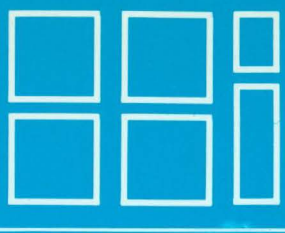


Heinrich-Hertz-Institut
für Nachrichtentechnik
Berlin GmbH

Technischer Bericht Nr. 204

GRENZEN UND MÖGLICHKEITEN DER AUTOMATISCHEN SPRECHERERKENNUNG

von

Peter Jesorsky

Technischer Bericht Nr. 204

GRENZEN UND MÖGLICHKEITEN DER AUTOMATISCHEN SPRECHERERKENNUNG

von

Peter Jesorsky

Dr.-Ing. P. Jesorsky
(Bearbeiter)

Dr. H. Ohnsorge
(Wiss.Techn. Geschäftsführer)

Dr. K.H. Vöge
(Abteilungsleiter)

Berlin-Charlottenburg, den 5. Februar 1979
ISBN 0342-6815

INHALT

1. EINLEITUNG

1.1 Definition "Sprechererkennung"

Unter dem Begriff "Sprechererkennung" versteht man die Erkennung einer Person anhand ihrer Stimme. Soll dieser Erkennungsvorgang automatisch durchgeführt werden, dann muß die Stimme meßtechnisch erfaßt, analysiert und klassifiziert werden. Im allgemeinen wird der zeitliche Schalldruckverlauf, der sich bei einer sprachlichen Äußerung ergibt, durch ein Mikrofon in ein elektrisches Signal s(t) umgewandelt, das hier kurz als Sprachsignal bezeichnet werden soll. Die Aufgabe der automatischen Sprechererkennung ist es also, aus einem Sprachsignal s(t) eine Entscheidung über den Sprecher abzuleiten.

1.2 Sprecheridentifizierung und -verifizierung, Anwendungen

Automatische Sprechererkennungssysteme können aus heutiger Sicht in zwei Bereichen eingesetzt werden, im kriminalistischen Bereich zur Täterermittlung und in Sicherheitssystemen (Zugangskontrolle zu Sicherheitsbereichen, Zugriffskontrolle zu vertraulichen Informationen, Legitimation von finanziellen Transaktionen) zur Überprüfung von Personen.

Die bei diesen Anwendungen auftretenden Randbedingungen und damit auch die verwendeten Methoden weichen z.T. voneinander ab und sollen deshalb diskutiert werden. Für ausführliche grundlegende Darstellungen zur automatischen Sprechererkennung wird auf die Referenzen /1, 2, 3/ verwiesen.

Im kriminalistischen Bereich legt die zu erkennende Person - sofern sie eine Straftat begeht - größten Wert darauf, anonym zu bleiben. Aus diesem Grund wird die Stimme z.B. bei Telefongesprächen (Erpressungsversuchen, Bombendrohungen usw.) häufig verstellt. Ist der Täter einmal erfaßt, so wird er alles tun, damit

eine von ihm geforderte Sprachprobe möglichst wenig Ähnlichkeit mit der als Beweis vorliegenden Tonbandaufnahme aufweist. Das Problem, das als Sprecheridentifizierung bezeichnet wird, ist es also, unter Berücksichtigung einer mutmaßlichen Verstellung aus einem eingeschränkten Personenkreis (Tatverdächtige) diejenige Person herauszufinden, deren Stimme die größte Ähnlichkeit mit der gespeicherten Bandaufnahme hat. Erschwerend kommt hinzu, daß meist weder kontrollierbare noch reproduzierbare äußere Bedingungen wie Hintergrundgeräushce, durch die Telefonübertragung verursachte Verzerrungen usw. die sprecherspezifischen Eigenschaften überdecken und die Sicherheit der Identifizierung beeinträchtigen.

Im Gegensatz zur Anwendung in der Kriminalistik kann man bei Verwendung der Stimme als Legitimation in Sicherheitssystemen davon ausgehen, daß die sprechende Person daran interessiert ist, erkannt zu werden, sich also kooperativ verhält. Diese Kooperationsbereitschaft läßt sich ausnutzen. Wenn die zu erkennende Person gleichzeitig mit der Sprachprobe auch ihren Namen, d.h. ihr Identitätsziel angibt, dann muß das Erkennungssystem nicht mehr entscheiden, welchem (von vielen) Sprechern diese Sprachprobe zugeordnet werden soll, sondern lediglich, ob die Sprachprobe (bei genügender Ähnlichkeit) dem Identitätsziel zugeordnet werden kann oder nicht. Diese Vorgehensweise wird als Sprecherverifizierung bezeichnet.

Die Unterschiede zwischen Identifizierung und Verifizierung sollen verdeutlicht werden. Bei der Sprecheridentifizierung geht es darum, eine unbekannte Sprachprobe einem von K Sprechern zuzuordnen (Fig. 1.1), es liegt ein K-Klassenproblem der Mustererkennung vor (Sprecher ω_1 oder Sprecher ω_2 oder ... oder Sprecher ω_K). Das System hat dabei keinerlei Vorinformation über die Identität des Sprechers.

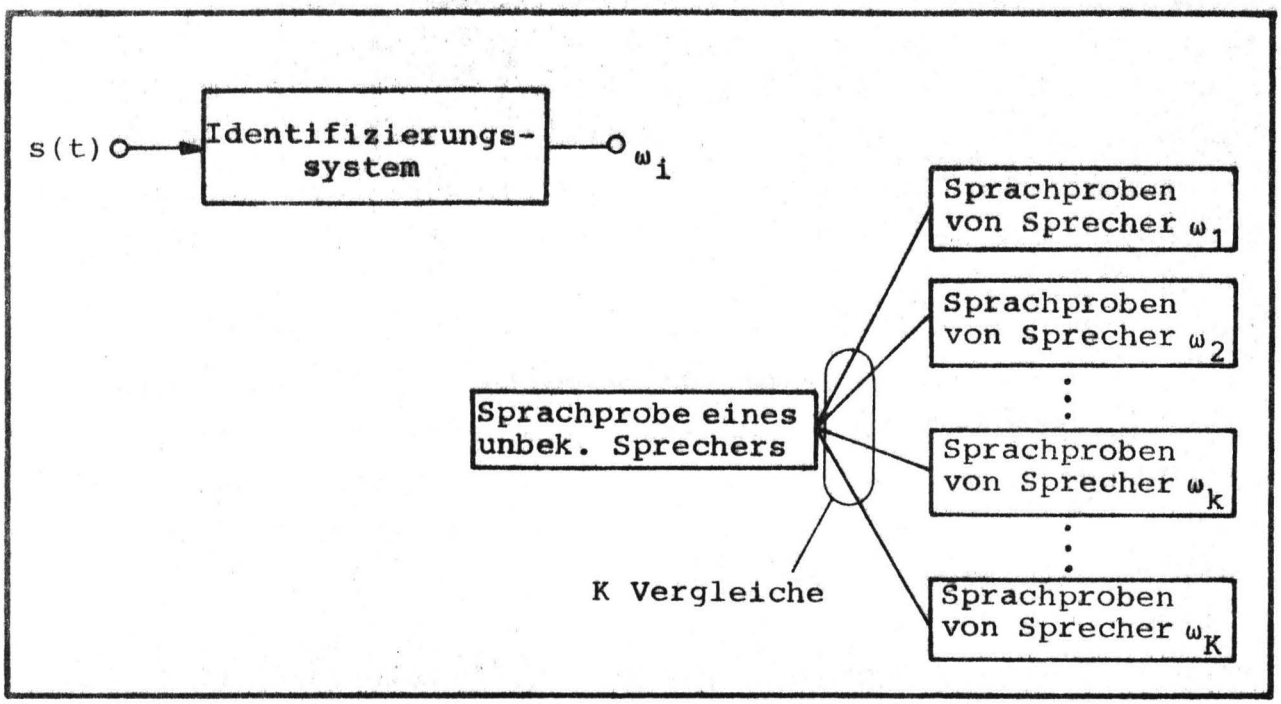


Fig. 1.1 Prinzip der Sprecheridentifizierung

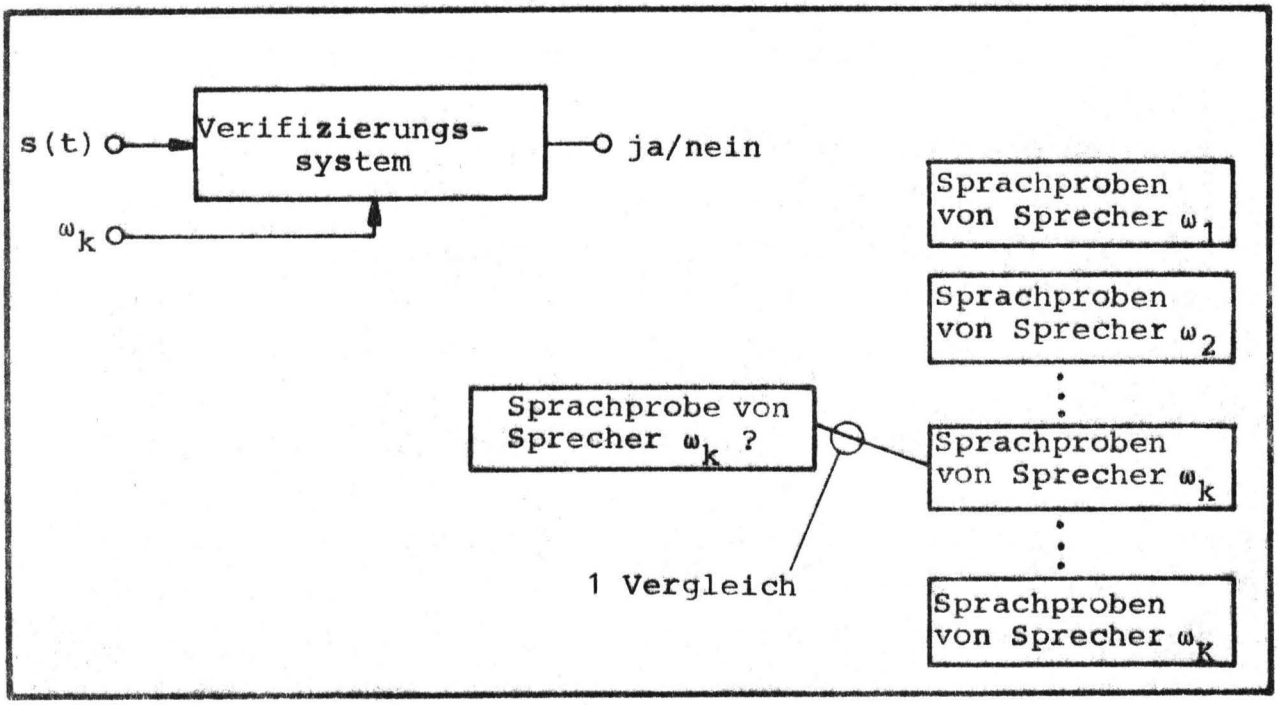


Fig. 1.2 Prinzip der Sprecherverifizierung

Bei der Sprecherverifizierung dagegen wird nur die Ähnlichkeit der Sprachprobe mit den Vergleichsdaten des Identitätsziels ω_k ermittelt (ein Vergleich!) und danach die Sprachprobe entweder angenommen oder zurückgewiesen (Fig. 1.2), es liegt ein 2-Klassenproblem der Mustererkennung vor (Sprecher ω_k, nicht Sprecher ω_k).

Die Leistungsfähigkeit eines Identifizierungssystems läßt sich durch eine einzige Zahl, die Fehlerrate ε beschreiben, die aus einer vorgegebenen Stichprobe als Quotient der Falschklassifizierungen zur Gesamtzahl der Klassifizierungen abgeschätzt werden kann.

Bei der Sprecherverifizierung muß man dagegen zwischen zwei Arten von Fehlern unterscheiden. Ein Fehler tritt auf, wenn

a) ein "wahrer Sprecher" nicht vom System akzeptiert wird. Als Fehlerrate ε_{FR} (False Reject) definiert man hier die Zahl der Rückweisungen wahrer Sprecher im Verhältnis zur Gesamtzahl der Verifizierungsversuche wahrer Sprecher.

b) ein Täuschungsversuche erfolgreich ist, d.h. ein Sprecher mit dem Identitätsziel eines anderen Sprechers vom System akzeptiert wird. Als Fehlerrate ε_{FA} (False Accept) definiert man die Zahl der erfolgreichen Täuschungsversuche bezogen auf die Gesamtzahl der Täuschungsversuche.

Beide Fehlertypen ε_{FR} und ε_{FA} sind eine Funktion der vom System geforderten "hinreichenden Ähnlichkeit". Ist diese Forderung sehr streng, dann werden zwar kaum Täuschungsversuche gelingen, aber auch wahre Sprecher werden wegen der mangelnden Reproduzierbarkeit ihrer Stimmen häufiger zurückgewiesen. Läßt man dagegen auch größere Abweichungen zu, dann werden zwar kaum noch wahre Sprecher zurückgewiesen, das wird aber mit einer größeren Zahl von erfolgreichen Täuschungsversuchen erkauft.

Ein weiterer Unterschied zwischen Identifizierung und Verifizierung besteht darin, daß die Erkennungssicherheit bei der Identifizierung kontinuierlich mit der Zahl der Sprecher abnimmt, also bei großem Personenkreis sehr unzuverlässig wird (wegen der endlichen Verwechslungswahrscheinlichkeit zwischen 2 beliebigen Sprechern). Bei der Verifizierung ist die Erkennungsrate dagegen von der Zahl der Sprecher unabhängig (immer nur 1 Vergleich). Deshalb können Verifizierungssysteme praktisch beliebig viele Benutzer versorgen.

Viele geschäftliche Vorgänge, deren Absicherung bisher durch eine Unterschrift (Kontenbewegungen, verbindliche Bestellungen) oder durch geheime Schlüsselwörter (geschützte Informationen einer Datenbank, autorisierte Befehlsgabe) erfolgte, lassen sich mit einem Sprecherverifizierungssystem durch einen "Stimmabdruck" absichern. Ein besonderer Vorteil solch einer Sicherung ist darin zu sehen, daß mit der Stimme ein personengebundenes Kennzeichen zur Legitimation verwendet wird, bei dem die persönliche Anwesenheit des zu erkennenden Sprechers nicht erforderlich ist. Der "Stimmabdruck" kann also auch über einen Telefonkanal übertragen werden. Dies vereinfacht die Bearbeitung von eiligen geschäftlichen Vorgängen und bewirkt einen Zeitvorteil. Der Nachteil von Schlüsselwörtern, die unter Umständen auch von Unbefugten verwendet werden können, wird durch die Personengebundenheit des "Stimmabdrucks" vermieden.

