commonwealth v. Lykus, 367 Mass. 191, 327 N.E.2d 671 (Mass. 1975); People v. Bein, 114 Misc.2d 1021, 453 N.¥.S.2d 343 (Sup. Ct.1982); United States v. Maivia, 728 F.Supp. 1471 (D.C.Hawai'i 1990).

7. United States v. Wright, 37 C.M.R. 447 (1967); Worley v. State, 263 So.2d 613 (Fla. Dist.Ct.App.1972); State v. Olderman, 44 Ohio App.2d 130, 336 N.E.2d 442 (Ohio Ct. App. 1975); People v. Rogers, 86 Misc.2d 868, 385 N.¥.S.2d 228 (Sup.Ct.1976); United States v. Franks, 511 F.2d 25 (6th Cir.1975); United States v. Williams, 443 F.Supp. 269 (S.D.N.¥. 1977); State v. Williams, 388 A.2d 500 (Me. 1978); State v. Williams, 4 Ohio St.3d 53, 446 N.E.2d 444 (Ohio 1983); People v. Bein, 114 Misc.2d 1021, 453 N.¥.S.2d 343 (Sup. Ct. 1982); United States v. Smith, 869 F.2d 348 (7th Cir.1989); United States v. Maivia, 728 F.Supp. 1471 (D.C.Hawai'i 1990).

8. People v. Collins, 94 Misc.2d 704, 405 N.¥.S.2d 365 (Sup. Ct. 1978); State v. Free, 493 So.2d 781 (La.Ct.App.1986); D'Arc v. D'Arc, 157 N.J.Super. 553, 385 A.2d 278 (N.J. Super. Ct. Ch. Div. 1978).

9. State v. Coon, 974 P.2d 386 (Alaska 1999), discussed in some detail, *infra* § 1.2.

10. United States v. Smith, 869 F.2d 348 (7th Cir.1989).

11. State v; Free, 493 So.2d 781 (La. Ct. App.1986).

12. People v. Collins, 94 Misc.2d 704, 405 N.¥.8.2d 365 (Sup. Ct. 1978). Some excerpts from the opinion:

> It should be pointed out that although many ofthe Courts which admitted Spectrographic Voice Identification have done so based largely on the Tosi study, this study has not been replicated, and there seems to be no other formal experimentation in this area upon which the scientific community can make an informed judgment.
>
> It is certainly reasonable to expect science to withhold judgment on a new theory until it has been well tested in the crucible of controlled experimentation and study.
>
> [T]he entire technique is based substantially on the premise that intraspeaker variability is never as great as inter-speaker variability therefore, while each speaker's voice will be somewhat different each time he renders the same utterance, that difference will never be as great as the difference between the utter

Other courts employed the McCormick test, weighing the proffered evidence's scientific acceptability against the risks of opaqueness, error, or an exaggerated popular opinion of the technique. Every court employing this test found talker identification expert testimony admissible.[13]

Only one *post-Daubert* opinion exists, and that it discussed at length *infra* § 1.2.

Finally, one opinion reached its conclusion without employing a discerni ble legal test.[14]

The refusal of some courts to admit talker identification expert evidence is an exception to the traditional receptiveness of the courts to forensic individuation techniques. Why has talker identification been treated differently? Several interconnected explanations are plausible.

One may be that judges have gradually grown more thoughtful and discerning and less credulous about scientific offerings than their judicial ancestors' had been. Numerous courts evaluating talker identification expertise were critical of witnesses testifying on behalf of the technique who were mere technicians rather than educated scientists;[15] or whose livelihoods depended upon continued admission of the technique;[16] or who came from a very small circle of proponents of the technique.[17]

Another factor is that the literature of scientific talker identification, both supporting and questioning the technique, was more quantified and qualified1s than [18] earlier courts had received about earlier forensic individuation techniques. This is because most of the people involved in talker identification came from fields that had a tradition of empirical testing of their ideas. Indeed, more research was available to the courts about talker identification expertise than for any forensic individuation field that preceded it. This immediately provided the courts with unusual resources with which to comprehend the shortcomings of the technique.[19] When a field provides rigorous

ances of any two different speakers. It would seem reasonable to suppose that this is true, but this fact has not been proven to the Court's satisfaction.

The testimony however, reveals that there has. been no experimentation to show that two different voices will always appear different spectrographically.

Without additional independent proof the Court cannot accept the assumption that in. ter-speaker variability is always greater than intra-speaker variability.

13. United States v. Sample, 378 F.Supp. 44 (E.D.Pa.1974); United States v. Baller, 519 F.2d 463 (4th Cir.1975); United States v. Jenkins, 525 F.2d 819 (6th Cir.1975); United States v. Franks, 511 F.2d 25 (6th Cir.1975); People v. Rogers, 86 Misc.2d 868, 385 N.Y.S.2d 228 (Sup. Ct. 1976); United States v. Williams, 583 F.2d 1194 (2d Cir.1978); State v. Wheeler, 496 A.2d 1382 (R.I.1985); United States v. Leon, 966 F.2d 1455 (6th Cir.1992) (unpublished) .

14. State ex reI. Trimble v. Hedman, 291 Minn. 442, 192 N.W.2d 432 (Minn. 1971).

15. "[The expert witness's] qualifications are those of a techniciari and law enforcement officer, not a scientist." People v. Kelly, 17. Cal.3d 24, 130 Cal.Rptr. 144, 549 P.2d 1240 (Cal.1976).

16. [d. Compare this to the narrow version of the *Frye* test, which essentially asks the practitioners of a technique if they have sufficient confidence in their work that they should be allowed to continue to make a living at it.

17. Of course, these shortcomings do not distinguish talker identification from most other forensic individuation techniques when they were gaining admission to the courts. Indeed, all b.ut the third criticism continues to be true for them.

18. In the sense of limited, restricted, cir cumspect.

19. The same was true for DNA typing, and was not true for most other forensic individuation techniques.

self-critiques of its own concepts and techniques, it greatly aids the courts in making a more informed and sober assessment of the field and its likely contribution to the factfinding process.[20] Moreover, controversy tends to precipitate still more research, and a greater volume of research tends to produce a more complex and skeptical impression of the technique in the mind of the court.[21]

In the face of actual data, the courts had a real choice to make. Although the technique could reduce uncertainty in identification, it also was less than perfect. Errors were going to be made, and, unlike some other fields of forensic individuation, talker identification proponents said so.[22] The courts had concrete error rates to evaluate. How good is good enough? How much error is too much? The law provides no standards for making that assessment. Ten percent error may have been viewed by some courts as quite adequate and by other courts as not nearly good enough.

Finally, the courts may have been overwhelmed by the studies. Although more research means a greater potential to understand the scientific questions at issue, it also may have confused some courts, which had limited capacity to interpret and evaluate the empirical studies. If this was the problem, help was on the way.

Unique assistance in evaluating the available data came into being only a decade after talker identification made its first appearance in the courts.[23] Help came in the form of a careful review of scientific talker identification by the National Academy of Sciences.[24] A panel of highly knowledgeable scientists and other experts from diverse relevant fields carefully reviewed the relevant scientific literature and concluded:

> [The assumption] that intraspeaker variability is less than ... interspeaker variability ... is not adequately supported by scientific data.
>
> Estimates of error rates now available pertain to only a few of the many combinations of situations encountered in real-life situations. These estimates do not constitute a generally adequate basis for a judicial or

20. When other fields lack such critiques, is that because there is nothing to question? Or because an uninformed and unquestioning consensus developed among members of the field? And how can courts distinguish between the two possibilities?

21. This presents a paradox: All else equal, it appears that the better a field studies and critiques itself, the more skeptical the courts seem likely to be of it. The less a field tests its ideas and the more confidently it asserts them, the more positive an impression the courts develop of the field. For a number of the more conventional forensic individuation techniques, there still is no tradition of self-scrutiny or a literature reporting the results of rigorous testing which can inform the courts. At least in terms of their continued acceptance by the courts, those fields have nothing to gain and much to lose by adopting a tradition of inquiry, testing, and skepticism.

22. "Possibly, no combination of methods may ever produce-absolutely positive identification or eliminations in 100% of the cases submitted." Oscar Tosi, *The Problem of Speaker Identification and Elimination, in* Measurement PROCEDURES IN SPEECH, HEARING. AND LAN. GUAGE 399, 428 (Sadanand Singh ed., 1975).

23. Up until that time. There have been two NAS panels formed to review the data on the technique of DNA typing. *See* Chapter 25.

24. The NAS was created during the administration of Abraham Lincoln to provide any agency of the federal government with first rate scientific advice on issues of concern to those agencies. In this instance, the FBI made the request for a review.

legislative body to use in making judgments concerning the reliability and acceptability of aural-visual voice identification in forensic applications.[25]

Upon publication of the Report, the FBI ceased performing talker identification for the purpose of offering testimony in Court,[26] and it was expected[27] that the courts would stop admitting talker identification expert testimony, at least until the scientific support for it improved.[28] However, of the 12 judicial opinions written since release of the NAS Report,[29] seven admitted the expert testimony while five excluded it. Still more curious, only four cite the Report at all and only one seems to have actually read and learned what the Report had to say. Thus, for the most part, the courts decided the post-NAS cases as if the NAS Report did not exist. [30]

§ **31-1.2** State Decisions *Post-Daubert*

Only one case by a court following *Daubert* has considered the admissibility of expert evidence using voice spectrography. That case, *State v. Coon* (1999),[31] is also the case through which Alaska adopted *Daubert* as its state law.

The defendant in this case was accused of making terroristic telephone calls to the husband of his ex-daughter-in-law. Part of the evidence introduced against him was expert testimony based on voice spectrograph comparisons. The trial court had held this evidence admissible under *Frye* as "generally accepted by courts," and the jury had found the defendant guilty. On appeal Alaska's intermediate appellate court held that the support for admission under *Frye* was inadequate, and remanded for further proceedings on the admissibility issue. The State appealed to the Alaska Supreme Court, which retained jurisdiction but remanded for findings under both *Frye* and *Daubert.* In its decision, the Alaska Supreme Court explicitly adopted *Daubert,* adopted a deferential standard of review, and held the voice spectrograph evidence admissible under the *Daubert* test.

Query whether, when making rulings on the admissibility of scientific evidence as a general matter (that is, whether the science is sufficiently dependable to be admitted, not whether it has sufficient fit to the facts of the particular case at bar), the trial court is in a better position to make the decision than an appellate court. Is verbal testimony by a few witnesses (the typical mode of information gathering by a trial court) a more or a less illuminating method of learning about the underlying basis of the expertise than reading the relevant research literature, with the guidance of counsel in the form of briefs and arguments (the mode of information gathering more

25. BOLT ET AL., ON THE THEORY AND PRACTICE OF VOICE IDENTIFICATION (1979).

26. But, as with the polygraph, they continued doing voice spectrographic tests for investigative purposes.

27. See ANDRE A MOENSSENS ET AL., 8cIENTI. FIC EVIDENCE IN CIVIL AND CRIMINAL CASES 645 (4th ed.1995).

28. Few if any of the scientific shortcomings raised by the Report have been solved by subsequent research. *See* discussion *infra* § 2.4.

29. See Table 1.

30. Whether this reflects the shortcomings of counsel (for not drawing the courts' attention to the NAS study) or the courts (for not finding it themselves, or not appreciating its value to their decisions), we are unable to say.

31. 974 P.2d 386 (Alaska 1999).

I

often used by an appellate court).[32] The Court suggested that the main advantage of a deferential 'Standard of review lies in the notion that a trial court would have at its disposal more up-to-date information than an appellate court could.[33]

If the Alaska Supreme Court believed that the trial court was in a better position to gather the evidence, why didn't it make the *Daubert* versus *Frye* decision, remand for the trial court to complete the case consistent with that holding, and let that specific admissibility decision be appealed if and when the parties chose to do so? Since the Supreme Court reviewed the trial court ruling on admissibility for abuse of discretion following the United States Supreme Court's opinion in *Joiner,*[34] (and ruled that the trial court's conclusions were "not an abuse of discretion"), does that mean that the admissibility of voice spectrographic evidence is not settled as a matter of precedent in Alaska, and that the State's trial courts are free to make contrary decisions when the same question of admissibility presents itself in future cases, so long as they make their rulings under the *Daubert* test? Apparently so. From the opinion it appears that the Alaska Supreme Court expects trial courts to make these decisions case-by-case, to contradict each other from time to time, and to be reviewed for abuse of discretion-yet the court hints that somehow (notwithstanding the announced rule) appellate courts will resolve the contradictions before .they became an embarrassment, and that in any event the court did not expect this problem to occur very often, The court justifies its approach in part by treating all applications of science as so highly case specific, that the contradictions will be attributable to differences in the case facts.

Oddly, the opinion relied on Rule 703, rather than 702, as the foundation for its *Daubert* analysis, noting that the "commentary to the Alaska Rules of Evidence provides support for the State's view that ... Rule 703 is also a source for an approach broader than the Frye standard."[35] The basic points the Court makes about the dependability of scientific knowledge are entirely reasonable, but finding them in Rule 703 makes little sense. Rule 703 pertains to the facts or data relied on in the *particular* case (that is, the adjudicative

32. For an analysis of the special problems scientific evidence presents to determining the proper standard of review, *see* Chapter 1.

33. In the present case, this clearly is not what happened. As the opinion states, ". no scientific literature was submitted to the trial court for review, but [the voice identification expert) testified about several articles and studies addressing voice spectrographic analysis, and conceded that the reliability of the technique was disputed among members of the relevant scientific community." *Coon,* at 402. A visit to a library by a judicial clerk could unearth a far more complete review of the relevant scientific research than the selective, self-serving, and, in this instance, out-of-date sampling of research literature referred to verbally from the witness stand.

34. General Electric Co. v. Joiner, 522 U.S. 136, 118 S.Ct. 512, 139 L.Ed.2d 508 (1997). The Alaska Supreme Court adopted that same position, with one of the four justices dissenting. The dissent emphasized the trans-case nature of scientific evidence, in contrast to the usual adjudicative evidence whose admissibility is being ruled upon. For further discussion of this problem, *see* Chapter 1.

