Speech Recognition Using Surface Electromyography

Lena Maier-Hein

Diplomarbeit
Juli 2005

Universität Karlsruhe
76131 Karlsruhe
Deutschland

Supervisors:
Dipl. Phys. Florian Metze
Dr. Tanja Schultz
Prof. Dr. Alex Waibel

Declarations

Hiermit erkläre ich, dass ich diese Diplomarbeit selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Die verwendeten Literaturquellen sind vollständig im Literaturverzeichnis aufgeführt.

Karlsruhe, den 31.7.2005

Lena Maier-Hein

Permission is herewith granted to the Universität Karlsruhe (TH), Germany, and the Carnegie Mellon University, PA, USA, to circulate copies from this thesis for non-commercial purposes.

Karlsruhe, den 31.7.2005

Lena Maier-Hein

Abstract

The applicabilty of conventional automatic speech recognition in everyday life is still limited. It is neither suitable in situations requiring silence (e.g. in a meeting) nor in noisy environments. For this reason, we introduce a speech recognition system based on myoelectric signals. The system handles audible and non-audible speech and outperforms previously developed electromyography (EMG) based speech recognition systems on the ten English digits vocabulary. It deploys seven surface electrodes placed in various positions on the face and the neck and uses Hidden Markov Models (HMMs) as classifiers.

Major challenges in surface electromyography based speech recognition ensue from repositioning electrodes between recording sessions and from varying qualities of the electrode/skin interface. In order to reduce the impact of these factors, we investigate a variety of signal normalization and model adaptation methods. An average word accuracy of 97.3% on the ten English digits vocabulary is achieved using the exact same electrode positions for training and recognition (*within-session testing*). The performance drops to 76.2% when the electrodes are removed after training and re-applied for testing (*across-sessions testing*). By applying our adaptation and normalization methods we manage to raise across-sessions recognition rates to 87.1%.

Furthermore, we compare audibly to non-audibly spoken speech. The results suggest that large differences exist between the corresponding muscle movements, yet, it is possible to merge training data to obtain a recognizer that deals accurately with both speech manners.

As a first step towards continuous speech recognition we further investigate connected digits recognition. The results indicate that segmentation and context dependency are major issues in EMG based continuous speech recognition. Furthermore, initial experiments on phoneme based approaches suggest that classical phoneme models are not the appropriate choice for the recognition of non-audible speech.

In order to demonstrate the potential of speech recognition based on myoelectric signals we introduce two online recognition systems showing different applications of the new technology: a prototype “silent” mobile phone suitable for conducting non-disturbing private conversations in situations requiring silence and a prototype lecture translater designed to translate a non-audibly spoken talk in a given language into a language of one’s choice.

Zusammenfassung

Automatische Spracherkennung hat inzwischen breite Anwendung in unserem Alltagsleben gefunden. Auskunftssysteme, Mobiltelefone und Diktiersysteme bedienen sich beispielsweise längst der neuen Technologie. Dennoch ist die Anwendbarkeit konventioneller Spracherkennung auf bestimmte Situationen beschränkt. Das akustische Signal verhindert nicht nur eine vertrauliche Kommunikation durch ein auf automatischer Spracherkennung basierendes elektronisches Gerät, sondern kann zudem auch sehr störend sein (z.B. im Meeting, in Bus und Bahn, in der Bibliothek). Hinzu kommt, dass konventionelle Spracherkennung in geräuschbehafteten Umgebungen sowie in veränderten atmosphärischen Bedingungen wie unter Wasser oder im All nur mäßig gut funktioniert. Sprachbehinderte können derartige Systeme ebenfalls nicht nutzen.

Um diese Einschränkungen zu überwinden, wurden alternative Methoden entwickelt, welche nicht vom akustischen Signal abhängen. Nachdem Morse et. al 1986 den Beweis erbrachten, dass die myoelektrischen Signale gewisser Hals- und Gesichtsmuskeln Sprachinformation enthalten [Morse M., 1986], nutzten verschiedenste Forschergruppen in den vergangen fünfzehn Jahren elektromyographische Signale für isolierte Worterkennung. Jorgensen et al. zeigten zudem, dass eine Klassifikation auch bei nicht hörbarer Sprache möglich ist, das heißt, wenn kein akustisches Signal erzeugt wird, aber Lippen und Zunge wie bei normaler Spracherzeugung bewegt werden [Jorgensen et al., 2003].

Die Anwendbarkeit Elektromyographie (EMG) basierter Spracherkenner ist zur Zeit allerdings noch aufgrund folgender Probleme beschränkt: Erstens benötigen Oberflächenelektroden, die zum Messen der Muskelaktivität notwending sind, physikalischen Kontakt mit der Haut, was die Benutzerfreundlichkeit vorhandener Systeme verringert. Zweitens wurden bislang nur isoliert gesprochene einzelne Wörter mit zufrieden stellenden Ergebnissen erkannt. Drittens sind heutige Systeme alles andere als robust, da sie auf identische Trainings- und Testkonditionen angewiesen sind. Myoelektrische Signale hängen nämlich nicht nur vom jeweiligen Sprecher mit seinem individuellem Sprechstil ab, sondern zusätzlich von der genauen Elektrodenpositionierung sowie Temperatur, Hautleitfähigkeit und anderen Konditionen die sich von Aufnahmesession zu Aufnahmesession ändern können. Letzteres Phänomen bezeichnen wir als *Sessionabhängigkeit* in Analogie zur *Kanalabhängigkeit*, die sich in konventioneller Spracherkennung durch veränderte Mikrofonqualität und Umgebungsgeräusche ergibt.

Um die neue Technologie voranzutreiben, entwickelten wir einen EMG basierten Einzelworterkenner, der dem Stand der Technik entspricht. Anschließend beschäftigten wir uns mit Themen, die bislang nicht in der Fachliteratur behandelt wurden, und zwar mit Sessionabhängigkeit, mit dem Vergleich von hörbarer und nicht hörbarer gesprochener Sprache sowie mit initialen Experimenten zu kontinuierlicher Spracherkennung. Zuletzt implementierten wir zwei Echtzeit-Systeme, um das Potential EMG basierter Spracherkennung zu demonstrieren.

Unser Baseline-System benutzt sieben Oberflächenelektroden und Hidden-Markov-Model Einzelwort-Klassifikatoren zur Einzelworterkennung auf einem festen Vokabular. Vorverarbeitung, HMM Topologien und Elektrodenpositionierung wurden in verschiedenen Experimenten optimiert. Die Worterkennungsrate auf dem Zehn-Ziffern-Vokabular beträgt 97.3%.
Unserer Erfahrung nach sind die durch Sessionabhängigkeit verursachten Leistungseinbußen in EMG basierten Spracherkennungssystemen höher als die durch Kanalabhängigkeit entstehende Verringerung der Wortakkuratheit in konventionellen Systemen. Bislang wurden jedoch laut Literatur nur sessionabhängige Systeme entwickelt. Aus diesem Grund untersuchten wir verschiedene Normalisierungs- und Adaptionsmethoden zum Anpassen der Signale einer neuen Aufnahmekonfiguration an gegebenes Trainingsmaterial. Unsere Ergebnisse zeigten, dass Methoden, die in konventioneller Spracherkennung angewandt werden, um Sprecherunabhängigkeit zu erreichen, in EMG basierter Spracherkennung gegen Sessionabhängigkeit eingesetzt werden können. Varianznormalisierung, Feature Space Adaption sowie das Trainieren auf mehreren Sessions verbesserten sessionübergreifende Erkennung. Wir erhielten eine Erkennung von 97.3% bei identischen Trainings- und Testbedingungen. Sessionübergreifendes Testen ohne Normalisierungen und Adaptionen verursachte einen Erkennungsabfall um 20.9% auf 76.2%. Durch Anwendung unserer Adaptionsmethoden konnten wir diese Erkennungsrate um 14.3% auf 87.1% steigern.
Einer der größten Vorteile EMG basierter Spracherkennung ist die Tatsache, dass keine Erzeugung eines akustischen Signals erforderlich ist. Da unseres Wissens nach bislang keine Studie über die für die Spracherkennung relevanten Unterschiede von hörbarer und nicht hörbarer Sprache durchgeführt wurde, beschäftigten wir uns mit dem Vergleich der beiden Sprachmodi. Unsere Experimente zeigten signifikante Unterschiede zwischen den korrespondierenden Muskelbewegungen. Dennoch war es möglich, *einen* Erkenner zu trainieren, der für *beide* Sprechmodi zufrieden stellende Ergebnisse liefert. Zudem erhielten wir bei mit dem System nicht vertrauten Sprechern signifikant bessere Erkennungsergebnisse für hörbar gesprochene Sprache. Allerdings konnte im Laufe der Zeit ein Lerneffekt festgestellt werden.
Da Einzelworterkennung nur bedingte Anwendbarkeit hat, beschäftigten wir uns ausserdem mit initialen Experimenten zur kontinuierlichen Spracherkennung. Untersuchungen zur Ziffernfolgenerkennung ergaben, dass Segmentierung und Kontextabhängigkeit die Hauptprobleme bei der Umstellung von isolierter Worterkennung auf die Erkennung von Wortfolgen sind. Unseren Ergebnissen zufolge sind klassische Phonem-Modelle zudem nicht die optimalen Spracheinheiten bei der Erkennung nicht hörbarer Sprache mit Hilfe von Elektromyography.
Unsere beiden Demo-Systeme illustrieren zwei verschiedene potentielle Anwendungen, die ausschließlich für *EMG* basierte Spracherkenner geeignet sind: ein "stilles Mobiltelefon"' zum Führen privater, nicht störender Telefongespräche sowie einen automatischen "Übersetzer" zum Übersetzen einer nicht hörbar gesprochenen Rede in eine Sprache nach Wahl. Beide Demo-Systeme werden jeweils auf einer Menge von für die zugehörige Anwendung geeigneten *Sätzen* trainiert. Die Ausgabe erfolgt sowohl auf dem Bildschirm als auch über eine Sprachsynthese.
Als zukünftige Forschungsschwerpunkte schlagen wir die Umstellung auf kontinuierliche Sprache sowie die Entwicklung benutzerfreundlicherer Sensoren vor.

Acknowledgements

I would like to thank my supervisors Dr. Tanja Schultz and Florian Metze for their constant guidance and support during the research conducted for this thesis. Both of them were a valuable help. Furthermore, I want to thank Prof. Alex Waibel for encouraging me to run the risk of approaching a relatively new research field. Many thanks also go to Christoph Mayer, Marcus Warga, Peter Osztotics and Artus Krohn-Grimberghe who helped me with the data recording as well as with the design and implementation of parts of the software. I would also like to thank Christian Fügen and Sebastian Stüker for their introduction to the Janus Recognition Toolkit at the beginning of this project. Finally, I want to express my gratitude to my family, my boyfriend and my friends for cheering me up when the project did not progress as quickly as I had planned.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Automatic Speech Recognition (ASR) has developed into a popular technology and is being deployed in a wide variety of every day life applications, including personal dictation systems, call centers and mobile phones. Despite the various benefits a conventional speech-driven interface provides to humans, there are three major drawbacks: Firstly, the audible (i.e. acoustic) speech signal prohibits a confidential conversation with or through a device. Besides that, talking can be extremely disturbing to others, especially in libraries or during meetings. Secondly, speech recognition performance degrades drastically in adverse environmental conditions such as in restaurants, cars, or trains. Acoustic model adaptation can compensate for these effects to some degree, however the pervasive nature of mobile phones challenges this approach. Performance is also poor when sound production limitations occur, like under water. Last but not least, conventional speech-driven interfaces cannot be used by speech handicapped people, for example those without vocal cords.

To overcome these limitations, alternative methods are being investigated, which do not rely on an acoustic signal for ASR. Chan et al. [Chan et al., 2002b] proved that the myoelectric signal (MES) from articulatory face muscles contains sufficient information to discriminate a given set of words accurately (>90% word accuracy on the ten English digits). This holds even when the words are spoken non-audibly, i.e. when no acoustic signal is produced [Jorgensen et al., 2003].

The potential of electromyography (EMG) based speech recognition lies primarily in the fact that it does not rely on the transmission of an acoustic signal: it allows private, non-disturbing communication in any situation and could possibly be deployed by speech handicapped people. Moreover, it is robust to environmental noise.

To date, however, the practicability of MES based speech recognition is still limited. Firstly, the surface electrodes require physical contact with the speaker's skin. Secondly, experiments are still restricted to isolated word recognition. Finally, today's systems are far from being robust, since they only work in matching training and test conditions. Just like conventional speech recognizers, the MES based systems are heavily influenced by speaker dependencies, such as speaking style, speaking rate, and pronunciation ideosyncrasies. Beyond that, the myoelectric signal is affected by even slight changes in electrode positions, temperature or tissue properties [Lamb and Hobart, 1992]. We will refer to this phenomenon as *session dependence* in analogy to the *channel dependence* of a conventional speech recognizer resulting from microphone quality, environmental noise, and signal transmission of the acoustic signal.

Thus, significant challenges remain. In the next section we introduce the goals we have set for this work in order to contribute to the development of the new technology.

1.2 Goals of this work

The goals of this work were to (1) build a state-of-the-art speech recognition system based on myoelectric signals, (2) to address major issues in the novel technology that have not yet been addressed in literature and (3) to demonstrate the practicability of EMG based speech recognition.

State-of-the-Art-System

State-of-the-art EMG based speech recognizers perform *session dependent, isolated word* recognition (section 3). One goal of this work was to explore appropriate feature extraction and classification methods in order to develop an EMG based discrete speech recognizer that achieves recognition results comparable to those that have so far been reported in literature.

Novel Issues

According to our experience the loss in performance caused by session dependence in MES based speech recognition is significantly higher than that resulting from channel conditions in conventional systems. Despite this, only session dependent MES based speech recognition systems have been developed so far. A second goal of this work was therefore to address the issue of session dependence by exploring methods for adjusting data from a new recording session to given training material from previous recording sessions.

The most important advantage of using the MES for speech recognition is the fact that it does not rely on the speaker to pronounce words audibly. Coleman et al have established that the speech motor control plans for whispered speech and vocalized speech are similar [Coleman et al., 2002]. Yet, no study has investigated the differences between audible and non-audible speech relevant for MES based speech recognition. This issue was therefore another focus of our work.

Isolated word recognition is only of limited practical applicability. In order to move towards continuous speech recognition on large vocabularies we planned to perform initial experiments on connected digits recognition and to examine phoneme model based approaches.

Demo System

Another aim of our work was to implement an online recognition demo system showing the potential and practicability of EMG based speech recognition.

1.3 Approach

Our approach for solving the goals we presented in the previous section can be be divided into the following six steps.

Step one: Session Dependent Isolated Word Recognizer for Audible Speech

In order to build an initial EMG based speech recognizer and to gain experience with the new technology we used a small set of electrodes and a small set of words for recognition. Audible speech was produced to allow the recording of the *acoustic* signal which could be used to support the analysis of the EMG channels. We developed a baseline isolated word recognizer by exploring various features for an EMG speech recognition frontend, examining several HMM topologies and investigating useful segmentation methods.

Step two: Session *Independent* Isolated Word Recognizer for Audible Speech

In a second step we investigated session *independent* recognition of audible speech, that is, we conducted *across-sessions* experiments by testing a system on data from recording sessions that were not used for training the system. Optimizing across-sessions results involved investigating methods to obtain reliable electrode placement and exploring feature normalization as well as model adaptation schemes.

Step three: Isolated Word Recognizer for *Non-audible* Speech

After having gained experience in the recognition of audible speech we approached non-audible speech recognition. We examined differences between non-audible and audible speech, compared the corresponding recognition rates and investigated implementing a recognizer that works accurately on both speech manners.

Step four: *Optimization* of Isolated Word Recognizer

In order to achieve state-of-the-art recognition rates for isolated word recognition we purchased a second physiological data recording system providing eight EMG channels instead of two. Experiments on electrode placement yielded a set of positions which was then used to optimize the frontend and the Hidden Markov classifier of the previously obtained baseline system.

Step five: *Connected* Words Recognition

In order to move towards continuous speech recognition we performed initial experiments on connected words recognition and and explored phoneme model based approaches. The goal was to identify the main challenges associated with large vocabulary continuous speech recognition and to compare recognition results for isolated digits and connected digits.

Step six: Implementation of Demo Systems

Finally, we implemented two demo systems showing different applications of EMG based speech recognition: a prototype “silent” mobile phone suitable for conducting non-disturbing, non-audibly spoken phone calls in situations requiring silence (e.g. in a meeting) and a prototype lecture translation system that can be applied for an online-translation of a non-audibly spoken talk into a language of one’s choice. Both systems were trained on a set of *sentences* suitable for the corresponding application.

1.4 Structure of the Report

The remaining part of this thesis is structured as follows:
Chapter 2 gives background information necessary for understanding the concepts of this work. It provides both, basics of electromyography from the physiological and measurement point of view as well as an introduction to speech production and automatic speech recognition.
Related work on EMG based speech recognition is presented in chapter 3.
In chapter 4 we give an overview of our speech recognition system which includes a description of the hardware and software we deployed and implemented.
The corpora we produced for our work are introduced in chapter 5.
Chapter 6 is the core chapter of this work, presenting the most relevant experiments we conducted on EMG based speech recognition. The first part describes the development of a state-of-the-art system presenting experiments on electrode positioning, feature extraction, HMM classification and segmentation methods. In the second part of the chapter, we deal with issues in EMG based speech recognition that have not yet been addressed in literature, namely with session independence, the comparison of audible and non-audible speech, and connected digits recognition. Furthermore, the performance of EMG based speech recognizers is compared to the performance of conventional speech recognition systems.
In the following chapter 7 we describe the two demo systems we have implemented to show the applicability of EMG based speech recognition.
We end this thesis with chapter 8 which summarizes our work and provides suggestions for future directions.

Chapter 2

Background

The purpose of this chapter is to provide background information necessary for understanding the concepts of this work. Section 2.1 explains anatomical and physiological basics of surface Electromyography (sEMG) while section 2.2 gives an introduction to sEMG measurement and processing. The nature of speech production is described in section 2.3 while section 2.4 presents basic concepts of state of the art speech recognition systems.

2.1 Anatomical and Physiological Basics

This section aims to provide information fundamental to understanding the recording of the electrical activity of muscle using surface electromyography. After an introduction of the anatomy of skeletal muscle in section 2.1.1 the origin of the electromyographic signal is explained in section 2.1.2.

2.1.1 Anatomy of the Skeletal Muscles

Skeletal muscle - as opposed to heart muscle and smooth muscle - is the muscle attached to the skeleton. Its structural unit is the muscle fiber - an elongated cell ranging from 10 to 100 microns in diameter and from a few millimeters to 40cm in length [Soderberg, 1992]. Each muscle fiber is surrounded by a layer of connective tissue called *endomysium*. Groups of muscle fibers are wrapped by another layer of connective tissue called *perimysium* to form muscle bundles or *fascicles*. Skeletal muscles are composed of numerous fascicles. They are usually attached to bones via *tendons* composed of epimysium (figure 2.1).

The contraction of skeletal muscle is controlled by the nervous system. Each muscle fiber can be activated by one *motor neuron* (i.e. by one nerve) yet, one motor neuron can branch in up to several thousand branches, each one terminating in a different muscle fiber. A motor neuron and all the fibers it innervates is called a *motor unit* (figure 2.2). The term *neuromuscular junction* refers to the junction between a muscle fiber and the terminal of the motor neuron it is recruited by.