Zusammenfassend läßt sich sagen, daß die automatische Sprecherverifizierung eine wesentlich einfachere Aufgabe darstellt als die Sprecheridentifizierung. Wegen der Kooperationsbereitschaft des zu verifizierenden Sprechers können z.B. fest Codesätze verabredet werden, aus denen sprechertypische Informationen viel leichter als aus einem beliebigen Text gewonnen werden können. Die Nennung eines Identitätszieles ermöglicht die Verwendung spezieller Klassifizierer (2-Klassenproblem), die besonders bei einer großen Zahl von Sprechern zuverlässiger arbeiten. Schließlich können die äußeren Bedingungen (z.B. Störgeräusche) unter Mitwir-

kung des Benutzers so weit dem System angepaßt werden, daß eine ausreichende Zuverlässigkeit des Verifizierungssystems gewährleistet ist.

1.3 Menschliche Leistungsfähigkeit

Beim Entwurf eines Systems zur automatischen Sprechererkennung drängt sich natürlich sofort die Frage auf, wie groß die Leistungsfähigkeit eines solchen Systems gegenüber der menschlichen Fähigkeit ist, Personen anhand ihrer Stimmen zu erkennen. Jeder Leser hat sicherlich die Erfahrung gemacht, daß er eine ihm bekannte Person am Telefon anhand ihrer Stimme erkannt hat, obwohl sie ihren Namen noch nicht genannt hatte. Auf der anderen Seite sind aber durch Funk und Fernsehen professionelle Stimmimitatoren bekannt, die die Stimmen anderer Personen täuschend ähnlich nachahmen können und auf diese Weise die Grenzen der menschlichen Leistungsfähigkeit aufzeigen. Ein Vergleich zwischen dem Menschen und einem automatischen System wurde von Rosenberg für die Sprecherverifizierung durchgeführt /4/. Diese Untersuchung soll kurz erläutert werden und die wesentlichen Ergebnisse sollen zusammengefaßt werden.

Es wurden 3 Arten von Sprachproben verwendet

(a) Sprachproben von 8 wahren Sprechern
(b) Sprachproben von 4 anerkannten Imitatoren (natürliche Sprechweise)
(c) Sprachproben von 4 anerkannten Imitatoren (gezielte Täuschungsversuche)

Die Sprachproben wurden einmal automatisch klassifiziert und zum anderen in einem subjektiven Hörtest von einer Gruppe von Personen klassifiziert, denen jeweils 2 Sprachproben der Form

(a) - (a) reguläre Verifizierung
(a) - (b) zufälliger Täuschungsversuch
(a) - (c) beabsichtiger Täuschungsversuch

angeboten wurden und die dann zu entscheiden hatten, ob beide Sprachproben vom selben Sprecher stammten. Die Ergebnisse sind in Tab. 1.1 zusammengefaßt:

	reguläre Verifizierung (a)-(a) ε_{FR} (%)	zufälliger Täuschungsversuch (a)-(b) ε_{FA} (%)	beabsichtigter Täuschungsversuch (a)-(c) ε_{FA} (%)
automatisches Verifizierungssystem	0	1	14
subjektiver Hörtest (gemittelt über alle Hörer)	2,6	4,2	22

Tab. 1.1 Vergleich der menschlichen Leistungsfähigkeit mit der eines automatischen Erkennungssystems

Es zeigt sich, daß die Leistungsfähigkeit des automatischen Systems dem subjektiven Urteil des Menschen überlegen ist, bei allen 3 Kombinationen sind die Fehlerraten kleiner. Die Fehlerraten sind dabei über alle Sprecher und alle Hörer gemittelt. Interessant sind dabei natürlich auch die Streuungen sowohl bei den Hörern als auch bei den Imitatoren. So waren z.B. beim besten Hörer nur 4% der beabsichtigten Täuschungsversuche erfolgreich (Mittelwert 22%). Auch die Leistungsfähigkeit der Imitatoren war recht unterschiedlich. Während es dem erfolgreichsten Imitator gelang, die Hörer in 38% aller Fälle zu täuschen, lag diese Ra-

te beim schlechtesten Imitator bei nur 7% (Mittelwert 22%).

Aus Tab. 1.1 geht deutlich hervor, wie groß der Unterschied zwischen zufälligen und beabsichtigten Täuschungsversuchen - sowohl im subjektiven Test wie auch bei der automatischen Erkennung - ist. Die Bedeutung dieser Angabe sollte man sich bei der Beurteilung der Leistungsfähigkeit von Erkennungssystemen anhand von Fehlerraten stets vor Augen halten.

Abschließend soll darauf hingewiesen werden, daß das mit dem Erkennungssystem erzielte Verhältnis der Fehlerrate von beabsichtigten zu zufälligen Täuschungsversuchen (14%/1%) sich auf ein spezielles System bezieht und nicht verallgemeinert werden darf, da es stark von denjenigen Eigenschaften des Sprachsignals und deren Imitierbarkeit abhängt, die zur Klassifizierung herangezogen werden.

2. PRINZIPIELLER AUFBAU EINES SPRECHERERKENNUNGSSYSTEMS

Da die Art und Weise, in der der Mensch andere Personen anhand ihrer Stimmen erkennt, bis heute nicht befriedigend geklärt ist, und außerdem der Mensch einem technischen System bei dieser Aufgabe nicht notwendig überlegen sein muß (vgl. Kap. 1.3), kann man das Problem zunächst als rein mathematisch-statistische Aufgabe erfassen. Dabei wird sich allerdings herausstellen, daß man die bei der Sprachproduktion und -perzeption auftretenden Gesetzmässigkeiten durchaus einbeziehen kann bzw. sogar einbeziehen muß, um zu einer realisierbaren Lösung zu gelangen.

Vom mathematischen Standpunkt aus gesehen kann die Sprachprobe eines Sprechers ω_i nach der Abtastung als Vektor

$$V = (s(1), s(2), \ldots s(K))^T$$

aufgefaßt werden, der insgesamt K Komponenten besitzt (K Abtast-

werte). Da verschiedene Äußerungen eines Sprechers im allgemeinen unterschiedliche, nicht vorhersagbare Signalformen aufweisen, können diese als Realisierungen eines stochastischen Vektors **V** betrachtet werden, dessen statistische Eigenschaften durch die (für jeden Sprecher ω_i unterschiedliche) multivariante Wahrscheinlichkeitsdichten $p(V|\omega_i)$ beschrieben werden.

Die mathematische Optimierungsaufgabe besteht nun darin, einer Realisierung V mit unbekannter Klassenzugehörigkeit eine Klasse derart zuzuordnen, daß im Mittel möglichst wenig Fehler auftreten. Denkt man an die bei Sprachsignalen übliche Abtastrate von 8 kHz und eine Sprachprobe von einigen Sekunden Dauer, dann stellt man sofort fest, daß eine Schätzung der Wahrscheinlichkeitsdichten $p(V|\omega_i)$ völlig unrealistisch ist, außerdem ist die Satzlänge K keine Konstante, was zu prinzipiellen mathematischen Schwierigkeiten führt.

Deshalb teilt man den gesamten Verarbeitungsprozeß in 2 Schritte, die Merkmalsgewinnung und die Klassifizierung, auf. Dieser klassische Aufbau eines Mustererkennungssystems ist in Fig. 2.1 dargestellt. In der ersten Stufe werden aus dem hochdimensionalen Vektor V bzw. dem ursprünglich analogen Signal $s(t)$ möglichst wenige Merkmale $X = (x_1, x_2, .., x_N)^T$ abgeleitet. Diese Merkmale x_i sollen - vereinfacht ausgedrückt - die Eigenschaft besitzen, daß sie für eine Klasse (einen Sprecher) möglichst reproduzierbare Werte annehmen, während sie für verschiedene Klassen (Sprecher) möglichst breit gestreute Werte aufweisen.

Im zweiten Schritt wird dann mit dem Merkmalsvektor X (mit konstanter, möglichst geringer Komponentenzahl N) die Klassenzugehörigkeit ω_i geschätzt.

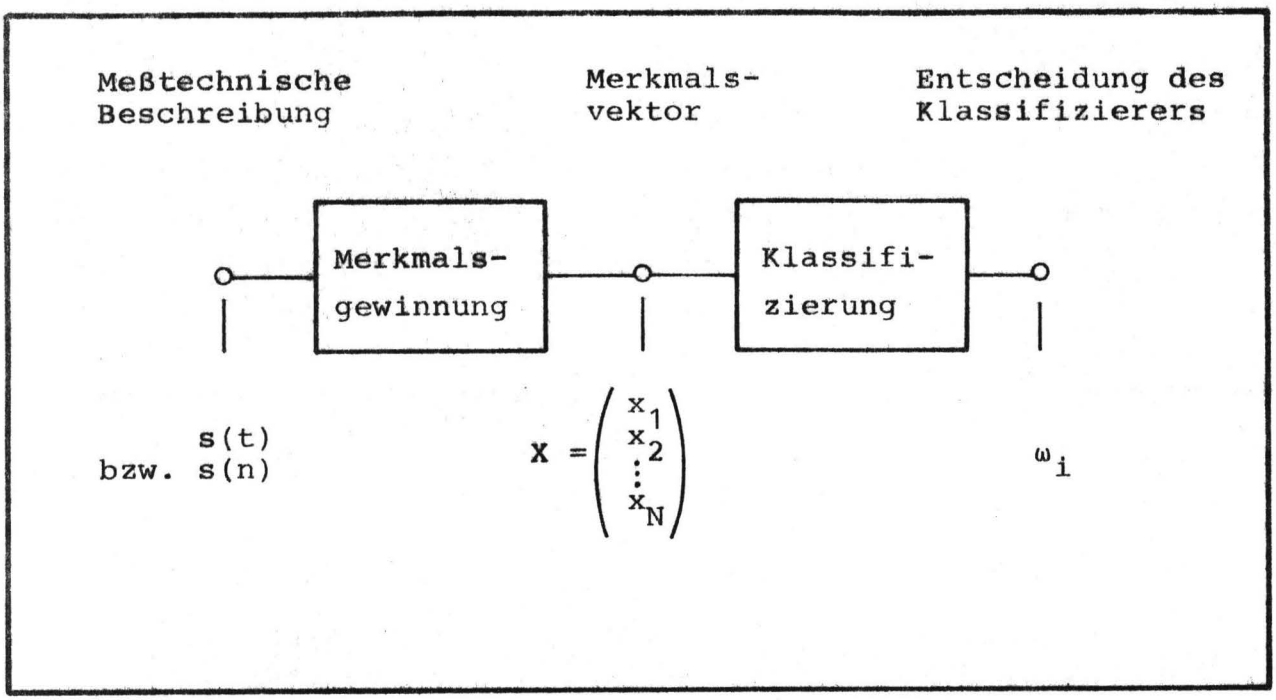


Fig. 2.1 Klassischer Aufbau eines Systems zur Mustererkennung

Die Merkmalsgewinnung, d.h. die Abbildung V→X kann dabei (außer in trivialen Fällen) kaum nach mathematischen Gesichtspunkten durchgeführt werden, sondern hängt stark und im allgemeinen völlig nichtlinear von dem physikalischen Prozeß ab, der den Vektor V erzeugt. Um hier zu befriedigenden Ergebnissen zu gelangen, müssen die Kenntnisse über die Eigenschaften des Sprachproduktionsprozesses ausgenutzt werden.

Aus der Vocodertechnik sind schon seit längerer Zeit Verfahren bekannt, die es gestatten, das Sprachsignal parametrisch in der Form $R(t) = (r_1(t), r_2(t), \ldots, r_M(t))^T$ zu beschreiben. Während bei der Vocodertechnik die mit dieser Beschreibung erzielbare Reduktion der für die Übertragung benötigte Bandbreite im Vordergrund steht, interessiert bei der Sprechererkennung (und auch bei der Spracherkennung) mehr die Tatsache, daß die für die Beschreibung verwendeten Parameter in direkter Beziehung zum Sprachproduktionsprozeß stehen. Deshalb können die für die Klassifizierung ver-

wendeten Merkmale aus ihnen viel einfacher gewonnen werden als aus dem ursprünglichen Sprachsignal s(t).

Unter Berücksichtigung dieser Überlegungen gelangt man zu einem 3-stufigen System für die Sprechererkennung (s. Fig. 2.2) (das übrigens in der gleichen Struktur für die Spracherkennung verwendet werden kann).