35. Coon, 974 P.2d 386 (Alaska 1999). Alaska Rule of Evidence 703 provides: "The facts or data in the particular case upon which an expert bases an opinion or inference may be those perceived by or made known to the expert at or before the hearing. Facts or data need not be admissible in evidence, but must be of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject."

facts), not the general scientific background being relied upon (more akin to legislative facts, or empirical authority) and the methods by which the expert may come into possession of those case-specific facts. In addition, query whether *Daubert* really is "broader" than *Frye.36* At the same time, the opinion clearly recognizes that its adoption of *Daubert* would lead both to admitting previously inadmissible evidence and excluding previously admissible evidence (and therefore in some situations *Daubert* is "narrower" than *Frye).* The court rejected a number of arguments against the adoption of *Daubert.* It rejected the argument that *Daubert* would place too heavy a burden on trial courts, noting that courts can obtain help by appointing their own experts under Rule 706. It also rejected concerns about adversely affecting the admissibility of traditional forensic evidence like fingerprinting, hand.: writing, and hair comparison analyses,37 and about opening the doors to "junk science."38

In examining the evidence underlying the claims of voice spectrographic identification, the Alaska Supreme Court conducted a limited and. superficial review of the research on which such a decision must depend, doing little more than quoting the trial court's conclusory assertions.39 Given that no research literature was "submitted" to the trial court, and that court did not ask for any or do any research on its own, unless the court recognized any duty to look beyond the four corners of the record from the trial's hearing on the issue, then by definition the supreme court's review will be limited to the limited review of the science coriducted below. As noted above and in the original chapter, few courts have cited the National Academy of Sciences' authoritative review of voice spectrography research, the findings of which led the FBI to withdraw from offering such evidence in courts. *State v. Coon* joins that list of cases that overlooked the major scientific review of the question before them. Thus, despite the *Coon* court's own discussion of the heightened analysis of the science that is called for under a *Daubert* review, its own first outing offers a review of the scientific claims, and a review of the adequacy of the trial court's gatekeeping, that is remarkably meager.

§ 31-1.3 Federal Decisions *Post-Daubert*

No cases involving disputes over "classical" voice spectrography have been reported from the federal courts subsequent to *Daubert.* But other types of voice identification expertise and some more peripheral issues were discussed and debated.

The defendant in *United States v. Salimonu*[40] was found guilty of importing heroin. Among the issues he raised on appeal was the trial court's decision

36. Recent judicial experience and scholarly analysis have eroded that simple equation. *See* Chapter 1.

37. Consult the appropriate chapters in this treatise to see how those asserted expertises have fared,. or are expected to fare, under a *Daubert* analysis.

38. Notice that these two arguments-that *Daubert's* standard is so low that it will lead to the admission of junk science and so high that it will exclude forensic science-are at war with each other. They cannot both be true.

39. The opinion gives a more detailed recitation of the expert's background and experience than it does the data on the underpinnings of the technique at issue (for which any facts about the particular expert are irrelevant).

40. 182 F.3d 63 (1st Cir.1999).

to exclude expert testimony that the voice on the inculpatory tape recordings was not his. The First Circuit affirmed. The trial judge had admitted defense testimony about voice spectrographs, but excluded a linguist's testimony that was based on simply listening to the tapes in question. This expert "admitted that he had no training or special certification in voice identification or comparison, and that he had only engaged in voice recognition procedures two or three times before." Moreover, he "knew of no studies to determine the rate of error for this kind of identification," and conceded that a lay person would be able to discern the same differences between the tapes that he heard.

The defendant in *Virgin Islands v. Sanes*[41] was convicted of robbery and rape. Part of the evidence introduced against him at trial was the victim's identification of his voice. She selected his voice from recordings of several voices. The defendant sought to introduce the testimony of an expert who would have testified about why voice identification is not as accurate as eyewitness identification. The expert was not allowed to give this testimony, but was allowed to testify regarding the distinguishing characteristics of the defendant's voice. With little analysis, the Third Circuit held the trial court had not abused its discretion. Concerning research relevant to the scientific issues in this case, see the discussion of Earwitness Research.[42]

The defendant in *United States v. Jones*[43] had been convicted for distributing cocaine. On appeal, he argued that the trial court had improperly excluded expert testimony on voice identification. The trial court had applied the *Frye* Test, but the Ninth Circuit found that even under *Daubert* the evidence should have been excluded. The expert had developed his voice comparison technique himself, and could not cite any scientific basis for it. He conceded that no scientific studies or published research supported his theory.

The defendant in *United States v. Drones,*[44] sought, and had obtained from the district court, relief for his claim of ineffective assistance of counsel on the grounds that his attorney had failed even to investigate the availability of voice identification expert testimony to evaluate a tape that the government asserted contained the defendant's voice. The court of appeals reinstated the state court verdict. At the habeas hearing, the petitioner's expert stated that he found from his examination that there was "probably elimination" of the defendant as a source of the voice on the recording. According to the expert, this meant that "80% of the comparable words in the samples were dissimilar aurally and spectrographically." The petitioner's expert conceded, however, that there were sundry weaknesses with this technology and that no objective criteria existed by which to check the accuracy of any conclusions an examiner might reach. Also testifying at the hearing, the government's expert echoed these cautionary words, noting that very little research had been done to validate the courtroom use of this technology. The court of appeals concluded that voice identification expertise is not competent evidence. "Given the current state of the law regarding the admissibility of expert voice

41. . 57 F.3d 338 (3d Cir.1995). *42. See infra* § 2.2.5.

43. 24 F.3d 1177 (9th Cir.1994).

44. 218 F.3d 496 (5th Cir.2000).

identification testimony and the expert testimony presented at the evidentiary hearing, we cannot say that counsel's choice of strategy was unreasonable and therefore deficient."

One of the defendants in *United States v. Bahena,*[45] complained that the district court erred in excluding his expert on voice spectrography. The appellate court rejected this argument, and affirmed the convictions of all of the defendants. The court of appeals found that the lower court had not abused its discretion in excluding this particular witness, noting that the expert here had no college degree, was not a member of any professional association and was not familiar with the standard practices in the field of voice identification.

B. SCIENTIFIC STATUS

by

Raymond D. Kent* & Michael R. Chial**

§ 31-2.0 THE SCIENTIFIC BASIS OF EXPERT TESTIMONY ON TALKER IDENTIFICATION

§ 31-2.1 Introductory Discussion of the Science

§ 31-2.1.1 The Scientific Questions [1]

Terminology and Basic Concepts

Most people can easily recognize family members, friends, coworkers, and popular figures from the sounds of their voices. This familiar form of personal identification finds forensic application in situations where a voice has been heard by a witness or, even better, a recording has been made of the voice in question. Talker identification may be broadly defined as a decision-making process that relies on properties of the talker's speech signal. The decision maker's objective is to identify an individual by the characteristics of that individual's speech. The term *talker identification* is used in this chapter

45. 223 F.3d 797 (8th Cir.2000).

* Raymond D. Kent is Professor of Speech Science in the Department of Communicative Disorders, University of Wisconsin-Madison. His doctorate is from University of Iowa and he did postdoctoral work in speech analysis and synthesis at the Massachusetts Institute of Technology. He has edited or written eleven books, including THE ACOUSTICAL ANALYSIS OF SPEECH (with Charles Read, 1992), and is past editor of the JOURNAL. OF SPEECH AND HEARING RESEARCH. He holds an honorary doctorate from the University of Montreal Faculty of Medi
cine, is a Fellow of the Acoustical Society of America, the International Society of Phonetic Sciences, and the American Speech-LanguageHearing Association, and has earned Honors of the American Speech-Language-Hearing Association.

** Michael R. Chial is Professor of Audiology in the Department of Communicative Disorders at the University of Wisconsin-Madison. His doctorate is from the University of Wisconsin-Madison. For 20 years he has worked with the American National Standards Institute and is currently working with the Audio Engineering Society to develop technical standards for forensic applications of audio recordings. He is past associate editor. of the JOURNAL OF ‚SPEECH AND HEARiNG RESEARCH, and Fellow of The American Speech-Language-Hearing Association and the American Academy of Audiology.

The authors thank Lonnie L. Smrkovski for his comments on an earlier draft of this chapter. The opinions expressed herein are solely those of the authors.

because it denotes the task of trying to identify a human talker. Other terms used for this application are *speaker identification* and *voice identification.*

Traunmuller1 listed four kinds of information contained in the speech signal:

1. *Phonetic quality* refers to the linguistic content of the spoken message, i.e., the essential material from which we derive the information intended by the talker.

2. Affective quality is paralinguistic information, meaning that it accompanies the linguistic message of speech and may contribute to the interpretation of that message. Emotional attributes fall into this category.

3. Personal quality is extralinguistic, meaning that it is outside the ordinary linguistic aspects of speech. Personal quality is informative about the talker, but not the message. The information can include the talker's gender, age, state of health, and individual characteristics.

4. Transmittal quality gives perspectival information about the talker's location, including the distance from the one who hears the signal, orientation in space, presence of background noise, and influence of environmental acoustics that may introduce effects such as reverberation.

Talker identification rests on the assumption that intratalker variability (e.g., the variability associated with multiple productions of the same speech sample by a given talker) is less than intertalker variability (e.g., the variability associated with productions of the same speech sample by different talkers). The capability of recognizing a talker is based on two primary sources of intertalker differences: (1) anatomic differences in the size and shape of the speech organs, and (2) subtle individual differences in how speech sounds are made. The former are sometimes called *physiological differences* and the latter *behavioral differences.* Physiological differences generally are not subject to learning effects, whereas behavioral differences often are. A hardware software analogy also has been used to distinguish these two types of differences among talkers,[2] with physiological factors being compared with the hardware and behavioral factors (including sociolinguistic and psycliological factors) with the software. Presumably, the hardware is less easily altered than the software. The speech pattern produced by anyone individual is a combination of physiological and behavioral factors. Differences among talkers are therefore a combination of the same factors.

Talker recognition may be subdivided into various approaches: talker recognition by listening (aural recognition), by machine (automatic recognition), and by visual inspection of spectrograms (also known as "voiceprints" or "voicegrams"). These are not necessarily mutually exclusive procedures. Forensic applications commonly make use of both aural recognition and

§ 31-2.0

1. Hartmut Traunmuller, *Conventional, Biological, and Environmental Factors in Speech Communication: A Modulation Theory,* 18 PERILUS (PHONETIC EXPERJMENTAL RESEARCH, IN. STITUTE OF LINGUISTICS. UNIVERSITY OF STOCKHOLM) 1 (1994).

2. Hisao Kuwahara & Y oshinori Sagisaka, *Acoustic Characteristics of Speaker Individuality: Control and Conversion,* 16 SPEECH COMMU. NICATION 165 (1995).

spectrograms, and it is possible to use all three methods in reaching a decision.

This chapter concentrates on the third approach, visual inspection of spectrograms, but some comments will be included on the first and second approaches as well. Visual inspection of spectrograms is the major source of evidence provided by trained examiners. The overarching scientific question is whether an individual talker can be distinguished from a larger group of talkers on the basis of visual patterns in a spectrogram. Because the properties of the spectrogram are essential to an understanding of their use in talker identification, some general comments on spectrograms are in order.

Spectrogram is a generic term for the conventional analysis of sound according to the three dimensions of frequency, time, and intensity. In the typical spectrogram, time is represented along the horizontal axis, frequency (the rate of vibration of a sound component, heard as pitch) along the vertical axis, and intensity (magnitude of a sound component, heard as loudness) as a gray (or darkness) scale. An example of a spectrogram is shown in Figure 1. These visual patterns were introduced as a practical laboratory technique in the 1940s and have been a major source of information in the study of speech. Terms synonymous with *spectrogram* are *voiceprint, voicegram,* and *Sonagram* TM.

The term "voiceprint" was coined by Gray and Kopp3 and reintroduced by Kersta4, an early proponent of talker recognition through comparisons of visual patterns. Some writers viewed the "voiceprint" as analogous to the "fingerprint." The term "voicegram" was substituted for voiceprint by others who believed that the term *voiceprint* could be misleading. Whereas a finger can leave a direct physical impression when it is pressed against a surface (hence leaving a genuine "print") the voice is given visual representation only by a series of transformations in which acoustic energy is eventually represented on papers The term *voicegram* is preferable to *voiceprint* though both have the technical disadvantage of emphasizing "voice" rather than "speech." Although voice certainly is important as the primary energy source of speech, speech is really more than voice. This is one reason why the term *talker identification* is used in preference to *voice identification* in this chapter. Energy from the voice is modified by the speech organs through resonance and other influences. Talker identification generally relies on patterns of speech including characteristics of voice, resonance, and articulation.6 *Speaker* identification *is* a *frequently used term, which may be gaining prominence.* 7

3. C. H. Gray & G. A. Kopp, *Voiceprint Identification,* BELL TELEPHONE LABORATORIES RE. PORT 1 (1944).

4. Lawrence G. Kersta, *Voiceprint Identifi cation,* 196 NATURE 1253 (1962).

5. "Voiceprint" also was used as a trademark by Voiceprint Laboratories, Inc. a manufacturer of speech spectrography equipment. A successor fllm, Voice Identification, Inc., retains rights to that trademark and manufactures an analog device (Model 700) favored by some forensic practitioners.

6. The term *Sona-gram* is a trademark of a manufacturer (Kay Elemetrics Corporation) that currently markets two digital spectro graphs-the Model 5500 (a dedicated, standalone device) and the model 4300B (designed for use with general-purpose..personal computers). A number of other sys~ms, especially computer programs designed for clinical research and treatment, geophysical. and bioacoustical research, music and audio engineering purposes, produce spectrograms as one analysis alternative. *See* Charles Read et al., *Speech Analysis Systems: A Survey,* 33 J. SPEECH & HEARING RESEARCH 363 (1990); Charles Read et al., *Speech Analysis Systems: An EValuation,* 35 J. SPEECH & HEARING RESEARCH 314 (1992).

7. A disadvantage to this term is that the word *speaker* has two prominent meanings,

[a] Instrumentation and Display

Instruments for this type of speech analysis differ in several respects, but all include. components that acquire sound (generally via microphones or audio recorders), edit stored signals, analyze sounds to produce spectrograms, and display results (analog units by means of facsimile technology; digital units by means of video monitors and laser or video printers). Because several systems can be used to make spectrograms, it is reasonable to ask if there are any differences among them that should be considered in the accuracy of talker identification. Unfortunately, very few comparisons of this type have been made, but" Hazen[8] reported no differences between the Voiceprint Laboratories 4691C Sound Spectrograph and the Kay Elemetrics Corporation 6061A Sonagraph (neither of these analog devices is now manufactured). While contemporary digital instruments offer greater flexibility and precision than earlier analog machines, the older devices produced hard-copy records of superior resolution. It has not been studied whether important differences exist among the various devices and computer systems currently used to make spectrograms, a problem complicated by the lack of appropriate recorded reference material (speech and speech-like signals) designed to compare alternative systems.

In this chapter, the term *spectrogram* will be used in the broad sense to include all varieties of visual displays of speech that rely on a conventional three-dimensional analysis of time, frequency and intensity.[9] In customary practice, two types of spectrograms have been used: wide-band and narrowband. These two types are distinguished by the width of the analyzing filter which can result in different kinds of spectrograms. The bandwidth of the analyzing filter can be likened to a kind of acoustic "window" that is passed along the signal to determine the energy in various frequency regions. The narrower the bandwidth of the analyzing filter, the better the resolution of frequency but the poorer the resolution of time. Briefly, the wide-band spectrogram uses either a 250 or 300 Hz bandwidth filter and IS especially useful for speech analysis because it reveals certain acoustic features that have been important in distinguishing among various types of sounds and among different talkers. In particular, the wide-band spectrogram usually is effective in displaying *formants* (acoustical energy constrained to frequency

one being a human talker and the other being an electroacoustic..device such as a loudspeaker.

