Karlsruhe Institute of Technology

F0-Erkennung bei elektromyographischer Sprachsynthese mit Elektrodenarrays

Bachelorarbeit am Cognitive Systems Lab
Prof. Dr.-Ing. Tanja Schultz
Fakultät für Informatik
Karlsruher Institut für Technologie

von

Luben Alexandrov

Betreuer:
Prof. Dr.-Ing. Tanja Schultz
Dipl. Inform. Matthias Janke
Dipl.-Math. Michael Wand

Tag der Anmeldung: 13. Dezember 2011
Tag der Abgabe: 13. Februar 2012

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 16.04.2012

Zusammenfassung

Elektromyographische Sprachsynthese bezeichnet den Prozess, bei dem elektrische Signale der artikulatorischen Gesichtsmuskeln gemessen werden und mit deren Hilfe Sprache synthetisiert wird.

Eines der größten Probleme in der Sprachsynthese ist die Rekonstruierung der Fundamentalfrequenz F0. Diese Arbeit befasst sich mit der Aufgabe, Verfahren für Vorverarbeitung und Merkmalsextraktion für elektromyographische (EMG) Sprachsynthese, zu analysieren und evaluieren. Dabei wird besondere Rücksicht auf bessere Resultate bei der F0-Erkennung genommen. Für die Experimente wird ein Sprachsynthesesystem, das auf Gaussian Mixture Models (GMM) basiert, verwendet. Spezifisch in dieser Arbeit ist, dass man für die Datenerfassung nicht die bisher gewöhnlichen einzelnen Ag-AgCl Elektroden verwendet. Die Aufnahmen sind mit Elektrodenarrays gemacht.

Als Vorverarbeitungsmethoden werden Principal Component Analysis (PCA), Independent Component Analysis (ICA) und Second Order Blind Identification (SOBI) durchgeführt. Zusätzlich untersuchen wir den Einfluss der Fenstergröße und die Anzahl der Gauß-Glocken auf die Ergebnisse. Wir päsentieren auch ein neues Featureset, welches eine Kombination von zeitlichen und spektralen Merkmale ist.

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Einleitung

Die künstliche Synthese der menschlichen Sprache ist ein Forschungsgebiet, das seit mehreren Jahrhunderten die Wissenschaftler reizt. 1779 baute Christian Kratzenstein eine Orgel mit Lingualpfeifen und Rohre, die die Resonanzfrequenzen des menschlichen Vokaltraktes reproduzieren konnten und somit die Vokale a, e, i, o und u synthetisierten. Es folgten weitere Entwicklungen von unterschiedlichen Forschern, die immer bessere "sprechende Maschinen" gebaut haben.

Auf immer mehr Popularität freut sich in den letzten Jahren die elektromyographische (EMG) Sprachsynthese. Man misst dabei mit Elektroden die elektrischen Spannungen an der Hautoberfläche des Gesichtes, die bei der Kontraktion der artikulatorischen Muskeln entstehen. Die so gesammelten Daten werden vorverarbeitet, wobei eine Merkmalsextraktion (feature extraction) durchgeführt wird. Danach erfolgt eine Dimensionsreduzierung, wobei nur die wichtigsten und aussagekräftigsten Feature genommen werden. Mit einem Teil der Daten wird ein Synthese-System trainiert. Mit dem anderen Teil der Daten wird das System getestet, indem man die vorverarbeiteten EMG-Daten auf Sprache abbildet. Da die Muskelaktivität auch bei lautloser Sprache[1] existiert, ermöglicht dieses Verfahren eine Verwendung als Silent Speech Interface, das in vielen Gebieten nützlich sein kann. Leute mit Sprachbehinderungen (z.B. Kehlkopfkrebs) können dadurch kommunizieren. Dieses Interface würde auch die Sprachkommunikation in Räumen mit Hintergrundgeräusche ermöglichen. Es kann auch für nichtstörende Sprachkommunikation in leisen Umgebungen, wie Bibliotheken, genutzt werden. Weiterer Einsatz des Systems wäre für ein Schutz der Vertraulichkeit der privaten Gespräche.

Ein sehr wichtiger Aspekt der Sprache ist die Fundamentalfrequenz (F0). Diese beinhaltet Information über die Timbre des Sprechers und über die Intonation und die Lebendigkeit der Aussprache. Wie wir in dem Kapitel Grundlagen sehen werden, steckt in der Fundamentalfrequenz sogar Information über die emotionale Stimmung des Sprechers. Zusammenfassend kann man sagen, dass für eine anspruchsvolle

[1] Lautlose Sprache ist, wenn der Sprecher seinen Sprechapparat (Mund, Zunge) bewegt, genau wie er es bei normaler Sprache machen würde, jedoch ohne Töne zu produzieren (silent speech versus audible speech).

Sprachkommunikation die Fundamentalfrequenz der Sprecher essentiell ist. Deswegen wird ein Sprachsynthesesystem menschlichnahe Sprache produzieren, nur wenn es die F0 richtig erkennen kann. Das ist auch die Motivation für die Erforschung und das Suchen der F0 in den EMG Signalen. Die korrekt erkannte Fundamentalfrequenz wird ermöglichen, nicht nur die Übermittlung der bloßen Botschaft, sondern auch die Synthese von natürlich klingender Sprache.

In dieser Arbeit wird ein auf Gaussian Mixture Model (GMM) basiertes Sprachsynthesesystem verwendet. Die EMG-Signale wurden parallel mit Audiosignalen aufgenommen. Für die Evaluierung wird die Korrelation zwischen zwei F0-Konturen verwendet, die entsprechend aus dem Audiosignal und aus dem EMG-Signal gewonnen werden.

Der benutzte Datencorpus besteht aus insgesamt 11 audible Sessions, die mit vier unterschiedlichen Sprechern aufgenommen wurden. Dabei beinhaltet jede Session 50 englische Sätze, die laut ausgesprochen wurden, aus Fernsehnachrichten. 45 der Sätze sind für Training der GMMs verwendet und 5 zum Testen.

1.1 Zielsetzung der Arbeit

Diese Bachelorarbeit erweitert die Konzepte, die in [16] präsentiert wurden. In dieser Arbeit wurde das EMG-Signal mit sechs Kanälen durch einzelne Ag-AgCl Elektroden erfasst. Ziel in dieser Arbeit ist Elektrodenarrays für die Datenaufnahme zu verwenden und ihre Performanz im Hinblick auf F0-Schätzung zu untersuchen. Die Elektrodenarrays wurden noch nicht für Sprachsynthese untersucht. Deswegen setzen wir uns das Ziel, diese neue Art von Aufnahmehardware zu erforschen. Ein zweites Ziel ist unterschiedliche Vorverarbeitungsmethode auf den EMG-Daten durchzuführen und die entstehenden Ergebnisse zu evaluieren.

1.2 Gliederung der Arbeit

In dem Kapitel Grundlagen erläutern wir die Einzelheiten über die Fundamentalfrequenz in der Sprache. Es werden auch die einzelnen technischen Schritte, die wir bei der Sprachsynthese durchführen, beschrieben. Danach berichten wir über die verwendete Hardware und Software in dem Kapitel Implementierung. Als nächstes stellen wir in Kapitel 4 den Datencorpus vor. In dem Kapitel Durchgeführte Experimente werden die Resultate von unterschiedlichen Vorverarbeitungsmethoden präsentiert und evaluiert. Die folgenden Punkte werden untersucht:

- Suchen nach besseren Featuresets
- Variation der Fensterlänge
- Einfluss der Anzahl der Gauss-Glocken auf die Resultate
- Principal component analysis (PCA)
- Independent component analysis (ICA)
- Second order blind identification (SOBI)

- Resultate mit weniger EMG-Kanälen

Im Kapitel Zusammenfassung und Ausblick werden schließlich die Ergebnisse resümiert. Es sind auch konkrete Vorschläge für weitere Forschungsmöglichkeiten gegeben.

2. Grundlagen

In diesem Abschnitt werden wir die Wissensgrundlagen für die Elektromyographie (EMG) und für die Fundamentalfrequenz (F0) präsentieren. In den ersten zwei Unterabschnitten wird der anatomische und psychologische Ursprung der EMG und der F0 erläutert. Im Folgenden gibt es auch eine Beschreibung wie die Mensch-Maschine Kommunikation von richtig geschätzter und synthesierter F0 profitieren kann. Der dritte Unterabschnitt beschreibt das verwendete Sprachsynthesesystem und die technische Vorgehensweise, die in dieser Arbeit für die F0-Schätzung verwendet wurden. *Verwandte Arbeiten* kommentiert die bisherigen Ansätze für die EMG-zu-F0 Transformation.

2.1 EMG Grundlagen

In diesem Abschnitt beschreiben wir die physiologischen Grundlagen der EMG-Signale. Mit Elektromyographie werden die Muskelaktivitäten aufgenommen. Die menschliche Muskulatur ist in zwei Typen unterteilt - glatte und Skelettmuskulatur[1]. Die glatte Muskulatur kommt in den inneren Organen vor und man kann sie nicht willkürlich kontrollieren. Die glatten Muskeln werden von dem vegetativen Nervensystem gesteuert. Andererseits werden die Skelettmuskeln vom somatischen Nervensystem beeinflusst, die willkürlich kontrollierbar ist. Die Skelettmuskeln werden über Sehnen am Skelett befestigt. Sie sind zuständig für die Bewegungen der Körperglieder aber auch für die Mimik und für die Bewegung der Zunge und des Kehlkopfes. An der Sprachproduktion sind hauptsächlich die Skelettmuskeln beteiligt, deswegen werden wir an dieser Stelle ihren Aufbau kurz beschreiben.

In Abbildung 2.1 sind die einzelnen Bestandteile eines Muskels visualisiert. Ein Skelettmuskel ist ein Bündel von mehreren Muskelfasern. Die Muskelfasern bestehen sich aus Tausenden kleineren Strukturen, die Myofibrillen genannt sind. Die Myofibrillen von ihrer Seite sind von Myofilamenten gebildet. Diese Myofilamenten sind fadenförmigen Protein-Ketten, die einen Durchmesser von unter $6nm$ haben.

[1] Auch quergestreiften Muskulatur

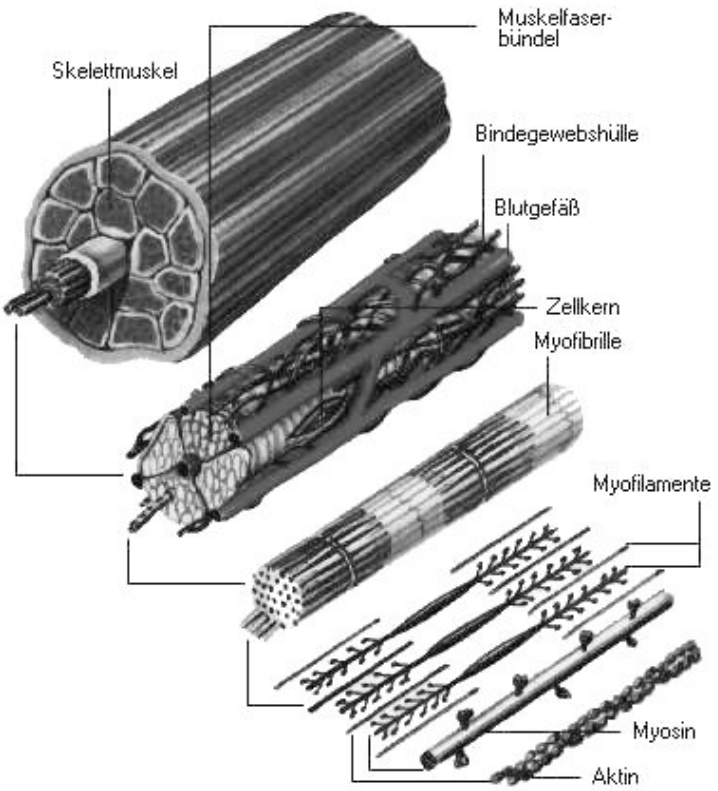


Abbildung 2.1: Aufbau eines Muskels, aus [3]

Ein Aktionspotential ist die Abweichung des Zellenmembranenspotentials von seinem Ruhezustand. In der Plasmamembran der biologischen Zellen gibt es Ionenkanälen, die bei bestimmten Bedingungen Ionenströmen durchlassen oder stoppen[1]. Dadurch können die Aktionspotentiale zwischen den Zellen transportiert werden. Das Nervensystem funktioniert mit solchen Sendungen von Aktionspotentialen. Bei dem Durchlass von Ionenströmen durch die Zellenmembran gilt das Alles-oder-Nichts-Prinzip. Nur wenn bestimmten Schwellenwert des Zellenmembranenspotentials erreicht ist, werden Ionen durchgelassen. Also es gibt nicht so was als ein "starkes" oder ein "schwaches" Aktionspotential. Die Stärke des Signals wird von der Frequenz der nacheinanderfolgenden Aktionspotentialen gebildet.

Zu jeder Muskelfaser wird ein Axon des Nervensystems gekoppelt. Aktionspotentiale wandern das Axon entlang und gehen in die Muskelfaser über. Das verursacht eine Veränderung der chemischen Umgebung in der Muskelfaser. Die Myofilamenten (Myosin und Aktin) reagieren an diese Veränderung und gleiten sich enger aneinander. Das führt zu einer Verkürzung der Muskelfaser und daher auch zu Muskelkontraktion.

Die Aktionspotentiale wandern durch die Muskelfaser. Bei EMG messen wir mit Elektroden auf der Hautoberfläche diese Aktionspotentiale, die im Millivolt Bereich sind. Durch Helmholtz-Doppelschicht wird der Ionenstrom in einem Elektronenstrom umgewandelt. Es existieren auch Methoden mit Nadelelektroden, wo die Elektrode

[1] Siehe dazu Natrium-Kalium-Pumpe

direkt in Muskelfaser gestochen wird. Für ein Interface sind aber diese Elektroden nicht geeignet. Obwohl die Nadelelektroden mehr exakte Signale liefern, sind sie sehr unbequem für den Probanden. Bei dieser Art von Elektroden besteht auch die Gefahr von Verletzungen und Verbreitung von Krankheiten. Deswegen haben wir uns von der Benutzung dieser Elektroden bei der EMG-Sprachsynthese verzichtet.

Für weitere physiologischen Grundlagen der EMG und für Information über die Sprachproduktion bei Menschen wird der Leser auf [14, Seiten 3-17] verwiesen.

2.2 F0 Grundlagen

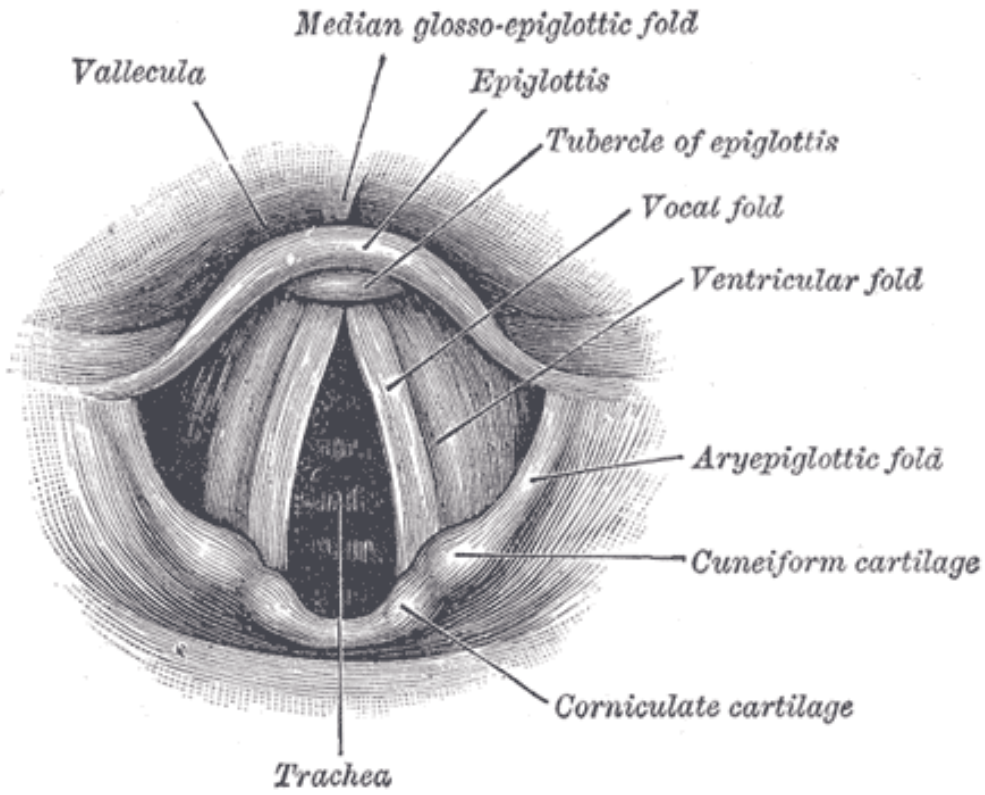
Resonanz bezeichnet die Neigung eines Objektes, in manchen bestimmten Frequenzen, mit größerer Amplitude zu oszillieren. Diese Frequenzen sind die resonanten Frequenzen des Objektes. Die physikalischen Eigenschaften des Objektes - wie Dichte und Größe - bestimmen die Resonanzfrequenz. Es ist leichter ein Objekt in seiner Resonanzfrequenz in Oszillation zu setzen als in irgendeiner anderen Frequenz. Folglich ist die Resonanzfrequenz die natürliche Oszillationsfrequenz von einem Objekt.