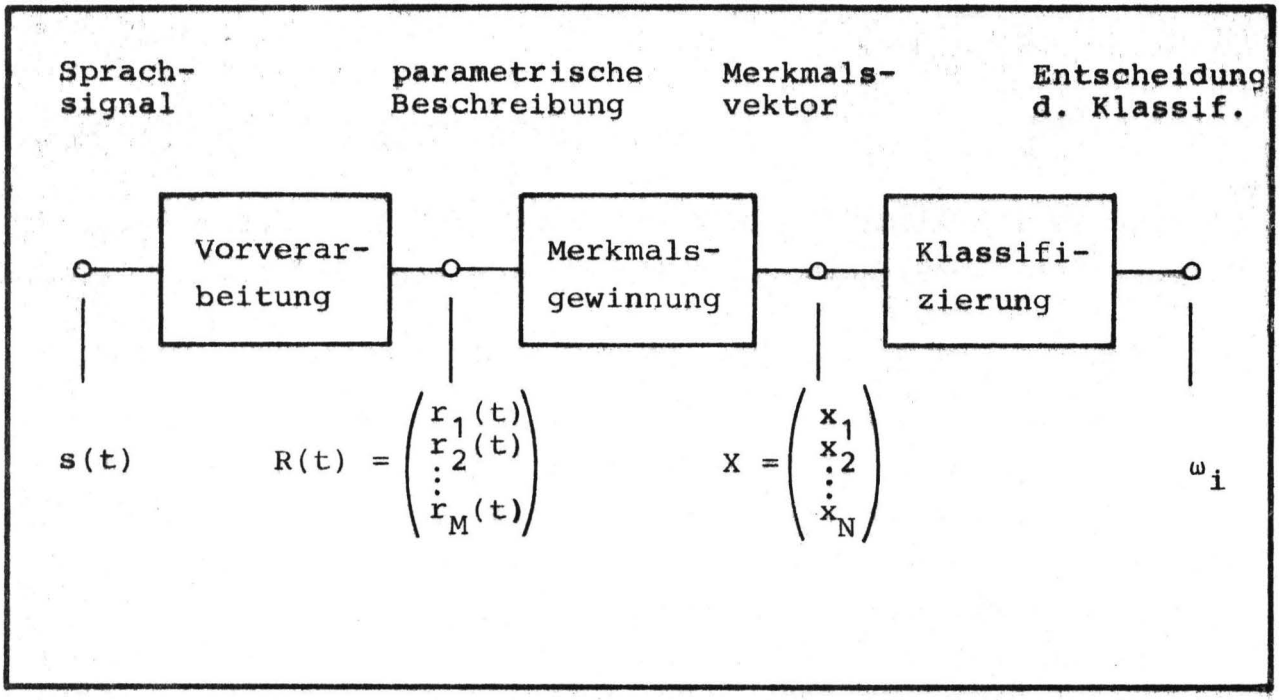


Fig. 2.2 Struktur eines Sprechererkennungssystems

In der Vorverarbeitungsstufe wird das Sprachsignal s(t) parametrisch als $R(t) = (r_1(t), r_2(t), \ldots, r_M(t)^T$ beschrieben, dadurch wird die Datenrate reduziert,und man erhält Parameter, die sich direkt auf Eigenschaften des Sprachproduktionsprozesses beziehen. Diese Verarbeitungsstufe wird in Kap. 2.1 näher erläutert. Aus den Parameterverläufen wird dann in der Merkmalsgewinnungsstufe ein Merkmalsvektor $X = (x_1, x_2, \ldots, x_N)^T$ gewonnen, der einem Klassifizierer angeboten wird. (Bei der Merkmalsgewinnung liegen

die wesentlichen Unterschiede zwischen Sprecher- und Spracherkennung). Die hier verwendeten Methoden werden in Kap. 2.2 dargestellt.

Der letzte Verarbeitungsschritt, die Schätzung der Klassenzugehörigkeit mit Hilfe des Merkmalsvektors X, erfolgt im Gegensatz zu den bisher dargestellten, mehr empirischen Verfahren nach weitgehend bekannten Algorithmen der Mustererkennung und wird in Kap. 2.3 ausführlicher erläutert.

2.1 Vorverarbeitung von Sprachsignalen

Um die Vorteile einer parametrischen Beschreibung des Sprachsignals zu erläutern, soll kurz der Sprachproduktionsprozeß erläutert werden.

Sprachsignale werden erzeugt durch Schwingungen der Stimmbänder (stimmhafte Laute) oder durch Turbulenzen infolge von Querschnittsverengungen des Vokaltraktes (stimmlose Laute). Der Vokaltrakt (Rachen-, Mund- und Nasenraum) wirkt als einstellbares akustisches Filter, das einzelne Frequenzbereiche des Quellspektrums mehr oder weniger unterdrückt. Durch Variation der Filterparameter werden die verschiedenen Sprachlaute gebildet. Die Tonhöhe wird dabei durch die Grundfrequenz der Stimmbänder f_o festgelegt. Ausführlichere Beschreibungen sind z.B. zu finden bei Fant /6/ und Flanagan /7/.

Ein vereinfachtes Modell der Sprachproduktion ist in Fig. 2.3 dargestellt.

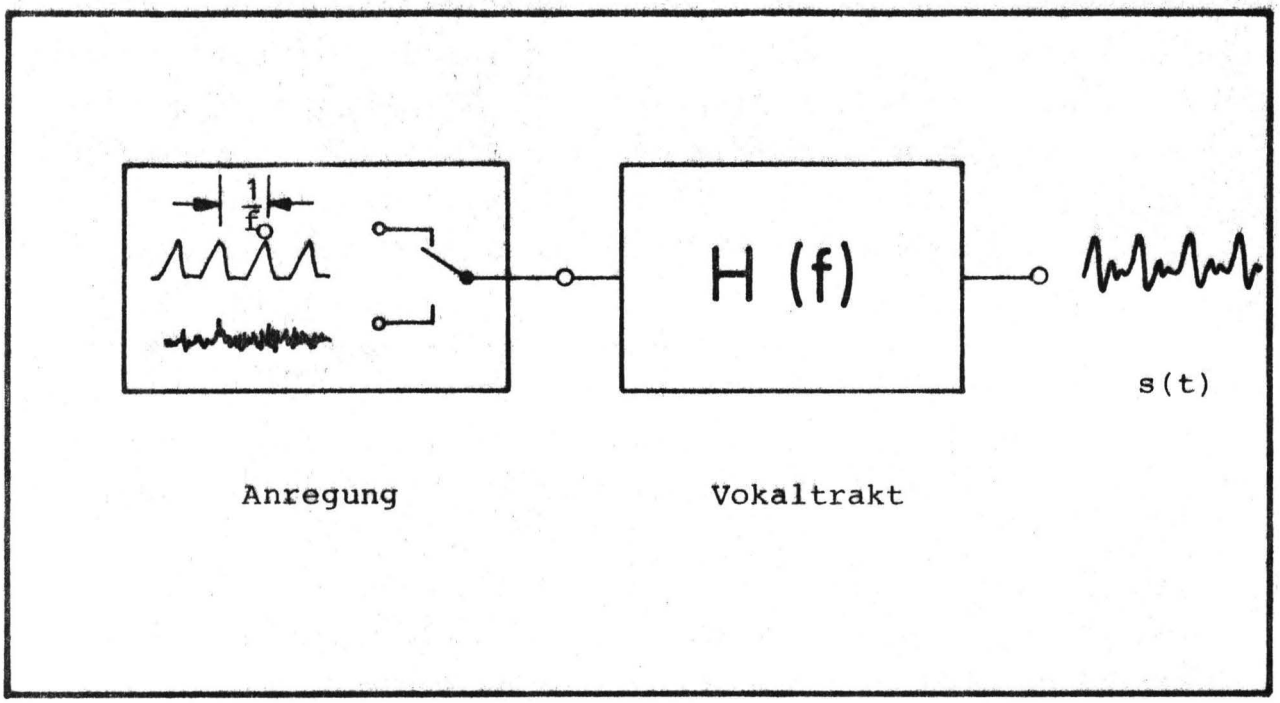


Fig. 2.3 Vereinfachtes Modell der Sprachproduktion

Darin ist der Wechsel von stimmhafter zu stimmloser Anregung durch einen Schalter berücksichtigt, das Vokaltraktfilter wird durch seine Übertragungsfunktion H(f) beschrieben. Dieses Modell läßt sich parametrisch beschreiben durch die stimmhaft/stimmlos-Entscheidung, die Sprachgrundfrequenz f_o (nur bei stimmhaften Lauten) sowie einen Parametersatz zur Einstellung des Filters H(f). Die Einstellung der Lautstärke kann sowohl dem Filter zugeordnet werden als auch der Anregung.

Diese Beschreibung des Sprachsignals durch Parameter ist so vollständig, daß mit ihnen ein verständliches Sprachsignal erzeugt werden kann. Dies wird durch den Einsatz von Vocoderverfahren bewiesen, bei denen lediglich die Sprachparameter vom Sender zum Empfänger übertragen werden.

Der Vorteil der parametrischen Beschreibung liegt nun darin, daß die Parameter in direkter Beziehung zu anatomischen Größen der Sprachproduktion stehen. Da sowohl die Art Anregung als auch die geometrische Form des Vokaltraktes sich nur relativ langsam ändern, kann man die Sprachparameter mit einer sehr viel kleineren Abtastfrequenz abtasten, und zwar mit ca. 50-100 Hz (gegenüber mindestens 8000 Hz beim ursprünglichen Sprachsignal).

Verschiedene Vocoderverfahren unterscheiden sich im wesentlichen durch die parametrische Beschreibung des Vokaltraktes H(f), es gibt sowohl analoge als auch digitale Vocoderverfahren. An dieser Stelle sollen nur 2 Beispiele erörtert werden.

Beim Kanalvocoder wird das Spektrum des Sprachsignals (und damit auch die Übertragungsfunktion H(f) des Filters) durch eine Filterbank in ca. 20 einzelne Kanäle zerlegt und deren Intensität übertragen.

Eine ausführliche Darstellung der Spektralanalyse in dieser Anwendung wird von Talmi /7/ gegeben.

Beim Prädiktionsvocoder, der in den letzten Jahren stark an Bedeutung gewonnen hat, wird der Vokaltrakt durch ein digitales Filter (Prädiktor) nachgebildet, übertragen werden hier die Filterkoeffizienten. Es gibt darüberhinaus eine Reihe von weiteren Beschreibungsmöglichkeiten, die sich direkt aus diesen Filterkoeffizienten ableiten lassen (Parcorkoeffizienten, Cepstrum, Areafunction usw.). Ein Vergleich dieser Parametersätze wurde für die Sprechererkennung von Höfker /8/ durchgeführt. Dabei zeigte sich, daß die spektrale Beschreibung des Sprachsignals durch eine Filterbank recht gut für die Weiterverarbeitung geeignet ist. Da sie im Gegensatz zur Prädiktionsanalyse mit begrenztem Aufwand in Echtzeit durchgeführt werden kann, wurde diese Art der parametrischen Beschreibung in dem vom Heinrich-Hertz-Institut entwickelten SPREE-System zur Sprecherverifizierung (s. Kap. 6.) verwendet.

Zusammenfassend läß sich sagen, daß sich das Sprachsignal s(t) parametrisch durch einen Vektor R(t) mit ca. M = 20 Komponenten beschreiben läßt. Wird dieser Vektor z.B. alle 20 ms abgetastet, dann wird einmal die Datenrate gesenkt (1000 Werte pro Sekunde gegenüber 8000 Werten beim Originalsignal s(t)), außerdem stehen die Parameter in direktem Zusammenhang zum Sprachproduktionsprozeß, so daß sprecherspezifische Merkmale aus ihnen sehr viel leichter gewonnen werden können als aus dem ursprünglichen Sprachsignal.

Nach der Abtastung wird der Parametervektor mit R(l) bezeichnet, eine sprachliche Äußerung der Länge L liegt dann in der Form vor:

$$\{R(1),R(2),\ldots,R(L)\} = \begin{pmatrix} r_1(1), & r_1(2), \ldots, & r_1(L) \\ r_2(1), & r_2(2), \ldots, & r_2(L) \\ \vdots & \vdots & \vdots \\ r_M(1), & r_M(2), \ldots, & r_M(L) \end{pmatrix}$$

2.2 Merkmalsgewinnung

Im Parametervektor R(l) sind wie im ursprünglichen Sprachsignal die Informationen

wer spricht
was wird gesprochen
wie wird gesprochen

enthalten, allerdings in einer komprimierten, dem Sprachproduktionsprozeß besonders angepaßten Form.

Die Methoden zur Gewinnung sprecherspezifischer Merkmale zielen deshalb im wesentlichen darauf ab, den störenden Einfluß des Sprachinhalts ("Was wird gesprochen") zu reduzieren und die Randbedingungen, unter denen gesprochen wird, möglichst konstant zu halten, damit auch die Sprechweise ("Wie wird gesprochen") möglichst frei von Veränderungen ist.

Eine Möglichkeit der Merkmalsgewinnung ist die Mittelung der Sprachparameter über verschiedene Sprachinhalte, die zur sogenannten statistischen Analyse führt, eine zweite die Beschränkung auf vorgegebene, definierte Sprachinhalte, die entweder zur Segmentanalyse oder zur Konturanalyse führt. Beide Methoden können sowohl textabhängig als auch textunabhängig durchgeführt werden. Die textunabhängige Erkennung ist besonders einfach mit der statistischen Analyse möglich. Von dem gesprochenen Text muß hier lediglich eine ausreichende Länge gefordert werden, damit die ermittelten Merkmale unabhängig vom Text und damit sprechertypisch sind. Sollen bei der textunabhängigen Erkennung nur bestimmte Sprachinhalte analysiert werden, dann muß dem eigentlichen Sprechererkennungssystem ein Spracherkennungssystem vorgeschaltet werden. Das Spracherkennungssystem markiert die gesuchten Sprachinhalte, anschließend kann dann eine Segmentanalyse zur Sprechererkennung durchgeführt werden. Dabei muß sichergestellt sein, daß die gesuchten Sprachinhalte in dem zu analysierenden Text enthalten sind.

Durch die Vereinbarung eines festen Textes wird die Erkennung von Sprechern wesentlich vereinfacht. Die statistische Analyse kann ebenso angewandt werden wie bei unbekanntem Text, jedoch darf die zu analysierende Sprachprobe dabei wesentlich kürzer sein. Bei der Segmentanalyse wird das Auffinden bestimmter Segmente vereinfacht, da man die Reihenfolge der artikulierten Sprachlaute kennt und man die Algorithmen zum Markieren bestimmter Segmente an den vorgegebenen Codesatz anpassen kann. Schließlich kann man bei fest vereinbartem Text die zeitliche Struktur des Codesatzes ausnutzen und direkt die Zeitverläufe eines oder mehrerer Parameter für die Sprechererkennung verwenden, wie es in der Konturanalyse geschieht. Die drei Methoden zur Merkmalsgewinnung, statistische Analyse, Segmentanalyse und Konturanalyse wurden von Höfker /8/ und Jesorsky /1/ ausführlicher behandelt, außerdem wird in Kap. 6 noch näher darauf eingegangen werden, so daß an dieser Stelle auf eine ausführliche Diskussion verzichtet werden kann. Bei der Konturanalyse ist im allgemeinen eine nichtlineare zeitliche Anpassung

der Parameterverläufe (zum Ausgleich unterschiedlicher Sprechgewohnheiten) notwendig, einen Überblick über die dafür verwendeten Methoden findet man bei Kriener /9/.

Allen Verfahren der Merkmalsgewinnung ist gemeinsam, daß sie aus der parametrischen Beschreibung einen niedrigdimensionalen Merkmalsvektor X erzeugen,

$$R(1), R(2), \ldots, R(L) \longrightarrow X$$

der für die Klassifizierung verwendet wird. Natürlich können diese Verfahren auch kombiniert werden, um die Leistungsfähigkeit eines Systems zu verbessern (s. Kap. 6.).

