VOCALISE:

eine gemeinsame Plattform für die Anwendung automatischer und semiautomatischer Methoden in forensischen Stimmenvergleichen

Marianne Jessen, Oscar Forth & Anil Alexander

1. Einleitung

Im Bereich der forensischen Fachdisziplin, die als Sprechererkennung und Tonträgeranalyse oder auf vergleichbare Weise bezeichnet wird, nimmt der Stimmenvergleich eine zentrale Rolle ein. In einem forensischen Stimmenvergleich besteht die Aufgabe darin, die Stimme eines Täters zu vergleichen mit der Stimme eines Verdächtigen, von denen jeweils Audioaufnahmen existieren. Bei der Aufnahme des Täters kann es sich beispielsweise um einen Droh- oder Erpressungsanruf handeln, um eine Kommunikation im Rahmen von Drogenhandel oder um Absprachen im Rahmen spezifischer Wirtschaftkriminalitätsdelikte. Die Stimme des Verdächtigen kann entweder explizit von einem Experten für Stimmenvergleiche im Rahmen einer Vergleichsaufnahme erhoben werden oder es gibt aus anderen Quellen bereits eine oder mehrere Aufnahmen, die dem Verdächtigen zugeordnet werden, beispielsweise im Rahmen von Telekommunikationsüberwachungen (TKÜ). Die Aufgabe des Experten besteht nun darin, Tataufnahme(n) und Vergleichsaufnahme(n) einer genauen Analyse zu unterziehen, diese zu vergleichen, und zum Abschluss dieses Prozesses zu einer wahrscheinlichkeitsbasierten Schlussfolgerung zu gelangen, die relevant ist für die Frage, ob der Sprecher in der Tataufnahme identisch oder nichtidentisch ist mit dem Sprecher des Vergleichsmaterials oder ob in dem Tatmaterial eine andere Person gesprochen hat.

Die Historie des Stimmenvergleichs reicht international zurück bis mindestens in die 1960er Jahre und es gibt

Zusammenfassung

Ein neues Computerprogramm namens VOCALISE (VOice Comparison and Analysis of the LIkelihood of Speech Evidence) wird vorgestellt, das gemeinsam von den Autoren entwickelt wurde. VOCALISE stellt eine gemeinsame Plattform dar, mit der im Rahmen von forensischen Stimmenvergleichen erstmals und auf benutzerfreundliche Weise sowohl mit der Methode der automatischen Sprechererkennung als auch der Methode der semiautomatischen Sprechererkennung gearbeitet werden kann. Diese beiden Methoden werden anhand von anonymisierten Falldaten und von Forschungsdaten im Detail vorgestellt und diskutiert. Anwendungsaspekte, wie den Umgang mit umfangreichen TKÜ-Daten, werden ebenso thematisiert wie Forschungsaspekte, z. B. die Anzahl erforderlicher Gaußmodule bei der Sprechermodellierung von Langzeitformanten.

Forensischer Stimmenvergleich; automatische Sprechererkennung; semiautomatische Sprechererkennung; Gaussian Mixture Models; TKÜ (Telekommunikationsüberwachung)

Abstract

A new computer program called VOCALISE (VOice Comparison and Analysis of the LIkelihood of Speech Evidence) is introduced that has been developed jointly by the authors of this paper. VOCALISE offers a common environment in which for the first time both automatic speaker and semiautomatic methods in forensic voice comparison can be applied in a user-friendly manner. These two methods will be demonstrated and discussed based on anonymised case data and on a research corpus. Both practical and research-oriented topics will be discussed, including the handling of large amounts of telephone interception data (TKÜ) and the investigation of the number of Gaussian distributions necessary for the speaker modeling of long-term formant frequencies.

Forensic voice comparison; automatic speaker recognition; semiautomatic speaker recognition; Gaussian Mixture Models; telephone interception

Hinweise, dass bereits in der Zeit des zweiten Weltkriegs und kurz danach (vor allem in der Sowjetunion) geheime Projekte zur Sprechererkennung existierten (Hollien 1990, 2002). Richtig konsolidiert und als allgemein bekannte und anerkannte forensische Disziplin etabliert hat sich der Stimmenvergleich dann erst in den 1980er Jahren (Nolan 1983, Künzel 1987, Baldwin & French 1990, Hollien 1990). In diesen Jahren hat sich auch ein Verfahren etabliert, das als auditiv-akustische Methode bezeichnet wird. Wie der Name bereits sagt, wird im Rahmen der auditiv-akustischen Methode das Tat- und Vergleichsmaterial sowohl auditiv als auch akustisch analysiert. Zur auditiven Analyse gehört unter anderem das Identifizieren von regionalen und fremdsprachakzentuierten Merkmalen, von Stimmqualitäten (z. B. raue, behauchte, nasale Stimme) und von sprachlichen und nichtsprachlichen Angewohnheiten in Bereichen wie Redefluss, Pausenverhalten, Artikulationsgenauigkeit und Atmen. Im akustischen Teil werden Methoden der akustischen Phonetik verwendet, u. a. die Grundfrequenzanalyse, die Formantenmessung und die Messung von Silbendauerwerten bzw. des Sprechtempos. Auf das Thema der Formantenmessung wird im Rahmen dieses Artikels noch genauer eingegangen. Ziel der auditiv-akustischen Methode ist es, ein möglichst breit angelegtes und vielfältiges Inventar von sprecherunterscheidenden Merkmalen zu erfassen, um auf diese Weise die Zuverlässigkeit der Schlussfolgerungen in Hinblick auf Identität bzw. Nichtidentität der zu vergleichenden Sprecher zu verbessern. Diese Merkmale sollten möglichst wenig miteinander korreliert sein, um auch mit einer beschränkten Anzahl analysierter Merkmale einen hohen Informationsgehalt zu erzielen (Rose 2002). Einen aktuellen Gesamtüberblick über die auditiv-akustische Methode liefert Jessen (2012).

Die auditiv-akustische Methode hat ihre Tradition und Fundierung in den Fachdisziplinen Phonetik und Linguistik. Parallel zu der Entwicklung dieser Methode gab es im Verlaufe der letzten Jahrzehnte Entwicklungen im Rahmen dessen, was als automatische Sprechererkennung bezeichnet wird. Automatische Sprechererkennung hat ihre Tradition in der Sprachtechnologie (zu der u. a. auch Bereiche wie automatische Spracherkennung und Text-to-Speech-Sprachsynthese gehört) und basiert auf Methoden der Signalverarbeitung, der Mustererkennung und des maschinellen Lernens. Die ersten Erfolge der automatischen Sprechererkennung liegen in Bereichen wie der Sprecherverifikation bei Zugangskontrollen zu sensiblen Bereichen oder Daten (z. B. Telebanking). Diese Erfolge liegen u. a. darin begründet, dass in solchen Anwendungen die Qualität der Aufnahmen in der Regel sehr gut ist, dass die Sprecher kooperativ sind und dass ggf. auch bestimmte vorgegebene Wortlaute gesprochen werden. Forensische Situationen sind für Methoden der automatischen Sprechererkennung deutlich schwieriger zu bewältigen. Hier ist die Qualität (und Dauer) der Aufnahmen oft gering, die Sprecher oft unkooperativ und die Wortlaute und Sprechstile meist nicht kontrollierbar. Zwar wurde bereits in den 1970er und 1980er Jahren versucht, die automatische Sprechererkennung für forensische Zwecke nutzbar zu machen, aber die genannten Schwierigkeiten führten dazu, dass diese Versuche nicht erfolgreich genug waren. Hinzu kam, dass die entsprechende Computertechnik in diesen Zeiten noch nicht ausgereift genug war und dass bestimmte effektive Methoden der Sprechermodellierung (z. B. das Gaussian Mixture Modelling, das hier noch zur Sprache kommen wird) noch nicht entwickelt waren. Mit Fortschreiten der technologischen Möglichkeiten wurde die Anwendung der automatischen Sprechererkennung für forensische Zwecke immer mehr in Erwägung gezogen. So wird die automatische Sprechererkennung in der nationalen und internationalen Sprechererkennung in zunehmendem Maße angewendet. Im Bundeskriminalamt beispielsweise kommt sie seit 2006 in Form des Systems SPES („Sprechererkennungssystem“) zum Einsatz, das zusammen mit der Fachhochschule Koblenz, Lehrstuhl Franz Broß, entwickelt wurde. Becker (2012) gibt einen Überblick über die frühen und neuen Ansätze der forensischen automatischen Sprechererkennung.

Neben der auditiv-akustischen Methode und der automatischen Sprechererkennung gibt einen dritten Ansatz, der als semiautomatische Sprechererkennung bezeichnet wird. Die semiautomatische Methode enthält Elemente aus den beiden erstgenannten Methoden. Aus der auditiv-akustischen Methode enthält sie die akustisch-phonetischen Anteile. Dies sind zurzeit in erster Linie die Formantenmessungen, aber auch andere Bereiche wie die Grundfrequenzanalyse oder Dauer- und Tempomessungen fallen in diesen Rahmen. Aus der automatischen Sprechererkennung übernimmt sie bestimmte Verfahren der Sprechermodellierung (wie das Gaussian Mixture Modelling) und Verfahren der Quantifizierung von Sprecherähnlichkeiten (sog. likelihood scores). Das Präfix „semi“ in dem Namen der Methode bezieht sich auf die akustisch-phonetischen und die damit verbundenen manuellen und auf phonetischem Wissen basierenden Anteile, wie z. B. das Korrigieren von Formantenspuren (Engl. formant tracks) oder das Auswählen und

Segmentieren von Vokalen, Konsonanten oder Silben. Der Wortstamm „automatisch“ bezieht sich auf die daran anschließenden voll automatisierbaren Schritte Sprechermodellierung und Ähnlichkeitsberechnung. Erste Ansätze der semiautomatischen Sprechererkennung gibt es aus den 1990er und frühen 2000er Jahren, insbesondere in Form des Programms SAUSI aus den USA (Hollien 1990, 2002), dem Programm IDEM aus Italien, SIVE aus Litauen und DIALECT aus Russland (siehe Überblick von Broeders 2001). Diese Systeme sind allerdings entweder nur bedingt ausgereift oder nur sehr geringfügig und wenig detailliert in Form von Publikationen dokumentiert. Aktueller, ausgereifter und sehr detailliert beschrieben hingegen sind die Forschungen der australischen Forschergruppe um Phil Rose und Geoffrey Morrison (z. B. Rose 2002, 2010, Morrison 2011, Morrison et al. 2011).[1] Die semiautomatische Sprechererkennung findet derzeit, soweit bekannt ist, in der forensischen Fallarbeit nur sehr vereinzelt Anwendung, wahrscheinlich in erster Linie in Australien sowie in einigen der erwähnen osteuropäischen Länder, u. a. Litauen. Außerdem gab es bis jetzt für die moderne semiautomatische Sprechererkennung keine benutzerfreundliche Anwenderoberfläche, d.h. es waren Spezialkenntnisse in Programmen zur mathematischen Modellierung wie MATLAB oder zur statistischen Verarbeitung wie R erforderlich, was mit dazu beigetragen haben mag, dass die Verbreitung dieser Methode in forensischen Instituten bis jetzt nur sehr eingeschränkt ist. Das von den Autoren entwickelte Programm VOCALISE enthält nun erstmals einen solchen benutzerfreundlichen Zugang zur semiautomatischen Sprechererkennung. Gleichzeitig und im Rahmen der gleichen Oberfläche ermöglicht das Programm auch die Anwendung der automatischen Sprechererkennung.

