Karlsruhe Institute of Technology

Konvertierung von myoelektrischen Signalen der Gesichtsmuskulatur zu Sprache: Ein Unit Selection-Ansatz

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

von

Marlene Zahner

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

Tag der Anmeldung: 01. Februar 2014
Tag der Abgabe: 31. Juli 2014

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

Karlsruhe, den 31. Juli 2014

Zusammenfassung

In dieser Masterarbeit wird ein neuer Ansatz zur Konvertierung von Muskelsignalen zu hörbarer Sprache vorgestellt. Allein die elektrischen Signale der Gesichtsmuskulatur werden mittels Oberflächenelektroden aufgezeichnet, ohne dass der Sprecher hörbare Laute artikulieren muss. Anschließend werden die Signale in kurze Abschnitte unterteilt und mit bereits aufgezeichneten EMG-Signalen verglichen, die wiederum mit Sprachsegmenten verknüpft sind. Die gewählten Sprachsegmente werden dann erneut zu Sprache zusammengesetzt. Diese Vorgehensweise des Wählens und Zusammensetzens bereits vorhandener Sprachabschnitte wird *Unit Selection* genannt.

Der Vorteil, Muskelsignale in Sprache umzuwandeln liegt darin, dass man auch in lauten Umgebungen oder für einen Sprecher mit Sprachbehinderungen hörbare akustische Signale erzeugen kann, die zu einem bestimmten Empfänger geleitet werden können. Beistehende werden hierbei nicht belästigt und die Privatsphäre wird ebenfalls gewahrt.

Frühere Experimente haben die Machbarkeit der Umwandlung myoelektrischer Signale zu Sprache gezeigt, jedoch existieren Probleme vor allem bei der Natürlichkeit der erzeugten Sprache. Durch das hier vorgestellte Verfahren sollen sowohl Verständlichkeit als auch Natürlichkeit der konvertierten Sprache gegenüber bereits existierenden Ansätzen verbessert werden.

Die Evaluation der konvertierten Sprache zeigt in einer objektiven Evaluationsmethode eine Steigerung von durchschnittlich 13% gegenüber einem bereits existierenden Ansatz. Ein Hörtest zur subjektiven Bewertung der Natürlichkeit bestätigt ebenfalls die Verbesserung der konvertierten Sprache durch den hier vorgestellten Ansatz.

Inhaltsverzeichnis

1 Einleitung **1**
1.1 Zielsetzung der Arbeit . . . 2
1.2 Gliederung der Arbeit . . . 2

2 Grundlagen **3**
2.1 Elektromyographie . . . 3
2.1.1 Physiologische Grundlagen der Muskelaktivität . . . 3
2.1.2 Ableitung myoelektrischer Signale . . . 5
2.1.3 EMG-Features . . . 7
2.2 Sprache . . . 9
2.2.1 Sprachproduktion . . . 9
2.2.2 Repräsentation von Sprache . . . 10
2.2.3 Geflüsterte Sprache . . . 13
2.2.4 Phonem-Labels . . . 13
2.3 Evaluation der konvertierten Sprache . . . 13
2.3.1 Mel Cepstral Distortion . . . 13
2.3.2 Erkennung der Stimmhaftigkeit . . . 14
2.4 Verwandte Arbeiten . . . 14
2.4.1 EMG-basierte Spracherkennung . . . 14
2.4.2 EMG-basierte Sprachsynthese . . . 15

3 Unit Selection **17**
3.1 Units . . . 17
3.2 Codebook . . . 18
3.3 Test-Segmente . . . 19
3.4 Kostenfunktionen . . . 20
3.4.1 Target Cost . . . 20
3.4.2 Concatenation Cost . . . 21
3.4.3 Gewichtung der Kostenfunktionen . . . 22
3.4.4 Distanzmaße . . . 22
3.5 Suche nach optimaler Unit-Sequenz . . . 23
3.5.1 Greedy-Suche . . . 23
3.5.2 Viterbi-Suche . . . 24
3.5.3 Separater Suchpfad für F0 . . . 26

4 Experimente **27**
4.1 EMG-Daten . . . 27
4.1.1 Aufnahmesetup . . . 27
4.1.2 Datenkorpus . . . 30

4.2 Whisper/Audible-Konvertierung 30
4.2.1 Units 30
4.2.2 Alignment 31
4.3 Oracle-Experiment: Audible/Audible-Konvertierung 32
4.3.1 Menge der Trainingsdaten 32
4.3.2 Framegröße 33
4.4 EMG/Audible-Konvertierung 33
4.4.1 Units 34
4.5 Synthese des akustischen Ziel-Signals 34
4.5.1 Mel Log Spectrum Approximation (MLSA) 34
4.5.2 Direkte Synthese aus akustischem Signal 35

5 Evaluation 37
5.1 Parameter der Unit Selection 37
5.1.1 Stacking-Höhe 37
5.1.2 Segmentgröße 37
5.1.3 Überlappung der Units 39
5.1.4 Distanzmaße 41
5.1.5 Kostenfunktionen 41
5.1.6 Suche nach der optimalen Unit-Sequenz 43
5.1.7 Gewichtsfunktion 43
5.2 Evaluationsmethoden 44
5.2.1 Test auf Trainingsdaten 44
5.2.2 Erkennung der Stimmhaftigkeit 44
5.2.3 Phonem-Labels 45
5.2.4 Vergleich mit GMM-basiertem System 45
5.2.5 Hörtest 47

6 Zusammenfassung und Ausblick 51

Literaturverzeichnis 53

1. Einleitung

Sprache ist die natürlichste Form der Kommunikation für Menschen. Ihr Erwerb beginnt im Kindesalter, aber sie entwickelt sich das Leben lang weiter. Dabei dient sie nicht allein der Übermittlung von Wissen und Fakten, sondern kann, besser als geschriebener Text, insbesondere auch durch Variationen in der Sprachmelodie und Betonung dem Gesprächspartner Emotionen und Meinungen vermitteln.

Allerdings gibt es Situationen, in denen eine direkte Kommunikation über Sprache für einen Menschen nicht möglich ist, obwohl er die Ausdrucksmöglichkeiten, die sich mit der Sprache bieten, gerne nutzen würde. Beispiele hierfür sind Sprachstörungen, die eine hörbare Artikulation verhindern; die Notwendigkeit, sich geräuscharm zu verhalten; oder das Bedürfnis nach Wahrung der Privatsphäre, indem Beistehende das Gespräch nicht akustisch mitverfolgen können.

Ein *Silent Speech Interface* (SSI) verwendet die Bewegung der Gesichtsmuskulatur, als ob ein Sprecher tatsächlich spricht, selbst wenn er nur lautlos die Worte artikuliert. Es kann in diesen Situationen Abhilfe schaffen, indem es lautlose Sprache in Text oder hörbare Sprache umwandelt, aber es sollte verschiedene Bedingungen erfüllen, um praktikabel zu sein: Es sollte nicht-invasiv sein, um dem Benutzer keine Schmerzen zuzufügen; es sollte möglichst einfach zu bedienen sein; und vor allem sollte die Ausgabe des Interfaces qualitativ möglichst hochwertig sein.

Zu diesem Zweck wird der Einsatz von *Elektromyographie* (EMG) zur Erfassung von elektrischen Signalen der Gesichtsmuskulatur erforscht, woran auch diese Arbeit anknüpft. Im Gegensatz zur Sprach*erkennung* ist das Ziel der Sprach*synthese* nicht die Erkennung des Gesagten und dessen Umwandlung in Text, sondern die direkte Generierung eines hörbaren Audiosignals aus dem EMG-Signal. Dies hat für einen menschlichen Empfänger den Vorteil, dass es nicht die Fähigkeit oder Möglichkeit zu Lesen voraussetzt. Außerdem können, wie schon erwähnt, prosodische Informationen in Sprache besser transportiert werden als in geschriebenem Text.

Wenn die Muskelsignale mittels Oberflächenelektroden erfasst werden, ist das Kriterium der Nichtinvasivität erfüllt. Da der Sprecher als Eingabe für das Interface nur stumm sprechen muss, also seinen Sprachapparat wie gewohnt bewegt, ist es ebenfalls intuitiv in der Anwendung.

Das bestehende Problem liegt jedoch in der Qualität der Ausgabe. Da Sprache nicht nur inhaltliche, also textuelle, Informationen enthält, sollten möglichst viele der beabsichtigten unterschwelligeren Informationen wie Betonung, Stimmlage etc. bei der Konvertierung von EMG-Daten zu Sprache beibehalten werden, wobei die Verständlichkeit der Ausgabe nicht vernachlässigt werden darf.

1.1 Zielsetzung der Arbeit

Das Ziel dieser Arbeit ist es EMG-Daten in möglichst natürliche und zugleich verständliche hörbare Sprache zu konvertieren, indem Unit Selection angewendet wird um die Eingabe in Form kurzer Segmente mit bereits vorhandenen Segmenten zu vergleichen und die bestmögliche Sequenz wieder zusammenzusetzen. Die Ausgabe des Verfahrens soll in dieser Arbeit ebenfalls evaluiert werden.
Bereits existierende Ansätze haben Schwierigkeiten mit der Generierung einer natürlichen klingenden Prosodie für die konvertierte Sprache. Dieses Problem soll durch den Unit Selection-Ansatz behoben werden, da die verwendeten Units bereits inhärente Informationen darüber enthalten. Außerdem soll die Verständlichkeit gegenüber den vorhandenen Ansätzen verbessert werden.
Die Konvertierung von EMG zu Sprache soll trotz einer begrenzten Trainingsdatenmenge zu einer akzeptablen Ausgabequalität führen.

1.2 Gliederung der Arbeit

Zunächst werden in Kapitel 2 die Grundlagen der Elektromyographie beschrieben, sowie die Entstehung und Repräsentation von Sprache, wie sie in dieser Arbeit genutzt werden. Außerdem werden die Methoden und Maße für die Evaluation der konvertierten Sprache eingeführt und die verwandten Arbeiten zu diesem Thema vorgestellt, um diese Masterarbeit in die vorhandene Forschung einzuordnen und eine Vergleichsmöglichkeit zu anderen Ansätzen zu schaffen.
Ein gesondertes Kapitel der Grundlagen stellt die Einführung der Unit Selection in Kapitel 3 dar, da diese das Hauptthema der Arbeit bildet. Es werden die benötigten Elemente sowie die Vorgehensweise bei der Unit Selection beschrieben.
In Kapitel 4 werden zunächst die verschiedenen Vorstudien und ihre Ergebnisse sowie die Durchführung der Hauptstudie erörtert, anschließend werden zwei Methoden zur Synthese der konvertierten Sprache vorgestellt.
Im Anschluss an die Experimentbeschreibung werden in Kapitel 5 die Ergebnisse der Experimente, vor allem in Hinblick auf verwendete Parameter und im Vergleich mit anderen Arbeiten evaluiert.
Im abschließenden Kapitel 6 werden die Ergebnisse zusammengefasst und es wird ein Ausblick auf mögliche Folgearbeiten gegeben.

2. Grundlagen

Dieses Kapitel erläutert die Grundlagen der Elektromyographie und der Sprachproduktion und -repräsentation, die für das Verständnis dieser Arbeit erforderlich sind. Außerdem werden die Methoden beschrieben, die zur Evaluation der konvertierten Sprache eingesetzt wurden. Zuletzt werden die Arbeiten vorgestellt, welche die Grundlage für diese Arbeit bilden.

2.1 Elektromyographie

Dieser Abschnitt beschreibt die Grundlagen der Elektromyographie, also der Messung von Muskelaktivität anhand ihrer elektrischen Potentiale.

2.1.1 Physiologische Grundlagen der Muskelaktivität

Für das Verständnis der Elektromyographie ist ein grundlegendes Verständnis der physiologischen Abläufe bei Bewegungen des menschlichen Körpers, in diesem Fall speziell der Bewegungen des Artikulationsapparates, erforderlich.

Die für diese Arbeit relevante Muskulatur ist die *Skelettmuskulatur*, die vom Menschen willkürlich gesteuert werden kann. Der Aufbau eines Skelettmuskels wird in Grafik 2.1 dargestellt.

Ohne Bewegung befinden sich die Muskelzellen der Skelettmuskulatur in einem entspannten Zustand. Hierbei ist das Zellinnere gegenüber dem Äußeren negativ geladen. Diese Spannungsdifferenz wird als *Ruhepotential* bezeichnet.
Wenn der Mensch sich zu einer Bewegung entschließt, werden elektrische Nervenimpulse vom Gehirn zu den entsprechenden Muskelfasern geleitet. Der Impuls gelangt zunächst zur Schnittstelle zwischen Nervenfaser und Muskelfaser, die *motorische Endplatte* genannt wird. Wenn der Impuls ausreichend stark ist, entsteht an dieser Stelle ein *Aktionspotential*, was bedeutet, dass sich die Spannungsdifferenz zwischen Zellinnerem und -äußerem umkehrt; die Zelle ist nun also positiv geladen. Diese Erregung breitet sich anschließend in Längsrichtung des Muskels aus. Die Spannungsänderung löst eine Reihe von chemischen Prozessen aus, die bewirken, dass sich *Aktin-* und *Myosinfilamente* gegeneinander verschieben und sich der Muskel dadurch verkürzt.

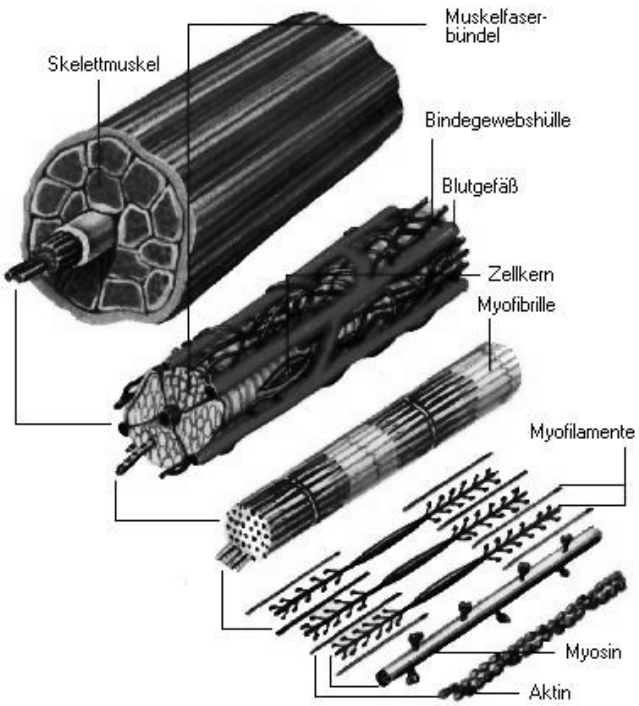

Abbildung 2.1: Aufbau eines Skelettmuskels, aus [Ber, 2001]

Um die Verschiebung auszuführen, binden die sogenannten *Myosinköpfe* (s. Grafik 2.2) an das Aktin und klappen um, wodurch sich die Filamente um eine kleine Strecke gegeneinander verschieben.

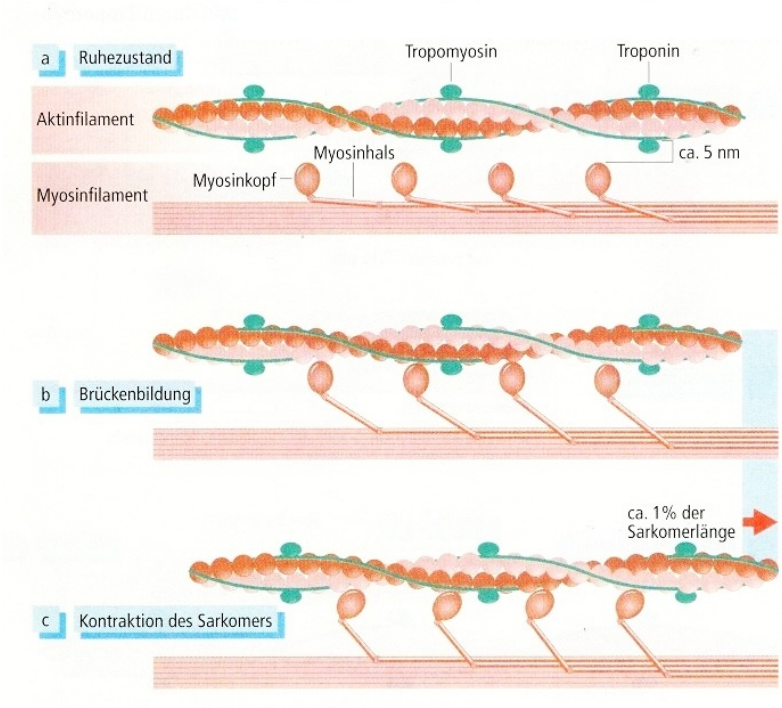

Abbildung 2.2: Kontraktion eines Muskels auf Mikrofilament-Ebene, aus [Solbach, 2000]

Die Verschiebung wird gegebenenfalls mehrfach für jedes Aktin- und Myosinfilament durchgeführt, indem die Myosinköpfe sich wieder vom Aktin lösen, in ihre Ausgangslage zurückklappen und neu ansetzen. Die Summe dieser winzigen Verschiebungen führt dadurch, dass in der Muskelfaser viele Filamente aneinandergereiht sind, letzten Endes zur Kontraktion des Muskels.

2.1.2 Ableitung myoelektrischer Signale

Bei der Elektromyographie werden mittels Nadel- oder Oberflächenelektroden die Potentialänderungen der Muskelzellen gemessen, genauer gesagt wird die Potentialdifferenz zwischen zwei Elektroden erfasst. Es gibt hierbei zwei Arten der Ableitung: unipolare und bipolare Ableitung.

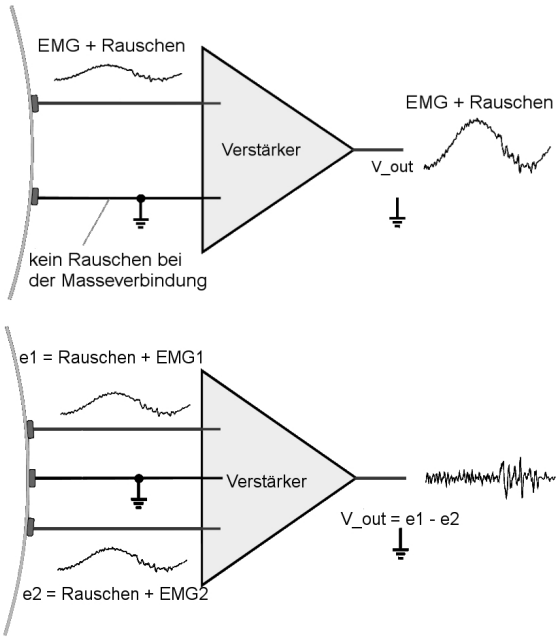

Abbildung 2.3: Vergleich unipolarer und bipolarer Ableitung bei Rauschen, nach [Winter, 1990]

Bei der *unipolaren* Ableitung wird die Spannungsdifferenz zwischen der relevanten Elektrode und einer Referenzelektrode gemessen, die auf elektrisch inaktivem Gebiet platziert wurde.
Bei der *bipolaren* Ableitung hingehen wird die Spannungsdifferenz zwischen zwei Elektroden auf aktivem Gebiet berechnet, die nebeneinander über dem gleichen Muskel angebracht wurden.
Der Vorteil der bipolaren Ableitung ist, dass eventuelle Störsignale mit hoher Wahrscheinlichkeit von beiden Elektroden erfasst und durch die Differenzbildung eliminiert werden, wie Grafik 2.3 zeigt.

Grafik 2.4 illustriert den zeitlichen Verlauf der Potentialdifferenz, die zwischen zwei Elektroden gemessen wird. Da beide Elektroden die Aktivität des Muskels erfassen, entspricht dies einer bipolaren Ableitung.

Die Messung der Spannungsdifferenz zwischen jeweils zwei Elektroden ergibt einen sogenannten *Kanal*, unabhängig davon, was als Referenzelektrode gewählt wurde.

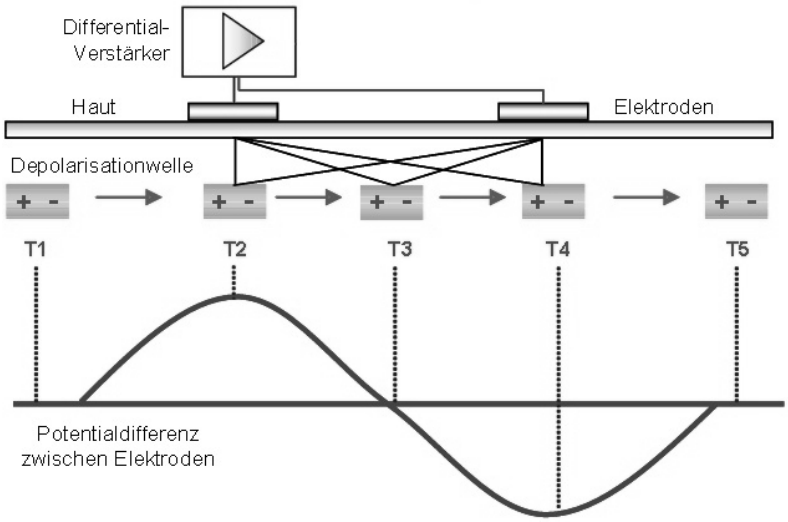

Abbildung 2.4: Zeitlicher Verlauf der Potentialdifferenz zwischen zwei Elektroden während der Erregungsausbreitung, nach [Kumar und Mital, 1996]

Nadelelektroden erlauben zwar eine Messung von Potentialänderungen mit hoher Ortsauflösung, sie sind jedoch invasiv und daher für den Einsatz in Silent Speech Interfaces ungeeignet. In dieser Arbeit werden ausschließlich Oberflächenelektroden verwendet, welche die Summe der Signale eines größeren Bereichs erfassen, wobei sich auch Potentiale verschiedener motorischer Einheiten überlagern können, wie in Grafik 2.5 dargestellt.