Jedes Objekt hat eine Resonanzfrequenz und manche Objekte besitzen sogar mehrere Resonanzfrequenzen. Die Fundamentalfrequenz (F0) bezeichnet die niedrigste von diesen Frequenzen. [1]

Der anatomische Sprachapparat bei Menschen können wir auch als ein physikalisches Objekt untersuchen, das auch Resonanzfrequenzen besitzt. Er beinhaltet die Stimmbänder[2], die Zunge, den Mundraum, den Rachenraum, den Nasenraum, die Zähne und die Lippen. Die Grundfrequenz der Sprache wird am meisten von den Stimmbändern beeinflusst, weil sie, mit seiner Vibration, den Grundton produzieren. Dieser Grundton wird danach von den anderen Sprachorganen moduliert, sodass unterschiedliche Laute geformt werden können. Wenn wir die F0 mit einer positiven ganzen Zahl multiplizieren, bekommen wir eine *Harmonische* (in der Musik auch Oberton). Der Mundraum, Nasenraum und Rachenraum, die als Resonanzräume agieren, heben manche von den Harmonischen besonders hervor. Diese Harmonischen werden *Formanten* genannt. In einem Spektrogramm sind die Formanten die Frequenzen mit höherer Amplitude - Bereiche mit Konzentration von akustischer Energie. Beim Sprechen verändert der Mundraum ständig seine Form. Entsprechend wird auch seine Resonanzfrequenz verändert und folglich werden unterschiedliche Formanten hervorgehoben, was zu unterschiedlichen Lauten führt. Dabei unterscheidet man zwei Typen von Lauten - Vokalen und Konsonanten. Bei den Vokalen steht der Mund ständig offen und die Stimmbänder vibrieren. Die Vokalen sind immer stimmhafte Laute und ihre Formanten sind üblicherweise deutlich in einem Spektrogramm erkennbar. Demgegenüber sind die Konsonanten Lauten, die durch die teilweise oder ganze Blockierung des Luftstroms entstehen. Die können stimmhaft oder stimmlos sein. Ein Konsonant ist stimmlos, falls er bei seiner Bildung nur den ausgeatmeten Luftstrom und nicht die Vibration der Stimmbänder benutzt. In einem Spektrogramm sind Formanten bei der Konsonanten sehr schwer oder gar nicht erkennbar.

Der Kehlkopf und die Stimmbänder werden von unterschiedlichen Skeletmuskeln bewegt und gespannt. Die Vagus Nerve ist für die Koordination dieser Muskeln

[1] F0 wird auch als die niedrigste Frequenz in einem Frequenzgemisch definiert.
[2] Auch Stimmlippen genannt.



[h]

Abbildung 2.2: Stimmbänder im Kehlkopf (Larynx), aus [1]

verantwortlich. Die komplexe Kehlkopfmuskulatur wird von den folgenden Muskeln gebildet:

- *cricothyroideus* - spannt die Stimmbänder
- *cricoarytaenoideus posterior* - öffnet die Stimmritze (die Distanz zwischen den Stimmbändern)
- *cricoarytaenoideus lateralis* - schließt die Stimmritze
- *cricoarytaenoideus thyroarytaenoideus* - schließt die Stimmritze
- *cricoarytaenoideus arytaenoideus transversus* - schließt die Stimmritze
- *cricoarytaenoideus arytaenoideus obliquus* - schließt die Stimmritze
- *cricoarytaenoideus thyroepiglotticus* - adduziert die Stimmbänder
- *cricoarytaenoideus aryepiglotticus* - adduziert die Stimmbänder
- *cricoarytaenoideus vocalis* - bewirkt Eigenspannung der Stimmbänder und verschließt die Stimmritze

Also das Produzieren des Grundtons der Stimme ist mit vielen Muskelaktivitäten verbunden. Das schlägt, zumindest theoretisch, vielversprechende Möglichkeiten für elektromyographische F0-Schätzung vor.

Die F0, die auch Einfluss auf das Timbre hat, variiert zwischen unterschiedlichen Sprechern. Es ist auch eine deutliche Trennung zwischen den zwei Geschlechtern zu beobachten. Grund dafür ist, dass die Länge der Stimmbänder für die F0 eines Sprechers entscheidend ist. Bei Frauen sind die Stimmbänder zwichen $12,5mm$ und $17,5mm$ lang. Bei Männern beträgt diese Länge $17,5mm$ - $25mm$, was zu tieferer Fundamentalfrequenz führt. Die F0 bei Frauen bewegt sich in dem Frequenzband $165Hz$ - $255Hz$ und bei Männern liegt die Fundamentalfrequenz im Intervall $85Hz$ - $180Hz$. In Abbildung 2.3 sind die durchschnittlichen Werte für F0 von unterschiedlichen Sprechern dargestellt.

Schall	**F0**
Baby Weinen	500Hz
Sprache bei Kindern	250Hz-400Hz
Sprache bei adulten Frauen	200Hz
Sprache bei adulten Männern	125Hz

Abbildung 2.3: Fundamentalfrequenzen bei unterschiedlichen Sprechern

Es existiert auch eine starke Variation in F0 zwischen den unterschiedlichen Sprachen. In [25] wurde gezeigt, dass F0 Werte bei Männern, die *Wu* chinesischen Dialekt sprechen, ungewöhnlich hoch sind - $170Hz$. Also kann man behaupten, dass die kulturelle Umgebung und selbst die Sprache deutlich die F0 beeinflussen.

Die Fundamentalfrequenz hängt auch von der Stimmung des Sprechers ab. Emotionale Faktoren verursachen eine Steigerung des Muskeltonus, was zur Steigerung der F0 führt. Zum Beispiel weisen Sprecher bei einer Rede, Schauspielerei oder Vortrag vor Publikum eine höhere F0 auf als bei einer Unterhaltung im Freundeskreis auf. Es wurde auch festgestellt, dass deprimierte, traurige oder beschämte Sprecher sehr kleine F0-Variation zeigen. Im Gegensatz dazu haben emotional aufgeregte Sprecher, die Überraschung, Freude oder Ärger fühlen, große Variation in der F0 ihrer Äußerungen. Die Variation von F0 ist stark mit der Prosodie und dem Satzrythmus der Sprache verbunden.

Zusamenfassend kann man sagen, dass F0 von viele Faktoren abhängt. Die Sprache, der Typ der Rede (Vortrag, Unterhaltung) und die emotionale Lage des Sprechers haben Einfluss auf die Fundamentalfrequenz des Sprechers.

Von einer präzisen F0-Schätzung kann die Mensch-Maschine Kommunikation viel profitieren. Da in der F0 auch die Stimmung des Sprechers steckt, kann die Maschine durch eine korrekte F0-Schätzung diese nachvollziehen und sich entsprechend anpassen. Das bietet eine Kommunikation an emotionalem Niveau, etwas was für die Mensch-Maschine Interaktion noch ganz fremd ist.

Das richtige Synthetisieren der Fundamentalfrequenz ist für die Sprachsynthesesysteme sehr wichtig. Es macht die produzierten Äußerungen lebendig, nicht mechanisch klingend und nicht monoton. Das anspruchsvolle Reproduzieren von variierender F0, wird die Qualität der lautlosen Mensch-Mensch-Kommunikation stiegern.

2.3 Grundlagen der EMG-basierten Sprachsynthese

Im Folgenden werden die Techniken erläutert, die in dem Sprachsynthese-Prozess in dieser Arbeit vorgenommen wurden. Die Unterabschnitten beschreiben die Schritten der EMG-zu-F0 Mapping.

2.3.1 Datenaufnahme und Elektrodenposition

In diesem Abschnitt beschreiben wir die verwendete Aufnahmekonfiguration. Die EMG-Signale wurden mit Elektrodenarrays aufgenommen, die an der Hautoberfläche geklebt wurden. Die Arrays beinhalten 8 Elektroden in einer Reihe, mit einem Abstand von $0,5mm$ zwischen den einzelnen Elektroden. Die Untersuchungen wurden mit einer *monopolaren* Elektrodenkonfiguration durchgeführt. Als Referenz wurde eine EKG-Elektrode am Nacken des Sprechers benutzt. Die Aufnahmen wurden mit den folgenden zwei Einstellungen gemacht:

1. Mit zwei 8-Elektrodenarrays - einer am Kehlkopf und einer an der Wange neben dem Mundwinkel
2. Mit drei 8-Elektrodenarrays - einer am Kehlkopf und zwei an der Wange

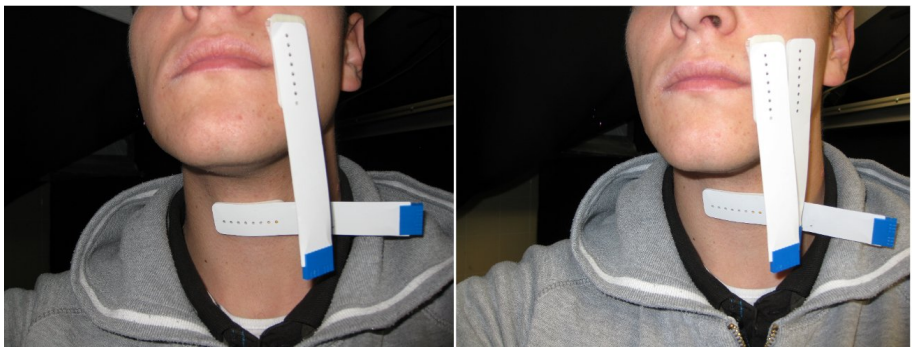


Abbildung 2.4: Die zwei Elektrodenkonfigurationen, die für die Aufnahmen verwendet wurden

Bei den Experimenten wurde festgestellt, dass man bessere Ergebnisse erzielt, wenn der Halselektrodenarray gerade am Kehlkopf steht. Eine Positionierung am Unterkiefer hat zu schlechterer F0-Schätzung geführt. Der Array, der am Wange steht, muss möglichst nah an dem Mundwinkel geklebt werden sein.

Das aufgenommene EMG-Signal wird zuerst verstärkt und mit einem Bandpassfilter gefiltert. Die Frequenzen zwischen 10Hz und 500Hz werden durchgelassen. Damit man das Analogsignal im Rechner repräsentieren und verarbeiten kann, wird es durch Sampling und Quantisierung digitalisiert. Man muss das Nyquist-Shannon Abtasttheorem beachten. Um ein Analogsignal fehlerlos zu digitalisieren, muss die Abtastfrequenz größer als das Doppelte der größten Frequenz des Signals sein. Wenn die größte Frequenzkomponente in unserem Analogsignal F_{max} ist und die Abtastfrequenz F_{abt} ist muss also:

$$F_{abt} > 2 * F_{max} \tag{2.1}$$

Falls die Abtastfrequenz diese Bedingung nicht erfüllt, entsteht Aliasing und man kann aus dem digitalen Signal das ursprüngliche Signal nicht fehlerlos rekonstruieren.

Für die Filterung, Verstärkung und Digitalisierung der EMG-Signale wurde den Rekorder EMG-USB2 von OT Bioelettronica [2] verwendet. Die Elektrodenarrays wurden auch von diesem Hersteller geliefert. Mehr Einzelheiten für die benutzte Hardware sind in dem Kapitel Implementierung erläutert.

2.3.2 Merkmalsextraktion und Fensterung

Auf den aufgenommenen EMG-Signalen wird eine Merkmalsextraktion durchgeführt. Ziel ist möglichst viele Informationen aus den EMG-Daten zu gewinnen, so dass die Abbildung auf Sprache, oder F0, danach so gut wie möglich akkurat ist.

Das *TD15* Featureset aus [23] wurde in den Mehrheit der Experimenten in dieser Arbeit verwendet. Das EMG-Signal wurde auf Stücke - auch *Fenster* genannt - zerlegt. Im *TD15* Featureset beträgt die Zeitdauer eines Fensters $27ms$. Das nächste Fenster wird gebildet, indem das jetzige Fenster um $10ms$ verschoben wird. Also die *Fensterverschiebung* beträgt $10ms$. Die Fensterung erfolgt so, dass das Signal mit einer Fensterfunktion multipliziert wird. Beispiele für Fensterfunktionen sind Hammingfenster und Rechteckfenster. Bei der Multiplikation gehen manche Frequenzen aus dem ursprünglichen Signal verloren (es exisitiert keine perfekte Fensterfunktion). Die Überlappung der Fenster sorgt für die Bewahrung dieser Frequenzen.

Das über 9 Werte doppelt gemittelte Signal $w[n]$ ist für ein EMG-Signal mit normaliesierten Mittelwert $x[n]$, definiert als

$$w[n] = \frac{1}{9} \sum_{k=-4}^{4} v[n+k], \text{ mit } v[n] = \frac{1}{9} \sum_{k=-4}^{4} x[n+k] \tag{2.2}$$

Diese Art von Filterung ermöglicht das hochfrequente Signal als $p[n] = x[n] - w[n]$ zu definieren.
Der Absolutbetrag des hochfrequenten Signals ist entsprechend $r[n] = |x[n] - w[n]|$.

Die Merkmalsextraktion in der *TD15* wird fensterweise durchgeführt. Die folgenden fünf Zeitbereich-Merkmale werden für jedes Fenster extrahiert (dabei sei N die Anzahl der Werte in dem gegebenen Fenster):

- Mittelwert von dem gefilterte Signal $w[n]$

$$\overline{w} = \frac{1}{N} \sum_{n=0}^{N-1} w[n] \tag{2.3}$$

- Energie von dem gefilterte Signal $w[n]$

$$P_w = \sum_{n=0}^{N-1} |w[n]|^2 \tag{2.4}$$

- Mittelwert von dem Absolutbetrag des hochfrequenten Signal $r[n]$

$$\overline{r} = \frac{1}{N}\sum_{n=0}^{N-1} r[n] \tag{2.5}$$

- Energie des hochfrequenten Signales $p[n]$

$$P_r = \sum_{n=0}^{N-1} |p[n]|^2 \tag{2.6}$$

- Null-Durchgangsrate - die Anzahl der Nulldurchgänge von hochfrequenten Signal $p[n]$

$$z_p = \text{ Anzahl der Nulldurchgänge in } (p[0], p[1], ..., p[N-1]) \tag{2.7}$$

Ein Schritt für Kontextsensivität ist die Zusammenfassung der Werte einzelner Merkmale für benachbarten Fenster. $S(\mathbf{f}, n)$ bezeichnet die Stapelung von $2n+1$ Fenster (von $-n$ bis n) für ein gegebenes Merkmal $\mathbf{f}$. Das *TD15* Featureset ist definiert als

$$\mathbf{TD15} = S(\mathbf{f2}, 15), \text{ mit } \mathbf{f2} = [\overline{w}, P_w, \overline{r}, P_r, z_p] \tag{2.8}$$

Eines der Ziele dieser Arbeit ist ein besseres Featureset zu finden, das für die Arrays mehr geeignet ist. Die Ergebnisse dazu werden in dem Kapitel *Neues Featureset* vorgestellt.
Außerdem werden auch Experimente mit unterschiedlichen Fensterlängen und Fensterverschiebungen durchgeführt. In dem Kapitel *Untersuchungen der Fensterlänge* sind die Resultate dazu erläutert. *TD15* war das verwendete für diese Experimente Featureset.

2.3.3 Merkmalsreduzierung

Die Merkmalsextraktion wird für alle aufgenommenen Signalkanäle durchgeführt. So wird für jedes einzelnes Fenster einen Merkmalsvektor(Featurevektor) gebildet. Die Dimensionalität dieses Vektors ist durch die folgende Gleichung beschrieben:

$$\text{Dimensionen} = \text{Anzahl Kanäle * Anzahl Merkmale * Kontextweite} \tag{2.9}$$

Hier ist die Kontextweite die Anzahl der gestapelten Fenster. In unserem Fall mit zwei 8-Elektrodenarrays und dem Featureset *TD15* haben wir $16 * 5 * 31 = 2480$ Dimensionen. Mit der zweiten Elektrodenkonfiguration, die wir verwenden - drei 8-Elektrodenarrays, besitzt der Merkmalsvektor sogar $24 * 5 * 31 = 3720$ Werte.

Um die Daten effizient weiter zu verarbeiten ist eine Technik für Reduzierung der großen Dimensionalität notwendig. Ziel ist die Dimensionsanzahl zu reduzieren, wobei gleichzeitig die charakteristische Information des Merkmalsvektors behalten wird. Die Anzal der Dimensionen wurde in dieser Arbeit, bei den beiden Elektrodenkonfigurationen, auf 32 Werte pro Fenster reduziert.

Für diese Aufgabe wurde die Linear Discriminant Analysis (LDA) verwendet. Zuerst werden hierbei die Daten in unterschiedlichen Klassen geordnet. Die einzelnen Klassen ergeben sich aus Labels, die vorher berechnet sind. Danach wird eine lineare Transfomationsmatrix (LDA-Matrix) berechnet. Diese Matrix bildet so die unterschiedlichen Klassen ab, dass sie möglichst deutlich voneinander trennbar sind. Gleichzeitig werden die Samples einer Klasse möglichst "eng" zu einander abgebildet. Bei der Multiplikation von dem ursprünglich hochdimensionalen Merkmalsvektor mit dieser Matrix wird die Dimensionsanzahl reduziert und ein neuer Merkmalsvektor gebildet.

Im Folgenden wird anhand eines einfachen Beispieles die Einzelheiten über die Berechnung der LDA-Matrix erläutert.