Die Funktionsweise von VOCALISE sowie einige Aspekte aus der Praxis und Forschung werden in diesem Artikel vorgestellt. In Abschnitt 2 werden einige Rahmeninformationen über die Entwicklung von VOCALISE gegeben. In Abschnitt 3 wird auf die automatische Sprechererkennung in VOCALISE eingegangen und in Abschnitt 4 die semiautomatische Methode vorgestellt. Abschließend wird in Abschnitt 5 der Funktionsumfang und die praktische Anwendung von VOCALISE noch einmal im Gesamtzusammenhang diskutiert.

1 Rose und Morrison (und weitere ihrer Studenten und Mitarbeiter in Australien) verwenden zwar nicht die Bezeichnung semiautomatisch , aber die Prinzipien ihrer Ansätze entsprechen der Charakterisierung der semiautomatischen Methode, wie sie hier gegeben wird.

2. Rahmeninformationen zu VOCALISE

VOCALISE ist ein Acronym und steht für VOice Comparison and Analysis of the LIkelihood of Speech Evidence. Dieser Name besagt, dass es in dem Programm um Stimmenvergleiche (Engl.: voice comparison) geht, dass sprachliche Muster analysiert werden, die als Beweismittel (Engl.: evidence) verwendet werden, und dass diese Analyse unter Anwendung von likelihood scores geschieht, was im Folgeabschnitt noch genauer zu erläutern ist. Das Programm VOCALISE ist eine gemeinsame Entwicklung zwischen Marianne Jessen, die gutachterlich im Bereich der Sprechererkennung tätig ist, und Anil Alexander und Oscar Forth, Gründer der Firma der Oxford Wave Research. Oxford Wave Research ist eine Forschungs- und Entwicklungsfirma mit Sitz in Oxford, UK, die sich darauf spezialisiert hat, Lösungen für Behörden und Firmen im Bereich Forensik, Polizeiarbeit und Militär in Großbritannien und weltweit anzubieten, die im Zusammenhang mit Fragen der Audiosignalverarbeitung und Mustererkennung stehen. Unmittelbares Ziel der Kooperation war es, ein System zur automatischen Sprechererkennung zu realisieren, um auf dem aktuellen Stand der forensischen Sprechererkennung die Zuverlässigkeit und Beweiskraft der Stimmenvergleichsgutachten zu verbessern. Die Kernmethode der Stimmenvergleichsbegutachtung sollte weiterhin die auditiv-akustische Methode bleiben, aber diese sollte ergänzt werden durch die automatische Sprechererkennung – dort, wo diese bei ausreichender Materialqualität und -quantität anwendbar ist. Während es verschiedene Verfahren zur automatischen Sprechererkennung gibt, hat sich eines als besonders robust erwiesen und sich in der Literatur etabliert. Hierbei handelt es sich um ein sog. GMM-UBM-System (diese Abkürzungen werden im Folgeabschnitt erklärt). Ziel der Kooperation war es, ein solches GMM-UBM-System zu realisieren. Dieses sollte möglichst transparent gestaltet werden, so dass der Benutzer zu jeder Zeit den Zusammenhang zu den analysierten Sprachsignalen behält (diese also in jeder Phase der Analyse anhören kann) und die Möglichkeit erhält, verschiedene Parametereinstellungen selbst zu setzen, anstatt dass solche Einstellungen fest und unveränderlich im System verankert sind.

Im Verlaufe der Kooperation zeigte sich, dass die Möglichkeit bestand, neben der automatischen Sprechererkennung durch ein GMM-UBM System auch eine semiautomatische Sprechererkennung zu realisieren, die in Hinblick auf Sprechermodellierung und Ähnlichkeitsberechnung nach den gleichen Prinzipien arbeitet wie die

automatische Sprechererkennung. Diese Funktionalität wurde dann im Rahmen eines Ergänzungsauftrages von Marianne Jessen an die Firma Oxford Wave Research in das System VOCALISE integriert.

VOCALISE ermöglicht es, Stimmenvergleiche auf automatischer und semiautomatischer Grundlage durchzuführen. Für jede einzelne Gegenüberstellung einer Tataufnahme mit einer Vergleichsaufnahme gibt das System einen Ähnlichkeitswert aus, der als likelihood score, oder einfach score bezeichnet wird. Damit klar wird, welchen Beweiswert verschiedene scores in Hinblick auf die Frage haben, ob die untersuchten Sprecher identisch oder verschieden sind, ist es erforderlich, einen Systemtest durchzuführen. In einem Systemtest werden viele einzelne Stimmenvergleiche durchgeführt, von denen die Ergebnisse bereits bekannt sind. Der Systemtest zeigt dann an, wie gut das System (also in diesem Fall das automatische oder semiautomatische Modul von VOCALISE) in der Lage ist, Paare mit gleichen Sprechern von Paaren mit verschiedenen Sprechern zu unterscheiden. Für solche Systemtests hat die Firma Oxford Wave Research ein Programm namens Bio-Metrics entwickelt. Bio-Metrics existierte bereits, bevor VOCALISE entwickelt wurde, es wurde allerdings in jüngster Zeit in Hinblick auf das Zusammenwirken mit VOCALISE noch etwas optimiert. Wenn im Verlaufe des Artikels Ergebnisse aus der praktischen Arbeit und Forschung mit VOCALISE vorgestellt werden, wird gleichzeitig auch Bezug genommen auf die Systemtest-Funktionalitäten von Bio-Metrics.

Ein wichtiger Einfluss in der Entwicklung von VOCALISE ergibt sich auch aus Erfahrungen des Autors Anil Alexander in der Entwicklung des Systems ASPIC (Automatic Speaker Individualisation by Computer), die unter der Leitung von Andrzej Drygaljo an der technischen Universität (EPFL) in Lausanne durchgeführt wurden (Alexander 2005). Wesentliche Aspekte jenes Systems sind auch in die Entwicklung des oben genannten SPES-Systems eingegangen.

3. Automatische Sprechererkennung

Sowohl die automatische als auch die semiautomatische Sprechererkennung bestehen aus (zumindest) den folgenden drei nacheinander ausgeführten Komponenten: 1. Merkmalsableitung, 2. Sprechermodellierung und 3. die Quantifizierung von Ähnlichkeiten zwischen den zu vergleichenden Stimmproben. Die zweite und dritte Komponente, also die Sprechermodellierung und die Ähnlichkeitsberechnung, gelten in sehr ähnlicher Weise sowohl für die automatische als auch die semiautomatische Methode. Dies gilt teilweise bereits allgemein, teilweise ist es aber auch das spezifische Merkmal des hier vorgestellten Programms VOCALISE, dass dort die konzeptionellen und anwendungsbezogenen Ähnlichkeiten in der zweiten und dritten Komponente besonders deutlich herausgearbeitet worden sind.

Der wesentliche Unterschied zwischen der automatischen und der semiautomatischen Methode liegt also in der ersten Komponente, d.h. der Merkmalsableitung. Die Art der Merkmale, die in der automatischen Methode erhoben werden, unterscheidet sich von den Merkmalen aus der semiautomatischen Methode. In der semiautomatischen Methode handelt es sich, wie schon in Abschnitt 1 erklärt, um Merkmale, die aus der akustischen Phonetik stammen. In diesem Artikel wird als akustisch-phonetisches Merkmal mit sog. Langzeitformanten gearbeitet. In der automatischen Sprechererkennung hingegen wird mit sog. MFCCs (Mel Frequency Cepstral Coefficients) bearbeitet.

Bei den MFCC handelt es sich um akustische Kennwerte (sog. Cepstralkoeffizienten), die mithilfe von Spektralanalyse automatisch in kurzen Zeitabständen aus dem Signal extrahiert werden. Die Cepstralkoeffizienten sind fast komplett de-korreliert und haben deshalb einen hohen Informationsgehalt. Mit Hilfe von nur ca. 13 Cepstralkoeffizienten (Standardeinstellung in VOCALISE, aber vom Benutzer veränderbar) lasst sich der größte Teil der ursprünglichen spektralen Gestalt des Sprachsignals wiederherstellen, d.h. die meiste der im Signal enthaltenen Information über die Sprachlaute und deren Sprecher wird mit solch einer kleinen Anzahl von Koeffizienten erfasst. Leider werden mit den MFCC nicht nur Sprach- und Sprecherinformationen erfasst, sondern auch Störanteile. Wenn also beispielweise in einem Signal ein starker Rauschanteil vorhanden ist und sich dieser auf die Cepstralkoeffizienten auswirkt, „weiß“ das auf MFCC basierende System nicht, ob diese Rauschanteile zu den stimmlich/sprachlichen Eigenschaften gehören oder aber aus Störgeräuschen aus der Umwelt bzw. einer mangelhaften Aufzeichnungstechnik resultieren. Zumindest teilweise ist dieses Problem dadurch lösbar, dass durch Verfahren wie „mean subtraction“ oder „mean variance normalisation“, wie sie in VOCALISE ausgewählt werden können, solche Anteile als „Kanaleigenschaften“ (als Oberbegriff für Anteile, die nicht Träger von stimmlich/sprachlichen Informationen sind) identifiziert und

danach vom Ergebnis der Cepstralanalyse abgezogen werden können. Genaueres zur Cepstralanalyse wird von Rose (2013) erläutert.

Der zweite Schritt der automatischen Sprechererkennung, also die Sprechermodellierung, erfolgt mit sog. GMM (Gaussian Mixture Models). GMM ist eine Methode der Modellierung von Werteverteilungen. Viele statistische Verfahren basieren darauf, dass Verteilungen mit einfachen Gaußverteilungen (Normalverteilungskurve) modelliert werden. Solche Verteilungsmodelle sind aus nur zwei Kennwerten herleitbar, dem Mittelwert und der Varianz (bzw. Standardabweichung). Im Unterschied hierzu wird bei den GMM eine Werteverteilung nicht nur durch eine Gaußverteilung modelliert, sondern durch viele Gaußverteilungen (auch Gaußmodule genannt), die alle einen anderen Wert für Mittelwert und Varianz haben. In VOCALISE beträgt die (veränderbare) Anzahl der Gaußverteilungen 32. Die Werteverteilungen, um die es in der automatischen Sprechererkennung geht, sind die Zusammenstellungen aller Cepstralkoeffizienten, die über den Zeitverlauf einer Aufnahme automatisch gesammelt werden. Dadurch, dass pro Zeitfenster nicht nur ein Koeffizient extrahiert wird, sondern (bei Standardeinstellung) 13, ist die Verteilung, die entsteht, 13-dimensional. Die Modellierung mit GMM muss also entsprechend ebenfalls multivariat sein, d.h. alle 13 Dimensionen gleichzeitig modellieren. Es wird also in der automatischen Sprechererkennung mit VOCALISE bei Standardeinstellung eine multivariate Verteilung von 13 Cepstralkoeffizienten mit 32 Gaußverteilungen angenähert.