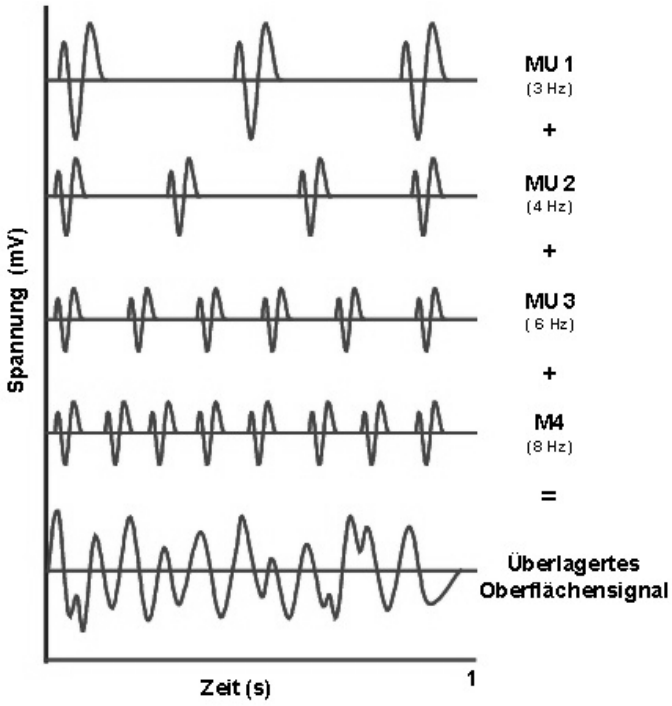

Abbildung 2.5: Überlagerung der Signale mehrerer motorischer Einheiten (Motor Units, MUs), nach [Kumar und Mital, 1996]

Ein weiteres Problem bei der Messung myoelektrischer Signale mittels Oberflächenelektroden neben der Überlagerung ist die Anfälligkeit für Störfaktoren, wie z.B.

Schwitzen. Außerdem hängt die Ausprägung des Signals von dem Gewebe ab, welches den Muskel von der Messelektrode trennt.

2.1.3 EMG-Features

Die Rohsignale der EMG-Kanäle unterliegen vielen Einflussfaktoren und sind daher sehr komplex. Um aus den Rohsignalen geeignete Features für die weiteren Berechnungen zu erhalten, wurden für jedes Rohsignal folgende Schritte durchgeführt (nach [Jou et al., 2006]):

1. Verstärkung des Signals
2. Gleichstrom-Normalisierung, um die Grundspannung zu entfernen
3. Hoch- und Tiefpassfilterung zur Artefaktentfernung
4. Sampling des gefilterten Signals
5. Berechnung der Features im Zeitbereich für den jeweiligen Kanal
6. Anwendung der *Linearen Diskriminanz-Analyse* (LDA), welche die Features auf diejenigen Dimensionen reduziert, welche am meisten zur Unterscheidbarkeit der verschiedenen Phoneme (s. Abschnitt 2.2.1) beitragen

Aus dem diskretisierten Rohsignal $x[n]$ können unterschiedliche sogenannte *EMG-Features* berechnet werden, die den zeitlichen Verlauf des Signals als eine Sequenz von *Frames* abbilden. Diese Frames bilden je nach Feature unterschiedliche Eigenschaften des Signals ab, wobei die Features ein- oder mehrdimensional sein können.

Das Rohsignal $x[n]$ wird vor der Berechnung mancher Features auch selbst vorverarbeitet. Das doppelt über eine Breite von 9 Samples gemittelte Signal $w[n]$ ist definiert als:

$$w[n] = \frac{1}{9} \sum_{k=-4}^{4} v[n+k] \text{ mit } v[n] = \frac{1}{9} \sum_{k=-4}^{4} x[n+k] \tag{2.1}$$

Weiterhin ist noch das Signal $p[n]$ definiert, zu dessen Berechnung die gemittelten Werte vom Rohsignal abgezogen werden, sodass nur hochfrequente Anteile übrig bleiben:

$$p[n] = x[n] - w[n] \tag{2.2}$$

und das Signal $r[n]$, das dem Absolutwert von $p[n]$ entspricht:

$$r[n] = |p[n]| \tag{2.3}$$

Die folgende Übersicht zeigt die Berechnung verschiedener Features für einen Frame mit Framegröße N:

Mean-Feature

Das *Mean-Feature* gibt den euklidischen Mittelwert des diskreten EMG-Signals an:

$$\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \tag{2.4}$$

Analog kann auch der euklidische Mittelwert $\bar{r}$ von $r[n]$ berechnet werden.

Power-Feature

Power-Features werden sowohl für das geglättete Signal $w[n]$ als auch für das hochfrequente Signal $r[n]$ berechnet:

$$P_w = \sum_{n=0}^{N-1} w[n]^2 \text{ bzw. } P_r = \sum_{n=0}^{N-1} r[n]^2 \tag{2.5}$$

Nulldurchgangsrate

Die *Nulldurchgangsrate* von $p[n]$ bildet ein weiteres für diese Arbeit relevantes Feature.

$$z = \text{Anzahl der Nulldurchgänge von } (p[0], \ldots, p[N-1]) \tag{2.6}$$

Time Domain-Feature

Alle vorher genannten Features werden in einem kombinierten Feature, genannt *Time Domain-Feature* (*TD*-Feature) zusammengefasst:

$$TD0 = (\bar{w}, P_w, P_r, z, \bar{r}) \tag{2.7}$$

Bei Voruntersuchungen wurde festgestellt, dass sich einzelne Features (Mean, log Power) für die Unit Selection nicht so gut eignen wie die kombinierten TD-Features. Durch die Verwendung der TD-Features wird außerdem die Vergleichbarkeit zu [Janke et al., 2012] geschaffen. Sie werden daher in dieser Arbeit grundsätzlich als EMG-Feature gewählt.

Um zeitkontextabhängige Informationen abzubilden, kann auf jegliche Features ein sogenanntes *Stacking* S der Stacking-Breite n angewendet werden. Dabei werden zu einem Frame jeweils die n vorausgehenden und folgenden Frames hinzugenommen, sodass ein kombinierter Featurevektor entsteht:

$$S(f_i, n) = (f_{i-n}, \ldots, f_{i+n}) \tag{2.8}$$

Für die TD-Features z.B. ergibt sich durch ein Stacking der Breite 15:

$$TD15 = S(TD0_i, 15) = (TD0_{i-15}, \ldots, TD0_{i+15}) \tag{2.9}$$

Nach dem Stacking haben die TD15-Features aus den Aufnahmen mit Einzelelektroden $6 \cdot (1 + 2 \cdot 15) = 186$ Dimensionen. Bei den Aufzeichnungen mit den Elektroden-Arrays wurden Störkanäle manuell entfernt, sodass von 35 Kanälen 18 übrig blieben. Dadurch ergaben sich für die Array-Sessions $18 \cdot (1 + 2 \cdot 15) = 558$ Dimensionen.
Da diese Dimensionalität einen sehr hohen Rechenaufwand mit sich bringt, wird zur Reduktion der Dimensionalität der Features eine LDA durchgeführt, die nur die k Dimensionen der Features beibehält, die am meisten zu ihrer Unterscheidbarkeit beitragen. In dieser Arbeit wurde $k = 25$ gewählt.

Da die Muskelsignale physiologisch vor dem akustischen Signal auftreten, werden sie auch entsprechend früher erfasst. Von [Jou et al., 2006] wurde festgestellt, dass dieser Versatz bei den gegebenen Aufnahmesessions ungefähr bei 50 ms liegt, er wurde daher durch die Verzögerung der EMG-Signale um diesen Wert ausgeglichen.

2.2 Sprache

In diesem Abschnitt werden die Grundlagen der Sprachproduktion, sowie die Repräsentation von Sprache mit Hilfe von diskreten Features erläutert.

2.2.1 Sprachproduktion

Hörbare Sprache besteht aus Schallwellen, die von Menschen auditiv wahrgenommen werden können. Diese Schallwellen werden von Menschen erzeugt, indem Luft aus der Lunge durch den Vokaltrakt strömt und dabei angeregt und moduliert wird.
Der menschliche Artikulationsapparat ist für die Erzeugung differenzierter Laute angepasst. Die Luft, die aus der Lunge durch die Luftröhre strömt, kann durch die einzelnen Komponenten des Artikulationsapparates verschieden moduliert werden, sodass unterschiedliche Laute entstehen. Die unterschiedlichen Artikulationsorte, wie in Grafik 2.6 gezeigt, sind nach [Catford, 1977]:

- Lippen (1,2)
- Zähne (3)
- Zahndamm (4,5)
- harter Gaumen (6,7)
- weicher Gaumen (8,9)
- Rachen (10)
- Stimmlippen (11)
- Kehldeckel (12)
- Zunge (13,14,15,16,17)

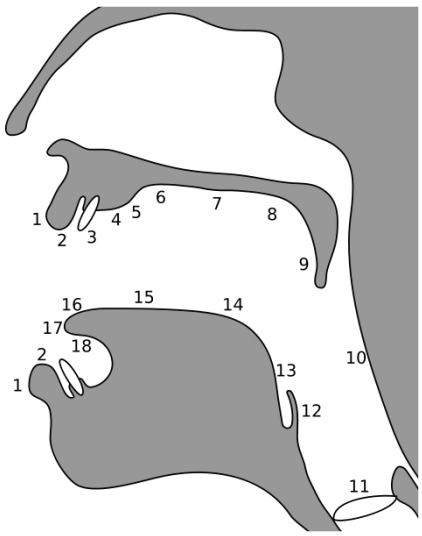

Abbildung 2.6: Schnitt des Vokaltrakts mit Kennzeichnung der Artikulationsorte, nach [Minifie et al., 1973] und [Catford, 1977]

Bei Frikativen (Reibelauten) wird beispielsweise an einem Artikulationsort eine Engstelle gebildet, durch die die Luft geleitet und verwirbelt wird, wie unter anderem bei f und s; bei Plosiven wird der Luftstrom erst vollständig unterbrochen und dann explosionsartig fortgesetzt, wie zum Beispiel beim k, wobei die Zunge und der Gaumen den Verschluss bilden, und beim p, bei dem die Lippen geschlossen sind.

Indem Luft aus der Lunge durch die Stimmritze gepresst wird, werden die Stimmlippen im Kehlkopf in Schwingungen versetzt und *stimmhafte* Sprache entsteht. Die Weite der Öffnung wird hierbei durch die *Stellknorpel* reguliert und führt zu einer Variation in der Höhe des erzeugten Tons. Diese Steuerung erfolgt willkürlich durch den Sprecher. Wenn die Stimmritze so weit geöffnet ist, dass die Luft ungehindert hindurchströmen kann und die Stimmlippen nicht schwingen, entstehen *stimmlose* Laute. Ihre Anregung ist dann lediglich ein ungeformtes Rauschen.

Die verschiedenen Laute einer Sprache werden *Phone* genannt und über Artikulationsort und -art definiert. Das Phon $[f]$ ist beispielsweise ein stimmloser labiodentaler Frikativ, d.h. die Engstelle zur Erzeugung des Reibelauts wird durch Lippen und Zähne gebildet.
Diejenigen Spracheinheiten einer Sprache, die zur Bedeutungsunterscheidung eines Wortes beitragen, werden als *Phoneme* bezeichnet. Das bedeutet, dass unterschiedliche *Phone*, z.B. ein gerolltes r und ein am Gaumen artikuliertes Zäpfchen-r, das selbe *Phonem* realisieren können, da Wörtern die selbe Bedeutung beigemessen wird, unabhängig davon ob sie mit gerolltem r oder mit Zäpfchen-r ausgesprochen werden.

2.2.2 Repräsentation von Sprache

Die Frequenz, mit der die Stimmlippen bei stimmhafter Sprache schwingen, wird als *Grundfrequenz* $F0$ der Sprache bezeichnet. Wenn stimmlose Laute ausgesprochen werden, ist die Grundfrequenz immer 0, da die Stimmlippen gar nicht in periodische Schwingungen versetzt werden.
Die Erzeugung der Grundanregung (periodische Schwingungen oder chaotisches Rauschen), kann als Signalquelle verstanden werden, deren Anregung mit einem Filter - dem Vokaltrakt - gefaltet wird, da das Signal nach Bedarf durch den Vokaltrakt moduliert wird. Hierbei werden manche Frequenzen verstärkt und andere abgeschwächt.
Bei der digitalen Repräsentation von Sprache, die in dieser Arbeit nach [Imai, 1983] verwendet wird, wird ebenfalls die Grundfrequenz als Signalquelle betrachtet. Die Filterung durch den Vokaltrakt wird durch die sogenannten *Mel-Frequenz-Cepstral-Koeffizienten* (engl. *Mel Frequency Cepstral Coefficients*, *MFCCs*) modelliert.
Beide Bestandteile gemeinsam geben folglich für einen Laut an, was bzw. welcher Laut (MFCCs) in welcher Tonhöhe (F0) gesprochen wurde. Für die Unterscheidung von Lauten ist der genau Wert der Grundfrequenz also meistens nicht relevant, sondern nur für die Sprachmelodie. Lediglich bei Phonen, die am gleichen Ort artikuliert werden, wobei eines stimmhaft und das andere stimmlos ist, ist die Grundfrequenz zur Unterscheidung relevant.

Die Berechnung der MFCCs aus dem Rohsignal besteht aus folgenden Schritten:

Wandlung vom analogen zum digitalen Signal
Das Signal wird sowohl im zeitlichen Verlauf digitalisiert (*Abtastung* oder *Sampling*) als auch im Wertebereich (*Quantisierung*).

Für das Sampling wird meist eine Rate von 16 kHz gewählt. Die Quantisierung wird mit 16 bit vorgenommen.

Fensterung

Bei der Fensterung wird das kontinuierliche Signal in eine Abfolge von einzelnen *Frames* umgewandelt. Ein Frame repräsentiert das Signal in einem festgelegten Zeitintervall entsprechend der *Framelänge*. Durch die Aneinanderreihung von Frames, deren Intervalle lückenlos aneinandergrenzen, wird der komplette Verlauf des Signals abgedeckt.
Die zeitliche Differenz zwischen dem Beginn zweier aufeinanderfolgender Frames wird *Frameshift* genannt. Es ist möglich, dass sich die Frames überlappen, wenn der Frameshift kleiner ist als die Framelänge.

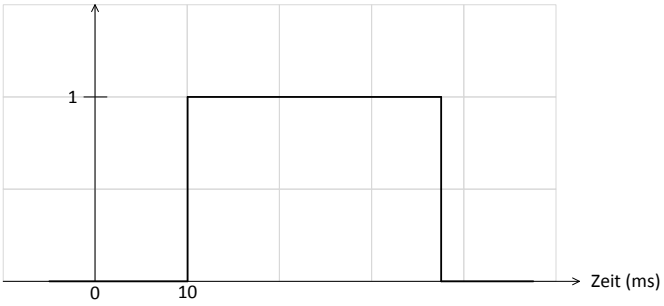

Abbildung 2.7: Rechteckfunktion, mit der der zweite Frame eines Signals berechnet wird bei einer Framegröße von 27 ms und einem Frameshift von 10 ms

Zur Berechnung der einzelnen Frames kann das Signal mit verschiedenen *Fensterfunktionen* multipliziert werden. Die direkte Aufteilung in Blöcke entspricht der Multiplikation mit einer Rechteckfunktion wie in Grafik 2.7, die in dem gewünschten Bereich 1 ist und sonst 0. Da das Signal nur endlich lang ist, entsteht bei der Fensterung ein fälschlicherweise verbreitertes Frequenzspektrum. Dieser *Leakage* genannte Effekt kann durch andere Fensterformen reduziert, jedoch nicht ganz vermieden werden.

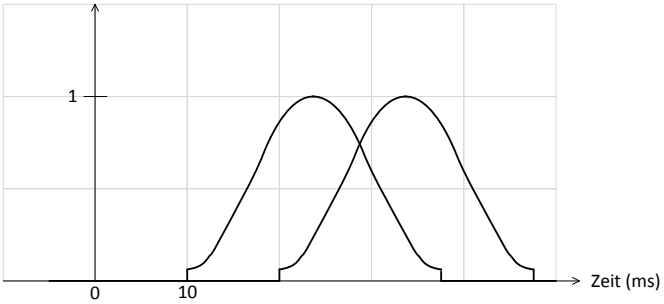

Abbildung 2.8: Zwei überlappende Hammingfenster, mit denen der zweite und dritte Frame eines Signals berechnet werden

Da die direkte Einteilung des Signals in aneinandergrenzende Blöcke diesen Effekt mit sich bringt und folglich meist andere Fenster als die Rechteckfunktion verwendet werden, überlappen sich die Frames oft, da es sonst zu einem Informationsverlust an den Verbindungsstellen kommt.

In dieser Arbeit wurden Hamming-Fenster mit einer Framelänge von 27 ms und einem Frameshift von 10 ms verwendet (s. Grafik 2.8).

Frequenzanalyse der Frames

Zur Frequenzanalyse werden die Frames mittels Diskreter Cosinus-Transformation (DCT) vom Zeit- in den Frequenzbereich transformiert.
Es wird lediglich das Betragsspektrum der Fouriertransformierten verwendet, da das Phasenspektrum der Frequenzen in der gegebenen Größenordnung für die menschliche Wahrnehmung nicht relevant ist.

Dimensionalitätsreduzierung durch Mel-Filterbank

Ein einzelner Frame mit 27 ms bildet bei einer Samplingrate von 16 kHz das Frequenzspektrum auf 432 Werte ab. Da diese Auflösung nicht erforderlich ist und viel redundante Information enthält, wird die Dimensionalität der Frames mittels einer *Filterbank* reduziert.

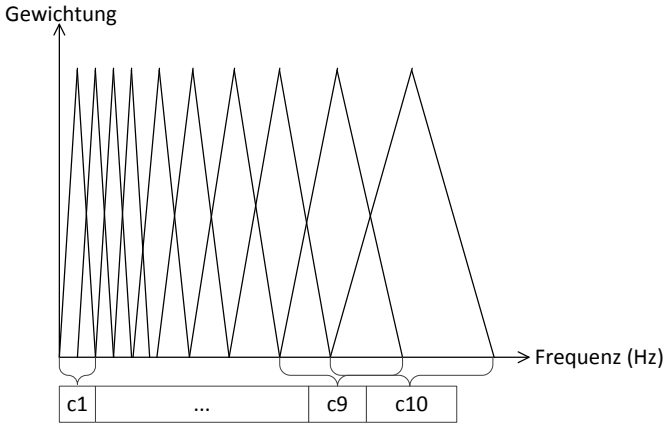

Abbildung 2.9: Vereinfachte Filterbank mit 10 einzelnen Dreiecksfiltern

Diese fasst jeweils bestimmte Frequenzbereiche zu einem Wert zusammen, indem die einzelnen Werte innerhalb des Bereichs gewichtet aufaddiert werden (s. Grafik 2.9). Die resultierenden Werte werden als *Mel-Spektral-Koeffizienten* bezeichnet (in der Grafik $c1$ bis $c10$). "Mel" steht hierbei für "Melody", die wahrgenommene Tonhöhe, da die Filterbank die Frequenzanteile angepasst an das menschliche Gehör skaliert.

Berechnung des Cepstrums

Da das menschliche Gehör Frequenzen logarithmisch unterscheidet, wird zunächst der Logarithmus des Spektrums berechnet. Anschließend wird mittels inverser DCT aus dem Spektrum das sogenannte *Cepstrum* im Zeitbereich gebildet, welches aus den Mel-Cepstral-Koeffizienten besteht.
Durch die Logarithmierung wird aus der Multiplikation von Grundfrequenz und Modulation durch den Vokaltrakt eine Addition. Die Grundfrequenz kann dadurch leicht als Spitze im Cepstrum erkannt und herausgefiltert werden, da sie in den MFCCs, die ja den Filter für die Anregung repräsentieren, nicht enthalten sein soll.

Aus den entstandenen F0s und MFCCs kann die Sprache in hinreichender Qualität rekonstruiert werden. In dieser Arbeit wurde das Frequenzspektrum mit der

Mel-Filterbank auf 25 Koeffizienten reduziert, da die Ausgabequalität durch die Einschränkung für das menschliche Gehör kaum merklich gemindert wird, wohingegen die Komplexität und Rechenzeit durch die Dimensionalitätsreduzierung deutlich sinken.

2.2.3 Geflüsterte Sprache

Geflüsterte Sprache unterscheidet sich von hörbarer Sprache darin, dass sie vollständig stimmlos ist. Die Stimmlippen sind geschlossen und schwingen nicht, lediglich die Stellknorpel der Stimmlippen bilden eine kleine Öffnung, durch die die Luft strömt, deren Anregung dann ein Rauschen ist. In der Folge entsteht wahrnehmbare, aber stimmlose Sprache, die als Flüstern bezeichnet wird.

In dieser Arbeit wurde geflüsterte Sprache für eine Vorstudie verwendet, bei der geflüsterte Sprache zu hörbarer Sprache konvertiert wurde. Diese Vorstudie wird in Abschnitt 4.2 beschrieben.

2.2.4 Phonem-Labels

Indem die akustischen Aufnahmen durch einen Spracherkenner in Phoneme, transkribiert werden, kann jedem Sprach-Frame ein Phonem-Label zugeordnet werden, das ausdrückt, welcher Laut in diesem Frame artikuliert wurde.
Für die Unit Selection selbst werden die Phonem-Labels nicht verwendet, sie werden lediglich zur Evaluation der Ausgabe benötigt (s. Abschnitt 5.2.3).

2.3 Evaluation der konvertierten Sprache

Dieser Abschnitt erläutert die Evaluationsmethoden und -maße, die zur Bewertung der Ausgabequalität herangezogen wurden.

2.3.1 Mel Cepstral Distortion

Ein objektives Maß für die Verzerrung von Lauten ist die *Mel Cepstral Distortion* (*MCD*) [Kubichek, 1993]. Sie wird verwendet um die MFCCs einer Äußerung frameweise mit denen einer Referenz zu vergleichen.
Mathematisch entspricht die MCD dem skalierten quadratischen Mittel der Differenzen zwischen den einzelnen d Dimensionen zweier Frames $f^{(1)} = (f_1^{(1)}, \ldots, f_d^{(1)})$ und $f^{(2)} = (f_1^{(2)}, \ldots, f_d^{(2)})$:

$$MCD(f^{(1)}, f^{(2)}) = \frac{10}{ln10}\sqrt{2 \cdot \sum_{i=1}^{d}(f_i^{(1)} - f_i^{(2)})^2} \tag{2.10}$$

In dieser Arbeit wurde d=25 für die MFCCs gewählt.

