Bundeskriminalamt

Forensischer Stimmenvergleich mit Formanten und Cepstralkoeffizienten:
Aktuelle Methoden und Ergebnisse

Mittwochs-Kolloquium am Institut für Phonetik und Sprachverarbeitung der Ludwig-Maximilians-Universität München

29.01.2014

Michael Jessen
BKA, Kriminaltechnisches Institut, Fachbereich für Sprechererkennung, Tonträgerauswertung und Autorenerkennung (KT54)

Michael.Jessen@bka.bund.de

Themen der forensischen Phonetik I: Sprechererkennung

Gebiet	Kurzbeschreibung
Stimmen-vergleich	Audioaufnahme des Täters und eines Verdächtigen: Aussagen zur Frage, ob es sich um den gleichen Sprecher handelt.
Stimmen-analyse (=Stimmenprofil)	Audioaufnahme des Täters, aber kein Verdächtiger: Angaben zu u.a. Geschlecht, Alter, Regiolekt, Soziolekt, L1-Einfluss; Angaben von Sprechereigenschaften, die von Laien verstanden werden, z.B. hohe Stimme, schnelles Sprechen, undeutliches Sprechen.
Stimmen-Wahlgegen-überstellung	Keine Audioaufnahme, aber Zeugenwahrnehmung der Täterstimme und Verdächtiger existiert, ist dem Zeugen aber nicht bekannt: Präsentation der Verdächtigenstimme zusammen mit den Stimmen anderer Unbeteiligter; der Zeuge gibt an, ob es sich jeweils um die Stimme des Täters handelt.

Bundeskriminalamt

Themen der forensischen Phonetik II: Tonträgerauswertung

Gebiet	Kurzbeschreibung
Qualitätsverbesserung	Verbesserung der Sprachverständlichkeit oder der stressfreien Anhörbarkeit mit Hilfe von Sprachsignalverarbeitungsverfahren
Phonetische Textanalyse	Bestimmung des Wortlauts schwer verständlicher Sprachanteile
Authentisierung	Prüfung von Aufzeichnungen/Geräten im Hinblick auf forensisch relevante mögliche Manipulationen
Nicht-sprachliche akust. Ereignisse	z.B. Analyse von Hintergrundgeräuschen, Ereignisabfolge anhand Cockpit-Stimmrekordern, Analyse von Vogelstimmen
ENF-Analyse	Analyse der akustischen Spur der Stromnetzfrequenz (**E**lectric **N**etwork **F**requency) in Aufnahmen zum Zweck der Authentisierung und der zeitlichen Einordnung eines Tatereignisses
Fallrekonstruktion / Perzeptionsexperimente	Beispiele: War das impulsartige Geräusch ein Schuss? Hätte man einen Schrei zwei Räume weiter wahrnehmen können, wenn man unter der Dusche war? Wird Sprache schlechter verstanden, wenn spezielle Textilien vor dem Mund sind?
LADO	**L**anguage **A**nalysis for the **D**etermination of **O**rigin: Analyse der Sprache von Asylbewerbern bei Zweifel über die Korrektheit der Angabe über das Herkunftsland (in Deutschland beim BAMF)

Forensischer Stimmenvergleich

Situation: Audio-Aufnahme der Täterstimme und Audio-Aufnahme eines Tatverdächtigen liegen vor.
Aufgabe: Die Aufzeichnungen vergleichen und eine Wahrscheinlichkeitsaussage darüber treffen, inwieweit und wie stark die Identitätshypothese (gleicher Sprecher) oder die Nicht-Identitätshypothese (verschiedene Sprecher) unterstützt wird.

Beispiele:	Täter	Verdächtiger
Drogenhandel		 (.wav)
Erpressung	 (.wav)	 (.wav)
Kindesentführung	 (.wav)	 (.wav)

(Selektive Kurz-) Geschichte des forensischen Stimmenvergleichs

1. Frühphase prä-1980er
2. Konsolidierung der forensischen Phonetik als Disziplin der angewandten Phonetik/Linguistik (1980er bis heute)
3. Einzug der forensischen Statistik und der forensischen automatischen Sprechererkennung (ca. 2000 bis heute)

Frühphase: Schlechter Start in den USA mit der Voiceprint-Methode I

Sinnvoller Grundgedanke: In Spektrogrammen befindet sich sprecherunterscheidende Information, z.B. in den Formanten (siehe z.B. die Patterson & Barney-Studie; frühe Sprechernormalisierungsforschung wie Ladefoged & Broadbent);

aber Probleme in der Realisierung/Propagierung durch Kersta:

1. Zu optimistisch in Hinblick auf die **inter-individuelle Variation** (diese ist geringer als angenommen)
2. Zu naiv in Hinblick auf die **intra-individuelle Variation** (diese ist größer als angenommen)
3. Zu naiv/fahrlässig in Hinblick auf die Kompetenz der forensischen Experten (visuelles Pattern-Matching reicht nicht, umfangreiche Kenntnisse in akustischer Phonetik sind erforderlich)
4. Zu einseitig an Akustik orientiert (wurde später etwas gelockert)

Frühphase: Schlechter Start in den USA mit der Voiceprint-Methode II

Resultierende Probleme/Konsequenzen:

1. Gefahr der Fehlbegutachtung bei unausgereifter Methode
2. Situation der verpassten Möglichkeiten für die wissenschaftliche Entwicklung der forensischen Sprechererkennung in den USA (Kommentar von Hollien 2013, Vol 1, No 1 von http://lesli-journal.org)
3. Danach (in Abgrenzung von der Methode) übertriebener Skeptizismus, was die Verwendung von Spektrogrammen und die Messung von Formanten betrifft (Kind-mit-dem-Bad-Ausschütt-Reaktion)

Es gibt Resolutionen gegen die Voiceprint-Methode, sie wird aber offenbar weiterhin in einigen Ländern und einigen US-Staaten verwendet.

Frühphase: Weitere internationale Schwerpunkte

- **Großbritannien: Sinnvoll aber einseitig auditiv**. Auditive Analyse mit besonderer Berücksichtigung feiner dialektaler Analyse (z.B. Stanley Ellis, John Baldwin)
  https://www.leeds.ac.uk/secretariat/obituaries/2009/ellis_stanley.html
- **UDSSR und Ostblock: Vielseitig aber mysteriös**. Sowohl akustische als auch auditive/linguistische Ansätze werden berücksichtigt, aber es wird sehr wenig publiziert. In der DDR aber ein interessantes Werk von Christian Koristka (1968): *Magnettonaufzeichnungen und kriminalistische Praxis*.
- **Deutschland (West): Zu früh gefreut.** Automatische Sprechererkennung am BKA in den 1970ern (Ernst Bunge) war noch nicht leistungsfähig genug und wurde in den 1980ern durch die auditiv-akustische Methode abgelöst (Hermann Künzel).