In der dritten Komponente eines Systems zur automatischen Sprechererkennung werden die akustischen Ähnlichkeitswerte (sog. scores) zwischen Tat- und Vergleichsmaterial bestimmt. Dabei werden von der Aufnahme des Vergleichssprechers die Cepstralkoeffizienten extrahiert und daraus ein GMM errechnet. Von der Aufnahme des Tatsprechers werden ebenfalls die Cepstralkoeffizienten extrahiert, daraus aber kein GMM errechnet, sondern die Cepstralkoeffizienten als Rohwerte belassen. Nun wird für jeden multivariaten Rohwert des Tatmaterials automatisch quantifiziert, wie gut dieser mit dem multivariaten GMM des Vergleichssprechers übereinstimmt. Parallel zu diesem Vergleich wird ein zweiter Vergleich durchgeführt, in dem jeder Rohwert des Tatmaterials mit einem sog. UBM (Universal Background Model) verglichen wird. Ein UBM ist ein GMM, bei dem nicht nur die Aufnahme eines einzelnen Sprechers verwendet wird, sondern die Aufnahmen einer ganzen Anzahl von Sprechern. Diese Menge von anderen Sprechern dient zur Quantifizierung der Möglichkeit, dass der Tatsprecher nicht mit dem Vergleichssprecher identisch ist, sondern mit einer anderen Person (gleichen Geschlechts). Für das UBM werden die Aufnahmen von mindestens ca. 20 Sprechern benötigt. Nun werden in einem letzten Schritt die Ergebnisse des jeweils ersten Vergleichs (Tatmaterial mit Vergleichsmaterial) durch die Ergebnisse des jeweils zweiten Vergleichs (Tatmaterial mit UBM) geteilt und über alle Rohwert-Einzelvergleiche gemittelt. So ergibt sich pro Vergleich zwischen einer Tataufnahme und einer Vergleichsaufnahme eine Zahl, die als likelihood score bezeichnet wird. Diese Zahl kann im weiteren Sinn als Index für den Grad der Ähnlichkeit zwischen Tat- und Vergleichsmaterial betrachtet werden.[2]

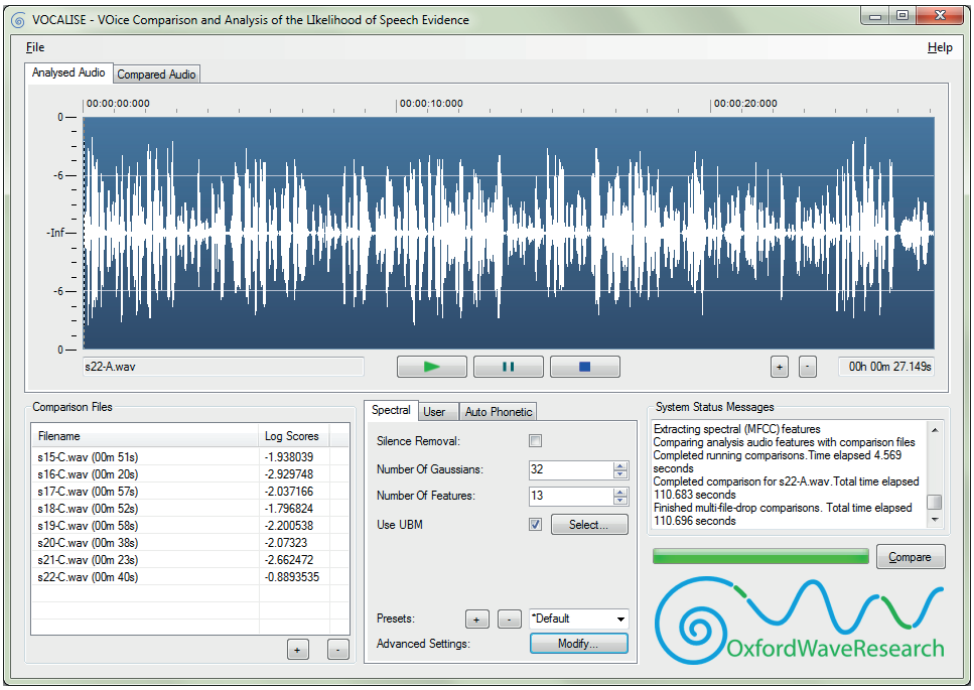
Ausführlichere Erläuterungen der automatischen Sprechererkennung werden u. a. von Alexander (2005), Reynolds & Campbell (2008) und Becker (2012) geliefert.

Die Funktionsweise der automatischen Sprechererkennung mit VOCALISE wird im Folgenden anhand von anonymisierten Fallbeispielen illustriert, in denen Deutsch gesprochen wurde. Gezeigt wird eine Systemevaluation, in der 22 Paare gleicher Sprecher und eine entsprechend hohe Anzahl verschiedener Sprecher miteinander verglichen wurden (näheres wird später erklärt). Die Aufnahmen stammen aus TKÜ-Maßnahmen (TKÜ bedeutet Telekommunikationsüberwachung) und haben eine Nettodauer (Dauer unter Abzug von Pausen und anderen sprachinaktiven Phasen des Sprechers) von mindestens 20 Sekunden. Sprechstilistisch und technisch sind die beiden Aufnahmen des jeweils gleichen Sprechers ähnlich. Eine solche Konstellation mit ähnlichen sprechstilistischen und technischen Bedingungen wird auch als matching conditions bezeichnet. Dies steht im Unterschied zu sog. mismatched conditions, in denen es deutliche Unterschiede in sprechstilistischer und/oder technischer Hinsicht gibt. Die Absprache zwischen den Autoren bestand darin, dass VOCALISE zunächst in der Lage sein sollte, unter Bedingungen von matching condition einsetzbar zu sein und dass die Nettosprachdauer ungefähr mindestens 20 Sekunden betragen sollte. Abbildung 1 zeigt die Benutzeroberfläche von VOCALISE, wobei 1a die Ansicht der Hauptseite zeigt und 1b die Ansicht des Untermenüs „Advanced Settings“.

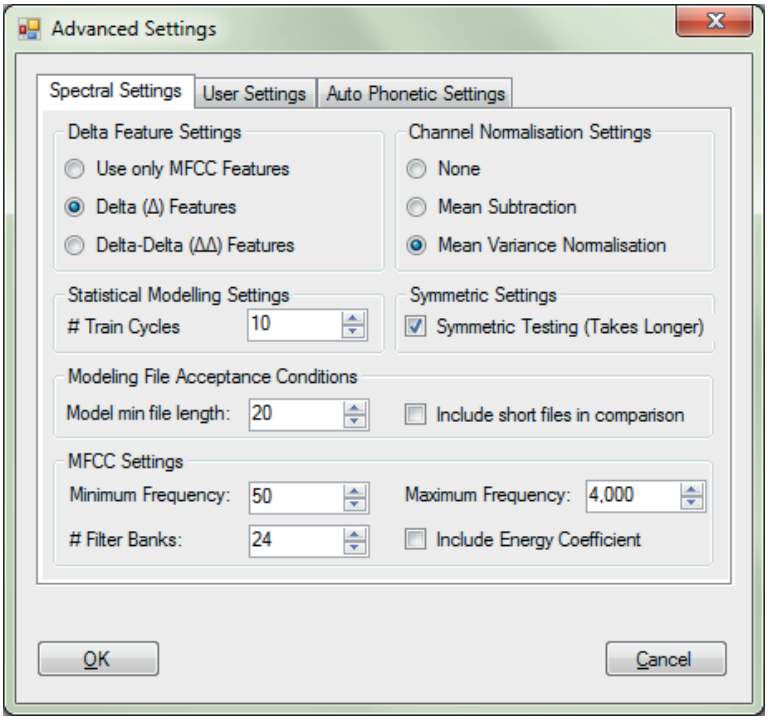
2 Genauer betrachtet handelt es sich vom Prinzip her um einen sog. Likelihood Ratio. Eine genaue Diskussion von Likelihood Ratios in forensischen Stimmenvergleichen kann hier aus Platzgründen nicht geführt werden. Genauere Informationen hierzu liefert u. a. Rose (2002).

Abbildung 1: Benutzeroberfläche von VOCALISE bei der automatischen Sprechererkennung. Teil (a) der Abbildung zeigt die Ansicht der Hauptseite von VOCALISE und Teil (b) die Ansicht des Untermenüs „Advanced Settings“.

(a)



(b)



Im oberen Teil der Hauptseite, abgebildet in Teil (a) von Abb. 1, ist das Zeitsignal einer Tataufnahme mit Namen s22-A zu sehen. Durch die Funktionsfelder unterhalb des Zeitsignals lässt sich die jeweilige Aufnahme beliebig vergrößern, navigieren und abspielen. Dies gilt sowohl für die Tataufnahmen, die in VOCALISE als Analysed Audio bezeichnet werden, als auch für die Vergleichsaufnahmen, die Compared Audio genannt werden. Unten links in Abb. 1a ist der Bereich zu sehen, in den beliebig viele Vergleichsaufnahmen durch einfaches drag&drop bewegt werden können. Bei Anklicken des Symbols „Compare“ (rechts, oberhalb des Firmenlogos) wird das Signal in Analyse Audio (hier im Bild) mit allen Signalen verglichen, die in den Comparison-Files-Bereich bewegt wurden. Zu jedem dieser Vergleiche zeigt das System nach Abschluss der Berechnungen einen (log-)likelihood score an (hier bezeichnet als Log Scores, rechts im Comparison Files-Bereich zu erkennen). In der Mitte befinden sich unter der Registerkarte „Spectral“ einige der Einstellungen für die automatische Sprechererkennung. (Die zweite Registerkarte „Formant“ betrifft die semiautomatische Sprechererkennung und wird im folgenden Abschnitt besprochen; die dritte Registerkarte „Auto Phonetic“ ist eine weitere Neuentwicklung in VOCALISE, die in diesem Artikel aus Platzgründen nicht besprochen wird.) Zu diesen Einstellungen, die von dem Benutzer frei verändert werden können, zählen die Anzahl der Gaußverteilungen (hier 32) und die Anzahl der Cepstralkoeffizienten (hier 13). In dem Feld „Use UBM“ kann angegeben werden, ob ein UBM verwendet werden soll (dies ist zu empfehlen, ist aber nicht zwingend erforderlich), und durch Betätigung von Select wird dann das Verzeichnis spezifiziert, in dem sich die für das UBM zu verwendeten Audiodateien befinden. Ist dieses Verzeichnis spezifiziert, beginnt VOCALISE, das GMM für das UBM zu berechnen, es sei denn ein solches UBM mit genau den gleichen Systemeinstellungen existiert bereits aus vorangehenden Durchgängen. Im Falle der hier beschriebenen Systemevaluation wurden für das UBM 25 Sprecher aus authentischen Telefonaten verwendet.[3] Das Fenster mit Namen „System Status Messages“ zeigt jeweils das Abarbeiten der einzelnen Analyseschritte des Systems an.