2.3 Klassifizierung

Der Klassifizierer bildet den letzten Baustein des Sprechererkennungssystems. Ihm wird ein Merkmalsvektor angeboten, der durch eines der im letzten Kapitel vorgestellten Verfahren gewonnen wurde. So wie der Mensch die Stimme einer anderen Person erst kennenlernen muß, ehe er sie wiedererkennen kann, so muß der Klassifizierer erst in einer Lernphase die statistischen Eigenschaften der zu klassifizierenden Muster X auswerten, bevor er in der Testphase ihm unbekannte Muster klassifizieren kann.

Es soll hier nicht versucht werden, einen allgemeinen Überblick über Klassifizierungsverfahren zu geben; hier kann auf eine breite Literatur verwiesen werden, z.B. /10, 11, 12/. Stattdessen soll die prinzipielle Wirkungsweise anhand eines sehr einfach strukturierten "Abstandsklassifizierers" erläutert werden, der sich, wie schon aus dem Namen hervorgeht, geometrisch interpretieren läßt. Diese Interpretation soll für den Fall erfolgen, daß der Merkmalsvektor nur N = 2 Komponenten $X = (x_1, x_2)^T$ hat, und außerdem nur K = 2 Sprecher unterschieden werden sollen. Die mathematische Beschreibung ist dagegen allgemein, gilt also für

beliebig viele Sprecher K. Zuerst wird die Sprecheridentifizierung betrachtet, danach die -verifizierung. In der Lernphase wird dem Klassifizierer eine Lernstichprobe mit bekannter Klassenzugehörigkeit zur Verfügung gestellt, die Muster des Sprechers ω_k seien durch $X_1^{(k)}, X_2^{(k)}, \ldots, X_J^{(k)}$ gekennzeichnet. Jedes Muster der Lernstichprobe kann durch einen Punkt in der x_1, x_2 - Ebene dargestellt werden, ein Beispiel ist in Fig. 2.4 angegeben.

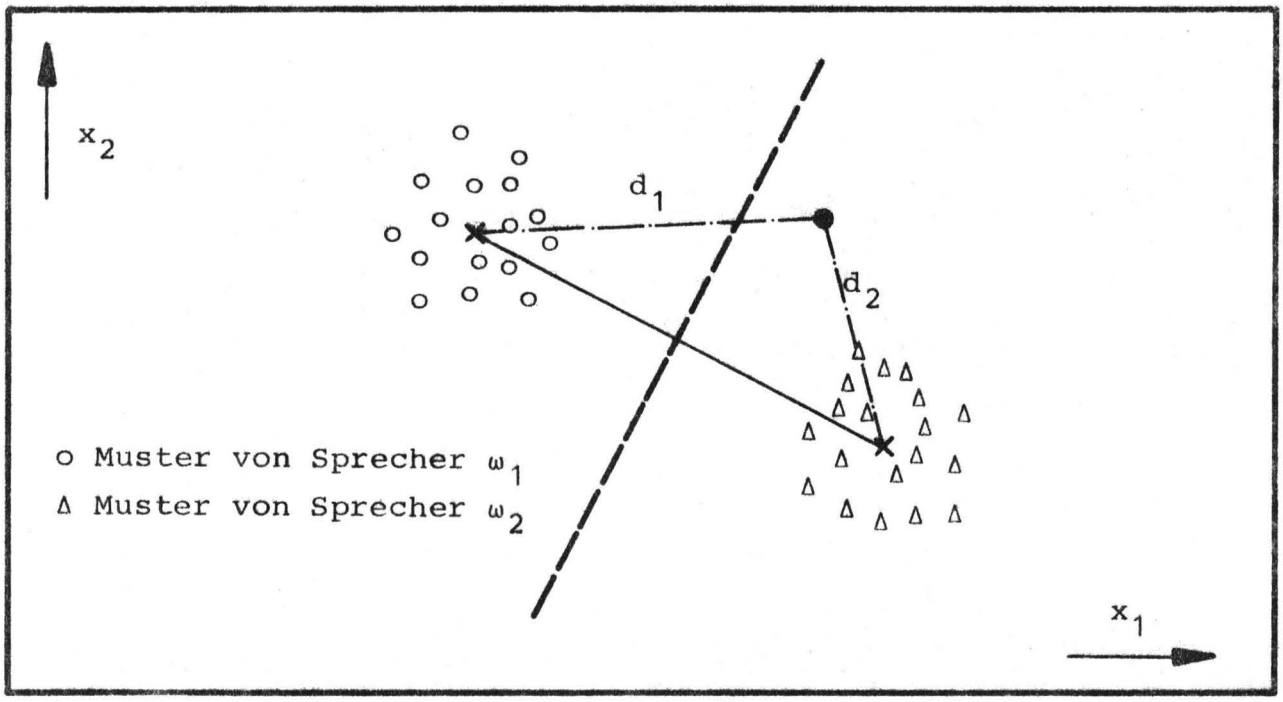


Fig. 2.4 Prinzip des Abstandsklassifizierers bei der Sprecheridentifizierung

In dieser Darstellung wurde angenommen, daß die gewonnenen Merkmale x_1 und x_2 zur Sprechererkennung geeignet sind und sich deshalb die Bereiche, in denen sich die Muster der beiden Sprecher befinden, nicht überlappen.

In der Lernphase des Abstandsklassifizierers wird aus den J Lernmustern jedes Sprechers ω_k ein typisches Referenzmuster $M^{(k)}$, z.B. als Mittelwert, berechnet:

$$M^{(k)} = \frac{1}{J} \sum_{i=1}^{J} x_i^{(k)}$$

Diese Referenzmuster sind in Fig. 2.4 durch Kreuze gekennzeichnet.

In der Testphase wird ein unbekanntes Muster X dann der Klasse zugeordnet, zu deren Referenzmuster es den kleinsten Abstand

$$d_k^2 (X, M^{(k)}) = (X - M^{(k)})^T (X - M^{(k)})$$

hat, es ergibt sich also folgende Klassifizierungsvorschrift:

$$\min_{k=1}^{K} \{ d_k^2 (X, M^{(k)}) \} = d_i^2 (X, M^{(i)}) \curvearrowright X \rightarrow \omega_i$$

Durch diese Vorschrift ergibt sich die Klassengrenze implizit als Mittelsenkrechte auf der Verbindungslinie der Referenzvektoren, wie in Fig. 2.4 gestrichelt eingezeichnet.

Es ist leicht zu erkennen, daß der Abstandsklassifizierer nur dann optimal arbeitet, wenn die Muster einer Klasse konzentrisch um ihren Referenzvektor versammelt sind. Diese Voraussetzung ist im allgemeinen nicht gegeben. Um die Leistungsfähigkeit des Abstandsklassifizierers in solchen Fällen zu verbessern, können statt des euklidischen Abstands andere verfeinerte Abstandsmaße, z.B. der Mahalanobis-Abstand (s./1/) verwendet werden. Auf diesen Themenkreis soll hier jedoch nicht näher eingegangen werden.

Bei der bisher dargestellten Sprecheridentifizierung geht es darum, die Muster der Stichprobe genau einem von K Sprechern zuzuordnen. Die Entscheidung ist entweder richtig oder falsch, und die Fehlerrate ε als Verhältnis der Falschklassifizierungen zur Gesamtanzahl der Muster gibt Aufschluß über die Leistungsfähigkeit des Systems.

Bei der Sprecherverifizierung treten dagegen 2 Fehlertypen auf wie schon in Kap. 1.2 festgestellt. Der wahre Sprecher kann fälschlich zurückgewiesen werden (ε_{FR}), und ein anderer Sprecher kann das System erfolgreich täuschen und als wahrer Sprecher akzeptiert werden (ε_{FA}). Deshalb ist die Fragestellung bei der Klassifizierung eine andere. Geht es bei der Identifizierung um die Ermittlung einer "größten" Ähnlichkeit bzw. eines kleinsten Abstands, dann ist das Kernproblem der Verifizierung zu ermitteln, ob die Ähnlichkeit "genügend groß" bzw. der Abstand "genügend klein" ist.

Die Ähnlichkeit zwischen dem Referenzmuster des "wahren Sprechers" ω_k und einem unbekannten Muster X, das als zum Sprecher ω_k zugehörig verifiziert werden soll, wird wie bisher durch den euklidischen Abstand $d_k = d(X, M^{(k)})$ beschrieben, um die Klassifizierung geometrisch interpretieren zu können. Akzeptiert man das Muster nur, wenn der Abstand kleiner als eine vorgegebene Schwelle d_s ist, dann lautet die Klassifizierungsvorschrift für die Verifizierung

$$d(X, M^{(k)}) \gtrless d_s \curvearrowright \begin{cases} \text{X zurückgewiesen} \\ \\ \text{X akzeptiert} \end{cases}$$

Das durch die Schwelle d_s definierte Toleranzgebiet ist eine Hyperkugel mit dem Referenzvektor $M^{(k)}$ im Zentrum und dem Radius d_s. Für den 2-dimensionalen Fall ist ein Toleranzgebiet für den Sprecher ω_2 in Fig. 2.5 dargestellt.

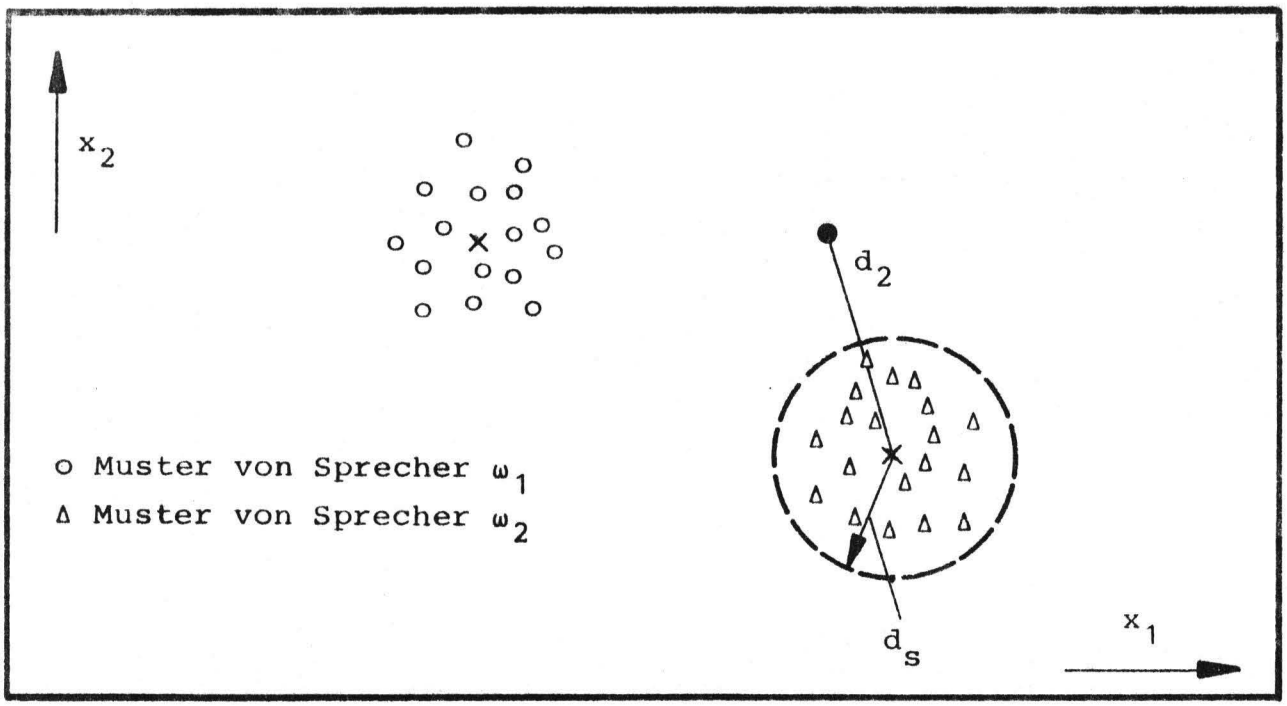


Fig. 2.5 Prinzip des Abstandsklassifizierers bei der Sprecherverifizierung

Durch die Größe der Schwelle d_s wird das Verhältnis der beiden Fehlertypen ε_{FR} und ε_{FA} bestimmt. Ist die Schwelle d_s zu klein, dann ist das System zwar sehr sicher gegenüber Täuschungsversuchen, aber der wahre Sprecher wird selbst häufig zurückgewiesen. Bei zu groß gewähltem d_s wird der wahre Sprecher immer akzeptiert, aber gleichzeitig werden erfolgreiche Täuschungsversuche erleichtert. Diese prinzipielle Abhängigkeit ist für ein Klassifizierungsexperiment, auf das hier nicht näher eingegangen werden soll, in Fig. 2.6 dargestellt.

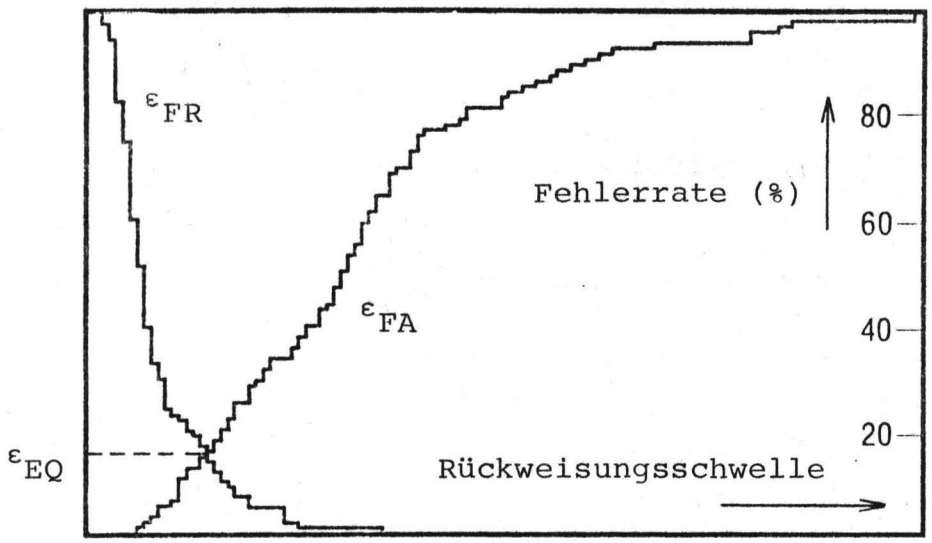


Fig. 2.6 Beispiel für die Abhängigkeit der Fehlerraten ε_{FR} und ε_{FA} von der Rückweisungsschwelle d_s

Um die Leistungsfähigkeit der Verifizierung wie die der Identifizierung durch eine einzige Zahl beschreiben zu können, bezieht man sich häufig auf den Schnittpunkt beider Fehlerraten, der als "equal error rate" ε_{EQ} bezeichnet wird:

$$\varepsilon_{EQ} = \varepsilon_{FA}\Big|_{\varepsilon_{FA}=\varepsilon_{FR}}$$

Die in Fig. 2.6 benutzte Darstellung von Fehlerraten wird in Zukunft sehr häufig zur Charakterisierung der Leistungsfähigkeit von Verifizierungssystemen verwendet werden.

