Likelihood-Ratio-Statistik durch, die auch in anderen Wissenschaftsdisziplinen (z.B. in der Biologie zum DNS-Abgleich) angewandt wird.

18.4 Die grundlegende Heuristik

Akzeptiert man, dass die Unterscheidbarkeit aller Menschen aufgrund ihrer Stimme/Sprechweise (nur) axiomatisch postuliert werden kann (vgl. French et al., 2007), so ist beim Stimmenvergleich ein weiterer Aspekt mündlicher Äußerungen zu beachten: Natürliche Äußerungen sind *nie* akustisch gleich, auch nicht, wenn sich ein und derselbe Sprecher allergrößte Mühe gibt, das Gleiche wiederholt zu sagen. Das aufgezeichnete elektro-akustische Abbild wird immer Unterschiede aufweisen. Dies trifft sogar auf nahezu alle einzelnen akustischen Parameter zu, die aus Äußerungen extrahiert werden können. Die Frage, ob zwei Aufnahmen vom selben Sprecher stammen, muss also immer auf der Grundlage variierender Messungen erfolgen. Und ein unbekannter Teil der Faktoren dieser Variation (z.B. das Aufnahmemikrophon, der Aufnahmeraum, etc.) ist und bleibt unklar. Darüber hinaus ist nahezu immer anzunehmen, dass sich der Variationsbereich eines akustischen Merkmals eines Sprechers mit denen der zu vergleichenden anderen Sprecher überschneidet.

Trotztdem besteht die Möglichkeit, durch Kombination mehrerer akustischer Parameter im multi-dimensionalen Merkmalsraum für einzelne Sprecher separate Räume zu finden, sie also von anderen unterscheiden zu können (vgl. Abbildung 18.1): M_1 und M_2 seien zwei an zwei Sprechern (Dreieck und Kreis) jeweils 3-mal gemessene (akustisch-phonetische) Merkmale, anhand derer man – nimmt man jedes für sich alleine – die Sprecher nicht unterscheiden kann. Als Dimensionen spannen sie jedoch eine Ebene auf, in der eine Linie (hier sogar eine Gerade) gefunden werden kann, die die Sprecher voneinander trennt. Man braucht also theoretisch nur genügend Merkmale, um für jeden Sprecher einen separaten Bereich im x-dimensionalen Merkmalsraum zu finden. Voraussetzung für eine Modellierung der Merkmale als aufeinander orthogonale Dimensionen ist allerdings, dass die Merkmale *unabhängig* voneinander sind. Aber kaum zwei Merkmale, die an mündlichen Äußerungen erhoben werden können, sind völlig unabhängig voneinander! Zwar bedeutet Abhängigkeit für das Modell „nur“, dass die Merkmalsachsen eben Winkel kleiner 90° einschließen; allerdings ist regelmäßig das Ausmaß der Abhängigkeit unklar und praktisch ist eine solche Modellierung ungleich komplizierter.

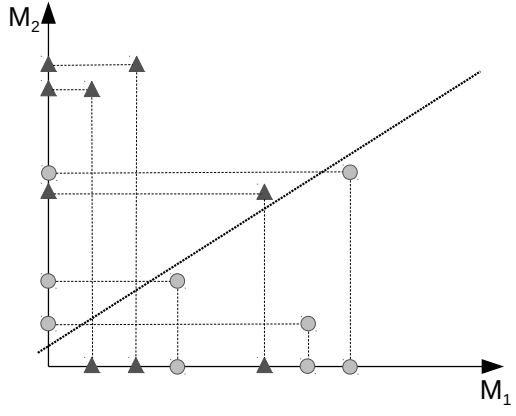

Abbildung 18.1: Die Grundidee des multi-dimensionalen Merkmalsraumes.

Aus Abbildung 18.1 ist auch ersichtlich, dass es nur sinnvoll ist, Merkmale zu verwenden, die im Mittel zwischen den Sprechern unterschiedliche Ausprägungen aufweisen, die also eine möglichst große Varianz zwischen verschiedenen Sprechern haben. Gleichzeitig wäre wünschenswert, dass das Merkmal innerhalb eines Sprechers möglichst einheitlich ausgeprägt ist, also eine kleine intra-Sprecher-Varianz aufweist.