In Teil (b) von Abb. 1 ist der Aufbau des Untermenüs „Advanced Settings“ zu sehen. Hier werden weitere Details angezeigt, die der Benutzer einstellen kann. Die Kategorie Delta Feature Settings lässt eine Auswahl zu, ob nur die regulären Cepstralkoeffizienten (MFCC) verwendet oder ob auch die Delta bzw. Doppel-Delta-Werte verwendet werden sollen. Da die MFCC in kurzen Zeitabständen erhoben werden, entsteht gewissermaßen für jeden Cepstralkoeffizienten eine Zeitverlaufskurve. An dieser Zeitverlaufskurve können nun die erste Ableitung (Delta Features) oder auch noch die zweite Ableitung (Delta-Delta Features) berechnet werden. Diese Berechnungen stellen gegenüber den einfachen MFCC zusätzliche Information dar, da mit denen nicht nur die statischen spektralen Muster der Stimme erfasst werden, sondern auch die Veränderungen der spektralen Muster im Zeitverlauf. Dies hat allgemein in der Entwicklung der automatischen Sprechererkennung zu Verbesserungen geführt und steht in einem Zusammenhang mit dem Thema der Koartikulation und der Formantenbewegungen (Nolan 1983; McDougall 2006), es kann aber auch zu zusätzlichen Anfälligkeiten gegenüber Materialproblemen führen (insbesondere die Doppel-Delta-Merkmale), die für das Gesamtergebnis eher schädlich sind. Weil aber die meisten Systeme zur automatischen Sprechererkennung anhand von Material entwickelt wurden, das im Durchschnitt besser und homogener ist als forensisches Fallmaterial, sind die Optionen, die von den Systementwicklern verwendet werden, und die in der Regel unveränderbar in das System integriert sind, nicht zwangsläufig auch die besten Optionen für die forensische Fallarbeit. Deswegen ist es sinnvoll, wenn der Benutzer die Möglichkeit erhält, eigene Systemtests mit unterschiedlichen Einstellungen des Delta Feature Settings durchzuführen. Dieses Prinzip gilt ebenso auch für andere Systemeinstellungen und wird im letzten Abschnitt noch einmal diskutiert.

In den Channel Normalisation Settings kann eingestellt werden, ob eine Kompensation von Kanaleigenschaften durchgeführt werden soll und wenn ja, welche der beiden Optionen Mean Subtraction oder Mean Variance Normalisation durchgeführt werden soll. Auf diesen Aspekt der Kanalkompensation wurde oben bereits eingegangen. Wenn die Option Symmetric Testing (in Symmetric Settings) ausgewählt wird, bedeutet dies, dass in einem Einzelvergleich die Rolle von Tatmaterial und Vergleichsmaterial vertauscht wird. Das heißt, in einem Durchgang

3 Das Silence Removal-Werkzeug, auf das hier nicht weiter eingegangen wird, dient der automatischen Eliminierung von Pausen (welche nicht in die automatische Sprechererkennung einfließen sollten) und befindet sich noch in der Entwicklung. Derzeit werden Pausen im Rahmen der Extraktion der Nettosprachanteile, die auch aus unabhängigen Gründen erforderlich ist, manuell vom Experten durchgeführt. Bei den teilweise qualitativ eingeschränkten oder qualitativ wechselhaften Aufnahmen, die in der Forensik auftreten, ist ein automatisches Erkennen von Pausen bzw. komplementär dazu das Erkennen von Sprache eine sehr anspruchsvolle Aufgabe; die Fachbezeichnung hierzu lautet Voice Activity Detection. Sehr einfache Verfahren der automatischen Pausenerkennung, wie sie in einigen Sprechererkennungssystemen vorkommen, können bei forensischem Material unter Umständen problematisch sein.

wird die Datei in Analysed Audio als Tatmaterial und die in Compared Audio (also in der Comparison Files-Liste) als Vergleichsmaterial gehandhabt (d.h. von letzerem ein GMM berechnet, gegen das die Merkmalsvektoren von ersterem verglichen werden), so wie es regulär gemacht wird. In einem zweiten Durchgang wird dann die Datei in Analysed Audio als Vergleichsmaterial und die in Compared Audio als Tatmaterial behandelt. Zum Abschluss werden die Ergebnisse aus beiden Analysen zu einem Gesamtergebnis gemittelt. Dieses symmetrische Vorgehen hat teilweise zu Verbesserungen, teilweise auch zu Reduzierungen der Sprecherunterscheidungsleistung geführt; es war deshalb sinnvoll, es in VOCALISE als Option für eigene Tests zur Verfügung zu stellen.

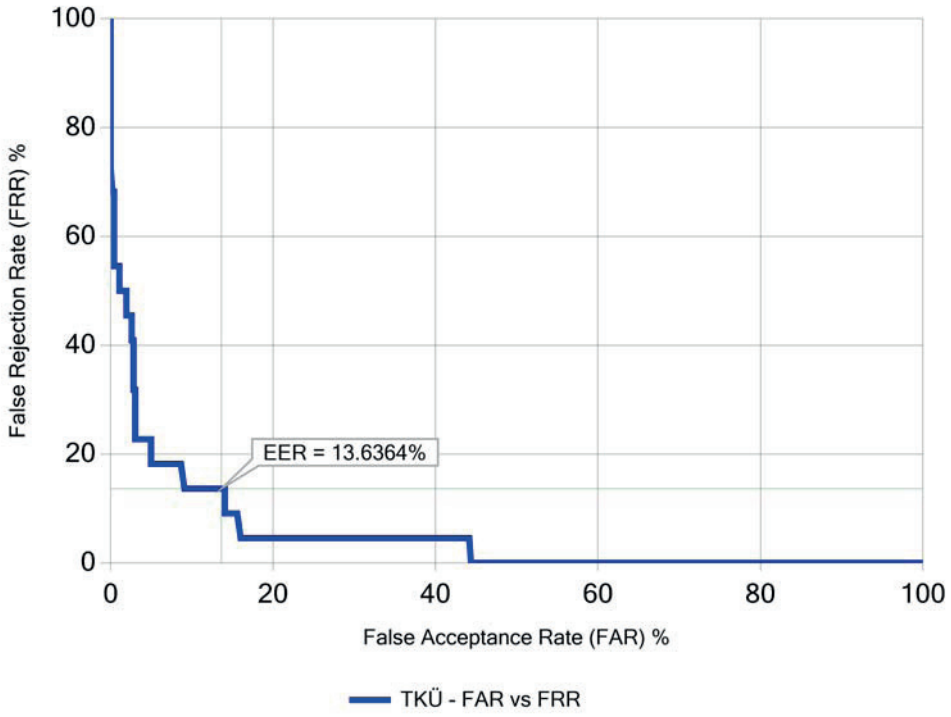
Die verbleibenden Optionen in Abb. 1b beziehen sich auf weitere spezifische technische Details. Die Anzahl der Train Cycles (in Statistical Modeling Settings) beziehen sich darauf, wie der sog. EM-Algorithmus (Expectation Maximisation) eingesetzt wird, mit dem die GMM berechnet werden. Mit jedem Berechnungszyklus wird die vorgegebene Anzahl an Gaußverteilungen *(32 in Abb. 1a)* genauer an die MFCC-Rohdaten angepasst. Bei ca. 10 Zyklen ist die Anpassung in der Regel gut genug, d.h. weitere Zyklen würden kaum eine Verbesserung bringen, aber mehr Zeit in Anspruch nehmen. Modeling File Acceptance Conditions erlaubt es dem Benutzer anzugeben, wie lang die zu modellierenden Dateien mindestens sein müssen; alle Dateien, die kürzer sind als der angegebene Wert, werden ignoriert (mit include short files in comparison kann diese Bedingung wieder ausgesetzt werden). Minimum Frequency und Maximum Frequency geben an, welcher Frequenzbereich analysiert werden soll. Es macht bei forensischen Stimmenvergleichen in der Regel keinen Sinn, mehr als 4000 Hz als Obergrenze zu wählen. Gegebenenfalls können weitere Einschränkungen des Frequenzgangs als der hier angegebene Bereich von 50 bis 4000 Hz zu Verbesserungen führen. Die Anzahl der Filter Banks gibt an, wie genau das Signal für die Gewinnung der Cepstralkoeffizienten gefiltert wird. Schließlich wird dem Benutzer noch die Option gegeben, den Energy Coefficient zu verwenden. Dies ist der niedrigste aller Cepstralkoeffizienten und enthält in der Regel keine oder kaum irgendwelche für die Unterscheidung von Stimmen relevanten Informationen, ist es aber wert, als Testoption freigestellt zu werden.

Mit den Parametereinstellung, die in Abb. 1 zu erkennen sind, wurde, wie erwähnt, ein Systemtest durchgeführt, an dem 22 Paare von jeweils gleichen Sprechern aus TKÜ-Aufnahmen beteiligt sind. Diese Vergleiche von Stimmen, von denen bekannt ist, dass sie vom gleichen Sprecher stammen (soweit dies im forensischen Kontext möglich ist), werden im Englischen als same-speaker comparisons bezeichnet. Diese Bezeichnung wird hier teilweise Englisch belassen und als same-speaker-Vergleiche bezeichnet. Wenn man 22 Paare gleicher Sprecher zur Verfügung hat, kann man neben der 22 same-speaker-Vergleiche auch eine große Anzahl von different-speaker-Vergleichen durchführen, also solche, bei denen die Nicht-Identität bekannt ist. Dies ergibt hier eine Anzahl von 462 different-speaker-Vergleichen, also 222 minus die 22 same-speaker-Vergleiche. Dass in Systemtests die Anzahl der different-speaker-Vergleiche deutlich größer ist als die der same-speaker-Vergleiche, ist der Normalfall und müsste andernfalls künstlich geändert werden, wenn es anders gewünscht wird (beispielweise, indem aus den vielen different-speaker-Vergleichen eine Zufallsauswahl oder Durchschnittsbildung mehrerer Vergleiche getroffen wird, wofür es aber kaum Argumente gibt).

Wie im Zusammenhang mit Abb. 1a erklärt wurde, gibt VOCALISE zu jedem Einzelvergleich einen Ähnlichkeitswert wieder, der als likelihood score bezeichnet wird und dessen Entstehen bereits erläutert wurde. In der vorliegenden Untersuchung berechnet VOCALISE den score von 484 Vergleichen, darunter 22 same-speaker und 462 different-speaker-Vergleiche. Damit diese scores im Gesamtzusammenhang einer Systemevaluation interpretiert werden können, können sie in das Programm Bio-Metrics von Oxford Wave Research eingelesen werden. Bio-Metrics ist vom Prinzip her unabhängig von VOCALISE, aber gut auf die Ausgabeformate von VOCALISE angepasst. Zwei Darstellungsmöglichkeiten von Bio-Metrics sollen hier vorgestellt werden – der DET-Plot und der Tippett-Plot *(siehe Abbildungen 2 und 3)*.