Da die MCD ausschließlich auf den MFCCs berechnet wird, ist sie ein gutes Maß für die *Verständlichkeit* einer Äußerung, sie sagt jedoch nichts über den Verlauf der Sprachmelodie aus. Daher wurde für die Evaluation ein zweites objektives Maß eingesetzt, das im folgenden Abschnitt beschrieben wird.

2.3.2 Erkennung der Stimmhaftigkeit

Um die Grundfrequenz-Komponente der Sprachausgabe zu evaluieren, wurde als Maß die Erkennungsrate der Stimmhaftigkeit verwendet. Sie gibt an, welcher Anteil der stimmhaften bzw. stimmlosen Frames im Vergleich mit der Referenz korrekt konvertiert wurde.

2.4 Verwandte Arbeiten

Im Bereich der Spracherkennung gibt es viele verschiedene Ansätze zur Erkennung von Sprache ohne Verwendung des Audiosignals (u.a. Silent Speech Interfaces, [Denby et al., 2010]), wie zum Beispiel die Erfassung der Bewegungen mittels im Mundbereich angebrachter Magneten in einem Magnetfeld ([Fagan et al., 2008]) oder mittels eines Mikrophons, das feinste Vibrationen an der Hautoberfläche erfasst ([Nakajima et al., 2006]). In dieser Arbeit wird jedoch nur die Verwendung von Elektromyographie mit Oberflächenelektroden behandelt.

2.4.1 EMG-basierte Spracherkennung

Die Erkennung von Sprache anhand der elektrischen Signale der Gesichtsmuskulatur wird schon seit den 80er Jahren erforscht.
[Morse und O'Brien, 1986] verwendeten 4 Edelstahl-Elektroden an der Schläfe und dem Hals um zu erforschen, ob aus den erfassten Signalen überhaupt Informationen gewonnen werden können. Hierzu wurden zwei Probanden untersucht, die in nicht-kontinuierlicher Sprache zweisilbige Wörter artikulierten. Bei den Experimenten mit 12 verschiedenen Wörtern konnte eine Erkennungsrate erreicht werden, die die *a priori*-Wahrscheinlichkeit der korrekten Worterkennung um das fünffache übersteigt.

[Sugie und Tsunoda, 1985] untersuchten die Erkennung von 5 japanischen Vokalen anhand der Signale dreier Gesichtsmuskeln in der Mundregion. Es wurde hierbei eine Erkennungsrate von im Schnitt 64% erreicht, was bedeutet, dass knapp zwei Drittel der artikulierten Vokale korrekt erkannt wurden.

Da ähnlich *klingende* Laute durch sehr unterschiedliche Stellungen des Artikulationsapparates erreicht werden können, wurde von [Chan et al., 2001] durch die Hinzunahme von EMG-Daten zu akustischen Aufnahmen eine bessere Erkennungsleistung, insbesondere in lauter Umgebung, erhofft. Als Vokabular für die Experimente wurden die englischen Ziffern verwendet und die einzelnen Wörter innerhalb des kontinuierlichen Sprachsignals anhand der Audioaufnahmen isoliert.
Die Fehlerraten lagen zwischen 2,3% bei kontinuierlicher und 6,5% bei zufälliger Wiederholung der Ziffern und damit wesentlich besser als in den früheren Experimenten.

[Jorgensen et al., 2003] untersuchten als Erste die Erkennung von Wörtern ausschließlich mit EMG-Daten, wobei die Probanden ein Vokabular von 6 Wörtern lautlos artikulierten. Dabei wurde mit nur zwei Elektroden bei Verwendung eines neuronalen Netzes eine Erkennungsleistung von 92% erreicht.

[Jou et al., 2006] und [Walliczek et al., 2006] verwendeten kleinere Spracheinheiten (Phoneme) anstelle ganzer Wörter, wodurch mit einer kleineren Menge an Trainingsdaten bessere Erkennungsraten erzielt werden können, da alle Wörter in die gleichen Bestandteile zerlegt werden können.

Der Phonem-basierte Ansatz wurde von [Schultz und Wand, 2010] um sogenannte *Phonetic Features* erweitert, die Eigenschaften wie Stimmhaftigkeit oder Artikulationsort modellieren. Auf einem kleinen Vokabular konnte dadurch wiederum eine Verbesserung erzielt werden.

2.4.2 EMG-basierte Sprachsynthese

Im Bereich der EMG-basierten Sprach*synthese* gibt es weniger Arbeiten, da auch die Möglichkeit besteht, mittels *Text-to-Speech* (*TTS*), anschließend an die EMG-basierte Spracherkennung, aus EMG-Daten Sprache zu erzeugen. Bei dieser Vorgehensweise gibt es jedoch durch die Erkennung und anschließende Synthese zwei potenzielle Fehlerquellen. Außerdem gehen prosodische Informationen bei der Umwandlung in Phoneme oder Wörter verloren.

[Toth et al., 2009] verwendete einen auf *Voice Conversion* (*VC*) basierenden Ansatz. Bei der VC wird die Stimme eines Menschen so modifiziert, dass sie wie die eines anderen Menschen klingt, es wird also Sprache auf Sprache abgebildet. Die VC wurde bei diesem Ansatz auf EMG-Daten angewendet, um diese auf Sprache abzubilden, da die Transformation von einer Quell- in eine Zielmenge ein Problem aus dem Bereich des Maschinellen Lernens darstellt und von der Art der Daten unabhängig ist. In diesem Fall wurde das Lern-Verfahren mittels *Gauss-Mixtur-Modellen* (GMMs) umgesetzt.
Die Generierung einer natürlichen Grundfrequenz aus EMG-Daten wurde bei dieser Arbeit übergangen, indem die Grundfrequenz eines parallel aufgenommenen Audiosignals zur Sprachsynthese verwendet wurde.

[Nakamura et al., 2011] untersuchte daher die Erzeugung von F0 aus EMG-Daten, ebenfalls mittels GMMs. Hierfür wurden die Frames zunächst mittels einer *Support Vector Machine* in stimmhaft und stimmlos unterteilt und anschließend mit den stimmhaften Frames das GMM trainiert. Durch diese Vorgehensweise wurden zwischen 73% und 84% der Frames korrekt als stimmhaft oder stimmlos konvertiert.

Der Unit Selection-Ansatz wurde von [Hunt und Black, 1996] entwickelt, um Sprache aus Text zu synthetisieren. Er besteht in der Auswahl passender kurzer Spracheinheiten und deren Verknüpfung. Es wurden auch phonetische Informationen für die Auswahl der passendsten Einheiten verwendet.
Mit kleineren Änderungen, wie z.B. das Weglassen der phonetischen Informationen, kann dieser Ansatz analog zu [Toth et al., 2009] ebenfalls verwendet werden, um EMG in Sprache zu transformieren, da die Quell- und Zielmengen auch unabhängig sind.

3. Unit Selection

In diesem Kapitel wird die Unit Selection nach [Hunt und Black, 1996] beschrieben. Hierzu gehören die Einführung der wichtigsten Begriffe sowie ein Überblick über die Vorgehensweise von der Erstellung der Units bis hin zur Auswahl der optimalen Unit-Sequenz. Die anschließende Synthese des Ziel-Sprachsignals wird in Abschnitt 4.5 erläutert.

3.1 Units

Die wichtigste Einheit bei der Unit Selection bilden die sogenannten Units. Sie verknüpfen Abschnitte von Quell- und Zielsignal (z.B. Muskelaktivität und Sprache) und ermöglichen dadurch die Konvertierung. Quell- und Zielsignal müssen hierfür einander zugeordnet werden können, entweder indem sie gleichzeitig aufgezeichnet werden, oder durch Alignment (s. Abschnitt 4.2.2), falls eine gleichzeitige Aufzeichnung nicht möglich ist.

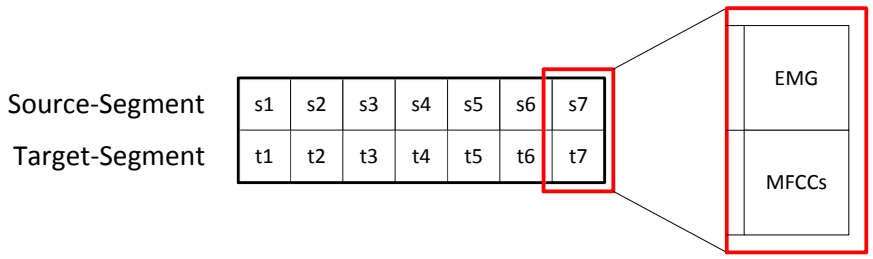

Abbildung 3.1: Unit mit Source-Segment aus Muskelsignalen und Target-Segment aus Sprache

Nachdem die Signale wie in Abschnitt 2.1.3 bzw. 2.2.2 beschrieben vorverarbeitet wurden, werden die entstandenen Feature-Sequenzen in kurze Blöcke, sogenannte *Segmente*, eingeteilt. Diese bestehen aus einer festen Anzahl Frames, also aufeinanderfolgenden Featurevektoren. Die so entstandene Sequenz von Featurevektoren

des Quellsignals, genannt *Source-Segment*, und die des Zielsignals, genannt *Target-Segment*, bilden zusammen eine *Unit*. So kann z.B. nachgebildet werden, welche Muskelaktivität welcher Lautäußerung entspricht.

Die Anzahl an Featurevektoren in einem Segment wird *Segmentgröße* (engl. *segment width*) w genannt, ihr Einfluss wird in Abschnitt 5.1.2 erörtert.

3.2 Codebook

Um für die Unit Selection eine möglichst große Datenbasis zu schaffen, werden alle Äußerungen der Trainingsdatenmenge folgendermaßen in Units aufgespalten:
Es wird eine Unit erzeugt, deren Source- und Target-Segment mit dem ersten Frame der jeweiligen Feature-Sequenz beginnen und die folgenden w Frames beinhalten, wobei w der Segmentgröße entspricht. Anschließend wird der Start-Index um einen Frame verschoben und der Vorgang so lange wiederholt, bis die übrigen Frames der Äußerung nicht mehr für ein vollständiges Segment ausreichen. Durch diese Vorgehensweise werden möglichst viele Units erzeugt, sodass die Auswahl für die Unit Selection später so vielfältig wie möglich ist.

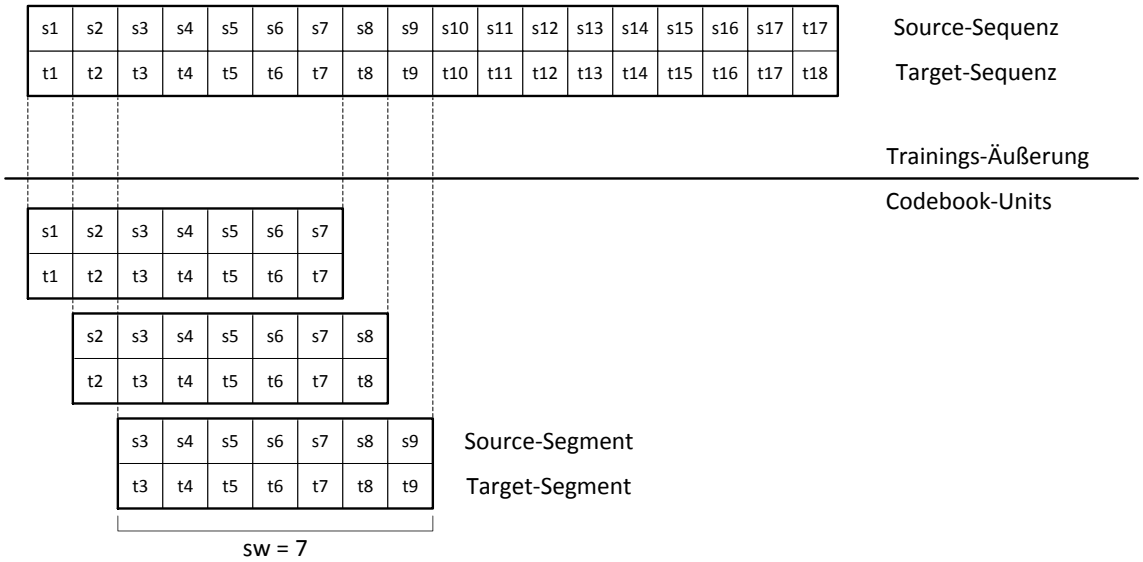

Abbildung 3.2: Erzeugung des Codebooks am Beispiel einer Trainings-Äußerung mit 17 Frames

Die Units, die auf diese Weise aus allen Trainingsäußerungen entstehen, werden im sogenannten *Codebook* gesammelt. Es stellt die Datenbank dar, aus der bei der Unit Selection die Units für die Erzeugung des akustischen Zielsignals ausgewählt werden. Die Größe des Codebooks hängt von der Menge der zur Verfügung stehenden Daten ab. Bei einer Menge von Trainingsdaten von ca. 30 Minuten und einem Frameshift von 27 ms enthält das Codebook beispielsweise ungefähr 150.000 Units.

Neben den für die Unit Selection notwendigen Segmenten enthalten die Units im Codebook zusätzliche Informationen:

Start- und Endindizes

Für die Synthese des Ziel-Sprachsignals direkt aus den akustischen Rohsigna-

len der Trainingsäußerungen (s. Abschnitt 4.5.2) werden in den Codebook-Units die Indizes von Start- und Endframe des Target-Segments innerhalb der Trainingsäußerung gespeichert. Zusätzlich zu den Indizes wird die ID der Trainingsäußerung in den Units gespeichert, um damit auf das ursprüngliche akustische Signal zugreifen zu können.

Phonem-Labels
Zur Evaluation der Unit Selection werden in den Codebook-Units diejenigen Phonem-Labels gespeichert, deren akustische Repräsentation das Target-Segment je Frame darstellt. So kann die Phonemsequenz ermittelt werden, die bei der Unit Selection ausgewählt wurde.
Die Labels werden allerdings nur zur Evaluation verwendet und haben keinen Einfluss auf die Unit Selection, da diese auch ohne Transkriptionen möglich sein soll.

Wenn das Zielsignal aus Sprache besteht, enthalten die Units zusätzlich zu den MFCCs ein weiteres, dem Target-Segment zugehöriges, Segment, das die Grundfrequenz $F0$ der akustischen Frames beinhaltet. Dieses Segment spielt für den Vorgang der Unit Selection keine Rolle; es wird lediglich zur Speicherung des F0-Verlaufs verwendet, der später für die Synthese benötigt wird.

3.3 Test-Segmente

Für die Unit Selection wird eine weitere Einheit benötigt, die den Codebook-Units ähnlich ist: die Test-Segmente.

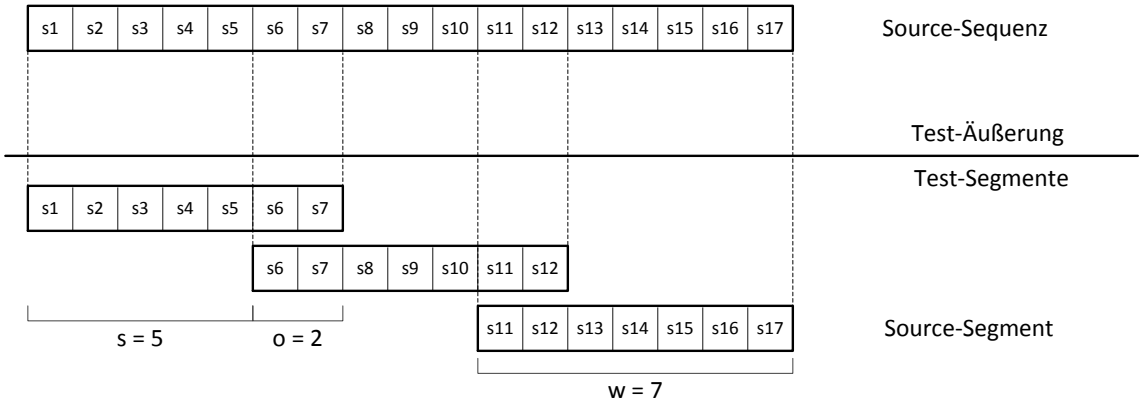

Abbildung 3.3: Erzeugung der Test-Segmente am Beispiel einer Test-Äußerung mit 16 Frames

Sie werden ebenfalls durch Aufteilung der Test-Feature-Sequenzen in Segmente der gleichen Größe w erzeugt, da aber die Target-Features bei den Test-Äußerungen unbekannt sind, bestehen die "Units" nur aus einem einzigen Segment, das aus der Source-Feature-Sequenz gewonnen wird. Die Segmente jeder Test-Äußerung werden geordnet in einem Vektor gespeichert, um damit später die Unit Selection durchzuführen.

Anders als bei den Codebook-Units wird bei der Aufteilung der Startindex für das nächste Segment gegebenenfalls nicht nur um einen einzigen Frame verschoben, son-

dern um eine bestimmte Anzahl Frames, den *Segment Shift* s. Zwei aufeinanderfolgende Segmente überlappen sich also um $o = (w - s)$ Frames, wobei o für die *Überlappung* (engl. *overlap*) steht.

3.4 Kostenfunktionen

Um diejenige Unit-Sequenz zu finden, die am besten geeignet ist um die gegebene EMG-Sequenz zu repräsentieren, werden zwei Arten von Kostenfunktionen verwendet, die im folgenden Abschnitt näher erläutert werden sollen: *Target Cost* und *Concatenation Cost*.

3.4.1 Target Cost

Die Target Cost dient als Maß für die Ähnlichkeit zweier *Source*-Segmente. Sie wird zwischen einem Test-Segment, das z.B. aus einem EMG-Featurevektor besteht, und dem Source-Segment einer Codebook-Unit berechnet und gibt also an, wie gut eine Codebook-Unit das gegebene Test-Segment repräsentiert. Je kleiner die Target Cost ist, desto ähnlicher sind sich die Segmente.

In dieser Arbeit wird die Target Cost folgendermaßen berechnet:
Zuerst wird für alle w Frames der Segmente die Distanz tc_i zwischen dem i-ten Frame des Test-Segments $s^{(1)}$ und dem i-ten Frame des Source-Segments $s^{(2)}$ der Codebook-Unit gemäß einer Distanzfunktion df aus Abschnitt 3.4.4 berechnet:

$$tc_i = df(s_i^{(1)}, s_i^{(2)}), i = 1, \ldots, w \tag{3.1}$$

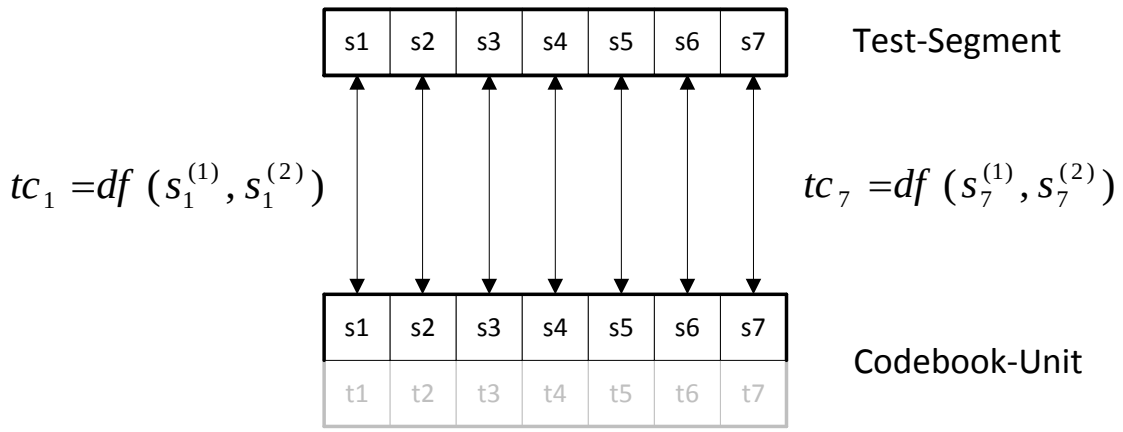

Abbildung 3.4: Berechnung der Target Cost zwischen einem Test-Segment und einer Codebook-Unit mit Segmentgröße $w = 7$

Anschließend wird aus diesen Distanzen der Mittelwert und damit die gesamte Target Cost tc gebildet:

$$tc = \frac{1}{w} \sum_{i=1}^{w} tc_i \tag{3.2}$$

Durch die Mittelwertbildung ist die Target Cost unabhängig von der Segmentgröße, was unter anderem für die Gewichtung der Kosten, die in Abschnitt 3.4.3 erläutert wird, von Bedeutung ist.