Konsolidierungsphase: Prinzipielles und Organisatorisches

- Methodische Einseitigkeiten werden beseitigt, es werden jetzt sowohl auditive als auch akustische Aspekte berücksichtigt (z.B. durch Francis Nolan, Hermann Künzel, Peter French, Harry Hollien); es entsteht die **auditiv-akustische Methode** des forensischen Stimmenvergleichs.
- Eine wissenschaftliche Gesellschaft wird 1991 gegründet (*IAFP*; später *IAFPA*) und eine assoziierte Zeitschrift erscheint (*Forensic Linguistics*; später *International Journal of Speech, Language and the Law*).
- Zur forensischen Phonetik erscheinen die ersten Monographien (Nolan 1983) und Textbücher (Künzel 1987, Baldwin & French 1990, Hollien 1990). Dort wird die auditiv-akustische Methode besonders hervorgehoben.
- Forensische Phonetik tritt seit 1991 regelmäßig als Thema auf dem ICPhS auf.

Konsolidierungsphase: Starke deskriptive Orientierung

Zum Beispiel Publikationen zu:

- Fallstudien; Illustration und Klassifikation typisch forensisch-phonetischer Probleme, z.B. Stimmverstellung
- Untersuchungen zur **intra**-individuellen Variation in Gebieten, die in der allgemeinen Phonetik vergleichsweise wenig behandelt wurden, z.B. Einfluss Alkohol, Einfluss Stress, Einfluss Sprechlautstärke
- Erstellung von Korpora mit rel. vielen Sprechern und daran Untersuchungen zur **inter**-individuellen Variation. Zum Beispiel Untersuchungen zur mittleren Grundfrequenz bei
  - Künzel (1987, 1989, 2000) anhand von 100 männl. und 50 weiblichen (erwachsenen) Sprechern des Deutschen an Lesesprache
  - Jessen et al. (2005), Jessen (2009) anhand von 100 männl. Sprechern des Deutschen (spontan/gelesen; neutral/Lombard).
  - Hudson et al. (2007) [Cambridge-Gruppe um Francis Nolan] anhand von 100 männl. Sprechern des SBE (versch. an forensische Settings angelehnte Spontan- und Leseaufgaben)

Phase 3: Einzug der forensischen Statistik und der automatischen Sprechererkennung

- Späte 1990er: Allgemeine Methoden der **automatischen Sprechererkennung** werden für die Forensik entdeckt bzw. adaptiert (siehe insbesondere die Lausanne-Gruppe um Andrzej Drygajlo und Didier Meuwly; Überblick von Drygajlo 2012).
  - Voraussetzung 1: Als Zwischenschritt dorthin war es erforderlich, Methoden zu entwickeln, die textunabhängig arbeiten und die eine gewisse Kanalrobustheit haben, z.B. Telefonkanal. Wichtig hierzu ist die Dissertation von Douglas Reynolds (1992), der auch das Gaussian Mixture Modeling (GMM) für die Sprechererkennung eingeführt hat.
  - Voraussetzung 2: In den 1990ern wurde in den forensic sciences die Verwendung der Bayes'schen Statistik, einschließlich der Verwendung von Likelihood Ratios propagiert. Dieser (häufig so genannt) **Likelihood Ratio-Ansatz** wurde insbesondere durch Drygajlo und Meuwly für die automatische Sprechererkennung eingeführt.
- Unabhängig davon wurde durch Phil Rose dieser Likelihood Ratio-Ansatz auch für die forensische Phonetik eingeführt (zunächst für Formantenmessungen) und international bekannt durch das Textbuch Rose (2002).
- Ab dann war die forensisch-phonetische Methodologie nicht mehr nur deskriptiv orientiert, sondern hatte ihre eigene Form von moderner Statistik (es brauchte aber lange, bis diese Nachricht durchdrang, und viele hat sie bis heute noch nicht erreicht).

Zwei Grundfragen

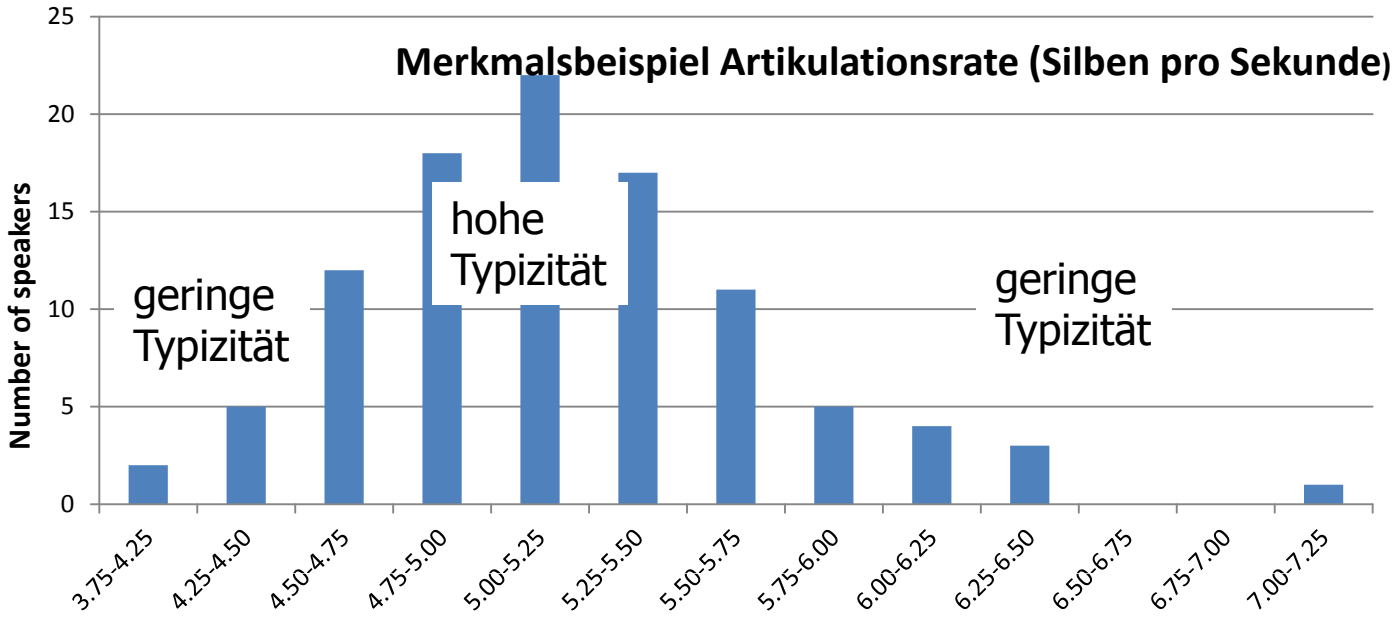
1. Wie sprecherdiskriminativ (sprecherspezifisch) ist ein auditives/akustisches Merkmal M im Allgemeinen?
2. Wie hoch ist die Beweisstärke für oder gegen Identität in einem spezifischen Stimmenvergleichsfall, wenn beim Tatsprecher die Merkmalsausprägung M_i und beim Vergleichssprecher die Merkmalsausprägung M_k festgestellt (z.B. gemessen) wird?