Abbildung 2 zeigt eine Darstellung, die in der Forschung zur Sprechererkennung bekannt ist, den DET-Plot (Detection Error Tradeoff). Der DET-Plot liefert eine Darstellung von verschiedenen Fehlerwahrscheinlichkeiten, die entstehen, wenn verschiedene scores als Schwellwerte für eine Entscheidung gewählt werden, ob ein gegebenes Paar von Stimmen vom gleichen Sprecher oder von verschiedenen Sprechern produziert wurde. Auf die genaue Herleitung und Interpretation von DET-Plots soll in diesem Zusammenhang nicht eingegangen werden (siehe van Leeuwen et al. 2007 für weitere Details). Eine wichtige Kennzahl, die im Rahmen von DET-Plots dargestellt werden kann, ist die Gleichfehlerrate (Englisch: Equal

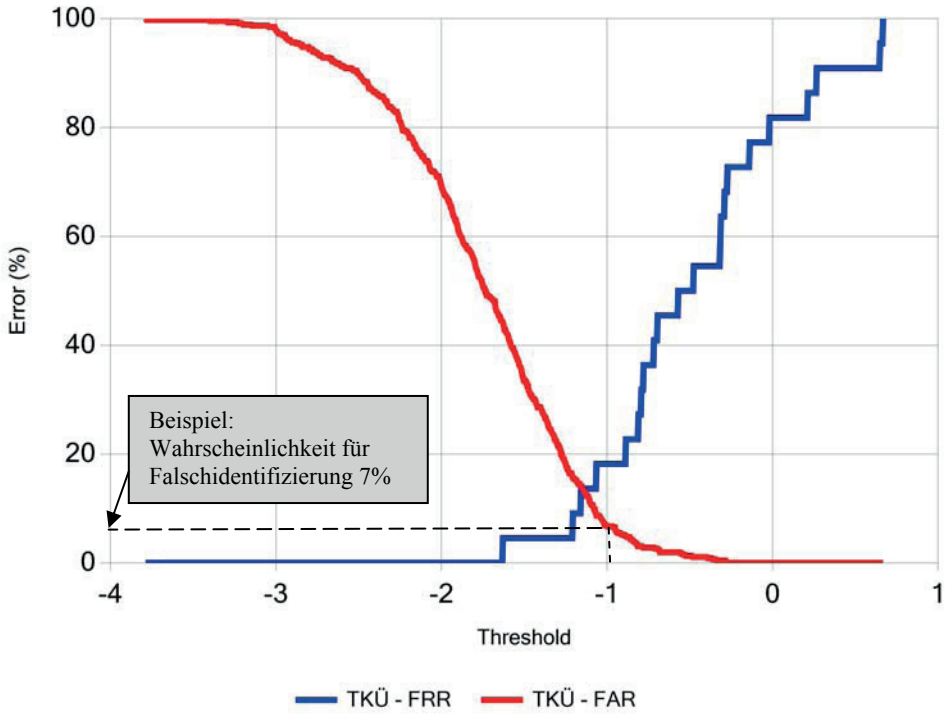
Abbildung 2: DET-Plot für eine Systemevaluation bei automatischer Sprechererkennung anhand von TKÜ-Daten. Horizontale Achse: Fehlerwahrscheinlichkeit einer falschen Identifizierung; Vertikale Achse: Fehlerwahrscheinlichkeit einer falschen Zurückweisung. Die Gleichfehlerrate (EER) wird grafisch angezeigt und beträgt 13,6 %.



Error Rate, abgekürzt als EER). Die Gleichfehlerrate gibt die Fehlerwahrscheinlichkeit an, die für beide Fehlerarten gleich groß ist. Im Fall der hier durchgeführten Systemevaluation beträgt die Gleichfehlerrate 13,6 %, d.h. mit der hier durchgeführten Methode der automatischen Sprechererkennung, der gewählten Parameter *(siehe Abb. 1)* und der verwendeten Datengrundlage, besteht die Wahrscheinlichkeit für eine falsche Identifizierung (d.h. different-speaker-Vergleiche werden für same-speaker-Vergleiche gehalten) 13,6 %, und ebenso beträgt die Wahrscheinlichkeit für eine falsche Zurückweisung (d.h. same-speaker-Vergleiche werden für different-speaker-Vergleiche gehalten) 13,6 %. Hierbei ist daran zu erinnern, dass dies die alleinigen Fehlerwahrscheinlichkeiten für die automatische Sprechererkennung sind. In forensischen Stimmenvergleichen werden aber auch andere Merkmale aus dem Bereich der auditiv-akustischen Methode verwendet, so dass die endgültigen Fehlerwahrscheinlichkeiten in Wirklichkeit geringer sind.

Abbildung 3 zeigt eine weitere bekannte Darstellung, den Tippett-Plot (Tippett ist der Name eines englischen Statistikers). Ein Tippett-Plot zeigt die kumulierten Wahrscheinlichkeitsdichtefunktionen von zwei Verteilungen. Zum einen ist dies die Verteilung der scores für alle different-speaker-Vergleiche; dies ist die Kurve, die von links nach rechts abfällt (typischerweise in rot dargestellt). Zum anderen ist dies die Verteilung der scores aller same-speaker-Vergleiche; dies ist die Kurve, die von links nach rechts ansteigt (typischerweise in blau dargestellt). Der Tippett-Plot in Abb. 3 zeigt, dass die different-speaker-Vergleiche insgesamt geringere scores aufweisen als die same-speaker-Vergleiche. Dies entspricht der Erwartung, denn wie oben erklärt wurde, können die scores als Maß für die Ähnlichkeit von Stimmproben betrachtet werden und bei jeweils gleichen Sprechern würde man insgesamt eine größere Ähnlichkeit erwarten als bei jeweils verschiedenen Sprechern. Der Tippett-Plot zeigt aber auch, dass die Trennung von Paaren gleicher und verschiedener Sprecher nicht perfekt ist. Beispielsweise gibt

Abbildung 3: Tippett-Plot für eine Systemevaluation bei automatischer Sprechererkennung anhand von TKÜ-Daten: Horizontale Achse: Sprecherähnlichkeiten (log10-likelihood scores); Vertikale Achse: kumulierte Fehlerwahrscheinlichkeit.



es einzelne different-speaker-Paare mit einem score, der so groß ist, dass er eher für die Ähnlichkeiten zwischen gleichen Sprechern typisch ist. Dies gilt z. B. für das different-speaker-Paar, das einen score von -1 hat *(siehe Abb. 3)*. Umgekehrt gibt es einzelne same-speaker-Paare mit einem score, der so klein ist, dass er eher für die Ähnlichkeiten zwischen verschiedenen Sprechern typisch ist. Problematisch sind also vor allem solche same-speaker-Paare, die links vom Überschneidungspunkt der beiden Kurven liegen, sowie solche different-speaker-Paare, die rechts vom Überschneidungspunkt liegen. Übrigens zeigt der Überschneidungspunkt die Gleichfehlerrate an, d.h. auf der vertikalen Achse liegt der Überschneidungspunkt bei ca. 13 %.

Tippett-Plots spielen eine große Rolle bei der Interpretation der Ergebnisse in konkreten einzelnen forensischen Stimmenvergleichen. Ein solcher Einzelvergleich zwischen Tatmaterial und Vergleichsmaterial ergibt einen score. Angenommen ein solcher score beträgt -1, dann spricht dieser score auf Grundlage des Tippett-Plots in Abb. 3 eher dafür, dass es sich um den gleichen Sprecher handelt, als dass es sich um verschiedene Sprecher handelt (Wert liegt rechts vom Schnittpunkt). Dennoch verbleibt eine gewisse, wenn auch eher geringe Fehlerwahrscheinlichkeit, dass es sich dennoch um verschiedene Sprecher handelt. Diese Wahrscheinlichkeit einer Falschidentifizierung kann anhand des Tippett-Plots jetzt beziffert werden; sie beträgt ca. 7 %. Dies ist der Wert auf der Vertikalachse für den Schnittpunkt zwischen der Vertikallinie bei -1 und der different-speaker-Verteilung (siehe die gestrichelte Linie in Abb. 3 und das graue Kästchen). Umgekehrt, wenn der score beispielsweise -1,6 beträgt, spricht dieser eher gegen Identität (Wert liegt links vom Schnittpunkt). Die Wahrscheinlichkeit, dass es sich dann dennoch um den gleichen Sprecher handelt, beträgt ca. 5 %.

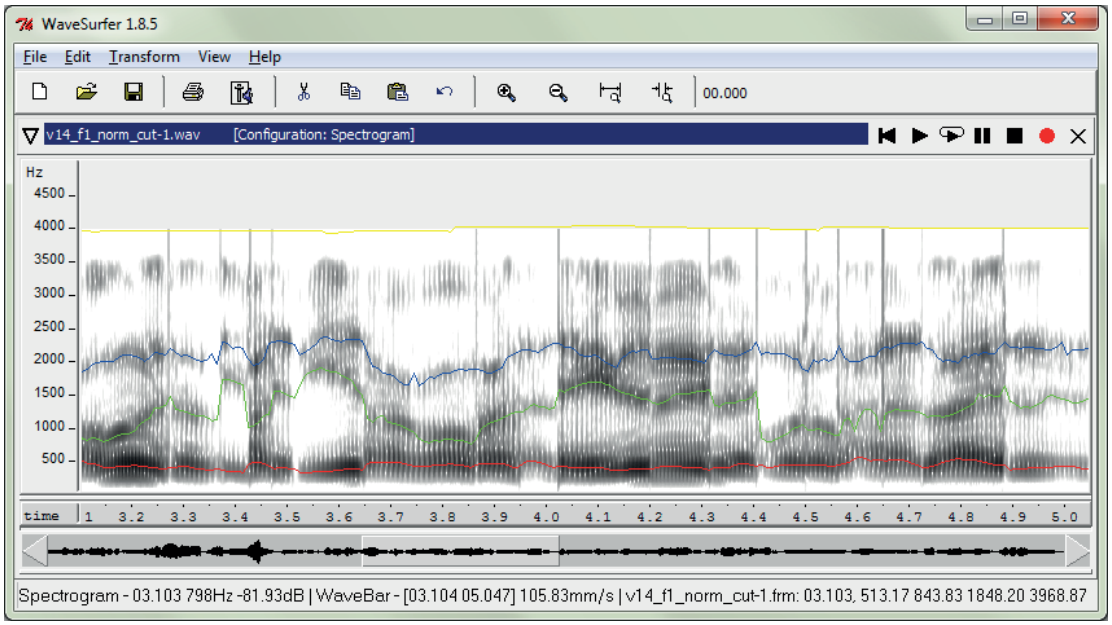
4. Semiautomatische Sprechererkennung: Langzeitformanten

Während im vorangehenden Abschnitt die automatische Sprechererkennung erklärt wurde, wird in diesem Abschnitt darauf eingegangen, wie in VOCALISE semiautomatische Sprechererkennung durchgeführt werden kann. Die semiautomatische Sprechererkennung wird hier anhand der Langzeitformantenanalyse vorgestellt. Diese wird im Folgenden als LTF-Analyse bezeichnet, wobei LTF für Engl. Long-Term Formants steht. Diese Methode wurde erstmals von Nolan & Grigoras (2005) vorgestellt. Weitere Untersuchungen mit dieser Methode wurden u. a. von Moos (2010) durchgeführt.

Bei der LTF-Analyse werden aus einem Signal nur solche Anteile extrahiert, in denen es sich um Vokale handelt und in denen die Formanten F1, F2 und F3 gut zu erkennen sind. Bei den Formanten handelt es sich um Resonanzfrequenzen, die sich aus der räumlichen Beschaffenheit des Vokaltraktes ergeben, d.h. der anatomischen Strukturen im Mundraum vom Kehlkopf bis zur Mundöffnung. Die extrahierten Anteile werden zu einer neuen Audiodatei zusammengefügt. Anschließend wird über diese Audiodatei eine automatische Formantenextraktion appliziert. Im Englischen wird diese als formant tracking bezeichnet. Formantenextraktionsverfahren sind in der Regel sehr leistungsfähig, aber insbesondere bei Aufnahmen mit geringer Qualität, wie sie in der Forensik teilweise auftreten, sowie bei einigen Stimmen, in denen einige Formanten nur schwach ausgeprägt sind, produzieren diese Verfahren Fehler, die durch den Phonetiker korrigiert werden müssen. Abb. 4 zeigt einen Ausschnitt einer solchen extrahierten Audiodatei zusammen mit den korrigierten Spuren der Formanten F1, F2 und F3.

Von dem formant tracking-Verfahren wird alle 10 Millisekunden eine Extraktion der numerischen Werte für die Formantenfrequenzen F1, F2 und F3 durchgeführt.