2.1.2 Origin of the Electromyographic Signal

Resting Membrane Potential

The membrane of a cell serves to separate the extra cellular space from the inner cellular space. The plasma membrane of a muscle cell is called *sarcolemma*. It is composed of a lipid bilayer that has channels by which certain ions can enter and leave the cell. Due to the fact that the membrane can regulate the movement of ions between the extracellular fluid and the intracellular fluid the intracellular and extracellular ion concentration differ from each other (table 2.1 and figure 2.3). The composition of the fluids can be explained by the following phenomena:

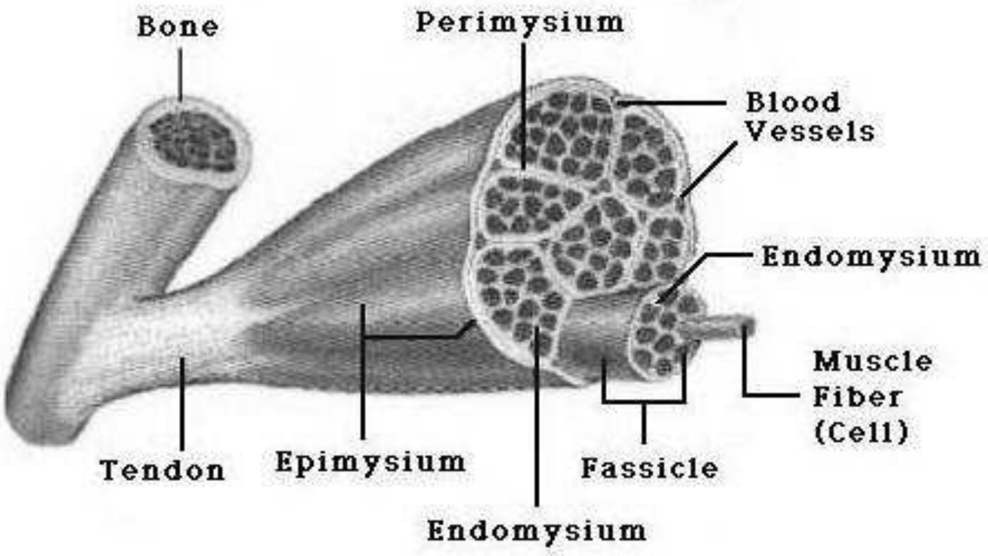


Figure 2.1: Composition of Skeletal Muscle [Ritchison, 2005]

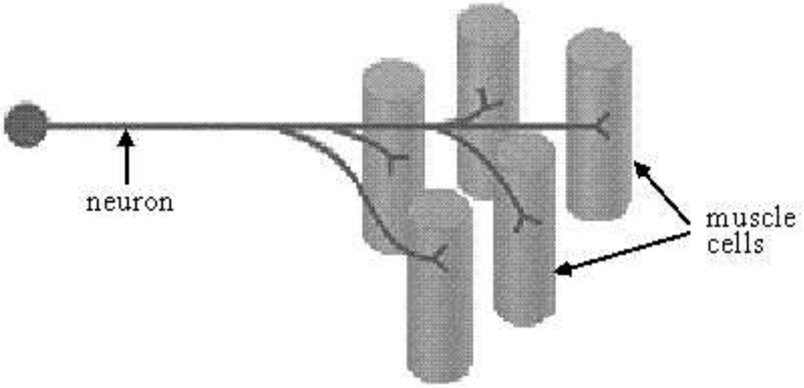


Figure 2.2: Motor Unit [Ritchison, 2005]

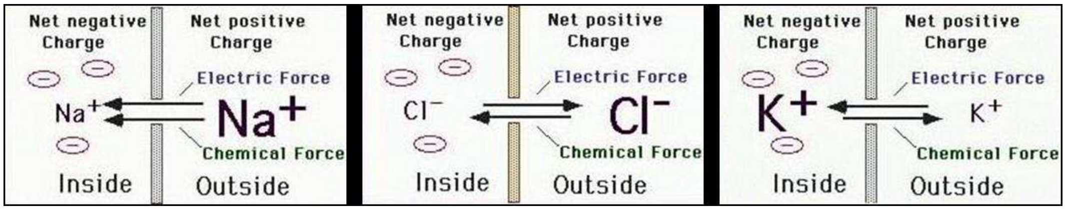


Figure 2.3: Electrical and chemical forces acting on Na^{2+}, K^+ and Cl^- ions in the intracellular and extracellular fluids [Ritchison, 2005]

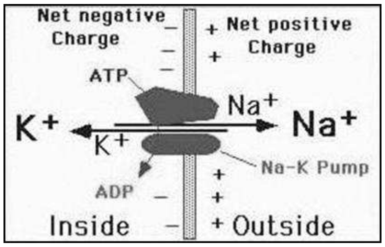


Figure 2.4: The Sodium-Potassium pump [Ritchison, 2005] constantly transports Na^{2+} ions from the inner cellular space to the extracellular space and K^+ ions back into the cell.

- *Electrical gradient*: a difference in potential between the intracellular and the extracellular fluid draws negative ions (such as Cl^-) to the location of higher potential and positive ions (such as Na^{2+} and K^+) to the location of lower potential (figure 2.3).
- *Chemical gradient*: *Diffusion* is the passive movement of a substance from an area of high concentration to an area of lower concentration by means of random molecular motion. The *concentration gradient* (i.e. the difference in concentration between two regions) draws ions to the region of lower concentration (figure 2.3).
- *Ion pumps*: The membrane contains a number of ion pumps that actively transport certain ions to the other side of the membrane using metabolic energy. The so-called sodium-potassium pump is the most relevant one in this context. It constantly transports Na^{2+} ions from the inner cellular space to the extra cellular space and K^+ ions back into the cell (figure 2.4).
- *Membrane permeability*: Cell membranes are semipermeable, that is, they act as barriers to some, but not all, substances. Moreover, some molecules can pass the membrane more easily than others. The sarcolemma, for instance, is much more permeable for Cl^- ions than for Na^{2+} ions.

Ion	Intracellular Fluid	Extracellular Fluid
K^+	140	4
Na^{2+}	14	142
Cl^-	4	125
HCO_3^-	8	28
A^-	150	-

Table 2.1: Intracellular and Extracellular Ion Concentration for Mammalian Muscle (mEQ/L) [Lamb and Hobart, 1992]

The effect of the electrical and chemical gradient and the active transport system results in a constant movement of ions between the intracellular and the extracellular fluid. When the muscle is in a resting

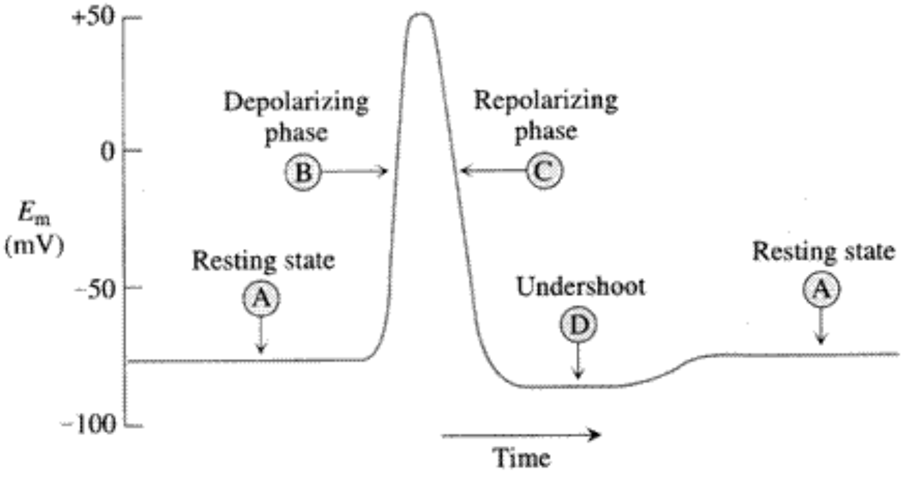


Figure 2.5: Action Potential involving the phases depolarization (B), repolarization(C), and hyperpolarization(D) [Matthews, 1991]

state, an equilibrium is reached where the concentration of ions is stable - that is, the number of ions leaving the cell is equal to the number of ions entering the cell in any given time interval. For example, there is a constant movement of Na^{2+} ions towards the intracellular space because of the higher Na^{2+} concentration outside the cell and the negative potential inside the cell (with respect to the outside) (figure 2.3). However, no *netto* movement takes place because at the same time the sodium-potassium pump causes Na^{2+} ions to leave the cell. The potential difference across the membrane corresponding to the equilibrium when the muscle fiber is in a resting state is referred to as *resting membrane potential.* It typically measures about -80mV inside the cell with respect to the outside.

Action Potential

An action potential is the rapid change in membrane potential occurring in nerve or muscle cells when excitation occurs. It minimally involves the phases *depolarization*, *repolarization* and *hyperpolarization* (figure 2.5).

Depolarization: Action potentials are triggered by an initial depolarization of the cell: due to a chemical, electrical or physical stimulation the membrane potential increases (i.e. becomes less negative). When a certain threshold value is exceeded, voltage-gated ion channels are opened and thus change the permeability of the membrane to specific ions - namely to sodium (Na^{2+}), calcium (Ca^{2+})) or both. Consequently, positively charged ions move into the cell along the concentration gradient. As a result, the membrane potential increases and temporarily even changes its polarity reaching about +20mV.

Repolarization: The voltage-gated sodium/calcium channels close after a fixed period of time (typically after about 1ms in skeletal muscle fiber) Moreover voltage-gated potassium channels are opened as a response to the cell's depolarization causing K^+ ions to leave the cell (because of the concentration gradient *and* the electrical gradient). Consequently, the membrane potential returns to a negative inside potential.

Hyperpolarization: The relatively slow closing of the potassium channels causes a period of hyperpolarization, when the membrane potential is more negative than in the resting state.

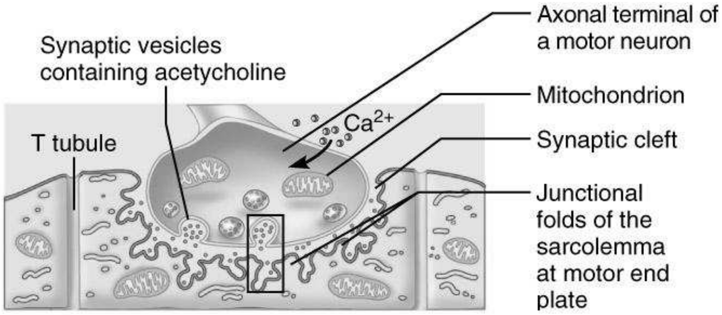

Axonal terminal of a motor neuron
Synaptic vesicles containing acetycholine
Mitochondrion
Ca2+
T tubule
Synaptic cleft
Junctional folds of the sarcolemma at motor end plate

Figure 2.6: Neuromuscular Junction [Noffke, 2005]

Finally, the resting membrane potential is re-established. Figure 2.5 illustrates the process. It is worth mentioning here, that an action potential is always followed by a so-called *refractory period*, where the cell cannot respond to another stimulus. More detailed explanations on the genesis of the action potential can be found in most physiology books (for example [S.Silbernagel and Despopoulos, 2003]).

Muscle Contraction

In order for a muscle fiber to contract the central nervous system has to activate the corresponding motor neuron by initiating a depolarization. The depolarization is conducted along the motor neuron and finally reaches the neuromuscular junction where it causes the release of a chemical substance called Acetylcholin (ACh) (figure 2.6).

Acting as a so-called neurotransmitter, ACh causes an action potential in the corresponding muscle fiber under the motor endplate. As a result, a potential difference is established between the active region and the adjacent inactive regions of the muscle fiber (figure 2.7). Consequently, ions are exchanged between the active and inactive regions causing depolarization (and thus action potentials) in the adjacent regions. This way, action potentials are propagated away from the motor endplate in both directions of the fiber. They spread along the sarcolemma and deep into the muscle fiber through the so-called *transverse tubules* (figure 2.8). Transverse tubules are invaginations of the cell membrane that almost touch the *sarcoplasmic reticulum* - a structure within the cell that serves as a storage for Ca^{2+}. Action potentials in the transverse tubules stimulate the release of Ca^{2+} from the sarcoplasmic reticulum and the increase in Ca2+ in the inner cellular fluid leads to the contraction of the muscle fiber. For a detailed explanation of the underlying mechanisms please refer to [S.Silbernagel and Despopoulos, 2003]).

Extracellular Recording of Action Potentials

A single muscle action potential can only be seen in isolation using microelectrode techniques. Surface EMG is "the temporal and spatial summation of *all* active motor units within the recording area" of the electrodes being used [Lamb and Hobart, 1992]. In order to understand how extracellular electrodes can be used to detect action potentials consider placing two electrodes, A and B, on the surface of a muscle fiber as illustrated in figure 2.9. When the muscle is in equilibrium, there is no potential difference between the electrodes because both are placed on the outside of the cell. When an action potential is initiated to the left of electrode A, however, it reaches the region under electrode A before it reaches the region under electrode B. As a result, a difference in potential can be detected between the two electrodes. Figure 2.9 shows the course of the potential differences between the electrodes.

The following section explains how surface EMG measurements are conducted in practice.

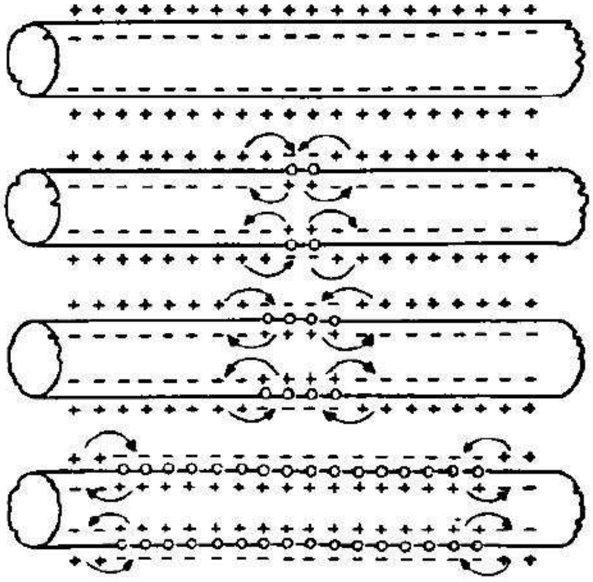


Figure 2.7: Propagation of action potentials in both directions along a conductive fiber [Lamb and Hobart, 1992]

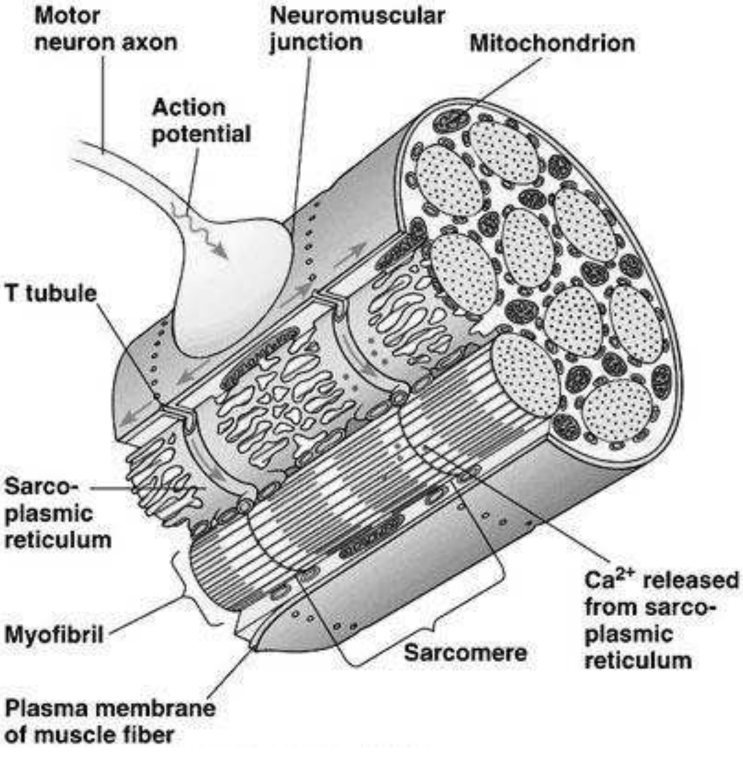


Figure 2.8: Propagation of action potentials along the sarcolemma into the muscle fiber through the transverse tubuli (t-tubuli) [Noffke, 2005].

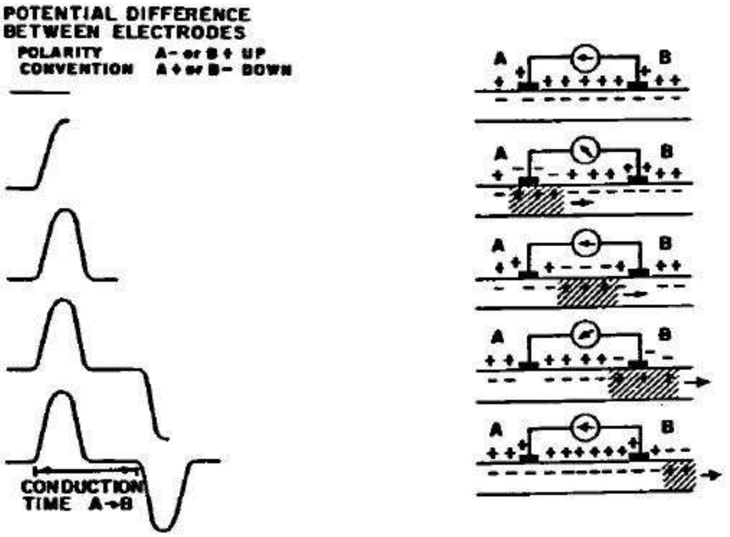


Figure 2.9: The measurement of action potentials with electrodes placed on isolated irritable tissue [Lamb and Hobart, 1992]

2.2 Surface EMG Measurement

Electromyography is the process of recording the electrical activity of a muscle. As explained in section 2.1.2 muscle fibers generate small electrical currents as part of the signaling process for the muscle fibers to contract. There are two basic methods to measure the signal called Electromyogram: invasively using fine wire electrodes that are inserted directly into the muscle or non-invasively by applying the electrodes to the skin surface.

Fine wire electrodes allow the testing of deep or small muscles and have a more specific pick-up area than surface electrodes. However, the needles may cause discomfort and the measurements should only be carried out by a medical doctor. Moreover, it is extremely difficult to identify the same point of insertion in consecutive recording sessions. As a result, surface EMG (sEMG) is the more common method of measurement. There is more potential for *cross-talk* (section 2.2.3) from adjacent muscles and only the signals from surface muscles can be adequately measured, yet, surface electrodes are easy to apply and their application does not involve physical pain. In this chapter, we will therefore focus on surface EMG measurement. The reader may refer to [Ankrum, 2000] and [Luca, 2002] for a more detailed introduction into the subject.

2.2.1 Equipment

The following equipment is necessary for surface EMG recordings:

Electrodes: Generally speaking, surface electrodes convert the ionic currents generated by muscle contraction into electronic currents that can be fed into electronic devices. While the *detection electrodes* serve to pick up the desired signal, the *ground* electrode provides a common reference to the differential input of the preamplifier. Refer to section 2.2.2 for more details on properties of surface electrodes.