3.4.2 Concatenation Cost

Während die Target Cost dazu dient, die passendsten Codebook-Units für die *einzelnen* Segmente einer Test-Sequenz zu finden, wird durch die Concatenation Cost die Güte ihrer *Verbindungen* bewertet, d.h. wie natürlich der Übergang an der Schnittstelle zweier Segmenten ist.
Sie wird für zwei Codebook-Units folgendermaßen berechnet: Zunächst werden die einzelnen Distanzen cc_i zwischen zwei Frames mittels einer Distanzfunktion df aus Abschnitt 3.4.4 berechnet; jeweils zwischen den o überlappenden Frames der Target-Segmente $t^{(1)} = (t_1^{(1)}, \ldots, t_w^{(1)})$ und $t^{(2)} = (t_1^{(2)}, \ldots, t_w^{(2)})$:

$$cc_t = df(t_{s+i}^{(1)}, t_i^{(2)}), i = 1, \ldots, o \tag{3.3}$$

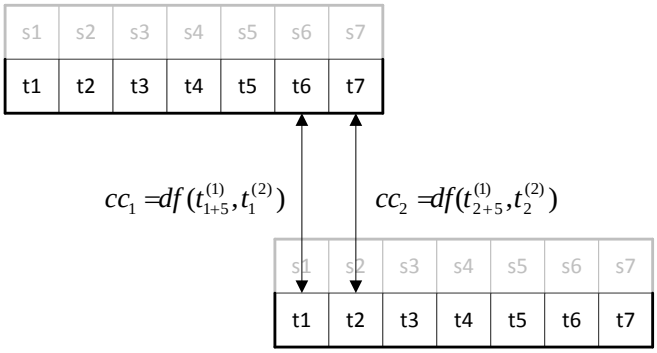

Abbildung 3.5: Berechnung der Concatenation Cost zwischen zwei Units mit $w = 7$ und $o = 2$

dann wird analog zur Target Cost der Mittelwert der Distanzen gebildet. Dadurch ist die Concatenation Cost ebenfalls unabhängig von der Segmentgröße w.
In dieser Arbeit wird für Target und Concatenation Cost die selbe Distanzfunktion verwendet, jedoch sind ihre Werte, wie in Abschnitt 3.4.3 gezeigt wird, nicht ohne Weiteres vergleichbar.
Die Concatenation Cost wird auf den *Target-Segmenten* berechnet, da es die Ausgabe ist, die möglichst flüssig und natürlich klingen soll. Je niedriger die Concatenation Cost ist, desto natürlicher ist der Übergang zwischen den Segmenten.
Für zwei Codebook-Units, die aus der selben Äußerung stammen und durch eine Verschiebung um den Segment Shift s entstanden sind (s. Grafik 3.6), ist die Concatenation Cost bei den verwendeten Distanzmaßen $cc = 0$, da die Frames, die verglichen werden, die selben sind. Mit

- i: Index des Frames innerhalb der Unit
- t: Startindex des ersten Segments innerhalb der Target-Sequenz der Trainingsäußerung
- s: Verschiebung des zweiten Segments gegenüber dem ersten

ist offensichtlich:

$$t_{t+(s+i)}^{(1)} = t_{(t+s)+i}^{(2)}, i = 1, \ldots, o \tag{3.4}$$

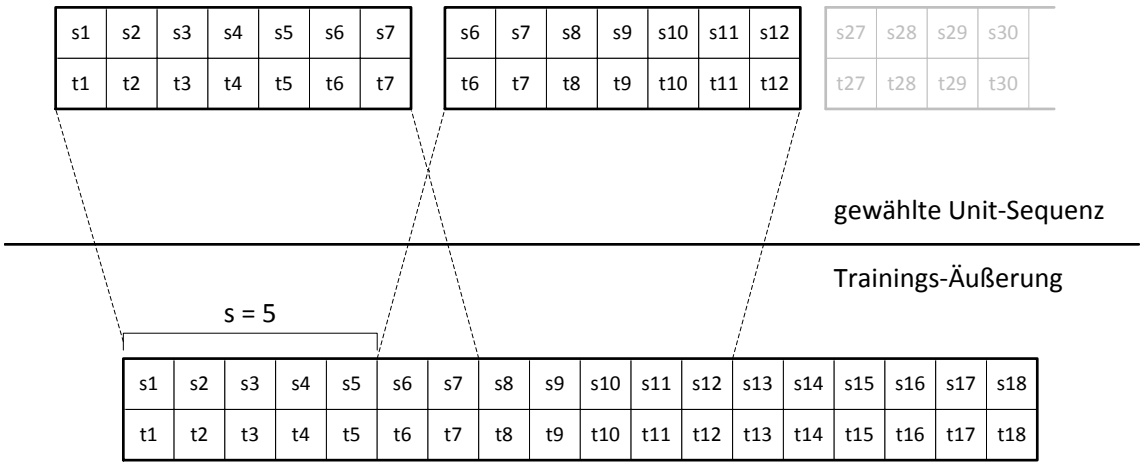

Abbildung 3.6: Zwei Units, die um s Frames verschoben aus einer Äußerung entstanden sind

Die dadurch entstehende Bevorzugung ist erwünscht, da der Übergang zwischen solchen Units die natürlichste Verbindung darstellt, die möglich ist.

Die Überlappung o der Segmente muss mindestens 1 Frame betragen, da die Concatenation Cost ansonsten nicht berechnet werden kann. Welche Auswirkungen die Anzahl der überlappenden Frames hat, wird in Abschnitt 5.1.3 erörtert.

3.4.3 Gewichtung der Kostenfunktionen

Da die Kostenfunktionen gegebenenfalls auf unterschiedlichen Arten von Features berechnet werden - die Target Cost auf den EMG-Features und die Concatenation Cost auf den akustischen Features - liegen die jeweiligen Kosten in unterschiedlichen Wertebereichen. Dadurch kann es passieren, dass eine der Kostenfunktionen durch ihre geringen Werte keinen Einfluss mehr auf die Wahl der Unit-Sequenz hat, was nicht erwünscht ist.

Um dieses Problem zu lösen, wurde in den Experimenten eine Gewichtung für die Kostenfunktionen bestimmt. Welche der Kostenfunktionen gewichtet wird ist irrelevant, da die absoluten Kosten für die Suche nach der optimalen Unit-Sequenz (s. Abschnitt 3.5) nicht von Belang sind, sondern lediglich ihr relatives Verhältnis.

Da Target- und Concatenation Cost als Mittelwert über das gesamte Segment gebildet werden, ist das Gewicht unabhängig von der Segmentbreite und muss nicht für jede Segmentbreite neu bestimmt werden.

Es kann jedoch kein eindeutiges Gewicht für alle Features festgelegt werden, da sich die EMG-Features in ihrem Wertebereich ebenfalls unterscheiden. Welche Gewichtung die besten Ergebnisse in den Experimenten geliefert hat, wird in Abschnitt 5.1.5 beschrieben.

3.4.4 Distanzmaße

Als Distanzmaße wurden für die Untersuchungen zwei verschiedene Größen verwendet: Die euklidische und die Cosinus-Distanz.

Die *euklidische Distanz* wird für zwei Featurevektoren $f^{(1)} = (f_1^{(1)}, \ldots, f_d^{(1)})$ und $f^{(2)} = (f_1^{(2)}, \ldots, f_d^{(2)})$ mit jeweils d Dimensionen berechnet als:

$$df_{euc} = \sqrt{\sum_{i=1}^{d} (f_i^{(1)} - f_i^{(2)})^2} \tag{3.5}$$

Sie ist also ähnlich zur Mel Cepstral Distortion, die in Abschnitt 2.3.1 beschrieben wird.

Die Formel für die *Cosinus-Distanz* ist:

$$df_{cos} = 1 - \frac{\sum_{i=1}^{d} f_i^{(1)} \cdot f_i^{(2)}}{\sqrt{\sum_{i=1}^{d} f_i^{(1)2}} \cdot \sqrt{\sum_{i=1}^{d} f_i^{(2)2}}} \tag{3.6}$$

Ihr Wertebereich liegt zwischen 0 und 2. Die Cosinus-Distanz ist zudem skaleninvariant, da sie den Cosinus des Winkels zwischen den beiden Featurevektoren im d-dimensionalen Raum misst, nicht aber deren Länge.

Welche Unterschiede sich bei der Verwendung der jeweiligen Distanzmaße ergeben, wird in Abschnitt 5.1.4 beschrieben.

3.5 Suche nach optimaler Unit-Sequenz

Mit Hilfe der nun bekannten Kostenfunktionen kann die Suche nach der optimalen Unit-Sequenz durchgeführt werden. In dieser Arbeit wurden zwei Ansätze verwendet: die einfache *Greedy*-Suche und die differenzierte *Viterbi*-Suche. Beide sollen im folgenden Abschnitt genauer beschrieben werden.
Die Suche kann in beiden Fällen als Suche in einem vollständig verbundenen Netzwerk verstanden werden, in dem die Units die Zustände bilden und die Target Cost die Kosten für eine Zustandsbelegung und die Concatenation Cost die Kosten für einen Zustandsübergang darstellen. So wird sowohl modelliert, wie gut die gewählten Segmente die Test-Sequenz abbilden, als auch wie gut ihre Verbindungen untereinander sind.

Abbildung 3.7 zeigt ein einfaches Netzwerk aus Codebook-Units und der zu konvertierenden Test-Sequenz. Aus Gründen der Übersichtlichkeit werden nur wenige Units und Segmente gezeigt. Die gestrichelten Linien markieren die Target Cost zwischen Test-Segment und Unit-Codebook und die durchgezogenen Linien repräsentieren die Concatenation Cost zwischen den Codebook-Units.
Schwarze Units und Linien markieren die gewählte Sequenz, graue Units und Linien zeigen mögliche Sequenzen an, die jedoch nicht gewählt wurden.
Da jede Codebook-Unit auf jede Codebook-Unit folgen kann, ist es auch möglich, dass zwei mal die selbe Unit aufeinanderfolgt. In diesem Beispiel jedoch wurden drei unterschiedliche Units als optimale Sequenz gewählt.

3.5.1 Greedy-Suche

Das wesentliche Merkmal der Greedy-Suche ist, dass sie in jedem Suchschritt nur jeweils *eine einzige* Unit-Sequenz beibehält und alle anderen verwirft.

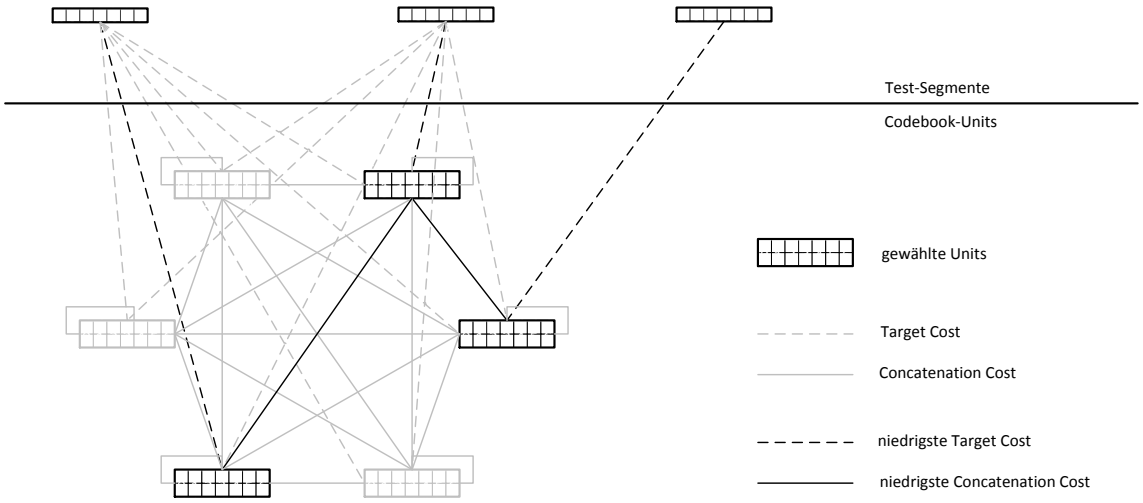

Abbildung 3.7: Netzwerk aus Codebook-Units und Test-Segmenten

Dazu wird in einem ersten Schritt zunächst die Codebook-Unit mit der niedrigsten Target Cost zum ersten Test-Segment bestimmt. Da diese Unit keinen Vorgänger hat, ist die Concatenation Cost in diesem Schritt nicht relevant. Folglich wird diese Unit als erste in der optimalen Sequenz gewählt. Die folgenden Schritte werden dann für die übrigen Test-Segmente gemäß ihrer Reihenfolge wiederholt:

1. Die k Codebook-Units mit der niedrigsten Target Cost tc_k zum aktuellen Test-Segment werden als mögliche Kandidaten für die optimale Unit bestimmt.

2. Für diese k Kandidaten wird jeweils die Concatenation Cost cc_k zu der im vorigen Suchschritt gewählten Unit berechnet.

3. Die Unit mit den geringsten Gesamtkosten $c_k = tc_k + w_{cc} \cdot cc_k$ wird als optimale Unit für den aktuellen Suchschritt gewählt, wobei w_{cc} das Gewicht für die Concatenation Cost bezeichnet, das in Abschnitt 3.4.3 erläutert wird.

Schließlich wird die Sequenz der schrittweise optimalen Units als Ergebnis ausgegeben.

3.5.2 Viterbi-Suche

Der Viterbi-Suchalgorithmus ermöglicht es theoretisch, den tatsächlich optimalen Pfad durch das Netzwerk zu finden, allerdings ist eine vollständige Abdeckung aufgrund der Ausmaße des Netzwerks nicht sinnvoll, da die benötigte Rechenzeit dann stark ansteigt. Deshalb wird die Suche in der praktischen Umsetzung auf eine bestimmte Anzahl Pfade eingeschränkt.

Die Vorgehensweise bei der Suche ist wie folgt:
Zunächst wird für jede Codebook-Unit die Target Cost tc zum aktuellen Test-Segment berechnet.

1. Die k Units mit den niedrigsten Pfad-Gesamtkosten bis zu diesem Punkt werden als Zustände $u_k^{(t)}$ im Suchgraphen beibehalten, alle anderen werden zur Minimierung der Rechenzeit deaktiviert und nicht weiter betrachtet.

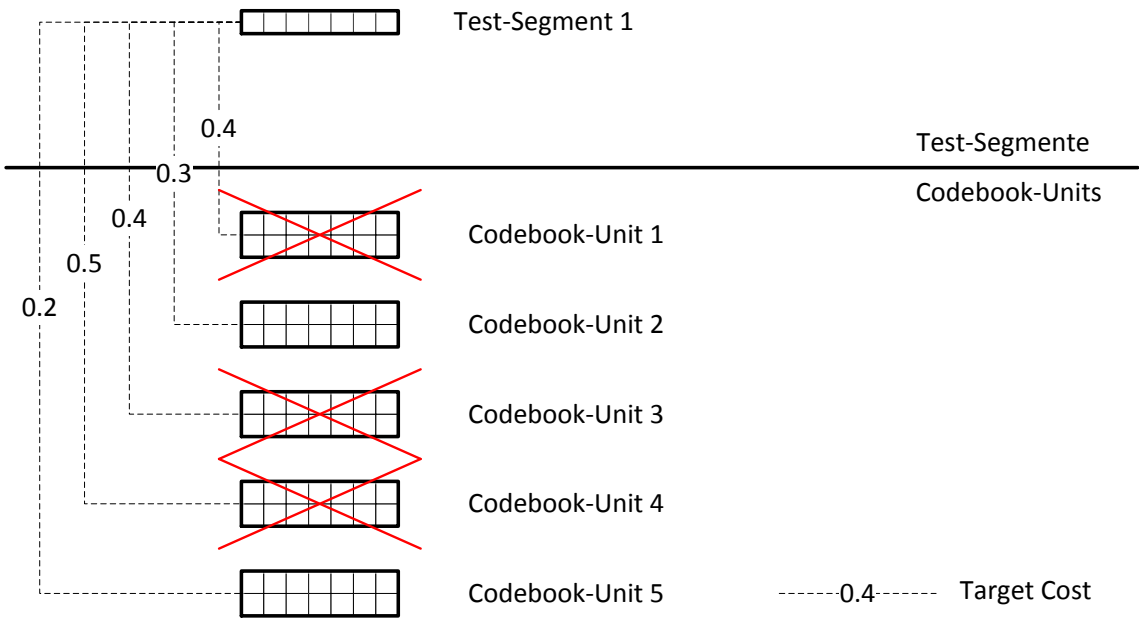

Abbildung 3.8: Viterbi-Graph nach dem ersten Schritt: die Pfade mit den höchsten Gesamtkosten werden deaktiviert

2. Für jede Codebook-Unit (Zustand u_k^{t+1}) wird die Concatenation Cost cc zu jedem der k möglichen Vorgängerzustände berechnet.

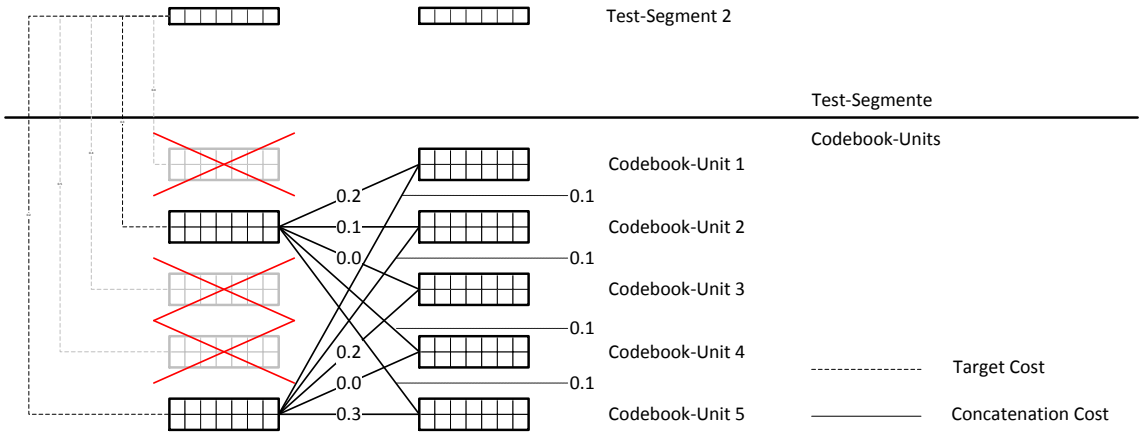

Abbildung 3.9: Viterbi-Graph nach dem zweiten Schritt: die Concatenation Cost wird berechnet

3. Der Vorgängerzustand zu dem die Concatenation Cost minimal ist und seine jeweiligen Vorgänger werden als bestmöglicher Pfad zu dem aktuellen Zustand festgelegt und die Gesamtkosten des Pfades um die Target Cost tc zum aktuellen Test-Segment und die Concatenation Cost erhöht.
   Die Target Cost zum aktuellen Test-Segment ist an dieser Stelle nicht entscheidungsrelevant, da sie von den potenziellen Vorgängern unabhängig ist und es

nur darum geht, auf welchem Pfad die Codebook-Unit für dieses Segment am besten erreicht werden kann.

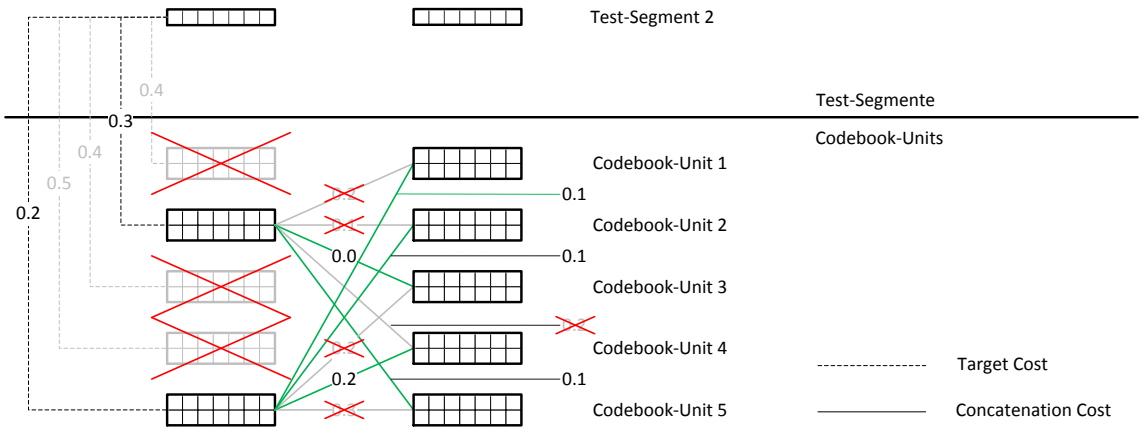

Abbildung 3.10: Viterbi-Graph nach dem dritten Schritt: die optimalen Pfade zu allen Codebook-Units werden festgelegt

Wenn die Suche den Graphen vollständig aufgebaut hat, wird der Pfad mit den niedrigsten Gesamtkosten als optimale Unit-Sequenz gewählt. Da immer der Pfad mit den geringsten Kosten zu jedem Zustand gewählt wird, ergibt sich nur durch die Einschränkung auf k Pfade ein Fehler. Dessen Einfluss wird in Abschnitt 5.1.6 evaluiert.

Bei einem Gewicht der Concatenation Cost von $w_{cc} = 0$ entspricht die Viterbi-Suche mit n Pfaden einer Greedy-Suche mit n Kandidaten, da in beiden Fällen in jedem Suchschritt die Unit mit der geringsten Target Cost gewählt wird.

3.5.3 Separater Suchpfad für F0

Als weiterer Ansatz wurde überlegt, die Grundfrequenz ebenfalls in die Suche mit einzubeziehen und eine separate Viterbi-Suche auf einem Netzwerk auf F0-Segmenten durchzuführen.
Diese Vorgehensweise führt aber dazu, dass überproportional viele Frames als stimmlos klassifiziert werden. Dies ist darauf zurückzuführen, dass die Concatenation Cost zwischen zwei stimmlosen Frames 0 beträgt, wohingegen sie zwischen zwei stimmhaften Frames meistens > 0 ist, da diese selten genau den selben Wert haben. Dadurch werden Folgen von stimmlosen Frames bevorzugt, was zu der genannten Überpriorisierung der stimmlosen Segmente führt. Der Ansatz wurde daher nicht weiter verfolgt.

4. Experimente

In diesem Kapitel werden zunächst die Aufzeichnung der EMG-Signale und deren Vorverarbeitung beschrieben, sowie ein Überblick über die verwendeten Daten gegeben. Anschließend werden die verschiedenen Experimente erläutert, die im Rahmen dieser Arbeit durchgeführt wurden.

4.1 EMG-Daten

Dieser Abschnitt behandelt den Aufbau der EMG-Daten und den Daten-Korpus, der in dieser Arbeit verwendet wurde.

4.1.1 Aufnahmesetup

In dieser Arbeit wurden zweierlei Elektrodensetups für die EMG-Aufzeichnung genutzt:

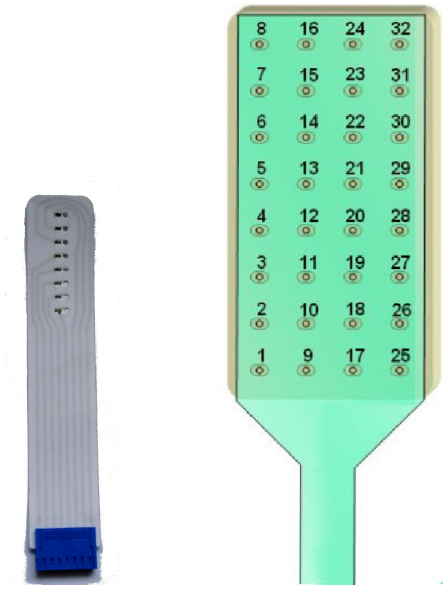

Abbildung 4.1: Zwei Elektroden-Arrays, mit 8 (links) bzw. 32 Elektroden (rechts)

Elektroden-Array

Für das Array-Setup aus [Wand et al., 2013] werden zwei Arrays eingesetzt, von denen das größere eine Matrix von 8x4 Elektroden beinhaltet und auf der Wange angebracht wird. Zwischen den einzelnen Elektroden ist jeweils ein Abstand von 10 mm. Ein kleineres Array mit einer Reihe aus 8 Elektroden wird unter dem Kinn befestigt, dessen Elektroden 5 mm voneinander entfernt sind. Bei der bipolaren Ableitung ergeben sich aus dem 8x4-Array 28 Signal-Kanäle und aus dem kleineren Array 7 Kanäle, da die Spannungsdifferenz jeweils zwischen zwei benachbarten Elektroden in einer Spalte gemessen wird.