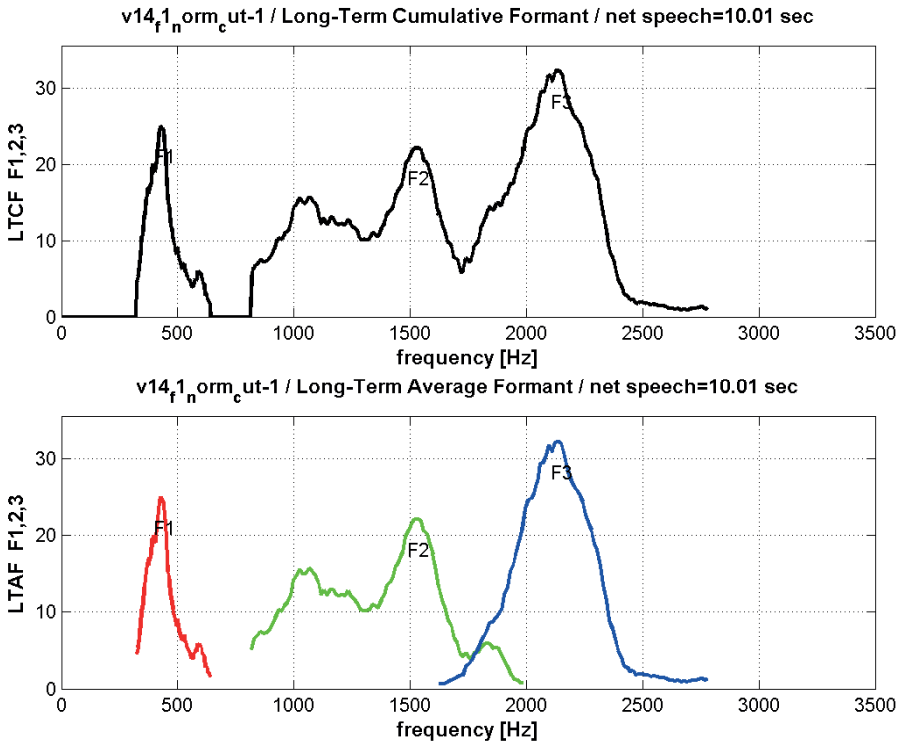
Abbildung 4: Illustration der LTF-Analyse. Der Hauptteil der Abbildung zeigt ein Spektrogramm, in dem die extrahierten Spuren des ersten Formanten F1 (unterste Spur; rot in Farbe), des zweiten Formanten F2 (mittlere Spur; grün in Farbe) und des dritten Formanten F3 (oberste Spur; blau in Farbe) dargestellt sind. In einem Spektrogramm zeigt die horizontale Achse den Zeitverlauf an (in Sekunden), während die vertikale Achse die Frequenz (in Hz) anzeigt. Die Amplitude wird durch den Schwärzungsgrad angegeben. Dabei zeigen sich die Formanten in Form von dunklen, horizontal verlaufenden Bändern. Verwendet für die LTF-Analyse wurde das Programm Wavesurfer, das frei im Internet zur Verfügung steht.



Außerdem werden die Formantenbandbreiten extrahiert. Formantenbandbreiten sind ein Maß dafür, wie schmal (geringe Bandbreite) oder breit (große Bandbreite) die Formanten sind. Sprecher können sich sowohl in den Formantenfrequenzen als auch den Formantenbandbreiten unterscheiden, allerdings ist die Extraktion der Bandbreiten insbesondere bei qualitativ vermindertem Material wahrscheinlich unzuverlässiger als die der Formantenfrequenzen (zu diesem Thema ist noch weitere Forschung erforderlich). Die von dem formant tracking-Verfahren (und manuell korrigierte) bereitgestellte Information wird über den Gesamtverlauf der extrahierten Audiodatei gesammelt. Diese gesammelte Information kann beispielsweise in Form eines Histogramms dargestellt werden *(Abb. 5)*.

Abb. 5 zeigt, dass die Verteilungen der Formantenfrequenzen teilweise so komplex sind, dass man sie nicht hinreichend genau mit einzelnen Gauß-Verteilungen annähern könnte. Dies gilt in Abb. 5 vor allem für den zweiten Formanten, der dort eine dreigipflige Verteilung annimmt, aber auch die Verteilungen der anderen Formanten sind nicht gänzlich regulär im Sinne einer Normalverteilung. Als Methode für die Modellierung solcher komplexen Verteilungen würde sich das Gaussian Mixture Modeling (GMM) besonders anbieten, das ebenfalls in der automatischen Sprechererkennung zum Einsatz kommt. Erste Untersuchungen von Langzeitformanten mit GMM sind von Becker et al. (2008) präsentiert worden. Ein wesentliches Design-Merkmal von VOCALISE besteht darin, dass mit diesem Programm alle Arten

Abbildung 5: Verteilung der Formantenfrequenzen F1, F2 und F2 in der Aufnahme, die in Auszügen in Abb. 4 dargestellt wurde. Die horizontale Achse zeigt die Frequenzwerte der Formanten (in Hz), die Vertikalachse die relativen Häufigkeiten der Formentenfrequenzen (mit entsprechender Glättung). In der oberen Darstellung sind alle Formanten zusammengefasst. In der unteren (etwas nützlicheren) Darstellung werden die Verteilungen der Formanten separat für F1, F2 und F3 dargestellt. Die Formantenbandbreiten werden hier nicht dargestellt. Die Software zur Erstellung dieser Grafiken wurde von Catalin Grigoras geschrieben.



von stimmlichen Eingabemerkmalen mit GMM modelliert und entsprechend weiterverarbeitet werden können, somit auch die Formantenwerte, die während der LTF-Analyse entstehen.

Zur Darstellung der Möglichkeiten, die sich aus der Verwendung der semiautomatischen Sprechererkennung anhand von Langzeitformanten ergeben, wurde ein Experiment anhand von Daten aus dem „Pool 2010" durchgeführt.[4] Beim Pool 2010 handelt es sich um ein Korpus von 100 männlichen erwachsenen Sprechern, die unter verschiedenen Bedingungen gesprochen haben (Jessen et al. 2005). In dem hier gezeigten Experiment wurden die Daten von 22 Sprechern verwendet. Für das Tatmaterial wurde Sprache aus einer Bedingung gewählt, in der die Versuchspersonen ihre Eindrücke und Gedanken über den Versuchsablauf mitgeteilt hatten. Für das Vergleichsmaterial wurde Sprache aus einer Bedingung gewählt, in der die Versuchspersonen einem Gesprächspartner Bilder zu beschreiben hatten, aber dabei bestimmte Wörter zu vermeiden hatten. Die Dauer der extrahierten Audiodateien (Vokale mit gut erkennbarer Formantenstruktur) betrugen jeweils für das Tatmaterial ca. 10 Sekunden. Für das Vergleichsmaterial stand eine längere Dauer zur Verfügung, so dass die extrahierten Audiodateien dort jeweils in zwei Teile aufgeteilt wurden und dadurch die doppelte Menge an Vergleichen mit dem Tatmaterial möglich wurde (dies führt zu einer genaueren Struktur der same-speaker-Verteilung, da jetzt 44, statt 22 same-speaker-Vergleiche möglich waren). Die Dauer dieser einzelnen Teile des Vergleichsmaterials reichte von ca. 10 Sekunden bis maximal ca. 40 Sekunden reiner vokalischer Anteile. Als UBM wurden vergleichbar lange extrahierte Audiodateien von 22 anderen Sprechern verwendet. Nachdem für alle diese Dateien (Tatmaterial, Vergleichsmaterial und UBM) die korrigierten Formantenwerte extrahiert wurden, wurden diese Werte mit VOCALISE analysiert. Abbildung 6 zeigt die Benutzersettings, die dabei verwendet wurden.

Der Aufbau der Oberfläche für semiautomatische Sprechererkennung (Reiter „User") ist in den meisten Aspekten identisch mit dem Aufbau der Oberfläche für automatische Sprechererkennung (Reiter „Spectral"). Wie schon erwähnt, war es einer der wesentlichen Intentionen der Entwickler, die automatische und semiautomatische Sprechererkennung in ihren methodischen Abläufen möglichst parallel zu gestalten. Ein Unterschied besteht in dem Fenster in Abb. 6a, in dem verschiedene Spalten (columns) ausgewählt werden können. Diese Spalten beziehen sich auf das benutzerdefinierte Eingabeformat der importierten Dateien. In diesem Fall, in dem es um Langzeitformanten geht, beziehen sich die in Abb. 6a ausgewählten Spalten 1 bis 3 auf die Formantenfrequenzen F1, F2 und F3.

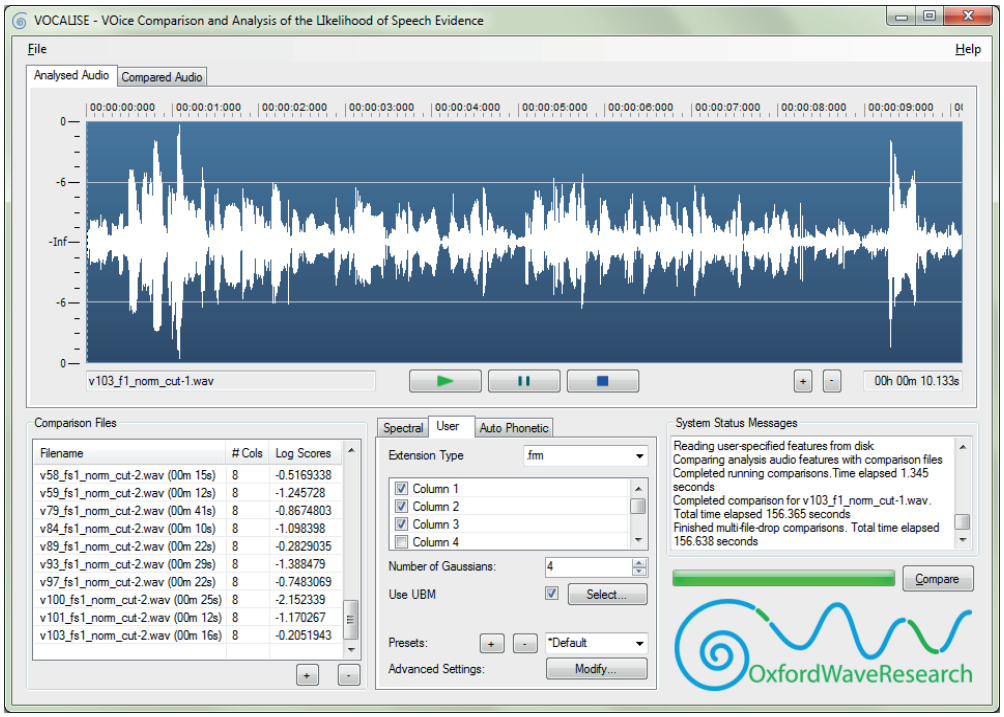
In Abb. 6b ist zu erkennen, dass die Anzahl der hier enthaltenen Optionen gegenüber denen in der automatischen Sprechererkennung (Abb. 1b) reduziert ist. Die meisten Reduktionen sind prinzipieller Natur. Beispielsweise ist es nicht sinnvoll, obere und untere Frequenzgrenzen einzugeben, da die Formantenfrequenzen bereits durch die LTF-Analyse extrahiert wurden. Lediglich die Modeling File Acceptance conditions sind nicht prinzipiell motiviert, sondern könnten in einer zukünftigen Version des Programms hinzugefügt werden.