Differential Amplifier: When detecting an EMG signal, amplification is necessary to optimize the resolution of the digitizing equipment [Scott, 2003]. Moreover, an amplifier can also be used to maximize the signal-to-noise ratio - that is, the ratio of the energy of the wanted EMG signal to the energy of unwanted noise contributions of the environment. For that reason sEMG recordings generally involve a differential detecting configuration as schematically shown in figure 2.10. The

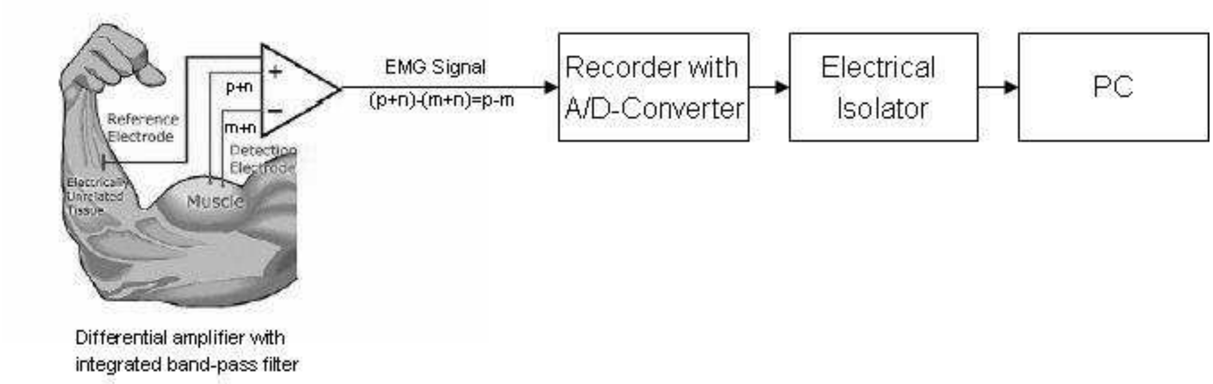

Reference Electrode
Electrically Unrelated Tissue
p+n
m+n
+
−
Detection Electrode
Muscle
EMG Signal
(p+n)-(m+n)=p-m
Recorder with A/D-Converter
Electrical Isolator
PC
Differential amplifier with integrated band-pass filter

Figure 2.10: Equipment required for sEMG measurements. The EMG signals are represented by "p" and "m" and the noise signals by "n" [Luca, 2002]

idea is simple: A differential amplifier subtracts the signals from two detection sites and amplifies the difference voltage between its two input terminals. As a consequence, signals common to both electrodes - such as noise originating far away from the detection sites - should ideally produce a zero output, whereas local EMG signals are amplified. The *Common Mode Rejection Ratio (CMRR)* is a measure of the degree to which this ideal is realized in practical designs. It is defined as the difference signal gain divided by the common mode signal gain. An ideal differential amplifier would thus have a CMRR of infinity, yet, in practice, only amplifiers with a maximum CMRR of approximately 120dB are available. As a result, it is not possible to obtain a signal free from noise, however, a CMRR of 90dB ($CMRR(x)[dB] = 20 * log_{10}CMRR(x)$) normally results in sufficient noise suppression [Soderberg, 1992]. Please refer to [Scott, 2003] or [Soderberg, 1992] for a more detailed description of the properties of ideal and realistic amplifiers.

Electrical Isolator: The failure of any electrical device that has galvanic contact with the subject can cause a potentially harmful current to pass through the skin. In order to ensure safety the subject must therefore be electrically isolated from any electrical connection to the power source. This can be achieved by placing an optical isolater between the amplifier and the devices that are connected to the power point (e.g. the computer) [Luca, 2002].

A/D-converter: EMG signals usually need to be digitized for further processing and data analysis. The analog-to-digital converter transforms an analog signal into a discrete number of data points representing the amplitude of the input signal at particular instances in time.

Recorder: The purpose of the recorder is to generate a time record of the input EMG signal that can be reviewed later for data analysis.

Figure 2.10 schematically shows how the individual components work together.

2.2.2 Electrodes

Electrodes serve as converters of the ionic currents produced in muscles into electronic currents that can be manipulated in electronic circuits. This section gives background knowledge on the use of surface electrodes in EMG measurements.

Dry vs Gelled Electrodes

There are two main types of surface electrodes: dry electrodes that have direct contact with the skin, and gelled electrodes, where an electrolytic gel is placed between the metallic part of the electrode and the skin to decrease the skin-electrode impedance. [Scott, 2003]

Dry electrodes are typically used when the constitution of the electrodes does not allow the use of gel (e.g. bar electrodes). Due to the high electrode-skin impedance it is common to have the preamplifier circuitry at the electrode site. This makes the dry electrodes considerably heavier than gelled electrodes (about 20g vs. 1g) such that electrode fixation becomes an issue.
Gelled electrodes are therefore the common choice. Oxidative and reductive chemical reactions in the contact region of the metal surface and the gel allow an exchange between the ionic current generated by muscle contraction and the electron current flow of the recording instrumentation. It is worth mentioning here, that the quality of an electrode depends almost exclusively on its ability to exchange ions for electrons and vice versa [Soderberg, 1992]. A general explanation of the mode of operation of surface electrodes from the chemical point of view may be found in [Meyer-Waarden, 1985] and [Soderberg, 1992].

Electrode Properties

Surface electrodes differ in shape, size and material. Moreover, inter-electrode distance also plays a crucial role for EMG measurements. The SENIAM initiative (Surface Electromyography for the Non-Invasive Assessment of Muscles) is a European project that “has resulted in European recommendations for sensors and sensor placement procedures and signal processing methods for SEMG” [Hermens and Freriks, 2005]. The recommendations for the use of bipolar sEMG electrodes include

- Electrode material: Pre-gelled Ag/AgCl electrodes. (disposable or reusable)
- Electrode size: size of the electrode in direction of the muscle fiber should not exceed 10mm
- inter-electrode distance: 20mm

For more details on the SENIAM recommendations refer to [Scott, 2003] and [Hermens and Freriks, 2005].

Electrode Placement

When determining electrode positions it is desirable to identify locations where a good and stable sEMG signal can be obtained. Despite the use of surface EMG in wide variety of disciplines, only little information on optimal electrode positions is available in literature. The SENIAM initiative set forth some guidelines for determining electrode placements. Generally speaking, a location should be defined in relation to a line between two anatomical landmarks (e.g. bones). When using bipolar electrodes, the general recommendation is to choose an arrangement longitudinal to the long axis of the muscle of interest. The electrode should be placed “between a motor point and the tendon insertion or between two motor points”, where the motor point is defined as “ that point on the muscle where the introduction of minimal electrical current causes a perceptible twitch of the surface muscle fibers” [Luca, 2002]. A detailed description can be found in [Luca, 2002].
The ground electrode should be placed on electrically neutral (e.g. bony) tissue relatively far away from the detection site. This way it can serve as common reference to the differential input of the preamplifier.

2.2.3 Signal Characteristics

The raw EMG signal detected by a differential amplifier using surface electrodes is a “bipolar signal whose random fluctuations, if summed over a significantly long time period, would produce a zero result” [Lamb and Hobart, 1992]. Its amplitude typically ranges from 0.01 to 5 mV [Lamb and Hobart, 1992]. Figure 2.11 shows an example of an EMG signal and the corresponding frequency spectrum. The usable energy of the signal is contained in the 0 to 500Hz frequency range, that is, the signal energy is above the electrical noise level in that frequency band. In fact, the dominant energy lies in the 50-150Hz frequency range [Luca, 2002].

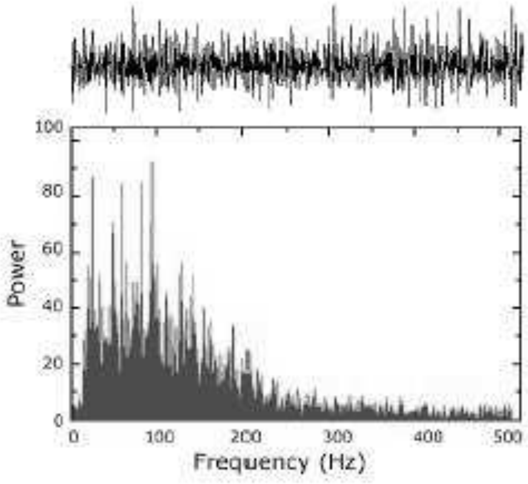


Figure 2.11: EMG signal and corresponding frequency spectrum detected from the Tibialis Anterior muscle during a constant force contraction [Luca, 2002]

Factors that influence the sEMG signal

The sEMG signal is the result of many anatomical, physiological and technical factors. Generally speaking, the effect of these factors on the electromyographic signal can be *qualitatively* characterized, yet, there exists no complex model that allows the deduction of *quantitative* relationships. According to [Scott, 2003] the time and frequency domain properties of the sEMG signal are dependent on the following factors:

- the timing and intensity of muscle contraction
- the distance of the electrode from the active muscle area
- the properties of the overlying tissue
- the electrode and amplifier properties
- the quality of contact between the electrode and the skin

In most applications, only the information on the timing and intensity of muscle contraction is desired whereas the rest of the factors merely increase the variability of EMG records. In order to ensure the comparability of consecutive recordings, it is therefore necessary to work with the same equipment and to place the electrodes on the same skin location in consecutive recording sessions. For a more detailed description on the influence of a number of factors on the electromyographic signal refer to [Soderberg, 1992].

Noise

Noise is defined as any unwanted signal collected along side the wanted signal. Sources of noise include [Luca, 2002] [Scott, 2003]:

- *Ambient noise*: This noise is generated from electromagnetic devices such as computers and power lines. It has a wide range of frequency components, yet, the dominant frequency component is 50Hz or 60Hz depending on the frequency of the power supply. The exposure to ambient noise can be reduced by carrying out the recordings in a room that contains a minimum of electronic equipment.

- *Inherent noise in the equipment*: Electronic equipment always generates electrical noise with frequency components ranging from 0Hz to several thousand Hz. This noise can not be eliminated but the use of high quality devices and appropriate circuit design can reduce it.

- *Electrode contact*: The signal to noise ratio is particularly determined by the properties of the electrode-electrolyte-skin contact [Scott, 2003]. For this reason, it is recommended to prepare the skin (e.g. clean it with alcohol) prior to recording and to ensure a proper electrode-skin contact.

- *Cross-Talk*: Cross-talk is “interference of the EMG signals from adjacent muscles or deeper muscles that are within the pick-up area of the electrode” [Rash, 1999]. It can be reduced by choosing an appropriate electrode size and inter-electrode distance [Scott, 2003].

- *Movement Artifacts*: these artifacts can result from a movement disturbance of the electrode-electrolyte interface or from cable movement. Again, most of the generated energy lies in the frequency components between 0 and 20Hz so that the use of a highpass filter is advisable (see below).

2.2.4 Signal Processing

Filtering

It is well established that the bipolar electrode configuration has a bandpass filtering effect in the spectral frequency region of the EMG signal. It can be explained by the differences in the time of arrival of the signal at each detection site: due to the fact that the differential amplifier merely amplifies differences in potential, a signal frequency “whose wavelength is equal to the interelectrode distance or is an integer multiple of that frequency would be cancelled” [Soderberg, 1992]. See [Soderberg, 1992] for further details.
Despite this, it is advisable to apply a low-pass filter in order to avoid aliasing. High-pass filters with a cut-off frequency between 10 and 20Hz are also often used to remove movement artifacts.
In the past, it was common to use a 60Hz (50Hz respectively) notch filter for power-line noise removal. Yet, due to the fact that notch filtering results in the loss of important EMG signal information (a great part of the EMG energy lies in the 50/60Hz region), it is nowadays often avoided.

Normalization

As explained in section 2.1.2 the sEMG signal serves as a measurement of the electrical activity in a muscle during contraction. However, slight changes in electrode position, temperature or tissue properties may alter the signal significantly. In order to make comparisons of amplitudes possible it is therefore advisable to apply a normalization procedure at each recording that compensates for these changes. The most widely used method of normalization is to perform a reference contraction - the so-called *Isometric Maximal Voluntary Contraction (MVC)* - and to express all myoelectric values obtained as a percentage of the MVC [Leveau and Andersson, 1992]. Again, further details can be found in [Leveau and Andersson, 1992].

Signal Interpretation

In the past, the most common way to interpret EMG was by visual inspection of the unprocessed signal. However, the raw signal only provides limited information. For this reason, many *time* domain representations of the myoelectric signal have been introduced for data analysis. The *linear envelope* can be used to provide a profile of the activity of the muscle over time while the *root-mean-squared (RMS)* voltage is applied to measure the electrical power in the signal. Both methods make use of absolute values instead of the actual time domain values. Moreover, pattern recognition system based on myoelectric signals often deploy a *frequency analysis* of the signal [Leveau and Andersson, 1992]. Englehart et al. reported that feature sets based on the *Short Time Fourier Tranform (STFT)*, the

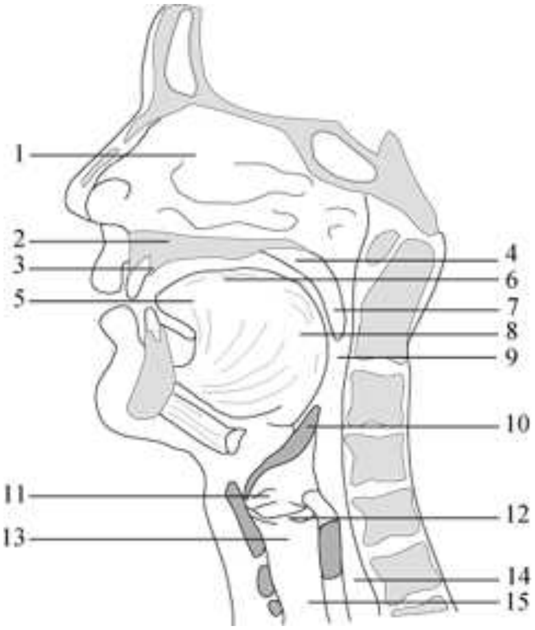


Figure 2.12: The human vocal organs. (1) Nasal cavity, (2) Hard palate, (3) Alveoral ridge, (4) Soft palate (Velum), (5) Tip of the tongue (Apex), (6) Dorsum, (7) Uvula, (8) Radix, (9) Pharynx, (10) Epiglottis, (11) False vocal cords, (12) Vocal cords, (13) Larynx, (14) Esophagus, and (15) Trachea. [Lemmetty, 1999]

Wavelet Transform, or the *Wavelet Packet Transform* provide an effective representation for classification provided that they are subject to an appropriate form of dimensionality reduction [Englehart et al., 1999]. A detailed introduction into EMG signal processing methods can be found in [Leveau and Andersson, 1992].

2.3 Speech Production

2.3.1 Human Speech Organs

Human speech is produced in cooperation by the vocal organs presented in Figure 2.12. When speaking, an airstream is forced through the *glottis* (the space between the *vocal cords*) (12) and the *larynx* (13) to the three main cavities of the vocal tract: the *pharynx* (9) and the *oral* and *nasal cavities* (1). From the oral and nasal cavities the air flow exits through the nose and mouth, respectively.

Different sounds result from different modifications of the airstream. They can roughly be divided into *voiced* and *unvoiced* sounds. When a voiced sound is produced, the vocal cords vibrate (i.e. close and open rapidly) and modulate the air flow such that a quasi-periodic pressure wave is produced. When there is no vibration of the vocal cords the resulting sound is called unvoiced. In this case the vocal cords can for example be completely open (like for the unvoiced consonants /s/ or /f/) or change from a closed position to an open position to produce a stop consonant like /p/. The voiced sounds determine the speech "melody" [Lemmetty, 1999].

The oral cavity plays a major role in sound production because its size and shape can be varied considerably by movements of the tongue, the lips, the cheeks and the teeth to modify the incoming airstream.

Despite the fact that the speech organs are in constant motion during the act of speaking it is possible to segment a speech signal by identifying points where linguistically relevant changes occur. In order to describe the pronunciation of every possible word in a given language a minimal set of symbols

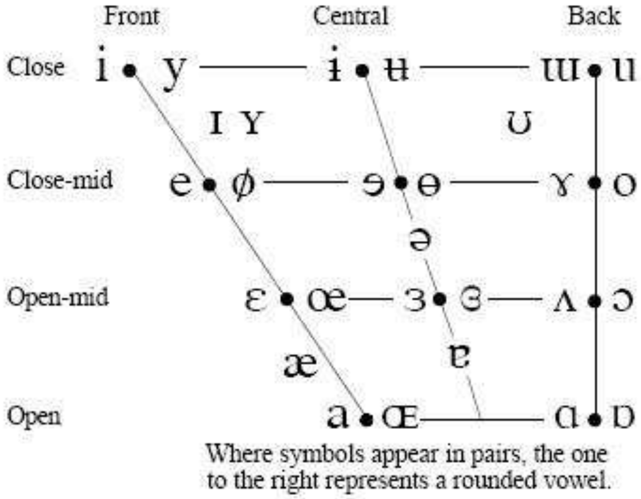


Figure 2.13: The vowel quadriliteral from the IPA chart [Stüker, 2003]. Vertical position of the dorsum (row), horizontal position of the dorsum (column).

each representing a certain sound can be defined. The following section addresses this issue.

2.3.2 Phonetic Alphabet

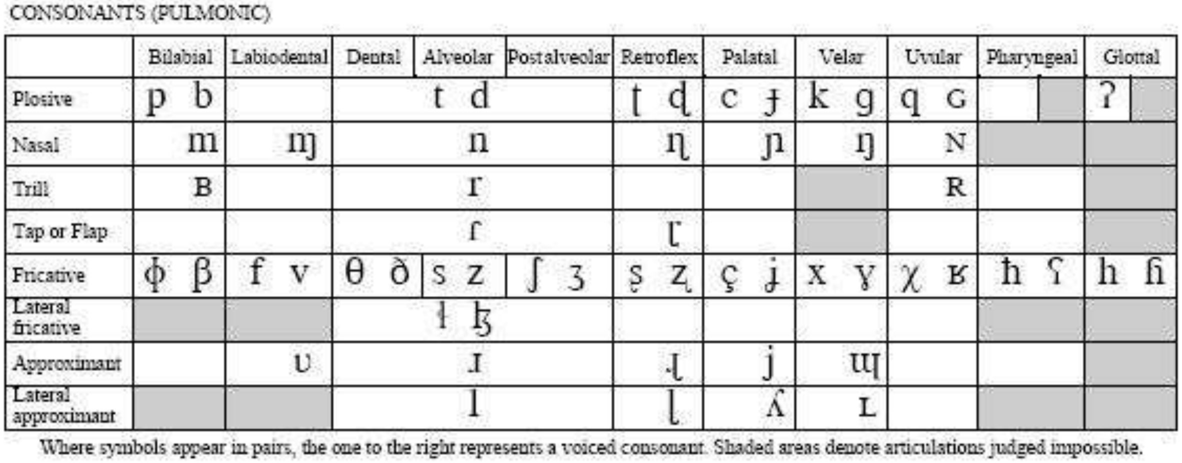
A phoneme is the smallest contrastive unit in the sound system of a language. In other words, the set of phonemes corresponding to a given language is the minimum number of symbols needed to describe the pronunciation of every possible word in that language. The number of phonetic symbols ranges from 20 to 60 for different languages. The IPA (International Phonetic Association) - the major as well as the oldest representative organisation for phoneticians - constructed a language-independent phonetic alphabet, the *International Phonetic Alphabet* (also IPA).

The pronunciation of a particular phoneme depends on contextual effects, speaker's characteristics, and emotions. When continuous speech is produced the articulators are in different positions depending on the preceding and following phoneme. The variations in pronunciation in individual phonemes are called *allophones*. Thus, each allophone is a specialization of a phoneme.

A phonetic alphabet is usually divided into two main categories: *vowels* and *consonants*. Vowels are voiced sounds where the air flows inhibited through the vocal tract. Consonants, on the other hand, are produced by a narrow or closed vocal tract and may be either voiced or unvoiced [Lemmetty, 1999].

Vowels are mainly distinguished by the horizontal and vertical position of the highest point of the tongue, called the *dorsum*. Figure 2.13 shows the so-called *vowel quadrilateral* which represents the space of all possible vowels. The *height* parameter refers to the notional height of the dorsum during production of a vowel and can take the values *close, close-mit, open-mid, and open*. The parameter *backness* describes how far forward or back the dorsum lies (*front, central, or back*). Finally, the *rounding* parameter refers to the position of the *lips* during vowel production. Refer to [Bowden and HAJEK, 1999] for further information.