Somit sind wichtige Schritte eines phonetisch-forensischen Stimmenvergleichs:

1. das auditive Auffinden von im Tatmaterial besonders/extrem ausgeprägten Merkmalen und die Bestimmung der entsprechenden Parameter als relevant für das Gutachten;

2. das auditive/qualitative Bestimmen von Signalteilen im Tat-, Vergleichs- und Hintergrundmaterial, in denen sich diese Merkmale finden, als gleich bzw. vergleichbar (z.B. durch Segmentation und enge phonetische Transkription), um diese sinnvoll einer akustischen Analyse zuführen zu können;

3. die vorzugsweise instrumental-akustische Messung dieser Merkmale im Tat- und im Vergleichsmaterial als Grundlage für die Schätzung der sprecher-internen Verteilungen (z.B. über Mittelwert und Varianz unter Normalverteilungsannahme);

4. die Abschätzung der Verteilung des Merkmals in der (Hintergrund-) Population (siehe Abschnitt 18.8);

5. die inferenz-statistische Bewertung der erhobenen Merkmale als Indizien durch Bestimmung der konkreten mathematischen Modellierung, sowie Schätzung der gegenseitigen Abhängigkeiten der akustischen Parameter in der Hintergrundpopulation und schließlich Inbezugsetzung des Täter-, Beschuldigten- und Hintergrundmodells.

18.5 Besonders geeignete Merkmale

Eigentlich müssen die zu einem forensischen Sprechervergleich „besonders geeigneten“ Merkmale in jedem neuen Fall neu bestimmt werden. Das ist geboten, weil jeder Sprecher „seine Einzigartigkeit“ nur durch individuelle Ausprägungen oftmals anderer Parameter (-kombinationen) erzielen kann – sofern er überhaupt eine einzigartige Stimme/Sprechweise hat. Das liegt daran, dass es (bislang) keine stimmlichen Merkmale gibt, die immer und bei jedermann einzigartige Muster ergäben, so wie das bei der DNS oder dem Fingerabdruck der Fall ist. Ganz im Gegenteil: es ist vielmehr so, dass nicht nur alle Merkmale der Stimme und Sprechweise von Mal zu Mal innerhalb eines Sprechers variieren, sondern dass diese Varianz oft annähernd so groß ist wie die Varianz dieser Merkmale zwischen den Sprechern. Hinzu kommt, dass Tataufnahmen regelmäßig eher zu kurz und von zu schlechter Qualität sind, als dass sie durchwegs verlässliche Merkmalsextraktionen erlauben würden. Weiterhin ist zu berücksichtigen, dass bestimmte Teile ihres Zustandekommens regelmäßig unbekannt bleiben (z.B. Mikrophon und Raum der Aufnahme bei einem Erpresseranruf) und dass davon auszugehen ist, dass sich ein Täter verstellt haben könnte und dass dabei unklar ist, ob und wie er das gemacht hat. All diese speziell forensisch-phonetischen Bedingungen machen es erforderlich, dass alles, was da ist und verlässlich erscheint, genutzt werden muss. Und das ist nun einmal von Fall zu Fall ziemlich verschieden.

Davon abgesehen gibt es natürlich Kriterien, nach welchen die Eignung bestimmter Parameter(-sätze) zur Sprecheridentifikation abgeschätzt werden kann:

Die geeigneteren Merkmale sind die, welche

- willentlich schwer zu verändern sind, also eine hohe Resistenz gegenüber Verstellung aufweisen;
- eher anatomische Gegebenheiten des Sprechapparats abbilden, weniger Erlerntes;
- die Stimme abbilden; weniger geeignet sind Merkmale, welche die Sprechweise (Prosodie, Idiolekt, Dialekt) erfassen, noch weniger geeignet sind Merkmale der speziellen Verwendung der gerade benutzten Sprache;
- verlässlicher zu erheben (reliabler) sind;
- stringenter und valider zu deuten sind, da deren Variation möglichst eindeutig auf möglichst wenige Faktoren zurückgeführt werden kann;
- hohe Robustheit bei verschiedenen (also auch unbekannten) Übertragungswegen aufweisen.