8. Barry Hazen, *Effects of Differing Phonetic Contexts* on *Spectrographic Speaker Identification,* 54 J. ACOUSTICAL SOC'Y AM. 650 (1973).

9. Time is represented from left to right. Frequency is the' rate of vibration of a sound stimulus and is expressed in the unit of hertz (Hz), which is the number of cycles of vibration per second. The acoustic energy in adult male speech is essentially contained in a frequency range (bandwidth) of about 50 to 8,000 Hz. However, speech can be understood even with much narrower bandwidths. For instance, telephone bandwidth is on the order of 3,000 Hz (500 to 3,500 Hz). The greatest concentration of speech energy for adult male voices is in the range of about 100 to 2500 Hz. Intensity is one measure of the magnitude or strength of sound energy. It is usually expressed in decibels (dB), a logarithmic scale.

regions by vocal tract resonances). The wide-band spectrogram is particularly useful for the analysis of highly dynamic signals such as speech.1o Figure 1 illustrates a wide-band speech spectrogram of the utterance "I said stop"; Figure 2 notes landmarks typical of speech sounds in the word "stop." The horizontal axis of both panels is time and the vertical axis is frequel1CY in Hertz (Hz) or cycles per second. The third dimension of the spectrogram is intensity: darker areas represent greater intensity.

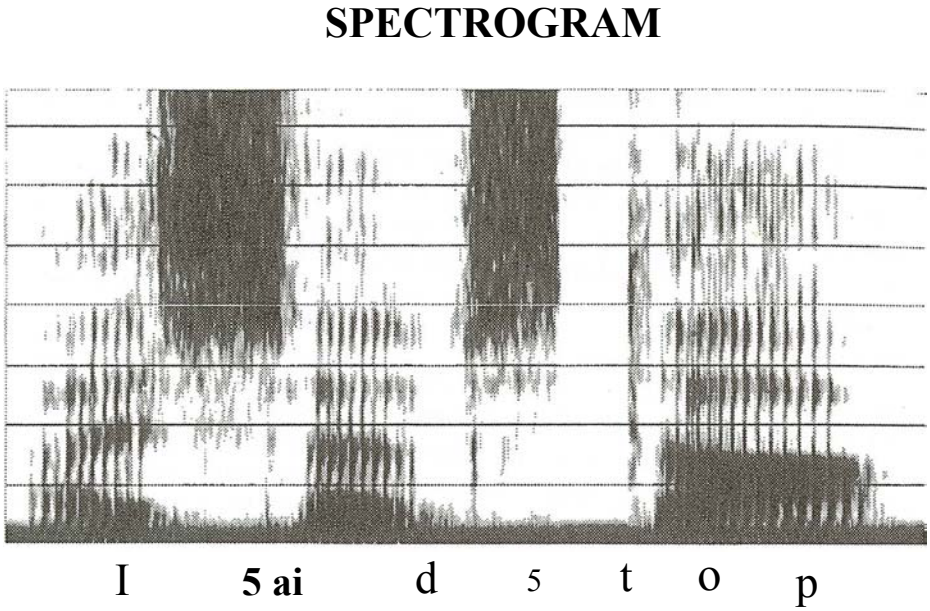

Figure 1. A spectrogram of the utterance, "I say stop." Time is represented on the horizontal axis, frequency on the vertical axis, and intensity as variations in darkness. The horizontal lines indicate frequency intervals of 1 kHz (1000 Hz).

10. Roel Smits, *Accuracy of Quasistationary Analysis of Highly Dynamic Speech Signals, 96* J. ACOUSTICAL SOC'Y AM. 3401 (1994).

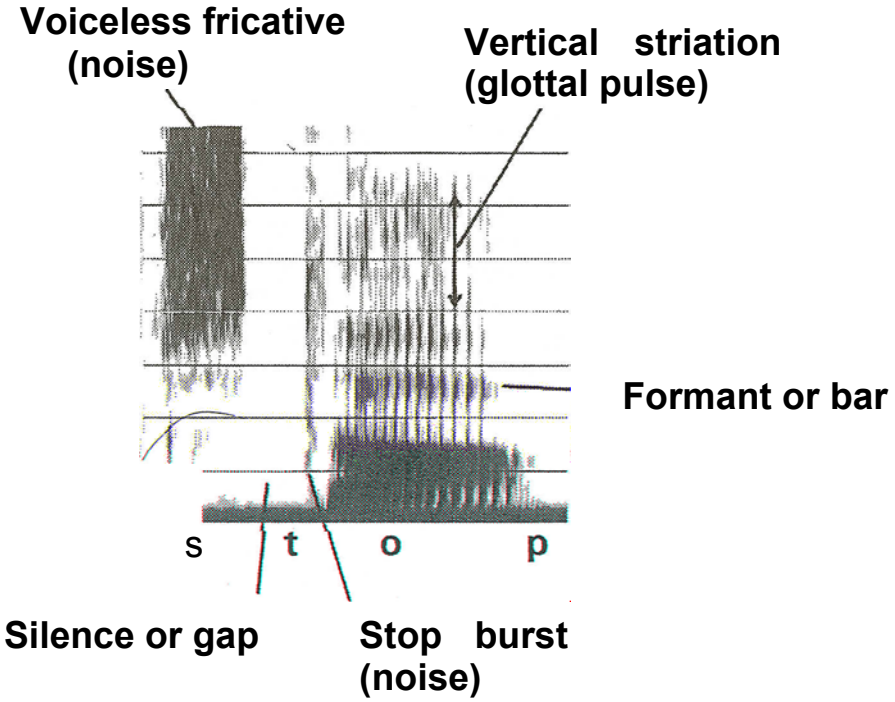

Figure 2. A spectrogram of the word "stop" from Fig. 1, labeled with some acoustic features that might be used in talker identification by spectrogram.

Most of the published scientific papers are based on wide-band analyses of selected words produced by adult male talkers. In this chapter, the wide-band analysis is assumed unless otherwise noted. This is also the common form of analysis used in talker identification from the visual examination of spectrograms. An advantage of many of the contemporary computer-based speech analysis systems is that they offer a range of choices of analysis filter bandwidths. The clarity of formants is influenced by interactions between bandwidth and the frequency characteristics of the speech to be analyzed. Conventional filters 250 or 300 Hz in width are generally suitable for analyzing the speech of adult males, but other bandwidths may be preferable for certain groups, including women, children, some adolescent males, and men with unusually high vocal fundamental frequency (the physical attribute most closely related to what we hear as vocal pitch). Scientific interest in the features of formants that distinguish males, females and children began in the 1940s and continues to the present.[11] Buder,[12] for example, has summarized

11. W. Keonig et aI., *The Sound Spectrograph,* 17 J. ACOUSTICAL SOC'y AM. 19 (1946); Gordon E. Peterson & Harold L. Barney, *Control Methods f,lsed in the Study of the Vowels,* 24 J. ACOUSTICAL SOC'y AM. 175 (1952); James Hillenbrand et aI., *Acoustic Characteristics of American English Vowels,* 97 J. ACOUSTICAL SOC'y AM. 3099 (1995).

12. Eugene H. Buder, *Acoustic Analysis of Voice Quality: A Tabulation of Algorithms 1902-1990, in* VOICE QUALITY MEAsUREMENT (Raymond D. Kent & Martin J. Ball eds., 2000).

nearly 90 years of quantitative descriptions of vocal quality. These include simple statistical summaries of fundamental frequency and voice amplitude long and short term perturbations and covariations in both, and various spectral measures. While many of these pertain to clinical voice disorders some have been incorporated into currently available instrumentation used in forensic practice. Differences between controlled clinical recording environments and those common to forensic practice may limit application of recent innovations in measurement, but some may prove useful. The essential point is that characteristics of the laboratory analysis should be selected to match talker characteristics in ways that recognize the limitations of recording methods to capture those characteristics.

[b] Decision Objectives

Talker recognition embraces procedures with different decision objectives and assumptions. These include *talker verification* (or authentication), *talker identification,* and *talker elimination.* Talker verification is a test of an identity claim in which a speech sample from an individual is compared to a stored reference sample previously obtained from the individual whose identity is claimed. A common application is security access. A person making the identity claim will be granted access if this person's speech sample is a satisfactory match to a stored pattern. Talker identification is a decision process in which an utterance from an unknown speaker is attributed to one speaker in a known population, such as employees in a high-security facility. Talker elimination is the inverse process of deciding that an utterance from an unknown talker cannot be attributed to a particular speaker in a known population. Most forensic applications involve speaker identification or elimination and these will be the central issues in this chapter.

Talker identification and elimination can be studied experimentally using three different comparison procedures: *closed-set, open-set,* and *discrimination.* Different patterns of correct and incorrect decisions are possible for each procedure. In the closed and open procedures, a sample (exemplar) of speech from an unknown talker (D) is compared to exemplars from some number (N) of talkers whose identities are known (K). Normally, each known talker (K_n) is represented by a single exemplar-in other words, the known talkers are independent of each other. If one of the known talkers (assume it is K_3) is indeed the same as talker D, and if (prior to the experiment) the examiner is informed a match exits, then the procedure is closed. The procedure is open if the examiner is told that the population of known talkers mayor may not actually include talker U. A common experimental strategy is to organize comparisons of unknown and known exemplars as pairs following the form: U vs. K_1, U vs. K_2, U vs. K_N Assuming the examiner is required to consider each pair only once, a total of N paired comparisons is possible. .Each pair-wise comparison is constrained to one of two decisions: $U = K_N$, (a claimed identification) or $U \neq K_N$, (a claimed elimination). Examiner claims are compared to the actual configuration of pairs (known to the experimenter, but not the examiner), If a claimed identification is wrong, the decision is called a false identification or false positive. If a claimed elimination is wrong, the decision is called a false elimination or false negative.

In the closed-set procedure, the examiner is asked *which* known exemplar matches the unknown sample. Only one claimed identification ($U = K_3$ in this example) can be correct and no more than N-l claimed eliminations can be correct. There can be only one false elimination because the pairing of U vs. K_3 occurs only once. Up to N-l false identifications are possible, but because the examiner knows that only one match exists, and because at least one identity claim must be made, the closed-set procedure effectively limits the possible number of false identifications to one.

In the open-set procedure, the examiner is asked *whether* one of the known samples matches the unknown exemplar and (if so), *which* one. If the target is included among the known exemplars, there may be one correct identification

N-l correct eliminations, one false identification, and one false elimination. In the open procedure with the target talker (K_3 in this example) absent, however, there can be as many as N correct eliminations and one false identification. There can be no correct identification and no false elimination because the target talker is not available for comparison. Comparison of results from open-and closed-set procedures allow study of examiner preferences for claims of elimination or identification, as well as the impact of the spectrographic cues available to the examiner upon decision-making behavior:

Systematic variations of closed and open procedures are possible in which the examiner is either allowed or required to consider each pair more than once, - and in which the experimenter manipulates the size and nature of exemplars, the prior probabilities of correct identifications, the examiner's knowledge of those probabilities, or the costs assigned to false identifications and false eliminations. A common variation employs *match trials* in which coded versions of the unknown exemplar are included among the set of known samples, resulting in a comparison of the form U_a vs. U_b. Match trials are single-blind experimental controls intended to index correct identification and false elimination. Experimental variations modify the numbers of possible correct and incorrect decisions, but not the types of error. Distinctions between closed and open procedures are pertinent to laboratory studies, but in forensic practice it may not be possible to know which condition applies.

The discrimination procedure differs from open-set and closed-set procedures in that the examiner is provided with several exemplars produced by one unknown talker and several exemplars produced by a single known talker ($N = 1$). The examiner's task is to determine whether the two groups of exemplars are *sufficiently similar* to have been produced by the same individual. Match trials can be used in discrimination procedures for the purposes noted above. This procedure can produce correct identifications, correct eliminations, false identifications and false eliminations. Most scientific studies can be classified according to the terms introduced to this point.

Under controlled experimental conditions, correct and incorrect decision outcomes can be described for different identification procedures based upon various data. Rigorous quantitative comparison of decision methods is possible using techniques drawn from signal detection theory[13] and Bayesian statistics.

13. See John Swets, *Measuring the Accuracy of Diagnostic Systems,* 240 SCIENCE 1285.

These scientific techniques are similar to those used in research on medical diagnosis. [14]

[c] Acoustic Characteristics

A number of different acoustic characteristics are potentially useful in talker identification. An extensive and detailed listing is not possible in this brief chapter, but some commonly used characteristics can be cited as examples. Tosi et al.111 considered the parameters of mean frequencies and bandwidths of vowel formants, gaps and type of vertical striations, slopes of formants, duration of similar phonetic elements and plosive gaps, energy distribution of fricatives, plosives, and interformant spaces.16 Buder17 identifies other parameters of broader scientific interest. The Voice Identification and Acoustic Analysis Subcommittee of the International Association for Identification (VIAAS-IAI) guidelines IS specifically mentioned the following: general formant shaping and positioning, pitch striations, energy distribution, word duration, and coupling of the oral and nasal cavities. Other possibilities include inhalation noise, repetitious throat clearing, and vocalized pauses. It should be noted that these are broad categories of acoustic differences and each can include a number of variations or subtypes. The degree to which a given acoustic characteristic may contribute to an identification can vary with the talkers under examination, the quality of the recordings, and the speech sample available for inspection. The various sounds of a language differ in terms of their distinguishing acoustic characteristics, and some research studies indicate that some phonemes (the basic sound elements that distinguish among words) are better for discriminating among speakers than others.[19]

As the preceding discussion reveals, the scientific study. of speech represents a number of[20] different disciplines, including physics, physiology, anatomy, and psychology.

[2] A Model of Talker Identification Variables

A general conceptual model, or theory, would be useful in integrating the data from various studies and in understanding the potential interactions of

(1988); JOHN SWETS & RONALD PICKET, EVALUATION OF DIAGNOSTIC SYSTEMS: MODELS FROM SIGNAL DETECTION THEORY (1982).

14. HELENA C. KRAEMER, EVALUATING MEDICAL TESTS: OBJECTIVE AND QUANTITATIVE GUIDELINES (1992).

15. Oscar Tosi et al., *Experiment on Voice Identification,* 51 J. ACOUSTICAL SOC'Y AM. 2030 (1972).

16. Individual vowel formants have two primary characteristics: the center frequency of the formant and its bandwidth (spread of energy). Vertical striations relate to the vocal pitch and to irregularities in vocal fold vibrations. Formant slopes refer to changes in the frequency of a formant during a specified time interval. Durations can be determined for a variety of acoustic segments, each of which is defined in terms of one or more acoustic features. Energy distribution typically is described.- in terms of the major frequency regions of sound energy, e.g., the fricative "s" in the wild "stop" has the most high-frequency energy as shown in Figure 1.

17. Buder, *supra* note 12.

18. VIAAS-IAI: Voice Identification and Acoustic Analysis Subcommittee (V1AAS) of the International Association for Identification (IAI), *Voice C07J1-parison Standards,* 41 FORENSIC IDENT. 373 (1991).