Am Anfang haben wir eine Menge von mehrdimensionalen Samples $\{x_1, x_2, ..., x_N\}$, die in Einfachheit halber, in zwei Klassen C_1 und C_2 unterteilt sind (für mehrere Klassen sind die Berechnungen analog). Die Anzahl der Samples, die zu der Klasse C_i gehören bezeichnen wir mit N_i, wo $i \in \{1, 2\}$.
Wir wollen unsere urspünglichen Samples auf eine Dimension reduzieren. Das heißt nach der Transformierung werden die Samples Punkten auf einer Gerade sein. Also suchen wir ein W^T, das ist unsere LDA-Matrix, so dass

$$y_i = W^T x_i, \text{ mit } i \in \{1, .., N\} \tag{2.10}$$

und y_i einen Skalar ist. Der Mittelwert innerhalb einer Klasse ist definiert durch

$$\mu_i = \frac{1}{N_i} \sum_{x \in C_i} x \tag{2.11}$$

Weiter sei S_W die Diskriminanz Matrix innerhalb der Klassen (within-class scatter matrix). Diese repräsentiert die Streuung der Daten in den Klassen.

$$S_W = S_1 + S_2, , \text{ mit } S_i = \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^T \tag{2.12}$$

S_B ist die Diskriminanz Matrix zwischen den Klassen (between-class scatter matrix).

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{2.13}$$

Die LDA Matrix w^T wird durch die Maximierung der Funktion $J(w)$ gewonnen:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{2.14}$$

2.3.4 Akustische Merkmalsextraktion

Da wir die EMG-Signale auf Audiosignale abbilden wollen, findet parallel zu der EMG-Aufnahme auch eine Audioaufnahme statt. Aus dem Audiosignal werden auch die LDA-Labels gebildet. Aus den gespeicherten Audiodateien werden die folgenden Merkmale extrahiert:

- F0 - das sind die Zielwerte für die Fundamentalfrequenz. Diese werden durch Fixed-Point Analyse ermittelt.
- Logarithmiertes F0.

Für eine Sprachsynthese, bei der Ende ein Audiosignal produziert wird, sind natürlich nur die F0-Werte nicht ausreichend. Deswegen wurden in diesem Fall auch die Energie und die Mel-Frequenz-Cepstrum-Koeffizienten als akustische Merkmale aus dem Audiosignal extrahiert.

Ein Markersignal wird in der Audio- und in der EMG-Datei gespeichert. Die Aufgabe dieses Signals ist das Audio- und das EMG-Signal zu synchronisieren. Anhand dieses Markersignals werden danach die Dateien geschnitten, sodass die EMG-Datei und die Audiodatei genau die selbe Zeitdauer besitzen. Jedoch existiert eine geringe Verzögerung zwischen den beiden Signalen. [23] und [8] haben gezeigt, dass das EMG-Signal mit 40-50ms das EMG-Signal überholt. Die selben Fensterlänge und Fensterverschiebung aus der EMG-Mermalsextraktion wurden auch in der akustischen Merkmalsextraktion verwendet. So wird gewährleistet, dass die Anzahl der Fenster in den akustischen Merkmalen und EMG Merkmalen übereinstimmt. Das ermöglicht auch die zeitlich korrekte Abbildung von der EMG-Daten auf die Audiodaten.

In Abbildung 7.3 (Anhang) sind die parallel aufgenommenen Audio- und EMG-Signale für eine Äußerung zu sehen (Audiosignal oben, EMG unten). Die zwei Markersignale, in der Audiodatei und in der EMG-Datei, befinden sich entsprechend in dem zweiten und in dem siebzehnten Kanal. Der Sprung des Signals bezeichnet, sowohl in der Audio- als auch in der EMG-Datei, den Anfang der Aufnahme.

2.3.5 Training

Jede Session[1] in dieser Arbeit beinhaltet insgesamt 50 englische Sätze. Bei allen Experimenten wurden 45 von diesen Sätzen für das Training des Systems verwendet. Es wurden zwei Klassifizierungsalgorithmen im Training eingesetzt - Support Vector Machines (SVM) und Gaussian Mixture Model (GMM). Zuerst werden die SVM trainiert um stimmhafte und stimmlose Fenster in der Äußerungen zu unterscheiden. Diese Strategie wurde vorgenommen, weil die F0 nur in den stimmhaften Phonemen vorhanden ist. Danach werden die stimmhaften Fenster, mittels der schon trainierten SVM, genommen. Nur auf diesen Fenstern werden das GMM für F0-Schätzung trainiert. Auf der Abbildung 2.5 ist der Trainingprozess graphisch dargestellt. *Reference* dabei sind die Werten für F0 aus den akustischen Merkmalen.

2.3.5.1 SVM und GMM

Die SVM ist ein Klassifikator, der besonders geeignet für Klassifizierung von Datenmengen mit zwei Klassen ist. Wenn wir uns die vorliegenden Datenvektoren als Punkten in einem Raum vorstellen, versucht die SVM eine Hyperebene zu finden, die die Punkten in zwei Klassen separiert. In einem zweidimensionalen Fall und wenn die Daten linear trennbar sind ist die Hyperebene also eine Gerade. Ein zweites Ziel ist, dass der Abstand (Margin) von den Punkten zu dieser Hyperebene maximiert

[1] Session - parallele Aufnahme von Audio- und EMG-Daten von einem sprechenden Probanden.

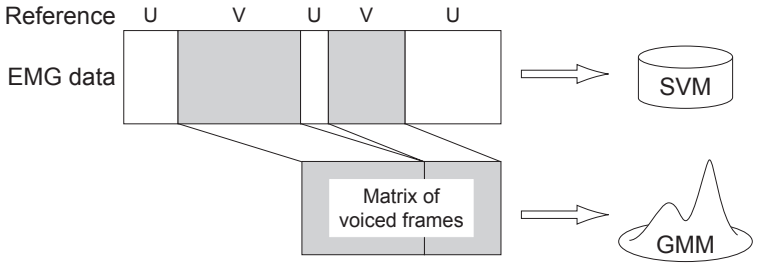


Abbildung 2.5: Der Trainingprozess. V und U steht entsprechend für voiced (stimmhaft) und unvoiced (stimmlos). Aus [16]

wird. Das hilft beim Testen, Objekte die sehr nah an der Klassengrenze liegen robust klassifizieren zu können. In der Praxis ist oft so, dass die zwei Klassen nicht linear trennbar sind. In diesem Fall wird der sogenannten *Kernel-Trick* eingesetzt. Die Trainingsvektoren werden dabei mit einer geeigneten Kernelfunktion in einen Raum mit höherer Dimensionalität überführt. In diesem Raum steigt die Wahrscheinlichkeit, dass die Klassen linear trennbar sind (Theorem von Cover). Wenn die Daten wieder nicht linear trennbar sind, werden die in einen Raum mit noch höherer Dimensionalität überführt. Falls eine Trennungshyperebene gefunden wird, werden die Datenvektoren und die Trennnungsebene zurück in den ursprünglichen Raum überführt. Die lineare Trenungsebene wird dann zu einer nichtlinearen Hyperfläche. Die SVM bekommt als Eingabe eine Menge von Training-Datenvektoren und die tatsächliche Einordnung dieser Datenvektoren zu den zwei Klassen.

$$\{(x_1, y_1), \ldots, (x_m, y_m) | x_i \in X, y_i \in \{-1, 1\}\} \tag{2.15}$$

In unserem Fall beinhaltet X alle Fenster des Trainingsets und -1 und 1 sind als stimmlos bzw. stimmhaft zu interpretieren (diese Information haben wir aus dem Audiosignal). Es wird die Entscheidungsfunktion 2.16 gebildet, wo w ein Normalenvektor ist und b ein Bias.

$$y_i = sgn(\langle w, x_i \rangle + b) \tag{2.16}$$

Um die Trennungsebene zu finden, wird das folgende Optimierungsproblem berechnet:

$$\text{Minimiere bezüglich } w \text{ und } \frac{1}{2}\|w\|_2^2, \text{ so dass } y_i(\langle w, x_i \rangle + b) \geq 1 \text{ für alle } 1 \leq i \leq m \text{ gilt.} \tag{2.17}$$

Aufgrund der stimmhaften Fenster wird das GMM trainiert. Das GMM ermöglicht eine komplexere Modellierung der Daten. Dieser Klassifikator hat die Möglichkeit die Daten in mehr als zwei Klassen einzuordnen. Die unterschiedlichen Klassen sind als einzelne Gauß-Verteilungen repräsentiert. Bei dem Training des GMMs haben wir eine Menge von Datenvektoren (Samples) $\{x_1, \ldots, x_N\}$, die alle zu einer Klasse gehören. Für diese Klasse bilden wir eine Gauß-Verteilung (eine Gauß-Glocke) in unserem GMM. Der Mittelwert und die Kovarianzmatrix dieser Verteilung optimieren wir so, dass die Verteilung alle Samples beinhaltet. Der Mittelwert μ und die Kovarianzmatrix $\sum$ der Verteilung sind dann gegeben durch:

$$\mu := \frac{1}{N}\sum_{n=1}^{N}(x_n) \tag{2.18}$$

$$\sum := \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T \tag{2.19}$$

Diese Abschätzung des Mittelwertes und der Varianz wird für alle Klassen durchgeführt. Die Klassifizierung eines Datenvektors x erfolgt unter zu Hilfenahme der Dichtefunktionen der Gauß-Verteilungen. Man berechnet den Wert der Dichtefunktion für x für jede Klasse. Der Datenvektor x wird zu dieser Klasse zugeordnet, wo der Wert der Dichtefunktion am größten ist.

Die SVM und das GMM werden mit einem iterativen EM-Algorithmus trainiert. Dieser Algorithmus besteht aus zwei Schritten. In dem E-Schritt (Expectation-Schritt) werden die Daten besser zu den einzelnen Klassen zugeordnet. Danach in dem M-Schritt (Maximization-Schritt) wird das Modell so verändert, dass es besser zu den neuklassifizierten Daten passt. Dann erfolgt wieder einen E-Schritt. Wenn das Model sich nur "wenig" verändert, terminiert der Algorithmus. Der Wissenschaftler muss "wenig" selbst definieren.

2.3.6 Testen

Das Testen des Systems wurde auf den restlichen 5 Sätzen durchgeführt. Zu bemerken ist, dass diese Sätze für das System völlig unbekannt sind, da diese im Training nicht genommen wurden. Die trainierte SVM und GMM wurden entsprechend für die Separierung der stimmhaften und stimmlosen Fenster und für die Schätzung der F0-Kontur eingesetzt.

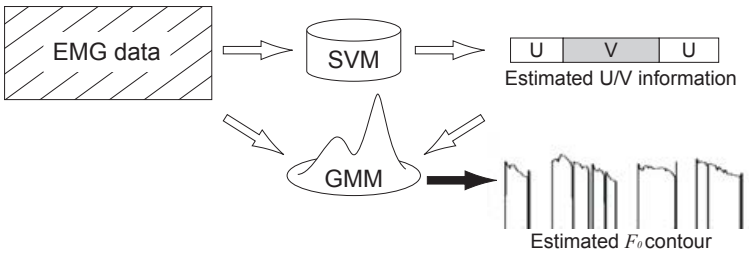


Abbildung 2.6: Der Testprozess. Aus [16]

Für die Evaluierung der Ergebnisse wurden zwei Maße berechnet.

1. Die Prozent der korrekt erkannten stimmlosen/stimmhaften Fenster.

2. Die Genauigkeit der F0-Erkennung - hier wurde der Korrelationskoeffizient zwischen geschätztem und tatsächlichem F0 ermittelt und auch die Standardabweichung dieses Korrelationskoeffizientes für die 5 Testäußerungen.

Der Korrelationskoeffizient zwei Zufallsvariablen X und Y ist in Formel 2.20 definiert. Wobei $Cov(X, Y)$ bezeichnet die Kovarianz zwischen X und Y und $\sigma()$ bezeichnet die Standardabweichung.

$$Korr(X, Y) := \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)} \tag{2.20}$$

In der Abbildung 2.7 ist die tatsächliche und die geschätzte F0-Kontur für eine Äußerung visualisiert. Der ausgesprochene Satz war "I think she just doesn't understand". Die Fenster, die keine Werte für F0 besitzen, sind dabei Teil von stimmlosen Phonemen. Die Korrelation wird auf allen Fenstern berechnet.

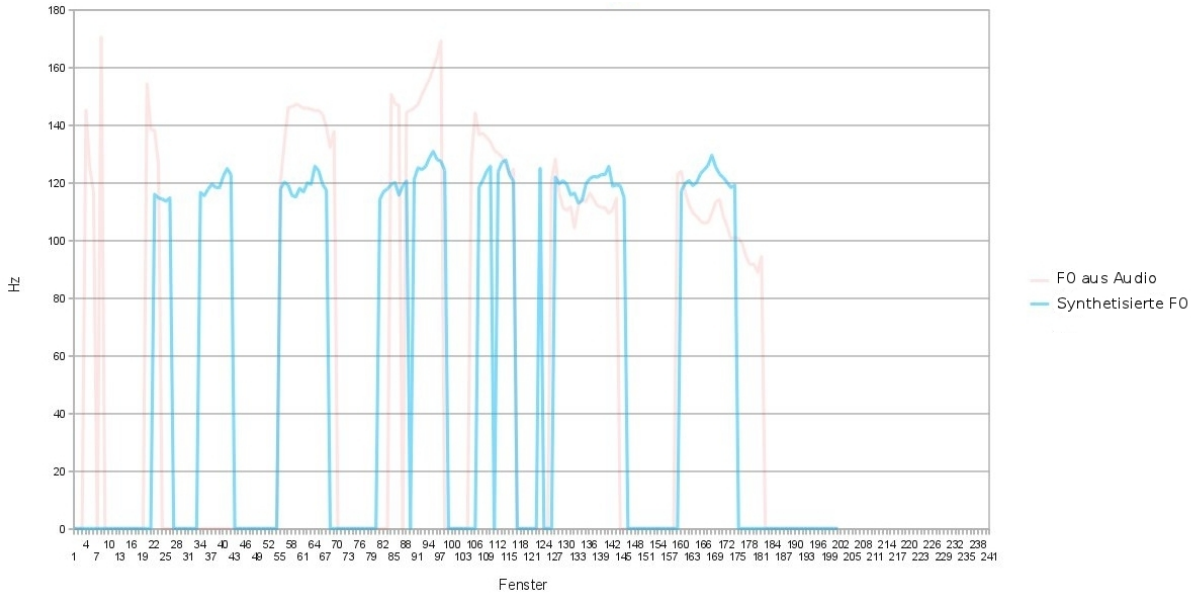


Abbildung 2.7: Tatsächlicher und geschätzter F0-Verlauf

2.4 Verwandte Arbeiten

Die Forschung bezüglich Fundamentalfrequenz in der herkömmlichen akustischen Sprache freut sich seit mehreren Jahren über das Interesse der Wissenschaftler. Es sind viele technische Berichte mit Untersuchungen dazu vorhanden. [25] präsentiert ausführliche F0-Ergebnisse für die unterschiedlichen Geschlechter, Altersgruppen und Sprachen. In [4] wurde die Beziehung zwischen der emotionalen Lage des Sprechers und der Fundamentalfrequenz bewiesen. Es existieren mehrere Methoden für die Extraktion von F0 aus der akustischen Sprache (eine davon ist in [12] beschrieben).

Das Suchen nach F0 in EMG-Signalen ist jedoch ein relativ neues Forschungsgebiet und derzeit existieren nicht viele wissenschaftliche Arbeiten oder Ergebnisse in dieser Richtung.
Diese Bachelorarbeit wurde auf den Ansätzen und den Ergebnissen aus [16] aufgebaut. In den beiden Arbeiten wurde das gleiche GMM-basierte Sprachsynthesesystem verwendet.

In [16] wurden 6 EMG-Kanäle mit einzelnen Ag-AgCl Elektroden erfasst. Die zum Trainieren und Testen benutzten Datenmengen sind wesentlich größer als hier. Die Experimente wurden über 3 Sprecher gemacht, wobei von jedem Sprecher 500 englische Äußerungen vorhanden waren. Das System wurde bei einem der Sprecher (spk-1) mit 380 der Sätze trainiert und mit den restlichen 120 Sätzen getestet. Bei den anderen zwei Sprechern (spk-2 und spk-3) wurden jeweils 250 Sätze für Training und Testen verwendet. Die Merkmalsextraktion hat man mit *TD15* Featureset gemacht und die Fensterlänge und die Fensterverschiebung waren entsprechend $27ms$ und $10ms$. Es wurden 2 Gauß-Glocken bei dem Training des GMMs eingesetzt. Am Anfang von Kapitel 5 stellen wir unsere Ergebnisse mit genau den selben Einstellungen vor. Die Resultate die in [16] erzielt wurden, sind in der folgenden Abbildung zu sehen.

Sprecher	**Korrelation**	**Standardabweichung**	**Stimmlos/stimmhaft Erkennungsrate [%]**
spk-1	0.49	0.19	84
spk-2	0.3	0.23	73
spk-3	0.25	0.18	82

Abbildung 2.8: Resultaten der F0-Schätzung aus [16]

.

3. Implementierung

Einer der Aufgaben dieser Arbeit war den Einsatz der Elektrodenarrays für EMG-basierte Sprachsynthese zu untersuchen. In diesem Kapitel werden technische Einzelheiten über diese Arrays und auch über die gesamte Hardware erläutert.