Antworten im Kontext von Phase 2

1. Entweder ergibt sich aus Erfahrung, wie wertvoll ein Merkmal für die Sprechererkennung ist, oder (bei einigen Merkmalen) wird mit Mitteln klassischer statistischer Verfahren die Sprecherdiskriminationsleistung quantifiziert, z.B. der f-Ratio im Rahmen einer ANOVA (Nolan 1983).
2. Einschätzungen oder deskriptive Analyse von **Ähnlichkeit** und **Typizität**. Manchmal auch Ähnlichkeitsanalyse durch Signifikanztests („frequentistischer Ansatz" mit „cliff-edge effect")

Bundeskriminalamt

Ähnlichkeit und Typizität



Beispielfall 1:	**T**äter **V**erdächtiger	Geringe Ähnlichkeit: Evidenz gegen Identität
Beispielfall 2:	**T V**	Große Ähnlichkeit, hohe Typizität: moderate Evidenz für Identität
Beispielfall 3:	**T V**	Große Ähnlichkeit, geringe Typizität: hohe Evidenz für Identität

Antworten im Kontext von Phase 3

1. Die sprecherdiskriminative Leistung eines Merkmals wird quantifiziert mit Kennzahlen wie der Equal Error Rate (**EER**) oder mit Häufigkeitsdarstellungen wie dem **Tippett-Plot** (mehr dazu später).
2. Ähnlichkeit und Typizität werden in Form des Likelihood Ratio (**LR**) quantifiziert, wobei die Ähnlichkeit im Zähler und die Typizität im Nenner des LR ausgedrückt wird.

Bayes Theorem und der Likelihood Ratio

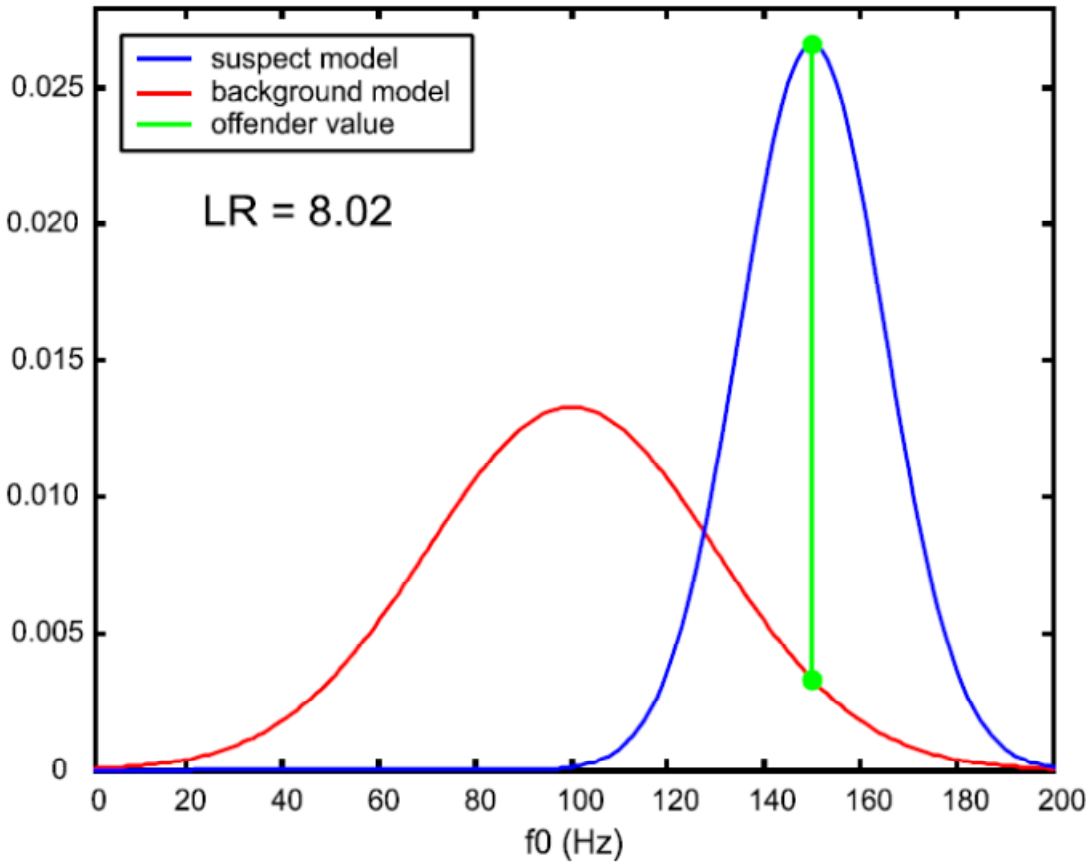
$$\frac{p(H_{so}|E)}{p(H_{do}|E)} = \frac{p(E|H_{so})}{p(E|H_{do})} \times \frac{p(H_{so})}{p(H_{do})}$$

posterior odds | likelihood ratio | prior odds

SO (same origin) = gleicher Sprecher
DO (different origin) = verschiedene Sprecher