3. SPRECHERIDENTIFIZIERUNG

Unter dem Begriff Sprecheridentifizierung versteht man, wie in den Kapiteln 1.2 und 2.3 erläutert, die Zuordnung einer Sprachprobe (bzw. des daraus gewonnenen Merkmalsvektors) zu einem von K Sprechern, ohne daß weitere Information über den mutmaßlichen Sprecher (Identitätsziel) vorliegt.

Da die Hauptanwendung für dieses Verfahren die Kriminalistik ist, sollen nur die Randbedingungen und Probleme für diesen Fall diskutiert werden.

Gehen wir davon aus, daß im Zusammenhang mit einer strafbaren Handlung eine Sprachprobe des Täters aufgezeichnet werden kann. Dabei wird es sich in der Regel um die Aufzeichnung eines Telefongesprächs (Bombendrohung, Erpressung) handeln, auch die akustische Aufzeichnung eines Banküberfalls kommt dafür infrage. Aus dieser Sprachprobe kann nun ein Merkmalsvektor gewonnen werden, der nach Kap. 2.3 mit den Referenzvektoren der K infrage kommenden Sprecher vergleichen werden muß.

An dieser Stelle sind jetzt 2 Anwendungen möglich. Wenn sich die Technik der automatischen Sprechererkennung bei den Sprachverfolgungsbehörden durchgesetzt hat, dann werden vermutlich bei der erkennungsdientlichen Behandlung eines Straftäters nicht nur Fotos und Fingerabdrücke, sondern auch Sprachproben gesammelt.

In diesem Fall verläuft die Sprechererkennung ähnlich wie die Erkennung anhand von Fingerabdrücken, das Testmuster (Tatgespräch) wird mit den abgespeicherten Sprachproben aller karteimäßig erfaßten Personen verglichen. Wegen der im Vergleich zum Fingerabdruck hohen Fehlerrate des Stimmabdrucks (nach Meinung des Autors unter günstigen Randbedingungen zwischen 1‰ und 10%) wird das automatische Erkennungssystem aber vermutlich keine endgültige Entscheidung treffen, sondern als Ergebnis eine Liste der mit größerer Wahrscheinlichkeit infrage kommenden Personen aufstellen (Scree-

ning-Verfahren).

Der Umfang dieser Liste muß dann mit anderen Methoden (z.B. Überprüfung des Alibis)reduziert werden.

Ist eine umfassende "Stimmdatenkartei" noch nicht aufgebaut, dann kann ein Sprechererkennungssystem nur sekundär eingesetzt werden. Erst nachdem mit üblichen Methoden der Strafverfolgung ein kleiner Kreis von Tatverdächtigen ermittelt worden ist, kann unter diesem eine Entscheidung getroffen werden. Dazu ist es dann notwendig, daß von diesen Personen Sprachproben aufgenommen werden und Referenzmerkmalsvektoren gebildet werden. Die Lernphase des Klassifizierers liegt also zeitlich nach dem Tatgespräch, natürlich erfolgt die Klassifizierung des Tatgesprächs aber erst nach der durchgeführten Lernphase.

Da in den meisten Fällen nicht sichergestellt werden kann, daß der wahre Täter im Kreise der Tatverdächtigten enthalten ist, muß man zusätzlich zur Entscheidung Sprecher ω_k (k=1,..,K) auch die Entscheidung "keiner der K Sprecher" zulassen. Dies kann durch Kombination von Elementen der Sprecheridentifizierung und -verifizierung berücksichtigt werden. Ist z.B. Sprecher ω_k am meisten verdächtig, da der Merkmalsvektor X des Tatgesprächs seinem Referenzvektor $M^{(k)}$ am ähnlichsten ist, also den kleinsten Abstand $d_k = d\,(X, M^{(k)})$ aufweist, dann kann dieser Sprecher dadurch entlastet werden, daß der Abstand größer ist als eine Schwelle d_s, deren Bedeutung in Kap. 2.3 in Zusammenhang mit der Verifizierung erläutert wurde.

Schließlich tritt als Sonderfall die "reine" Sprecherverifizierung auf, wenn lediglich ein Tatverdächtigter ermittelt worden ist. Gegenüber der üblichen Anwendung der Sprecherverifizierung, bei der der Sprecher darum bemüht ist, erkannt zu werden (s. ausführlich in Kap. 4.), liegt hier jedoch gerade die entgegengesetzte Zielrichtung vor (es sei denn, eine Übereinstimmung der Sprachproben führte zur Entlastung des Verdächtigten).

Damit sind wir auch schon von den möglichen Aufgabenstellungen zu den Problemen der kriminalistischen Anwendung übergegangen. Generell liegt keine Kooperationsbereitschaft der infrage kommenden Sprecher vor! Schon heutzutage wird die Stimme beim Tatgespräch häufig verstellt (monotone Aussprache oder Taschentuch vor dem Mund); diese Verhaltensweise wird bei einem breiten Einsatz von zuverlässigen Sprechererkennungssystemen sicherlich verstärkt auftreten.

Ist trotz dieser Schwierigkeiten ein Kreis von Tatverdächtigten ermittelt, dann werden sich die unschuldig Betroffenen bei der Abgabe ihrer Sprachprobe zwar neutral verhalten, dies kann man aber vom wahren Täter auf keinen Fall erwarten. Deshalb ist es fraglich, ob das System für Lern- und Teststichprobe den gleichen gesprochenen Text zur Verfügung gestellt bekommt. Weigert sich ein Betroffener zum Beispiel, den Text des Tatgesprächs zu wiederholen, dann können nur textunabhängige Verfahren der Merkmalsgewinnung verwendet werden, deren Anwendung bei sehr kurzen Sprachproben kritisch sein kann /1, 8/.

Neben diesen Schwierigkeiten, die am zu erkennenden Spracher selber liegen, treten bei der kriminalistischen Anwendung eine Reihe von äußeren Randbedingungen erschwerend hinzu. In vielen Fällen sind starke Hintergrundgeräusche vorhanden (z.B. Straßenlärm bei Anrufen von Telefonzellen aus). Dazu kommen Verzerrungen durch das Übertragungsmedium (Klirrfaktor der Mikrofonkapsel, lineare und nichtlineare Verzerrungen der Leitung, Wählergeräusche usw.), die zum Teil nicht reproduzierbar sind,und schließlich sind für die Sprachaufzeichnung verwendeten Geräte nicht immer von ausreichender Qualität oder fachgerecht angeschlossen (z.B. Netzbrummen).

All diese Schwierigkeiten tragen dazu bei, daß es bis heute nicht gelungen ist, ein leistungsfähiges Sprecheridentifizierungssystem zu entwickeln. Es ist bisher im Auftrag des U.S. Justizministeriums lediglich ein halbautomatisches Erkennungssystem SASIS (Semi-Automatic Speaker Identification System) entwickelt worden (1973-1976). Dieses System verwendet zur Merkmalsgewinnung die Segment-

analyse (vgl. Kap. 2.2), wobei die Segmentgrenzen manuell festgelegt werden.

Mit diesem System wurden zwar in der Laboratmosphäre (hochqualitative Sprachproben) Erkennungsraten von 97% erzielt /13/, in einem von der Polizei von Los Angeles durchgeführten Feldtest hat sich das System aber nicht bewährt. Dieses Versagen wurde von den Entwicklern darauf zurückgeführt, daß zur Erkennung nur stationäre Sprachlaute herangezogen wurden, die empfindlich gegenüber den Verzerrungen des Telefonkanals sind. Deshalb schlugen sie (leider erst nach Abschluß des Projektes) vor, den Merkmalsvektor vorwiegend aus dynamischen Eigenschaften des Sprachsignals zu gewinnen. Solche Merkmale werden z.B. in unserem SPREE-System verwendet (s. Kap. 6.).

4. SPRECHERVERIFIZIERUNG

Wie in den letzten Kapiteln schon erläutert, dient die Sprecherverifizierung zur Überprüfung der vorgegebenen Identität einer Person. Da weitere, von der Person beabsichtigte Aktionen (z.B. Betreten eines Sicherheitsbereichs) vom positiven Ausgang dieser Überprüfung abhängig sind, kann man dabei generell von der Kooperationsbereitschaft der zu überprüfenden Person ausgehen.

Ähnlich wie man bei einer Unterschrift einen eingeübten, möglichst reproduzierbaren Kurvenzug erwartet, kann man bei der "akustischen Unterschrift" vom Sprecher fordern, daß er für die Verifizierung z.B. einen ganz bestimmten Satz auf möglichst reproduzierbare Art und Weise spricht. Dies ermöglicht den Einsatz von sehr wirkungsvollen Verfahren der Merkmalsgewinnung (s. /1/, /2/ und Kap. 6.). Darüberhinaus benötigt das Erkennungssystem das Identitätsziel des Sprechers, etwa durch alphanumerische Eingabe des Namens, einer persönlichen Kennzahl oder durch maschinelles Lesen einer auf seiner Scheckkarte aufgebrachten Magnetspur.

Um die Leistungsfähigkeit eines Verifizierungssystems zu ermitteln, muß man 2 Typen von Erkennungsexperimenten durchführen. Bei den regulären Verifizierungen wird der Prozentsatz ε_{FR} der wahren Sprecher ermittelt, der vom System zurückgewiesen wird, und bei den Täuschungsversuchen wird die Erfolgsquote mit ε_{FA} angegeben. An dieser Stelle soll ausdrücklich betont werden, daß die Fehlerraten ε_{FR} und ε_{FA} sich auf die jeweilige Gesamtheit der Versuche wahrer Sprecher bzw. der Täuschungsversuche beziehen. Die Angabe einer Fehlerrate etwa in der Form "Zahl der Falschklassifizierungen zur Gesamtzahl aller Klassifizierungen" ist wenig hilfreich, wenn die a priori Wahrscheinlichkeit für das Auftreten von Täuschungsversuchen nicht bekannt ist.

Liegen zum Beispiel bei einem System die Fehlerraten ε_{FR} = 1% und ε_{FA} = 10% vor, dann würde sich bei Auftreten von 50% Täuschungsversuchen eine Gesamtfehlerrate von 5,5% ergeben, beim Auftreten von 1‰ Täuschungsversuchen (realistisch bei Strafandrohung!) aber eine Fehlerrate von nur 1,009%.

Allgemeingültige Aussagen, wie man das Verhältnis der Fehlertypen ε_{FR} und ε_{FA} (durch Variation der Rückweisungsschwelle d_s, vgl. Kap. 2.3) am günstigsten einstellt, lassen sich nicht treffen. Je nach der speziellen Anwendung muß man abschätzen, welcher Schaden durch ε_{FR} (System wird nach mehrmaliger Rückweisung vom Benutzer abgelehnt) und ε_{FA} (unbefugte Ausnutzung von Informationen, Sicherheitsrisiko) entsteht und wie hoch die a priori Wahrscheinlichkeit für Täuschungsversuche anzusetzen ist.

Besondere Bedeutung muß man den Randbedingungen beimessen, unter denen die Erfolgsquote von Täuschungsversuchen gemessen wird (vgl. Kap. 1.3). Man kann hier 3 Stufen unterschiedlicher Befähigung des Imitators unterscheiden. In der Stufe I werden zwar die vom System gestellten Forderungen (z.B. Sprechen eines bestimmten Codesatzes) erfüllt, der Sprecher gibt sich aber keine Mühe, eine bestimmte Person nachzuahmen, sondern spricht in seiner gewohnten Weise.

In der Stufe II stellt sich der Imitator völlig auf die nachzuahmende Stimme ein. Neben der individuell unterschiedlichen Fähigkeit des Imitators ist hier von Bedeutung, in welcher Form die Rückkopplung, d.h. die Interrichtung des Imitators über den Erfolg seines Täuschungsversuches erfolgt. Mögliche Formen sind z.B. keine Unterrichtung, Angabe des Ergebnisses Akzeptanz/Rückweisung, und reichen bis hin zur Angabe der Ähnlichkeit ganz bestimmter Merkmale, die den Imitator in die Lage versetzen, seine Sprechweise ganz gezielt zu variieren.

In der Stufe III setzt der Imitator schließlich nicht nur seine Stimme, sondern zusätzlich technische Hilfsmittel ein, um das System zu überlisten, z.B. durch Verwendung einer Tonbandaufnahme des wahren Sprechers oder durch Modifikation seiner Stimme durch ein einstellbares Filter. Für diese Art der Täuschungsversuche liegen bisher kaum Erfahrungen vor, eine im Auftrag des U.S. Verteidigungsministeriums durchgeführte Studie wurde leider als geheim eingestuft.

Die Gefahr des relativ einfach realisierbaren Täuschungsversuchs mit einer Tonbandaufnahme des wahren Sprechers kann man aber dadurch verringern, daß man den zu sprechenden Codesatz erst unmittelbar vor dem Verifizierungsversuch bekannt gibt (z.B. als zufällige Ziffernfolge).

Die meisten in der Literatur angegebenen Fehlerraten für ε_{FA} beziehen sich auf die oben erläuterte Stufe I, auch der vom US-Verteidigungsministerium für Verifizierungssysteme geforderte Standard (ε_{FR} = 1%, ε_{FA} = 2%) bezieht sich auf diese Stufe. Zwar ist der experimentelle Aufwand dafür - verglichen mit Stufe II und III recht gering, da die Täuschungsversuche im allgemeinen nicht explizit durchgeführt zu werden brauchen (Verwendung von Sprachproben anderer wahrer Sprecher), diese Experimente haben dafür aber auch nur begrenzte Aussagekraft, da später im praktischen Einsatz auftretende Täuschungsversuche sicherlich beabsichtigt sind (die anderen stellen kaum ein Sicherheitsrisiko dar) und zu

wesentlich höheren Fehlerraten führen (vergl. Kap. 1.3, Tab. 1.1).