19. FRANCIS NOLAN, THE PHONETIC BASEs OF SPEAKER RECOGNITION (1983).

20. A very readable account of this multi-disciplinary endeavor is provided by PETER B. DENES & EILIOTT N. PINSON, THE SPEECH CHAIN (2nd ed., 1993).

the factors that influence talker identification. No unified theory of talker identification has been offered in the scientific literature, but we suggest what one might look like here.

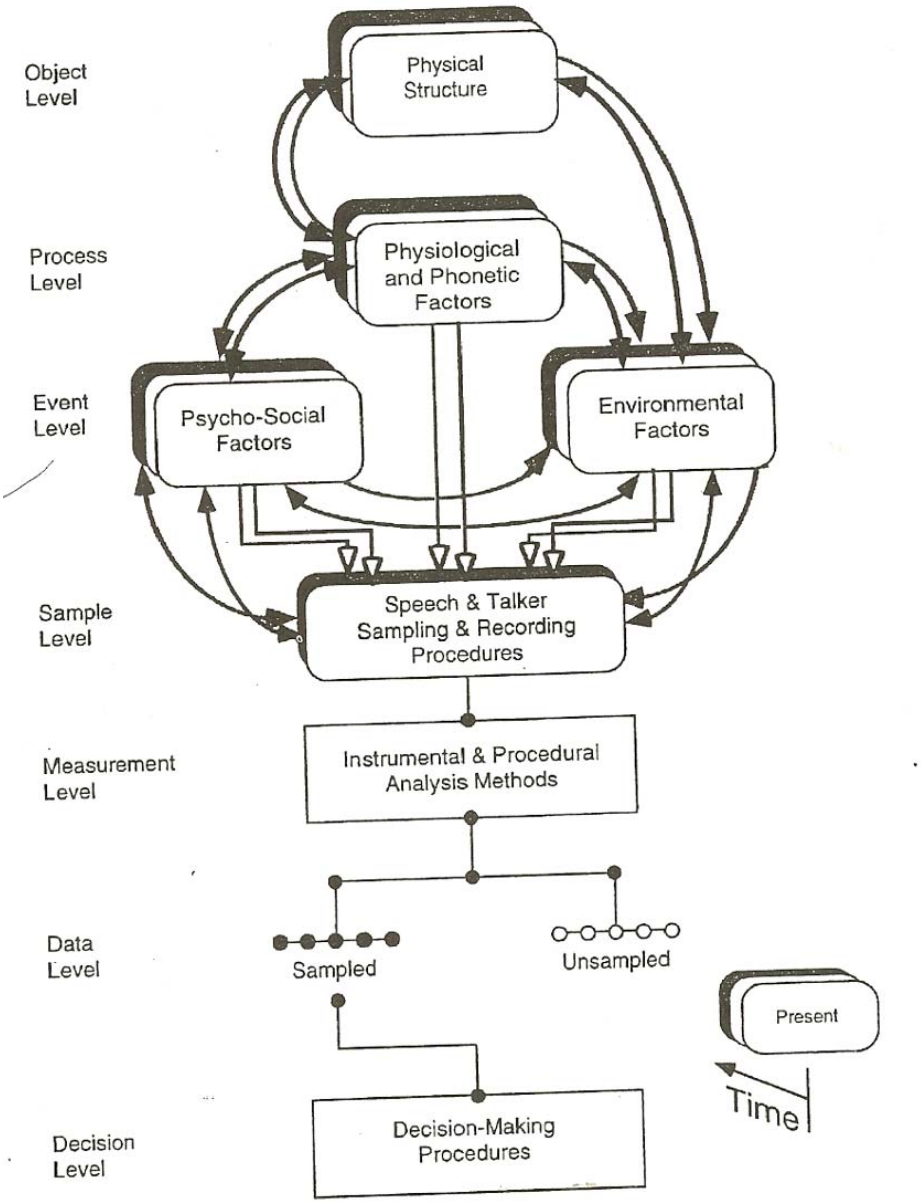

Figure 3. A conceptual model of talker identification by spectrogram, proceeding from events that occur early in the process (top) to those occurring later in the process (bottom). Curved arrows represent mutual influences (interactions) among components of the model and curved boxes represent components for which variability's are known to exist. The model considers talker identification in regard to several levels, ranging from an object level (anatomy of talkers) to a decision level (at which outcomes of talker identification or elimination are desired). Note the third dimension of time. See text for explanation.

Figure 3 considers talker identification with respect to seven major levels. The first three of these (object, process, and event levels) pertain to spontaneous speech production, and the next (sample level) pertains to actions by which speech is captured or elicited for later comparison. The fifth (measurement) level relates to the methods by which recorded speech is analyzed, resulting in reports of the outcomes of such analyses (data level). The last (decision) level deals with the methods and rules by which various data are compared for the purpose of making a decision. For these purposes, decision outcomes are limited to identification and elimination, with or without qualifications of relative certainty.

The object, process, and event levels all have temporal extent, i.e., they can change with the passage of time. The curved lines with solid arrows linking the components of these levels are intended to suggest interactions or mutual influence among the components. These interactions also can change over time. Process and event levels combine to yield spoken language, depicted here with uncurved lines ending in open arrows.

Object, process, and event level components also are subject to major sources of variability, indicated in Figure 3 by curved boxes. Some of these variances change with time. For example, over relatively, long periods of time, the anatomical structures involved in speech production change. An obvious change occurs at puberty when males undergo major changes in laryngeal anatomy. Less obvious are structural changes related to habits of vocal use, environmental agents, aging, and other life events (e.g., changes in dentition). These sources of variability exist both within individuals (over short and long periods of time) and between individuals at any given moment in time. Factors affecting physiological and phonetic variability include general health, state of sobriety or intoxication, and the "rules" underlying various spoken languages. Psycho-social factors include dialect, education and social status, communicative intent, speaking style, and emotional state. Another is the tendency of talkers to reflect the vocal style (changes in vocal loudness, pitch, speaking rate, and patterns of pauses) of those with whom they are communicating.

Environmental factors also can influence speech production. One such effect is the tendency of talkers to increase vocal effort (hence, vocal loudness) as the amplitude of ambient noise increases. This effect (known as the Lombard voice reflex) appears to differ between males and females.21 Talkers also tend to change certain articulation characteristics in the presence of high levels of background noise or other interference.

The absence of temporal invariance and the presence of variability within and between talkers do not necessarily obviate talker identification if their net impact is small compared to the sources of consistency within the time frames typical of forensic identification. Two conditions seem scientifically necessary to justify talker identification by spectrographic (or any other) methods. First, the effect of interactions among physical structure, physiological and phonetic

21. Jean-Claude Junqua, *The Lombard Re- Automatic Speech Recognizers,* 93 J. ACOUSTICAL *flex and its Role on Human Listeners and* SOC'Y AM. 510 (1993).

factors, and psycho-social factors must be consistent within an individual, at least over relatively short periods of time. Further, the patterns of these interactions must be sufficiently idiosyncratic to produce measurable differences among individuals under field conditions. This consistency presumably results in spoken language that differs demonstrably among individuals and is transparent to such factors as vocal disguise.

The fourth major level, sample level, is directly relevant to research design and forensic practice, and therefore is discussed *infra.22*

The decision level involves the ways in which data from different speech samples are compared for the purpose of identifying or eliminating individual talkers. Although standards exist to guide the procedures by which data are compared and the decision options that are allowed, standards are not always followed. Moreover, the standards themselves are somewhat arbitrary. The quality of decision-making ultimately is influenced by formal decision rules, the sufficiency of data, the accuracy and precision of measurement, and the thoroughness of procedural controls invoked while sampling speech materials and talkers.

Talker identification may resemble handwriting identification more than it does fingerprint identification. Both speech and handwriting are dynamic events, influenced by structural (anatomical) and physiological factors; both are influenced by developmental changes and by psycho-social factors. Because fingerprints are static (rather than dynamic) events, they are uninfluenced by .physiology and psycho-social factors. Fingerprints are not subject to developmental change and (barring serious physical trauma) they do not change with age.

§ 31-2.1.2 Scientific Methods Applied in Talker Identification Research

The scientific evidence on talker identification has been obtained primarily by laboratory experimentation. The basic design of a talker identification experiment involves selection of (a) a group of talkers, (b) a type of speech material (e.g., isolated words, short sentences, or conversation), (c) one or more examiners (who may be trained or untrained), and (d) decision categories used by the examiners. However, many additional details of the experimental design need to be specified.

Because the acoustic signal of speech is shaped by a large number of factors, some of which are not under direct experimental control, the results of experiments often must be interpreted with regard to several interacting factors. The conceptual model or framework, such as that presented in Figure 3, may serve to clarify the major issues underlying talker identification by means of spectrographic comparison.[23] Figure 3 summarizes a number of variables and potential interactions that conceivably influence talker identification by spectrogram comparison.

22. See infra § 2.1.2[3] Sampling of Talkers

23. With minor modifications, structurally similar models could be constructed for other systems of forensic identification such as those based on fingerprints and handwriting.

[1] Variations in Methods

As a result of the manifold decisions that have to be made in designing a single experiment, it is rare that any two published reports on talker identification can be directly compared with respect to size of error. As one major contributor to this field remarked, "the reader should be warned that most of the laboratory experiments on voice identification performed to date share one common characteristic-their results are very hard to compare because experimental conditions among them differ widely and data were reported differently."[24] The matter becomes further complicated in attempted extrapolations from the laboratory experiments to practical application in forensics (as discussed earlier in this chapter). These differences among laboratory experiments and between laboratory experiments and forensic examinations are major obstacles in the evaluation of the accuracy of talker identification.

[2] Quantity of Sound Studied

One reason for the difficulty of comparing various published studies is that each experiment must severely limit the quantity of evidence examined in comparison to the potential amount of evidence. Speech consists of a number of speech sounds that can be combined to form words. Words, in turn, can be selected to form a potentially infinite number of sentences. On any occasion, a talker may use just a few words. The words available for examination may not be ideal for the purpose of talker identification, because some words contain better cues for identification than others. Some of the most frequently occurring words in English are short in duration and therefore listed in their acoustic cues. These words include: the, of, I, and, you, to, a, in, that, it, is, he, we. Furthermore, a word produced in different phonetic contexts (such as different sentences) will not necessarily have the same acoustic appearance from one context to the next. Finally, the production of a word even in the same context (produced in the same phrase or sentence) may vary from one time to another.

[3] Sampling of Talkers

At the sample level of the model discussed, *supra,*[25] are those procedures used to select talkers to be recorded for forensic comparison, and specification of the speech to be produced by those talkers. Such selections are undertaken in circumstances in which it is normally possible to exercise some control over recording environment, language, and' 'Speaking style. If these controls are managed skillfully and thoroughly, some of the event level variances can be minimized, thus providing samples for comparison that are "fair" in the sense that the remaining variances are dominated by the presumably idiosyncratic interactions noted above. Because multiple speech and talker samples cannot be collected at exactly the same moment, temporal variables exist at the sample level. If samples are collected with proper attention to investigative technique, the identities of all but one of the talkers should be known with certainty. (This assumes, of course, that recordings are authentic and have not been edited in any way.)

24. OSCAR TOSI. VOICE IDENTIFICATION: THEORY AND LEGAL APPLICATIONS 57 (1979).

25. *See supra* § 2.1.1[2] A Model of Talker Identification Variables.

Once appropriate voice samples have been collected and authenticated, a host of formal observations can be made. Some of these observations result in quantitative measurements based on visual displays of speech, others exist in the form of non-numerical observations of the visual displays themselves, and still others exist in the form of perceptual observations about the audio material under consideration. Sample duration and recording quality limit the observations that are possible. The reports of quantitative and non-quantitative observations form the data used in talker identification. Of the universe of potential data based on speech spectrograms, only a subset is used in most forensic situations. Because the subject of most forensic speech analysis is recorded material, and because the method of analysis is instrumental, temporal variances are of minimal importance.

[4] Differences Between Scientific Research and Forensic Application

Finally, it would be helpful to compare the methods used in scientific research and those used in forensic practice. Before doing so, however, we should note that general standards or recommendations have been published for forensic examination.26 The procedures followed by individual examiners may vary. The following discussion, however, assumes forensic procedures generally consistent with the aforementioned sources. In particular, frequent reference will be made to the guidelines of the Voice Identification and Acoustic Analysis Subcommittee of the International Association for Identification (VIAAS-IAI). These guidelines, published in 1991, are the most recent recommendations for forensic examination.

Published scientific research differs from forensic application in four potentially important respects:

[a] Decision Categories

One difference between scientific studies and typical forensic application is that forensic examiners generally use categories of "no decision" or "uncertain" whereas most published research has required that subjects judging spectrograms choose between identification and elimination. According to VIAAS-IAI, "every examination conducted can only produce one of seven (7) decisions: Identification, Probable Identification, Possible Identification, Inconclusive, Possible Elimination, Probable Elimination, or Elimination".[27] Quantitative criteria based on decisions for the number of comparable words are associated with each of the seven decisions. Apparently, no published large-scale research project has used these decision categories. Consequently, some writers have argued that the accuracy rates in the experimental

26. See RICHARD H. BOLT ET AL., ON THE THEORY AND PRACTICE OF VOICE IDENTIFICATION (1979); Bruce E. Koenig, *Speaker Identification,* 49 FBI Law ENFORCEMENT BULLETIN 20 (1980); Bruce E. Koenig, *Spectrographic Voice Identification: A Forensic Survey,* 79 J. ACOUSTICAL Soc'y .AM. 2088 (1986) Getter); Bruce E. Koenig, *Spectrographic Voice Identification,* 13 CRIME LAB DIGEST 105 (1986); LONNIE L. SMRKOVSKI, FORENSIC VOICE IDENTIFICATION (Michigan Department of State Police, 1983); TOSI, *supra* note 24; VIAAS-IAI, VOICE IDENTIFICATION IN. STRUCTION MANuAL (n.d.).

27. VIAAS-IAI, *supra* note 18 at 387.

studies are conservative estimates of the accuracy to be expected in forensic examinations.

[b] Visual vs. Visual-Aural Identification

A second difference between the scientific research on talker identification and the forensic situation is that forensic examiners routinely use aural recognition and spectrograms in reaching a decision, but published research articles have rarely, if ever, studied the joint use of aural and spectrographic procedures. Stevens et al.[28] compared spectrographic and aural presentations of stimuli. They found that the percentages of correct responses in both closed and open tests were significantly higher for aural than for visual examination. However, the examiners were not specially trained for visual examination of spectrograms, and the speech samples were brief (although exceeding the minimum number of words recommended by the VIAAS-IAI). Apparently, no published scientific report used a procedure that accords with the customary practice in forensic examination of using trained personnel to conduct combined aural and spectrographic examinations in reaching a decision of identification or elimination. Experiments have been conducted on aural and spectrographic identification separately, but it has not been established how the two methods of examination complement one another in joint application.