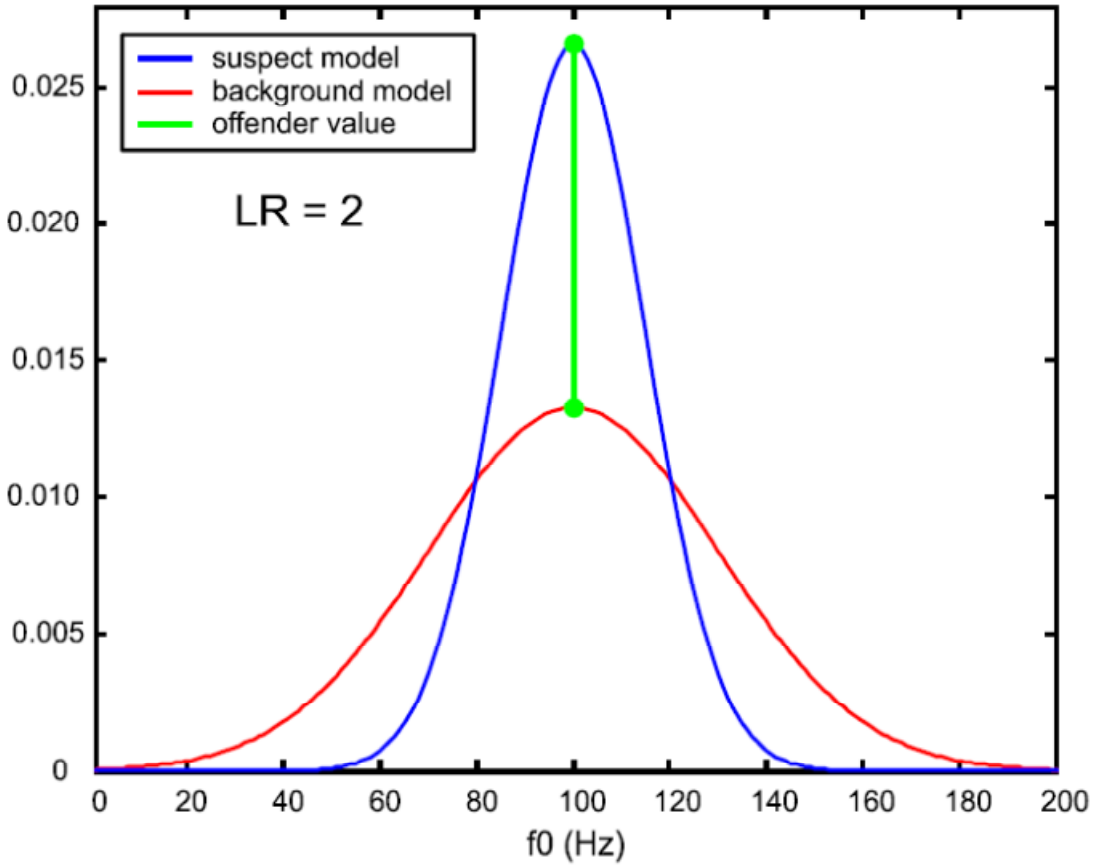
Morrison (2010)

Abstraktes Bsp. einer LR-Berechnung. **Hier:** Große Ähnlichkeit, geringe Typizität



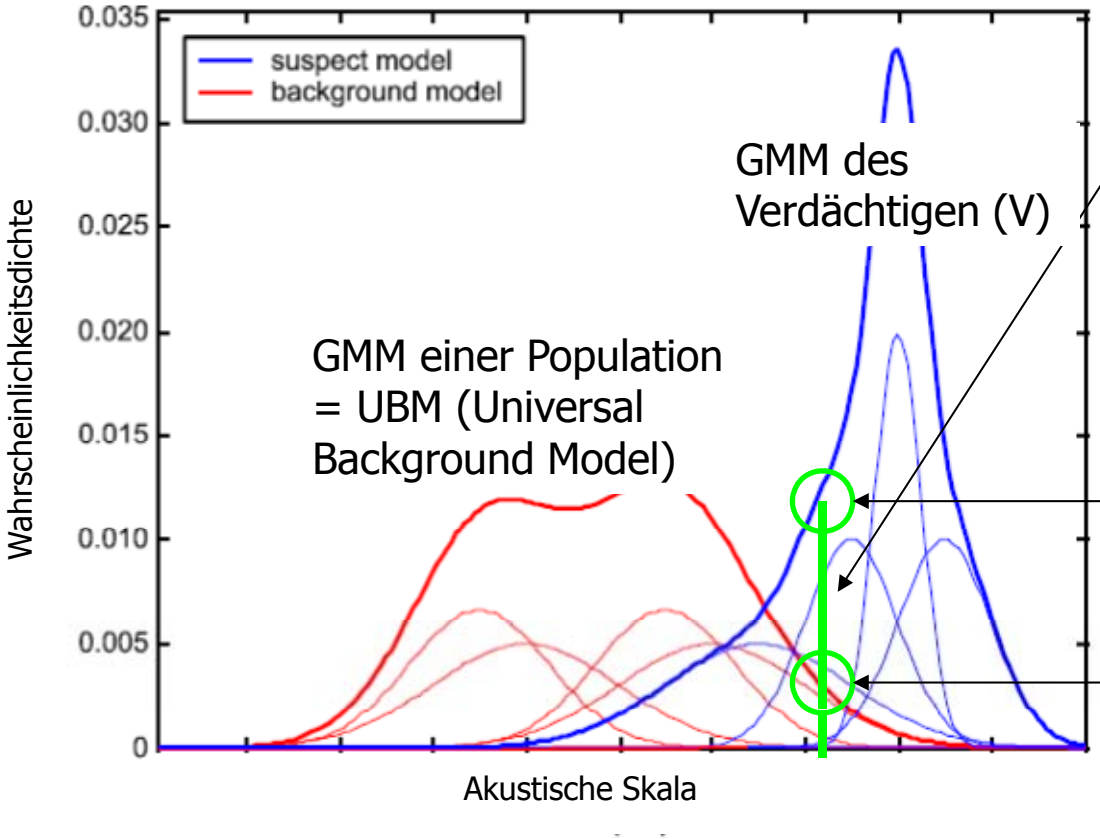
Morrison (2010)

Abstraktes Bsp. einer LR-Berechnung. **Hier:** Große Ähnlichkeit, hohe Typizität



Morrison (2010)

Noch ein abstraktes Beispiel, diesmal mit Gaussian Mixture Models (GMM)



Evidenz = Merkmalsvektor des Tatsprechers (T)

$p(\text{Merkmalsvektor}_T|\text{GMM}_V)$ (1)