Der hier durchgeführte F0-Syntheseprozess besteht aus mehreren Schritten (siehe Kapitel Grundlagen dazu). Im Abschnitt 3.2 beschreiben wir die für diese Schritte eingesetzte Software.

3.1 Verwendete Hardware

Für die Aufnahme des EMG-Signals wurde der Rekorder EMG-USB2 von OT Bioelettronica verwendet. Mit dessen Hilfe wurde das EMG-Signal digitalisiert, gefiltert und verstärkt. Die Abtastfrequenz des Gerätes ist $2048Hz$, was eine Aufnahme mit hoher Auflösung ermöglicht. Die Signale wurden zwischen 10Hz und 500Hz gefiltert, da die EMG-Signale mit größter Energie sich in diesem Frequenzbereich befinden [6]. Die Verstärkung wurde mit Gain 500 gemacht.

Abbildung 3.1: Biosignal Recorder EMG-USB2, aus [19]

Die EMG-Signale wurden mit Elektrodenarrays aufgenommen. Jedes Array besitzt 8 in Reihe geordnete Elektroden. Die Distanz zwischen den einzelnen Elektroden beträgt $5mm$. Die Elektroden wurden mit Hilfe von Schaumstoffpads auf die Haut geklebt. Zur Verringerung des Elektrode-Haut-Widerstands wird ein leitendes Elektrolyt-Gel verwendet. In Abbildung 3.2 ist das in dieser Arbeit verwendete Elektrodenarray zu sehen.

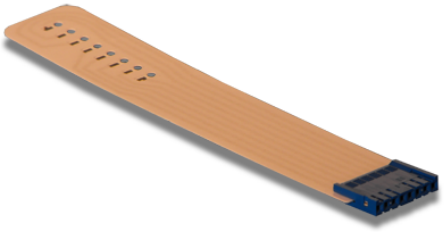


Abbildung 3.2: Elektrodenarray mit 8 Elektrode, aus [2]

Im Gegensatz zu den einzelnen Ag-AgCl Elektroden, sind die Elektrodenarrays schwierig am Gesicht anzubringen. In einer Session mit 50 Äußerungen, die ungefähr 10 Minuten dauert, kam es öfter zu einem Abkleben der letzten Elektrode des Arrays. Das verursachte entsprechend Verrauschung des Signals in diesem Kanal. Außerdem ist die Bewegungsfreiheit des Probanden, durch die große Verkabelung, sehr begrenzt, was zu unnatürliche Sprachweise führen kann. Drahtlose Elektroden würden an dieser Stelle sehr nützlich sein. Sie würden auch den praktischen Einsatz des hier beschriebenen Sprachsynthese-Systemes fördern.

Es wurden auch Aufnahmen mit dem in Abbildung 3.3 dargestellten 8x8 Elektrodenarray durchgeführt. Die Zwischenelektrodendistanz bei diesem Array ist $10mm$. Dieses Array war aber zu groß, unflexibel und konnte sich nicht an den komplexen Formen des Gesichtes anpassen. Es gab viele Elektroden, die keinen guten Kontakt mit der Haut hatten. Deswegen sind auch bei den Experimenten in dieser Arbeit nur Aufnahmen mit 8 Elektrodenarrays verwendet worden.

3.2 Verwendete Software

Für die Visualisierung der EMG-Signale während einer Aufnahme wurde die Software *OT BioLab* verwendet. Die Signale wurden mit der Software *Biosignalstudio* [26] aufgenommen. Diese Aufnahmeanwendung wurde am Cognitive Systems Lab Karlsruhe entwickelt. Die Anwendung besitzt eine modulare Architektur und wurde in der Programmiersprache Phyton programmiert. Für die Verwendung mit dem neuen Verstärker und mit den Arrays, musste jedoch die Software angepasst werden. Die Einzelheiten über die vorgenommenen Veränderungen sind in [20] beschrieben. Für die EMG-Merkmalsextraktion und LDA-Merkmalsreduzierung haben wir das Janus Recognition Toolkit eingesetzt [9]. Das Toolkit bietet ein Tcl Interface. Die akustischen Merkmale wurden mit dem Sprachsynthese-System aus [16] extrahiert. Dieses

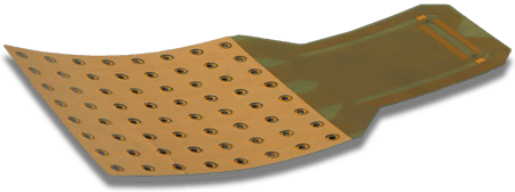


Abbildung 3.3: Elektrodenarray mit 64 Elektroden, aus [2]

System hat man auch für Training und Testen verwendet. Die einzelnen Verfahren zur Datenvorverarbeitung (PCA, ICA, SOBI) wurden mit *MATLAB* durchgeführt. Wir haben *MATLAB* auch für die Visualisierung und Signalanalyse der aufgenommenen Signale verwendet.

4. Der Datenkorpus

In diesem Kapitel wird der für die Untersuchungen verwendete Datenkorpus beschrieben. Die Daten wurden am Cognitive System Lab Karlsruhe aufgenommen. Alle 4 Probanden sind männlich, zwischen 22 und 29 Jahren. Jede Session besteht aus 50 englischen Sätzen aus Fernsehnachrichten. 45 von diesen Sätzen sind für Training verwendet und die restliche 5 für Testen. Die Testsätze waren dabei überall die folgenden:

- The state of Florida has a tough policy against ambulance chasing.
- The convicted murderer has avoided execution by lodging repeated appeals.
- Meanwhile Republicans are looking forward to separating the Senate agenda from the Presidential agenda.
- The administration announced steps to coordinate state and federal efforts.
- The national transportation safety board calls these canisters the number one issue in its investigation.

Zu bemerken ist, dass für alle Sprecher Englisch nicht die Muttersprache ist. Die Sprecher 702, 551 und 601 haben als Muttersprache Deutsch und Sprecher 701 Bulgarisch.

In Tabelle 4.1 sind die einzelnen Sessions mit der entsprechenden Elektrodenkonfiguration beschrieben. Details über die Elektrodenkonfiguration mit 6 einzelnen Elektroden sind in [14] erläutert. In der Tabelle ist auch die Zeitdauer der Trainings- und Testdaten zu sehen. Bei Session 007 von Sprecher 701 waren 6 Äußerungen fehlerhaft von der Aufnahmesoftware erfasst. Diese wurden bei dem Training- und Testprozess weggelassen. Wir können bemerken, dass die beiden Sessions von Sprecher 702 deutlich kürzer als bei den anderen Sprechern sind. Dieser Sprecher hat schneller gesprochen.

Es ist schwierig die hier präsentierten Ergebnisse mit [16] zu vergleichen, da dort eine deutlich größere Datenmenge sowohl für das Testen als auch für das Training verwendet wurde.

Sprecher	Session	Elektrodenkonfiguration	Dauer [sek]	
			Train	Test
Sprecher 701	*006*	2 x 8-Elektrodenarray	158	32
	007	3 x 8-Elektrodenarray	176	24
	008	3 x 8-Elektrodenarray	150	18
Sprecher 702	*003*	2 x 8-Elektrodenarray	111	17
	004	2 x 8-Elektrodenarray	109	14
Sprecher 551	*060*	2 x 8-Elektrodenarray	144	23
	063	3 x 8-Elektrodenarray	137	23
	006	6 einzelne Elektroden	140	24
Sprecher 601	*060*	2 x 8-Elektrodenarray	170	25
	061	3 x 8-Elektrodenarray	141	17
	005	6 einzelne Elektroden	180	19

Abbildung 4.1: Der in den Experimenten verwendete Datenkorpus.

Bei dem Aufnahmeprozess wurde festgestellt, dass die Signale merkbar störungsanfällig sind. Schlucken und Husten verrauschen deutlich die EMG-Kanäle aus dem Kehlkopf. Das benutzte System ist auch gegen elektromagnetische Störung empfindlich. Das bloße Berühren einer PC-Tastaturtaste war in dem EMG-Signal bemerkbar.

Im Folgenden sind die F0-Frequenzbereiche von den vier untersuchten Sprechern abgebildet. Diese Werte wurden von den aufgenommenen Audiosignalen extrahiert. Dabei wurde für jeden Sprecher die F0 für alle Fenster in einer Session berechnet. Viele von den Fenstern sind ein Teil von stimmlosen Lauten und da ist entsprechend die Fundamentalfrequenz 0. Diese Fenster wurden aber in den Grafiken nicht dargestellt. Es ist merkwürdig, dass die F0 von Sprecher 702 in einem breiteren Frequenzband liegt. Das heißt, dass dieser Sprecher eine größere Variation von F0 besitzt.

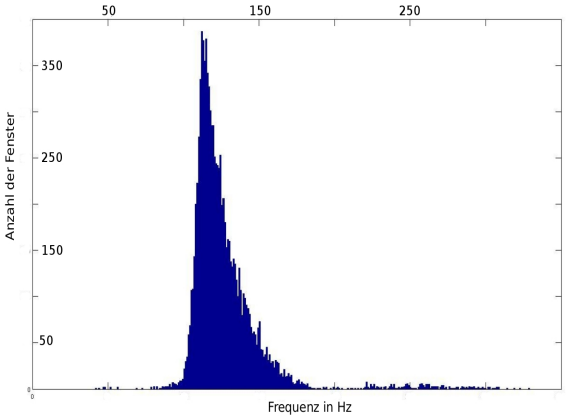


Abbildung 4.2: F0 von 701

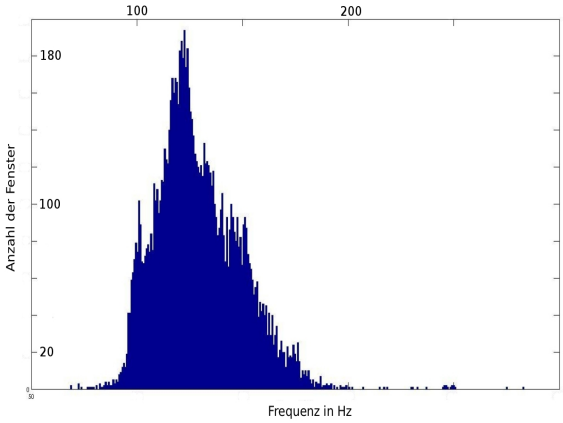


Abbildung 4.3: F0 von 702

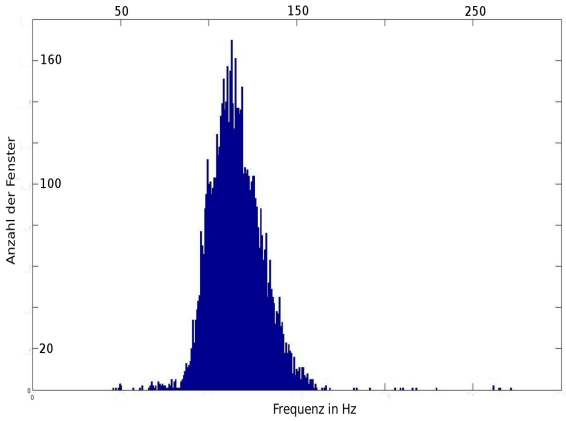


Abbildung 4.4: F0 von 551

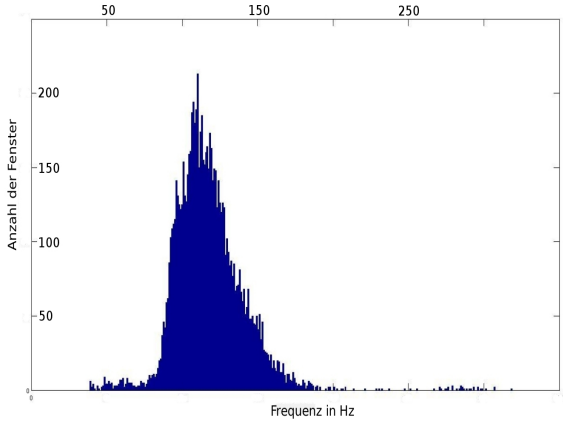


Abbildung 4.5: F0 von 601

5. Durchgeführte Experimente

In diesem Abschnitt werden die durchgeführten Untersuchungen beschrieben und kommentiert. Wir verändern unterschiedliche Parameter des Sprachsynthesesystems und evaluieren die erzielten Ergebnisse. Die Resultaten werden mit einer Baseline Konfiguration verglichen. Es wird auch der wissenschaftliche Hintergrund der genutzten Vorverarbeitungsmethoden (PCA, ICA und SOBI) kurz erläutert.

5.1 Baseline Ergebnisse

Zuerst präsentieren wir die Ergebnisse der F0-Schätzung bei einer Basiskonfiguration. Für die EMG-Merkmalsextraktion ist hier das originale *TD15* Featureset eingesetzt. Fensterlänge und Fensterverschiebung sind entsprechend $27ms$ und $10ms$ (wie in [23]). Das Gaussian Mixture Model wird mit 2 Gauß-Glocken trainiert. Die Ergebnisse in der Tabelle 5.1 sollten als Vergleichspunkt für die Evaluierung der anderen Experimente dienen. Sei im Folgenden die hier verwendete Einstellungen *ursprüngliche Konfiguration* genannt.

Unser Maß für die Qualität der F0-Schätzung ist die Korrelation der geschätzten F0-Kontur mit der aus dem Audiosignal extrahierten F0-Kontur, die wir als Ground Truth verwenden. Die Korrelation in der Tabelle ist der Mittelwert der F0-Korrelationen der 5 Testsätze. In der Abbildung 5.2 sind die Korrelationen für die unterschiedlichen Sessions noch mal grafisch dargestellt. Außerdem ist für die akkurate Prosodiegenerierung eine korrekte Unterscheidung von stimmhaften und stimmlosen Phonemen notwendig, wir trainieren deshalb zusätzlich eine SVM zur Unterscheidung diese beiden Phonemklassen und vergleichen die Resultate wieder mit der aus dem akustischen Signal extrahierten Ground Truth. Ziel ist, dass die Korrelation und die stimmlos/stimmhaft Erkennungsrate möglichst groß sind. Das würde eine gute F0-Schätzung bedeuten. Im Gegensatz dazu muss die Streuung der Ergebnisse aus den 5 Testsätzen möglichst klein sein. Also einen kleinen Wert für die Standardabweichung ist erwünscht.

Für alle Untersuchungen sind im Anhang in dem Unterabschnitt 7.5 Tabellen mit detaillierter Information über die stimmlos/stimmhaft Erkennungsraten vorhanden. Dort sind die folgenden vier Werte für jede Session in Prozent gegeben:

Sprecher	Session	Elektrodenkonfiguration	Ergebnisse		
			Korrelation	Standard-abweichug	Stimmlos/stimm-haft Erkennungsrate [%]
Sprecher 701	*006*	2 x 8-Elektrodenarray	0.061	0.053	53.6
	007	3 x 8-Elektrodenarray	-0.01	0.046	50.4
	008	3 x 8-Elektrodenarray	0.03	0.071	47
Sprecher 702	*003*	2 x 8-Elektrodenarray	0.184	0.08	59.6
	004	2 x 8-Elektrodenarray	0.255	0.031	64.6
Sprecher 551	*060*	2 x 8-Elektrodenarray	0.21	0.076	61.7
	063	3 x 8-Elektrodenarray	0.257	0.08	64
	006	6 einzelnen Elektroden	0.485	0.062	77.6
Sprecher 601	*060*	2 x 8-Elektrodenarray	0.346	0.058	68.4
	061	3 x 8-Elektrodenarray	0.476	0.103	73.5
	005	6 einzelnen Elektroden	0.515	0.077	78.7

Abbildung 5.1: Die ersten Ergebnisse der F0-Schätzung mit der *ursprünglichen Konfiguration*

1. Fenster erkannt als stimmhaft → sind tatsächlich stimmhaft
2. Fenster erkannt als stimmhaft → sind tatsächlich stimmlos
3. Fenster erkannt als stimmlos → sind tatsächlich stimmhaft
4. Fenster erkannt als stimmlos → sind tatsächlich stimmlos

Die stimmlos/stimmhaft Erkennungsrate, die richtig erkannten Fenster, setzt sich aus den Prozenten in *1* und *4* zusammen. Die Summe der beiden Prozentanteile ergibt die gesamte stimmlos/stimmhaft Erkennungsrate.