Betrachtet man nun akustische Parameter, die oftmals im forensischen Kontext eingesetzt wurden, unter dem Licht dieser Kriterien, so wird deutlich, dass es bislang keine stets optimalen Parameter gibt:

Zu den generell vielversprechendsten zählen dabei die Formanten, insbesondere die Vokalformanten (vgl. die Abschnitte 10.4 und 10.5). In diesen Resonanzfrequenzen (in ihren Mittenfrequenzen, Bandbreiten und relativen Intensitäten) bilden sich die räumlichen Verhältnisse im Artikulationstrakt und die Schallhärte begrenzender Gewebeschichten ab, also relativ individuelle anatomische Gegebenheiten. Vor allem die tiefsten beiden Vokalformanten werden aber auch extensiv durch die linguistische Information von Redeäußerungen moduliert. Sofern der linguistische Gehalt über die zu vergleichenden Äußerungen hinweg konstant gehalten werden kann, findet sich in F_1 und F_2 also neben der Vokalqualität und der Ansatzrohrgeometrie gleichzeitig und untrennbar auch Information über dia-, sozio- und idiolektale Aussprachevarianten.

Höhere Vokalformanten sind weit weniger durch die verschiedenen Vokalqualitäten und Aussprachevarianten beeinflusst, tragen also potentiell mehr Information über die Sprecheridentität. Diese sind aber nicht mehr so reliabel zu bestimmen. Mehrere Formanten zu verwenden, um damit einen Merkmalsraum aufzuspannen, ist problematisch, weil sie alle natürlich in erster Linie durch die Ansatzrohrlänge beeinflusst, also hochgradig abhängig sind (vgl. Abbildung 10.6 und s.u.). Nasalformanten – also die Resonanzen, die die geometrischen Eigenschaften der Nasenhohlräume abbilden – sind generell zur Sprecheridentifikation noch besser geeignet, weil die Nasenräume noch individueller geformt sind. Sie sind aber fortwährend durch unterschiedliche Schleimhautmengen und -konsistenzen beeinflusst. Und: alle Formanten können durch Gegenstände im Mund- oder Nasenraum relativ einfach verfälscht werden.

Cepstral-Koeffizienten[2] mögen gegenüber den Formanten den Vorteil haben, dass sie den Filter (lt. Quelle-Filter-Modell, vgl. Abbildung 10.7), also optimalerweise (!) die räumlichen Verhältnisse im Ansatzrohr, noch besser abbilden als die Formanten und dabei sogar relativ unabhängig voneinander ausfallen. Gerade im forensischen Kontext sind sie aber sehr problematisch, weil konkrete Koeffizienten(-ausprägungen) kaum zu deuten sind und in die Gesamtheit der Koeffizientenausprägungen zusätzlich Raum, Mikrophon und Übertragungsweg einfließen, die im forensischen Fall gewöhnlich nicht zu kontrollieren sind. Dasselbe Problem betrifft auch LPC-Koeffizienten.

Die (mittlere) Grundfrequenz F_0 bildet zwar einerseits anatomische Gegebenheiten wie die Länge und die Masse der Stimmlippen ab, ist also prinzipiell geeignet, um zur Identifikation eingesetzt zu werden. Dennoch ist sie (vgl. Kap. 6) innerhalb eines Sprechers auch durch sehr viele Faktoren beeinflusst und somit sehr variabel[3]. Diese Varianz ist oft annähernd so groß wie die Varianz zwischen gleichgeschlechtlichen Sprechern, wodurch die F_0 nur bei einer sehr extremen Ausprägung beim Täter zu seiner Identifikation brauchbar ist. Zudem ist bekannt, dass sie in hohem Maße vom Grad der Erregung abhängt, die wiederum durch den emotionalen Zustand (vgl. Kap. 5 und 14) bedingt ist. Und Täter befinden sich während der Tat fast immer in einem anderen emotionalen Zustand als bei Vergleichsaufnahmen. Diese Problematik trifft sogar noch verstärkt auf die Intonation als einem Hauptbestandteil der Prosodie (vgl. Kap. 13 und 14) zu, wie auch auf Maße der Stimmqualität (vgl. Abschnitt 10.8).

Die Sprech- bzw. Artikulationsgeschwindigkeit ist zwar sehr reliabel zu erheben und auch sehr stringent zu deuten, jedoch vielfältigen Einflüssen unterworfen und leicht willentlich zu verändern.

---

[2] oder auch (Mel-frequency) cepstral coefficients, MFCCs

[3] Innerhalb einer Person variiert die F_0 abhängig vom Alter, Gesundheitszustand, Grad der Intoxikation durch Koffein, Alkohol, Nikotin, etc., vom Adressaten der Äußerung, emotionalen Zustand (hier v.a. von der Erregung), von der Tageszeit, Sprechlautstärke, usw.