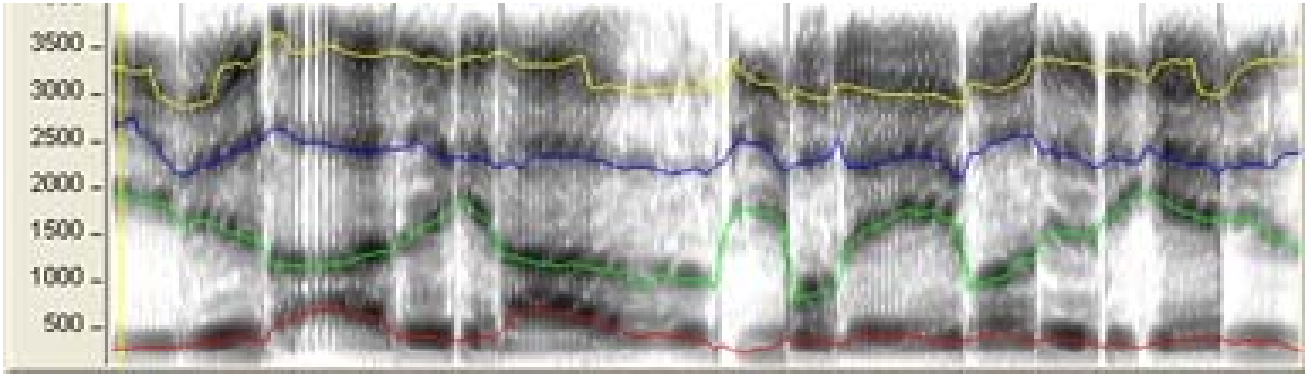
$p(\text{Merkmalsvektor}_T|\text{GMM}_{UBM})$ (2)

LR= (1) geteilt durch (2).
Hier LR>1: Unterstützung für die Identitätshypothese

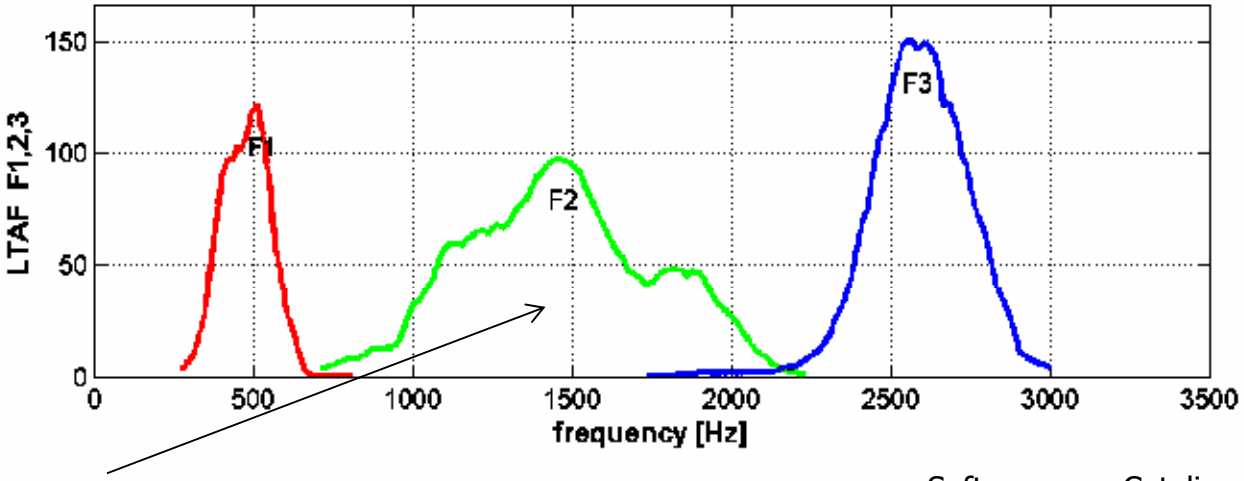
Bild basierend auf Morrison (2010)

Experimente mit Long-Term Formants (LTF)

Methode: Zusammenschnitt vokalischer Anteile, dann Formant tracking mit manueller Kontrolle



Beispiel einer LTF-Verteilung in einer Aufnahme



Siehe insb. die komplexe Verteilung von F2: Gaussian Mixture Modeling macht Sinn bei solchen Daten.