Abbildung 4.2: Probandin mit angebrachten Elektroden-Arrays

Einzelelektroden

Als “Einzelelektroden” werden Elektroden bezeichnet, die unabhängig voneinander im Gesicht angebracht werden können.

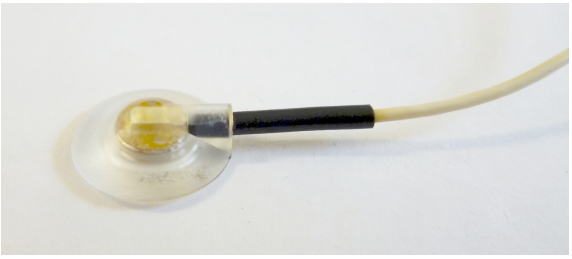

Abbildung 4.3: Einzelne Elektrode

In dieser Arbeit wurden die Einzelelektroden so angebracht, dass sie in Anlehnung an [Maier-Hein et al., 2005] die Signale folgender Muskeln erfassen, wie in Grafik 4.4 abgebildet:

- M. digastricus, Venter anterior (1-2)
- M. levator anguli oris (2,3)
- M. zygomaticus major (2,3)
- Platysma (4,5)
- M. depressor anguli oris (5)
- Zunge (1-2,6)

So wurden in dieser Arbeit aus dem Einzelelektroden-Setup mit 11 Elektroden 6 Kanäle gewonnen, indem bei zwei Elektrodenpaaren (weiß in der Grafik, 2,6) eine bipolare Ableitung und bei den übrigen eine unipolare Ableitung angewendet wurde. Die Referenzelektroden für die unipolare Ableitung wurden an der Nase (1-1) und hinter den Ohren (nicht bezeichnet) angebracht.

Abbildung 4.4: Probandin mit angebrachten Einzelelektroden; weiß markiert bipolare, schwarz unipolare Ableitung (Ausnahme: Elektrodenpaar 1-1 und 1-2)

Beim Einzelelektroden-Setup sowie auch beim Array-Setup können beide Arten der Ableitung eingesetzt werden. Wie in Abschnitt 2.1 beschrieben, ist die bipolare Ableitung jedoch vorteilhaft, da sie robuster gegen Störsignale ist.

4.1.2 Datenkorpus

Als Datengrundlage für diese Arbeit wurden 4 Sessions verwendet, die von zwei männlichen Sprechern aufgezeichnet wurden. Beide Sprecher sprechen Deutsch als Muttersprache und Englisch als Fremdsprache. Bei den Sprechern wurde für jeweils eine der Sessions ein Elektroden-Array und für die andere das Einzelelektroden-Setup eingesetzt.

Der Textkorpus basiert auf [Schultz und Wand, 2010] und besteht aus phonetisch balancierten Sätzen in englischer Sprache aus dem Nachrichtenbereich. Ca. 500 der Sätze wurden als Trainingsdaten verwendet. 10 weitere Sätze - für alle Sessions die selben Äußerungen - wurden als Testdaten verwendet und gegebenenfalls mit mehreren Wiederholungen aufgezeichnet. Mit diesen Aufnahmen ergeben sich die Größen für den Datenkorpus wie in Tabelle 4.1 gezeigt.

Sprecher - Setup	Dauer in min:sek		Anzahl Sätze	
	Training	Test	Training	Test
Sprecher 1 - Einzelelektroden	27:10	01:19	500	20
Sprecher 1 - Elektroden-Array	31:01	00:47	500	10
Sprecher 2 - Einzelelektroden	26:54	00:49	496	13
Sprecher 2 - Elektroden-Array	25:44	01:10	500	20
Gesamt	110:49	04:05	1996	63

Tabelle 4.1: Größe des vorhandenen Datenkorpus, nach Sprecher und Trainings-/Testmenge unterteilt

Für die Aufnahmen wurden die Sätze dem Sprecher in zufälliger Reihenfolge präsentiert und von diesem mit normaler, hörbarer Stimme vorgelesen. Die Sprache wurde hierbei mit einem Standard-Mikrophon aufgenommen. Zur Synchronisierung der Modalitäten wurde ein analoger Marker verwendet.
Für die Testäußerungen wurden ebenfalls sowohl EMG als auch Sprache aufgezeichnet, wobei die Sprache lediglich zur Evaluation als Referenz und nicht für die Unit Selection verwendet wird.

4.2 Whisper/Audible-Konvertierung

Als erstes Vorexperiment wurde die Konvertierung von geflüsterter zu hörbarer Sprache mit den Daten von [Janke et al., 2014] untersucht. Bei der Erzeugung der Units gibt es bei dieser Konvertierung eine Herausforderung, die in Abschnitt 4.2.2 beschrieben wird.

4.2.1 Units

Die Units bestehen bei dieser *Whis2Aud*-Konvertierung ausschließlich aus Sprach-Features. Das Source-Segment aus geflüsterter Sprache und das Target-Segment aus hörbarer Sprache werden jeweils wie in Abschnitt 3.1 beschrieben aus den Trainings-Äußerungen erzeugt. Da bei der geflüsterten Sprache die Grundfrequenz ohnehin immer 0 ist (s. Abschnitt 2.2.3), ist diese kein Bestandteil der Codebook-Units und der Test-Segmente, lediglich die Grundfrequenz der Target-Features wird für die spätere Synthese gespeichert.

4.2.2 Alignment

Im Gegensatz zu myoelektrischen Signalen und hörbarer Sprache können geflüsterte und hörbare Sprache nicht parallel aufgenommen werden. Um die Units aus geflüsterten und hörbaren Segmenten erzeugen zu können ist daher ein Alignment der geflüsterten Daten auf die hörbaren Äußerungen nötig. Dieses wurde mittels *Dynamic Time Warping* (DTW) berechnet. So entstehen zwei alignte Äußerungen mit gleicher Frameanzahl, die wiederum in alignte Segmente unterteilt werden können. Das Problem bei dieser Vorgehensweise liegt in der Ausgabequalität des DTW-Algorithmus. Wenn man die alignten akustischen Features zu Sprache resynthetisiert, klingt diese verzerrt, da manche Abschnitte besonders in die Länge gezogen werden. Entsprechend ist es nicht möglich, durch die Unit Selection bei *Whis2Aud* eine fehlerfreie Ausgabe zu erreichen, da schon die Eingabe fehlerbehaftet ist.

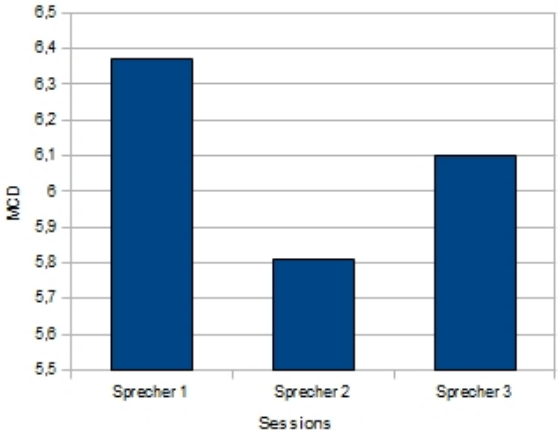

Abbildung 4.5: MCDs der Sessions beim *Whis2Aud*-Verfahren

Dies zeigt sich in Grafik 4.5, aus der zu erkennen ist, dass die MCD vergleichsweise hoch ist (vgl. Grafik 4.7). Auch die Spektrogramme in Grafik 4.6 weisen sichtliche Unterschiede auf.

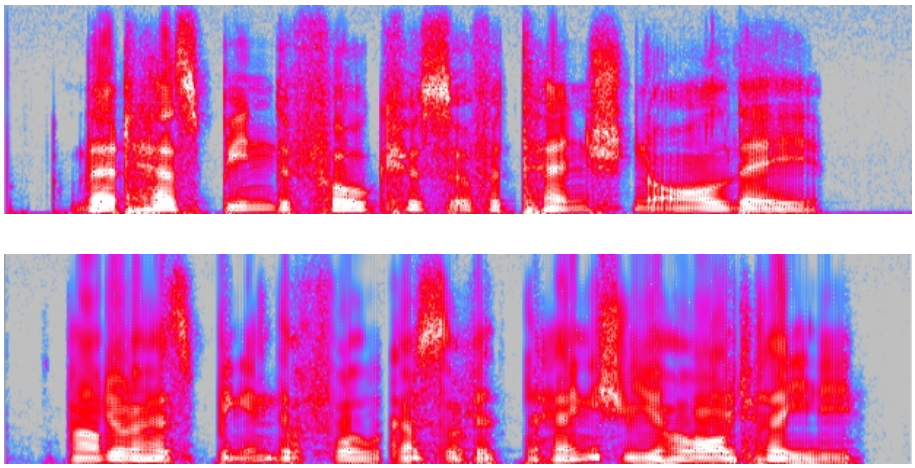

Abbildung 4.6: Spektrogramme der Äußerung *"That has been the focus of attention all day"*: Original (oben), mit *Whis2Aud* konvertiert (unten)

4.3 Oracle-Experiment: Audible/Audible-Konvertierung

Um nach den Problemen mit dem Alignment der Äußerungen zu überprüfen, ob der Unit Selection-Ansatz sinnvoll und machbar ist, wurde als Oracle-Experiment eine Audible/Audible-Konvertierung, genannt *Aud2Aud*-Verfahren, durchgeführt.
Hierbei bestehen sowohl Source- als auch Target-Segment der Units aus *MFCC-Featurevektoren*. Die Test-Segmente werden ebenfalls aus hörbarer Sprache gebildet. So wird eine Test-Äußerung, die in hörbarer Sprache vorliegt, aus den akustischen Daten im Codebook neu synthetisiert, ohne dass die Äußerung in der Trainingsdatenmenge enthalten war.
Diese Vorstudie sollte zeigen, ob es möglich ist mit dem gewählten Vorgehen und der gegebenen Datenmenge eine akzeptable Ausgabequalität zu erzielen. Der Vorteil darin, hörbare Sprache zu hörbarer Sprache zu konvertieren besteht in der höheren Robustheit akustischer Aufnahmen gegenüber EMG-Aufnahmen, wo es mehr Störquellen gibt und in der besseren Vergleichbarkeit der Daten.

Das Codebook wird wie in Abschnitt 3.2 beschrieben erzeugt, allerdings wird das Segment mit den akustischen MFCC-Features sowohl als Target als auch als Source verwendet. Im Codebook werden also kurze Sprach-Segmente mit sich selbst verknüpft. Die Suche kann hierdurch wie in 3.5 durchgeführt werden. Da die Test-Segmente ebenfalls Vektoren akustischer Features sind, wird hörbare Sprache in hörbare Sprache umgewandelt.

4.3.1 Menge der Trainingsdaten

Für die *Aud2Aud*-Konvertierung wurden die selben Sessions wie in Abschnitt 4.1.2 verwendet, jedoch nur jeweils die Sprachaufnahmen.

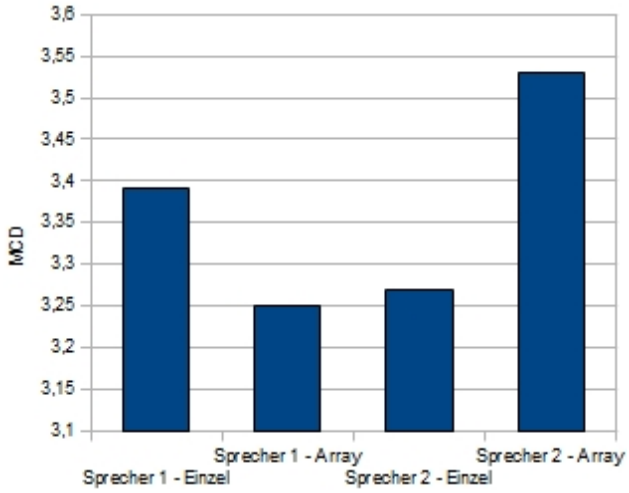

Abbildung 4.7: MCDs der Sessions beim *Aud2Aud*-Verfahren

Wie aus Grafik 4.7 ersichtlich ist, ergibt sich bei der *Aud2Aud*-Konvertierung eine durchschnittliche MCD zwischen 3,25 und 3,55. Dieser Wert kann als mäßige Verzerrung interpretiert werden, die die Verständlichkeit nur gering einschränkt.

Im Vergleich der Spektrogramme in Grafik 4.8 kann man erkennen, dass die Frequenzbänder durch die Konvertierung gut beibehalten werden, wenngleich das Spek-

trum bedingt durch die angewandte Syntheseart (s. Abschnitt 4.5.1) verwischt wurde.

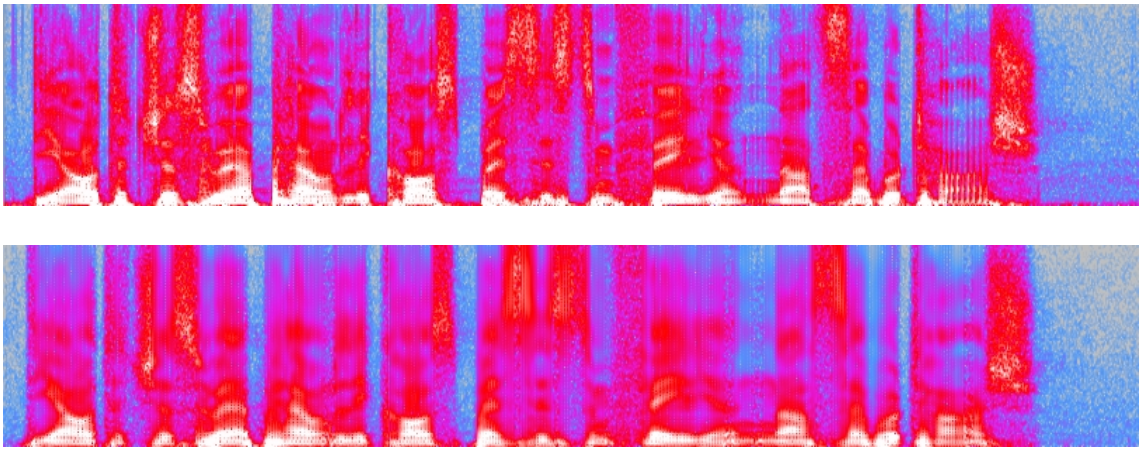

Abbildung 4.8: Spektrogramme der Äußerung *"The outages were apparently caused by system failure not sabotage"*: Original (oben), mit *Aud2Aud* konvertiert (unten)

Die Ergebnisse der *Aud2Aud*-Konvertierung wurden so interpretiert, dass die vorhandene Datenmenge für eine akzeptable Ausgabequalität bei der Unit Selection ausreichend ist.

4.3.2 Framegröße

Des Weiteren wurde die *Aud2Aud*-Konvertierung durchgeführt um zu überprüfen, welche Framegröße bei der Berechnung der akustischen Features (s. Abschnitt 2.2.2) am geeignetsten für die Synthese ist.
Hierbei ergab sich, dass die berechnete MCD mit steigender Framegröße kleiner wurde. Dies ist jedoch darauf zurückzuführen, dass sich die Feature-Vektoren in jeder Dimension einem Mittelwert annähern, wenn die Frames ein zu breites Fenster abdecken. Die Features werden sozusagen geglättet, wodurch sich eine scheinbar bessere MCD ergibt, da sich sowohl die Referenz als auch die konvertierten Features dem Mittelwert annähern. Die konvertierte Sprache klingt jedoch entsprechend verwaschen; im Extremfall mit einer Framebreite, die der Länge der Äußerung entspricht, wäre jeder Frame gleich. Folglich ist die MCD zwischen Referenz- und synthetisierter Äußerung mit gleicher Framegröße kein geeignetes Maß.
Wenn man die MCD hingegen zwischen der Referenz mit einer geringen Framegröße und den konvertierten Features mit einer höheren Framegröße berechnet, um diesen Glättungseffekt zu vermeiden, zeigt sich, dass die MCD bei einer Framegröße von 27 ms am geringsten ist. Daher wurde diese Framegröße für die weiteren Experimente beibehalten.

4.4 EMG/Audible-Konvertierung

Nach den Vorstudien, in denen hörbare Sprache zu hörbarer Sprache (*Aud2Aud*) und geflüsterte Sprache zu hörbarer Sprache (*Whis2Aud*) konvertiert wurden, wurde der Unit Selection-Ansatz auch auf myoelektrische Features angewendet (*Emg2Aud*-Konvertierung).

4.4.1 Units

Indem EMG und hörbare Sprache eines Sprechers gleichzeitig aufgezeichnet werden, entstehen Signalaufnahmen, die einander zugeordnet werden können. Es ist also im Gegensatz zur *Whis2Aud*-Konvertierung kein Alignment nötig.
Die Source-Segmente der Codebook-Units und die Test-Segmente bestehen bei der *Emg2Aud*-Konvertierung aus den myoelektrischen Featurevektoren. Die Target-Segmente werden analog zu denen der *Aud2Aud*- bzw. *Whis2Aud*-Konvertierung erzeugt, wiederum mit dem F0-Segment zur Speicherung der Grundfrequenz für die Synthese, die im nächsten Abschnitt beschrieben wird.

4.5 Synthese des akustischen Ziel-Signals

Um aus der gewählten Unit-Sequenz Sprache zu synthetisieren, wurden zwei Verfahren eingesetzt, die im Folgenden erläutert werden: Die *Mel Log Spectrum Approximation* (MLSA) nach [Imai, 1983] und die direkte Synthese aus den originalen akustischen Signalen der Trainingsaufnahmen.

4.5.1 Mel Log Spectrum Approximation (MLSA)

Bei der MLSA wird aus MFCC-Featurevektoren und F0 Sprache synthetisiert. Die Überlappung der Units um mindestens einen Frame macht jedoch eine gesonderte Behandlung der überlappenden Frames erforderlich.
In Anlehnung an [Wu et al., 2013] wurde eine Gewichtsfunktion gewählt, die mittels normalisierter Gewichte die einzelnen MFCC-Frames in jeder Dimension gewichtet aufaddiert, wie in Grafik 4.9 dargestellt wird. So entsteht für die überlappenden Frames ein einzelner vereinigter Featurevektor mit der selben Dimensionalität, wodurch gleichzeitig die Übergänge zwischen den Units geglättet werden.

Die Gewichte, mit denen die Frames multipliziert werden, werden folgendermaßen berechnet:

$$w[i] = \frac{exp(-\lambda \mathrm{a}[i])}{\hat{w}}, i = 1 \dots n \tag{4.1}$$

mit

$$\mathrm{a}[i] = \begin{cases} [\frac{n}{2}, \frac{n}{2} - 1, \dots, 1, 1, \dots, \frac{n}{2} - 1, \frac{n}{2}] & n \text{ gerade} \\ [\lceil \frac{n}{2} \rceil, \lceil \frac{n}{2} \rceil - 1, \dots, 1, \dots, \lceil \frac{n}{2} \rceil - 1, \lceil \frac{n}{2} \rceil] & n \text{ ungerade} \end{cases} \tag{4.2}$$

wobei n die Anzahl der überlappenden Units an der betroffenen Stelle und exp die Exponentialfunktion bezeichnet.

Die Gewichte werden mittels des Faktors $\hat{w}$ normalisiert:

$$\hat{w} = \sum_{i=1}^{n} exp(-\lambda \mathrm{a}[i]) \tag{4.3}$$

Dadurch ergibt die Summe der Gewichte 1 und die akustischen Features werden nicht verzerrt.

Der Gewichtungsfaktor λ dient der variablen Gewichtung von Frames. Für $\lambda > 0$ werden Frames stärker gewichtet, je näher sie an der Mitte der enthaltenden Unit

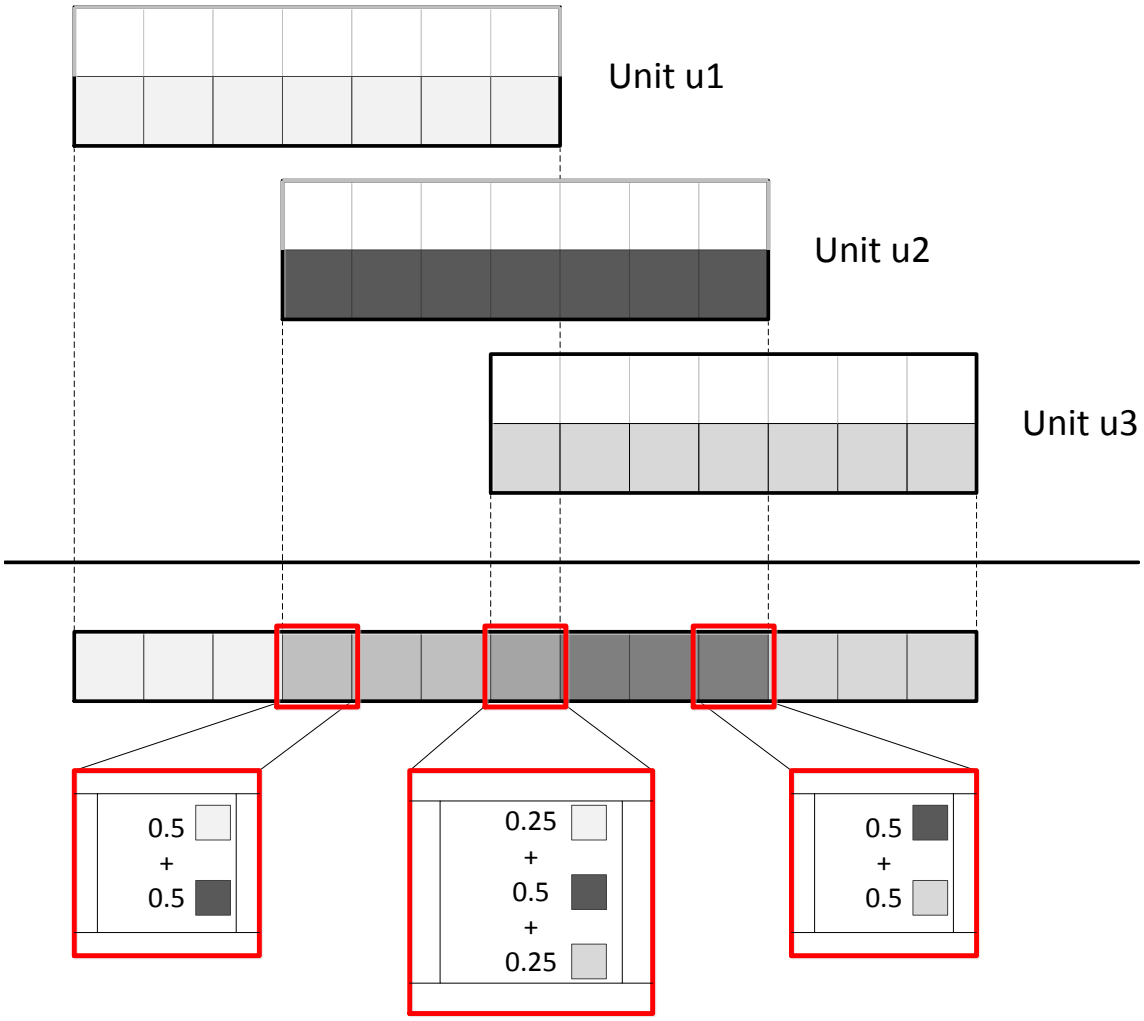

Abbildung 4.9: Gewichtete Aufaddierung überlappender Target-Frames der gewählten Unit-Sequenz zu einer geglätteten MFCC-Feature-Sequenz

liegen, wie auch in Grafik 4.9 dargestellt wird.
Für $\lambda = 0$ werden die Frames aller Units gleich gewichtet.