Consonants are defined by the *place of articulation* (describing the position of the constriction of the vocal tract on the mid-sagittal plane), the *manner of articulation* and by the fact if they are voiced or unvoiced (figure 2.14). The manner of articulation is defined by several different factors and is represented by different categories as shown in figure 2.14. The category plosiv, for instance refers to sounds that are produced by a complete closure of the vocal tract followed by a sudden release of the air. (e.g. /d/, /p/) while fricatives are associated with a near complete stoppage of air where friction occurs between the airstream and the speech organs (e.g. /f/). For a detailed description on



CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap				ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 2.14: The consonant table from the IPA chart [Stüker, 2003]. Manner of articulation (row), place of articulation (column).

the classification on human sounds refer to [Bowden and HAJEK, 1999].

2.3.3 Muscles Involved in Speech Production

A complete list of muscles involved in speech production, their exact location and functionality can be found in [Laboratory, 2002]. Figure 2.15 shows the muscles we have used for EMG based speech recognition experiments. Table 2.2 lists the corresponding functions.

2.4 Speech Recognition

The aim of this section is to give a brief overview of state-of-the-art speech recognition systems. It serves as a memory refresher rather than as a detailed introduction into the subject and requires some background knowledge on Artificial Intelligence (AI) and Digital Signal Processing (DSP). The reader may refer to [Rogina, 2003] or [Rabiner and Juang, 1993] for more detailed information.

2.4.1 Overview

The goal of Automatic Speech Recognition (ASR) is to translate an acoustic signal into a sequence of words. This process can generally be divided into two steps: The so-called *frontend* of the ASR system *preprocesses* the incoming acoustic signal: The signal is low-pass filtered (to avoid aliasing artifacts), digitized, quantized and transformed into a sequence of *feature vectors*. These feature vectors are designed to preserve the information important for phonetic distinction whilst omitting irrelevant information. In a second step, the *decoder* translates the sequence of feature vectors into a sequence of words. The search space is limited by a dictionary and a language model which define the words the sequence may be composed of and the order in which these words may occur respectively. Figure 2.16 gives a schematic overview of a typical speech recognition system.

A typical frontend and decoder used in conventional speech recognition systems are introduced in sections 2.4.2 and 2.4.3 respectively.

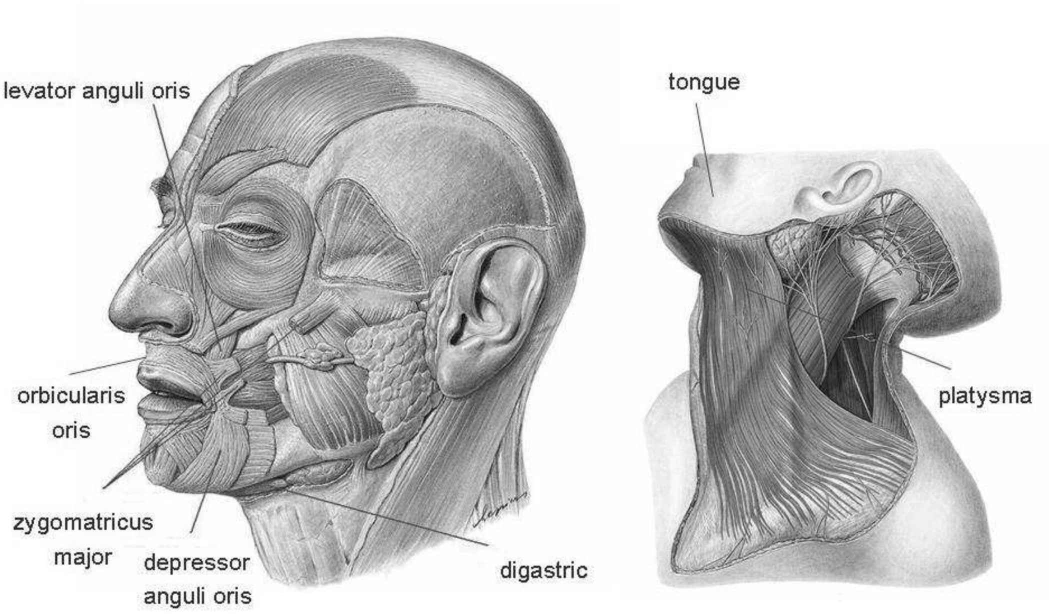


Figure 2.15: Muscles involved in speech production. [Matthews, 1991]

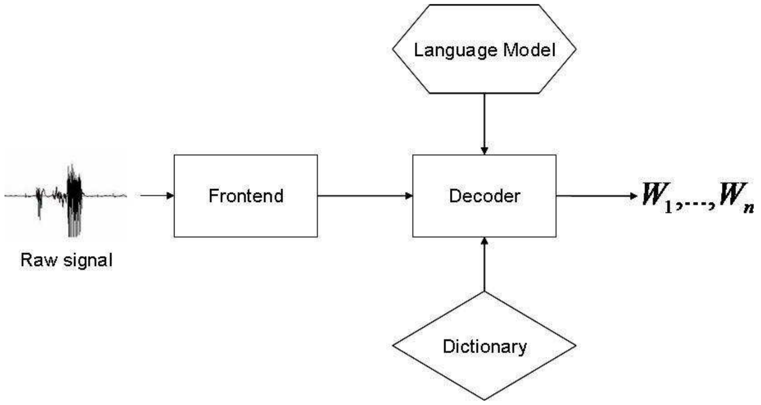


Figure 2.16: Schematic view of a typical speech recognizer

Muscle Name	Function
Orbicularis oris	On contraction, this muscle adducts the lips by drawing the lower lip up and the upper lip down, probably in conjunction with some of the other facial muscles. It may also pull the lips against the teeth. This muscle can also round the lips by its sphincter action.
Zygomaticus major	Raises upper lip for [f] along with the muscles that raise the angles of the mouth. On contraction, this muscle draws the angle of the mouth upward and laterally. The upward movement probably works with levator anguli oris to achieve the raised upper lip in labiodental fricatives. The lateral movement may be used in the production of [s].
Levator Anguli Oris	This muscle draws the corner of the mouth upwards and, because of the fibers that insert into the lower lip, may assist in closing the mouth by drawing the lower lip up, for the closure phase in bilabial consonants.
Depressor Anguli Oris	This muscle depresses the angles of the lips. This action may work with depressor labii inferioris to prevent the mouth from closing entirely when spreading for vowels like [i] and [e]. Because of the fibers that insert in the upper lip, this muscle may also aid in compressing lips by drawing the upper lip down.
Platysma	The platysma can aid depressor anguli oris and depressor labii inferioris to draw down and laterally the angles of the mouth.
Anterior Belly of the Digastric	The function of this muscle is to draw the hyoid bone up and forward. It also serves to bring the tongue forward and upward for alveolar and high front vowel articulations. In pulling up the hyoid bone, it may also pull up the larynx thereby tensing the stretching the vocal cords and raising the pitch. If the hyoid bone is fixed, the anterior belly of the digastric can serve to lower the jaw in conjunction with the geniohyoid, mylohyoid and lateral pterygoid muscles.

Table 2.2: Subset of muscles involved in speech production and their corresponding functions. [Laboratory, 2002]

2.4.2 Preprocessing

Figure 2.17 presents a typical frontend for state-of-the art speech recognition systems. [Wölfel, 2003] describes the module in detail. We give a brief summary of the most important components at this point:

Speech Waveform: The incoming acoustic signal is sampled at 16kHz and quantized at 16bit. A high-pass filter is deployed to avoid aliasing artifacts.

Windowing: Speech is a so-called *quasi-stationary* signal, that is, the vocal tract shape remains nearly constant over short periods of time (5-25ms). Typical preprocessing components like the Fourier Transform, however, assume *stationary* signals. Consequently, the incoming signal it transformed into a sequence of segments which can be assumed to be stationary. In order to achieve this, the complete signal is multiplied every 10ms (frame shift) with a window of typically 16ms length (frame size). The shape of the window determines the resolution of the speech segment in the frequency domain. The most commonly used windows are *Hamming windows* which are shown in figure 2.17.

FFT: The Fast Fourier Transform is computed for each window and its absolute value is squared such that the signal is decomposed into its frequency components.

VTLN: Vocal Tract Length normalization (VTLN) is applied to compensate for different anatomies of different speakers.

Mel Filterbank: The Fourier Coefficients are transformed into *Mel-scale filterbank coefficients* which imitate the frequency dependent spectral resolution of the human ear.

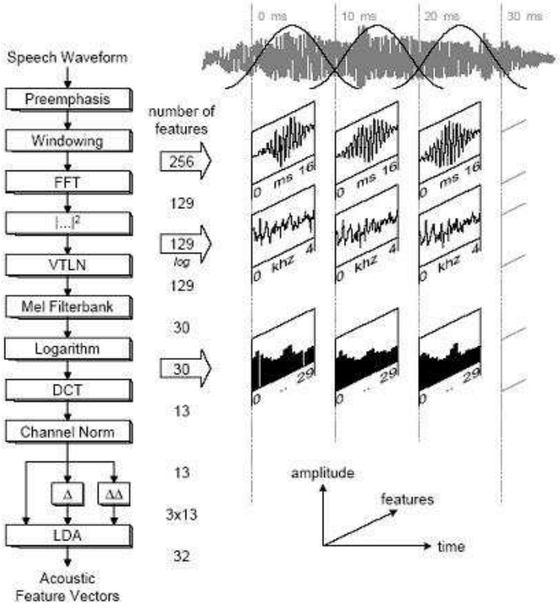


Figure 2.17: Typical frontend of a speech recognition system [Wölfel, 2003]

DCT: After applying the logarithm, the Mel Filterbank feature vectors are transformed by the *Discrete Cosine Transform (DCT)* into the ceptral space which is less sensitive to channel distortion and speaker variations such that *Mel-frequency cepstral coefficients (MFCC)* are obtained

LDA: A Linear Discriminant Analysis (LDA) is applied to further reduce the dimensionality of the feature vector. Each feature vector is assigned to a particular class (e.g. to a certain phoneme) and all vectors are linearly transformed into a space where vectors associated with the same class are as close as possible to each other (according to a similarity criterion) and vectors corresponding to different classes lie as far apart as possible.

After the preprocessing step the obtained sequence of feature vectors is translated into a sequence of words by the decoder. The decoding process is based on statistical models which are described in the following section.

2.4.3 Modelling and Classification

Maximum Likelihood Criterion

The *Decoder* translates a given sequence of feature vectors into a sequence of words according to the *Maximum Likelihood (ML)* criterion. That is, it identifies the most probable sequence of words $\hat{W}$ given the sequence of feature vectors X:

$$\hat{W} = \arg\max_{W} P(W|X) \tag{2.1}$$

By applying Bayes-Rule to the probability $P(W|X)$ we obtain:

$$P(W|X) = \frac{P(X|W) \cdot P(W)}{P(X)} \tag{2.2}$$

where $P(X)$ denotes the prior probability to observe the sequence of feature vectors X, $P(X|W)$ represents the probability that, given the sequence of words W, the feature vectors X are observed

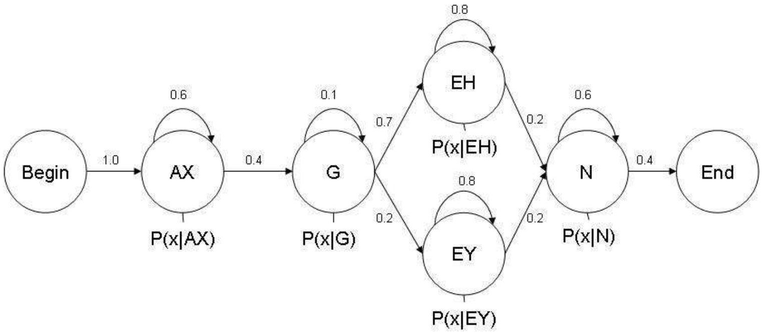


Figure 2.18: Hidden Markov Model for the word again.

and $P(W)$ represents the prior probability of observing W independently of the feature vector X. Due to the fact that $P(X)$ does not depend on W the equation can be reduced to

$$\hat{W} = \arg\max_{T} P(X|W) \cdot P(W) \tag{2.3}$$

This equation is known as the *Fundamental Equation of Speech Recognition.* $P(X|W)$ is called the *acoustic model* and P(W) is referred to as the *language model*.
$P(X|W)$ is most commonly determined using *Hidden Markov Models* (HMMs) as explained in the following section.

Hidden Markov Models for Speech Modelling

A Hidden Markov Model (HMM) is typically defined as a stochastic finite state automaton (SFSA) whose states Q are each associated with a specific probability distribution or probability density function, depending on the application. The states of a HMM are related by a certain *topology* which defines the possible *transitions* between the states.
In order to illustrate the use of Hidden Markov Models for speech modelling we give a simple example. Consider the word "again" with the two possible pronunciations

again(1): AX G EH N
again(2): AX G EY N

Let Λ_{again} be the Hidden Markov Model representing the word $W = again$. It is composed of five states each representing one of the occurring phonemes as shown in figure 2.18. The topology of the Hidden Markov Model defines the possible sequences of phonemes and the transition probabilities $P(q_j|q_i)$ represent the probability of one phoneme following another phoneme within the sequence of sounds. In our example the first pronunciation is more probable than the second one. The probability densities $P(x|q_i)$ present the probability of observing a feature vector x in state q_i. They are modelled by *Gaussian Mixture Models* (GMMs), that is, a weighted average of a set of multidimensional gaussians, each defined by a covariance matrix and a mean vector. Details can be found in [Rogina, 2003].
Let $X = x_1...x_n$ be a sequence of acoustic feature vectors observed for an utterance associated with the word "again". Each stationary speech segment x_i can be associated with one specific HMM state q_{x_i} such that X is represented by a succession of states $Q_x = q_{x_1}...q_{x_n}$ in Λ_{again}.
In order to decode an arbitrary utterance represented by a sequence of feature vectors $X = x_1, ..., x_n$ it is necessary to determine the probability $P(X|W)$ of observing X when W was spoken for every sequence of words W. For simplification, let us assume that the unknown word sequence consists of only one word. For each possible word W $P(X|W)$ can then be calculated from the transition probabilities and the mixture density functions of the corresponding Hidden Markov Model by iterating over all possible sequences of states:

$$P(X|W) = \sum_{S=s_1,...,s_n} P(X|W,S) \tag{2.4}$$

where $S = s_1, ..., s_n (s_i \in Q)$ represents a sequence of states of length n such that each x_i is associated with one state s_i. $P(X|W,S)$ can be optained by multiplying $P(x_i|s_i) * P(s_i|s_{i-1})$ for all vectors x_i. The Forward-Backward-Algorithm solves the problem efficiently. For ease of computation the sum in $P(X|W)$ is often replaced by the maximum to approximate $P(X|W)$ and save computational costs.

$$P(X|W) = \max_S P(X|W,S) \tag{2.5}$$

The latter equation is commonly solved by the so-called *Viterbi* algorithm. The path $\hat{S}$ for which 2.5 is maximized is called the *Viterbi path* of X for the word W.

Due to the fact that it is infeasible to construct a HMM for every possible utterance, a hierarchical scheme must be applied to reduce the number of models. One possible approach is to train one Hidden Markov Model with a particular topology (e.g. three states) for each *phoneme* and to then compose words of phonemes and sentences of words. Again, details can be found in [Rogina, 2003].

Hidden Markov Model Training

In order to construct a Hidden Markov Model for a given speech unit and a fixed topology it is necessary to define the gaussian mixture models associated with the individual states and find appropriate transition probabilities. The *Expectation-Maximation (EM) Algorithm* can be applied to set the required parameters based on statistics obtained from labelled training material. It is an iterative algorithm that can be used for finding Maximum Likelihood estimates of parameters in probabilistic models, where the model depends on *unobserved* or *hidden* variables.

In our case we are given a set of training feature vectors for each HMM state and the task is to estimate the mean vectors and covariance matrices of the gaussians associated with that state. Yet, we lack the information which gaussians the feature vectors correspond which is necessary to compute the missing parameters. Thus, in our example, the hidden variables represent the membership of feature vectors to gaussians.

The EM algorithm alternates between performing an “expectation (E) step, which computes the expected value of the unobserved latent variables, and a maximization (M) step, which computes the Maximum Likelihood estimates of the parameters given the data” and sets the unobserved variables to their expectation values. [Wikipedia, 2005]. Please refer to [Rogina, 2003] for a detailed explanation.

Context Dependency

As already mentioned above the pronunciation of a particular phoneme depends on its context. The words “two” and “true”, for instance, exemplify the possible acoustic differences of the phoneme /t/. State-of-the-art systems incorporate *context dependency* by using *triphones* as smallest speech units instead of phonemes. A triphone $X(Y|Z)$ represents the phoneme X occurring in the context Y (left) and Z (right) where Y and Z are also phonemes. When $n = 50$ is the size of the set of phonemes, there are $n^3 = 125000$ triphones. Due to the fact that it is infeasible to train one model per phoneme, a common approach is to *cluster* similar context and to represent a *set* of triphones by one model. Again, [Rogina, 2003] gives a detailed description.

Feature Space Adaptation

Feature Space Adaptation is a constrained Maximum Likelihood (ML) transformation of input features with the goal of adapting acoustic signals to a given speech recognizer. Incoming feature vectors are linearly transformed using an adaptation matrix, where the adaptation matrix is defined such that the overall probability of observing a given set of training utterances is maximized. The matrix can be computed in a supervised manner using labelled training data or in an unsupervised manner where hypothesis from a given recognizer are used for determining the matrix. A detailed description can be found in [Gales, 1997].

Chapter 3

Related Work

In this chapter we review the work that has so far been produced on speech recognition based on myoelectric signals. Section 3.1 gives an overview of the development of the new technology while sections 3.2 to 3.5 describe focuses of current research activity.

3.1 Overview

The use of myoelectric signals in speech dates back to the 1980's. In 1986, Morse et al. showed the availability of speech information in the myoelectric signals of neck and head muscles [Morse M., 1986]. In the following years, several research groups investigated the use of electromyography for isolated word recognition (e.g. [Morse et al., 1991]). The results were significantly better than chance but with maximum rates of 70% on a ten words vocabulary rather poor compared to conventional speech recognition standards. The first reasonable classification rates were reported in 2001 by Chan et al. who achieved an average word accuracy of 93% on the ten English digits vocabulary [Chan et al., 2001]. Moreover, the authors showed the potential of the myoelectric signal to augment conventional speech recognition systems [Chan et al., 2002a]. In 2003, Jorgensen et al. proved the applicability of the MES for *non-audible* speech recognition reporting 92% word accuracy on a set of six control words.

Current research focuses on vowel and consonant classification as a step towards continuous speech recognition [Jorgensen and Binsted, 2005] and on making EMG systems more user-friendly [Manabe and Z.Zhang, 2004]. The following sections give an introduction into state-of-the-art work.