Ein System zur Sprecherverifizierung kann in vielen Fällen dort eingesetzt werden, wo die Legitimation einer Person bisher durch Unterschriftsvergleich oder Ausweiskontrolle stattgefunden hat. Ein besonderer Vorteil dabei ist es, daß die persönliche Anwesenheit der zu überprüfenden Person nicht erforderlich ist.

Davon ausgehend kann man mögliche Erkennungssysteme in closed-line und open-line Systeme einteilen, bei denen unterschiedliche äußere Randbedingungen vorliegen.

Bei closed-line Systemen kann die Sprachrpobe ohne Qualitätseinbusse direkt dem Erkennungssystem übergeben werden, das meist an dem Ort aufgebaut ist, an dem die Verifizierung durchgeführt werden soll. Typische Anwendungen dafür sind Zugangskontrollen zu Sicherheitsbereichen (Rechenzentren, Atomkraftwerke oder militärische Sperrbezirke) /14/, automatische Geldausgabegeräte bei Banken (z.B. für Bagatellbeträge) /15/ oder (mit Einschränkungen) die Zugriffsberechtigung zu vertraulichen Informationen (Datenbanksysteme. Es müssen dabei keine einschränkenden Normen bestimmter Übertragungsmedien oder Endgeräte eingehalten werden, man kann also hochqualitative, breitbandige Mikrophone verwenden und einen breitbandigen störungsfreien Anschluß an das Erkennungssystem sicherstellen. Darüberhinaus kann man die Aufnahmebedingungen weitgehend beeinflussen, z.B. durch akustische Abschirmung des Sprechers (Reduktion des Hintergrundgeräusches). Die Vorteile dieser Gegebenheiten werden klarer, wenn die Randbedingungen für open-line Systeme diskutiert werden.

Bei open-line Systemen wird die Sprachprobe nicht an dem Ort abgegeben, an dem die eigentliche Verifizierung stattfindet, und die Verbindung zwischen dem Sprecher und dem Verifizierungssystem wird bei jeder Verbindung neu aufgebaut.

Typische Anwendungen, bei denen die Verifizierung über das Telefonnetz erfolgt, sind Vorgänge im wirtschaftlichen Bereich, bei denen der (nicht persönlich anwesende) Auftraggeber aus Haftungsgründen einwandfrei festgestellt werden muß, z.B. die Vergabe eines verbindlichen Auftrages (Bestellung bei einem Versandhaus, Buchung bei einem Reisebüro, Kontoüberweisung bei einer Bank) oder die autorisierte Befehlsgabe (Polizei, Militär usw.).

Bei diesen Diensten, die möglichst von jedem beliebigen Telefon erreichbar sein sollen, treten gegenüber den closed-line Systemen folgende erschwerende Randbedingungen (in der Reihenfolge der Verarbeitungskette) auf:

- Die Telefone sind in verschiedensten Umgebungen vorhanden, eine Abschirmung gegen Hintergrundgeräusche oder Maßnahmen gegen schlechte Raumakustik (Telefonzelle!) sind kaum möglich.

- Die weit verbreiteten Kohlekapselmikrophone weisen sowohl starke lineare als auch nichtlineare Verzerrungen auf (Eingeschränktes Frequenzband und großer Klirrfaktor). Auch die neuerdings verwendeten dynamischen Kapseln sind in ihrem Frequenzbereich stark eingeschränkt.

- Für die Übertragung steht nur ein begrenzter Frequenzbereich (im deutschen Fernsprechnetz 300-3400 Hz) zur Verfügung.

- Die Telefonverbindung hat - selbst bei mehrfacher Verifizierung vom gleichen Telefon aus - jedesmal ein anderes Übertragungsverhalten, da wegen des unterschiedlichen Belegungszustandes der zur Verfügung stehenden Leitungsbündel bei jedem Verbindungsaufbau eine andere Leitung zwischengeschaltet wird.

- Der Störabstand des Telefonnetzes ist - verglichen mit closed-line Systemen - sehr gering und auch abhängig vom Verbindungsaufbau.

- Die Übermittlung des Identitätsziels durch Eingabe einer Codezahl mit Hilfe der Wählscheibe kann (im deutschen Fernsprechnetz) zu Schwierigkeiten führen, wenn das Verifizierungssystem bei einem Endteilnehmer und nicht in einer Vermittlungsstation angeschlossen ist, da die Gleichstromwählimpulse durch Übertrager abgeblockt sein können.

Diese erschwerten Randbedingungen führen - verglichen mit einem closed-line System - bei gleicher Art der automatischen Verarbeitung auf jeden Fall zu einer verminderten Leistungsfähigkeit des Systems, da man den Einfluß z.B. der Bandbegrenzung und des geringen Störabstandes nicht beseitigen kann. Dagegen kann man den Einfluß unterschiedlicher Übertragungscharakteristika gewählter Telefonverbindung durch geeigneten Merkmalsgewinnung durchaus reduzieren.

Dazu muß kurz erwähnt werden, daß es - vereinfacht ausgedrückt - zwei Quellen sprecherindividuellen Verhaltens gibt. Einmal führen die natürlichen anatomischen Unterschiede der an der Spracherzeugung beteiligten Komponenten zu systematischen Veränderungen der sie beschreibenden Sprachparameter. Frauen besitzen z.B. im Mittel eine höhere Sprachgrundfrequenz als Männer, ebenso liegen die Resonanzfrequenzen des Vokaltraktes höher, da ihr Vokaltrakt kürzer ist. Solche individuellen Unterschiede liegen in der Sprache des Informatikers "hardwaremäßig" vor.

Zum anderen gibt es eine Reihe von sprecherindividuellen Besonderheiten, die sich aus der Art der dynamischen Abläufe des Sprachproduktionsprozesses erschließen lassen. Diese Abläufe sind nicht von der Natur vorgegeben, sondern werden beim Erlernen und praktischen Gebrauch einer Sprache eingeübt, liegen also "softwaremäßig" vor. Betroffen davon sind Sprechgeschwindigkeit, Betonung, Melodik, um nur einige Größen zu erwähnen.

Die Übertragung über verschiedene Telefonleitungen kann man sich nun als Kettenschaltung zweier Systeme, des Vokaltraktes $H(f)$ und

des Übertragungskanals G(f) vorstellen. Die Gesamtübertragungsfunktion der Eigenschaften des Sprachsignals beim Empfänger ist dann (bei vorausgesetzter Linearität) das Produkt beider Funktionen. Das Ergebnis ist, um es salopp auszudrücken, daß man nicht mehr weiß, ob man nun die Eigenschaften des Sprechers oder des Übertragungskanals erfaßt. Dies gilt aber nur, solange H(f) konstant ist, d.h. für einen stationären Sprachlaut. Betrachtet man dagegen längere Zeitabschnitte, z.B. die Dauer eines Codesatzes, dann sind die im Empfänger auftretenden spektralen Änderungen nur dem Sprecher zuzuschreiben,da die Übertragungsfunktion des Kanals für die Dauer einer Verbindung als konstant angenommen werden kann.

Das bedeutet nun, daß man hier die dynamischen Besonderheiten der Sprechweise verschiedener Sprecher stärker berücksichtigen muß als die stationären (durch die Anatomie bedingten) Besonderheiten (vergleiche auch die Bemerkungen über das SASIS-System in Kap. 2). Eine Möglichkeit, den Einfluß der Übertragungscharakteristik zu eliminieren, soll kurz erläutert werden. Besteht die parametrische Beschreibung R(1), ..., R(L) des Sprachsignals aus Stützstellen des Spektrums, dann lassen sich die Einzelspektren komponentenweise auf ihren Mittelwert normieren:

$$r'_i(l) = \frac{r_i(l)}{\frac{1}{L}\sum_{k=1}^{L} r_i(k)} \quad , \quad R'(l) = \begin{pmatrix} r'_1(l) \\ r'_2(l) \\ \vdots \\ r'_M(l) \end{pmatrix}$$

Bei der Übertragung des Signals wird jeder Wert des (Leistungsdichte-) Spektrums mit dem entsprechenden Wert der (Leistungs-) Übertragungsfunktion g_i multipliziert. Normiert man nun jede Komponente des übertragenen Spektrums wiederum auf seinen Mittelwert, dann ergibt sich

$$\frac{g_i \cdot r_i(l)}{\frac{1}{L}\sum_{k=1}^{L} g_i \cdot r_i(k)} = \frac{g_i \cdot r_i(l)}{g_i \cdot \frac{1}{L}\sum r_i(k)} = \frac{r_i}{\frac{1}{L}\sum r_i(k)} = r'_i(l)$$

Der auf die angegebene Art normierte Parametervektor R'(l) und damit auch der daraus gewonnene Merkmalsvektor ist also invariant gegenüber unterschiedlichen linearen Verzerrungen. Diese Eigenschaft ist bei der Merkmalsgewinnung im SPREE-System ausgenutzt worden (s. Kap. 6). Natürlich mußte dazu experimentell nachgewiesen werden, daß die normierten Spektren überhaupt für die Sprechererkennung geeignet sind /8, 16/.

Im Gegensatz zur Sprecheridentifizierung sind für die Sprecherverifizierung in den letzten Jahren mehrere Systeme labormäßig entwickelt worden. Neben dem SPREE-System des Heinrich-Hertz-Instituts, das in Kap. 6 ausführlicher vorgestellt wird, wurde bei Texas Instruments das "TI Entry Control System" entwickelt, das wohl die längste Erprobungsphase (Zugangskontrolle im TI-Rechenzentrum) hinter sich hat /14/. Bei diesem System treten Fehlerraten ε_{FR} und ε_{FA} (Stufe I) von etwa 1% auf. Ein weiteres closed-line System wird als low-cost Gerät von der Firma Philips für den Einsatz im Bankbereich (automatischer Kassenschalter) untersucht, hier traten bei ersten Laborversuchen Fehlerraten von ε_{FR} = 10% und ε_{FA} = 1% auf /18/.

In den Bell Laboratories wurde ein open-line System für den Einsatz über Telefon entwickelt, mit dem bei Erkennungsexperimenten Fehlerraten von 5% - 10% auftraten /17/. Allerdings wurde dabei eine recht unvollständige parametrische Beschreibung des Signals verwendet, lediglich die Intensitäts- und Sprachgrundfrequenzkontur eines Codesatzes wurden ausgewertet.

5. VERGLEICH ZWISCHEN STIMME, UNTERSCHRIFT UND FINGERABDRUCK

Bei der Präsentation unseres SPREE-Systems vor Fachkollegen und interessierten Laien entwickeln sich häufig Diskussionen, ob die Erkennungssicherheit eines Sprecherverifizierungssystems überhaupt ausreicht, um andere Verfahren, etwa die Legitimation durch Unterschrift oder Fingerabdruck, zu ersetzen. Abgesehen davon, daß die Sprecherverifizierung andere Verfahren nicht unbedingt erset-

zen soll, sondern auch in Kombination mit einem anderen Verfahren die Erkennungssicherheit erheblich verbessern kann, soll der Frage nach der grundsätzlichen Leistungsfähigkeit der verschiedenen Verfahren nachgegangen werden. Dabei sollen die wichtigsten Ergebnisse eines Vergleichs zwischen diesen Systemen vorgestellt werden, der im Auftrage des US-Department of Defense von der MITRE-Corporation in den Jahren 1976 und 1977 durchgeführt worden ist /19/. Die Struktur der in diesem Vergleich verwendeten Systeme,

- das Sprecherverifizierungssystem von Texas Instruments Corp.,
- ein Unterschriftsverifizierungssystem der Firma Veripen , Inc.,
- ein Fingerabdruckverifizierungssystem der Calspan Corp.,

soll kurz erläutert werden.

Beim Sprecherverifizierungssystem wird vorausgesetzt, daß der Codesatz aus vier einsilbigen Wörtern besteht, die vom System in zufälliger Reihenfolge vorgesprochen werden. Diese werden in der Vorverarbeitungsstufe durch eine Filterbank (16 Kanäle zwischen 300 und 3000 Hz) zerlegt. Als Merkmalsvektor wird pro Wort eine Folge von 6 Kurzzeitspektren aus dem Silbenkern (größte Intensität) verwendet, die mit den entsprechenden Referenzspektren des Sprechers verglichen werden.

Beim Unterschriftsverifizierungssystem wird der zeitliche Druckverlauf der Unterschrift, der zwischen Schreibgerät und Unterlage vorhanden ist, meßtechnisch erfaßt. Aus dieser Kontur werden dann 9 verschiedene Merkmale abgeleitet, die mit einer Referenz verglichen werden.

Das Fingerabdruckverifizierungssystem erzeugt zunächst ein binäres Bild der Fingerkuppen und bestimmt daraus Anfangs- und Endpunkte sowie Verzweigungen der Fingerrillen. Aus diesen Größen wird dann der Merkmalsvektor bestimmt, der wiederum durch eine Ähnlichkeitsmessung mit der abgespeicherten Referenz die endgültige Entscheidung bewirkt.

In Tab. 5.1 sind die Ergebnisse eines Feldtests angegeben, an dem über 200 Personen teilnahmen. Dabei wurden die drei Systeme jeweils gemeinsam benutzt. Die Fehlerraten ε_{FA} beziehen sich auf zufällige Täuschungsversuche (Stufe I, vergleiche Kap. 4).

	reguläre Verifizierung ε_{FA} (%)	zufälliger Täuschungsversuch ε_{FA} (%)
Sprecher-verifizierungs-system	1.1	3.3
Unterschrifts-verifizierungs-system	1.9	5.6
Fingerabdruck-verifizierungs-system	6.5	2.3

Tab. 5.1 Ergebnisse des Vergleichs zwischen Stimme, Unterschrift und Fingerabdruck

Es zeigt sich, daß das Sprecherverifizierungssystem die größte Leistungsfähigkeit hat, und daß die Ergebnisse alle in der Grössenordnung 1% ... 10% liegen. Eine weitere statistische Auswertung der Experimente ließ erkennen, daß die drei Systeme unabhängig voneinander arbeiten /19/, daß man also durch Kombination verschiedener Systeme die Erkennungssicherheit beträchtlich steigern kann.