Nach der Berechnung der Gewichte werden die jeweiligen Frames der überlappenden Units u_i aufaddiert, wobei jede Dimension des Frames mit $w[i]$ multipliziert wird, um die beschriebene Glättung zu erzielen.

Aus den F0-Segmenten wurden nur die jeweils mittleren s Frames jedes Segments für die Synthese verwendet anstatt eine Addition durchzuführen, da eine Glättung hier nicht sinnvoll gewesen wäre. Stattdessen wurden mit einem Median-Filter Ausreißer entfernt.

4.5.2 Direkte Synthese aus akustischem Signal

Für die direkte Synthese aus dem unverarbeiteten akustischen Signal der Trainingsäußerungen werden die Start- und Endindizes benötigt, die in Abschnitt 3.2 genannt wurden. Aus ihnen können anhand des Frameshifts der Start- und Endzeitpunkt des Target-Segments innerhalb des Signals berechnet werden. Anhand dieser Zeitpunkte wird dann für jede Unit der entsprechende Abschnitt des Signals extrahiert. Die

einzelnen Abschnitte werden wiederum zu einem einzelnen akustischen Signal zusammengefügt.
Bei dieser Syntheseart werden die Target-Segmente der Codebook-Units nur zur Berechnung der Concatenation Cost, nicht aber zur Synthese verwendet.
Da sich die Units um o Frames überlappen, werden von jeder Unit nur die $w - o = s$ mittleren Frames für die Synthese gewählt und die Start- und Endzeitpunkte entsprechend berechnet. Zu Beginn und Ende der Äußerung werden die führenden bzw. endenden Frames der ersten und letzten Unit ebenfalls mit einbezogen, wie Grafik 4.10 zeigt.

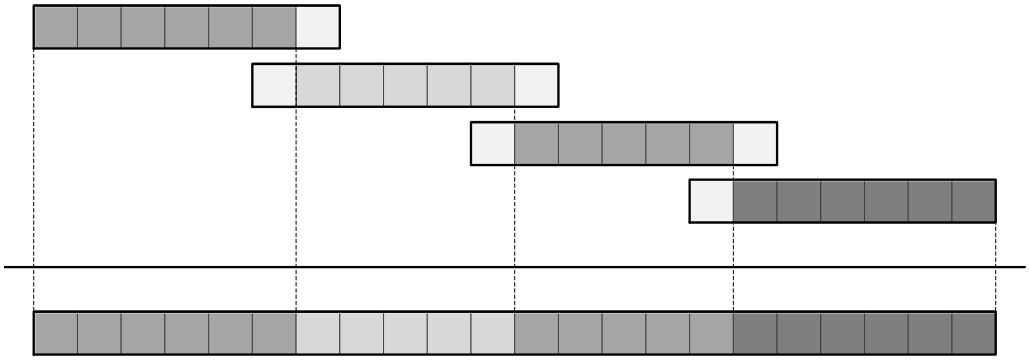

Abbildung 4.10: Verknüpfung von Abschnitten des akustischen Signals; die tatsächlich verwendeten Abschnitte sind dunkler markiert (mittlere s Frames der gewählten Units)

5. Evaluation

In diesem Kapitel sollen die Ergebnisse evaluiert werden. Zuerst werden verschiedene Parameter bei der Unit Selection beschrieben und deren Einfluss auf die Ausgabequalität erörtert, dann werden verschiedene Eigenschaften der Ausgabe bewertet, auch im Vergleich mit bereits vorhandenen Arbeiten zum Thema.

5.1 Parameter der Unit Selection

In diesem Abschnitt werden verschiedene Eigenschaften der Units und Parameter des Unit Selection-Algorithmus und deren Einfluss auf die Ausgabe evaluiert. Als Referenz für die Evaluation dient die parallel zu den Muskelsignalen aufgezeichnete Äußerung in hörbarer Sprache. Diese wird zur Konvertierung nicht verwendet und dient lediglich zum Vergleich.

5.1.1 Stacking-Höhe

Nachdem bei den Voruntersuchungen die TD-Features als beste Features für die Unit Selection ermittelt worden waren, wurden verschiedene Stacking-Höhen für diese Features evaluiert.
Wie in Grafik 5.1 zu sehen ist, führt das Stacking alleine schon zu einer deutlichen Verringerung der MCD.
Es kann keine global optimale Stacking-Höhe ermittelt werden, jedoch sind die TD15-Features immer unter den zwei besten MCDs, daher wurden sie für alle weiteren Experimente verwendet.

5.1.2 Segmentgröße

Eine Frage bei der Unit Selection ist die nach der Größe der verwendeten Segmente. Bei einer Segmentgröße, die ganze Silben oder Wörter umfasst, kann die konvertierte Sprache sehr natürlich klingen, da wenige Übergänge zwischen den Segmenten stattfinden. Dies würde jedoch auch eine sehr große Menge an Trainingsdaten erfordern, um die Silben oder Wörter in ausreichend vielen Variationen abbilden zu können.
In dieser Arbeit wurde die optimale Segmentgröße anhand der Mel Cepstral Distortion und der Erkennung der Stimmhaftigkeit ermittelt.

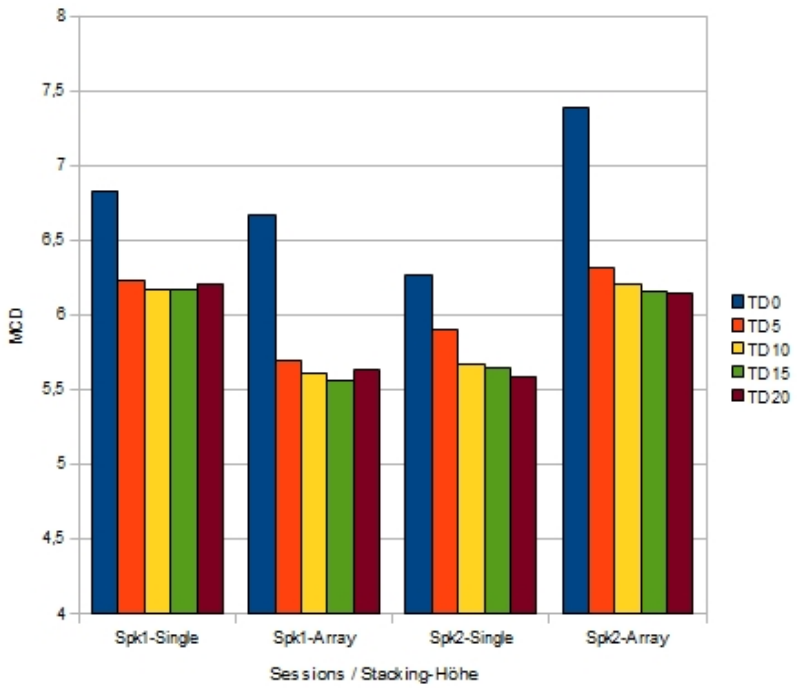

Abbildung 5.1: MCD-Werte für verschiedene Stacking-Höhen

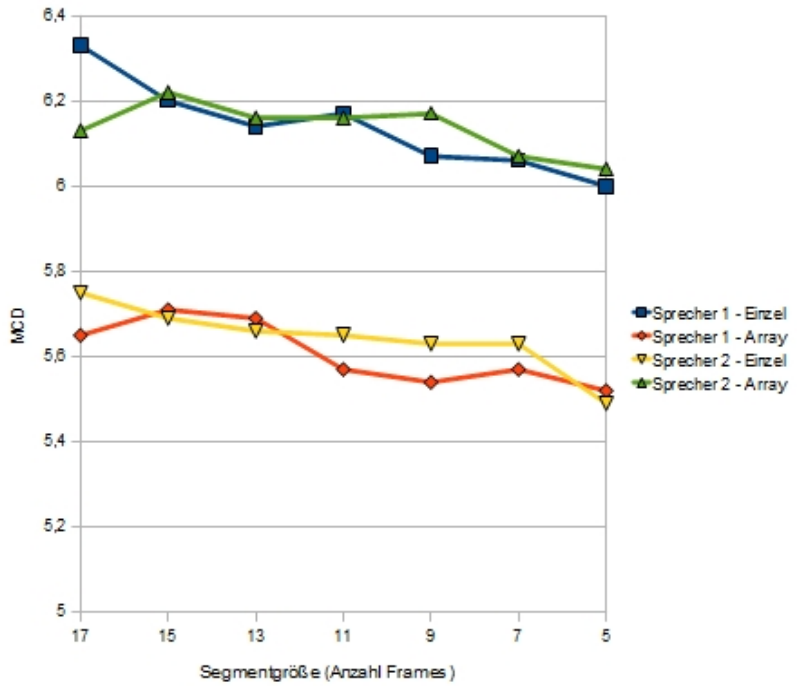

Abbildung 5.2: MCD-Werte für verschiedene Segmentgrößen (Verschiebung $s = 1$)

In Grafik 5.2 ist zu sehen, dass die Segmentgröße hinsichtlich der MCD möglichst klein sein sollte.

Bei der Erkennung der Stimmhaftigkeit hingegen ist in Grafik 5.3 kein eindeutiges Optimum zu erkennen, die Werte ergeben sich wohl eher aus zufälligen Schwankungen. Die Erkennungsrate stellt also kein gutes Kriterium zur Feststellung der optimalen Segmentgröße dar.

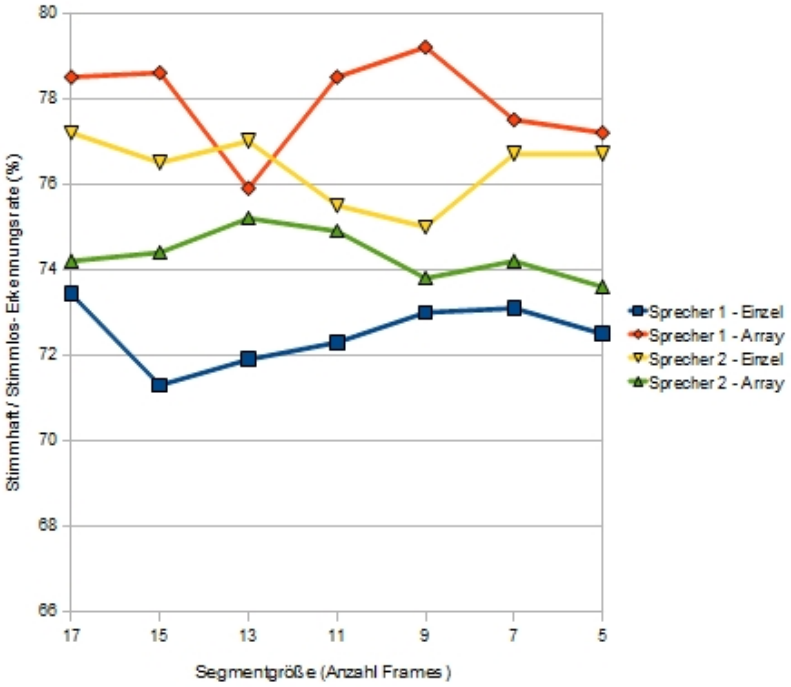

Abbildung 5.3: Erkennungsrate der Stimmhaftigkeit für verschiedene Segmentgrößen (Verschiebung $s = 1$)

Welche Segmentgröße letzten Endes optimal ist, hängt jedoch auch von weiteren Parametern ab, insbesondere von der Überlappung der Units, die im folgenden Abschnitt beschrieben wird.

Die Unterschiede in der MCD zwischen den Sessions ist auf die Qualität der Audiodaten zurückzuführen, wie in Abschnitt 4.3 zu sehen ist. Hier sind die relativen Unterschiede zwischen den Sessions ähnlich, auch wenn keine EMG-Daten verwendet wurden. Dies kann daran liegen, dass die Sprecher unterschiedlich konstant gesprochen haben, oder die Qualität der akustischen Aufzeichnungen durch andere äußere Faktoren beeinflusst wurden.

5.1.3 Überlappung der Units

Neben der Unit-Größe kann die Überlappung der gewählten Units entscheidend zur Verbesserung der Ausgabe beitragen. Hierfür wird die Test-Sequenz wie in Abschnitt 3.3 beschrieben in Segmente aufgeteilt, jedoch wird die Anzahl der überlappenden Frames zweier Segmente für diese Evaluation variiert.

Wie Grafik 5.4 beispielhaft für eine Session zeigt, führt eine geringere Verschiebung - also eine höhere Überlappung - zu einer Verringerung der MCD. Dies ist vermutlich darauf zurückzuführen, dass bei einer größeren Überlappung eine höhere Anzahl Codebook-Units für die Test-Sequenz ausgewählt werden (z.B. bei $w = 13$ und $s = 7$ doppelt so viele wie bei $w = 13$ und $s = 10$) und somit nicht optimale Units durch die anderen Units kompensiert werden.
Bei der MLSA-Synthese werden durch die gewichtete Aufaddierung der MFCCs an den überlappenden Stellen die Übergänge geglättet, was auch zur Kompensierung von suboptimalen Segmenten beiträgt.

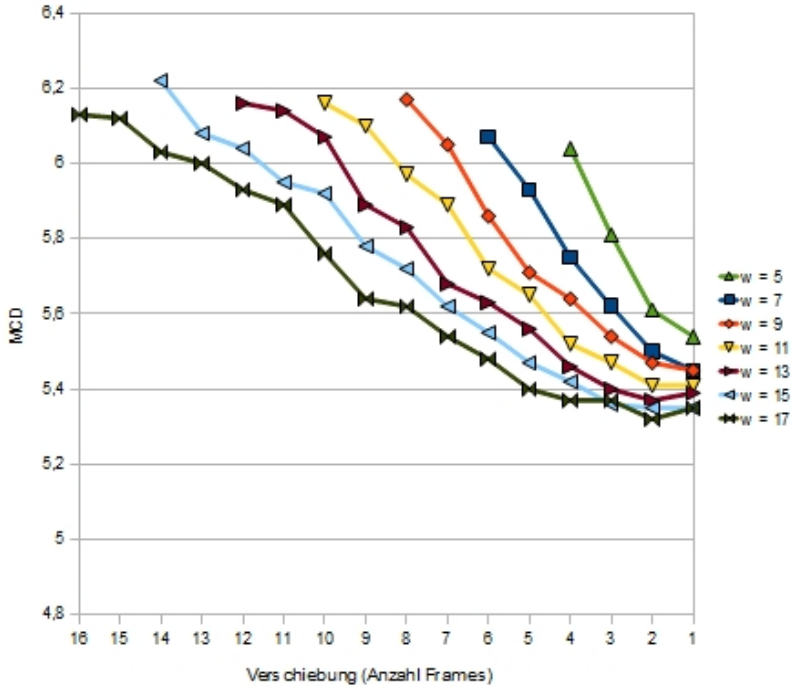

Abbildung 5.4: MCD-Werte für unterschiedliche Verschiebungen zwischen den Test-Segmenten (Session: Sprecher 2, Elektroden-Array)

Diese Werte zeigen auch, dass die optimale Segmentgröße aus dem vorigen Abschnitt bei einer höheren Überlappung nicht mehr optimal ist. Dieses Ergebnis weist ebenfalls darauf hin, dass die Wahl suboptimaler Units durch die Wahl einer größeren Menge von Units kompensiert werden kann, entweder durch kleinere Segmentgrößen oder durch größere Überlappung.

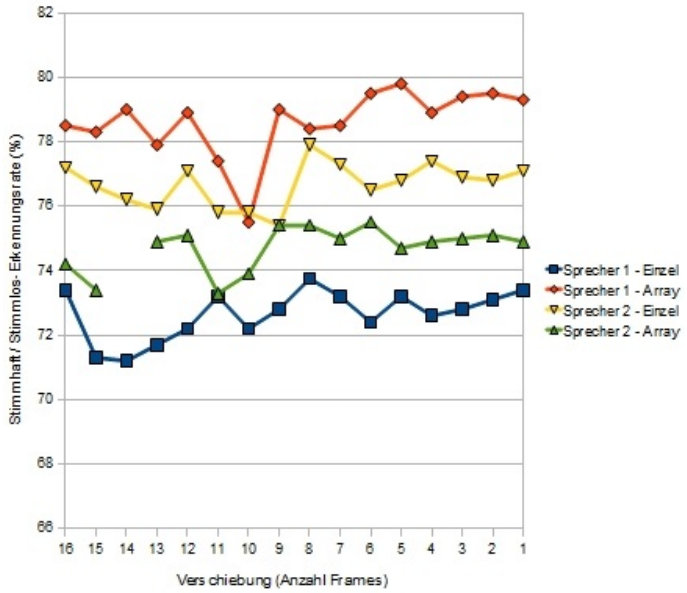

Abbildung 5.5: Erkennungsrate der Stimmhaftigkeit für verschiedene Verschiebungen (Segmentgröße $w = 17$). Die Berechnung des fehlenden Wertes von Sprecher 2 - Array war fehlerhaft und wurde daher ausgelassen

Die Auswertung der Erkennungsrate der Stimmhaftigkeit liefert, wie in Grafik 5.5 gezeigt, wieder kein eindeutiges Ergebnis und ist also kein Kriterium für die optimale Segmentgröße und Verschiebung.

Die geringste MCD in Abhängigkeit von Segmentgröße w und Verschiebung s lag für alle Sessions ungefähr bei $w = 15$ und $s = 2$. Die Länge eines nach diesem Kriterium optimalen Segments entspricht bei einem Frameshift von 10 ms und einer Framelänge von 27 ms also ca. 170 ms und die Verschiebung zwischen zwei Segmenten beträgt 20 ms.

5.1.4 Distanzmaße

Wie in Abschnitt 3.4.4 beschrieben, wurden verschiedene Distanzmaße für die Berechnung von Target Cost und Concatenation Cost verwendet. Bei den EMG2Aud-Experimenten war das Ergebnis unter Verwendung der Cosinus-Distanz konstant besser als mit der euklidischen Distanz, wie in Grafik 5.6 zu sehen ist. Sie wurde daher für alle weiteren Evaluationen als Distanzmaß verwendet.

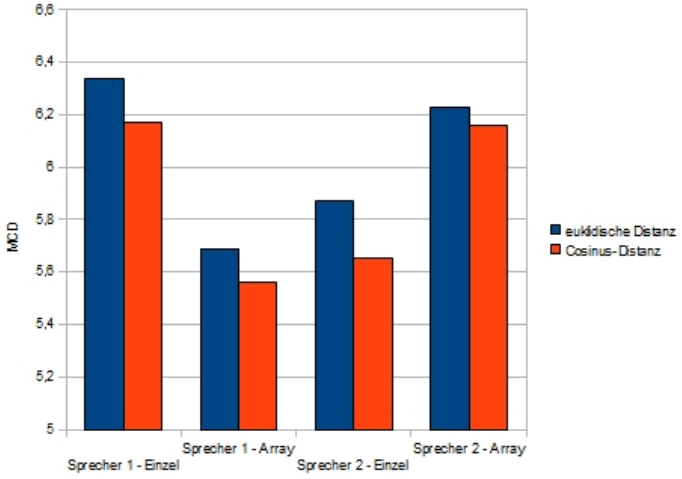

Abbildung 5.6: MCD-Werte für euklidische und Cosinus-Distanz (Segmentbreite $w = 11$, Verschiebung $s = 10$)

5.1.5 Kostenfunktionen

Die Kostenfunktionen für die Unit Selection wurden zunächst wie bei [Wu et al., 2013] verwendet. Es existieren jedoch Unterschiede zum dortigen Ansatz, wodurch eine erneute Evaluation der Kostenfunktionen nötig wurde:

1. Die Datenmenge, die bei dieser Arbeit verwendet wurde, ist eher gering. Möglicherweise ist dadurch bedingt die optimale Segmentgröße kleiner, als das bei einer großen Datenbasis der Fall wäre, wo genügend Variationen auch für größere EMG- oder Sprach-Abschnitte vorhanden wären. Hierdurch ergibt sich ein gesteigerter Einfluss der Concatenation Cost, da diese für jede Verbindung zwischen zwei Units berechnet wird und bei kleineren Units relativ mehr Verbindungen auftreten.

2. Die Test-Segmente bei [Wu et al., 2013] überlappen sich um $w - 1$ Frames, wohingegen die Überlappung bei dieser Arbeit zwischen 1 und $w - 1$ Frames betragen kann.

3. Target Cost und Concatenation Cost werden bei [Wu et al., 2013] jeweils auf den akustischen Features berechnet, die ungefähr im gleichen Wertebereich liegen. In dieser Arbeit haben Source- und Target-Features unterschiedliche Wertebereiche, was eine Gewichtung der Kostenfunktionen, wie in 3.4.3 beschrieben, erforderlich macht.

Die optimale Gewichtung der Concatenation Cost hängt von der Überlappung der Units ab und kann nicht eindeutig festgelegt werden, wie Grafik 5.7 zeigt.