3.2 Isolated word recognition

State of the art systems only achieve reasonable recognition results on *isolated* words. The best recognition rates have so far been reported by Chan et al. who proposed to apply ASR on the myoelectric signal in the context of aircraft pilot communication. Five bipolar electrodes were embedded in pilot oxygen masks and the myoelectric signals were recorded during audible pronunciation of the digits "zero" to "nine" (figure 3.1). A Linear Discriminant Analysis (LDA) classifier utilized on a set of wavelet transform features (reduced by principle component analysis (PCA)) yielded a word accuracy of 93% [Chan et al., 2001]. The LDA classifier required temporal alignment of the MES data which was accomplished by recording an additional acoustic channel. For each utterance a signal segment of size 1024ms beginning 500ms (pre-trigger value) before the start of audible speech was used for both, training and classification. It is worth mentioning here that pre-trigger values between 0s and 700s have been examined, yet a value of 500ms yielded the maximum classification rates. Due to the fact that the temporal position of articulation relative to the acoustic signal varies with the speaking rate, the authors proposed the use of a Hidden Markov Model (HMM) for MES based speech recognition in another paper [Chan et al., 2002b]. Even though the HMM classifier yielded worse maximum recognition rates (86%) than the LDA classifier for the same data, it was much less susceptiple to temporal

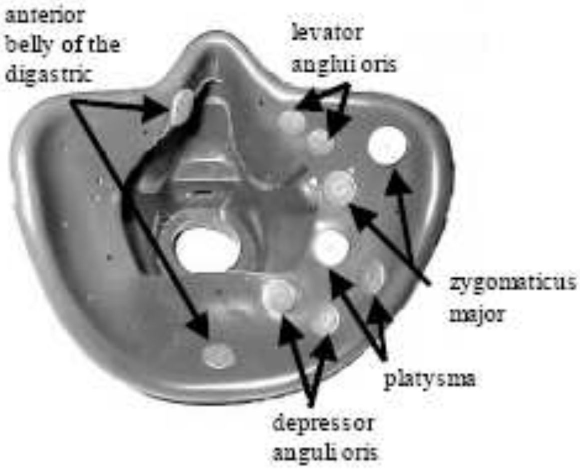


Figure 3.1: Oxygen mask with embedded electrodes [Chan et al., 2002b]

misalignment, that is, there was no dramatic decrease in performance when the pre-trigger value used for the training set was slightly different from the one used in the test set.

Jorgensen et al. demonstrated the applicability of electromyography for *non-audible* speech recognition. Their idea is to intercept nervous signal control signals sent to speech muscles using surface EMG electrodes placed on the larynx and sublingual areas below the jaw. The authors reported recognition rates of 92% on a set of six control words using a Neural Network classifier [Jorgensen et al., 2003]. They examined various feature extraction methods, including STFT coefficients, wavelets, and lpc coefficients and reported a maximum word accuracy for dual tree wavelets (92%) followed by Fourier coefficients (91%). In 2005, the authors extended the original six word vocabulary by the ten English digits and achieved a word accuracy of 73% [Jorgensen and Binsted, 2005].

Manabe et al. investigated conventional ASR techniques for isolated word recognition using three surface electrodes placed on different facial muscles (the *orbicularis oris*, the *zygomaticus major* and the *digastricus*) [Manabe and Z.Zhang, 2004]. Experiments with a *multi-stream* HMM indicated that it is effective to give different *weight* to different MES channels so that the corresponding muscles can contribute to a different extend to the classification. The authors reported a maximum recognition rate of 64% on the ten Japanese digits using delta filterbank coefficients and spectral subtraction.

3.3 Electromyography to Augment Conventional Speech Recognition Systems

Chan et al. proposed to use the myoelectric signal from articulatory muscles of the face as a secondary source of speech information in order to enhance conventional speech recognition systems [Chan et al., 2002a]. An EMG speech expert was combined with an acoustic speech expert. While recognition results of the acoustic expert decreased dramatically with increased noise level the EMG expert remained rather unaffected by noise. The multi-expert system which dynamically tracks the reliability of each expert yielded recognition rates better or near either individual expert at a wide variety of noise levels.

3.4 Vowel and Consonant Recognition

Recently, Jorgensen et al. expanded their earlier isolated word experiments to the recognition of vowels and consonants as a first step towards phoneme based speech recognition [Jorgensen and

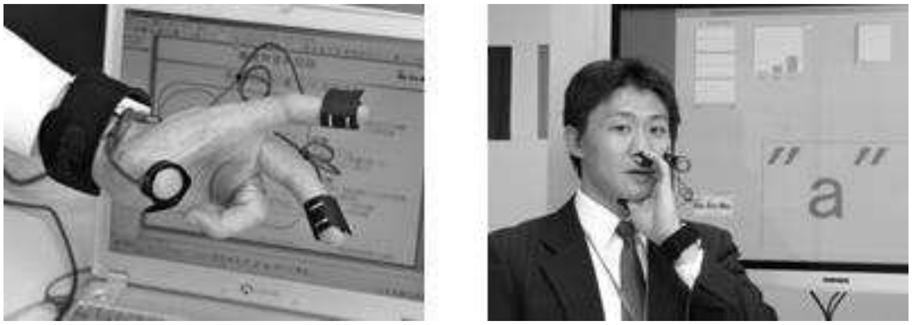


Figure 3.2: User-friendly electrodes wrapped around the fingers and the thumb [Manabe et al., 2003b]

Binsted, 2005]. They achieved recognition rates of 33% on a set of twenty-three consonants and eighteen vowels. Performance could be raised to 50% by excluding six alveolars (t,d,s,z,ch,j), that is, sounds where the tip of he tongue touches alveolar ridge. Furthermore, confusion pairs often only differed in the voicing feature. /d/ and /t/ for example, had a high confusion rate. The authors are currently working on sensor positioning in order to detect these problematic features as well as on the adaptation on context sensitive techniques commonly used in conventional speech recognition.

It is worth mentioning here that significantly better recognition results have been achieved for smaller sets of phonemes. Kumar et al. reported 88% word accuracy on the five English vowels (e,i,o,u,a) using Neural Networks [Kumar et al., 2004]. Manabe et al. achieved recognition rates of 95% for the five Japanese vowels [Manabe et al., 2003b]. According to our experience in conventional speech recognition systems, however, continuous speech recognition requires a larger set of vowels for accurate continuous speech recognition.

3.5 User-friendly systems

The physiological data recording systems available for surface electromyography recordings today are not user-friendly at all. Not only is the equipment extremely voluminous and unhandy but the surface electrodes need to be attached permanently to the skin. In order to adress this issue Manabe et al. proposed the use of ring-shaped electrodes wrapped around the thumb and two fingers for more user-friendly non-audible speech recognition [Manabe et al., 2003b] [Manabe and Z.Zhang, 2004]. In order for the electrodes to detect sEMG signals from facial muscles the fingers need to be pressed against the face in a specified manner as illustrated in figure 3.2. A detailed description of the system can be found in [Manabe et al., 2003a] and [Manabe, 2004]. The authors hope to perfect the system such that it develops to a mobile interface that can be used in both, silent and noisy environments.

Chapter 4

System Overview

In this chapter we give an overview of our EMG speech recognition system. Section 4.1 describes the hardware we deployed while section 4.2 introduces the software we have written and used for this project. The workflows for data collection, system training and recognition are presented in section 4.3.

4.1 Hardware

4.1.1 EMG Equipment

Section 2.2.1 introduced the equipment necessary for sEMG recordings. We used two different physiological data recording systems for data collection which we will refer to as VARIOPORT_II and VARIO-PORT_VIII depending on the number of EMG channels they provide (two and eight channels respectively) [Becker, 2003b].

VARIOPORT_II

The VARIOPORT_II data recording system was used for initial experiments on EMG based speech recognition. It consists of the following components:

Amplifier: The amplifier was originally designed for *EEG* measurements, i.e. for measuring the electrical activity of the brain. We made use of the following five inputs: one ground input (GND1), two reference inputs (REF1, REF2), and two EMG inputs (EMG1, EMG2). Table

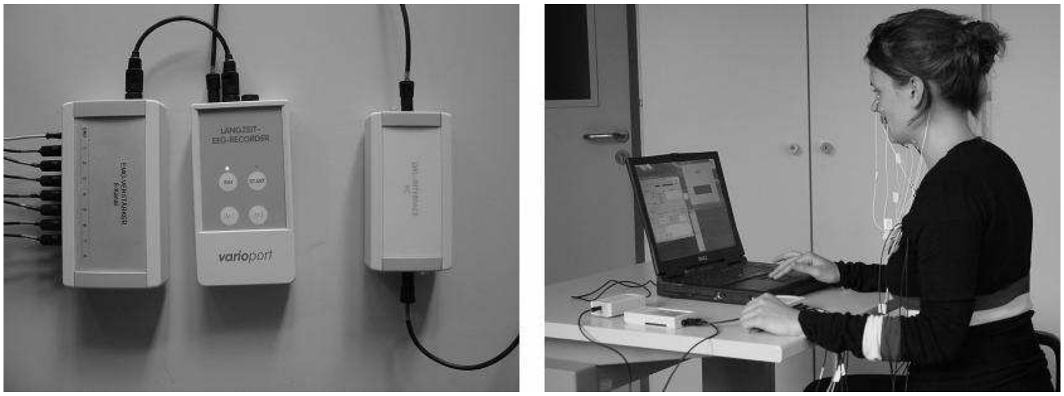


Figure 4.1: Physiological data recording system (Amplifier of the VARIOPORT_VIII system (left), recorder (middle), and electrical isolater (right)) and setup for data recording session

4.1 summarizes the amplification and filter properties for the EMG channels. Contrary to the VARIOPORT_VIII device the low-pass cutoff frequency is at 500Hz. Consequently, minimal sampling rates of 1000Hz are required for this system.

Recorder: The recorder we deployed (figure 4.1) communicates with the PC via the serial port. It has an integrated A/D-converter and can be configured with various sampling rates. When the data is directly transmitted to the PC (and not stored on a memory card) the serial port determines the maximal sampling rate. The laptop we deployed for data recording allows transmission rates of up to 115kBaud, yet, it is recommended not to exceed 80-90% of the total capacity. For two EMG channels and commonly used sampling rates (1000Hz), the serial port is no bottleneck.

Marker Channel: The recorder provides an additional *marker* channel input which can be connected to the LPC port of a PC for synchronizing an acoustic channel with the physiological data channels (when an audio recording is started the values of the marker channel are simply changed from 0 to a value $\neq 0$ by the software.)

Electrical Isolater: The electrical isolater is installed between the recorder and the PC (figure 4.1).

*Synchronization Channe*l A data package transmitted by the recorder over the serial port consists of one sample per selected channel. When the so-called synchronization channel is switched on, it additionally transmits a fixed number after each package which can be used by the software on the PC to detect lost data.

Electrodes: The EMG channels of the VARIOPORT_II system were designed to perform *unipolar* measurements, that is, voltages are recorded between one detection electrode placed on muscle tissue and one or two reference electrodes placed on electrically neutral tissue. When two reference electrodes are deployed, the potential difference between the detection electrode and the *average* potential of the two reference electrodes is amplified. We used reusable Ag/Ag-Cl electrodes which measured 5mm in diameter for the detection and reference electrodes and disposable self-adhesive Ag/Ag-Cl electrodes as ground electrodes (i.e. for the GND1 input) for the VARIOPORT_II system (figure 4.2). The disposable electrodes require the application of electrode gel for a reduction of impedance at the electrode/skin junction.

Name	EMG Channels	Frequency range	Amplification factor	Range	A/D conversion	Resolution
VARIOPORT_II	2	0.9Hz..500Hz	2400	$\pm 500\mu V$	12 bit	$0.25\mu V$ per Bit
VARIOPORT_VIII	8	EMG1-7: 0.9Hz..295Hz EMG8: 19..295Hz	1170	$\pm 1070\mu V$	16 bit	$0.033\mu V$ per Bit

Table 4.1: Properties of data recording systems VARIOPORT_II and VARIOPORT_VIII [Becker, 2003a] [Becker, 2004]

VARIOPORT_VIII

The VARIOPORT_VIII data recording system was used for the core experiments of this work. It consists of the following components (figure 4.1):

Amplifier: The amplifier provides nine inputs: one ground input (GND), and eight EMG inputs (EMG1-8). Table 4.1 summarizes its properties.

Recorder: The recorder of the VARIOPORT_VIII system is the same as for the VARIOPORT_II system (including marker channel and synchronization channel). Due to the fact that the amplifier

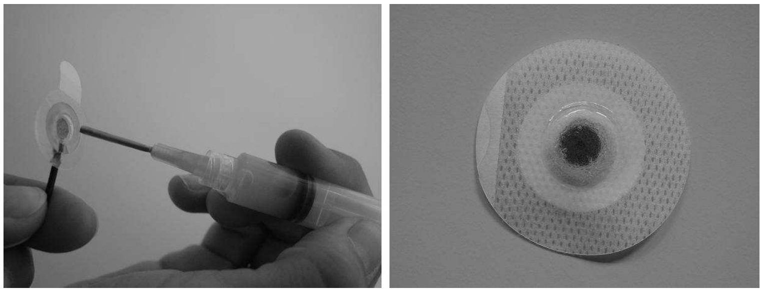


Figure 4.2: Reusable detection electrode filled with gel (left) and disposable ground electrode (right).

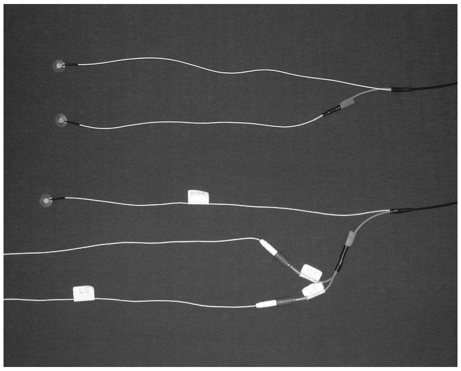


Figure 4.3: Detection electrodes of the VARIOPORT_III system. Bipolar (upper) and unipolar (lower) electrode configuration.

provides 8 EMG channels, however, there is now a tradeoff between choosing a high sampling rate and using as many channels as possible. A sampling rate of 1000Hz, for instance, allows to select a maximum of five channels ($80\% \cdot \frac{115kBaud/s}{16bit \cdot 1000Hz} = 5.75$). When the marker channel and the synchronization channel are both deployed this leaves only three EMG channels that can be used for recognition. Refer to section 6.3.4 for an analysis of this problem.

Electrical Isolater: The electrical isolater is the same as for the VARIOPORT_II system.

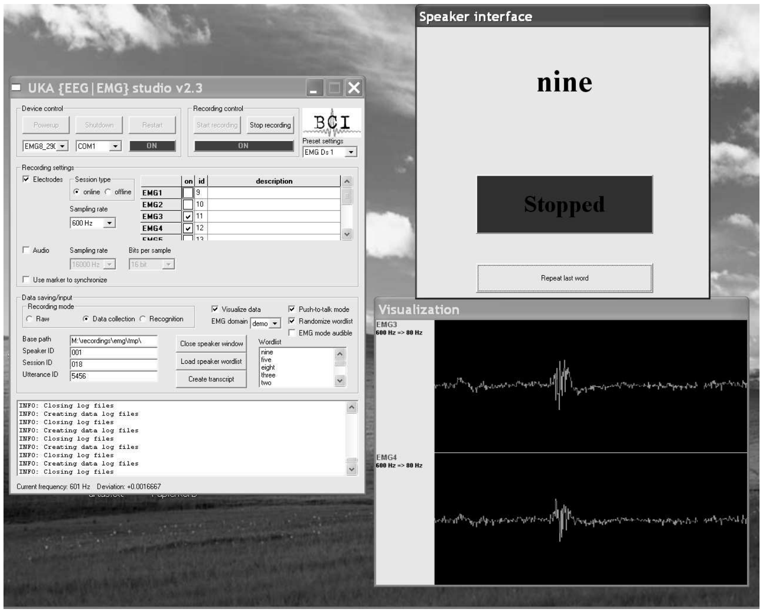
Electrodes: Each EMG channel is associated with two inputs that represent the plus and minus input of the corresponding amplifier. All channels can be measured in one of two ways as illustrated in figure 4.3:

- Bipolar measurements: each of the two inputs of the channel is connected to one surface electrode. The corresponding potentials are subtracted from each other.
- Unipolar measurements: One input channel is connected to a *detection* electrode placed on muscle tissue. The second input is connected to two *reference* electrodes that can for example be placed behind the left and the right ear. These two electrodes provide a special connector which allows the input of arbitrarily many EMG channels.

We used the same electrode types as in the VARIOPORT_II system.

4.1.2 Computers

Data collection and online recognition were generally performed on a Pentium III laptop (1000MHz, 512 MB RAM) with a Microsoft Windows operating system. For offline recognition and training we



EMG8_29(▾ COM1 ▾ ON

ON

Preset settings: EMG Ds 1

Recording settings

☑ Electrodes

Session type: ● online ○ offline

Sampling rate: 600 Hz

	on	id	description
EMG1	□	9	
EMG2	□	10	
EMG3	☑	11	
EMG4	☑	12	
EMG5	□	13	

□ Audio Sampling rate: 16000 Hz Bits per sample: 16 bit

□ Use marker to synchronize

Data saving/input

Recording mode: ○ Raw ● Data collection ○ Recognition

☑ Visualize data ☑ Push-to-talk mode

EMG domain: demo ☑ Randomize wordlist

□ EMG mode audible

Base path: M:\recordings\emg\tmp\
Speaker ID: 001
Session ID: 018
Utterance ID: 5456

Wordlist:
nine
five
eight
three
two

```
INFO: Closing log files
INFO: Creating data log files
INFO: Closing log files
INFO: Creating data log files
INFO: Closing log files
INFO: Creating data log files
INFO: Closing log files
INFO: Creating data log files
INFO: Closing log files
```

Current frequency: 601 Hz Deviation: +0.0016667

Figure 4.4: General user interface consisting of control, visualization, and speaker interface.

deployed different Linux based machines provided by the ITI Waibel of the Universität Karlsruhe. It is worth mentioning here that recognition rates varied slightly for different operating systems.

4.2 Software

The software we designed and implemented for this work consisted of two parts: (1) a Visual C++ project for data collection and demonstration purposes and (2) a JRTk based Tck/Tk script collection for recognizer training and classification. The following two sections describe the functionality of both units.

4.2.1 Visual C++ User Interface

The Visual C++ software was designed by Christoph Mayer, Marcus Warga, Marek Doniec, and Lena Maier-Hein. Most components were implemented by Christoph Mayer and Marcus Warga. For a detailed description of the software refer to [Mayer, 2005]. At this point we merely give a brief overview.

The functionality of the Visual C++ project is accessible via the general user interface which is shown in figure 4.4. It consists of the following components:

Control

The control component determines the general workflow. When the PC is connected to a physiological data recording system, the software can communicate with that system, initiate/stop data recording, and receive data. The control component allows to choose various settings. Refer to appendix A for a brief overview.

Speaker Interface

The speaker interface provides a push-to-talk button that can be pressed/released to initiate/end the storage of signal data. In the data collection mode an additional "repeat" button can delete previously stored utterances. The upper part of the speaker interface presents the word from the wordlist to be spoken next or the hypothesis of the previously recorded utterance depending on the mode the system is in (data collection or recognition).

Visualization

The visualization component visualizes all selected channels. The data can optionally be scaled and downsampled.

4.2.2 JRTk Recognition Engine

All speech recognition experiments presented in this thesis were conducted with the *Janus Recognition Toolkit (JRTk)* which is developed and maintained by the Interactive Systems Labs at the University of Karlsruhe (TH), Germany, and at the Carnegie Mellon University (CMU), Pennsylvania, USA. JRTk provides a flexible Tcl/Tk script based environment which allows rapid development of state-of-the-art speech recognition systems [Finke et al., 1997].
A JRTk based Tcl script collection was produced for this project which allows flexible training and recognition of EMG based speech recognition systems. Various frontends, HMM topologies, normalizations, adaptations and segmentation methods can be selected. A detailed description can be found in [Maier-Hein, 2005].
The following section presents the basic workflow for data collection, training, and recognition.