Auf Grundlage der erwähnen Daten von 22 Sprechern und mit den Settings, die in Abb. 6 gezeigt werden, wurden Systemtest durchgeführt und die Resultate mit Bio-Metrics errechnet und dargestellt. Dabei wurde eine ganze Serie von insgesamt 22 Systemtests durchgeführt (diese Zahl ist nur zufällig die gleiche wie die Zahl der untersuchten Sprecher, die ebenfalls 22 beträgt). In 11 dieser Tests wurde als Datengrundlage die Langzeitformanten F1, F2 und F3 verwendet. In weiteren 11 Tests wurden neben F1, F2 und F3 außerdem noch die Bandbreiten dieser Formanten verwendet, welche als B1, B2 und B3 bezeichnet werden. Innerhalb jeder dieser zwei Gruppen von 11 Tests wurde systematisch die Anzahl der Gaußverteilungen variiert, angefangen von einer Gaußverteilung bis hin zu einem Wert von 20 Gaußverteilungen (wobei von 10 auf 20 ein Sprung gemacht wurde). Wie oben im Zusammenhang mit Abb. 5 erläutert wurde, ist die Voraussage die, dass eine einzelne Gaußverteilung pro Formant zu wenig ist, um die Verteilungen der Langzeitformanten optimal für die Zwecke der Sprechererkennung zu modellieren, so dass die Sprechererkennungsleistung wahrscheinlich schlechter ist, wenn nur eine Gaußverteilung verwendet wird, als wenn mehrere verwendet werden. Becker et al. (2009) hatten sich mit dieser Frage beschäftigt und gefunden, dass 8 Gaußverteilungen optimal sind. Weiterhin ist zumindest bei qualitativ gutem Material, wie es hier verwendet wurde, zu erwarten, dass die Hinzufügung der Bandbreiten eine bessere Leistung ergibt, als wenn nur die Formantenfrequenzen verwendet werden; dies berichteten Becker et al. (2008).

4 Dem Bundeskriminalamt sei für die Überlassung von Aufnahmen aus dem Pool 2010-Korpus und für weitere Hinweise zum Umgang mit den Aufnahmen gedankt.

Abbildung 6: Benutzeroberfläche von VOCALISE bei der semiautomatischen Sprechererkennung. Teil (a) der Abbildung zeigt die Ansicht der Hauptseite von VOCALISE und Teil (b) die Ansicht des Untermenüs „Advanced Settings“..

(a)



(b)

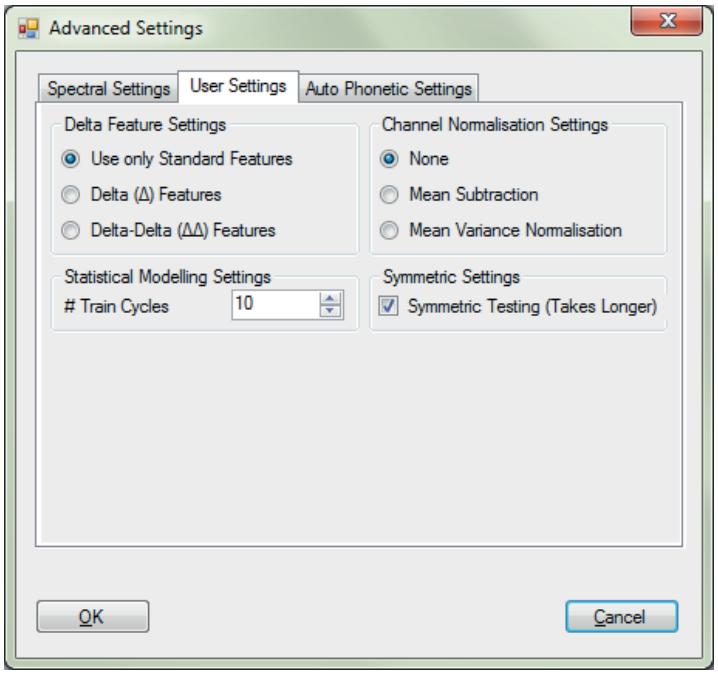


Abbildung 7: DET-Plots für drei Systemevaluationen bei Verwendung semiautomatischer Sprechererkennung anhand des Pool 2010. Horizontale Achse: Fehlerwahrscheinlichkeit einer falschen Identifizierung; Vertikale Achse: Fehlerwahrscheinlichkeit einer falschen Zurückweisung. Separate Kurven für die Versuchsbedingung mit einer Gaußverteilung (durchgezogene Linie), mit 4 Gaußverteilungen (gepunktet) und mit 20 Gaußverteilungen (Strichpunkte).

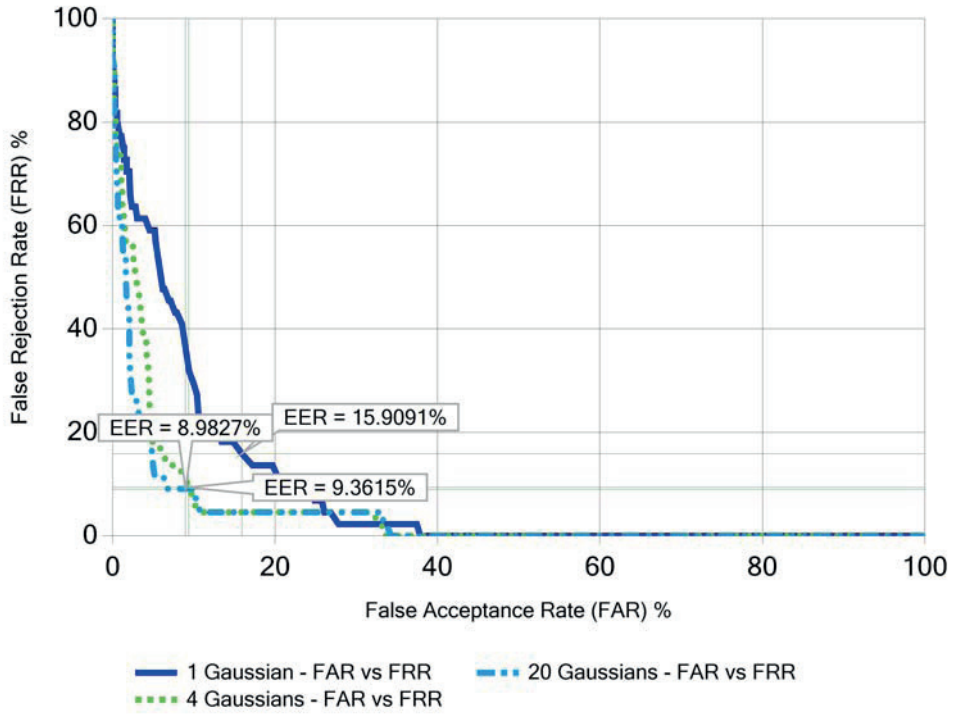


Abbildung 7 zeigt einen Teil der Resultate in Form eines DET-Plots bei drei der 22 Systemtests. Konkret sind dies diejenigen Tests, in denen mit den Formanten F1, F2 und F3 ohne die Bandbreiten gearbeitet wurde und in denen mit einer, vier und zwanzig Gaußverteilungen gearbeitet wurde.

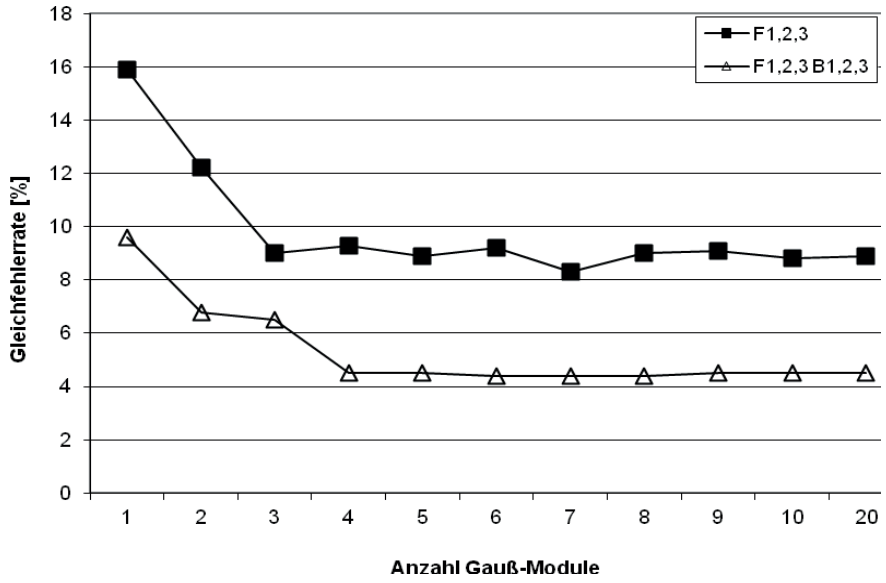
In Abb. 7 ist zu erkennen, dass insgesamt die Fehlerwahrscheinlichkeiten größer sind, wenn nur eine Gaußverteilung verwendet wird, als wenn mit einer der beiden größeren Anzahlen von Gaußverteilungen gearbeitet wird. Dieses Ergebnis bestätigt die oben diskutierte Voraussage, dass mit nur einer Gaußverteilung nicht das Maximum an möglicher Sprecherinformation erfasst wird.

Um das Gesamtbild von allen 22 Versuchen zu erhalten, werden in Abb. 8 die Gleichfehlerraten sämtlicher Versuche dargestellt.

Abb. 8 zeigt zurächst, dass die Gleichfehlerraten geringer sind, d.h. die Sprechererkennungsleistung besser ist, wenn die Frequenzen und die Bandbreiten der Formanten verwendet werden, als wenn nur die Frequenzen der Formanten zugrundegelegt werden. Dies entspricht der erwähnten Erwartung bei qualitativ gutem Material und bestätigt die Ergebnisse von Becker et al. (2008).

Zweites ist in Abb. 8 zu erkennen, dass die Gleichfehlerrate abnimmt, d.h. die Erkennungsleistung besser wird, wenn ausgehend von nur einer Gaußverteilung (linke Seite der Grafik) die Anzahl der Gaußverteilungen erhöht wird (von links nach rechts in der Grafik). Allerdings ist zu erkennen, dass bereits bei 4 Gaußverteilungen praktisch ein Plateau erreicht ist, d.h. bei mehr als 4 Gaußverteilungen ändert sich die Erkennungsleistung nicht mehr in relevanter Weise. Dieses Muster ist in den Versuchen, in denen nur mit den Frequenzen F1, F2 und F3 gearbeitet wurde, ähnlich zu denen, in denen außer-

Abbildung 8: Gleichfehlerraten in Prozent (Vertikalachse) bei der Modellierung von Langzeitformanten mit unterschiedlichen Anzahlen von Gaußverteilungen, auch Gauß-Modulen genannt (Horizontalachse). Gefüllte Quadrate: Ergebnisse bei Verwendung der Formantenfrequenzen F1, F2 und F3; ungefüllte Dreiecke: Ergebnisse bei Verwendung der Formantenfrequenzen F1, F2 und F3 sowie der Formantenbandbreiten B1, B2 und B3.



dem mit den Bandbreiten B1, B2, B3 gearbeitet wurde. Im ersten Fall (den alleinigen Formantenfrequenzen) ist das Plateau allerdings schon bei nur drei Gaußverteilungen erreicht. Diese Ergebnisse unterscheiden sich von denen bei Becker et al. (2009), wo erst bei 8 Gaußverteilungen das Maximum der Erkennungsleistung erreicht wurde.