Dialekte/Soziolekte werden zwar in der Kindheit und Pubertät erworben und sind in Deutschland sehr divergent, eignen sich somit bedingt als Identifikationsmerkmal. Allerdings ist ein „neuer Dialekt“ auch später erlernbar. Relativ viele Menschen können zudem zwischen verschiedenen Dia- oder Soziolekten (ähnlich wie zwischen Sprachen) wechseln. Zwar gelingt das oftmals nicht überzeugend, reicht aber, um evtl. Schlüsse auf die eigentliche Herkunft zu verschleiern oder um Vergleiche bzgl. segmenteller Messungen (z.B. von Formanten) zu erschweren.

Ein guter Parameter*satz* besteht aus möglichst vielen geeigneten Parametern, die zusätzlich – und jetzt wird's ein Dilemma – in der Hintergrund-Population an Sprechern voneinander *möglichst unabhängig* sein sollten, damit die mit Abbildung 18.1 veranschaulichte Modellierung durch einen multi-dimensionalen Merkmalsraum angewendet werden kann. Sind die verwendeten Parameter (hochgradig) abhängig, so führt dieses Modell ohne weitere Maßnahmen[4] zu einer hochgradigen Überschätzung der Diskriminationsfähigkeit zwischen den Sprechern und somit oftmals zu einer extremen Begünstigung der Sichtweise der Anklage! Und nochmals Vorsicht: die meisten Parameter der akustischen Phonetik sind voneinander abhängig!

18.6 Bayes'sche Statistik zur Bewertung von Indizien

Ist ein Parametersatz gefunden und die Werte im Tat- und Vergleichsmaterial bestimmt, so lautet die beim Stimmenvergleich für das Gericht zu beantwortende Frage: Wie wahrscheinlich stammt das Tatmaterial (z.B. ein Erpresseranruf) vom Beschuldigten, von dem akustisches Vergleichsmaterial vorliegt?

Diese Frage lässt sich formalisieren als Frage nach der Wahrscheinlichkeit p dafür, dass die Hypothese der Anklage H_A (Täter und Beschuldigter sind identisch) zutrifft, unter Berücksichtigung des Indizes I (der Entsprechung des Tat- und Vergleichsmaterials): $p(H_A|I)$. Das ist allerdings nur die „Modellierung der Anklage“. Genau betrachtet muss das Gericht aber zwischen zwei (zunächst) gleichwertigen Erklärungsmodellen abwägen: der H_A und der Hypothese der Verteidigung H_V (Täter und Beschuldigter sind nicht identisch – oder ganz genau: Irgendjemand sonst aus der „Hintergrundpopulation“ hat das Tatmaterial gesprochen). Das erforderliche Abwägen kann

[4] Um Abhängigkeiten berücksichtigen zu können, ist es nötig, den Grad der gegenseitigen Abhängigkeiten (z.B. über einen Korrelationskoeffizienten) der verwendeten Merkmale in der jeweiligen (Hintergrund-) Population zu schätzen.

durch ein Verhältnis der entsprechenden Wahrscheinlichkeiten formalisiert werden:

$$\frac{p(H_A|I)}{p(H_V|I)} = Odds_{post} \tag{18.1}$$

Ist die Hypothese der Anklage H_A (unter Berücksichtigung des Indizes I) wahrscheinlicher, so ist dieses Verhältnis größer als 1. Generell ist es umso größer, je wahrscheinlicher die H_A im Verhältnis zur H_V ist. Das ist also genau das, was das Gericht optimalerweise braucht, um zu einer Entscheidung zu kommen – optimalerweise, weil definitive Aussagen, wie oben (unter Abschnitt 18.2) erläutert, unmöglich sind. Mit der Odds-Form[5] des Satzes von Bayes (1763) ist dieses Verhältnis, die posterior odds ($Odds_{post}$), berechenbar:

$$\begin{aligned} Odds_{post} &= Odds_{prior} \cdot LR \\ \frac{p(H_A|I)}{p(H_V|I)} &= \frac{p_{prior}(H_A)}{p_{prior}(H_V)} \cdot \frac{p(I|H_A)}{p(I|H_V)} \end{aligned} \tag{18.2}$$

Die Odds für die Hypothesen nach der Berücksichtigung des Indizes ergeben sich aus dem Produkt der Odds von zuvor und dem Plausibilitätenverhältnis (Likelihood Ratio, LR). Zu dieser Berechnung sind also zwei Terme nötig: der LR und die $Odds_{prior}$, bestehend aus dem Verhältnis der a-priori-Wahrscheinlichkeiten für die Richtigkeit der Hypothesen. Betrachtet man den LR, so wird deutlich, dass er die logische Umkehrung der $Odds_{post}$ ist: das Verhältnis der bzw. die Wahrscheinlichkeiten dafür, dass man das Indiz I vorfindet, sofern die eine oder die andere Hypothese zutrifft. Das lässt sich berechnen oder zumindest abschätzen.