6. DAS SPRECHERVERIFIZIERUNGSSYSTEM SPREE

6.1 Allgemeine Struktur des Systems

In den Jahren 1976-1978 wurde am Heinrich-Hertz-Institut das Online System zur Sprecherverifizierung SPREE (SPREcherErkennung) entwickelt. An dieses System wurden folgende Anforderungen gestellt:

- benutzerfreundliche Abwicklung einer Verifizierung über ein Telefon ohne zusätzliche Bedienelemente, Bedienführung durch akustische Ansagen,

- kurze Reaktionszeit des Systems (< 2 s),

- Erkennung mit Hilfe von kurzen Sprachproben (< 2 s),

- Berücksichtigung von Langzeitveränderungen der Stimme.

Wegen der Forderung nach einer kurzen Reaktionszeit des Systems erfolgt die Vorverarbeitung des Sprachsignals in Echtzeit. Mit Hilfe einer analogen Filterbank (13 Kanäle im Bereich 300- 3400 Hz) wird eine Kurzzeitspektralanalyse der Sprachprobe durchgeführt und im Rechner abgespeichert. Wegen der bereits erwähnten Kooperationsbereitschaft kann als Sprachprobe ein fester Codesatz vereinbart werden. Fig. 6.1 zeigt als Beispiel die Spektralanalyse des Codesatzes "Sesam öffne dich" eines männlichen Sprechers, die nach Digitalisierung als Zeit-Frequenz-Matrix zur Verfügung steht.

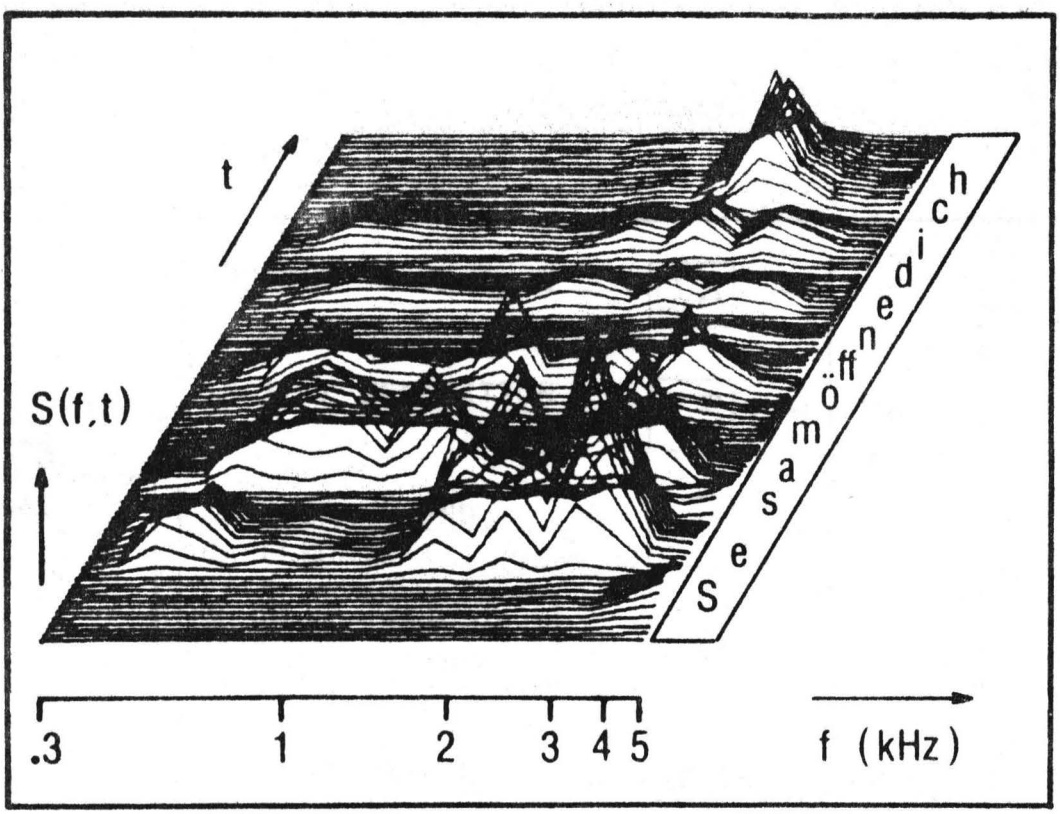


Fig. 6.1 Spektralanalyse des Codesatzes "Sesam öffne dich" eines männlichen Sprechers

Im nächsten Verarbeitungsschritt werden unterschiedliche Sprechgeschwindigkeiten, die bei fest eingestellter Abtastfrequenz zu unterschiedlichem Format der Zeit-Frequenz-Matrix führen, ausgeglichen; dazu ist eine nichtlineare Verzerrung der Zeitachse notwendig /9/. Aus der normalisierten Zeit-Frequenz-Matrix werden dann verschiedene Merkmalssätze gewonnen, die sowohl die individuelle Anatomie des Vokaltraktes als auch unterschiedliche Sprechgewohnheiten berücksichtigen; sie werden in den folgenden Abschnitten ausführlicher erläutert.

Die endgültige Entscheidung wird aufgrund von Abstandsmessungen zwischen den aktuellen Merkmalssätzen und den entsprechenden abgespeicherten Referenzsätzen getroffen. Bei erfolgreicher Verifi-

zierung werden die Referenzdaten des Sprechers durch die Daten der aktuellen Sprachprobe aufgefrischt, um den Einfluß von Langzeitveränderungen der Stimme zu berücksichtigen. Das Blockschaltbild des Systems ist in Fig. 6.2 dargestellt.

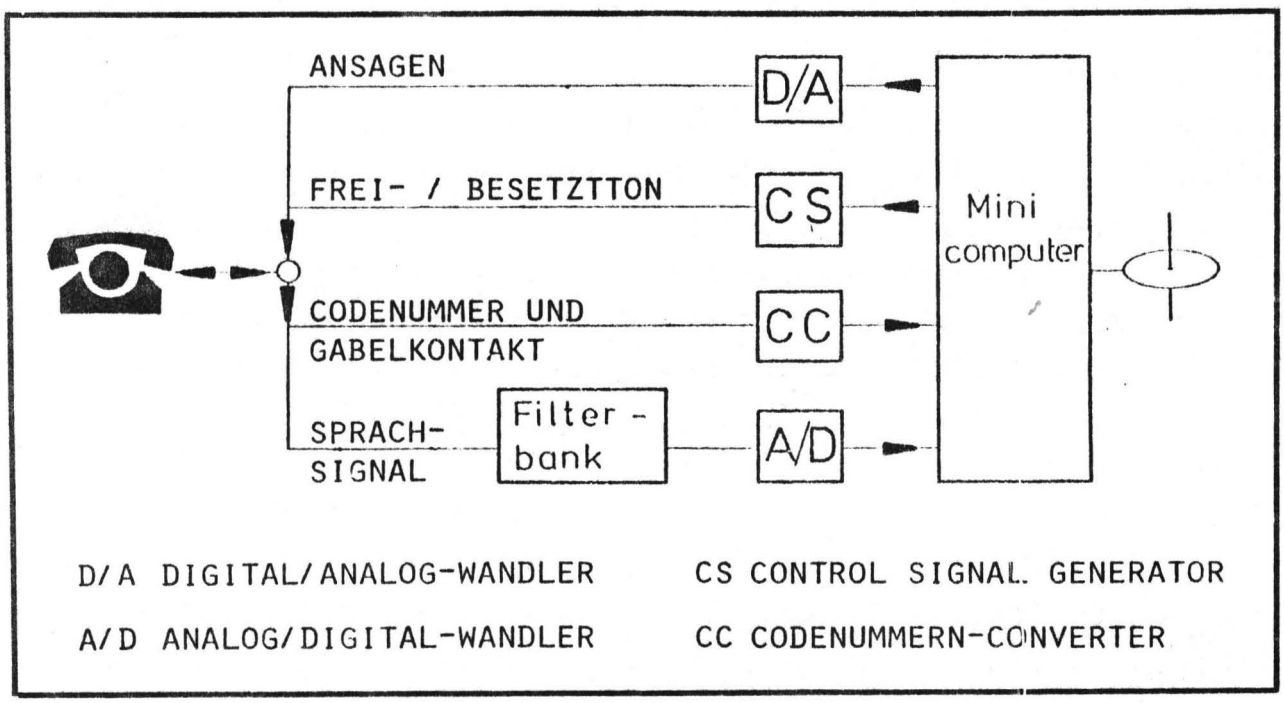


Fig. 6.2 Struktur des Sprecherverifizierungssystems SPREE

Der Benutzer gibt mit der Wählscheibe bzw. -tastatur eine ihm zugewiesene Codezahl (und damit sein Identitätsziel) an. Nach Prüfung dieser Codezahl meldet sich das System akustisch und fordert den Benutzer auf, einen vorgesprochenen Codesatz zu wiederholen (sämtliche Ansagen sind digital auf einem Hintergrundspeicher abgelegt). Der vom Benutzer nachgesprochene Codesatz wird von der Filterbank analysiert und dann im Rechner weiterverarbeitet. Das Ergebnis der Überprüfung wird dem Benutzer wiederum akustisch mitgeteilt.

6.2 Vorverarbeitung und Merkmalsgewinnung

Nach der Aufforderung durch das System muß der Sprecher einen Codesatz in einer bestimmten (durch einen Parameter wählbaren) Zeit fertig gesprochen haben. In diesem Zeitfenster werden die Kurzzeitspektren ermittelt und als Zeit-Frequenz-Matrix abgespeichert.

In dieser Matrix müssen nun zunächst Anfang und Ende des Codesatzes detektiert werden und unterschiedliche Sprechgeschwindigkeiten ausgeglichen werden. Beide Aufgaben werden mit einem "Dynamic Programming" Algorithmus gelöst, auf dessen Wirkensweise hier nicht näher eingegangen werden soll. Einen guten Überblick über diese Technik bietet White /20/, die spezielle Anwendung im SPREE-System ist in /9/ und /16/ dargestellt.

Nach der Durchführung dieser Zeitnormalisierung liegt die Zeit-Frequenz-Matrix des Codesatzes in einheitlichem Format vor. Sie bildet die Basis für die Gewinnung von einzelnen Merkmalsvektoren, die nun vorgestellt werden sollen:

Gemitteltes Spektrum: Durch Mittelung der Zeit-Frequenz-Matrix über die Zeit erhält man einen Merkmalssatz, der überwiegend individuelle anatomische Gegebenheiten der an der Sprachproduktion beteiligten Komponenten wiederspiegelt, und zwar umso besser, je länger die Sprachprobe ist.

Normierte Kurzzeitspektren: Ein einzelnes Spektrum der Zeit-Frequenz-Matrix beschreibt einen bestimmten Sprachlaut innerhalb des Codesatzes. Wie bereits früher gezeigt werden konnte /21/, sind verschiedene Sprachlaute in unterschiedlichem Maße zur Sprechererkennung geeignet, da sie die sprecherindividuelle Anatomie des Vokaltraktes in unterschiedlicher Weise auf das Sprachsignal abbilden. Als weitere Merkmalssätze werden deshalb Einzelspektren von besonders geeigneten Sprachlauten verwendet. Diese können auf das gemittelte Spektrum normiert werden. Die auf diese Weise gebilde-

ten Merkmalssätze sind (im Gegensatz zum gemittelten Spektrum selbst) invariant gegenüber wechselnden linearen Übertragungsverzerrungen, die z.B. bei der Übertragung von Codesätzen über gewählte Telefonverbindungen auftreten können (vergl. Kap. 4).

Intensitätskontur: Durch Mittelung der Zeit-Frequenz-Matrix über alle Frequenzen erhält man die Intensitätskontur. Sie gibt die sprecherindividuelle Sprachdynamik wieder und beschreibt somit weniger anatomische Unterschiede als vielmehr unterschiedliche Sprechgewohnheiten verschiedener Sprecher.

Stationaritätskontur: Die Abweichung zwischen 2 aufeinanderfolgenden Spektren, die sich z.B. durch den quadratischen Fehler kennzeichnen läßt, ist ein Maß für die Veränderung der Vokaltraktsgeometrie im betrachteten Zeitraum. Wird dieses Maß über die gesamte Zeit-Frequenz-Matrix berechnet, dann erhält man die sog. Stationaritätskontur, die vorwiegend individuelle Sprechgewohnheiten wiederspiegelt.

Die einzelnen aus einer Sprachprobe gewonnenen Merkmalsvektoren werden getrennt mit den jeweiligen Referenzvektoren verglichen. Anschließend werden die daraus resultierenden Ähnlichkeitsmaße entsprechend ihrer Zuverlässigkeit gewichtet und aufsummiert, um die endgültige Entscheidung durch Vergleich mit einer fest eingestellten Schwelle zu treffen. Einzelheiten zur Klassifizierung sind in /8/ dargestellt.

6.3 Ergebnisse

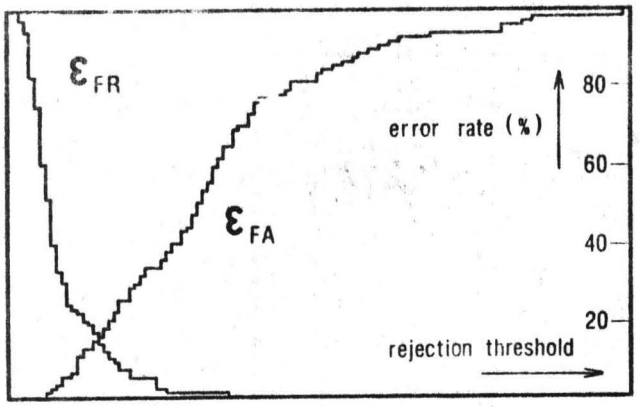
Um die Güte verschiedener Merkmalssätze zu beurteilen, wurden Erkennungsexperimente durchgeführt. Dazu wurde aus einer größeren Stichprobe von über 150 Sprechern eine Stichprobe von 5 Sprechern nach dem Kriterium ausgewählt, daß ihre Stimmen (nach subjektiver Beurteilung) eine möglichst große Ähnlichkeit aufweisen sollten. Von diesen Sprechern wurde zuerst eine "Sprecherkartei" aufgebaut, dazu wurden pro Sprecher 10 Sprachproben des Codesatzes

"Sesam öffne dich" verwendet. Anschließend wurden von jedem Sprecher 60 reguläre Verifizierungsversuche durchgeführt, außerdem 60 beabsichtigte Täuschungsversuche (Stufe II!), die sich gleichmässig zwischen die jeweils 4 verbleibenden Sprechern aufteilten.