4.3 Workflow

4.3.1 Data Collection

All signal data used for our experiments was collected in so-called *recording sessions*. A recording session is defined as a set of utterances collected in series by one particular speaker. All settings (channels, sampling rate, speech mode) remain constant during a session. Data collection consists of four steps:

1. Choosing settings: several settings have to be made prior to recording a session. Among other things a word list is selected containing all utterances a speaker has to record during that session. The list can optionally be randomized.

2. Data recording: The speaker records one file set (appendix B) for each word in the word list using the push-to-talk button of the speaker interface.

3. Generation of transcript file: When all utterances have been recorded a transcript is (automatically) created for the session.

4. Generation of settings file: All settings are stored in a file. A sample settings file is given in appendix C.

The session can then be transferred to the server and trained offline. Refer to appendix B or [Mayer, 2005] for a more detailed explanation on how to produce a recording session.

4.3.2 Training

Training is performed offline using the scripts collection introduced in section 4.2.2. It consistst of the following three steps:

1. Choosing settings. The following settings have to determined prior to training a system
   - one speaker and a set of n training recording sessions is selected.
   - the vocabulary and the desired number of training samples per speech unit is chosen

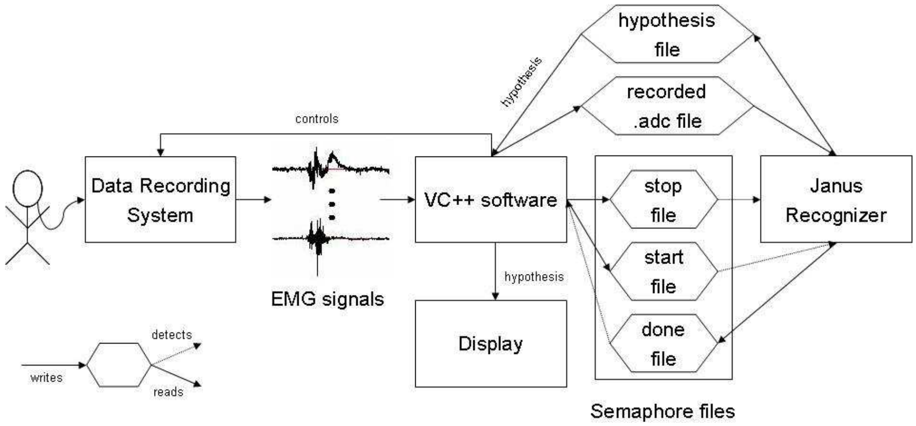


Figure 4.5: Principle of online recognition: The semaphore files (start, done, stop) initiate certain actions. The data files (hypothesis file and .adc file) contain the actual data.

- Frontend, HMM topologies and segmentation methods are selected. Normalizations/adaptations can optionally be used.

2. Model Initialization: Models are initialized as described in section 6.1.

3. EM-Training is performed.

4.3.3 Recognition

Recognition can either be performed online or offline.

Online recognition

Online testing was implemented for isolated words and phrase recognition only. First, a training session and a set of possible hypothesis is selected (subset of the set of training utterances). Next, the recorder and a janus recognition script are started by the VC++ software. Janus and the software communicate via semaphore files. When an utterance has been recorded (i.e. the push-to-talk "recording" button is released) a "start" file is created that is detected by janus. Janus deletes the file and reads the .adc file corresponding to the recorded signal data. It determines the Viterbi path for each hypothesis allowed in the vocabulary and the word/phrase yielding the best Viterbi score is written to a hypothesis file. A "done" file is then created by Janus which is detected by the VC++ software. The VC++ software reads the hypothesis and displays it on the speaker interface. When Janus detects a "stop" file (created by the VC++ software) it leaves the recognition loop. Figure 4.5 illustrates the communication between the different modules.

Offline testing

Contrary to online testing offline testing does not involve the VC++ software and is thus primarily conducted on Linux machines. The user can optionally

- perform training and testing on individual sessions and average the results over all sessions (within-session testing).

- perform training on $n-1$ sessions and test on the remaining session. This is performed n times so that each session is tested on once (leave-one-out).

- perform training on one session and test on another session. This is performed $n \cdot (n-1)$ for each tuple $(i, j); i! = j$.

- perform training and testing on the set of *all* sessions (i.e. each session occurs in training and testing).

In each case the round robin algorithm is applied such that a maximal number of test data is obtained: For within-session testing the complete session is split into a number of rN sets of equal size (each containing the same number of utterances per word). For each combination of $rN - 1$ sets a recognizer is trained and then tested on the remaining set. This way every single utterance is used for testing exactly once.

For across-sessions testing the training data is split into a disjoint set of training sets each containing the same number of samples. The test session is then tested on each of the corresponding recognizers. When *isolated* word recognition is performed all training sets contain the same number of utterances per word in the vocabulary.

Further details can be found in [Maier-Hein, 2005].

Chapter 5

Corpora

This chapter introduces the corpora we deal with in this work. An overview of all corpora, vocabulary domains and speakers is introduced in the following section 5.1. Sections 5.2 to section 5.6 describe each corpus in detail.

5.1 Overview

The term *corpus* refers to a collection of transcribed recorded speech data. We produced ten corpora for this work. Their distinguishing characteristics are (1) the vocabulary domain, (2) the physiological data recording system used for the recordings, and (3) their functionality.

Vocabulary domains

Name	**Words/Phrases in the Vocabulary**
phone	yes no accept deny wait
digits	zero, one, two, three, four, five, six, seven, eight, nine, [call], [end]
commands	stop, go, left, right, alpha, omega
meeting	hello, i'm in a meeting, is it urgent, i'll call back later, talk to you later, bye, hang on, [hi mom], [hi veronica], [ok]
lecture	good afternoon ladies and gentlemen, welcome to the interact center, my name is stan jou, let me introduce our new prototype, any questions, thank you for your attention, [my name is alex waibel], [good morning ladies and gentlemen], [my name is lena maier-hein], [thank you very much for your attention], [thank you very much] [thanks a lot]

Table 5.1: Vocabulary domains and corresponding words/phrases. When an utterance appears in brackets [] it is optional for the vocabulary and was not necessarily recorded in all sessions corresponding to the domain.

The following vocabulary domains were used in this work:

phone: the phone domain consists of five generic words suitable for accepting or denying a phone call. It was used for initial experiments on EMG based speech recognition.

digits: the digits domain consists of the ten English digits and the words "call" and "end". It was chosen as the standard isolated word vocabulary because it allows comparison of recognition rates to results previously reported in literature. The additional words call and end allow the recording of sequences of digits in the form "call [digit] ... [digit] end".

meeting: The meeting domain consists of a set of sentences typically used for answering a phone call during a meeting, for instance "I'm in a meeting", "is it urgent?" and "I'll call back later". It was deployed for the "silent mobile phone" demo introduced in section 7.1.

lecture: The lecture domain consists of a set of sentences typically used by somebody who is giving a talk, e.g. "good morning ladies and gentlemen", "my name is ...", "any questions?", "thank you for your attention". It was deployed for the lecture translation system demo introduced in section 7.2.

commands: The command set consists of a set of six commands that can for example be used to control a robot. It was developed by Jorgensen et al. for initial experiments on EMG based speech recognition [Jorgensen et al., 2003]. We used these words for experiments on connected words recognition.

The words and sentences corresponding to the different domains are shown in table 5.1.

Speakers

Seven speakers participated in recording the corpora for this work. None of them was a native English speaker, however their English was sufficient for our purposes. Table 5.2 presents sex, mother-tongue, age and command of English for each speaker.

Speaker ID	**mother-tongue**	**English**	**sex**	**age**
S0	German	fluent	male	23
S1	German	fluent	female	25
S2	German	fluent	male	26
S3	Hungarian	basic	male	27
S4	German	fluent	male	49
S5	Taiwanese	fluent	male	34
S7	German	fluent	male	25

Table 5.2: Speakers (ID S6 is missing because speaker S7 insisted on the ID S7 = 007)

Producing a corpus

The workflow for data collection was already introduced in section 4.3.1 and in appendix B. In each recording session a word list was selected with all utterances the speaker was to record in the session. The order of the words could be chosen to be either the same as in the loaded word list or to be a random permutation of the given list. In each recording session the words from the list were presented to the subject one at a time. A push-to-talk button controlled by the subject was used to mark the beginning and the end of each utterance. Subjects were asked to begin speaking approximately 1sec after pressing the button and to release the button about 1sec after finishing the utterance. When connected words were recorded they were asked to leave no silence between the words. When the pseudo-word silence appeared they were supposed keep all facial muscles relaxed for approximately 2sec.

EMG signal data was collected using one of the two data recording systems introduced in section 4.1 and the corresponding electrodes. The ground electrode was always placed on the left wrist and the reference electrodes for unipolar measurements were positioned behind the left and right ear. The positions of the detection electrodes varied.

In several sessions, the incoming data was visualized so that *biofeedback* was given to the speakers. The term biofeedback refers "the process of measuring and quantifying an aspect of a subject's physiology, analyzing the data, and then feeding back the information to the subject in a form that allows the subject to enact physiological change." [Wikipedia, 2005]. It has been shown that biofeedback can improve the performance of so-called *Brain Computer Interfaces*, which analyse the activity of the

Corpus Name	**Recording Device**	**Speakers (# sessions)**	**Function**
PHONE_INIT	VARIOPORT_II (1010Hz)	S0(4), S1(5), S2(2), S7(2)	Initial experiments on EMG based speech recognition.
PHONE	VARIOPORT_II (1010Hz)	S0(6), S1(5), S3(8)	Development of baseline system. Experiments on segmentation and on session independence.
DIGITS_II	VARIOPORT_II (1010Hz)	S3(16)	Experiments on session independence, non-audible speech recognition and connected digits recognition.
DIGITS_VIII_INIT	VARIOPORT_VIII (600Hz)	S3(11), S7(2)	Initial experiments with the VARIOPORT_VIII recording device and initial experiments on electrode positioning.
DIGITS_VIII	VARIOPORT_VIII (600Hz)	S1(5), S3(6), S7(8)	Optimization of baseline system. Final experiments on session independence and on the comparison of audible and non-audible speech.
DIGITS_VIII_POS	VARIOPORT_VIII (600Hz)	S1(4), S3(10)	Experiments on electrode positioning.
DIGITS_VIII_CON	VARIOPORT_VIII (600Hz)	S3(1), S5(1)	Final experiments on connected digits recognition.
COMMANDS	VARIOPORT_VIII (600Hz)	S3(1)	Experiments on connected words recognition.
MEETING	VARIOPORT_VIII (600Hz)	S0(3), S3(6) S4(2), S5(4)	sessions for practicing and presenting meeting demo.
LECTURE	VARIOPORT_VIII (600Hz)	S5(3)	sessions for practicing and presenting lecture demo.

Table 5.3: Overview of all corpora.

brain for classification tasks [Lehtonen, 2002]. We provided biofeedback to the speakers in all sessions corresponding to the VARIOPORT_VIII system.

Corpus structure

An overview of all corpora produced for this work is given in table 5.3. Each corpus is associated with a number of *test sets*. A test set corresponding to a given corpus consists of a set of sessions from that corpus that were used for a particular experiment discussed in chapter 6. For example, all non-corrupt sessions of the PHONE corpus make up Set I of the PHONE corpus while all non-corrupt *non-audible* sessions of the DIGITS_VIII corpus make up Set I of the DIGITS_VIII corpus. It is worth mentioning here that various reasons exist for a session not being part of any set:

- Hardware Problems: We encountered a variety of hardware problems. Several sessions, for example, were interrupted because the recording device crashed too often. In other sessions, too much data got lost. It should be pointed out that these problems have been solved.

- Synchronization problems: In certain sessions the acoustic channel and the EMG channels were badly synchronized because of system delays and buffer sizes. Some of these sessions had to be repeated in order to ensure a reliable evaluation of audio based segmentation methods. The synchronization problems have now been solved.

- Electrode problems: Several sessions contain bad quality data for certain channels which was either due to a bad electrode contact or hardware problems. These changes were sometimes not

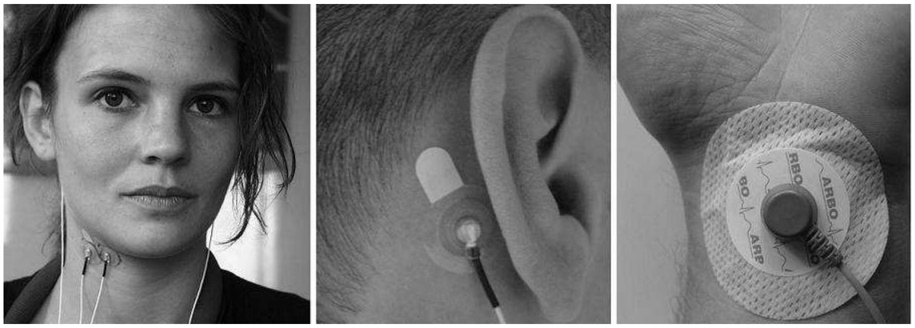


Figure 5.1: Positioning of the detection electrodes (P1: inner electrode, P2: outer electrode) and of the reference and ground electrodes for the PHONE corpus.

detected during the recording of the session because the visualization component was still being developed and had a relatively bad resolution.

- Hardware delivery delay: Due to the fact that the delivery of the VARIOPORT_VIII was delayed we recorded a large set of sessions with the VARIOPORT_II device which provides only two EMG channels. These sessions had to be repeated with the new eight-channel device so that better recognition results could be obtained.

Appendix D enumerates all sessions that were recorded for this work. The following sections 5.2 to 5.6 introduce each corpus and the corresponding test sets.

5.2 Corpora on the phone domain

We produced two corpora on the phone domain: one for initial experiments on EMG based speech recognition (PHONE_INIT) and one for the development of a baseline speech recognition system as well as for experiments on segmentation and session dependence (PHONE).

5.2.1 PHONE_INIT corpus

Four speakers, S0, S1, S2 and S7 participated in the recording sessions for the PHONE_INIT corpus. Various positions and numbers of utterances per word were chosen because the corpus merely served for experimental purposes.

5.2.2 PHONE corpus

The PHONE corpus consists of those sessions associated with the phone domain that were recorded for particular experiments discussed in chapter 6. Two speakers, S0 and S1, participated in these sessions. They recorded six and five audibly spoken sessions respectively with thirty to sixty repetitions per word including silence. The words were presented to the subjects in blocks of ten utterances of the same kind. Due to the fact that experiments on segmentation methods were performed on this data, the speakers were asked to record a relatively large amount of “silence” of 1-2sec before and after speaking a word from the list.
EMG signal data was collected using the two-channel device VARIOPORT_II introduced in section 4.1.1 and two Ag/Ag-Cl detection electrodes. Due to the fact that Jorgensen et al. reported encouraging results using electrodes placed on the left and right side of the larynx [Jorgensen et al., 2003] we selected similar positions for our two electrodes. We will refer to them as positions P1 and P2. However, instead of choosing a classical bipolar configuration with a 2cm inter-electrode spacing we referenced each electrode to both ears which is the standard for EEG measurements as already

Set	Sessions	Electrode Positions	Utterances per Word	audible	Comment
Set 1	S0: 004, 007, 008, 009 S1: 005, 006, 008, 009	EMG1: P1 EMG2: P2	30-60	yes	Set was used for developing a baseline system and for experiments on segmentation. Tape measure was used for position identification. A lot of "silence" in signals. No visualization.
Set 2	S0: 004, 006, 007	EMG1: P1 EMG2: P2	30-60	yes	Set was used to examine the effect of different factors on session dependency. Tape measure was used for position identification. A lot of "silence" in signals. No visualization.

Table 5.4: Test sets of the PHONE corpus.

explained in section 4.1.1. Figure 5.1 shows the exact positions for speaker S1. We used tape measure to identify the same positions in consecutive sessions.

Table 5.4 presents the test sets used from the PHONE corpus for the experiments in chapter 6.

5.3 Corpora on the digits Domain

Four corpora were produced for the digits domain: the DIGITS_II corpus for speech recognition experiments using the VARIOPORT_II device, the DIGITS_VIII_INIT corpus for initial experiments with the new recording device VARIOPORT_VIII, the DIGITS_VIII_POS for experiments on electrode posisioning, the DIGITS_VIII_CON corpus for connected digits recognition, and the DIGITS_VIII corpus for optimizing the baseline system, for comparing non-audible and audible speech and for performing experiments on session dependence and connected digits recognition.

5.3.1 DIGITS_II corpus

The DIGITS_II corpus consists of all sessions recorded on the digits domain with the recording device VARIOPORT_II. Speaker S3 recorded a total of seven audible sessions for isolated digits recognition, five non-audible sessions for isolated digits recognition, and four sessions for connected digits recognition. The isolated digit sessions contain fifty exemplars for each word including silence. The words were presented to speaker S1 in blocks of ten utterances of the same kind.

The same positions as for the PHONE corpus were used because serious experiments on electrode placement were planned to be conducted with the VARIOPORT_VIII device which was not yet available. A Gypsum mask produced for speaker S3 was deployed to identify the same positions in consecutive sessions. Table 5.5 presents the test sets of the DIGITS_II corpus that are referred to in chapter 6.

Set	**Sessions**	**Electrode Positions**	**Utterances per Word**	**audible**	**Comment**
Set I	S3: 018, 020 021	EMG1: P1 EMG2: P2	50	yes	Set was used for initial experiments on digits recognition and session independence. Gypsum masks were used for position identification.
Set II	S3: 010, 013, 014, 017	EMG1: P1 EMG2: P2	50	no	Set was used for initial experiments on non-audible speech recognition and on session independence. Gypsum masks were used for position identification.
Set III	S3: 018, 019	EMG1: P1 EMG2: P2	50	no	Sessions have exact same electrode placement. Isolated digits were recorded in session 018; utterances of the form “call [digit] end” were recorded in session 019.

Table 5.5: Test sets of the DIGITS_II corpus.

5.3.2 DIGITS_VIII_INIT corpus

Two speakers, S0, and S3 participated in the recording sessions for the DIGITS_III_INIT corpus. The sessions were recorded to test the new system VARIOPORT_VIII and to perform initial experiments on electrode placement.

5.3.3 DIGITS_VIII_POS corpus

The DIGITS_VIII_POS corpus consists of those sessions associated with the digits domain recorded with the VARIOPORT_VIII device that were used for systematic experiments on electrode positioning. Two speakers, S1 and S3, participated in these sessions. They recorded four and ten sessions respectively with thirty repetitions per word including silence. Refer to section 6.2 for a detailed description on the electrode positioning. Table 5.6 presents the test set of the DIGITS_VIII_POS corpus that is referred to in chapter 6.

5.3.4 DIGITS_VIII corpus

The DIGITS_VIII corpus consists of those sessions associated with the digits domain that were used to optimize the baseline system, to perform experiments on session dependence and to compare audible and non-audible speech. Three speakers, S0, S1 and S3, participated in these sessions. Each subject took part in audible and non-audible recording sessions on different days, in morning and afternoon sessions.

Set	Sessions	Electrode Positions	Utterances per Word	audible	Comment
Set I	S1: 010, 011, 012, 013 S3: 031, 032, 033, 034, 035, 036, 037, 038, 049, 050	refer to section 6.2	50	both	Set was used for optimizing electrode positioning. Gypsum masks were used for position identification.

Table 5.6: Test sets of the DIGITS_VIII_POS corpus.