Insgesamt zeigen diese Versuche zur semiautomatischen Sprechererkennung mit Langzeitformanten, dass sich die Absicht hinter der Entwicklung von VOCALISE, Methodiken, die in der automatischen Sprechererkennung bekannt sind, auch auf die semiautomatische Sprechererkennung auszuweiten, als sinnvoll erwiesen hat. Es wurde hier untersucht, ob das Gaussian Mixture Modeling aus der automatischen Sprechererkennung auch für die semiautomatische Sprechererkennung sinnvoll ist und konkret wie viele Gaußverteilungen verwendet werden sollten, wenn mit Langzeitformanten als semiautomatisches Merkmal gearbeitet wird. Im Sinne des Prinzips „so wenig wie möglich, so viel wie nötig“, ergab sich, dass je nach Merkmalskombination drei bis vier Gaußverteilungen ausreichen, um die maximale Sprecherinformation, die sich in den Langzeitformanten befindet, zu erfassen. Dieses „so wenig wie möglich, so viel wie nötig“-Prinzip ist deshalb sinnvoll, als dass eine Modellierung mit mehr Gaußverteilungen als erforderlich das Risiko in sich birgt, die GMM-Modelle zu sehr an spezifische Versuchskonstellationen anzupassen und die Generalisierbarkeit zu gefährden. Dass mit nur sehr wenigen Gaußverteilungen bereits das maximale Ergebnis erzielt wird, ist deshalb eine wichtige Erkenntnis.

5. Zusammenfassung, Schlussfolgerung und praktische Aspekte

In diesem Artikel wurde ein neues Programm namens VOCALISE vorgestellt. Dieses Programm, das von den Autoren entwickelt wurde, stellt eine Plattform dar, auf der zwei Verfahren zur forensischen Stimmenvergleichsanalyse durchgeführt werden können: die automatische Sprechererkennung und die semiautomatische Sprechererkennung.

Die Berücksichtigung automatischer Sprechererkennung entspricht dem aktuellen Stand der Forschung und Praxis in der forensischen Sprechererkennung. Sie wird seit ca. sechs Jahren im BKA und seit fast zwei Jahren in den deutschen LKÄ (Landeskriminalämtern) eingesetzt und findet auch in einigen internationalen forensischen Instituten und Praxen Verwendung. Automatische Sprechererkennung kann nur eingesetzt werden, wenn die Anforderungen an die Dauer und Qualität des Untersuchungsmaterials erfüllt sind. Wie diese Akzeptabilitätskriterien genau beschaffen sind, muss anhand von Systemtests ermittelt werden. In Abschnitt 3 wurde ein solcher Systemtest vorgestellt (siehe Abb. 2 und 3). Dort wurde beispielsweise die Voraussetzung an das Material gestellt, dass dieses aus natürlichen Telefonkommunikationssituationen stammt und dass es pro Aufnahme eine Mindest-Nettodauer von 20 Sekunden hat. Solche Systemtests liefern nicht nur Erkenntnisse über die Systemleistung allgemein (in Form der Gleichfehlerrate, die dort 13 % betrug), sondern auch über die Fehlerwahrscheinlichkeiten, die in einem konkreten Stimmenvergleich zu erwarten sind. Beispielsweise wurde anhand Abb. 3 gezeigt, dass bei einem einzelnen Stimmenvergleich (d.h. der Gegenüberstellung einer Tat- mit einer Vergleichsaufnahme), der einen Ähnlichkeitswert (score) von -1 ergibt, dieser Wert für Sprecheridentität spricht, allerdings mit einer Wahrscheinlichkeit von 7 % auch bei Nicht-Identität entsteht (diese Ergebnisse gelten wie gesagt vor dem Hintergrund dieses konkreten Systemtests). Die tatsächliche Fehlerwahrscheinlichkeit ist in der Regel noch deutlich geringer, weil eine Stimmenvergleichsanalyse keinesfalls nur auf automatischer Sprechererkennung basieren darf. Zahlreiche andere Merkmale und Methoden aus dem eingangs erwähnten auditiv-akustischen Ansatz kommen hinzu.

Hierbei haben verschiedene Methoden und Merkmale jeweils unterschiedliche Stärken und Schwächen und insgesamt ergänzen sie sich gegenseitig. Methoden aus der auditiven und akustischen Phonetik haben den Vorteil, dass über sie sehr viel publiziertes Wissen und theoretische Modelle existieren, beispielsweise über den Zusammenhang zwischen der Anatomie des Vokaltraktes und den Formantenfrequenzen. Die automatische Sprechererkennung ist in dieser Hinsicht stärker datengeleitet als theoriegeleitet und nicht alle Vergleichsergebnisse sind in der gleichen Weise erklärbar wie bei phonetischen (oder linguistischen) Merkmalen. Andererseits zeichnet sich die automatische Sprechererkennung durch ein sehr hohes Maß an Objektivität aus. Wenn beispielsweise ein Stimmenvergleich repliziert wird und dabei exakt die gleiche Datengrundlage und die gleichen Einstellungsparameter verwendet werden, ist auch das Ergebnis genau das gleiche. Ein weiterer Vorteil der automatischen Sprechererkennung, ist – wie der Name sagt – dessen Automatisierbarkeit und damit einhergehend die Fähigkeit sehr viele Aufnahmen zu bearbeiten. Gerade im Zusammenhang mit großen Strafprozessen, in denen TKÜ-Aufnahmen als Beweismittel relevant werden, treten oft größere Anzahlen von Aufnahmen auf. Bei der Bearbeitung solcher Fallkomplexe können automatische Verfahren eine wichtige Rolle spielen. Dabei ist bei der Bearbeitung solcher Komplexe genau zu planen, wie die automatischen Analysen mit den auditiv-akustischen zusammenwirken.

Die semiautomatische Sprechererkennung ist im Gegensatz zur automatischen derzeit nur sehr wenig verbreitet bzw. dort, wo sie seit längerem praktiziert wird, ist sie meist nur sehr geringfügig dokumentiert. Damit diese Methode weitere Akzeptanz findet, muss sie vor allem besser dokumentiert und wissenschaftlich evaluiert werden. Die Möglichkeiten, die VOCALISE bietet, gehen genau in diese Richtung. Wie an einem Beispiel gezeigt wurde, lassen sich mit dem Programm Experimente durchführen, in denen relevante Parameter systematisch variiert und auf ihre Auswirkung hin untersucht werden. Dabei wird in vielerlei Hinsicht Neuland betreten, wie hier in der Frage nach der erforderlichen Anzahl von Gaußmodulen bei der Langzeitformantenanalyse. Auf diese Weise wird die semiautomatische Methode für ihren Einsatz in der forensischen Sprechererkennung vorbereitet.

Wichtig in diesem Prozess der Forschung und Entwicklung ist die Möglichkeit, verschiedene Parameter, die einen Einfluss auf das Ergebnis haben können, selbst einstellen zu können. Diese Möglichkeit bereitzustellen, war von Anfang an Teil der Konzeption von VOCALISE. Eine solche Einflussnahme auf die Parametereinstellungen ist nicht nur relevant für die semiautomatische, sondern auch für die automatische Sprechererkennung. Denn bei Systemen, in denen alle Parameter fest vom Entwickler eingestellt sind, besteht das Risiko, dass die Datengrundlage, die der Entwicklung zugrunde lag, nicht repräsentativ ist für die forensischen Daten, mit denen der Anwender zu arbeiten hat. Insofern ist eine flexible Lösung, bei der die Parametereinstellungen verändert und anschließend anhand von forensisch „maßgeschneiderten“ Systemevaluationen getestet werden können, sinnvoller.

6. Literatur

Alexander, Anil (2005): Forensic Automatic Speaker Recognition Using Bayesian Interpretation and Statistical Compensation For Mismatched Conditions. Ph.D.-Dissertation, École Polytechnique Fédérale de Lausanne.

Baldwin, John & Peter French (1990): Forensic Phonetics. London: Pinter.

Becker, Timo (2012): Automatischer forensischer Stimmenvergleich. Norderstedt: Books on Demand.

Becker, Timo, Michael Jessen & Catalin Grigoras (2008): Forensic speaker verification using formant features and Gaussian mixture models. Proceedings of INTERSPEECH 2008, Brisbane, S. 1505-1508.

Becker, Timo, Michael Jessen & Catalin Grigoras (2009): Speaker verification based on formants using Gaussian mixture models. Proceedings of NAG/DAGA 2009, Rotterdam, 1640-1643.

Broeders, A.P.A. (2001): Forensic Speech and Audio Analysis, Forensic Linguistics 1998 to 2001: A review. Proceedings of the 13th INTERPOL Forensic Science Symposium, Lyon, France, 16-19 October 2001.

Hollien, Harry (1990): The Acoustics of Crime. The New Science of Forensic Phonetics. New York: Plenum Press.

Hollien, Harry (2002): Forensic Voice Identification. San Diego: Academic Press.

Jessen, Michael (2012): Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs. München: LINCOM.

Jessen, Michael, Olaf Köster & Stefan Gfroerer (2005): Influence of vocal effort on average and variability of fundamental frequency. The International Journal of Speech, Language and the Law 12: 174–213.

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

McDougall, Kirsty (2006): Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. The International Journal of Speech, Language and the Law 13: 89-126.

Moos, Anja (2010): Long-Term Formant Distribution as a measure of speaker characteristics in read and spontaneous speech. The Phonetician 101/102: 7-24.

Morrison, Geoffrey Stewart (2011): A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model - universal background model (GMM-UBM). Speech Communication 53: 242-256.

Morrison, Geoffrey Stewart, Cuiling Zhang & Philip Rose (2011): An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. Forensic Science International 208: 59-65.

Nolan, Francis (1983): The Phonetic Bases of Speaker Recognition. Cambridge: Cambridge University Press.

Nolan, Francis & Catalin Grigoras (2005): A case for formant analysis in forensic speaker identification. The International Journal of Speech, Language and the Law 12: 143-173.

Reynolds, D.A. & W.M. Campbell (2008): Text-independent speaker recognition. In: Jacob Benesty, M. Mohan Sondhi, Yiteng Huang (Hrsg.) Springer Handbook of Speech Processing. Berlin: Springer. S. 763-781.

Rose, Philip (2002): Forensic Speaker Identification. London: Taylor & Francis.

Rose, Phil (2010): The effect of correlation on strength of evidence estimates in Forensic Voice Comparison: uni- and multivariate Likelihood Ratio-based discrimination with Australian English vowel acoustics. International Journal of Biometrics 2: 316-329.

Rose, Phil (2013): More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. Erscheint in The International Journal of Speech, Language and the Law, Ausgabe 20.1.

Van Leeuwen, David A. & Niko Brümmer (2007): An introduction to application-independent evaluation of speaker recognition systems. In: Christian Müller (Hrsg.) Speaker Classification I: Fundamentals, Features, and Methods. Berlin: Springer. S. 330-353.

Kontakt

Dr. Marianne Jessen
Stimmenvergleich, Wiesbaden
Postfach 120309
65081 Wiesbaden

Email: jessen@stimmenvergleich.de

Oscar Forth
Oxford Wave Research Ltd, Oxford, United Kingdom

Email: oscar@oxfordwaveresearch.com

Anil Alexander
Oxford Wave Research Ltd, Oxford, United Kingdom

Email: anil@oxfordwaveresearch.com