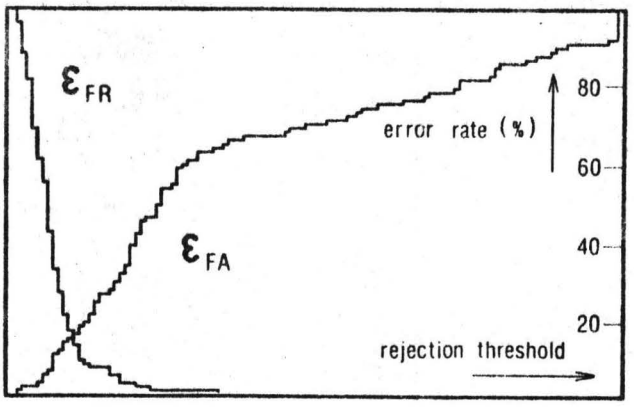
Den Täuschern wurde ihre Aufgabe so weit wie möglich erleichtert, sie konnten Tonbandaufzeichnungen und Originalsprachproben der wahren Sprecher abhören, außerdem wurden sie unmittelbar über den Erfolg des jeweiligen Täuschungsversuchs, d.h. über die Ähnlichkeit ihrer Sprachprobe quantitativ informiert, so daß sie gezielt Variationen in ihrer Sprechweise vornehmen konnten.

In Fig. 6.3 sind die Ergebnisse der Erkennungsexperimente für die einzelnen Merkmalssätze als Funktion der Rückweisungsschwelle dargestellt.

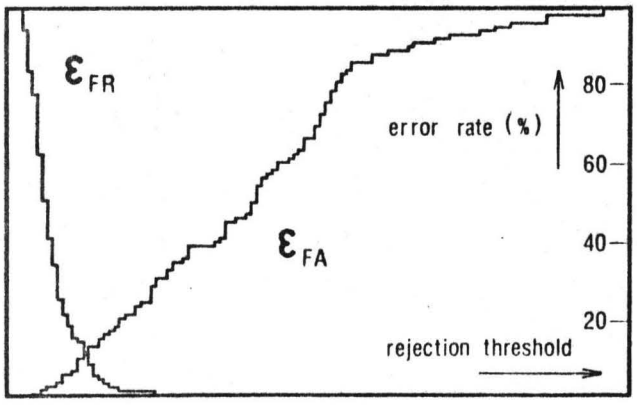
Da diese Ergebnisse an anderer Stelle von Höfker /8/ ausführlich diskutiert werden, soll hier lediglich hervorgehoben werden, daß mit keinem einzelnen Merkmalsvektor eine fehlerfreie Klassifizierung möglich ist. Kombiniert man die Merkmalssätze jedoch, um daraus eine Gesamtentscheidung abzuleiten, dann ist bei richtiger Einstellung der Rückweisungsschwelle eine fehlerfreie Klassifizierung möglich. Dies geht aus Fig. 6.4 hervor; hier sind die bei einer kombinierten Gesamtentscheidung auftretenden Fehlerraten wiederum als Funktion der Rückweisungsschwelle dargestellt.



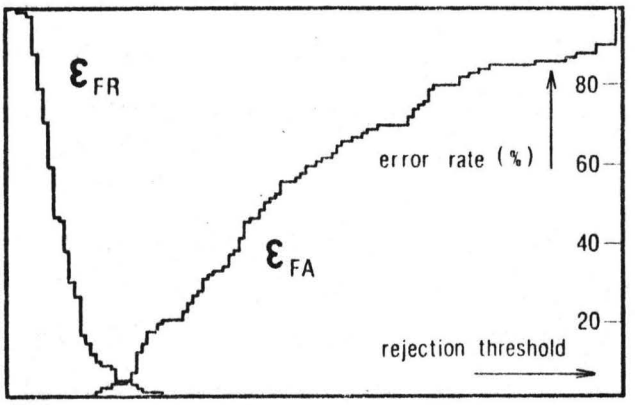
(a) Gemitteltes Spektrum



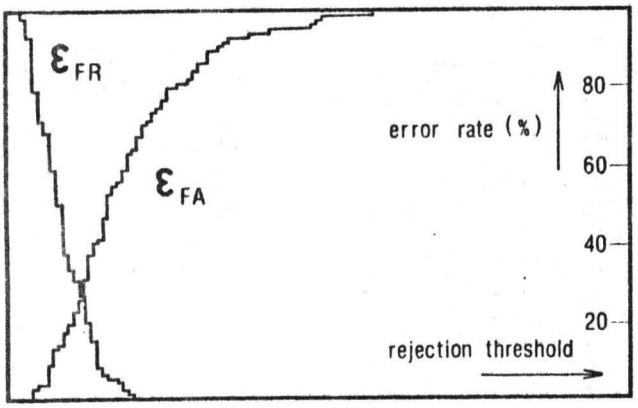
(b) Spektrum des /e/



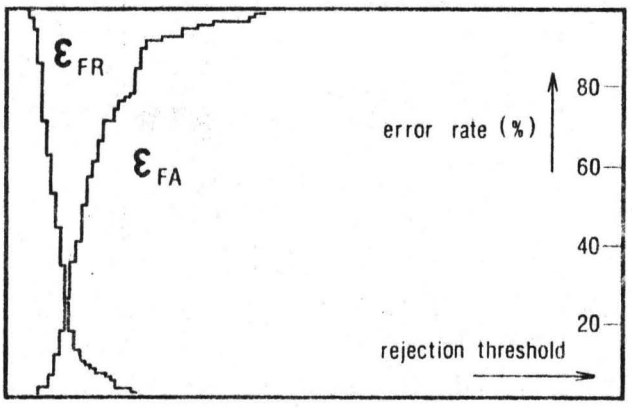
(c) Spektrum des /m/



(d) Spektrum des /ə/



(e) Intensitätskontur



(f) Stationaritätskontur

Fig. 6.3 Fehlerraten als Funktion der Rückweisungsschwelle für verschiedene Merkmalssätze

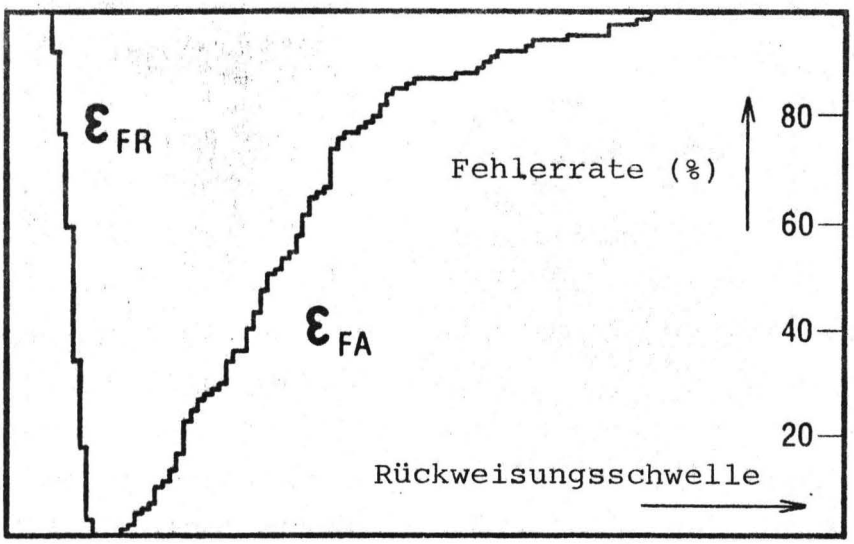


Fig. 6.4 Fehlerraten als Funktion der Rückweisungsschwelle bei Kombination der Merkmalssätze

Bei der Beurteilung dieses Ergebnisses sollte man sich vor Augen halten, daß die Experimente mit beabsichtigten Täuschungsversuchen durchgeführt wurden und die Täuscher außerdem interne Systeminformation über die auftretenden Ähnlichkeiten hatten.

7. Schlußbemerkungen

Die bisher durchgeführten Untersuchungen haben gezeigt, daß die automatische Erkennung von Sprechern durch Computer prinzipiell möglich ist. Eine dreistufige Verarbeitung hat sich dabei besonders bewährt. In der Vorverarbeitungsstufe werden aus dem Sprachsignal Parameter ermittelt, die unmittelbar mit dem Sprachproduktionsprozeß zusammenhängen. Aus diesen wird in der Merkmalsextraktionsstufe ein Merkmalsvektor abgeleitet, der für einen Sprecher möglichst reproduzierbare und für verschiedene Sprecher möglichst unterschiedliche Werte annehmen soll. Dabei existieren verschiedene Techniken, um die störende Abhängigkeit der Sprachparameter vom Sprachinhalt zu beseitigen. In der letzten Stufe wird der Merkmalsvektor automatisch klassifiziert, also einem Sprecher zugeordnet, hier werden meist Standardverfahren der Mustererkennung verwendet.

Die größten Fortschritte wurden erwartungsgemäß bei der Sprecherverifizierung erzielt, weil sich die Sprecher hier kooperativ verhalten und sich in gewisser Weise dem System anpassen, indem sie sich bemühen, ihren Codesatz bei jedem Erkennungsversuch gleichartig auszusprechen. Auch die äußeren Randbedingungen wie Umgebungsgeräusche usw. lassen sich weitgehend beherrschen. Die mit dem jetzigen Stand der Technik erzielte Leistungsfähigkeit ist für wirtschaftliche Anwendungen durchaus befriedigend.

Auch bei der Sprecheridentifizierung wurden hohe Erkennungsraten erreicht /22/. Diese vermitteln aber einen falschen Eindruck über den wahren Stand der Technik, da sie ausschließlich unter Laborbedingungen erzielt wurden und die Sprecher in keiner Weise versuchten, ihre Stimme zu verstellen. Diese Absicht muß aber bei der eigentlichen Anwendung der Sprecheridentifizierung, nämlich der Täterermittlung in der Kriminalistik, unterstellt werden. Hinzu kommt noch, daß sich bei dieser Anwendung die äußeren Randbedingungen nur selten kontrollieren bzw. reproduzieren lassen. Aufgrund dieser Schwierigkeiten erscheint ein vollautomatisches Sprecheridentifizierungssystem heute kaum realisierbar, jedoch kann

ein teilautomatisches interaktives System (wie z.B. das SASIS-System, s. Kap. 3) gute Hilfsdienste leisten. Es sind hier 2 Möglichkeiten denkbar. Zum einen kann der Mensch das System unterstützen, indem er z.B. bestimmte Signalsegmente markiert, die dann anschließend automatisch klassifiziert werden. Zum anderen kann, wenn eine letzte Beurteilung und Klassifizierung durch den Menschen vorgenommen werden soll, das System den Menschen unterstützen, indem es ihm Parameterverläufe und Ähnlichkeitsmaße in aufbereiteter Form übersichtlich zur Verfügung stellt.

LITERATUR

1. P. Jesorsky: "Principles of Automatic Speaker Recognition" in: Speech Communication with Computers, Ed.: L. Bolc, Hanser/Macmillan, 1978

2. A.E. Rosenberg: "Automatic Speaker Verification: A Review", Proc. of the IEEE vol. 64, No.4, pp. 475-487, 1976

3. B.S. Atal: "Automatic Recognition of Speakers from their Voices", Proc. of the IEEE vl. 64, No.4, pp. 460-475, 1976

4. A.E. Rosenberg: "Listener Performance in Speaker Verification Tasks", IEEE Transactions, vol. A 21, No.3, June 1973

5. G. Fant: "Acoustic Theory of Speech Production", The Hague, Netherlands, Mouton, 1970

6. J.L. Flanagan: "Speech Analysis, Synthesis and Perception", New York, Springer-Verlag 1972

7. M. Talmi: "Spektrale Vorverarbeitung von Sprachsignalen", Abschlußbericht Teil C des Projekts "Entwicklung eines Systems zur automatischen Sprechererkennung", HHI, Berlin 1979

8. U. Höfker: "Die Merkmalsgewinnung bei der Sprechererkennung", Abschlußbericht Teil E des Projektes "Entwicklung eines Systems zur automatischen Sprechererkennung", HHI, Berlin 1979

9. B. Kriener: "Die Zeitnormalisierung von Sprachparameterkonturen bei der Sprecherverifizierung", Abschlußbericht Teil D des Projektes "Entwicklung eines Systems zur automatischen Sprechererkennung", HHI, Berlin 1979

10. K. Fukunaga: "Introduction to Statistical Pattern Recognition", Academic Press, New York, London 1972

11. R.O. Duda und P.E. Hart: "Pattern Classification and Scene Analysis", John Wiley, New York, London 1973

12. E. A. Patrick: "Fundamentals of Pattern Recognition", Prentice Hall, London 1972

13. J. Meltzer: "Speaker Identification", Program 7907, Final Report, Aerospace Corporation 1977

14. G.R. Doddington und B.M. Hydrick: "Speaker Verification II", Texas Instruments Inc., RADC-TR-75-274, Final Technical Report, 1975

15. M. Kuhn: "Access Control by Means of Automatic Speaker Verification", Proc., Science and Security, Brighton, Sept. 1978

16. U. Höfker und P. Jesorsky: "Structure and Performance of an On-line Speaker Verification System", Proc., 1979 IEEE International Congress on Acoustics, Speech and Signal Processing, Washington, April 1979

17. A.E. Rosenberg: "Evaluation of an Automatic Speaker Verification System Over Telephone Lines", Bell Syst. Techn. Journ., Vol. 55, 1976

18. R. Geppert und H. Piotrowski: "Ein einsatzfähiges System zur Sprecherverifikation", Proc., DFG-Kolloquium Digitale Signalverarbeitung, Göttingen, Okt. 1978

19. A. Fejfar und Y. W. Myers: "The Testing of Three Automatic Identity Verification Techniques for Entry Control", Proc., Oxford Conference on Security, 1977

20. G.M. White: "Dynamic Programming, the Viterbi Algorithm, and Low Cost Speech Recognition", Proc. ICASSP 1978, April 1978

21. U. Höfker: "Die Eignung verschiedener Sprachlaute für die automatische Sprechererkennung", Proc., 5. IITB Koll. Mustererkennung, Karlsruhe, Febr. 1976

22. E. Bunge: "Vergleichende systematische Untersuchungen zur automatischen Identifikation und Verifikation kooperativer Sprecher", Dissertation, TH Darmstadt 1977

Heinrich-Hertz-Institut
für Nachrichtentechnik
Berlin GmbH