In [16] wurde eine andere Eveluierungsstrategie verwendet, wobei die F0-Korrelationen nur auf den als stimmhaft klassifizierten Fenstern berechnet wird (F0 liegt nur in den stimmhaften Fenstern). Falls wir diese Strategie auch auf unserem Datencorpus einsetzen, bekommen wir deutlich kleinere Korrelationen als diese in der Tabelle 5.1. Die Resultate dazu sind in der Tabelle 7.1 präsentiert. Auch da wurde die *ursprüngliche Konfiguration* genutzt. Die stimmlos/stimmhaft Erkennungsrate bleibt unverändert. Die kleinere Korrelationskoeffizienten, die wir mit dieser Eveluierungsstrategie bekommen, sind leicht zu erklären. Wegen dem kleinen Trainingsset und Testset, die entsprechend lediglich 45 bzw. 5 Sätze sind, liefert die SVM Klassifikation niedrige stimmlos/stimmhaft Erkennungsraten. Wie wir aus 5.1 entnehmen können, liegt die durchschnittliche Erkennungsrate von den 11 Sessions bei nur 63.5%. Das bedeutet, dass wenn wir die Menge der erkannten als stimmhaft Fenster betrachten, haben wir viele Fenster drin, die falsch klassifiziert wurden (erkannt als stimmhaft, sind aber in der Tat stimmlos). Alle diesen falsch erkannten Fenster bekommen von dem GMM einen Wert für F0 zugewiesen. Die Fenster, mit den man vergleicht, sind aber stimmlos und haben deswegen einen Wert für F0 von 0Hz. Die große Anzahl der falsch erkannten stimmlosen Fenster ist der Grund für die sehr kleine Korrelationen

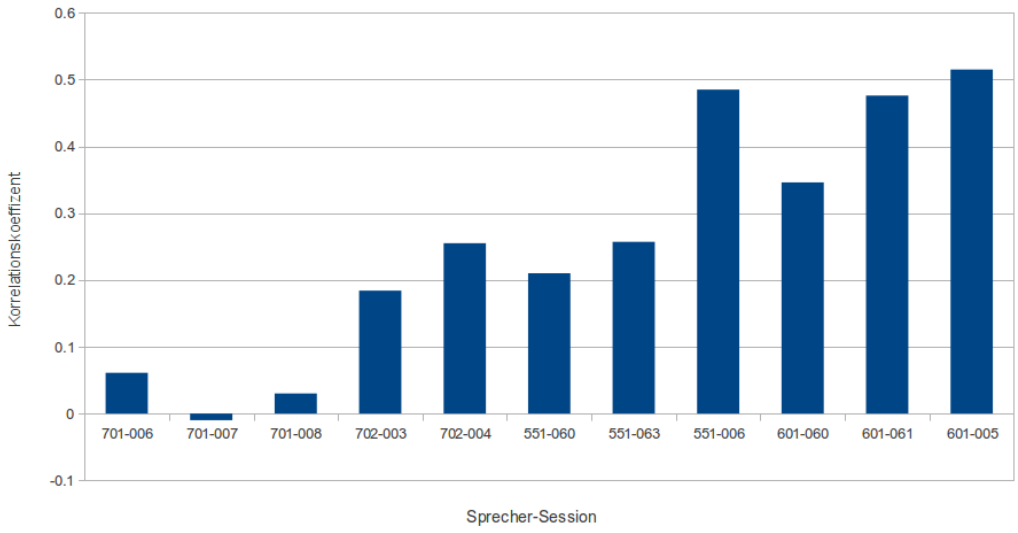


Abbildung 5.2: F0-Korrelationskoeffizenten für die unterschiedlichen Sessions

in 7.1. Wenn wir die Korrelationen auf allen Fenstern berechnen, wird der negative Effekt der falsch erkannten stimmhaft Fenster gemildert und wir bekommen größere (bessere) Korrelationen. Das Ganze ist noch einen Beweis dafür wie wichtig die SVM Klassifikation für das Endergebnis ist. Wenn die stimmlos/stimmhaft Erkennungsrate nicht präzis ist, arbeitet das GMM auf den falschen Fenstern und folglich sind die Korrelationen sehr niedrig.

Konzentrieren wir uns nun auf die Zahlen in 5.1. Wir können feststellen, dass eine deutliche Varianz der Resultaten sowohl zwischen den Sprechern als auch zwischen den einzelnen Sessions existiert. Man kann sehen, dass der Sprecher 701, bei allen seinen Sessions, deutlich schlechtere Ergebnisse aufweist. Die stimmhaft/stimmlos Erkennungsraten bei ihm sind sehr niedrig und wie beschrieben hat das eine entscheidende Bedeutung für die Korrelationsevaluierung. Demgegenüber erzielt der Sprecher 601 die besten Resultate. Sei es mit zwei Arrays, drei Arrays oder mit einzelnen Elektroden, weisen die Sessions von ihm vergleichsmäßig die besten Korrelationen. Dieser Sprecher hat auch die größte Erfahrung mit EMG-Experimenten. Das könnte auch Einfluss auf die Resultate haben. Es fällt auf, dass die zwei Sessions, wo einzelne Elektroden verwendet wurden (551-006 und 601-005), die beste Resultate erzielen. Die Werte in diesen Sessions sind sowohl für die Korrelation als auch für die stimmlos/stimmhaft Erkennungsrate signifikant besser.

Anhand der Daten aus der Tabelle 5.1 können wir noch schließen, dass die Ergebnisse sich für die Aufnahmen mit 2 Arrays und 3 Arrays unterscheiden. Trotzdem können wir nicht sofort festlegen welche von den beiden Konfigurationen besser geeignet ist. Die Ergebnisse in dieser Richtung widersprechen sich zwischen den einzelnen Probanden. Bei Sprecher 701 ist die Session mit 2 Arrays besser als die mit 3 Arrays. Bei den Sprechern 551 und 601 beobachten wir allerdings die umgekehrte Situation. Allgemein aber sind die Korrelationen bei dem Sprecher 701 extrem klein und deswegen können wir die Resultate bei ihm nicht für besonders aussagekräftig halten. Nach Analyse der zugrunde liegenden EMG-Signale, ist unsere Vermutung, dass bei diesem Sprecher der zweite Array an der Wange zu weit weg von dem Mund liegt.

Folglich wurde an den Kanälen dieses Arrays nicht viele Muskelaktivität gemessen. Anderseits liegen alle Arrays bei den Sprechern 551 und 601 auf Muskeln, die bei der Sprachproduktion beteiligt sind. Als Folge enthalten alle Kanäle wichtige EMG-Information. Wir können schlussfolgern, dass die Elektrodenpositionierung entscheidende Bedeutung für die Ergebnisse hat. Wenn die Arrays korrekt positioniert sind, wie bei 551-063 und 601-061, erzielt man mit drei Arrays bessere Korrelationen als mit zwei.

5.2 Anzahl der Training-Gauss-Mixturen

In diesem Abschnitt untersuchen wir den Einfluss der Anzahl der Gauß-Mixturen (auch Gauß-Glocken gennant) bei dem Training auf die Resultate. In dem Diagramm 5.3 sind die Ergebnisse mit 2, 4 und 8 Gauß-Glocken für 9 Sessions zu sehen. Bei den restlichen zwei Sessions (701-008 und 551-063) gab es Probleme beim Training mit mehreren Gauß-Glocken. Die Training-Software stürzte ab. Deswegen sind keine Ergebnisse für diese Sessions vorhanden. Das verwendete Featureset ist *TD15* und die Fensterlänge und Fensterverschiebung sind entsprechend $27ms$ und $10ms$. In der Tabelle 7.2 sind die Resultate aus dem untenstehenden Diagramm noch mal detailliert beschrieben.

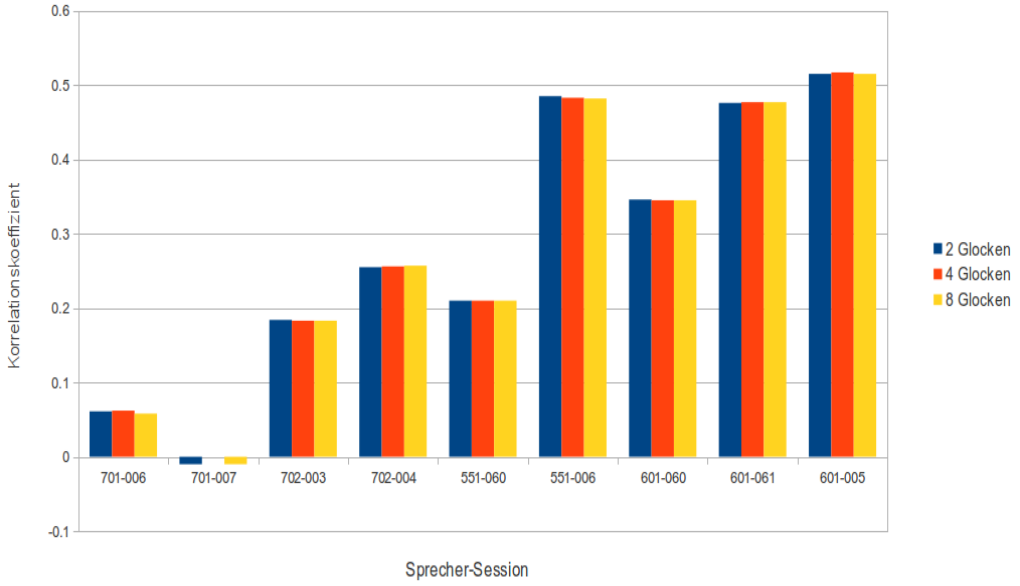


Abbildung 5.3: Einfluss der Anzahl der Gauß-Glocken bei der F0-Schätzung

Überall hat die Anzahl der Gauß-Glocken praktisch keine Auswirkung auf die Ergebnissen. Wir können jedoch nicht sagen, ob diese Tendenz bei mehreren Train- und Testsätzen auch bleibt. Unsere Hypothese war, dass die sehr eingeschränkte Trainingsmenge, lediglich 45 Sätze, die Ursache für die gleichen Resultate ist. Um eine deutlichere Tendenz festzustellen, müsste man die Experimente mit größeren Trainingsmengen und mit mehreren Gauß-Glocken durchführen.

Das GMM hat Zusammenhang nur mit der F0-Schätzung. Deswegen hat die Anzahl der Glocken keinen Einfluss auf die stimmlos/stimmhaft Erkennungsrate. Die bleibt bei allen Sessions hier unverändert.

5.3 Neues Featureset

Eins der Ziele dieser Arbeit war nach einem neuen Featureset zu suchen und dieses im Hinblick auf die F0-Schätzung zu evaluieren. Das *TD15* Featureset beinhaltet Merkmale nur aus dem Zeitbereich des EMG-Signals. Features aus dem Spektrum des Signals wurden bei der F0-Schätzung noch nicht eingesetzt. Das war die Motivation spektrale Merkmale zu untersuchen. Die vorgenommene Strategie war nicht ein komplett neues Featureset zu erzeugen, sondern manche Merkmale aus *TD15* zu verwenden und mit spektralen Merkmalen zu ergänzen.

Als neues Merkmal wurde die dominante Frequenz jedes Fensters genommen. Zuerst wird durch diskrete Fourier-Transformation das Spektrum des EMG-Signals gebildet, und zwar fensterweise. Für jedes Fenster wird ein Array gebildet mit 32 Koeffizienten, wobei ein Koeffizient für einen bestimmten Frequenzbereich steht. Dann wird der Index der Arrayzelle genommen, wo der Wert am größten ist. So wird die Frequenz mit der größten Amplitude aus dem Spektrum genommen. Also es wird die am meisten ausgeprägte Frequenz für jedes Fenster als Feature verwendet. Dieses neue Merkmal wird zu den 4 von den Merkmalen der *TD15* Featureset addiert. Die neu entstandene Featureset nennen wir *TD15-neu*. Die beinhaltet die folgende fünf Merkmale:

- Mittelwert des Signals
- Energie des Signals
- Energie des hochfrequenten Signals
- Null-Durchgangsrate des hochfrequenten Signals
- Dominante Frequenz des Signals

Der Unterschied zu der originalen *TD15* Featureset besteht darin, dass das Merkmal *Mittelwert des hochfrequenten Signals* mit dem Merkmal *dominante Frequenz* ersetzt wurde. Auch hier wurde eine Stapelung mit 31 Fenstern verwendet. In der Tabelle 5.4 sind die mit der neuen Featureset erzielten Ergebnisse beschrieben. Die Fensterlänge und die Fensterverschiebung sind entsprechend $27ms$ und $10ms$. Die Anzahl der Gaußglocken ist 2. In der Abbildung steht "v/u ER" für voiced/unvoiced (stimmhaft/stimmlos) Erkennungsrate.

Bei der Mehrheit der untersuchten Sessions führt die *TD15-neu* zur Verbesserung oder zu gleichen Resultaten bei der Korrelation. Die Korrelation bei Session 005 des Sprechers 601 hat sogar den Wert 0.53 erreicht. Das ist das beste Resultat, das überhaupt in dieser Arbeit erzielt wurde. Durchschnittlich ist eine Verbesserung in allen drei Kriterien - Korrelation, Standardabweichung, v/u Erkennungsrate - vorhanden. Durchschnittlich haben sich die Korrelation und die stimmlos/stimmhaft Erkennungsrate entsprechend um 5% und um 1% verbessert. Dabei wurde die durchschnittliche Verbesserung anhand der absoluten Verbesserung und nicht anhand der relativen Verbesserung berechnet, da sonst Session 701-006 sehr großen Einfluss auf das Endsresultat hätte. Diese Vorgehensweise wurde auch bei allen nachfolgenden Experimenten eingesetzt.

Sprecher	Session	Ergebnisse			Absolute Verbesserung		Relative Verbesserung	
		Korr.	SA	v/u ER	Korr.	v/u ER	Korr.	v/u ER
Sprecher 701	*006*	0.178	0.089	59%	0.117	5.4%	192%	10%
	007	0	0.041	51%	0	0.6%	0	1%
Sprecher 702	*003*	0.18	0.09	59%	0	-0.6%	0	-1%
	004	0.244	0.075	64%	-0.011	0	-4%	0%
Sprecher 551	*060*	0.24	0.017	63%	0.03	1.3%	14%	2%
	063	0.237	0.044	63%	-0.02	-1%	-8%	-1%
	006	0.464	0.113	76%	-0.021	-1.6%	-4%	-2%
Sprecher 601	*060*	0.317	0.046	67%	-0.029	-1.4	-8%	-2%
	061	0.521	0.041	76%	0.045	2.5%	10%	3%
	005	0.53	0.083	78%	0.015	-0.7%	3%	-1%
Durchschnittlich							**5%**	**1%**

Abbildung 5.4: Ergebnisse mit *TD15-neu* Featureset, SA steht für "Standard Abweichung".

Wir müssen nicht vergessen, dass unsere Sessions nicht viele Information beinhalten. Die Ergebnisse mit mehr Training- und Testdaten würden interessant sein. Es wird vermutet, dass die neue Featureset da ihre Vorteile deutlicher zeigen wird. Man kann auch die *TD15-neu* Featureset für Spracherkennung untersuchen.

Es wurden auch andere Merkmalskombinationen analysiert. Beispiele dazu sind:

- die dominante Frequenz nicht an der Stelle des Mittelwertes des hochfrequenten Signals setzen, sondern an der Stelle von anderem Merkmal aus *TD15*.
- Featureset mit 6 Merkmalen - komplett *TD15* und dominante Frequenz dazu.
- Featureset mit 7 Merkmalen - komplett *TD15*, dominante Frequenz und durschnittliche Frequenz dazu.

Alle diese Varianten haben aber schlechtere Ergebnisse als die ursprüngliche *TD15* Featureset erwiesen. Bei der Untersuchung der Merkmale hat sich herausgestellt, dass manchmal ein neues Merkmal sehr stark mit einem von den alten Merkmalen korreliert. Solche Merkmale sind nicht erwünscht, da sie keine neue Information mit sich bringen und folglich zu keiner Verbesserung der Schätzung führen konnten. Ähnliche Merkmale bauen auch Singularität in den Matrizen bei der LDA-Berechnung ein. Das erschwert und verlangsamt die Kalkulation der LDA-Matrix. Mit steigender Anzahl der Features steigt auch exponentiell der Bedarf an Trainingsdaten, um eine bestimmte Klassifikationsgüte zu erreichen (*Curse of Dimensionality*). Mehrere Features führen also nicht unbedingt zu besseren Ergebnissen.

5.4 Untersuchungen der Fensterlänge

In diesem Abschnitt untersuchen wir den Einfluss der Fensterlänge auf die F0-Schätzung. In Abbildung 5.5 sind die Korrelationskoeffizienten für drei unterschiedliche Fensterlängen dargestellt. Dabei beträgt die Fensterverschiebung überall $10ms$.

Bei den Experimenten wurde die *TD15* Featureset verwendet und das GMM wurde mit 2 Gauß-Glocken trainiert. In diesem Sinne repräsentiert die Linie $27ms/10ms$ die *ursprüngliche Konfiguration*. In Tabelle 7.4 (Anhang) sind die Resultaten detailliert vorgestellt.

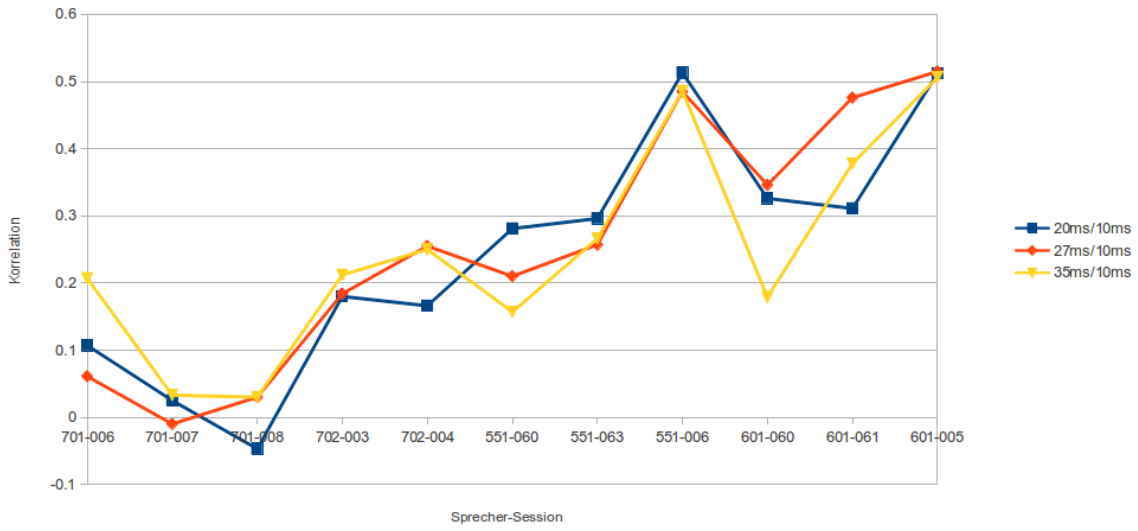


Abbildung 5.5: Ergebnisse für unterschiedliche Fensterlängen

Die bis jetzt verwendete Konfiguration $27ms/10ms$ hat im Durchschnitt die besten Ergebnisse, und zwar sowohl bei der Korrelation als auch bei der Standardabweichung und bei der stimmlos/stimmhaft Erkennungsrate. Bezüglich der Korrelation ist die ursprüngliche Konfiguration durchschnittlich um etwa 5% besser als die anderen zwei Varianten. Also die Fensterlänge von $27ms$ und die Verschiebung von $10ms$ haben sich als die günstigsten Einstellungen bewährt und werden deswegen in den weiteren Experimenten in dieser Arbeit eingesetzt. Die Fensterverschiebung haben wir nicht variiert, da Erfahrungen aus der Sprachverarbeitung zeigen, dass eine Fensterverschiebung von $10ms$ eine optimale Repräsentation von Phonemzuständen darstellt.