Es bleibt aber die Frage, woher man die $Odds_{prior}$ nehmen soll. Hätte man eine klare Vorstellung davon, so wäre doch die gesamte Rechnung hinfällig; dieser Umstand wurde oft als entscheidender Nachteil Bayes'scher Statistik gesehen. Hier aber ist es eher ein großer Vorteil, denn über diesen Term ist es *möglich*, bisheriges Wissen in eine vorläufig finale, kumulative Entscheidung einfließen zu lassen. Und genau dadurch lassen sich auch Erkenntnisse aus verschiedenen phonetischen Parametern zusammenfassen und dann auch noch mit Ergebnissen aus anderen Bereichen (wie Haaranalysen, Fingerabdrücken, DNS-Spuren, etc.) akkumulieren, sofern sie *voneinander unabhängig* sind. Ein weiterer Vorteil der Bayes'schen Statistik

[5] „Odds“ ist wohl am besten durch „(Wett-) Quoten“ oder auch „Chancen“ übersetzt, „Likelihood“ durch „Plausibilität“. Da diese Übersetzungen aber nicht gebräuchlich sind, werden fortan die englischen Originaltermini verwendet.

gegenüber der „frequentistischen“ im forensischen Kontext besteht in der formalen Gleichwertigkeit der beiden zu prüfenden Hypothesen. Hierdurch besteht keine Notwendigkeit, vorab irgendein beliebiges Signifikanzniveau (und evtl. eine minimale Effektgröße) zu bestimmen, bei dessen Unterschreitung dann kategorisch eine der Hypothesen, die Gleichheits-Hypothese, abzulehnen wäre.

In der konkreten forensischen Entscheidung kann das Gericht die $Odds_{prior}$ also aus anderen Indizien (DNS, Haaren, Fingerabdrücken, etc.) oder aus anderen Überlegungen heraus abschätzen. Zu Beginn eines Hauptverfahrens kann eine (knapp zugunsten der Anklage tendierende) 50:50-Wahrscheinlichkeit, also $Odds_{prior} \geq 1$ angenommen werden, da es nur zu einem Hauptverfahren kommt, sofern „ein hinreichender Tatverdacht“ besteht, eine Verurteilung also als (minimal) wahrscheinlicher gilt als keine. Umgekehrt folgt hieraus für den forensischen Sprachgutachter, dass weder die Bestimmung der $Odds_{prior}$ und schon gar nicht die der $Odds_{post}$ in seinen Zuständigkeitsbereich fallen. Seine Aufgabe ist die Berechnung des LR.

18.7 Die Berechnung eines Likelihood Ratios

Nehmen wir beispielsweise an, ein Erpresser hätte beim Tatanruf gestottert. Wenn auch der (oder die) Beschuldigte stottert (I sei „Stottern“), ist die Wahrscheinlichkeit, dass der Beschuldigte stottert, unter der Annahme, sie/er sei auch der Täter (H_A), trivialerweise $p(I|H_A) = 1$. Die Punktprävalenz des Stotterns liegt Schätzungen zufolge bei (ca.) 1% der Bevölkerung. Somit beträgt auch die Wahrscheinlichkeit, dass der/die Beschuldigte stottert, obwohl angenommen wird, dass sie/er nicht der Täter ist $p(I|H_V) = 1\%$. Hieraus folgt $LR_{Stottern} = 1/1\% = 100$ (vgl. Lucy, 2006).

Der absolute Wert des LR lässt sich auf die selbe Weise intuitiv interpretieren wie die $Odds_{post}$ (siehe oben), nur ist ein LR eben ein Maß der Evidenz[6], die I zugunsten der H_A liefert: Im Beispiel ist es also 100-mal plausibler, dass der Beschuldigte stottert, wenn man davon ausgeht, er sei auch der Täter (als dass der Beschuldigte stottert, wenn man annimmt, dass er nicht der Täter ist).

[6] Das englische „evidence“ meint dasselbe wie „Indiz(-ien)“ im Deutschen; das deutsche „Evidenz (eines Indizes)“ wird im Englischen als „strength (of evidence)“ bezeichnet!