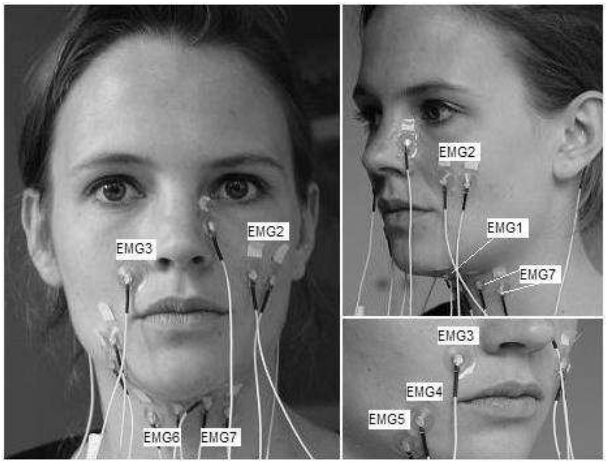


Figure 5.2: Positioning of electrodes EMG1-EMG7 for the DIGITS_VIII corpus.

In each recording session forty exemplars of each vocabulary word and forty exemplars of silence were recorded. The order of the words was randomly permuted.
EMG signal data was collected using the eight-channel device VARIOPORT_VIII introduced in section 4.1.1 and seven pairs of Ag/Ag-Cl electrodes. As shown in Figure 5.2 the electrodes were positioned such that they obtain the EMG signal of six articular muscles: the *levator anguli oris* (EMG2,3), the *zygomaticus major* (EMG2,3), the *platysma* (EMG4,5) the *depressor anguli oris* (EMG5), the *anterior belly* of the *digastric* (EMG1) and the *tongue* (EMG1,6,7) [Laboratory, 2002] [Chan et al., 2002b]. For three of the seven EMG channels (EMG2,6,7) a classical bipolar electrode configuration with a 2cm center-to-center inter-electrode spacing was used. For the remaining four channels one of the detection electrodes was placed directly on the articulatory muscles and was referenced to either the nose (EMG1) or to both ears (EMG3,4,5) (Figure 5.2). The gypsum masks we had produced for each speaker were used for position identification. The positioning of the electrodes was optimized in the experiments described in section 6.2.1.
Table 5.7 presents the test sets of the DIGITS_VIII corpus that are referred to in chapter 6.

5.3.5 DIGITS_VIII_CON corpus

The DIGITS_VIII_CON corpus consists of those sessions associated with the digits domain that were used for connected digits experiments. Two speakers, S3 and S5 recorded one session each.
In each session utterances of the form "*silence* [6 * [digit]] *silence*" were recorded such that the complete set of utterances could be divided into two disjoint sets each containing each digit *tripel* at least once. Thus each digit was contained a least a hundred times ($|V|$(left context) $\cdot|V|$(right context) $= 10 \cdot 10 = 100$; $|V|$: number of words in the vocabulary) in each set and at least once in any possible

Set	Sessions	Electrode Positions	Utterances per Word	audible	Comment
Set I	S1: 014, 015, 018, 019 S3: 051, 053, 055, 058 S7: 004, 008, 010, 011	EMG1: P22-P23 EMG2: P35-P38 EMG3: P17 EMG4: P42 EMG5: P28 EMG6: P46-P47 EMG7: P51-P52	40	no	Set was used for optimizing the baseline system and for experiments on session independence. Gypsum masks were used for position identification.
Set II	S1: 015, 016 S3: 051, 052 S7: 008, 009	EMG1: P22-P23 EMG2: P35-P38 EMG3: P17 EMG4: P42 EMG5: P28 EMG6: P46-P47 EMG7: P51-P52	40	no	The two sessions for each speaker were recorded in series without removing and re-applying the electrodes. Set was used for comparing audible and non-audible speech. Gypsum masks were used for position identification.
Set III	S1: 016 S3: 052 S7: 009	EMG1: P22-P23 EMG2: P35-P38 EMG3: P17 EMG4: P42 EMG5: P28 EMG6: P46-P47 EMG7: P51-P52	40	no	Set was used for comparing EMG based isolated digit recognition to audio based isolated digit recognition. Gypsum masks were used for position identification.

Table 5.7: Test sets of the DIGITS_VIII corpus.

left/right-context. The speakers were asked to speak continuously, that is, without a break between the individual digits.

Electrode positioning was the same as in the DIGITS_VIII corpus.

Table 5.8 presents the test set of the DIGITS_VIII_CON corpus that is referred to in chapter 6.

5.4 MEETING corpus

The MEETING corpus consists of those sessions that were recorded for practicing and presenting the silent phone demo introduced in section 7.1. The electrode positions were not the same in all sessions because these sessions were recorded at the same time as the sessions for optimizing electrode positions (the optimal position set was thus still being developed at that time). Our set of optimal positions was only used in the last recording sessions of this corpus. Several sessions were trained in a dialog to match training and testing conditions. Details can be found in appendix D.

Table 5.9 presents the test sets of the MEETING corpus that are referred to in chapter 6.

Set	Sessions	Electrode Positions	Utterances per Word	audible	Comment
Set I	S3: 056 S5: 006	EMG1: P22-P23 EMG2: P35-P38 EMG3: P17 EMG4: P42 EMG5: P28 EMG6: P46-P47 EMG7: P51-P52	40	no	Set was used for connected digits experiments. Gypsum masks were used for position identification.

Table 5.8: Test sets of the DIGITS_VIII_CON corpus.

Set	Sessions	Electrode Positions	Utterances per Word	audible	Comment
Set I	S5: 000, 004	EMG1: P22-P23 EMG2: P35-P38 EMG3: P17 EMG4: P42 EMG5: P28 EMG6: P46-P47 EMG7: P51-P52	40	no	Set was used for practicing the silent phone demo and for comparing filter settings. Tape measure was used for position identification.

Table 5.9: Test sets of the MEETING corpus

5.5 LECTURE corpus

The LECTURE corpus consists of those sessions that were recorded for practicing the lecture translation demo introduced in section 7.2. Electrode positions were the same as for the DIGITS_VIII corpus. The number of utterances per sentence varied. Details can be found in appendix D.
Table 5.10 presents the test set of the LECTURE corpus that chapter 6 refers to.

Set	Sessions	Electrode Positions	Utterances per Word	audible	Comment
Set I	S5: 001, 002, 003	EMG1: P22-P23 EMG2: P35-P38 EMG3: P17 EMG4: P42 EMG5: P28 EMG6: P46-P47 EMG7: P51-P52	40	no	Set was used for practicing the lecture translation demo. Tape measure was used for position identification.

Table 5.10: Test sets of the LECTURE corpus

5.6 COMMANDS corpus

The COMMANDS corpus consists of only one session. It was recorded for experiments on connected words recognition using the same electrode positions as were used for the DIGITS_VIII corpus. Utterances of the form "[command] [command] [command]" were recorded with no silence in between individual words. Each triple was recorded twice.

Set	**Sessions**	**Electrode Positions**	**Utterances per Word**	**audible**	**Comment**
Set I	S3: 054	EMG1: P22-P23 EMG2: P35-P38 EMG3: P17 EMG4: P42 EMG5: P28 EMG6: P46-P47 EMG7: P51-P52	40	no	Set was used for experiments on connected words recognition. Gypsum mask was used for position identification.

Table 5.11: Test sets of the COMMANDS corpus

Chapter 6

Experiments and Results

In this chapter we present the experiments and results we conducted on EMG based speech recognition. The first part of the chapter describes the development of a state-of-the-art system. Section 6.1 introduces our baseline system. Section 6.2 describes methods and experiments for identifying appropriate electrode positions and for repeating electrode placement based on this system. Experiments for optimizing the baseline system are presented in sections 6.3 and 6.4. The segmentation methods we examined are introduced in section 6.5.

In a second part of this chapter, we deal with issues in EMG based speech recognition that have not yet been addressed in the literature, namely with session independence (section 6.6), the comparison of audible and non-audible speech (section 6.7) and initial experiments on EMG based continuous speech recognition (6.8).

Finally, section 6.9 compares the performance of EMG based speech recognizers to the performance of conventional speech recognition systems.

6.1 Isolated Word Recognition - Baseline System

In this section we present the baseline system used for our isolated word recognition experiments. The system was developed based on various experiments on the PHONE_INIT and the PHONE corpus using two surface electrodes. Those experiments were conducted on different speakers, different sessions, and different numbers of training utterances and are thus not comparable. Despite this, we give a short summary of the system's development in the following section 6.1.1.

Sections 6.1.2 and 6.1.3 describe the system in more detail while section 6.1.4 presents the baseline results on the PHONE corpus and the DIGITS_VIII corpus.

Note that the baseline system serves as a reference system for all experiments we conducted on EMG based speech recognition. Among other things it was used to identify appropriate electrode positions for our eight-channel data recording system VARIOPORT_VIII. A systematic approach for *optimizing* the baseline system is presented in sections 6.3 and 6.4. However, we will show that parameter optimization does not lead to significant improvements.

6.1.1 Development of the Baseline System

Initial experiments on EMG based speech recognition were conducted on the PHONE corpus using two surface electrodes, 32ms observation windows with 8ms shift and Short Time Fourier Coefficients as features. One (context-independent) Hidden Markov Model with 8 gaussians per state was trained for each word in the vocabulary. The number of states of a Hidden Markov Model was equal to the number of phonemes in the corresponding word. We achieved word accuracies of approximately 80% after performing ten iterations of the Expectation-Maximization algorithm.

By optimizing HMM topologies, segmentation, window shift, the number of gaussians per state and the number of EM iterations we increased recognition rates by approximately 10% absolute to 90%. From the fact that mean subtraction led to a significant decrease in performance we deduced that the

Oth Fourier Coefficient - and therewith the mean of the time domain values - is of particular relevance for recognition. Consequently, we introduced an extra feature *mean* (containing the mean of the time domain values of each window) which led to further improvements (about 3% absolute). The use of delta coefficients instead of ordinary STFT coefficients raised recognition rates by approximately 2% absolute to about 95%. It should be pointed out that the use of cepstral coefficients, LPC coefficients, autoregressive coefficients, the zero crossing rate, the RMS value and the JRTk adjacent feature did not improve the system. Filterbank coefficients did not increase recognition rates either which is probably traceable back to the fact that the number of STFT coefficients was relatively small anyway (17 coefficients) because of the low sampling rate (1010Hz).

In summary, optimizations on our initial system yielded an improvement of approximately 15% absolute. The following paragraphs give a more detailed description of the obtained baseline system.

6.1.2 Feature Extraction

Training and Classification are performed on *feature vectors* using Hidden Markov Models (HMMs). For each channel, 18-dimensional channel feature vectors are extracted. We used a sampling rate of 1010Hz and a corresponding observation window size of 32ms for the two-channel device VARIOPORT_II. The window size for the eight-channel device VARIOPORT_VIII which was always configured with a 600Hz sampling rate was chosen to be 54ms so that the number of feature vector coefficients was the same as for the two channel device. Both systems used a window overlap of 4ms.

In order to obtain channel feature vector o_{ij} for channel j and observation window i the windowed Short Time Fourier Transform (STFT) is computed. Delta coefficients serve as the first 17 coefficients of o_{ij}. The 18th coefficient consists of the mean of the time domain values in the given observation window. The complete feature vector o_i for the observation window i is simply the concatenation of the channel feature vectors o_{ij}.

6.1.3 Training and Classification

A five-state left-to-right Hidden Markov Model λ_j with 12 Gaussians per state is trained for every word W_j in the vocabulary using the Expectation Maximization (EM) algorithm. Silence is modelled by a one-state HMM.

Training consists of the following steps:

1. For each utterance in the training data set (labelled with "*silence [word] silence*") a Hidden Markov Model is built. The signal data is transformed into feature vectors and the obtained sequence of feature vectors is split into k sets of equal size where k is the number of states in the HMM. The ith set is then assigned to the ith state in the HMM. It is worth mentioning here that each state of the complete HMM represents either one state in the five-state model of [word] or the (one-state) silence model.

2. The k-means algorithm is applied to every state of each word model using the previously assigned feature vectors. This way, an initial set of twelve codebooks is obtained for each state. We restricted the shape of gaussians by choosing diagonal covariance matrices.

3. EM Training with four iterations is performed to optimize codebooks and distributions.

To recognize an unknown signal the corresponding sequence of feature vectors (o_k) is computed. Next, the Viterbi alignment for each vocabulary word W_j is determined and the word corresponding to the best Viterbi score is output as the hypothesis. Feature extraction, HMM training, and signal recognition are performed using the *Janus Recognition Toolkit (JRTk)* which is introduced in section 4.2.2.

6.1.4 Recognition Results

Table 6.1 and table 6.2 summarize the baseline *within-session* results for isolated word recognition on the phone domain and on the digits domain respectively. The term *within-session* refers to a matching training/test condition, i.e. training and testing are performed on the same session. We always applied the round robin algorithm for within-session testing which works as follows: the complete session is split into a number of rN sets of equal size (each containing the same number of utterances per word). For each combination of $rN - 1$ sets a recognizer is trained and then tested on the remaining set. This way every single utterance is used for testing exactly once.

Baseline experiments for the phone domain were conducted on Set I of the PHONE corpus (section 5.2.2). We used 25 training samples per word, $rN = 6$ round robin sets and the segmentation method *Audio Speech Alignment* (section 6.5). The results are averaged over the four sessions from each speaker (S0 and S1). It should be pointed out that the number of *test* samples varied from session to session because different numbers of utterances were recorded (section 5.2.2). Moreover, several sessions contained a small number of corrupt IDs. The exact number of test samples for both speakers in Set I of the PHONE corpus is shown in table 6.1.

Baseline experiments for the digits domain were conducted on Set I of the DIGITS_VIII corpus (section 5.3.4). We used $rN = 4$ round robin sets which yielded $uN = \frac{rW \cdot |V|}{rN} = \frac{40 \cdot 10}{4} = 100$ utterances per round robin set ($|V|$: vocabulary size excluding silence; rW: repetitions per word) and thus $\frac{(rN-1) \cdot uN}{|V|} = 30$ training samples per word for each recognizer. The results for each speaker were averaged over the corresponding $sN = 4$ sessions which yielded a total of $sN \cdot rN \cdot uN = 4 \cdot 4 \cdot 100 = 1600$ test samples per speaker. No segmentation method was applied but pure silence was used in addition to the vocabulary words for system training. Refer to section 6.5 for an introduction into different segmentation methods.

Word	**S0**	**S1**	**S0 & S1**
Yes	97.4	90.2	94.1
No	**98.4**	**73.5**	86.9
Accept	**98.4**	**97.6**	98.0
Deny	98.9	97.0	92.1
Wait	**85.7**	**99.4**	92.1
Word accuracy	95.8	91.6	93.8
Number of test words	948	818	1766

Table 6.1: Within-session word accuracies (in %) for Set I of the PHONE corpus (four audible sessions per speaker). The segmentation method was *Audio Speech Alignment* (section 6.5).

Word	**S1**	**S3**	**S7**	**S1 & S3 & S7**
One	**100.0**	**99.4**	**100**	99.8
Two	**93.8**	100.0	**100.0**	97.9
Three	96.9	99.4	96.3	97.5
Four	98.8	99.4	96.3	98.1
Five	99.4	98.8	94.4	97.5
Six	91.3	98.8	93.1	94.4
Seven	**100.0**	98.1	**94.4**	97.5
Eight	96.3	98.8	98.1	97.7
Nine	96.3	98.1	91.1	95.4
Zero	99.4	99.4	95.6	98.1
Word Accuracy	97.2	98.8	96.0	97.3
Number of test words	1600	1600	1600	4800

Table 6.2: Within-session word accuracies (in %) for Set I of the DIGITS_VIII corpus (four non-audible sessions per speaker). No segmentation method was applied.

The following observations deserve mentioning: Firstly, recognition rates differ significantly across speakers. In the case of the PHONE corpus this may be traceable back to slightly different electrode

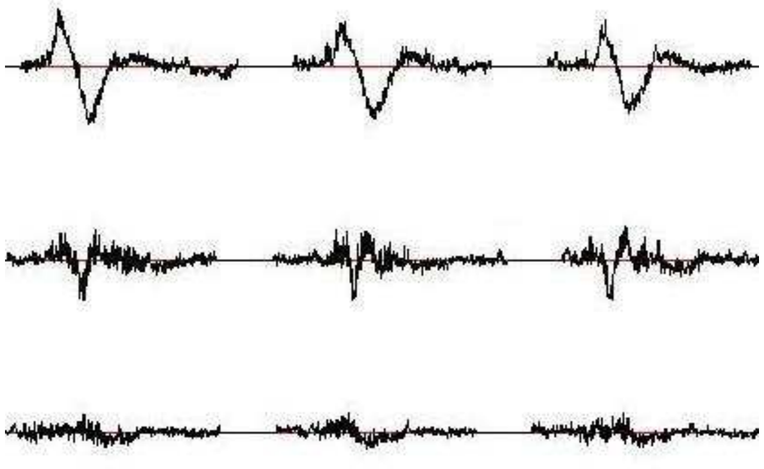


Figure 6.1: EMG signals for the words no (first), accept (middle) and yes (last) and speaker S0 (EMG1, session 008).

positioning whereas we attribute the variation to different degrees of experience in the case of the DIGITS_VIII corpus. Secondly, it is noticeable that several words are particularly well distinguishable for all speakers (e.g. accept, one). Others, however, show extremely high values for one speaker, and rather low values for another speaker (e.g. wait, no, two, seven). We deduce from this, that speaker dependence is an important issue in EMG based speech recognition.

Figures 6.1 and 6.2 show EMG signals for the words no, accept and yes for speakers S0 and S1 respectively. It is noticeable that the signals for the individual words look very different. Interestingly, the heartbeat can be seen in the signals from speaker S1 (regularly occurring peaks) but not in the signals from speaker S0.

In the following section we present our approach for finding a set of appropriate electrode positions using the baseline for comparison of recognition results for different positions. *All* settings (feature extraction, classification, number of training samples, number of round robin sets etc) described in this section will be used throughout this chapter unless explicitly stated otherwise.

6.2 Optimization of Electrode Positioning

As already mentioned in section 2.2.2, exact electrode positioning is an important issue for sEMG measurements. The main problems are to identify ideal electrode positions for individual muscles and to ensure repeatability of placements. When applying EMG for speech recognition another challenge is to choose appropriate muscles for recognition.

We used the eight-channel device VARIOPORT_VIII and the baseline system introduced in the previous section for optimizing electrode placement. Section 6.2.1 explains the methods we applied for determining appropriate electrode positions for EMG based speech recognition while different approaches for repeating placements are introduced in section 6.2.2.

6.2.1 Selecting Electrode Positions

In order to determine an appropriate set of electrode positions for EMG speech recognition with the eight-channel recording device VARIOPORT_VIII we considered the following problems:

1. Which muscles should be examined?

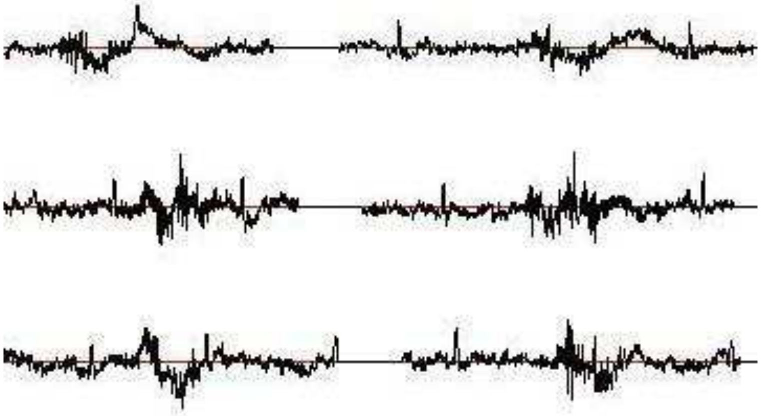


Figure 6.2: EMG signals for the words no (first), accept (middle) and yes (last) and speaker S1 (EMG1, session 008)

- Which muscles are involved in speech production?
- Which muscles were chosen by other research groups?