Software von Catalin Grigoras

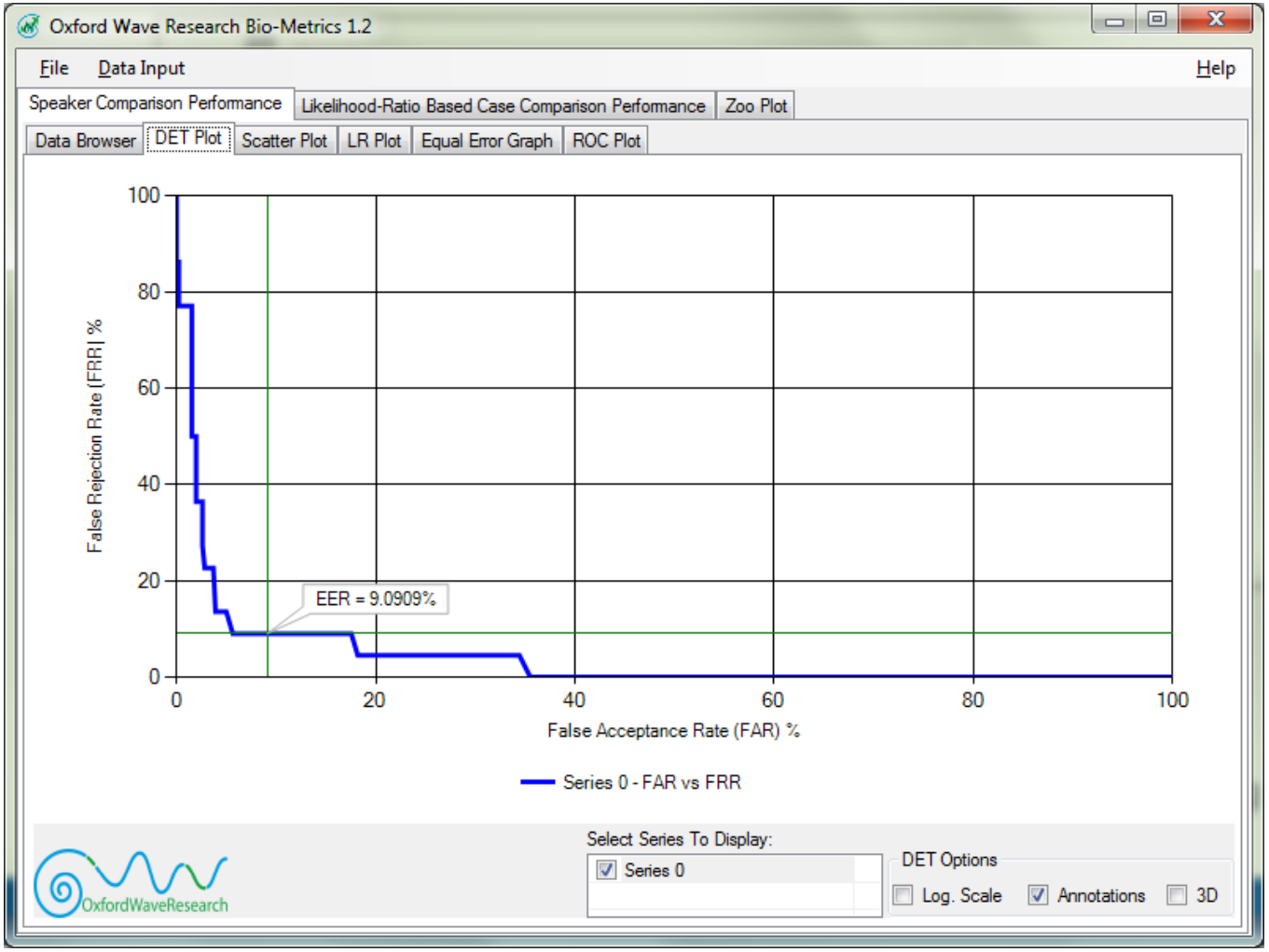
Versuchsbedingungen

- 22 männliche erwachsene Sprecher der (leicht gefärbten) mittelwestdeutschen regionalen Varietät des Deutschen (aus Pool 2010; Jessen et al. 2005)
- Von jedem Sprecher zwei Aufnahmen, so dass 22 same-speaker-Vergleiche und 462 different-speaker-Vergleiche. Studio-Aufnahmen, die später durch Molbiltelefon-Verbindungen übermittelt und neu aufgezeichnet wurden.
  - "Tataufnahmen" (analysis recording) aus einer (recht) spontanen Aufgabe (Erfahrungen und Gedanken während des Experiments kommentieren)
  - "Vergleichsaufnahmen" (comparison recording) aus einer (etwas weniger) spontanen Aufgabe (Bildbeschreibung unter Vermeidung bestimmter Wörter)
- UBM basierend auf 22 Sprechern im Sprechstil der Vergleichsaufnahmen


(.wav)

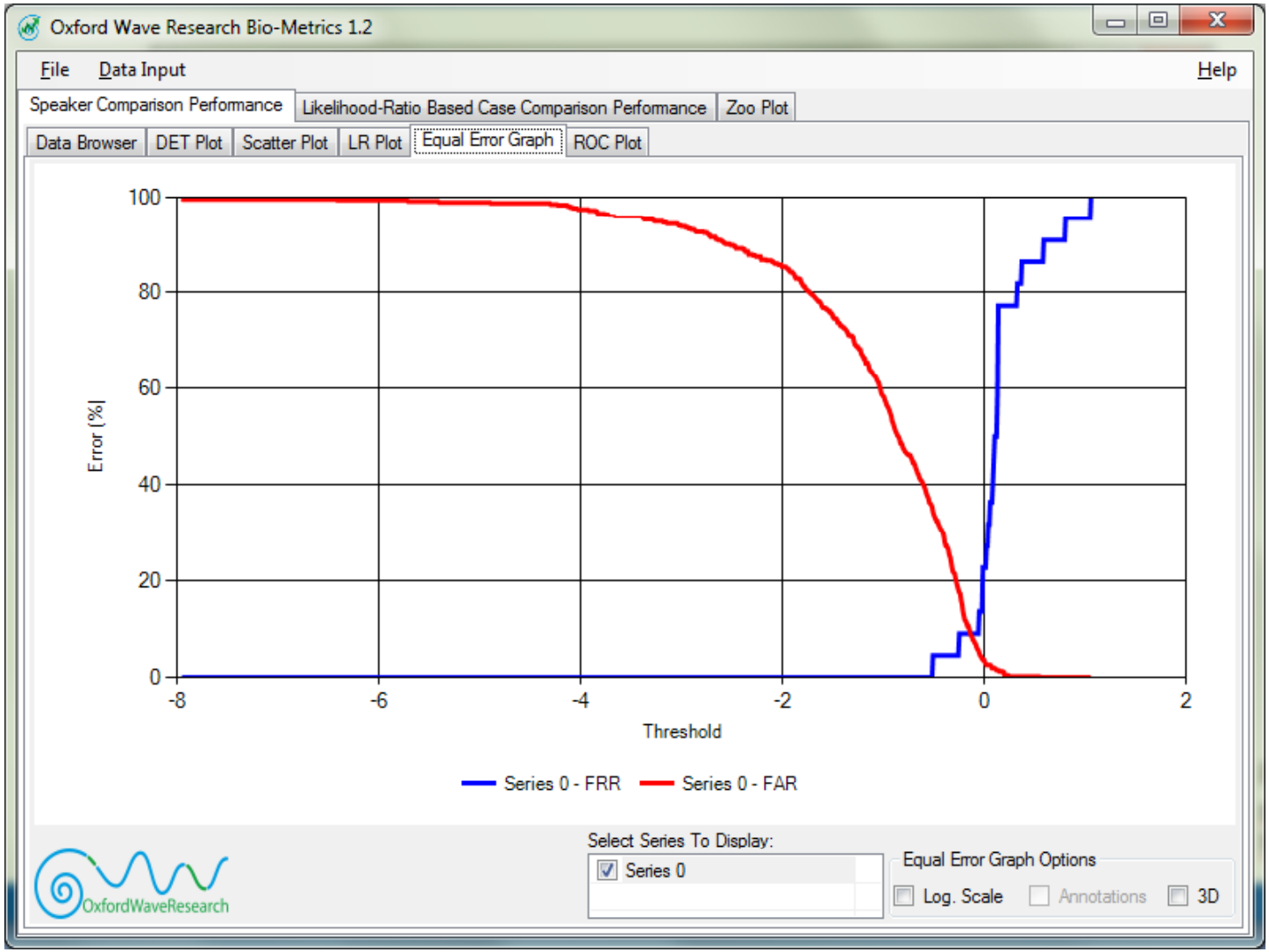

(.wav)