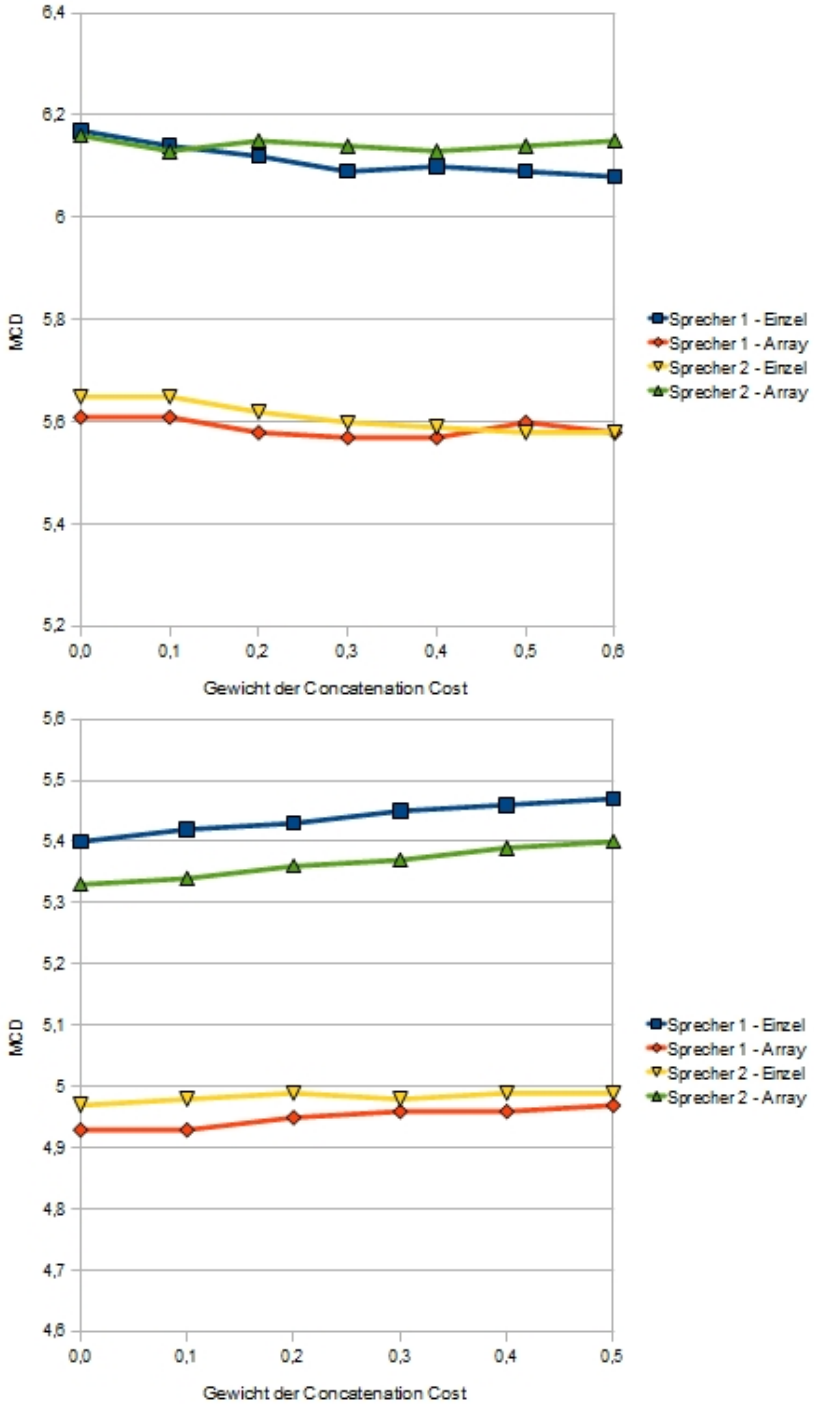

Abbildung 5.7: MCD-Werte für verschiedene Gewichtungen der Concatenation Cost, mit geringer Überlappung (oben) und höherer Überlappung (unten)

Dass die Concatenation Cost bei einer höheren Überlappung sogar einen nachteiligen Einfluss auf die Unit Selection hat, zeigt das untere Diagramm. Die besten Ergebnisse werden hier mit einer Gewichtung der Concatenation Cost von 0 erzielt. Auch bei der Überlappung von $o = 1$ führt die Verwendung der Concatenation Cost nur zu einer geringen Verbesserung. Dies ist vermutlich ebenfalls auf den kompensierenden Einfluss überlappender Units zurückzuführen, die einen natürlichen Übergang, auch durch das glättende Vorgehen bei der MLSA-Synthese, weniger wichtig machen.

5.1.6 Suche nach der optimalen Unit-Sequenz

Für die Suche nach der optimal Unit-Sequenz wurden zweierlei Suchalgorithmen verwendet. Die differenziertere *Viterbi*-Suche wurde durch die Beschränkung der maximal aktiven Suchpfade in ihrer Laufzeit verbessert, wodurch sich das Suchergebnis jedoch nicht wesentlich verschlechterte.

Die einfache *Greedy*-Suche hingegen ist ohnehin wesentlich schneller, da sie einer Viterbi-Suche mit nur einem aktiven Suchpfad entspricht. Die Ausgabequalität ist dennoch vergleichbar mit der Viterbi-Suche.
Da bei einer höheren Überlappung der Units ohnehin die Gewichtung der Concatenation Cost mit 0 optimal ist, wurde bei den meisten Experimenten eine noch einfachere Greedy-Suche verwendet, bei der als nächste Unit immer diejenige das mit der aktuell geringsten Target Cost gewählt wurde.

5.1.7 Gewichtsfunktion

In den meisten Experimenten wurde in Anlehnung an [Wu et al., 2013] $\lambda = 0.2$ als Faktor für die Gewichte gewählt, die wie in Abschnitt 4.5.1 beschrieben, berechnet werden.

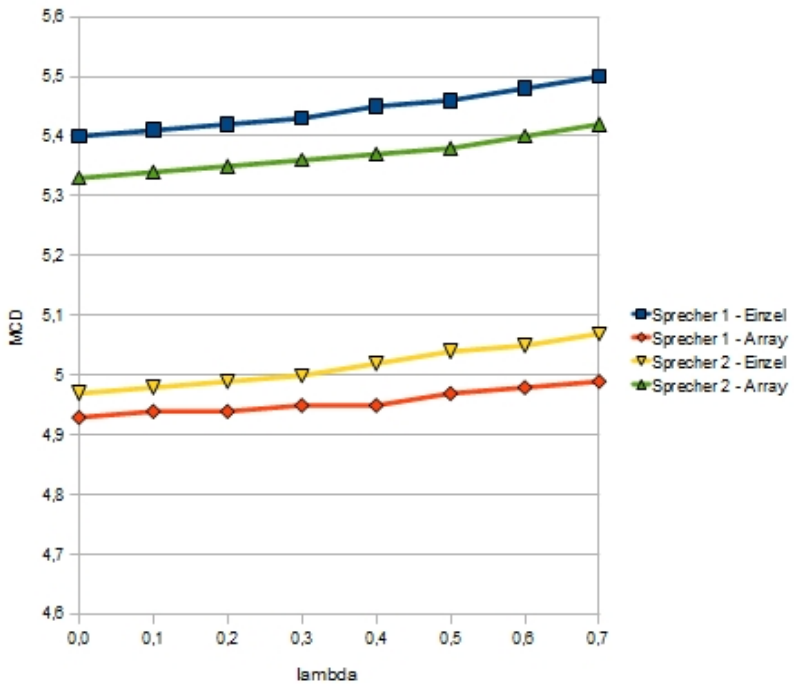

Abbildung 5.8: MCD-Werte für verschiedene Werte von λ

Bei der Evaluation des Gewichtsfaktors zeigte sich jedoch, dass die MCD niedriger ist, je kleiner λ gewählt wird; genauer gesagt ist die Gleichgewichtung aller beteiligten Frames, d.h. $\lambda = 0$, optimal, wie Grafik 5.8 zeigt.

5.2 Evaluationsmethoden

5.2.1 Test auf Trainingsdaten

Um zu überprüfen, ob der Unit Selection-Algorithmus die bestmöglichen Ergebnisse liefert, wurde ein Test auf Trainingsdaten durchgeführt, d.h. die Testäußerungen waren auch im Codebook enthalten. In diesem Fall sollte die Unit Selection die aus der Eingabe-Sequenz entstandenen Units direkt wieder ausgeben, da sowohl Target Cost als auch Concatenation Cost immer 0 sind.
Bei dieser Überprüfung war die MCD für alle Sessions $MCD = 0,0$ und die Erkennungsrate der Stimmhaftigkeit 100%, es wurde also tatsächlich immer die bestmögliche Sequenz gewählt.
Dieses Ergebnis bedeutet, dass eine nicht optimale Ausgabe nicht aus der Fehlerhaftigkeit des Algorithmus herrührt, sondern aus einer zu geringen Datenmenge im Codebook resultiert, oder aufgrund anderer äußerer Faktoren, wie z.B. Störsignale in den EMG-Features oder nicht-konstante Sprechweise des Sprechers.

5.2.2 Erkennung der Stimmhaftigkeit

Grafik 5.9 zeigt die erreichten Raten bei der Erkennung der Stimmhaftigkeit für die verschiedenen Sessions.
V/U gibt beispielsweise an, wie viele Frames zu stimmhaften (V für *"voiced"*) Frames konvertiert wurden, die in der Referenz aber stimmlos (U für *"unvoiced"*) waren.
Die Rate der korrekt erkannten Frames ergibt sich also aus V/V + U/U und beträgt im Durchschnitt über alle Sessions 77,4%.

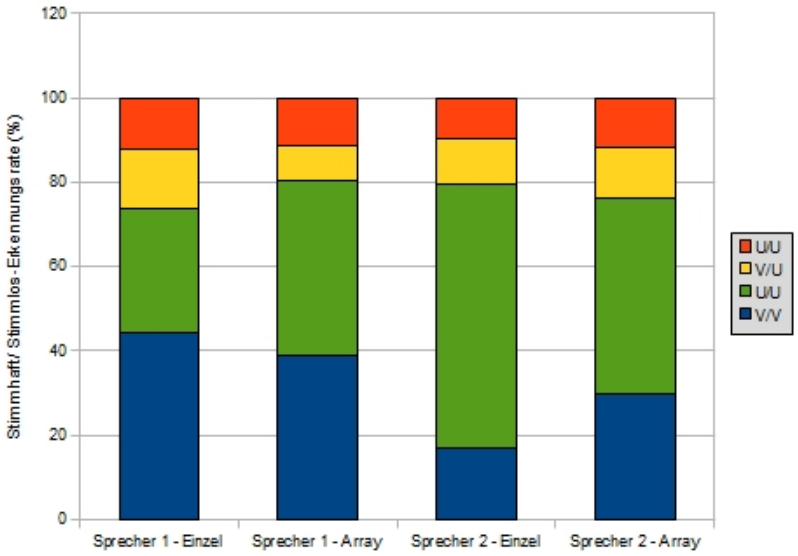

Abbildung 5.9: Erkennungsraten der Stimmhaftigkeit aller Sessions

5.2.3 Phonem-Labels

Um einen Überblick zu gewinnen, welche Phoneme durch die Unit Selection korrekt konvertiert wurden und welche mit anderen häufig verwechselt wurden, wurde eine Konfusionsmatrix für jede Session aufgestellt. Diese wurde mit Hilfe der Phonem-Labels in den Codebook-Units erstellt, indem aus der gewählten Sequenz analog zu Grafik 4.10 die Labels der mittleren Frames aneinandergereiht wurden.
In den Konfusionsmatrizen (Grafik 5.10) zeigt sich, dass die Unit Selection viele Phoneme korrekt konvertiert. Gängige Verwechslungen sind jedoch z.B. stimmhaftes s (z) und stimmloses s (s) oder t (t) und d (dh), welche anhand von EMG-Daten schwer zu unterscheiden sind.

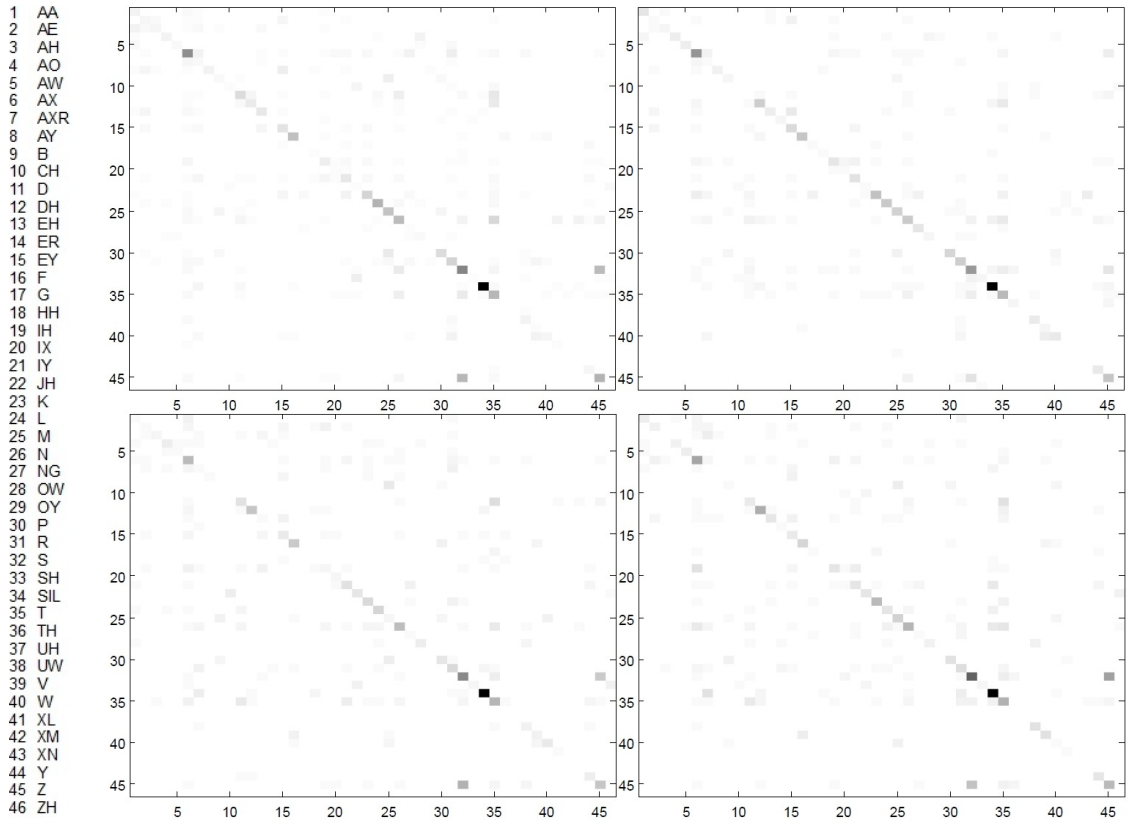

Abbildung 5.10: Konfusionsmatrizen der Phonem-Labels (Spalten: Referenz, Zeilen: Hypothese). Oben: Sprecher 1, unten: Sprecher 2. Links: Einzelelektroden-Sessions, rechts: Elektroden-Array-Sessions

5.2.4 Vergleich mit GMM-basiertem System

MCD-Werte

Tabelle 5.1 zeigt einen Überblick über die besten erreichten MCD-Werte der Unit Selection im Vergleich zum GMM-basierten System von [Janke et al., 2012].

Mit einem t-Test wurde nachgewiesen, dass die Verbesserung der Ausgabe durch die Unit Selection zum Signifikanzniveau $\alpha = 1\%$ statistisch signifikant ist.

Sprecher - Setup	GMM-Mapping [MCD]	Unit Selection [MCD]	Verbesserung
Sprecher 1 - Einzel	6,0	5,4	10,0%
Sprecher 1 - Array	5,9	4,9	16,9%
Sprecher 2 - Einzel	5,8	5,0	13,8%
Sprecher 2 - Array	6,3	5,3	15,9%
Durchschnitt	6,0	5,2	13,3%

Tabelle 5.1: Vergleich der besten erreichten MCD von GMM-Mapping und Unit Selection

Spektrogramme

Um den Unterschied zwischen den Verfahren zu veranschaulichen, zeigt Grafik 5.11 die Spektrogramme der Äußerung *"What follows is an old story more than fourty years old"*.

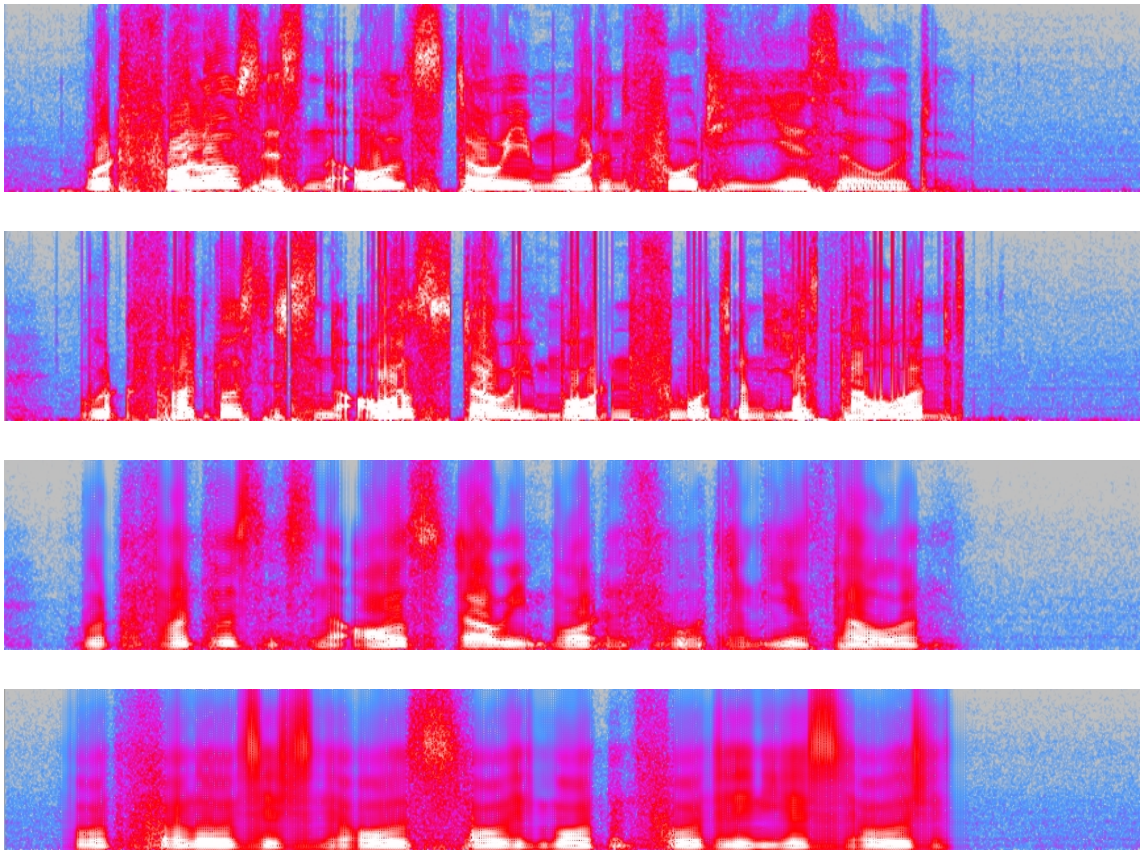

Abbildung 5.11: Spektrogramme der Äußerung *"What follows is an old story more than fourty years old"*: Original (oben), Unit Selection - Direktsynthese (2. v. o.), Unit Selection - MLSA-Synthese (3. v. o.), GMM-Mapping (unten)

Wie zu sehen ist, werden die Spektren durch den Unit Selection-Algorithmus mit der MLSA-Synthese und das GMM-Mapping verwischt, wobei die Unit Selection mehr Details und Variation der ursprünglichen Äußerung bewahrt.
Bei der Direktsynthese sind die einzelnen Audio-Abschnitte in Form von isolierten Spitzen als Bruchstücke erkennbar, analog dazu klingt auch die akustische Ausgabe

stückhaft. Die Frequenzbänder im Vergleich zum Original werden hingegen, von kurzen Unterbrechungen abgesehen, am besten wiedergegeben.

Verlauf von F0

In Grafik 5.12 sind die Verläufe der Grundfrequenz in drei Äußerungen zu sehen, die von dem Satz "The convicted murderer has avoided execution by lodging repeated appeals" stammen.

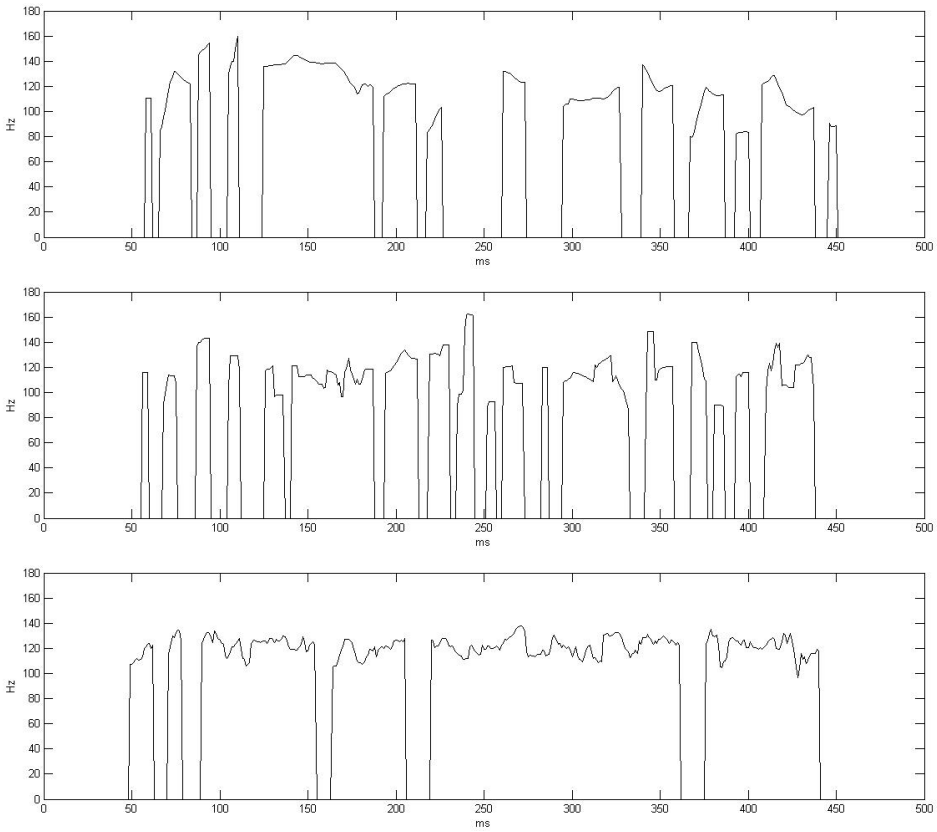

Abbildung 5.12: F0-Verläufe der Äußerung *"The convicted murderer has avoided execution by lodging repeated appeals"*: Original (oben), Unit Selection - MLSA-Synthese (Mitte), GMM-Mapping (unten)

Die oberste Kontur entspricht dem Verlauf in der Referenz-Äußerung.
Die Kontur in der Mitte stammt aus der Unit Selection und weist verglichen mit der Referenz mehr Übergänge zwischen stimmhaften und stimmlosen Abschnitten auf. Der grobe Verlauf wird durch das Verfahren jedoch gut angenähert.
Die letzte Grafik zeigt den F0-Verlauf beim GMM-Mapping. Hier werden viele Frames falsch konvertiert. Außerdem ist der Verlauf innerhalb der stimmhaften Abschnitte gleichzeitig monoton, da alle Werte zwischen 100 und 140 Hz liegen, und ungleichmäßig, wie der spezifische Verlauf der Linie zeigt.