As a partial answer to this question, Bolt et al.[29] compared the aural task performance reported by Stevens et al.[30] with the .visual spectrogram performance reported by Tosi et al.[31] Bolt et al. concluded, "These resultS would seem to support the idea that listening and visual examination of spectrograms are comparable as single-mode methods of speaker recognition" (at 118). Bolt et al. also concluded that listening-only experiments on talker identification show that (1) performance is less than perfect, with error scores for best conditions ranging from 5 to 15%; and (2) performance is fairly robust under certain types of degradation (such as filtering and addition of noise). It is not clear from this research to what degree the aural performance of experts in voice identification differs from that of laypersons. Carlson and Granstrom showed that accurate matching of unknown voices speaking different utterances can be done if the samples are sufficiently long. But they also report that listeners differ in this ability.[32] Representative studies on aural identification of talkers are given in the margin.[33] The results of selected studies are summarized in Table 2.

28. Kenneth N. Stevens et al., *Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material,* 44 J. ACOUSTICAL Soc'y AM. 1596 (1968).

29. BOLT ET AL., *supra* note 26.

30. Stevens et al., *supra* note 28.

31. Tosi et al., *supra* note 15.

32. Rolf Carlson & Bjorn Granstrom, *An Interactive Technique for Matching Speaker Identity,* 52 PHONETICA 236 (1995).

33. Peter D. Bricker & Sandra Pruzanski, *Effects of Stimulus Content and Duration on Talker Identification,* 40 J. ACOUSTICAL Soc'y AM. 1441 (1966); FRANK R. CLARKE ET AL., CHARACTERISTICS THAT DETERMINE SPEAKER RECOGNITION (Electronic Systems Division, U.S. Air Force Technical Report ESD-ti~36, 1966); Frances McGehee, *The Reliability of the Identification of the Human Voice,* 17 J. GEN. PSYCH. 249 (1937); Frances McGehee, *An Experimental Study in Voice Recognition,* 31 J. GEN. PSYCH. 53 (1944); Irwin Pollack et al., *On the Identification of Speakers by Voice,* 26 J. ACOUSTICAL SOC'y AM. 403 (1954); William Voiers, *Perceptual Ba. sis of Speaker Identity,* 36 J. ACOUSTICAL Soc'y AM. 1065 (1964).

Table 2

Summary of Selected Experiments in Talker Identification. Examples for Both Listening (Aural) and Visual (Spectrogram) Experiments Are Shown.. Note: Many Important Differences Among the Studies Are Not Represented in This Table, Which Is Intended to Show the General Sizes of Error in Selected Experiments.

Source	Talkers	Examiners	Error rate
LISTENING EXPERIMENTS			
McGehee(1937,1944)	31 males, 18 females, se- *Lected to form P31lN of 5* talkers.	740 college students (un- *Trained). :selected to form* 15 panels of 1isteners.	17 to 87%, depending on *Time elapsed between* first and second sessions.
Pollack et aI. (1954)	16, in groups of 2 to 8.	Listeners familiar with the talkers.	5% for normal voiced speech; 70% for whis- pered.
Bricker & Pruzansky (1966)"	10.	16 listeners. familiar with the talkers.	o to 8% across judges for best condition.
Stevens et aI. (1968)	24 highly homogeneous talkers.	6 (untrained).	Closed tests: 6 to 18% false identifica- tion; 6 to 8% faIse identi- fication and 8 to 12% faIse elimination.
VISUAL EXPERIMENTS			
Young & Campbell (1967)	5 men used as known talkers in each trial.	10 examiners with 2.5 hr. of training.	63% based on two short words excerpted from context.
Stevens et al. (1968)	24 highly homogeneous talkers.	4-6 (untrained).	Closed tests: 21 to 28% false identification; Open tests: 31 to 47% faIse identifica- tion and 10 to 20% false elimination.
Hazen (1973)	50 maIes.	7 two-person panels trained over severaI sessions.	Closed tests/same con- text: 2(>% Open tests/ same context: 7% faIse identification, 36% false elimination.
Smrkovski (1976)	7 male and 7 female.	4 experts, 4 trainees; 4 novices.	Match/no match deci- sions: 0% errors for ex- perts and trainees; 5% false identification and 25% false elimination for novices.
Tosi et aI. (1972)	Up to 40 in individuaI ex- periments; drawn from 250 men selected from 25,000	29 persons trained for one month, working indi- vidually or in two- or three-person teams.	For open tests, noncon- temporary spectrograms and continuous speech: 6% false identification 13% false elimination.

It is likely that a combination of aural and visual identification procedures would have a lower error rate than either procedure used alone, but the amount of error reduction is unknown. The VIAAS-IAI specifies that in spectrographic/aural analysis, an "aural short-term memory comparison must be conducted "[34] and this procedure appears consistent with common forensic procedure. Indeed, it is unlikely that examiners who prepare the spectrograms for identification purposes would not listen to the tape recordings as part of the process. Forensic standards developed at the German Bundeskriminalamt use a combination of listening and spectrographic analysis.[35] The procedure includes listening and analysis by a phonetician, who identifies and characterizes dialect, pathologic features, and any other idiosyncratic properties. Acoustic analysis is performed both to provide quantitative support to the characteristics determined by listening and to supply additional information. A careful co-registration of the two kinds of analysis would enhance the validity and reliability of identification judgments.'

The effects of combined auditory and visual information also have been assessed for the related purposes of speaker verification or identification. This research pertains to circumstances in which both facial and voice information are available. For speaker verification, a system using dual classifiers (~coustic features of the voice and visual features obtained from a lip tracker) outperformed single methods of classification and reduced the error rate of the acoustic classifier from 2.3% to 0.5%.[36] Fusion of audio and video information in a multi-expert decision making machine was accomplished by Duc, Bigun, Bigun, Maitre, and Fischer. [37] They reported a success rate of 99.5% for speaker verification. A general review of audio-visual integration by Chen and Rao also points to the advantages of using both sources of information. [38] These results may be particularly important in the use of videotapes containing both visual and auditory signals relevant to identification of individuals.

[c] Quality of Speech Samples and Authenticity of Recordings

A third difference is that with few exceptions laboratory research has used high-quality speech recording systems with well-established procedures and unquestioned authenticity. Forens;jc applications frequently must contend with recordings obtained under less than ideal conditions and sometimes of doubtful authenticity. Recordings based on telephone conversations exhibit

34. VIAAS-IAI, *supra* note 18 at 385.

35. IIERMANN J. KUNZEL, SPRECHERERKENNUNG: GRUNDZUGE FORENSISCHER SPRACHVERARBEITUNG (1987); Hermann J. Kunzel, *Current Approaches to Forensic Speaker Recognition,* PROCEEDINGS OF ESCA WORKSHOP ON SPEAKER REC0GNITION, IDENTIFICATION, AND VERIFICATION 135 (Martigny, Switzerland, April 5-7, 1994).

36. P. Jourlin et al., *Acoustic-labial Speaker Verification,* 18 PA'ITERN RECOGNITION LE'ITERS 853 (1997).

37. Benoit Due et al., *Fusion of Audio and Video Information for Multi Modal Person Authentication,* 18 PA'ITERN RECOGNITION LETTERS 835 (1997).

38. Tsuhan Chen & Ram R. Rao, *Audiovisual Integration in Multimodal Communication,* 86 PROCEEDINGS OF THE IEEE 837 (1998).

audio bandwidths ranging from 2000 Hz to about 3500 Hz, depending on a host of factors (telephone handset characteristics, mouth-to-microphone distance, the length of the transmission path, and intervening signal encryption, broadcasting, and switching technologies). Owen described the typical surveillance recording as having "an audio bandwidth of 300 Hz to 6000 Hz, with a maximum dynamic range of 30-50 dB."[39] These specifications are quite poor in comparison to laboratory equipment currently used in scientific studies of speech.[40] Although recent technical improvements permit surveillance recordings with much higher quality, many systems now in use remain very limited in quality. Problems with the quality of forensic recording equipment and methods may be complicated by difficulties in obtaining original recordings, for which enhancement is generally more successful.

Another problem facing the forensic specialist is the possibility that tapes submitted for analysis are nonauthentic or have been altered in some way. Owen,[41] Koenig,[42] and Hollien[43] describe procedures and criteria to ascertain the authenticity of audio tape recordings; Gruber, Poza & Pellicano[44] provide detailed treatment of legal and technical issues associated with authentication of audio recordings for evidentiary purposes. Technical authenticity analysis seeks to determine whether a particular recording was made of the events asserted by the parties who produced the recording, and in the manner claimed by those who produced it, and whether it is free from unexplained artifacts, alterations, deletions, or edits.

The issues of technical quality, enhancement and authentication go beyond the scope of this chapter but should be considered as potentially serious issues in forensic talker identification. Tape recordings should be evaluated to ascertain quality and determine authenticity before spectrograms are examined for talker identification. The VIAAS-IAI has recommended criteria for acceptable quality of speech recordings, including presence of speech energy above 2000 Hz.

[d] Selection of Talkers for Identification Task

A fourth difference pertains to the selection of talkers for whom an identification will be attempted. Scientific experiments generally select talker subsets randomly, that is, without regard to specific similarities to a given reference talker. In contrast, forensic examination usually involves talkers who are selected because their voices have a similarity to a suspect's voice. That is, the talkers are selected to form a reasonable "lineup" of voices.

39. Tom Owen, *An Introduction to Forensic Examination of Audio and Video Tapes,* 39 J. FORENSIC WENT. 75 (1989).

40. In some situations, recordings of poor quality can be enhanced by techniques such as amplitude compression or expansion, gating, simple filtering, or more complex processing (for example, adaptive predictive deconvolution, and adaptive noise cancellation) by which some background noise can be removed. *See* Bruce E. Koenig, *Enhancement of Forensic Audio Recordings,* 36 J. AUDIO ENGINEERING SOC'y 884 (1988)~ for a discussion of some of these procedures.

41. Owen, *supra* note 39.

42. Bruce E. Koenig, *Authentication of Forensic Audio Recordings,* 38 J. AUDIO ENGINEERING Soc'y 3 (1990).

43. HARRY HOLLIEN, THE ACOUSTICS OF Crime (1990).

44. Jordan S. Gruber et al., *Audio Recordings: Evidence, Experts and Technology,* 48 AM. JUR. TRIALS 1 (1993).

Similarity among voices is not easily determined or described but this factor becomes important in evaluating error rates in talker identification. If talkers are chosen because of a similarity criterion, identification or elimination is expected to be more difficult as compared to a: situation in which talkers are drawn randomly.

§ 31-2.2 Areas of Scientific Agreement

It is clear from studies of the acoustic properties of speech that marked differences may occur among various groups of talkers. For example, the acoustic patterns of speech are different among men, women and children, owing largely to differences in the size of the vocal tract, that is, the resonating system of speech that extends from the larynx up through the nose or mouth.45 Differences within a given age-gender group are not as conspicuous as the differences across age or gender groups, but some differences exist, at least in selected comparisons. It is possible that racial differences exist,[46] but such differences have not been studied extensively. As a general conclusion, one might say that there is a high likelihood that large subgroups of talkers, particularly age-gender subgroups and some dialect subgroups, can be distinguished from one another. It is also known that certain acoustic measures of speech are correlated with physical features such as age, gender, height, and weight of talkers.[47]

The gender of a talker can be determined with high accuracy using information on vocal fundamental frequency and vocal tract length.[48] When an isolated vowel segment was analyzed according to these features, classification by gender was nearly perfect. Classification of individual talkers required information on acoustic parameters associated with vocal tract f1ltering (formant patterns). Essentially the same pattern of results obtained for male and female talkers, but classification rates were consistently lower for females, who were more likely to be classified as males than males were to be classified as females. This study is one of the very few to address gender differences in the acoustic identification of talkers.

But the essential question for talker identification by spectrograms for the usual forensic application is whether the intertalker differences are sufficient to distinguish *one* individual from a group of talkers of the same gender and roughly the same age. The" relevant scientific literature pertains almost exclusively to men. Under conditions comparable to the typical forensic examination, unique identification (0% error rates for both false identification.

45. Donald G. Childers & Ke Wu, *Gender Recognition from Speech. Part II: Fine Analysis,* 90 J. ACOUSTICAL Soc'y AM. 1841 (1991); RAYMOND D. KENT & CHARLES READ, THE ACOUSTIC ANALYSIS OF SPEECH (1992); Ke Wu & Donald G. Childers, *Gender Recognition from Speech. Part 1: Coarse Analysis,* 90 J. ACOUSTICAL SOC'y AM. 1828 (1991).

46. Julie H. Walton & Robert F. Orlikoff, *Speaker Race Identification from Acoustic Cues in the Vocal Signal,* 37 J. SPEECH & HEAR ING RESEARCH 738 (1994).

47. J. Suzuki, *Correlation of Speaker's Physical Factors and Speech,* 41 J. ACOUSTICAL SOC'y OF JAPAN 895 (1985).

48. Jo-Anne Bachorowski &. Michael J. Owren, *Acoustic Correlates of Talker Sex and Individual Talker Identity Are Present in a Short Vowel Segment Produced in Running Speech,* 106 J. ACOUSTICAL Soc'y AM. 1054 (1999).

and false elimination in open tests) is unlikely. One expert in the area wrote, "no method or combination of methods will ever yield a positive result (identification or elimination) in 100% of the cases examined."[49] It is also well known that an individual talker never produces speech in *exactly* the same way in different repetitions of what is intended to be the same utterance. [50] Tokens that are recorded in the same session are more similar than tokens recorded in separate sessions.[51] For some acoustic features, intratalker variability can be as great as intertalker variability. [52]

Prospects for talker identification by spectrograms are improved if intertalker variations are large compared to intratalker variations. *As* will be discussed below, there are several sources of intratalker variation including emotional state, influence of intoxicants, and style of speaking. In addition, some short-term and long-term variations have been noted in the speech patterns of individual talkers. [53] The VIAAS-IAI recommends that the examiner compare "similarly spoken words within each voice sample to determine the range of intraspeaker variability,[54] and exclude the sample from comparison if considerable variability is observed. Each of the following factors can substantially affect the accuracy of spectrograms as a means of talker identification. In some cases, two or more of these factors may combine to limit the successful application of the method.

§ 31-2.2.1 Signal Transmission and Recording

It is generally agreed that the speech signal of an individual talker can be affected to some degree by a variety of factors and circumstances in addition to robust individual characteristics of the talker such as anatomy and learned speaking habits. [55] The spectrogram is a visual record of an acoustic signal that represents a series of transformations and passages. Speech is a perishable message. The acoustic vibrations of an utterance decay rapidly in the atmosphere and are lost forever except in the memory of a listener or a storage device such as a tape recorder. Before speech can be stored in a tape recorder or computer, it must be transduced from acoustic to electrical or electromagnetic energy. This transduction process itself can degrade the quality of the signal. In addition, if the speech to be recorded is produced in a background of other sounds, these extraneous sounds usually will be recorded along with the speech signal, reducing the signal-to-noise *(SIN)* ratio. Some general signal quality requirements have been recommended by the VIAAS-IAI.