Resultate: Equal Error Rate und DET-Plot

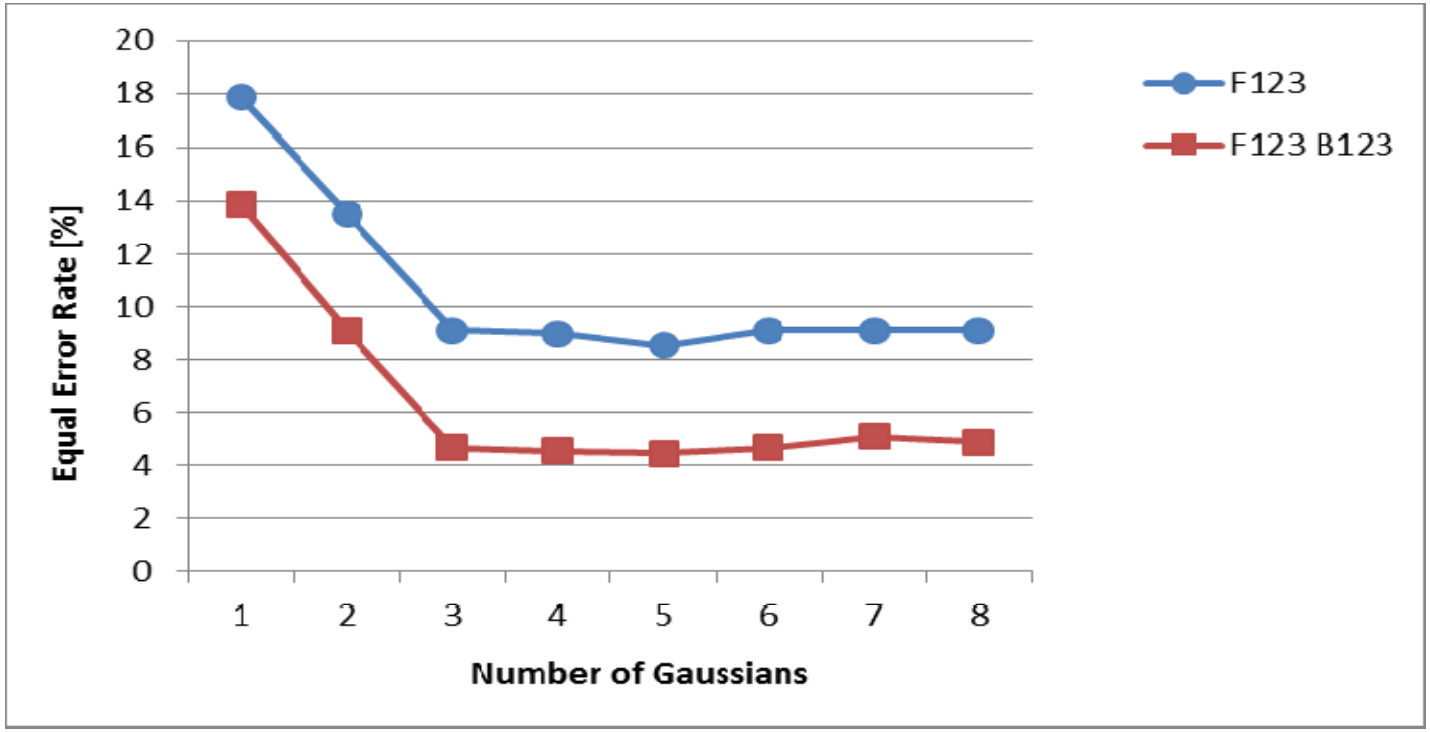


Mit Software Bio-Metrics

Resultate: Tippett-Plot

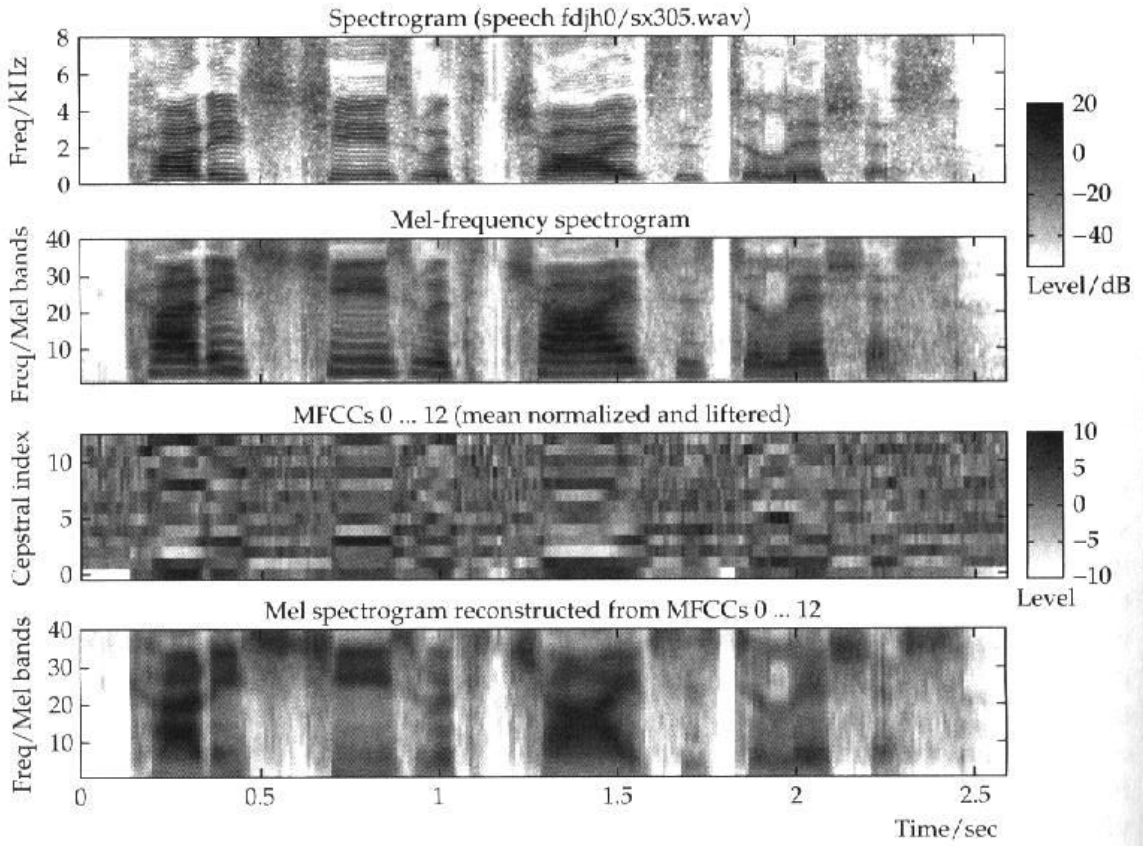


Versuchsreihe: verschiedene # Gauß-Module; + / - Formanten-Bandbreiten



- Man braucht bei den LTF-Daten drei Gauß-Module, um die Verteilungen angemessen für die Sprechererkennung zu modellieren, aber mehr sind nicht erforderlich.
- Die Hinzunahme der Bandbreiten führt zu einer Verbesserung der Ergebnisse.