Eine detaillierte Analyse der Zahlen in der Tabelle zeigt jedoch, dass sehr klare Unterschiede zwischen den Sprechern existieren. Nehmen wir den Sprecher 601. Alle seinen Sessions weisen die besten Korrelationen wenn man die mit einer Fensterlänge von $27ms$ verarbeitet. Im Gegensatz dazu verursacht eine kleinere Fensterlänge eine merkbare Verbesserung der F0-Schätzung bei Sprecher 551. Die Fensterlänge von $20ms$ ist für diesen Sprechern passender. Die Sprecher 702 und 701 haben die besten Ergebnisse mit einer größeren Fensterlänge, nähmlich von $35ms$. Es ist nicht auszuschließen, dass für die unterschiedlichen Probanden unterschiedliche Fensterlängen geeignet sind. Jeder besitzt unterschiedliche Frequenzen in seiner Stimme und deswegen könnte eine individuellere Vorgehensweise zur Verbesserung führen.

5.5 Principal Component Analysis

In diesem Abschnitt untersuchen wir die Ergebnisse nachdem die EMG-Signale mit Principal Component Analysis(PCA)[1] vorverarbeitet sind.

[1]Im deutschsprachigen Raum wird PCA auch Hauptkomponentenanalyse genannt.

PCA hat seine Wurzeln in der Statistik und wurde von Karl Pearson in 1901 eingeführt. PCA ist eine mathematische Methode, die durch orthogonale lineare Transformation eine Menge von Eingabevektoren umformt. Die Eingabedaten sind Vektoren von Zufallsvariablen, deren Komponenten möglicherweise miteinander korrelieren. In unserem Fall sind die Komponenten der Zufallsvariablen die unterschiedlichen Kanäle des EMG-Signals. Das Ziel bei der linearen Transformation ist es, neue Zufallsvariablen zu erzeugen, deren Komponenten nicht mehr miteinander korreliert sind. Diese neuen Zufallsvariablen werden *Principal Components* genannt. Die Anzahl der Principal Components ist gleich der Anzahl der ursprünglichen Zufallsvariablen. Die durchgeführte Transformation ist so gestaltet, dass die Varianz der Werte in den Principal Components maximiert wird. Die Werte in der ersten Principal Component besitzen die größte Varianz. Die zweite Principal Component hat die zweit größte Varianz und ist gleichzeitig zu der ersten Princip Component orthogonal (unkorreliert). Das ermöglicht auch die Verwendung von PCA für Dimensionsreduzierung, wobei die letzte Principal Components mit kleinsten Varianzen der Werten einfach weggelassen werden. Die Annahme ist, dass diese Compoments eher unwichtig sind.

In unseren Untersuchungen jedoch benutzen wir PCA lediglich für eine Transformation der Daten. Unsere Hypothese ist, dass vielleicht die Umformung der Werte die EMG-Signale aussagekräftiger macht und folglich die F0-Schätzung verbessert. In der Tabelle 5.6 sind die Ergebnisse nach einer Vorverarbeitung mit PCA erläutert. Anhand der zugrunde liegenden Werte der EMG-Signale wurden die PCA-Transformationsmatrizen berechnet. Da die Signale von dem Kehlkopf andere Bedeutung haben und unkorreliert zu diesen von der Wange sind, wird für jedes Elektrodenarray eine eigene PCA-Transformationsmatrix berechnet.
Die Experimente wurden mit der originalen *TD15* Featureset durchgeführt. Fensterlänge und Fensterverschiebung sind entsprechend $27ms$ und $10ms$. Das GMM wurde mit 2 Gauß-Glocken trainiert. Die Werte sind wie gewöhnlich mit der *ursprünglichen Konfiguration* verglichen.

Sprecher	Session	Ergebnisse nach PCA			Absolute Verbesserung		Relative Verbesserung	
		Korr.	SA	u/v ER	Korr.	u/v ER	Korr.	u/v ER
Sprecher 701	*006*	0.176	0.093	59%	0.115	5.4%	189%	10%
	007	-0.05	0.06	47%	-0.04	-3.4%	-400%	-7%
	008	-0.178	0.144	41%	-0.208	-6%	-700%	-13%
Sprecher 702	*003*	0.21	0.081	60%	0.026	0	14%	0
	004	0.26	0.067	65%	0.005	0	2%	0
Sprecher 551	*060*	0.193	0.067	61%	-0.017	-1%	-8%	-1%
	063	0.275	0.055	65%	0.018	1%	7%	2%
	006	0.48	0.116	77%	0	-0.6%	0%	-1%
Sprecher 601	*060*	0.23	0.078	64%	-0.116	-4.4%	-33%	-6%
	061	0.432	0.148	72%	-0.044	-1.5%	-10%	-2%
	005	0.482	0.097	77%	-0.033	-1.7%	-6%	-2%

Abbildung 5.6: Ergebnisse nach PCA. Eine PCA-Transformationsmatrix beinhaltet 8 Kanäle.

Die Ergebnisse der F0-Schätzung nach PCA sind kontrovers. Die Korrelation hat sich bei Sprechern 601 und 701 stark verschlechtert. Im Gegensatz dazu hat sich die Korrelation bei dem Probanden 702 verbessert und bei 551 merkt man kaum Veränderung im Durchschnitt. Anhand dieser Resultate können wir keine deutliche Aussage machen, ob PCA bei der F0-Schätzung hilft. Durchschnittlich hat sich die stimmlos/stimmhaft Erkennungsrate um 2% verschlechtert und bei der Korrelation ist eine Verschlechterung um 10% vorhanden.

Um die Untersuchung von PCA weiter zu vertiefen, wurden Experimenten durchgeführt, wo die PCA-Matrizen für weniger Kanäle berechnet sind. Tabelle 5.7 zeigt die Ergebnisse wenn für jedes Elektrodenarray zwei PCA-Matrizen verwendet werden. Jede Matrix wird dabei über die Daten von 4 EMG-Kanälen berechnet. Folglich wird PCA bei jedem Array separat für zwei Blöcke durchgeführt. Der erste Block beinhaltet die ersten 4 Kanäle des Arrays und der zweite Block die nächsten 4 Kanäle. Diese blockweise Strategie verhindert, dass die Werte aus weit voneinander liegenden Elektroden durch PCA zusammengeführt werden.

Sprecher	**Session**	**Ergebnisse nach blockweise PCA**			**Absolute Verbesserung**		**Relative Verbesserung**	
		Korr.	**SA**	**u/v ER**	**Korr.**	**u/v ER**	**Korr.**	**u/v ER**
Sprecher 701	*006*	0.19	0.059	60%	0.129	6.4%	211%	12%
	007	0.033	0.075	54%	0.043	3.4%	430%	7%
	008	0	0.086	41%	-0.03	-6%	-300%	-13%
Sprecher 702	*003*	0.176	0.066	59%	-0.008	0.6%	-4%	-1%
	004	0.288	0.058	66%	0.033	1.4%	13%	2%
Sprecher 551	*060*	0.267	0.065	65%	0.057	3.3%	27%	5%
	063	0.245	0.058	63%	-0.012	-1%	-4%	-2%
	006	0.48	0.116	77%	0	-0.6%	0%	-1%
Sprecher 601	*060*	0.257	0.091	66%	-0.089	-2.4%	-26%	-3%
	061	0.432	0.07	71%	-0.044	-2.5%	-9%	-3%
	005	0.482	0.097	77%	-0.033	-1.7%	-6%	-2%
Durchschnittlich							**2%**	**0%**

Abbildung 5.7: Ergebnisse nach blockweise PCA. Eine PCA-Transformationsmatrix beinhaltet 4 Kanäle.

Es ist zu beobachten, dass diese Einstellung zur Veränderung der Ergebnisse führt. Es ist merkwürdig, dass die Session 063 von Sprecher 551, die bei der vorherigen Konfiguration von PCA eine Verbesserung um 7% hatte, sich hier verschlechtert. Auch hier können wir eine Diskrepanz zwischen den Probanden feststellen. Die F0-Schätzung bei Sprecher 601 weist eine Verschlechterung bei allen Sessions auf. Andererseits verbessern sich im Durchschnitt die Korrelationen bei den Sprechern 551 und 702 deutlich. Auch bei Sprecher 701 ist eine Verbesserung zu sehen, jedoch müssen wir die Werte von diesem Sprecher skeptisch betrachten. Die ursprüngliche Werte bei ihm waren sehr niedrig und das führt zur sehr großen relative Verbesserungen/Verschlechterungen. Bei diesem Sprecher ist die absolute Verbesserung aussagekräftiger als die relative. In dem Sinne hat die blockweise PCA bei ihm die Korrelation für Session 007 zu einem höheren Niveau gebracht. Bei den anderen zwei

Sessions (007 und 008) ist F0-Schätzung immer noch mit sehr niedrige Genauigkeit. Der Leser sollte sich von den großen Zahlen bei der relativen Verbesserung nicht verwirren lassen.

Durchschnittlich hat die PCA bei der Korrelation zu einer Abwärtsentwicklung um 2% geführt. Die stimmlos/stimmhaft Erkennungsrate weist durschnittlich keine Veränderung auf. Der Nutzen der Methode für die F0-Erkennung zeigt sich auch hier als fraglich. Es existiert eine sehr große Varianz zwischen den Ergebnissen der unterschiedlichen Sessions. Anhand der durschnittlichen Resultate können wir aber schlussfolgern, dass die PCA für die F0-Schätzung schon hilfreich sein könnte.

5.6 Independent Component Analysis

In diesem Abschnitt führen wir eine Vorverarbeitung mit Independent Component Analysis[1](ICA) durch. Danach untersuchen wir die resultierenden Ergebnisse für die F0-Schätzung.

Ähnlich der PCA ist auch ICA eines mathematisches Verfahren aus der multivariaten Statistik. Ziel bei der ICA ist aus einem Signalgemisch, das aus Zufallsvariablen[2] besteht, unabhängige Komponenten zu ermitteln. Der große Unterschied zu PCA ist, dass die Varianz zwischen den berechneten Komponenten maximiert wird. Bei der PCA war die Varianz in den Komponenten maximiert.

Typischer Anwendungsbereich für die ICA sind Quelletrennung (Source Separation) Probleme. Ein Beispiel dazu ist das *Cocktail Party Problem*, wo viele Leute gleichzeitig in einem Raum sprechen (wie an einer Cocktailparty). Die Aufgabe dabei ist in dem Signalgemisch die einzelnen Sprecher zu unterscheiden und die richtigen Wörter zu dem richtigen Sprecher zuzuweisen.

In Abbildung 5.8 sind die Ergebnisse für F0-Erkennung nach einer ICA der EMG-Daten dargestellt. Es wurde die *TD15* Featureset verwendet. Fensterlänge und Fensterverschiebung sind entsprechend $27ms$ und $10ms$. Das GMM wurde mit 2 Gauß-Glocken trainiert.

Aus der Tabelle können wir auslesen, dass die ICA zu deutlicher Verbesserung der Resultate geführt hat. Auch in der stimmlos/stimmhaft Erkennungsrate ist eine relative Verbesserung (um 2%) vorhanden. Die Korrelation, die das wichtigste Maß ist, ist bei der Mehrheit der Sessions deutlich größer geworden. Mit stimmlos/stimmhaft Erkennungsrate von 78% und eine Korrelation von 0.517 ist dabei die Session 006 von Sprecher 551, die beste von dem Datencorpus. Im Durchschnitt hat sich die Korrelation um 23% verbessert. Diese klare Verbesserung zeichnet die ICA als die beste Methode in dieser Arbeit aus.

Nach einer Vorverarbeitung mit ICA konzentrieren sich in manchen EMG-Kanälen Frequenzen, die man als Rauschen klassifizieren kann. Das können wir auf Abbildung 5.9 ansehen. Wir vermuten, dass die Kanäle mit diesen Frequenzen von der LDA-MAtrix vernachlässigt werden. Dadurch bewirkt die ICA eine Bereinigung der Störung im Signal und führt entsprechend zu besseren Ergebnissen.

[1]Im deutschsprachigen Raum wird ICA auch Unabhängigkeitsanalyse genannt.
[2]Die Eingabezufallsvariablen sind in unserem Fall die einzelnen EMG-Kanäle.

Sprecher	Session	Ergebnisse nach ICA			Absolute Verbesserung		Relative Verbesserung	
		Korr.	SA	u/v ER	Korr.	u/v ER	Korr.	u/v ER
Sprecher 701	*006*	0.202	0.137	60%	0.141	6.4%	231%	12%
	007	0.028	0.056	51%	0.038	0.6%	380%	1%
	008	-0.033	0.114	41%	-0.63	-6%	-210%	-13%
Sprecher 702	*003*	0.302	0.152	66%	0.118	6.4%	64%	10%
	004	0.38	0.092	70%	0.125	5.4%	49%	8%
Sprecher 551	*060*	0.24	0.107	63%	0.03	1.3%	14%	2%
	063	0.458	0.081	64%	0.201	0%	78%	0%
	006	0.517	0.093	78%	0.032	0%	7%	0%
Sprecher 601	*060*	0.364	0.095	69%	0.018	0.6%	5%	1%
	061	0.509	0.104	75%	0.033	1.5%	7%	2%
	005	0.494	0.085	78%	0.021	0%	-4%	0%
Durchschnittlich							**23%**	**2%**

Abbildung 5.8: Ergebnisse nach ICA

Die blockweise Strategie aus dem Abschnitt Principal Component Analysis wurde auch mit ICA ausprobiert. Die F0-Evaluierung hat aber von diesem Verfahren nicht profitiert.

5.7 Second Order Blind Identification

Das letzte Vorverarbeitungsverfahren, das wir untersuchen, ist die Second Order Blind Identification (SOBI) Methode. Die ist ähnlich zu ICA in dem Sinne, dass auch hier für das Signalgemisch eine Quellenseparation durchgeführt wird. Das erfolgt durch die gemeinsame Diagonalisierung von unterschiedlichen Kovarianzmatrizen, die aus den ursprünglichen EMG-Daten berechnet werden. Man verwendet dabei die Kohärenz der Signale. Der SOBI Algorithmus besitzt die folgenden Schritten:

1. Sei $x(t)$ die Eingabe - eine Reihe von Werten. Wir berechnen die Kovarianzmatrix R. $\lambda_1, \lambda_2, \ldots, \lambda_n$ sind die n größte Eigenwerte von R und $h_1, \ldots, h_n$ die entsprechend dazugehörenden Eigenvektoren.

2. Sei σ^2 der Mittelwert der m-n kleinsten Eigenwerte. Dann

$$z(t) = [z_1(t), \ldots, z_n(t)]^T, \text{ wo } z_i(t) = (\lambda_i - \sigma^2)^{-\frac{1}{2}} \text{ für } 1 \leq i \leq n \tag{5.1}$$

$$W = [(\lambda_1 - \sigma^2)^{-\frac{1}{2}} h_1, \ldots, (\lambda_n - \sigma^2)^{-\frac{1}{2}} h_n]^H \tag{5.2}$$

   wobei H bezeichnet die komplex konjugierte und transponierte Matrix.

3. Seien $F(\tau)$ die Kovarianzmatrizen von $z(t)$ für unterschiedliche Lags $\tau \in \{\tau_j | j = 1, \ldots, K\}$

4. Die unitäre Matrix U ist berechnet bei einer gemeinsammen Diagonalisierung der Matrizen $\{ F(\tau_j) | j = 1. \ldots, K \}$

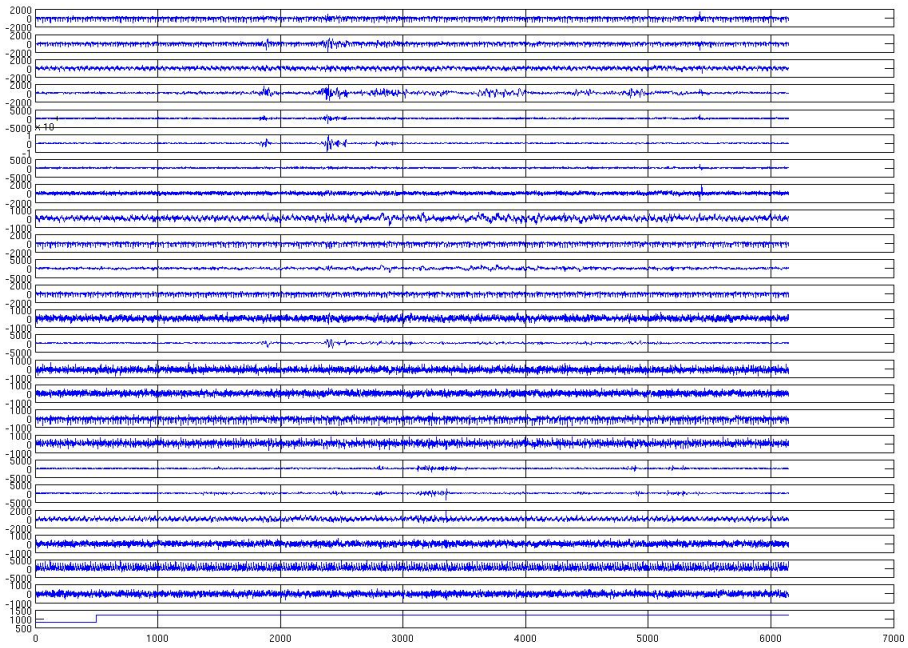


Abbildung 5.9: EMG-Signal einer Äußerung nach ICA.