.49. TOSI, *supra* note 24 at 148.

50. Richard H. Bolt et al., *Speaker Identification by Speech Spectrograms: Some Further Observations,* 54 J. ACOUSTICAL Soc'y AM. 531 (1973); BOLT ET AL., *supra* note 26.

51. Aaron E. Rosenberg & Frank K. Soong, *Recent Research on Automatic Speaker Identification, in* ADvANCES IN SPEECH SIGNAL PROCESSING 701 (Sadaoki Furui & M. Mohan Sondhi eds., 1992).

52. James E. Atkinson, *Inter-and Intraspeaker Variability in Fundamental Voice Fre" quency,* 60 J. ACOUSTICAL SOC'y AM 440 (1976).

53. Werner Endres et al., *Voice Spectrograms As a Function of Age, Voice Disguise, and Voice Imitation,* 49 J. ACOUSTICAL SoC'y AM. 1842 (1971-); Michael P. Rastatter & Richard D. Jaques, *Formant Frequency Structure of the Aging Male and Female Vocal Tract,* 42 FOLIA PHONlATRICA 312 (1990).

54. VIAAS-lAI, *supra* note 18 at 384.

55. See Richard H. Bolt et al., *Speaker Identification of Speech Spectrograms: A Scientists'View of Its Reliability for Legal Purposes,* 47 J. ACOUSTICAL SOC'y AM. 597 (1970); Bolt et al., *supra* note 50; BOLT ET AL., *supra* note 26; HOLLIEN, *supra* note 43; TOSI, *supra* note 24.

§ 31-2.2.2 Phonetic Context and Duration

It is commonly accepted that speech samples chosen for use in talker identification by spectrograms should be controlled for phonetic context and should meet minimal requirements for duration. Phonetic considerations are important because speech sounds can be highly adapted to individual phonetic contexts, particularly the adjacent sounds but also more global influences such as stress pattern (stress is the degree of prominence given to syllables in an utterance) and speaking rate. [56] Words that may seem to be similar from their orthographic spelling can in fact be very different in phonetic properties. The VIAAS-IAI requires a minimum of ten comparable words between two voice samples; the Federal Bureau of Investigation requires twenty words.[57] Words can vary greatly in their total length (measured in time), number of syllables, and number of phonetic constituents. Consider, for example, the monosyllable words a, *the, it* versus the polysyllable words *telephone, carefully, explosive.* Although the number of words is a convenient index for perceptual and linguistic reasons, it should be recognized that words are not necessarily equal in their potential for talker identification.

Control for phonetic context possibly can be relaxed if talker-sensitive parameters are studied over long durations of speech. However, it should be stressed that such long durations are not always available in forensic applications. The duration of a sample is important to ensure that a sufficient. number of distinguishing features can be identified in the visual pattern. Research on intertalker and intratalker variability among correlation matrices obtained from continuous speech indicates that about 30 seconds is a minimum duration of speech required for the stabilization of the correlation matrices. [58] Assuming an average speaking rate of about 170 words per minute, a 30-second sample of continuous speech should include about 85 words. [59]

§ 31-2.2.3 Disguises and Mimicking

Another point of consensus is that disguises or mimicking of speech can complicate the use of spectrograms in talker identification, and in some situations may preclude comparisons.[60] One of the most damaging disguises is falsetto (a high-pitched voice quality), which can greatly obscure formants in

56. BOLT ET AL., *supra* note 26.

57. Koenig, CRIME LAB DIGEST, *supra* note 26.

58. K. Li & G. Hughes, *Talker Differences As They Appear in Correlation Matrices of Continuous Speech Spectra,* 55 J. ACOUSTICAL SOC'y AM. 833 (1974).

59. This value is given only as an example based on average speech rates; the actual number of words in a time interval can vary widely across talkers and situations.

60. See Endres et al., *supra* note 53; Harry Hollien & Robert E. McGlone, *The Effect of Disguise on "Voiceprint" Identification, 1976* PROCEEDINGS OF THE CARNAHAN CONFERENCE ON CRIME CoUNTERMEASURES 30, *reprinted in* 2 NAT'L - J. CRIM. DEF. 117 (1976); Kathleen Houlihan, *The Effects of Disguise on Speaker Identification from Sound Spectrograms, in* ~ ISSUES IN THE PHONETIC SCIENCES 811 (Harry Hallien & Patricia Hollien eds., 1979); Alan Reich, *Detecting the Presence of Disguise in the Male Voice,* 69 J. ACOUSTICAL Soc'y AM. 1458 (1981); Alan Reich &: James Duke, *Effects of Selected Vocal Disguises Upon Speaker Identification by Listening,* 66 J. ACOUSTICAL Soc'y AM. 1023 (1979); Alan Reich et al., *Effects of Selected Disguises Upon Spectrographic Speaker Identification,* 60 J. ACOUSTICAL SOC'y AM. 919 (1976).

wide-band spectrograms.[61] This result occurs because the harmonics of the falsetto voice become so widely spaced that the conventional wide-band analysis of 250 or 300 Hz resolves the harmonic pattern of the voice rather than the formant pattern. Disguises differ in the degree to which they hinder talker identification by spectrograms. Self-selected vocal disguises have been among the most effective in published research on speaker identification.[62] The VIAAS-IAI recommends that speech samples containing disguises such as falsetto or whispering be eliminated from comparison consideration.

§ 31-2.2.4 Psychological or Physical State of the Talker

The psychological and physical state of the talker exerts a variety of influences on the speech signal. Although emotional state is not straightforwardly associated with a simple set of acqustic effects, different emotions can be distinguished by combinations of acoustic characteristics.[63] For example, Murray and Arnott[64] compared emotions such as anger, fear, sadness and disgust. Speaking rate tends to be accelerated in anger and fear but decelerated in sadness or disgust. The average pitch of the voice tends to be higher in anger and fear but lower in sadness and disgust. Vocal intensity is increased in anger but reduced in sadness and. disgust. The range of vocal pitch increases in anger and fear but decreases in sadness. Tartter and Braun[65] found that listeners can accurately discriminate recorded speech produced by the same talkers while frowning; smiling or assuming a neutral facial expression. Frowning has the effect of lowering formant frequencies and increasing syllable duration. It appears that emotions have a multidimensional effect on speech. Although rate, voice pitch and intensity changes are especially notable, alterations can occur in the precision of articulation as well.[66] Very few studies of talker identification have manipulated talker emotion as a factor that might influence the accuracy of talker identification. As an additional complication, phonetic characteristics of speech may change with changes in loudness.[67]

A topic related to a talker's emotional state is the psychological stress of an individual from whom speech is recorded. Note that the word *stress* can denote two different concepts in the study of speech. One is the degree of emphasis or accent placed on a particular word or syllable in an utterance. The other meaning is a general psychological or physiological state that occurs as the result of demanding, hazardous, or threatening conditions. Hecker et

61. Endres et aI., *supra* note 53.

62. Reich et aI., *supra* note 60.

63. Michael H. L. Hecker et aI., *Manifestations of Task-Induced Stress in the Acoustic Speech Signal,* 44 J. ACOUSTICAL SOC'y AM. 993 (1968); lain R. Murray & John L. Arnott, *Toward the Simulation' of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion,* 93 J. ACOUSTICAL SOC'y AM. 1097 (1993); Klaus R. Scherer, *Vocal Affect Expression: A Review and a Model for Future Research,* 99 PSYCH. BULLETIN 143 (1986); Vivien C. Tartter & David Braun, *Hearing Smiles and Frowns in Normal and Whisper Registers, 94* J. ACOUSTICAL SOC'y AM. 2101 (1994); Carl E. Williams & Kenneth N. Stevens, *Emotions and Speech: Some Acoustical Correlates,* 52 J. ACOUSTICAL SoC'y AM. 1238 (1972).

64. Murray & Arnott, *supra* note 63.

65. Tartter & Braun, *supra* note 63. 66.

Murray & Arnott, *supra* note 63.

67. D. Rostolland, *Phonetic Structure of Shouted Voice,* 51 ACOUSTICAL 80 (1982).

al.[68] reported that stress could be aurally detected in the speech of some individuals and that acoustic correlates of psychological stress varied across talkers. One of the most frequently observed acoustic correlates was voicing irregularity, or an irregularity in the pattern of glottal vibration (the vertical striations in a wide-band spectrogram). The understanding of the acoustic correlates of psychological stress is incomplete owing to limitations of published research.[69] Much of the research in this area overlaps with research on the study of the emotional correlates of speech patterns.

The influence of alcohol and other substances also may induce changes in speech patterns. Alcohol inebriation is known to influence speech .and it appears that some parameters are particularly sensitive to alcohol consumption.[70] Among the typical changes that have been observed in inebriated speech are misarticulations, reduced speaking rate, an increase in the ratio of voiceless to voiced sounds, a change in pitch range (either larger or smaller, depending on the talker) and an increase in pitch variability. It must be kept in mind that the group studies published in the scientific literature show that individual differences occur but the extent of these differences is not established for anything but small samples of talkers. Accordingly, care should be taken in generalizing from the limited data now in print. Moreover, very little is known about the interaction of factors such as the talker's emotional state, overall loudness, psychological stress, and the influence of intoxicants.

A talker's health, particularly the status of the upper respiratory tract, can influence speech patterns. One of the most efficient parameters proposed for talker identification relates to nasal sounds,[71] but these sounds are vulnerable to even a mild head cold.[72] Research on this topic is limited, but it seems likely that a talker's health can affect speech patterns. Normal aging processes also can influence speech production in various ways,[73] and it appears that the speech characteristics most predictive of aged speech also are frequently implicated in the speech of persons with neurologic diseases.[74]

68. Hecker et aI., *supra* note 63.

69. Douglas A. Cairns & John H. L. Han sen, *Nonlinear Analysis and Classification of Speech under Stressed Conditions,* 96 J. Acous-TICAL SOC'y AM. 3392 (1994).

70. F. Klingholz et aI., *Recognition of Low-Level Intoxication from the Speech Signal,* 84 J. ACOUSTICAL SOC'y AM. 929 (1988); L. Lester & R. Skousen, *The Phonology of Drunkenness,*. PA. PERS FOR THE P ARASESSION ON NATURAL PHONOWGY 233 (A. Bruck et aI. eds., 1974); G. Niedzielska et aI., *Acoustic Evaluation of Voice in Individuals with Alcohol Addiction,* 46 FOLIA PHONIATRI. CA 115 (1994); David B. Pisoni & Christopher S. Martin, *Effects of Alcohol on the AcousticPhonetic Properties of Speech: Perceptual and Acoustic Analyses,* 13.ALcOHOLISM: CLIN. & EXPERIMENTAL REs. 577 (1989); Linda C. Sobell & Mark B. Sobell, *Effects of Alcohol on the Speech of Alcoholics,* 15 J. SPEECH & HEARONG REs. 861 (1972); Linda C. Sobell et aI., *Alcohol-induced Dysfluency in Nonalcoholics,* 34 FOLIA PHONIATRICA 316 (1982); STEVEN B. CHIN & DAVID B. PISONI, .ALCOHOL AND SPEECH (1997).

71. Lo-Soun Su et aI., *Identification of Speakers by Use of Nasal Coarticulation,* 56 J. ACOUSTICAL SOC'Y AM. 1876 (1974); NOLAN, *supra* note 19.

72. M. R. Sambur, *Selection of Acoustic Features for Speaker Identification,* ASSP-23 IEEE TRANSACTIONS ON ACOUSTICS AND SPEECH SIGNAL PROCESSIN:G 176 (1975).

73. Raymond D. Kent & Robert Burkhard, *Changes in the Acoustic Correlates of Speech Production, in* AGING: COMMUNICATION PROCESSES AND DISORDERS 47 (Daniel S. Beasley & G. A. Davis eds., 1981); Gary Weismer & Julie M. Liss, *Aging and Speech Motor Control, in* HANDBOOK OF AGING AND COMMUNICATION 205 (D. Ripich ed., 1991).

74. W. Ryan & K. Burke, *Perceptual and Acoustic Correlates of Aging in the Speech of Males,* 7 J. COMMUNICATION DISORDERS 181 (1974).

Therefore, it has been suggested that normal aging and some neurological diseases have similar effects on speech production.[75]

The implication is that the emotional and physical state of a talker ideally would be controlled in forensic applications of talker identification. However, ascertaining these characteristics and replicating them may be extremely difficult and sometimes impossible. Perhaps the best that can be done in most practical situations is to acknowledge that these factors may limit the accuracy of the identification procedure. The number of potentially significant variables makes talker identification more complicated than fingerprint identification.

§ 31-2.2.5 Earwitness Testimony

There is substantial agreement on the ability of witnesses to identify individuals from the sound of their speech. This research does not speak directly to the abilities of forensic experts but rather to general earwitness testimony (i.e., testimony based on recall of auditory events, especially spoken messages uttered at the scene of a crime). Most published earwitness research focuses either on famous voices or unfamiliar voices. Recent research includes studies of memory for voices and studies of the effects of training on earwitness accuracy, both of which are relevant to gauging the role of audition in talker identification.

The recognition of famous voices improves with duration of the voice sample, with the largest gain in performance occurring within the first second and smaller gains thereafter. [76] A similar effect appears to occur for unfamiliar voices. Cook and Wilding reported that recognition memory improved with longer extracts of a stranger's voice but was not affected by increasing vowel variety (i.e., different types of vowel sounds).[77] These authors also concluded that memory for words spoken was not correlated with memory for an unknown voice. In considering these results, Cook and Wilding raise the possibility that there may be a general ability for memory-for-voices. Sheffert, however, reached a somewhat dissimilar conclusion in an experiment designed to determine whether memory for voices is distinct from episodic memory. The results of Sheffert's study were interpreted to mean that "voices are represented in long-term memory as episodic traces that contain talker specific perceptual information."[78] Essentially the same conclusion was reached by Remez, Fellowes, and Rubin in an experiment in which natural voices were radically transformed to sinewave replicas (i.e., the formants of speech were replaced by sinusoids).[79] The results of the experiment involving

75. Id.

76. Stephan R. Schweinberger et ai, *Recog nizing Famous Voices: Influence of Stimulus Duration and Different Types of Retrieual Cues,* 40 J. SPEECH, LANGUAGE, &: HEARiNG REs. 453 . (1997).

77. Susan Cook & John Wilding, *Earwit*ness *Testimony: Never Mind the Variety, Hear the Length,* 11 APPL. CooNITIVE PSYCH. 95 (1997).

78. Sonya M. Sheffert, *Contribution of Surface and Conceptual Information to Recogni tion Memory,* 60 PERCEPTION &: PSYCHOPHYSICS 1141, 1141 (1998).