Experiment mit automatischer Sprechererkennung: MFCC als Merkmal



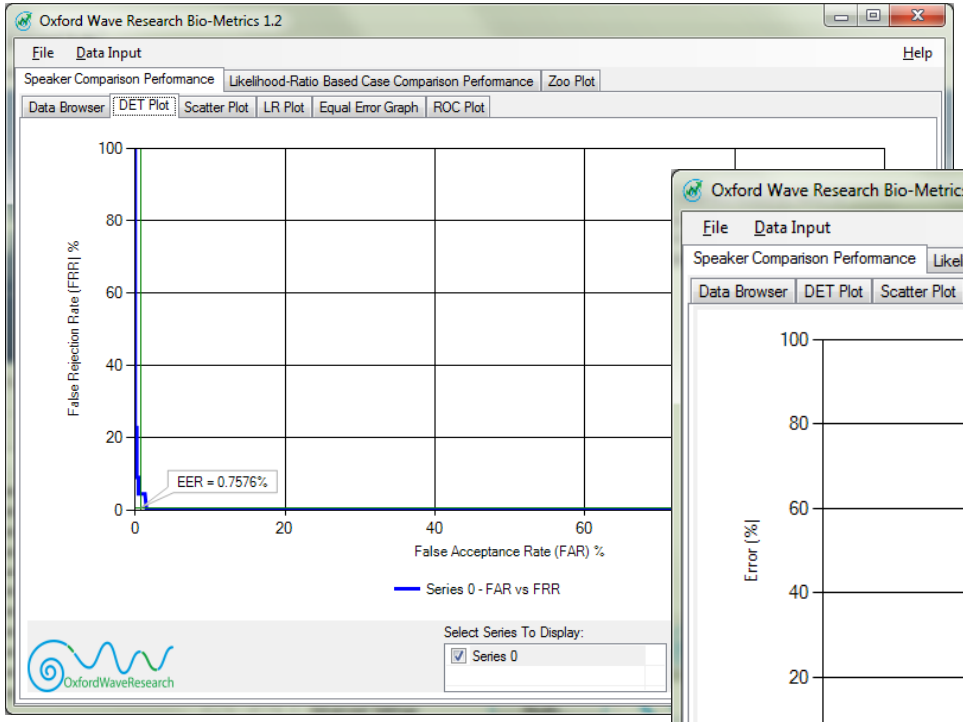
Ellis (2010, HB of Phonetic Sciences 2. ed)

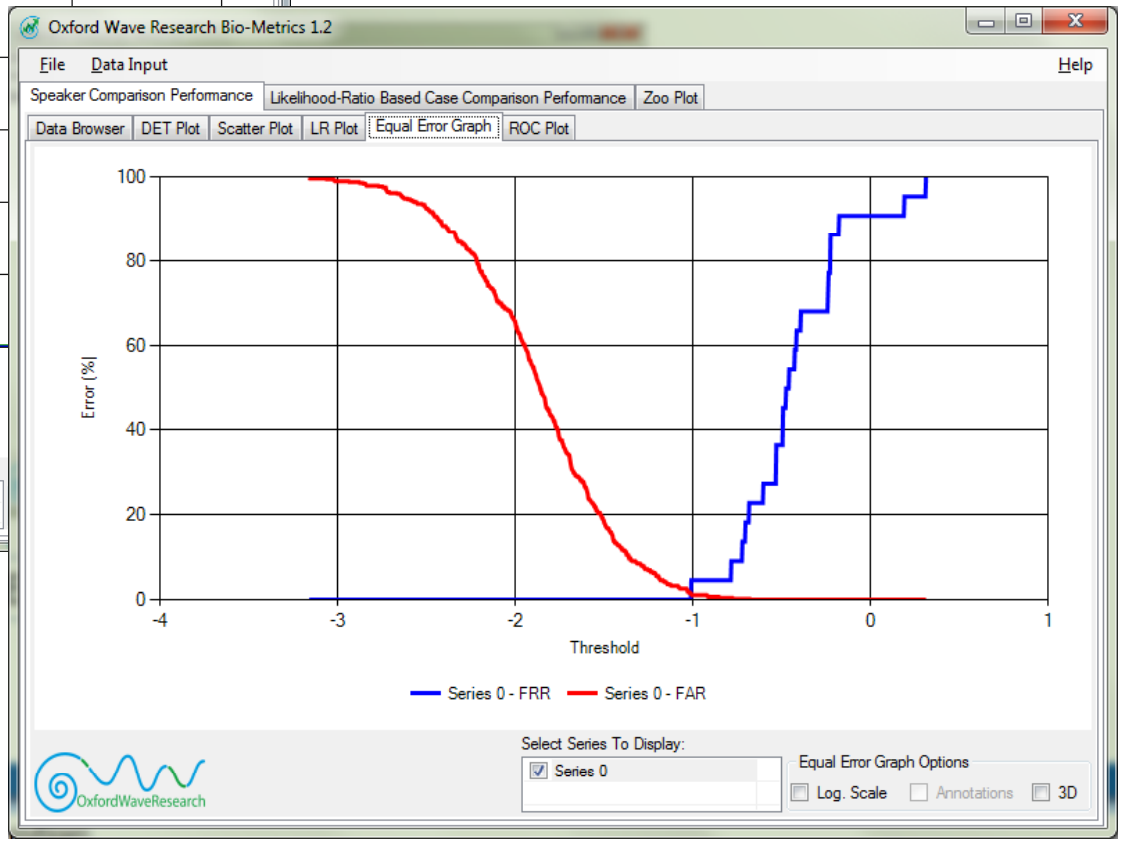
Einige Eigenschaften von MFCC:

- MFCC erfassen Vokaltrakteigenschaften (Filter) und ignorieren f0 (Quelle)
- Die MFCC (ca. 13) sind (fast) unkorreliert
- Unsupervidierte (automatische) Merkmals-extraktion. Störungen sind darin enthalten und müssen wegnormalisiert werden, z.B. durch Cepstral Mean Subtraction.

Bundeskriminalamt

Experiment mit automatischer Sprechererkennung: Plots





EER weniger als 1%

Ausgewählte Takehome-Messages

- Vielfalt an Methoden in der geschichtlichen Entwicklung des forensischen Stimmenvergleichs
- Heute: Koexistenz verschiedener Methoden aus allen drei genannten Phasen der Entwicklung (Gold & French 2011)
- Es gibt durch die Entwicklungen der automatischen Sprechererkennung und durch die Initiative einiger Phonetiker (insbesondere die Gruppe um Phil Rose in Australien) jetzt eine eigene statistische Methodologie für:
  - Die Bestimmung der Sprecherdiskriminationsleistung eines Merkmales (oder von Merkmalskombinationen)
  - Die Evidenzstärke eines einzelnen Stimmenvergleichs
- MFCC-basierte (automatische) Methoden ergeben an „gutem" (laborbasiertem) Material bessere Ergebnisse als (Langzeit-) Formantenfrequenzen, die Leistung der Formanten ist allerdings auch respektabel. Es ist noch zu erforschen, wie die Ergebnisse bei echten Falldaten aussehen.