5. Die einzelne Quellsignale sind definiert als $s(t) = U^H W x(t)$

Mehr Einzelheiten über den wissenschaftlichen Hintergrund der SOBI ist in [5] beschrieben.

In Tabelle 5.10 sind die F0-Ergebnisse für sieben Sessions nach einer Vorverarbeitung mit SOBI illustriert. Die Sessions mit 3 Elektrodenarrays konnte nach der SOBI Methode von dem Sprachsynthese-System nicht weiterverarbeitet werden. Es gab Fehler beim Training des GMMs. Deswegen sind keine Resultate für diese Sessions vorhanden. Auch hier wurden die Experimente mit der blockweisen Technik gemacht. Wenn man die SOBI Methode auf den ganzen Arrays durchführt, bekommt man schlechtere Ergebnisse. Die Anzahl und die Länge der Blöcke sind gleich wie in dem Abschnitt für Principal Component Analysis. Wir haben also 2 Blocks mit 4 Kanälen für jeden Elektrodenarray. Folglich berechnen wir bei den Sessions mit zwei 8-Elektrodenarrays 4 unterschiedliche SOBI-Transformationsmatrizen. Für die zwei Sessions mit einzelnen Elektroden wurden keine Blocks verwendet. Das verwendete Featureset ist *TD15* und die Fensterlänge und die Fensterverschiebung sind entsprechend $27ms$ und $10ms$. Das GMM wurde mit 2 Gauß-Glocken trainiert.

Wie gewöhnlich vergleichen wir die Ergebnisse mit der ursprünglichen Konfiguration. Aus den Werten in der Tabelle können wir schließen, dass SOBI zu keiner Verbesserung führt. Die Korrelation ist für 5 von den untersuchten 7 Sessions schlechter geworden. Bei manchen Sessions ist die Abwärtsentwicklung ganz deutlich. Zum Beispiel ist die Korrelation bei 601-060 zwischen tatsächliche F0 und geschätzte F0 um 29% kleiner geworden. Auch bei der stimmlos/stimmhaft Erkennungsrate gibt es bei der Mehrheit der Sessions eine Verschlechterung. Wie bei den anderen Experimenten

Sprecher	Session	Ergebnisse nach blockweise SOBI			Absolute Verbesserung		Relative Verbesserung	
		Korr.	SA	u/v ER	Korr.	u/v ER	Korr.	u/v ER
Sprecher 701	*006*	0.178	0.125	58%	0.117	4.4%	192%	8%
Sprecher 702	*003*	0.166	0.074	58%	-0.018	1.6%	-10%	3%
	004	0.277	0.053	66%	0.022	1.4%	8%	2%
Sprecher 551	*060*	0.182	0.077	60%	-0.028	-1.7%	-13%	-3%
	006	0.444	0.1	76%	-0.041	-1.7%	-8%	-2%
Sprecher 601	*060*	0.245	0.088	64%	-0.101	-4.4%	-29%	-6%
	005	0.495	0.118	78%	-0.02	-0.7%	-4%	-1%
Durchschnittlich							**-3%**	**0%**

Abbildung 5.10: Ergebnisse nach blockweise SOBI

ist eine große Verbesserung bei 701-006 vorhanden, jedoch ist die Korrelation bei dieser Session immer noch relativ niedrig.

Im Allgemeinen verschlechtert sich die Korrelation um durschnittlich 3%. Anzumerken ist, dass nur zwei Sessions (701-006 und 702-004) positiven Einfluss auf die durchschnittlichen Resultaten haben. Deswegen können wir sagen, dass die SOBI Methode für die F0-Synthese eher ungeeignet ist.

5.8 Experimente mit weniger Kanälen

Überall in den präsentierten Experimenten hat man mit den Elektrodenkofigurationen mit einzelnen Elektroden die besten Resultate erzielt. Das hat uns motiviert Untersuchungen mit den Arrays durchzuführen, bei dennen wir nur weniger der vorhandenen Kanälen nutzen. In Tabelle 5.11 sind die Ergebnisse mit 5 Kanälen dargestellt. Es wurden 2 Kanäle von der Wange und 3[1] Kanäle von dem Kehlkopf genommen. Die genommenen Kanäle sind:

- Kanal 6 und 7 von der Wange
- Kanal 10, 13 und 14 von dem Kehlkopf

Da wir in [20] mit diesem Elektrodensetting optimale Ergebnisse erzielten, wurde die Positionierung übernommen.

Bei drei der vier Probanden hat die neue Konfiguration zur durchschnittlichen Verbesserung der Korrelation geführt. Bei diesen Sprechern ist auch die u/v Erkennungsrate im Durchschnitt besser geworden. Nur bei Sprecher 601 ist eine Verschlechterung der Ergebnisse vorhanden.

Die Reduzierung der Kanalanzahl hat die durchschnittlichen Resultate nicht beeinflusst. Im Durchschnitt bleiben die Korrelation und die u/v Erkennungsrate auf denselben Werten wie bei der *urspünglichen Konfiguration*. Wir können also feststellen,

[1] Wir haben mehr Kanäle aus dem Kehlkopf genommen, denn vermutlich beinhalten die Muskelaktivitäten von diesem Bereich mehr F0-spezifische Information.

Sprecher	Session	Ergebnisse mit 5 Kanälen			Absolute Verbesserung		Relative Verbesserung	
		Korr.	SA	u/v ER	Korr.	u/v ER	Korr.	u/v ER
Sprecher 701	*006*	0.163	0.14	58%	0.102	4.4%	167%	8%
	007	-0.014	0.136	53%	0	2.6%	0%	5%
	008	0.058	0.066	45%	0.028	-2%	93%	-4%
Sprecher 702	*003*	0.289	0.188	64%	0.105	4.4%	57%	7%
	004	0.241	0.11	63%	-0.014	1.4%	-5%	2%
Sprecher 551	*060*	0.248	0.069	63%	0.038	1.3%	18%	2%
	063	0.256	0.055	64%	0	0%	0%	0%
Sprecher 601	*060*	0.29	0.104	66%	-0.056	-2.4%	-16%	-3%
	061	0.261	0.057	63%	-0.215	-7.5%	-45%	-10%
Durchschnittlich							**0%**	**0%**

Abbildung 5.11: Ergebnisse mit Reduktion auf 5 Kanälen

dass eine große Anzahl von Kanälen nicht unbedingt zu besseren Ergebnissen führt. Ganz im Gegenteil hat sich bei manchen Sessions gezeigt, dass die F0-Schätzung mit weniger Kanälen viel präziser ist.

Jedoch können wir nicht eindeutige Schlussfolgerungen für die Arrays machen. Man könnte Vorverarbeitungstechniken, die speziell für die Natur der Arrays geeignet sind, entwickeln und so die Leistungsfähigkeit steigern. Auf Abbildung 5.12 ist die Kreuzkorrelationsfunktion von zwei benachbarten Kanälen eines Arrays dargestellt. Es ist zu beobachten, dass eine sehr interessante Periodizität im EMG-Signal vorhanden ist. Vielleicht könnte man diese Information nutzen und zum Beispiel mit Autokorrelationsmethoden die F0-Schätzung verbessern.

5.9 Zusammenfassung

In diesem Kapitel haben wir die Resultate von unterschiedlichen Experimenten präsentiert. Wir haben festgestellt, dass die F0-Schätzung von einer größeren Anzahl der Gauß-Glocken nicht profitiert. Es wurde ein neues Feauterset vorgeführt. Diese Featureset beinhaltet eine Kombination von zeitlichen und spektralen Merkmalen und hat bei den Untersuchungen bessere F0-Erkennungsraten gezeigt. Wir haben noch gezeigt, dass eine Fensterlänge von $27ms$ für die F0-Schätzung besser als eine von $20ms$ oder $35ms$ ist. Es wurden noch die Vorverarbeitungsmethoden PCA, ICA und SOBI untersucht. Wir haben diese Methoden auch blockweise ausprobiert. Die Resultate nach einer Transformation mit PCA variieren stark zwischen den unterschiedlichen Sprechern. Durchschnittlich aber hat PCA eine leichte positive Auswirkung auf die F0-Schätzung. Bei den Untersuchungen der ICA, hat sich erwiesen, dass diese Methode für die F0-Erkennung passend ist. Durchschnittlich hat man mit ICA eine deutliche Verbesserung erreicht. Die Ergebnisse nach einer Vorverarbeitung mit SOBI zeigen, dass diese Methode für die F0-Synthese eher ungeeignet ist.

Die durchgeführten Experimente wurden aufgrund eines relativ begrenzten Datencorpus mit 11 Sessions durchgeführt. Dabei beinhaltete jede Session lediglich 50 Äußerungen[1]. Testen mit größeren Datenmengen ist notwendig.

[1] 45 Sätze für Training und die restlichen 5 für Testen

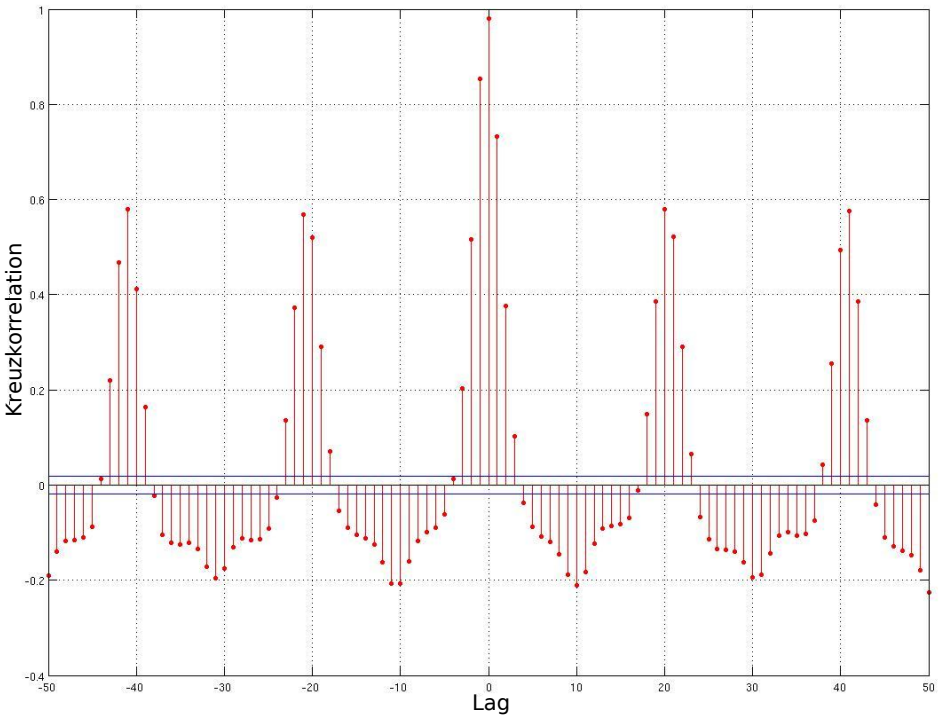


Abbildung 5.12: Korrelationskoeffizienten von zwei benachbarten EMG-Kanälen für unterschiedliche Verschiebungen

6. Zusammenfassung und Ausblick

Diese Arbeit präsentiert aktuelle Ergebnisse im Bereich der F0-Erkennung in der EMG-basierten Spracherkennung. Es wurden unterschiedlichen Verfahren bezüglich besserer F0-Schätzung untersucht und kommentiert. Allerdings ist zu bemerken, dass alle Untersuchungen auf sehr begrenztem Datencorpus durchgeführt wurden. Deswegen ist es auch schwierig die Ergebnisse zu evaluieren. Um deutlichere Vorstellung für den Nutzeffekt der einzelnen Methoden zu gewinnen, sollte man die Experimente mit mehreren Training- und Testdaten machen. Aufgrund der gemachten Untersuchungen konnten wir die folgende Resultate feststellen:

- die Anzahl der Gauß-Glocken bei dem Training hat keinen Einfluss auf die F0-Schätzung.
- eine Fensterlänge von $27ms$ hat sich für die F0-Schätzung als optimal erwiesen.
- eine neue Featureset namens *TD15-neu* hat zu einer durchschnittlichen Verbesserung von 5% bei der F0-Schätzung geführt.
- von drei untersuchten linearen Signaltransformationen (PCA, ICA, SOBI) erzielten wir mit der ICA die besten Ergebnisse (durchschnittliche Verbesserung um 23 %).
- nur mit 5 EMG-Kanälen erzielt man die gleichen Ergebnisse wie mit 16 oder 24 Kanälen.

In dieser Arbeit wurden die Elektrodenarrays im Bezug auf EMG-Sprachsynthese und besonders F0-Schätzung untersucht. Im Vergleich mit zwei Sessions, die mit 6 einzelne Ag-AgCl Elektroden durchgeführt wurden, zeigen die Arrays deutlich schlechtere Ergebnisse. F0 ist jedoch nur eine Komponente der Sprache. Man sollte die Elektrodenarrays auch für tatsächliche Sprachsynthese untersuchen.

Andere Forschungsmöglichkeit wäre die Kombination von Arrays mit einzelnen Elektroden. Eine Konfiguration mit einem Array am Hals und einzelnen Elektroden am Gesicht wäre interessant. Außerdem wird die F0 hauptsächlich von den Muskeln in

dem Kehlkopf beeinflusst. Deswegen wären Experimente, die die F0 nur aus dem Halsbereich extrahieren, angemessen. Es kann sein, dass die aus der Wange genommene EMG-Kanäle die F0-Schätzung stören.

In dieser Arbeit haben wir zwei Sessions mit einzelnen Elektroden analysiert. Bei den beiden Sessions wurden lediglich 6 EMG-Kanäle benutzt. Man konnte Konfigurationen untersuchen mit mehrere Ag-AgCl Elektroden. Dadurch wird die hohe Dimensionalität von den Arrays auch mit einzelnen Elektroden reproduziert. Außerdem konnte man unterschiedliche Einstellungen in dem Sprachsynthesesystem vornehmen und die entstehenden Resultate analysieren. Zum Beispiel wurde hier bei der Merkmalextraktion eine Stapelung von den 31 benachbarten Fenstern vorgenommen. Die Merkmalsvektoren wurden immer auf 32 Dimensionen reduziert. Man konnte diese Zahlen variieren.

Die F0-Erkennung ist sehr abhängig von der Präzision der aufgenommenen EMG-Signale. Die Exaktheit der Aufnahmehardware spielt also eine wichtige Rolle. Es ist festzustellen, dass die Arrays störungsanfällig sind. Sie sitzen nicht fest am Gesicht des Sprechers. Ihre Unflexibilität und die anspruchsvolle Verkabelung begrenzen stark die Bewegungsfreiheit des Probanden. Drahtlose Technik würde an dieser Stelle sehr hilfreich sein. Drahtlose Elektroden würden die Türe der Markteinführung für das hier beschriebenes Sprachsynthesesystem öffnen.

7. Anhang

In diesem Kapitel sind weitere detaillierte Ergebnisse und Grafiken vorgestellt. Das verwendete Qualitätsmaß sind die Korrelation zwischen wahrer und geschätzter F0 und die Erkennungsrate von stimmhaften und stimmlosen Phonemen (siehe Seite 16). Die Standardabweichung wurde über die Korrelationen der 5 Testsätze berechnet.