5.2.5 Hörtest

Da die objektiven Maße nur schlecht dazu geeignet sind, die Natürlichkeit der konvertierten Äußerungen zu bemessen, wurde ein Hörtest mit 21 Testpersonen durchgeführt.

Anweisungen

Zu Beginn des Hörtests wurden den Testpersonen die Anweisungen präsentiert. Sie lauteten folgendermaßen:

- Please listen to the audio files using headphones!
- Please rate the naturalness of the given utterance, meaning the melody of speech, emphasis, fluency etc. which can range from very robotic and monotonous to fluent and accentuated.
- Understanding the meaning of the sentences is not essential.

Test-Setup

Als Framework für die Durchführung des Hörtests wurde das BeaqleJS-System von [Kraft und Zoelzer] verwendet, das ursprünglich für die Bewertung der Qualität von Audioübertragungen entwickelt wurde. Dieses Framework ermöglichte es, einen *Mean Opinion Score* (MOS) für die einzelnen Äußerungen zu berechnen und gleichzeitig die verschiedenen Ansätze miteinander zu vergleichen.

Grafik 5.13 zeigt die graphische Benutzeroberfläche des Hörtests für eine Äußerung.

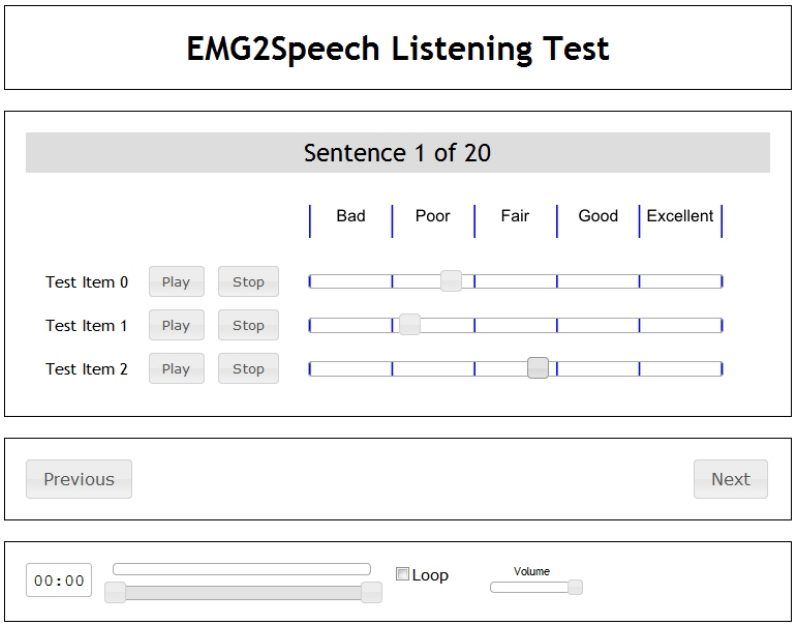

Abbildung 5.13: Screenshot des Hörtests

Die drei Test-Objekte, die zu sehen sind, wurden mit unterschiedlichen Methoden konvertiert:

- GMM-Mapping (s. Abschnitt 2.4)
- Unit Selection mit MLSA-Synthese (s. Abschnitt 4.5.1)
- Unit Selection mit Direktsynthese (s. Abschnitt 4.5.2)

Wie in der Grafik zu sehen ist, war die Konvertierungsmethode nicht angegeben. Zu den Äußerungen wurden außerdem keine Transkriptionen bereitgestellt, damit die Probanden nicht unbewusst die Verständlichkeit, sondern die Natürlichkeit bewerteten.

Die Probanden konnten die verschiedenen Varianten beliebig oft in beliebiger Reihenfolge abspielen und auch zu vorigen Äußerungen zurückkehren, um ihre Bewertung zu verändern.

Auswertung

Pro Testperson wurden 3 Varianten von 20 Äußerungen, also insgesamt 60 Sätze, bewertet, was insgesamt 420 Einzelbewertungen für jede Methode entspricht.
Das Bewertungsschema wurde im Backend als eine Skala von 0 (linkes Ende von *bad* in Grafik 5.13) bis 100 (rechtes Ende von *excellent*) umgesetzt.

Bei dem Hörtest zeigten sich insgesamt folgende Ergebnisse:

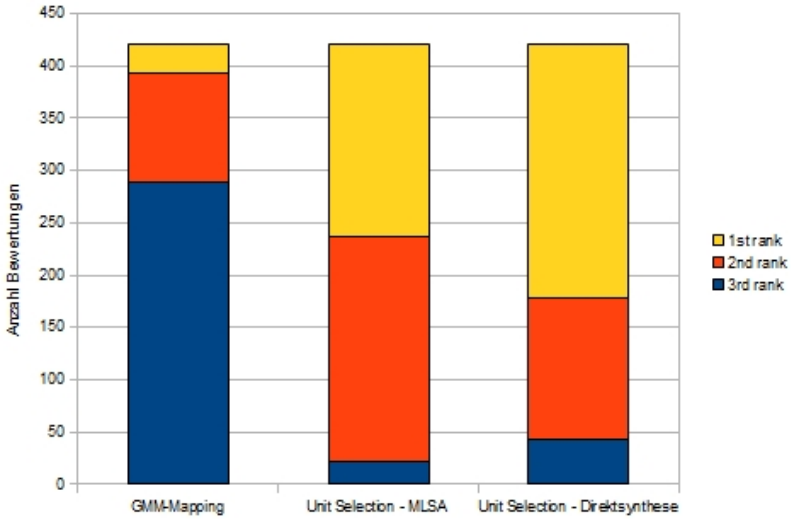

Abbildung 5.14: Ranking der einzelnen Systeme, erstellt anhand der Bewertungen aller Probanden

Grafik 5.14 zeigt einen Vergleich der drei Konvertierungs-Varianten. Es wurde bei jeder Äußerung aller Testpersonen verglichen, welches System für die aktuelle Äußerung die meisten, die mittleren, und am wenigsten Punkte zugewiesen bekommen hatte. Kumuliert über alle Testsätze und alle Testpersonen zeigt sich so, dass das GMM-Mapping in den meisten Fällen am schlechtesten bewertet wurde. MLSA- und Direktsynthese wurden ähnlicher bewertet, wobei die Direktsynthese im Vergleich öfter bevorzugt wurde.

In Grafik 5.15 ist erkennbar, dass die Testpersonen sehr unterschiedliche Bereiche die Bewertungsskala von 0 bis 100 Punkten für die Bewertung genutzt haben. Dementsprechend ist ein Vergleich der einzelnen Werte schwierig, jedoch kann man anhand der Mittelwertslinien erkennen, dass die beiden mit Unit Selection konvertierten Varianten einen MOS von 30 als Bewertung bekamen, wohingegen das GMM-basierte System mit ca. 12 Punkten bewertet wurde.

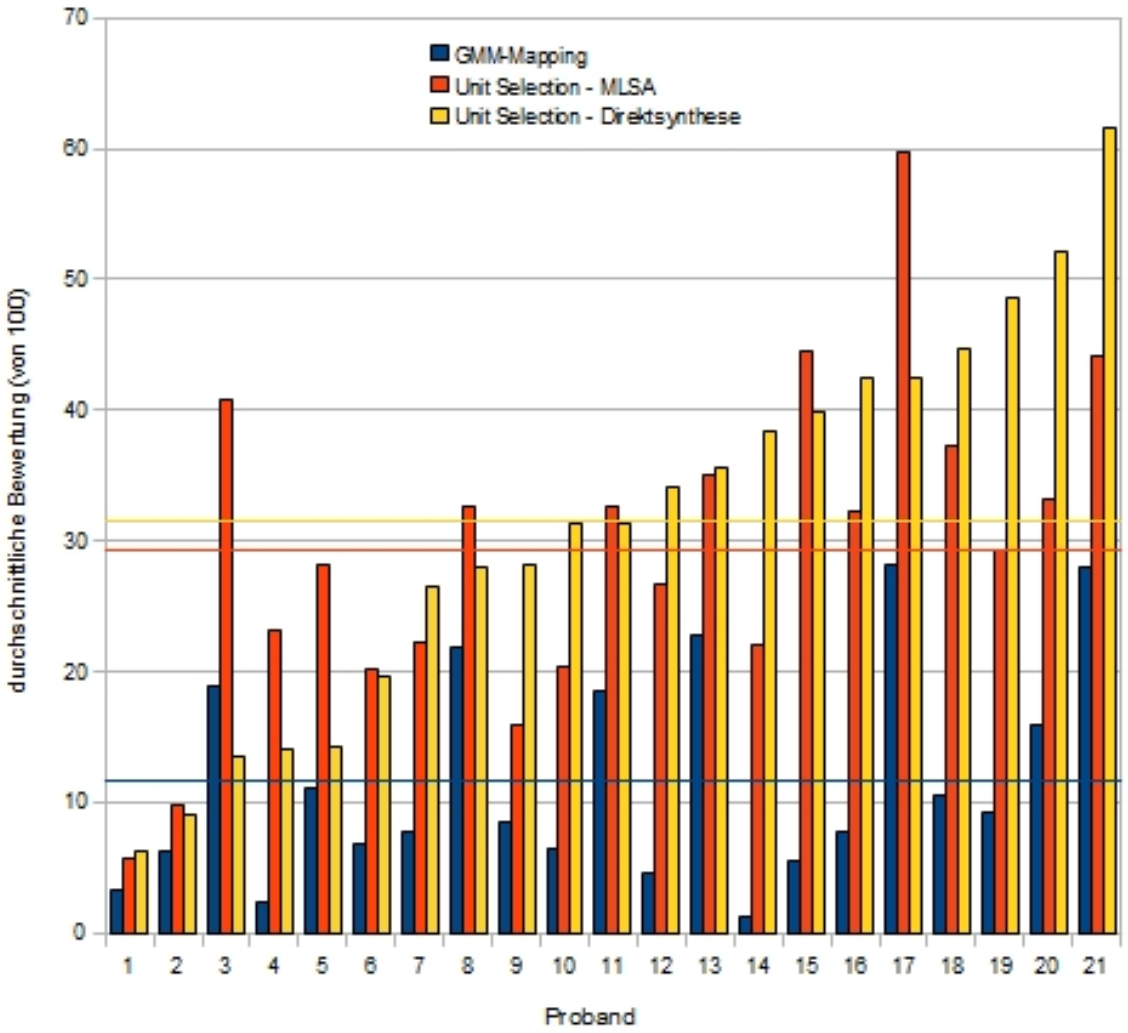

Abbildung 5.15: Durchschnittliche Bewertungen aller Probanden für die einzelnen Systeme

Ebenfalls zu erkennen ist die Bevorzugung der Direktsynthese gegenüber dem MLSA-Algorithmus, von wenigen Ausnahmen abgesehen. Dies ist vermutlich darauf zurückzuführen, dass bei der Direktsynthese unbearbeitete Abschnitte menschlicher Sprache aneinandergereiht werden, wohingegen die MLSA-Synthese gewisse Qualitätsverluste und Verzerrung mit sich bringt.

Da die Direktsynthese von den Testpersonen aber insgesamt bevorzugt wird, sollten die Übergänge zwischen den Units verbessert werden, da diese vermutlich die Ursache für die schlechtere Bewertung in manchen Fällen sind.

Insgesamt weniger gut bewertet wurden insbesondere Äußerungen, bei denen auch die MCD höher war als bei anderen. Es besteht also ein Zusammenhang zwischen der Verständlichkeit und der wahrgenommenen Natürlichkeit, was vermutlich darauf zurückzuführen ist, dass erkannte Wörter oder Satzteile als natürlicher empfunden werden als unverständliche Lautfolgen.
Diese Vermutung wird dadurch auch untermauert, dass bei dem Hörtest die Äußerungen mit der niedrigsten MCD, wo einzelne Wörter kaum auszumachen waren, auch am schlechtesten bewertet wurden.

6. Zusammenfassung und Ausblick

In dieser Arbeit wurde die Konvertierung myoelektrischer Signale der Gesichsmuskulatur zu hörbarer Sprache mittels eines Unit Selection-Verfahrens umgesetzt und evaluiert.
Hierbei zeigte sich bei einem Oracle-Experiment zur Konvertierung von hörbarer Sprache in hörbare Sprache (*Aud2Aud*-Experiment), dass eine geringe Datenmenge für die Unit Selection theoretisch ausreichend ist, jedoch wird bei der Konvertierung von EMG-Signalen zu Sprache oft eine weniger gute Ausgabe erzielt. Dies ist vermutlich darauf zurückzuführen, dass die EMG-Features für die Unit Selection weniger robust sind.
Dennoch wurde die Verständlichkeit der Sprachausgabe gegenüber einem GMM-basierten Ansatz, gemessen anhand der Mel Cepstral Distortion, um 13% verbessert. Die Natürlichkeit der Ausgabe wurde durch einen Hörtest überprüft und zeigt ebenfalls deutliche Verbesserungen gegenüber dem GMM-basierten System. Hierbei wurde die Synthese der konvertierten Sprache direkt aus dem Audiosignal gegenüber der Synthese mittels des MLSA-Algorithmus von den Testpersonen vorgezogen.
Die Qualität der Ausgabe lässt sich auch an der Ähnlichkeit der Spektrogramme von konvertierter Sprache und Referenz und deren F0-Konturen sowie Konfusionsmatrizen erkennen.

Obwohl mit dem Unit Selection-Ansatz schon gute Ergebnisse erzielt wurden, bestehen noch zahlreiche Möglichkeiten zur Weiterentwicklung des Verfahrens.

Da von den Testpersonen die Direktsynthese der Sprachausgabe aus dem Audiosignal gegenüber der MLSA-Synthese bevorzugt wurde, sollte dieses Syntheseverfahren gezielt verbessert werden. Das bestehende Problem sind die harten Übergänge zwischen den Units, welche durch Glättung verbessert werden könnten. Eventuell kann auch eine größere Datenmenge im Codebook zur Verbesserung beitragen. Weiterhin ist es denkbar, die Units nicht in eine feste Anzahl Frames einzuteilen, sondern anhand linguistischer Informationen, z.B. in Silben, Wörter oder Phoneme. Dadurch werden die Übergänge einfacher gestaltet, da sie nicht mehr willkürlich beginnen und enden. Um mehr Daten für die Unit Selection zur Verfügung zu haben, ist es auch möglich die Aufnahmen mehrere Sessions zu kombinieren, gegebenenfalls auch von verschiedenen Sprechern. Hierbei muss allerdings die Herausforderung der Neuplatzierung

der Elektroden beachtet werden, die zu veränderten EMG-Signalen führt, ebenso wie die unterschiedliche Leitfähigkeit der Haut und andere Einflussfaktoren.

Ebenfalls untersucht werden muss die Tauglichkeit des Unit Selection-Verfahrens für lautlose Sprache. In dieser Arbeit wurden die Sprachaufnahmen der Test-Äußerungen zwar nicht für die Konvertierung verwendet, jedoch ist bekannt, dass sich die Sprechweise bei hörbarer und lautloser Sprache unterscheidet. Dieser Unterschied muss gegebenenfalls kompensiert werden.

Zuletzt ist für die praktische Anwendung des Verfahrens die Echtzeitfähigkeit essentiell. Eine unmittelbare hörbare Ausgabe seiner stummen Sprechweise liefert dem Benutzer die Rückmeldung, die er benötigt um den Umgang mit dem Silent Speech Interface zu üben. Zeitliche Verzögerung erschwert das Erkennen von problematischen Bewegungen und sollte daher vermieden werden.

Literaturverzeichnis

[Ber, 2001] (2001). *Bertelsmann Lexikon*. Bertelsmann Lexikon Verlag, Gütersloh.

[Catford, 1977] Catford, J.C. (1977). *Fundamental problems in phonetics*. Indiana University Press.

[Chan et al., 2001] Chan, A.D.C., K. Englehart, B. Hudgins und D. Lovely (2001). *Myo-electric signals to augment speech recognition*. Medical and Biological Engineering and Computing, 39(4):500–504.

[Denby et al., 2010] Denby, B., T. Schultz, K. Honda, T. Hueber, J. M. Gilbert und J. S. Brumberg (2010). *Silent speech interfaces*. Speech Communication, 52(4):270–287.

[Fagan et al., 2008] Fagan, M.J., S. Ell, J. Gilbert, E. Sarrazin und P. Chapman (2008). *Development of a (silent) speech recognition system for patients following laryngectomy*. Medical Engineering & Physics, 30(4):419 – 425.

[Hunt und Black, 1996] Hunt, A.J. und A. Black (1996). *Unit selection in a concatenative speech synthesis system using a large speech database*. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Conference Proceedings.*, Bd. 1, S. 373–376.

[Imai, 1983] Imai, S. (1983). *Cepstral analysis synthesis on the mel frequency scale*. In: *IEEE International Conferenceon Acoustics, Speech, and Signal Processing (ICASSP). Conference Proceedings.*, Bd. 8, S. 93–96.

[Janke et al., 2014] Janke, M., M. Wand, T. Heistermann, T. Schultz und K. Prahallad (2014). *Fundamental Frequency Generation for Whisper-to-Audible Speech Conversion*. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Conference Proceedings.*.

[Janke et al., 2012] Janke, M., M. Wand, K. Nakamura und T. Schultz (2012). *Further investigations on EMG-to-speech conversion*. In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, S. 365–368. IEEE.

[Jorgensen et al., 2003] Jorgensen, C., D. Lee und S. Agabont (2003). *Sub auditory speech recognition based on EMG signals*. In: *Proceedings of the International Joint Conference on Neural Networks.*, Bd. 4, S. 3128–3133. IEEE.

[Jou et al., 2006] Jou, S.-C. S., T. Schultz, M. Walliczek, F. Kraft und A. Waibel (2006). *Towards continuous speech recognition using surface electromyography.*. In: *INTERSPEECH*.

[Kraft und Zoelzer] Kraft, S. und U. Zoelzer. *BeaqleJS: HTML5 and JavaScript based Framework for the Subjective Evaluation of Audio Quality*. Helmut-Schmidt-University Hamburg.

[Kubichek, 1993] Kubichek, R.F. (1993). *Mel-cepstral distance measure for objective speech quality assessment*. In: *IEEE PacificRim Conference on Communications, Computers and Signal Processing*, Bd. 1, S. 125–128 vol.1.

[Kumar und Mital, 1996] Kumar, S. und A. Mital (1996). *Electromyography in Ergonomics*. CRC Press.

[Maier-Hein et al., 2005] Maier-Hein, L., F. Metze, T. Schultz und A. Waibel (2005). *Session independent non-audible speech recognition using surface electromyography*. In: *IEEE Workshop on Automatic Speech Recognition and Understanding*, S. 331–336.

[Minifie et al., 1973] Minifie, F.D., T. Hixon, F. Williams und D. Broad (1973). *Normal aspects of speech, hearing, and language*. Prentice-Hall.

[Morse und O'Brien, 1986] Morse, M. S. und E. M. O'Brien (1986). *Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes*. Computers in Biology and Medicine, 16(6):399 – 410.

[Nakajima et al., 2006] Nakajima, Y., H. Kashioka, N. Campbell und K. Shikano (2006). *Non-audible murmur (NAM) recognition*. IEICE TRANSACTIONS on Information and Systems, 89(1):1–8.

[Nakamura et al., 2011] Nakamura, K., M. Janke, M. Wand und T. Schultz (2011). *Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Conference Proceedings.*, S. 573–576.

[Schultz und Wand, 2010] Schultz, T. und M. Wand (2010). *Modeling coarticulation in EMG-based continuous speech recognition*. Speech Communication, 52(4):341–353.

[Solbach, 2000] Solbach, H. (2000). *Vita nova: Biologie für die Sekundarstufe II*. Buchner.

[Sugie und Tsunoda, 1985] Sugie, N. und K. Tsunoda (1985). *A Speech Prosthesis Employing a Speech Synthesizer-Vowel Discrimination from Perioral Muscle Activities and Vowel Production*. IEEE Transactions on Biomedical Engineering, BME-32(7):485–490.

[Toth et al., 2009] Toth, A. R., M. Wand und T. Schultz (2009). *Synthesizing speech from electromyography using voice transformation techniques.*. In: *INTERSPEECH*, S. 652–655.

[Walliczek et al., 2006] Walliczek, M., F. Kraft, S.-C. S. Jou, T. Schultz und A. Waibel (2006). *Sub-word unit based non-audible speech recognition using surface electromyography.*. In: *INTERSPEECH*, S. 1487–1490.

[Wand et al., 2013] Wand, M., C. Schulte, M. Janke und T. Schultz (2013). *Array-based Electromyographic Silent Speech Interface.*. In: *BIOSIGNALS*, S. 89–96.

[Winter, 1990] Winter, D. A. (1990). *Biomechanics and Motor Control of Human Movement*. A Wiley-Interscience publication. Wiley, New York [u.a.], 2. Aufl.

[Wu et al., 2013] Wu, Z., T. Virtanen, T. Kinnunen, E. Chng und H. Li (2013). *Exemplar-based unit selection for voice conversion utilizing temporal information.*. In: *INTERSPEECH*, S. 3057–3061.