79. Robert E. Remez et al., *Talker Identification Based on Phonetic Information,* 23 J. ExPERIMENTAL PSYCH.: HUMAN PERcEPTION &: PERFORMANCE 651 (1997). A *formant* is a resonance of the vocal tract; a particular pattern of formants characterizes individual speech sounds. A sinusoid is a pure tone, or signal containing energy at a single frequency.

six voices indicated that word recognition was affected by changes in voice for several encoding tasks. It was concluded that even with these highly altered stimuli, listeners still were able to identify ten voices. The authors viewed their results as compatible with a model in which phonetic properties of utterances carry information on both words spoken and the individuals who speak them. Although it is premature to make definitive statements, the recent research is consistent with the idea that voice identification improves with the duration of the voice sample, and that talker information is to some degree combined with information on the words spoken in a composite memory trace. These laboratory studies provide some basic information about how well listeners can identify the voices of others.

Work by Elaad, Segev, and Tobin indicates that talker identification ability is affected by training.[80] They compared identification accuracy in a mock theft study involving three listener groups: voice identification experts,[81] individuals who are' totally blind, and sighted listeners without special training or experience. The highest accuracy was achieved by the voice identification experts. There also appears to be an effect of nationality on both face and voice recognition, with subjects showing better recognition for faces and voices of their own nationality.[82] The influence of interviewing technique was investigated by Memon and Yarmey in a mock abduction with 77 subjects who rated the abductor's voice for nine speech characteristics and attempted to identify his voice from an audiotaped lineup.[83] No effect of interview type was found.

In a comparison of eyewitness and earwitness identification with respect to calibration and diagnosticity analysis, Olsson, Juslin and Winman reported that earwitness identification is poorer and is characterized by overconfidence and low diagnosticity of confidence, meaning that the witness's level or feeling of confidence had a poor relationship with accuracy.[84]

A number of issues relevant to earwitness identification have yet to be studied, including the effects of competing signals, the effects of hearing loss and the effects of linguistic accent (in particular, similarities and differences in the accents of talker and listener). Yarmey reviewed the general topic of earwitness voice identification and recommended procedures for voice lineUpS.[85] Although careful attention to these procedures can enhance the reliability of voice lineups, the research on earwitness identification points to general concerns about its accuracy. These shortcomings make objective identification by acoustic analysis all the more important in forensic application. This is not

80. Eitan Elaad et aI., *Long-term Working Memory in Voice Identification,* 4 PSYCHOL., CRIME & LAW 73 (1998).

81. The subjects were lay individuals with no special training in talker identification.

82. Nathan D. Doty, *The Influence of Nationality on the Accuracy of Face and Voice Recognition,* 111 AM. J. PSYCH. 191 (1998).

83. Amina Memon & A. Daniel Yarmey, *Earwitness Recall and Identification: Com pari son of the Cognitive Interview and the Structured Interview,* 88 PERCEPT. & MOTOR SKILL'3 797 (1999).'

84. N. Olsson et aI., *Realism of ConfuJ.ence in Earwitness Versus Eyewitness Identification,* 4 J. EXPERIMENTAL PSYCHOL.: ,ApPUED 101 (1998).

85. A Daniel Yarmey, *Earwitness Speaker Identification,* 1 PSYCH" PUBLIC POL'y & LAw 792 (1995).

to detract from the combined use of perceptual and acoustic methods, as is commonly used in forensic talker identification.

§ 31-2.2.6 Summary

Reviews of talker recognition published within the last five years are cited in the margin.[86] Although these reviews document several steps of progress, they also make it clear that much remains to be done to establish the scientific validity and reliability of talker identification by acoustic methods. Gruber and Poza[87] list several areas in which there are open questions or a significant lack of scientific evidence. These include knowledge about (1) the uniqueness and consistency of voices, (2) the relative value of aural and visual methods of identification, (3) the relative accuracy of laypersons and experts using different methods of identification, (4) the specific features that pertain to visual. or aural pattern matching, (5) the criteria used by forensic specialists, (6) "error rates and accuracy in real-world forensic situations, and (7) effect of speaker situation on identification and elimination protocols. This is a substantial list of unresolved issues, and although some progress has been made since publication of the Gruber and Poza volume, the list can be repeated today virtually in its entirety as an agenda for future work.

§ 31-2.3 Areas of Scientific Disagreement

Several areas of disagreement have been discussed in the literature on talker identification. Much of the literature is polarized on the relative accuracy of this method for forensic application. Among the most important publications that are generally critical or guarded concerning speaker identification by spectrograms are Bolt et al.,[88] Gruber and Poza,[89] Hollien,[90] Poza,[91] and Shipp, Doherty and Hollien.[92] Papers that take a more positive view are Black et al.,[93] Kersta,[94] Koenig et al.,[95] and Tosi et al.[96] These papers provide an Overview of various disagreements concerning the use of spectrograms in voice identification or elimination.

86. Joseph P. Campbell, *Speaker Recognition-A Tutorial,* 85 PRocEEDINGS OF THE IEEE 1437 (1997); Sadaoki Furui, *Recent Advances in Speaker Recognition,* 18 PA'ITERN RECOGNITION LEITERS 859 (1997); Frederic A. Gruber & F. T. Poza, *Voicegram Identification Evidence,* 54 AM. JUR. TRIALs 1 (1995); W. Majewski, *Speaker Recognition in Forensic Applications,* 82 AcusTiCA (Supp. 1) S 230 (1996); Francis Nolan, *Speaker Recognition and Forensic Pho"netics, in* THE HANDBOOK OF PHONETIC Sc1ENCES 744 (William J. Hardcastle & John Laver eds., 1997).

87. Gruber & Poza, *supra* note 86.

88. Bolt et aI., *supra* note 50; Bolt et aI., *supra* note 55.

89. Gruber & Poza, *supra* note 86.

90. Harry Hollien, *The Peculiar Case of "Voiceprints,"* 56 J. ACOUSTICAL SOC'y AM. 210 (1971).

91. Fausto Poza, *Voiceprint Identification: Its Forensic Application,* 1974 PROCEEDINGS OF THE CARNAHAN CRIME COUNTERMEASURES CONFERENCE (University of Kentucky, Lexington, KY).

92. Thomas Shipp et aI., *Some Fundamental Considerations Regarding Voice Identification,* 82 J. ACOUSTICAL SoC'y AM. 687 (1987).

93. John W. Black et aI., *"Speaker Identification by_Speech Spectrograms: Some Further Considerations,"* 54 J. ACOUSTICAL Soc'y AM. 535 (1973).

94. Kersta, *supra* note 4.

95. Bruce E. Koenig et aI., *Reply to "Some Fundamental Considerations Regarding Voice Identification,"* 82 J. ACOUSTICAL Soc'y AM. 688 (1987).

96. Tosi et aI., *supra* note 15.

Perhaps the most significant area of disagreement concerns the degree to which research has established the validity and reliability of talker identification by spectrograms. Much of the disagreement centers on the interpretation of published data on error rates in identification and elimination. Table 2 reports error rates for selected published studies of talker identification that have used either aural or visual (spectrogram) methods. The studies differ considerably in many features, including number of talkers, number and sophistication of judges, exemplars available for examination, and type of decision. These considerable differences preclude direct comparisons among studies, and Table 2 must be interpreted cautiously. The error rates in this table range from 0% to 85%.

The largest and most comprehensive investigation of talker identification by spectrograms was reported by Tosi et al.[97] This study includes the largest talker sample reported in the published literature: 250 men drawn from what was considered by the investigators to be a "homogeneous population" of 25,000 males speaking general American English (students at Michigan State University). The examiners were 29 individuals given a month's training in spectrographic identification. A total of 34,996 identification trials were carried out. The trials were based on subsets of 10 or 40 speakers drawn from the larger sample of 250 talkers. The experiment included several aspects: comparison of open versus closed tests, use of noncontemporaneous reference and test samples, and context variation.

The condition that probably compares most closely with forensic tests is the open test with noncontemporaneous samples extracted from continuous speech. Forensic tests frequently must use conversational speech samples (continuous speech) obtained at different times (noncontemporaneous samples) and from a suspect who mayor may not be in the reference sample, or "lineup" of talkers (open test). For this condition, Tosi et al. reported error rates of 6.4% for false identification and 12.7% for false elimination. The examiners also were asked to rate the certainty of their judgments, and Tosi et al. pointed put that 60% of the incorrect judgments were associated with an "uncertain" rating. The authors suggested that if the examiners had been allowed to use a "no opinion" category of decision, the error rates would have fallen to 2.4% for false identification and 4.8% for false elimination.

Other published studies typically have used talker samples of 10-50 in open or closed tests. As noted earlier, it is difficult to make direct comparisons of reported error rates. However, as a general indication of the results, the lowest error rates (best examiner performance) across the experimental conditions used in major studies published in refereed journals were: 21.6% for comparison of single words spoken in isolation,[98] 18% for comparison of phrases and sentences,[99] 11.9% for open tests involving cue words in the same context. [100] Depending on which study one chooses, the accuracy of identifying one individual from a group of ten similar talkers under optimal conditions

97. Id.

98. Martin Young & Richard Campbell, *Effects of Context on Talker Identification,* 42 J. ACOUSTICAL Soc'y Ax 1250 (1967).

99. Stevens et ai., *supra* note 28.

100. Hazen, *supra* note 8.

may be as good as 100% or as poor as 80%. By contrast, in an open test under less than optimal conditions, the accuracy may be 40% or even lower.

In their influential paper, Bolt et al.101 remarked that, "Today's consensus suggests that ta1ker identification by voice patterns is subject to error at a high, and as yet undetermined, rate" (at 602). Bolt et al. concluded, "We find, in brief, that spectrographic voice identification has inherent difficulties and uncertainties. Anecdotal evidence given in support of the method is not scientifically convincing. The controlled experiments that have been reported give conflicting results. : .. We conclude that the available results are inadequate to establish the reliability of voice identification by spectrograms." Nearly 25 years have passed since publication of the Bolt et al.[102] reports, but no research papers that resolve the question of the reliability of spectrographic talker identification have been published. The debate over the reliability of this method centers on studies published before 1980.

OQ,e disagreement pertains to the nature of decisions reached in ta1ker identification or elimination. *As* noted earlier, scientific studies have relied almost exclusively on positive decisions (identification or elimination), but forensic specialists typically use categories such as "probable" or "inconclusive." Because of these differences in decision categories, it is not a Simple matter to relate accuracy rates reported in scientific reports to actual practice in forensic application. It has been argued that the accuracy rates in scientific reports are conservative estimates of the accuracy expected in forensic application, especially if a "no-decision" response is used in the latter. However, this is a matter of interpretation, about which controversy is not surprising. The opposite case also can be argued, given that forensic applications often contend with signals 'of poor quality, uncooperative subjects, and short samples of speech from which exemplars can be extracted.

Some disagreement continues over the vulnerability of the speech signal to some sources of intratalker variation. Some particular points of disagreement are over (a) the signal quality required for identification purposes, (b) the effect of vocal disguises and the emotional or physical state of the talker, (c) the selection of words or phonetic features for purposes of talker identification. But the importance of these disagreements is lessened by the use of forensic procedures such as those recommended by the VIAAS-IAI.

§ 31-2.4 Future Directions

Bolt et al.[103] recommended four main categories of research in talker identification: "the origins and characteristics of variability; the relations between intraspeaker and interspeaker variability; the relations between aural and visual eYRmin9tion; and the potential of developments in automated methods of voice identification to make contributions to the understanding and Improvement of identification performed by humans" (at 61). This research agenda might very well be reiterated today, only because limited progress has been made in these areas since the publication of the Bolt et al. report. .

101. Bolt et al.. *supra* note 50 at 602-603.
102. Id.; Bolt et al. *supra* note 55.
103. BoLT ET AL., *supra* note 26.

Clearly, the need remains for large-scale talker identification experiments that use procedures similar to those in forensic settings, specifically, joint Use of aural and spectrographic examinations, and the use of probable or inconclusive decisions. Experiments should use spontaneous speech samples[104] as well as samples gathered using the methods of experienced forensic investigators to elicit exemplars from known talkers, which simulate the rate, vocal effort, and emotional state of the "unknown" talker to whom comparisons will be made. The intensity of debate over the validity and reliability of talker identification by spectrograms is rooted largely in the lack of experimental data collected with appropriate procedures and adequate controls.

A likely direction of future research is continued exploration of automated (or semi-automated) talker identification, possibly aided by aural recognition. Automated identification usually is a computer-based procedure. It offers the advantage of a comparatively objective process in which specified parameters are examined algorithmically and, if appropriate, assigned individual levels of statistical confidence. Automated recognition may evolve as a useful supplement to other procedures that involve more direct human observation and decision making.

One example is the linear prediction model in which different parametric representations of speech are derived.[105] An important advantage of automated procedures is the capability for analyses of long samples of speech for which variation in phonetic characteristics may be less important than for brief samples.[106] Of course, long samples of speech are not necessarily available for forensic purposes. Another advantage is the capability for relatively more straightforward statistical comparison between the recorded speech of an. unknown talker and data bases drawn from large numbers of talkers and organized in ways that retain information about gender, age, dialect, speaking rate and effort level, and perhaps physiological and psychological state. This is not to say that visual examination of spectrograms should be superseded by automated methods. The pattern recognition abilities of the human visual system are not easily matched by automated systems. Yet visual recognition (like aural recognition) is a legitimate subject of study. Rosenberg and Soong[107] summarized several procedures for automated analysis, but noted that these procedures are more suited to applications in access control (i.e., security) than to speaker identification, primarily because of difficulties with uncontrolled variability in the latter. But the recent use of automated analysis with fairly large talker samples (e.g., 100 talkers in a study by Liu et al.[108]) holds promise that these techniques could have forensic relevance.

104. Hazen, *supra* note 8.

105. One of the most promising parameters was the cepstrum, which is the inverse Fourier transform of the amplitude spectrum. Bishu Atal, *Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification,* 55 J. ACOUSTICAL SOC'y AM. 1304 (1974). The cepstrum and a number of other computer-implemented analyses involve highly technical issues in mathematics and signal processing. It is not possible to review these matters in this chapter, but it should be noted that 'they can introduce considerable complexity in discussions of automated talker identification.

106. Li & Hughes, *supra* note 58.

107. Rosenberg & Soong, *supra* note 51.

108. Chi-Shi Liu et al., *A Study on Mini mum Error Discriminative Training for Speaker Recognition,* 97 J. ACOUSTICAL SOC'y AM. 637 (1995).

Scientific and statistical techniques used to study forensic talker identification are similar to those employed to assess decision-making in medicine. Research in clinical decision making suggests that performance depends on whether constituent information is combined in a serial or a parallel fashion.[109] In a serial decision strategy, discrete tests are conducted and evaluated one after another, typically using decision rules that dictate later steps on the basis of the outcomes of earlier steps. In a parallel strategy, discrete tests are made independently, and a decision is based on patterns of outcomes from the group of tests. It would be valuable to know how different types of information arising from talker identification can be combined most effectively to yield an ultimate decision of identification or elimination. For example, one approach is a parallel paradigm in which the outcomes of independent decisions are reached with aural identification, identification by spectrograms, and identification by automated methods. The separate decisions can be weighted, then combined into a final decision. Current approaches appear to be more serial in character: decisions are made based on if-then sequential logic suggested in part by procedural requirements. It appears that potential effects associated with the order in which comparisons are made have not been studied formally.