7.1 Korrelationskoeffizient auf stimmhaften Fenstern

Sprecher	**Session**	**Elektrodenkonfiguration**	**Ergebnisse**		
			Korrelation	**Standard-abweichug**	**Stimmlos/stimm-haft Erkennungsrate [%]**
Sprecher 701	*006*	2 x 8-Elektrodenarray	0.137	0.075	53.6
	007	3 x 8-Elektrodenarray	0.038	0.058	50.4
	008	3 x 8-Elektrodenarray	0.03	0.171	47
Sprecher 702	*003*	2 x 8-Elektrodenarray	-0.069	0.128	59.6
	004	2 x 8-Elektrodenarray	0.055	0.102	64.6
Sprecher 551	*060*	2 x 8-Elektrodenarray	0.054	0.152	61.7
	063	3 x 8-Elektrodenarray	0.113	0.133	64
	006	6 einzelnen Elektroden	0.283	0.286	77.6
Sprecher 601	*060*	2 x 8-Elektrodenarray	0.148	0.24	68.4
	061	3 x 8-Elektrodenarray	-0.041	0.293	73.5
	005	6 einzelnen Elektroden	0.249	0.14	78.7

Abbildung 7.1: Ergebnisse berechnet nur auf stimmhaften Fenstern

7.2 Gauß-Glocken

Sprecher	Session	Gauß-Glocken	Korrelation	Standardabweichung
Sprecher 701	*006*	2	0.061	0.053
		4	0.062	0.05
		8	0.058	0.05
	007	2	-0.01	0.046
		4	0	0.045
		8	-0.01	0.044
Sprecher 702	*003*	2	0.184	0.08
		4	0.183	0.08
		8	0.183	0.08
	004	2	0.255	0.031
		4	0.256	0.033
		8	0.257	0.033
Sprecher 551	*060*	2	0.21	0.076
		4	0.21	0.076
		8	0.21	0.076
	006	2	0.485	0.062
		4	0.483	0.063
		8	0.482	0.063
Sprecher 601	*060*	2	0.346	0.058
		4	0.345	0.06
		8	0.345	0.057
	061	2	0.476	0.103
		4	0.477	0.103
		8	0.477	0.102
	005	2	0.515	0.077
		4	0.517	0.077
		8	0.515	0.076

Abbildung 7.2: Einfluss der Anzahl der Gauß-Glocken auf die F0-Schätzung 5.3 - detailliert

7.3 Audio- und EMG-Signal

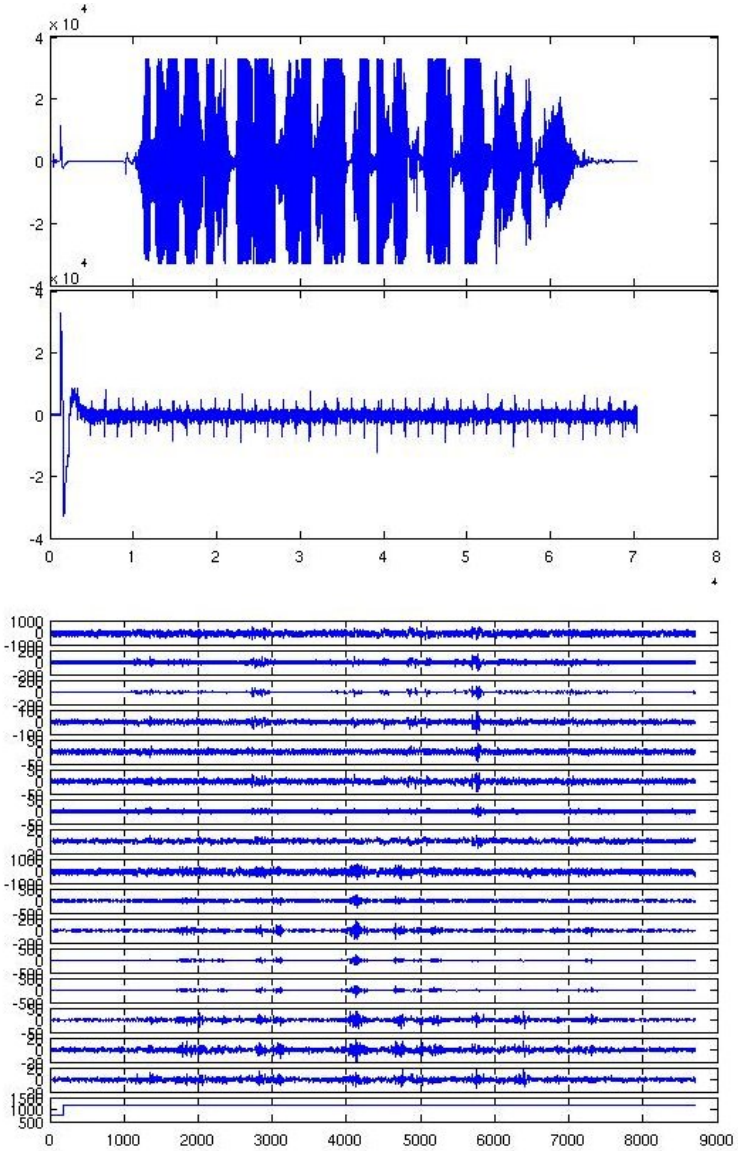


Abbildung 7.3: Parallel aufgenommenes Audiosignal (oben) und EMG-Signal (unten) für die Äußerung "Many states devise solutions to problems of welfare and health care."

7.4 Fensterlänge detaillierte Ergebnisse

Sprecher	Session	Fensterlänge/Fensterverschiebung								
		20ms/10ms			27ms/10ms			35ms/10ms		
		Korr.	SA	u/v ER	Korr.	SA	u/v ER	Korr.	SA	u/v ER
Sprecher 701	*006*	0.107	0.038	56%	0.061	0.053	53.6%	0.207	0.142	60%
	007	0.025	0.098	53%	-0.01	0.046	50.4%	0.033	0.066	53%
	008	-0.047	0.138	43%	0.03	0.071	47%	0.03	0.17	47%
Sprecher 702	*003*	0.18	0.09	59%	0.184	0.08	59.6%	0.212	0.071	61%
	004	0.166	0.037	61%	0.255	0.031	64.6%	0.25	0.052	64%
Sprecher 551	*060*	0.281	0.036	65%	0.21	0.076	61.7%	0.157	0.069	60%
	063	0.296	0.076	65%	0.257	0.08	64%	0.266	0.079	64%
	006	0.513	0.092	78%	0.485	0.062	77.6%	0.486	0.067	77%
Sprecher 601	*060*	0.326	0.047	67%	0.346	0.058	68.4%	0.179	0.041	62%
	061	0.311	0.116	66%	0.476	0.103	73.5%	0.378	0.128	69%
	005	0.512	0.075	78%	0.515	0.077	78.7%	0.506	0.078	78%
Durchschnittlich		**0.243**	**0.077**	**63%**	**0.255**	**0.067**	**64%**	**0.246**	**0.088**	**63%**

Abbildung 7.4: Ergebnisse für unterschiedliche Fensterlängen detailliert.

7.5 Stimmlos/stimmhaft Erkennungsraten

Hier sind die stimmlos/stimmhaft Erkennungsraten für alle Sessions und für alle untersuchten Modi in Einzelheiten beschrieben.

Sprecher	Session	u/v Erkennungsrate [%]		
		erkannt als	**sind v**	**sind u**
Sprecher 701	*006*	**v**	52	45
		u	1	2
	007	**v**	32	23
		u	26	19
	008	**v**	32	44
		u	9	15
Sprecher 702	*003*	**v**	37	23
		u	17	23
	004	**v**	47	24
		u	11	18
Sprecher 551	*060*	**v**	18	11
		u	27	44
	063	**v**	23	14
		u	21	42
	006	**v**	21	6
		u	12	62
Sprecher 601	*060*	**v**	19	9
		u	22	50
	061	**v**	32	13
		u	13	42
	005	**v**	26	9
		u	10	54

Abbildung 7.5: Detaillierte u/v Erkennungsraten für die *ursprüngliche Konfiguration* 5.1

Sprecher	Session	u/v Erkennungsrate [%]		
		erkannt als	**sind v**	**sind u**
Sprecher 701	*006*	**v**	41	28
		u	12	19
	007	**v**	32	24
		u	25	19
Sprecher 702	*003*	**v**	36	23
		u	18	23
	004	**v**	50	27
		u	8	15
Sprecher 551	*060*	**v**	22	13
		u	24	41
	063	**v**	24	16
		u	21	39
	006	**v**	18	7
		u	16	59
Sprecher 601	*060*	**v**	18	10
		u	22	50
	061	**v**	42	15
		u	9	34
	005	**v**	22	12
		u	10	56

Abbildung 7.6: Detaillierte u/v Erkennungsraten für das neue Featureset 5.4

Sprecher	Session	u/v Erkennungsrate [%]						
		erkannt als	20ms/10ms		27ms/10ms		35ms/10ms	
			sind v	sind u	sind v	sind u	sind v	sind u
Sprecher 701	*006*	**v**	52	45	52	45	41	28
		u	1	2	1	2	12	19
	007	**v**	35	24	32	23	32	21
		u	23	18	26	19	26	21
	008	**v**	37	53	32	44	32	44
		u	4	6	9	15	9	15
Sprecher 702	*003*	**v**	46	33	37	23	37	22
		u	8	13	17	23	17	24
	004	**v**	51	31	47	24	48	26
		u	7	11	11	18	10	16
Sprecher 551	*060*	**v**	24	13	18	11	16	10
		u	21	42	27	44	30	44
	063	**v**	27	17	23	14	23	14
		u	17	39	21	42	21	41
	006	**v**	21	9	21	6	21	9
		u	13	57	12	62	13	57
Sprecher 601	*060*	**v**	18	10	19	9	9	5
		u	23	49	22	50	32	54
	061	**v**	23	10	32	13	28	13
		u	23	44	13	42	18	41
	005	**v**	21	12	26	9	22	12
		u	10	57	10	54	10	56

Abbildung 7.7: Detaillierte u/v Erkennungsraten für die unterschiedlichen Fensterlängen 5.5

Sprecher	Session	u/v Erkennungsrate [%]		
		erkannt als	**sind v**	**sind u**
Sprecher 701	*006*	**v**	39	27
		u	14	20
	007	**v**	24	19
		u	34	23
	008	**v**	41	59
		u	0	0
Sprecher 702	*003*	**v**	37	23
		u	17	23
	004	**v**	47	24
		u	11	18
Sprecher 551	*060*	**v**	18	12
		u	27	43
	063	**v**	24	14
		u	21	41
	006	**v**	21	6
		u	12	62
Sprecher 601	*060*	**v**	15	10
		u	25	50
	061	**v**	33	15
		u	13	39
	005	**v**	26	9
		u	10	54

Abbildung 7.8: Detaillierte u/v Erkennungsraten für PCA 5.6

Sprecher	Session	u/v Erkennungsrate [%]		
		erkannt als	sind v	sind u
Sprecher 701	*006*	**v**	52	45
		u	1	2
	007	**v**	32	23
		u	26	19
	008	**v**	32	44
		u	9	15
Sprecher 702	*003*	**v**	37	23
		u	17	23
	004	**v**	47	24
		u	11	18
Sprecher 551	*060*	**v**	18	11
		u	27	44
	063	**v**	23	14
		u	21	42
	006	**v**	21	6
		u	12	62
Sprecher 601	*060*	**v**	19	9
		u	22	50
	061	**v**	32	13
		u	13	42
	005	**v**	26	9
		u	10	54

Abbildung 7.9: Detaillierte u/v Erkennungsraten für blockweise PCA 5.7

Sprecher	Session	u/v Erkennungsrate [%]		
		erkannt als	**sind v**	**sind u**
Sprecher 701	*006*	**v**	40	28
		u	12	20
	007	**v**	28	19
		u	30	23
	008	**v**	41	59
		u	0	0
Sprecher 702	*003*	**v**	40	21
		u	14	25
	004	**v**	47	19
		u	11	23
Sprecher 551	*060*	**v**	24	16
		u	21	39
	063	**v**	27	18
		u	18	38
	006	**v**	21	8
		u	13	58
Sprecher 601	*060*	**v**	24	14
		u	17	45
	061	**v**	36	15
		u	10	39
	005	**v**	21	12
		u	10	57

Abbildung 7.10: Detaillierte u/v Erkennungsraten für ICA 5.8

Sprecher	Session	u/v Erkennungsrate [%]		
		erkannt als	**sind v**	**sind u**
Sprecher 701	*006*	**v**	40	28
		u	13	19
Sprecher 702	*003*	**v**	37	25
		u	17	21
	004	**v**	47	23
		u	11	19
Sprecher 551	*060*	**v**	19	13
		u	26	42
	006	**v**	18	8
		u	16	58
Sprecher 601	*060*	**v**	18	13
		u	22	47
	005	**v**	20	11
		u	11	58

Abbildung 7.11: Detaillierte u/v Erkennungsraten für blockweise SOBI 5.10

Sprecher	Session	u/v Erkennungsrate [%]		
		erkannt als	**sind v**	**sind u**
Sprecher 701	*006*	**v**	41	30
		u	12	17
	007	**v**	40	29
		u	18	13
	008	**v**	37	52
		u	4	7
Sprecher 702	*003*	**v**	39	20
		u	15	26
	004	**v**	48	27
		u	10	15
Sprecher 551	*060*	**v**	19	11
		u	26	44
	063	**v**	22	14
		u	22	42
Sprecher 601	*060*	**v**	16	10
		u	24	50
	061	**v**	16	7
		u	30	47

Abbildung 7.12: Detaillierte u/v Erkennungsraten für die Experimente mit nur 5 Kanälen 5.11

Literaturverzeichnis

[1] *http://en.wikipedia.org/wiki/File:Gray956.png.*

[2] *http://www.otbioelettronica.it.*

[3] *Bertelsmann Lexikon. Bertelsmann Lexikon Verlag.* Gütersloh, 2001.

[4] A. Paeschke, M.Kienast, W.F. Sendkmeier: *F0-Countours in emotional speech.* Technischer Bericht, Technische Universität Berlin, Germany.

[5] Adel Belouchrani, Karim Abed-Meriam, Jean-Francois Eric Moulines: *A Blind Source Separation Technique Using Second-Order Statistics.* Technischer Bericht, IEEE Transactions on signal Processing, 1997.

[6] Andreassi, John L.: *Psychophysiology : human behavior and physiological response.* 2007.

[7] Arthur R. Toth, Michael Wand, Tanja Schultz: *Synthesizing Speech from Electromyography using Voice Transformation Techniques.* Technischer Bericht, Cognitive Systems Lab, Universität Karlsruhe, Germany.

[8] Chan, A., K. Englehart B. Hudgins und D. Lovely: *Myoelectric Signals to Augment Speech Recognition.* Technischer Bericht, Medical and Biological Engineering and Computing, 2001.

[9] Finke, Michael, P. Geutner H. Hild T. Kemp K. Ries und M. Westphal: *The Karlsruhe-Verbmobil Speech Recognition Engine*, 1997.

[10] Foad Ghaderi, Hamid R. Mohseni und Saeid Sanei: *A fast second order identification method for separation of periodic sources.* Technischer Bericht, Centre of Digital Signal Processing, School of Engineering, Cardiff University, UK.

[11] Gutierrez-Osuna, Ricardo: *LECTURE 10: Linear Discriminant Analysis.* Technischer Bericht, Texas AM University, USA.

[12] Hosom, John-Paul: *F0 Estimation for Adult and Children's Speech.* Technischer Bericht, Oregon Health and Science University, Portland, OR, USA.

[13] Hyvärinen, Aapo und Erkki Oja: *Independent Component Analysis:Algorithms and Applications.* Technischer Bericht, Helsinki University of Technology, FIN-02015 HUT, 2000.

[14] Janke, Matthias: *Spektrale Methoden zur EMG-basierten Erkennung lautloser Sprache.* Diplomarbeit, Karlsruhe Institut of Technology (KIT), Karlsruhe, 2010.

[15] Jieping Ye, Shuiwang Ji: *Discriminant Analysis for Dimensionality Reduction: An Overview of Recent Developments.* Technischer Bericht, Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA.

[16] Keigo Nakamura, Matthias Janke, Michael Wand und Tanja Schultz: *Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0.* Technischer Bericht, Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT) Karlsruhe, Germany, Karlsruhe, 2011.

[17] Komi, Paavo V. und Per Tesch: *EMG Frequency Spectrum, Muscle Structure, and Faitigue During Dynamic Contractions in Man.* Technischer Bericht, Kinesiology Laboratory, Department of Biology of Physical Activity, University of Jyvaskyla, Finland and Karolinska Hospital, department of Clinical Physiology, Stockholm, Sweden.

[18] Michael Wand, Szu-Chen Stan Jou, Arthur R. Toth Tanja Schultz: *Impact of Different Speaking Modes on EMG-based Speech Recognition.* Technischer Bericht, Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT) Karlsruhe, Germany and ATC, ICL, Industrial Technology Research Institute, Taiwan.

[19] OT Bioelettronica, C.so Unione Sovietica 312 10135 – Torino (TO) - ITALY: *User manual v 2.0 EMG-USB2 Multichannel Bioelectrical Signal Amplifier.* www.otbioelettronica.it.

[20] Schulte, Christopher: *Aufbau eines EMG-basierten Spracherkennungssystems unter Verwendung von Elektrodenarrays*, 2011.

[21] Schultz, Prof. Dr. Tanja und Dipl. Math. Michael Wand: *Biosignale und Benutzerschnittstellen; Digitale Signalverarbeitung.* 2010.

[22] Smith, Lindsay I: *A tutorial on Principal Components Analysis.* Technischer Bericht, February 26, 2002.

[23] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek Florian Kraft und Alex Waibel: *Towards Continuous Speech Recognition Using Surface Electromyography.* Technischer Bericht, International Center for Advanced Communication Technologies and Carnegie Mellon University, USA and University Karlsruhe, Germany.

[24] Tanja Schultz, Michael Wand: *Modeling Coarticulation in EMG-based Continuous Speech Recognition.* Technischer Bericht, Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT) Karlsruhe, Germany, Karlsruhe, 2010.

[25] Traunmüller, Hartmut und Anders Eriksson: *The frequency range of the voice fundamental in the speech of male and female adults.* Technischer Bericht, Institutionen för lingvistik, Stockholms universitet, Stockholm, Sweden.

[26] Wielatt, Thomas: *Entwicklung eines Werkzeuges zur Echtzeitvisualisierung von Biosignalen.* Diplomarbeit, Karlsruhe Institut of Technology (KIT), Karlsruhe, 2009.