Similarly, additional research is needed to determine the impact on summary decision outcomes of independent replications of talker identifications by examiners with varying degrees of experience. The VIAAS-IAI standards require that an independent opinion be obtained for all decisions other than inconclusive, possible identification, and possible elimination; a third opinion is required when the first and second examiners cannot resolve differences. Post hoc reports suggest such procedures produce very small proportions of errors, and that such errors as do occur tend to be related to poor recordings, high-pitched voices, or an insufficient number of words.[110] However, these observations are based on a limited number of comparisons (2000 talkers). The examiners were sufficiently similar in training and in the analysis tools used to make comparisons that generalizing beyond them is difficult.

Future work should lead to a better specification of the parameters that are most effective in identifying talkers. Much as fingerprint identification is based on distinguishing features such as whorls, talker identification examines particular features in the spectrogram. A number of features are potentially useful, but they probably are not equally sensitive to interspeaker differences. It would be important to know the relative value of various features for the purpose of speaker identification. One example is specification of formant tracks and other acoustic features that can be obtained from largely automatic procedures. Goldstein,[111] Sambur,[112] and Wolf [113] described

109. Martyn L: Hyde et aI., *Auditory Test Strategy, in* DIAGNOSTIC AUDIOLOGY (John T. Jacobson & Jerry L. Northern eds., 1991); Robert Turner, *Techniques to Determine Test Protocol Performance,* 9 EAR &: .HEARING 177 (1988); Robert Turner et aI., *Formulating* and *Evaluating Alldiological Test Protocols,* 5 EAR &: HIwuNG 321 (1984); Robert Turner & Donald Nielsen, *Application of Clinical Decision Analysis to A1.tdiological Tests,* 5 EAR &: HEARING 125 (1984).

110. Koenig, J. ACOUSTICAL Soc'y Ax., *supra* note 26.

111. Ursula G. Goldstein, *Speaker-Identifying Features Based on Formant Tracks,* 59 J. ACOUS'l'ICAL Soc'y Ax. 176 (1976).

certain acoustic characteristics that were useful in distinguishing among small sets of talkers but more work on larger numbers of talkers is needed. Another promising direction is to specify various phonetic characteristics that are candidates for talker identification. These might include formant patterns for vowels and vowel-like sounds, spectra of frication noise, durations of noise segments, perturbations (irregularities) in vocal fold vibrations, and covariation in selected frequency and intensity features. The large talker databases now available could be studied to construct statistical ensembles for some of these features. One recent example of such work produced long-term average speech spectra using standardized procedures and equipment for 13 languages as spoken by both male and female talkers.[114]

Kuwabara and Sagisaka[115] gave an example of a systematic approach to acoustic identification. They specified acoustic features at two levels, as follows:

Voice source:

1. average vocal fundamental frequency (voice pitch),
2. time-frequency pattern of vocal fundamental frequency (also known as the pitch contour),
3. fluctuations in fundamental frequency,
4. glottal wave shape.

Vocal tract resonances:

1. shape of spectral envelope and spectral tilt, 2.

absolute value of formant frequencies,

3. time-frequency pattern of formant frequencies, 4.

long-term average spectrum of speech,

5. formant bandwidths.

It should be possible to render identification or elimination decisions along different axes or dimensions, and then combine the independent decisions in a statistical decision model. Examples of different dimensions of analysis are the long-term averaged speech spectrum,[116] the cepstrum,[117] perturbation measures of vocal function,[118] aural recognition,[119] spectrograph

112. Sambur, *supra* note 72.

113. Jared Wolf, *Efficient Acoustic Parame ters for Speaker Recognition,* 51 J. ACOUSTICAL Soo'y AM. 2044 (1972).

114. Denis Byrne et al., *An International Comparison of Long-Term Average Speech Spectra,* 96 J. ACOUSTICAL Soc'y AM. 2108 (1994).

115. Kuwabara & Sagisaka, *supra* note 2. 116. AtaI, *supra* note 105; Mary Lou P. Gelfer et al., *The Effects of Sample Duration and Timing* on *Speaker Identification Accuracy by Means of Long-Term Spectra,* 17 J. PHONETICS 327 (1989); Hany Hollien & Wojciech Majewski, *Speaker Identification by Long-Term Spectra under Normal and Distorted Speech Conditions,* 62 J. ACOUSTICAL SOC'y AM. 975 (1977); H. Kiukaanniemi, *Individual Differences in the Long-Term Speech Spectrum,* 34 FOLIA PHONlATRICA 21 (1982); J. Zalewski et al., *Cross Correlation of Long-Term Spectra as a Speaker Identification Technique,* 34 AcouSTICA 20 (1975); Byrne, et al., *supra* note 114.

117. AtaI, *supra* note 105; James E. Luck, *Automatic Speaker Verification Using Cepstral Measurements,* 46 J. ACOUSTICAL Soc'y AM. 1026 (1969).

118. Philip Lieberman, *Some Acoustic Mea: sures of the Fundamental Periodicity of Normal and Pathologic Larynges,* 35 J. ACOUSTICAL Soc'y AM. 344 (1963); Ingo R. Titze et al., *Some*

ic examination of specific phonetic features including formant pattern, [120] temporal features,[121] nasal features,[122] and turbulence noise for fricatives.[123] Some examples of multiple-factor analyses are those of Doherty and Hollien.[124] Fuzzy logic is one potential approach to combining information from several features, no one of which in itself always yields an unambiguous decision.[125]

The use of particular features in spectrograms might be enhanced if standards were available for the determination of differences related to anatomical dissimilarities as opposed to differences that reflect learned speaking patterns. Presumably, anatomy is more difficult to change than learned patterns of speech. Research along these lines may eventually establish criteria for feature selection. Nolan listed the following criteria for the selection of parameters in talker identification: high between-speaker variability, low within-speaker variability, resistance to attempted disguise or mimicry, availability, robustness in transmission, and measurability. It would be a major advance in spectrograph talker identification if a consensus were reached on the degree to which features or parameters satisfy these criteria.[126]

Published research pertains almost exclusively to male speakers, especially young to middle-aged men. Very little scientific information has been collected for female speakers, children of either gender, or elderly men. It is interesting to note that individual identity can be determined fairly well even with infant cries.[127] At it stands now, the scientific validation of talker identification pertains to less than half of the population. Moreover, the total number of speakers examined in all published research on spectrogram identification remains small.

Audio tape authentication and enhancement will experience continued refinement through technological improvements. It is important that consensus be reached concerning criteria for authentication and acceptable procedures for enhancement. Efforts by organizations such as the Audio Engineer

Technical Considerations in Voice Perturbation Meeu;urements, 30 J. SPEECH 8 HEARING REsEARcH 252 (1987); Kim A. Wilcox & Yoshiyuki Horii, *Age and Changes in Vocal Jitter,* 35 J. GERONTOLOGY 194 (1980); Barbara J. Zyski et aI., *Perturbation Analysis of Normal and Pathological Larynges,* 36 FOLIA PHONJATRICA 190 (1984).

119. William. A. Hargraves & John A. Starkweather, *Recognition of Speaker Identity,* 6 LANGUAGE 8< SPEECH 63 (1963); Reich & Duke, *supra* note 60; Stevens et aI., *supra* note 28; Voiers, *supra* note 33.

120. Goldstein, *supra* note 111; Wiktor Jassem, *Formant Frequencies as Cues to Speaker Discrimination, in* 1 SPEECH ANALYSIS AND SYNTHESIS 41 (Wiktor Jassem ed., 1968); Conrad L. LaRiverie, *Contributions of Fundamental Frequency and Formant Frequencies to Speaker Identification,* 31 PHONETICA 185 (1975); Wolf, *supra* note 113.

121. C. C. Johnson et aI., *Speaker Identification Utilizing Selected Tempor.al Speech Features,* 12 J. PHONETICS 319 (1984).

122. Su et aI., *supra* note 71.

123. Conrad L. LaRiviere, *Speaker Identification for Turbulent portions of Fricatives, 29* PHONETICA 98 (1974).

124. E. Thomas Doherty, *An Evaluation of Selected Acoustic Parameters for Use in Speaker Identification,* 4 J. PHONETICS 321 (1976); E. Thomas Doherty & Harry Hollien, *Multiple-factor Speaker Identification of Normal and Distorted Speech,* 6 J. PHONETICS 1 (1978).

125. Manfred R. Schroeder, *Speech and Speaker Recognition,* 12 BIBIJOTHECA PHONETICA (1985).

126. NOLAN, *supra* note 19.

127. Gwen E. Gustafson et aI., *Robustness of Individual Identity in the Cries of Human Infants,* 27 DEVELOPMENTAL PSYCHOBIOLOGY 1 (1994).

ing Society to develop formal technical

Appendix

Glossary

Acoustic. of or relating to the science of sound, or to the sense of hearing.

Cepstrum. a type of acoustic analysis defined as the Fourier transform of the power spectrum of a signal. The transform is described in terms of *quefrency* (note the transliteration from frequency), which has time-like properties. The cepstrum often is used to determine the fundamental frequency of a speech signal. Voiced speech tends to have a strong cepstral peak, at the first *rahmonic* (note transliteration from harmonic).

Formant. a resonance of the vocal tract, that part of the speech production system that extends from the larynx to the lips or nostrils. A formant is specified by its center frequency (commonly called formant frequency) and bandwidth. Formants are denoted by integers that increase with the relative frequency location of the formants. F1 is the lowest-frequency formant, F2 is the next highest, and so on. The first three formants are most important for speech analysis and perception.

Fricative. a speech sound characterized by a long interval of turbulence noise. Examples are the initial sounds in the words *thin, van, see, zoo, shoe.*
Fricatives are often classified as *stridents* or *nonstridents,* depending on the degree of noise energy. Stridents have greater energy than nonstridents.

Harmonic. an integer multiple of the fundamental frequency in voiced sounds. Ideally, the voice source can be conceptualized as a line spectrum in which the energy appears as a series of harmonics. That is, each line represents one harmonic. The harmonics of the voice source are f1ltered by the vocal tract resonances.

lnterformant spaces. the low-energy regions that fall between adjacent formants. Depending on speech characteristics, these spaces can contain noise or other energy, but usually this energy is quite weak. Consequently, the interformant spaces will appear as white or light gray on a spectrogram.

Orthographic spelling. the written representation of a language. Orthographic spelling may be contrasted with a phonetic transcription. The former refers to the common spelling of words with the letters of the alphabet. The latter is a specialized system in which each symbol uniquely represents a speech sound.

Plosive gap. the acoustic interval corresponding to articulatory closure (vocal tract obstruction) for a plosive consonant; it is ideBtified on a spectrogram as an interval of low acoustic energy, conspicuously lacking in formant pattern or noise.

Plosives. speech sounds characterized by a complete obstruction of the vocal tract that may be followed by an abrupt release of air that produces a burst noise. Examples of plosives are the initial sounds in the words *bill, pill, dill, till, gill,* and *kill.*

Signal detection theory. a mathematical theory dealing with the problem of detecting the presence of a signal in a background of noise. Early applications of this theory sought to statistically measure the accuracy of instrumental and sensory systems, independent of biases arising from the decision criteria of observers. Decision criteria are influenced by the observer's knowledge of the prior probabilities of target events (signals), and by the values and costs associated with correct and incorrect decisions. Accuracy is indexed graphically by means of *relative operating characteristic* (ROC) curves, and quantitatively by measures of performance such as d_ and A_. Subsequent applications of the theory have included problems in medical diagnosis, aptitude testing, weather forecasting, automated information retrieval, and forensic identification.

Spectral envelope. the general shape of the spectrum (the distribution of signal energy as a function of frequency). The envelope often is drawn to represent the overall form of the spectrum and may neglect minor deviations from that form.

Spectral tilt. the general orientation of spectral energy, or the slant of the spectral envelope.

Spectrogram. a graph of acoustic analysis consisting of three dimensions. Typically, time is represented on the horizontal axis, the frequency (pitch) of sound on the vertical axis, and the intensity of sound on the gray scale (shades of darkness). This is a basic form of analysis used in acoustic studies of speech, including talker identification.

Vertical striations. the fine-grained lines that run vertically on a wide-band spectrogram and correspond to the vocal pulses of voicing energy. The presence of these striations is an indication of voicing, and their relative pattern describes properties of the voice source, such as vocal pitch and its regularity.

Vocal folds. muscles and associated tissues located in the larynx and capable of vibrating to produce voice.

Vocal tract. the cavities and structures above the vocal folds that are capable of modifying voice and airflow into distinctive sounds of speech; sometimes called the articulatory or articulatory-resonatory system.

Vocal tract resonances. the natUl:al or resonant frequencies of the vocal tract, which can be determined mathematically from a knowledge of the length and shape of the vocal tract; also called formants.

Voice. the tone produced by vibration of the vocal folds and modified by adjustments of the vocal tract; sometimes used in a general sense to refer to human speech.

Signal detection theory. a mathematical theory dealing with the problem of detecting the presence of a signal in a background of noise. Early applications of this theory sought to statistically measure the accuracy of instrumental and sensory systems, independent of biases arising from the decision. criteria of observers. Decision criteria. are influenced by the observer's knowledge of the prior probabilities of target events (signals), and by the values and costs associated with correct and incorrect decisions. Accuracy is indexed graphically by means of *relative operating characteristic* (ROC) curves, and quantitatively by measures of performance such as <L and A- Subsequent applications of the theory have included problems in medical diagnosis, aptitude testing, weather forecasting, automated information retrieval, and forensic identification.

Spectral envelope. the general shape of the spectrum (the distribution of signal energy as a function of frequency). The envelope often is drawn to represent the overall form of the spectrum and may neglect minor deviations from that form.

Spectral tilt. the general orientation of spectral energy, or the slant of the spectral envelope.

Spectrogram. a graph of acoustic analysis consisting of three dimensions. Typically, time is represented on the horizontal axis, the frequency (pitch) of sound on the vertical axis, and the intensity of sound on the gray scale (shades of darkness). This is a basic form of analysis used in acoustic studies of speech, including talker identification.

Vertical striations. the fine-grained lines that run vertically on a wide-band spectrogram and correspond to the vocal pulses of voicing energy. The presence of these striations is an indication. of voicing, and their relative pattern describes properties of the voice source, such as vocal pitch and its regularity.

Vocal folds. muscles and associated tissues located in the larynx and capable of vibrating to produce voice.

Vocal tract. the cavities and structures above the vocal folds that are capable of modifying voice and airflow into distinctive sounds of speech; sometimes called the articulatory or articulatory-resonatory system.

Vocal tract resonances. the natural or resonant frequencies of the vocal tract, which can be determined mathematically from a knowledge of the length and shape of the vocal tract; also called formants.

Voice. the tone produced by vibration of the vocal folds and modified by adjustments of the vocal tract; sometimes used in a general sense to refer to human speech.