2. What is the ideal position for measuring the activity of a particular muscle?
   - What is the ideal position according to EMG literature?
   - Which position yields the best signals (e.g.with little movement artifacts)?
   - Which position yields the best recognition results?
3. What is the best combination of electrode positions for EMG based speech recognition?
   - Which positions provide orthogonal information?
   - Which positions are the most practical ones?

Muscles

In order to solve the first problem we simply extracted appropriate information from literature. The following muscles were used for EMG based speech recognition by [Chan et al., 2002b], [Jorgensen et al., 2003] , [Manabe et al., 2003b]: the *levator anguli oris*, the *zygomaticus major*, the *platysma*, the *depressor anguli oris*, the *anterior belly* of the *digastric*, the *orbicularis oris* and the *tongue*.
The corresponding functions can be found in section 2.3. We decided to examine all of the enumerated muscles in our experiments.

Position Selection

The next step was to find an ideal electrode position for each of these muscles. We examined both, classical bipolar electrode configurations where voltages are recorded between pairs of neighbouring electrodes placed directly on a muscle and unipolar electrode configurations where voltages are recorded between one electrode placed on muscle tissue and a reference electrode placed on electrically neutral tissue. The second method is a standard approach for *EEG* measurements (i.e. for measuring brain activity) but is not commonly used for EMG measurements. Yet, the fact that we do not want to obtain an accurate measure for muscle force but simply want to minimize word error rates justifies exploring this method.

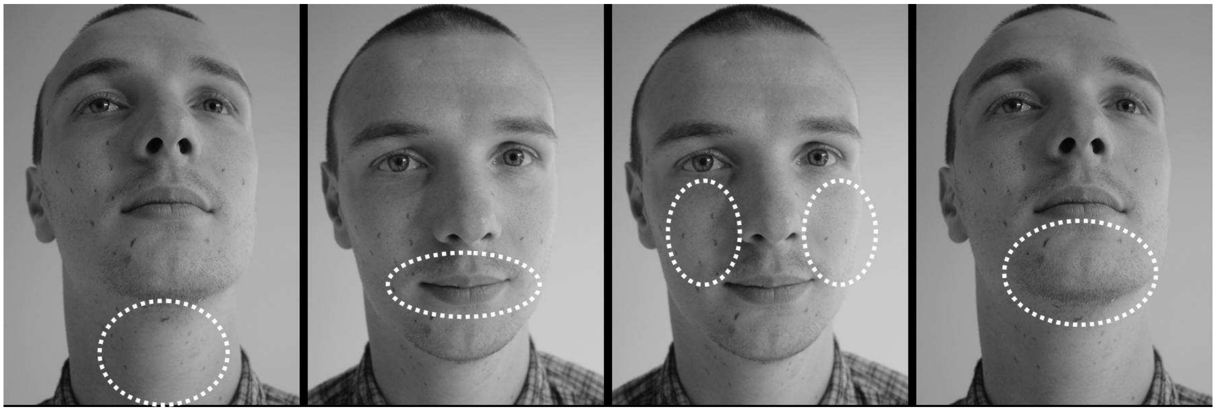


Figure 6.3: Muscles zones I-IV (left to right)

Zone	Associated Muscles	Associated Positions
Zone 1	Tongue	P3, P4, P5, P6, P7, P8, P46, P47, P51, P52
Zone 2	Orbicularis Oris	P31, P32, P33, P34, P40, P41, P42
Zone 3	Zygomaticus major, levator anguli oris	P15, P16, P17, P36, P37, P39, P43
Zone 4	Depressor anguli oris, digastric, tongue, platysma	P23, P24, P25, P27, P28

Table 6.3: Muscle zones, associated muscles and corresponding electrode positions. Figure 6.5 shows the enumerated positions.

Two speakers, S1 and S3, participated in the experiments on electrode placement. The resulting corpus was DIGITS_POS (5.3.3). Our original approach was as follows:

- Record three recording sessions per muscle: one audible and one non-audible session from the first speaker and one non-audible session from the second speaker
- in each session place two "electrode arrays" on the muscle to be examined - one for unipolar electrode configurations as illustrated in figure 6.4 (on one side of the face) and one for bipolar electrode configurations (on the other side of the face). Due to the fact that signals from these positions are measured *simultaneously* it is possible to compare signals (and recognition rates) for the *same* utterances but different positions.
- determine the best position (unipolar or bipolar) for each muscle for non-audible speech recognition by comparing signal quality and performance of the individual electrodes. It is worth mentioning here, that the classical bipolar electrode configuration is less susceptible to artifacts than the unipolar configuration because the detection electrodes are closer to each other and thus receive similar noise signals which are consequently eliminated by the differential amplifier. (refer to section 2.2 for more details). For this reason, bipolar configurations should always be preferred when similar recognition results can be achieved as with the unipolar configuration.
- select a subset of electrode positions from the set of "best positions" as final position set. Choose the combination of positions that maximizes word accuracies for non-audible speech.
- Additionally, compare the non-audible and audible sessions from the first speaker to examine which positions are particularly adequate for non-audible and audible speech respectively.

The following problems arose:

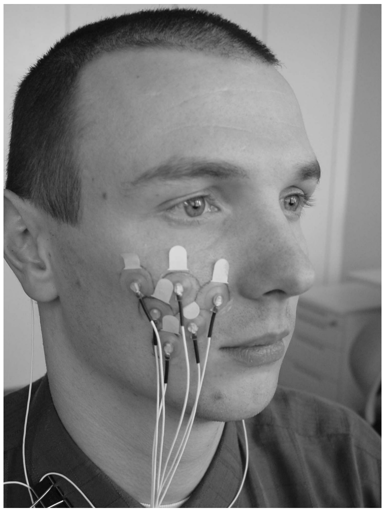


Figure 6.4: Data recording session for determining the best unipolar electrode configuration for *zone II.*

- our recording device only allowed simultaneous measurement of up to seven channels. Due to hardware problems, several sessions could only be recorded using six channels. As a result, it was impossible to measure two arrays, one for bipolar and one for unipolar electrodes, simultaneously.
- the electrodes could not be placed arbitrarily close to each other because of their own size. 1.5cm was the minimal possible inter-electrode distance. Consequently, not the complete area above a muscle could be covered by an array.
- identifying the exact location of a muscle without a medical expert was extremely challenging.
- An accurate assignment of *one* muscle to *one* position was impossible to find due to cross-talk and muscle overlap.

Due to these problems we came to the conclusion that our equipment and our medical background were not adequate for the task of identifying an *optimal* set of electrode positions for EMG based speech recognition. For this reason, we decided to find a (probably suboptimal) set of *appropriate* positions for each muscle *zone*. A muscle zone includes several muscles as illustrated in figure 6.3. Table 6.3 assigns muscles to zones.

We conducted several initial experiments on the DIGITS_INIT corpus to determine a set of electrode positions for further examination. Two speakers took part in these experiments. Among other things we examined where to optimally place bipolar electrodes by examining the signals resulting from different placement. We always included the positions chosen by other research groups in our experiments. Moreover, we recorded several short sessions to compare recognition results. The results are not reported here. Based on our experience we then chose a set of six to eight positions for each muscle zone for more systematic experiments on electrode placement (table 6.3). Figure 6.5 shows the corresponding positions. Note that position P22 merely serves as a reference point for unipolar measurements as proposed by [Chan et al., 2002b] (figure 3.1)).

Tables 6.4, 6.5, 6.6, and 6.7 summarize the results for our experiments on electrode positioning for each zone. The experiments were conducted on Set I of the DIGITS_VIII_POS corpus using six round robin sets and twenty-five training utterances per word in the vocabulary. The first column of each table contains the positions that were used. We used the format Px-Py to indicate which detection electrodes went to the plus and minus input of the amplifier respectively. When unipolar electrode configurations were used where both ears served as the reference Px served as an abbreviation for $P_x - average(P_{left\_ear} - P_{right\_ear})$. Columns 2,3 and 5 show the recognition results for the three

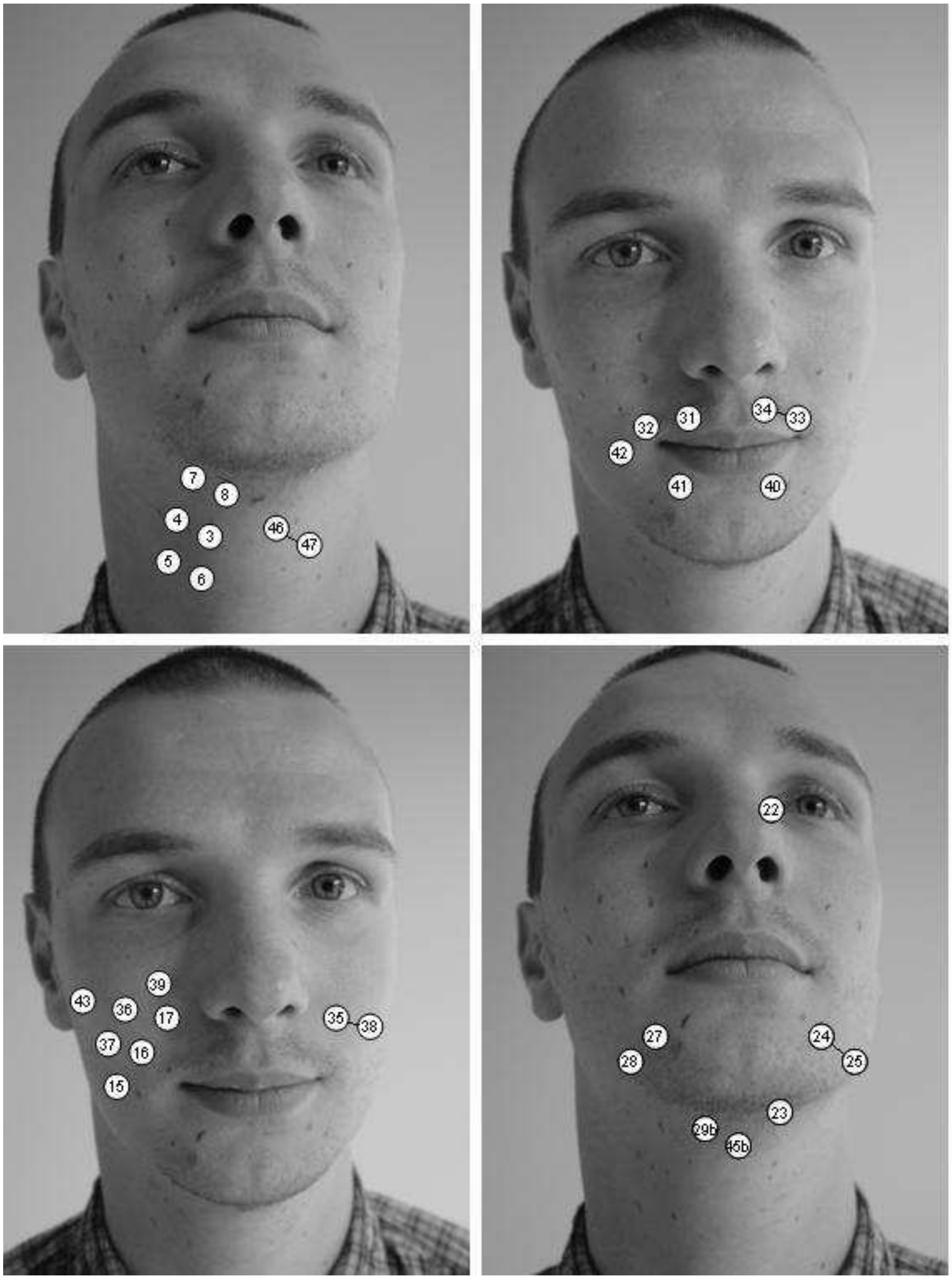


Figure 6.5: Set of electrode positions for zones I-IV. Connected positions were used for bipolar measurements. The remaining unipolarly measured positions were referenced to both ears except for P23 which was referenced to P22.

Position	S3 n-a.	S3 a.	Signal Quality S3	S1 n-a.	Signal Quality S1	candidate
P3	52.2%	66.3%	good; heartbeat visible	45.7%	weak signal; ear electrodes possibly had bad contact	no
P4	60.2%	77.7%	”	36.0%	”	no
P5	53.9%	79.3%	”	31.0%	”	no
P6	43.8%	57.0%	”	30.0%	”	no
P7	61.9%	77.3%	”	40.3%	”	no
P8	60.8%	73.7%	”	43.7%	”	no
P46-P47	-	-	-	61.3%	good	yes

Table 6.4: Within-session word-accuracies (in %) and description of the signal quality for the positions in *zone I*. The last column states whether the corresponding position is a candidate position for the final set of positions. n-a: non-audible session; a: audible session. The missing entries are referred to in the text.

Position	S3 n-a.	S3 a.	Signal Quality S3	S1 n-a.	Signal Quality S1	candidate
P42	82.0%	85.6%	good	47.0%	good	yes
P41	80.7%	85.6%	many artifacts	52.7%	many artifacts	no
P33-P34	83.7%	92.6%	many artifacts	45.0%	many artifacts	no
P31	83.7%	92.6%	many artifacts	45.0%	many artifacts	no
P32	86.0%	87.2%	good	48.0%	many artifacts	no
P40	76.3%	88.2%	many artifacts	45.0%	many artifacts	no
P35-P38	-	-	-	43.7%	good	yes

Table 6.5: Within-session word-accuracies (in %) and description of the signal quality for the positions in *zone II*. The last column states whether the corresponding position is a candidate position for the final set of positions. n-a: non-audible session; a: audible session. The missing entries are referred to in the text.

recorded sessions from speaker S1 (one non-audible (n-a.) session) and speaker S3 (one audible (a.) and one non-audible (n-a.) session). Feature extraction and classifier training were performed using our baseline system. A brief statement on the signal quality for each speaker is given in columns 4 and 6. Figure 6.6 shows sample signals for *zone IV*, and speaker S3. The third utterance contains a movement artifact in EMG6 (last signal sequences).

Depending on signal quality and recognition results the last column states whether or not the corresponding position was a *candidate* for our set of appropriate positions. Here, we have excluded positions that yielded bad signal quality or where a neighbouring position gave significantly better recognition results (e.g. P17 vs. P36 in table 6.6).

In *zone I* P46-P47 was chosen as candidate because it yielded the best recognition results. It is worth mentioning here that we assume that the reference (i.e. ear) electrodes had a rather bad contact to the skin in the recording session from speaker S1 because recognition results for all unipolar electrodes are extremely poor. For this reason we also included P4 in our set of candidates which had performed particularly well in initial experiments.

In the experiments for *zone II* P42 and P35-P38 were chosen as candidates because all other position yielded signals containing many artifacts. It is worth mentioning here that those positions were considered uncomfortable by the speakers.

In *zone III* the only positions that yielded acceptable signals for both speakers and had no neighbours with significantly better recognition results were P17 and P39 which were thus chosen as candidates for that zone.

Due to the fact that P22-P23 and P28 were the only positions with acceptable signal quality in *zone IV* they were picked as candidates for the fourth zone.

It should be pointed out that electrode configuration P35-P38 which does in fact correspond to *zone III*

Position	S3 n-a.	S3 a.	Signal Quality S3	S1 n-a.	Signal Quality S1	candidate
P17	88.7%	92.3%	good	74.3%	good	yes
P16	85.0%	89.3%	good	52.7%	good	no
P36	80.7%	86.0%	good	53.6%	good	no
P43	73.3%	79.3%	good	38.3%	weak signal	no
P37	83.3%	84.9%	good	56.0%	good; several artifacts	no
P15	80.3%	84.2%	good	60.3%	good; several artifacts	no
P39	60.8%	73.7%	good	66.3%	good	yes

Table 6.6: Within-session word-accuracies (in %) and description of the signal quality for the positions in *zone III*. The last column states whether the corresponding position is a candidate position for the final set of positions. n-a: non-audible session; a: audible session.

Position	S3 n-a.	S3 a.	Signal Quality S3	S1 n-a.	Signal Quality S1	candidate
P22-P23	84.6%	94.6%	good	71.0%	good	yes
P24-P25	81.9%	85.7%	good	50.0%	many artifacts	no
P27	74.6%	86.4%	many artifacts	57.3%	many artifacts	no
P28	78.9%	83.3%	good	58.3%	good	yes
P29b	73.9%	87.5%	several artifacts	70.0%	good	no
P45b	72.2%	86.4%	many artifacts	50.0%	good	no

Table 6.7: Within-session word-accuracies (in %) and description of the signal quality for the positions in *zone IV*. The last column states whether the corresponding position is a candidate position for the final set of positions. n-a: non-audible session; a: audible session.

was used in the recording session for *zone II* because our hardware device did not allow simultaneous measurement of more than seven channels. Moreover, two entries for speaker S3 (*zone I* and *zone II*) are missing which must again be attributed to hardware problems. A final session was recorded to compensate for these errors. Refer to the following paragraph for a more detailed explanation.

Our experiments yielded a set of candidates for appropriate electrode positions as shown in table 6.8. For each candidate the third column enumerates neighbouring positions that were also contained in the candidate set. We had to choose a subset of six positions from the set of candidates to ensure reliable recordings with our hardware device. We therefore recorded a final session with those candidate positions whose set of similar positions was not empty. The fact that positions P35-P38 and P46-P47 were recorded in these sessions compensates the problem addressed in the previous paragraph.

The results are shown in table 6.9. We excluded the positions P4 and P39 because similar positions (P46-P47 and P35-P38,P17 respectively) yielded better recognition results for non-audible speech. The set of appropriate electrode positions was thus: P22-P23 (EMG1), P35-P38 (EMG2), P17 (EMG3), P42 (EMG4) , P28 (EMG5), and P46-P47 (EMG6) as shown in figure 6.7.

For data collection and offline recognition (where we could run the risk of the recording system's crashing) we added another position to the list, namely position P51-P52 (EMG7) which is in fact the same as P46-P47, but on the other side of the larynx. We chose this position because it was recommended by [Jorgensen et al., 2003] and because positions in *zone I* are the most practical ones (the electrodes are placed on the neck rather than on the face). Moreover, we were going to examine if the two positions P46-P47 and P51-P52 would provide complementary information.

Results for Selected Positions

We evaluated our selected positions on Set I of the DIGITS_VIII corpus (3 speakers, 4 non-audible sessions each) using our baseline system described in section 6.1. Table 6.10 shows the word accuracies for within-session testing for each speaker using different numbers of channels for recognition. The

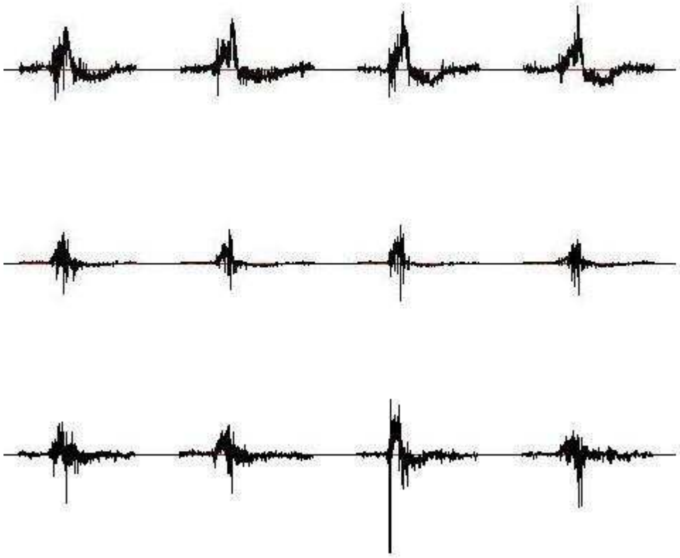


Figure 6.6: Sample signals for the word “three” (four repetitions) and EMG1 (P22-P23; upper), EMG2 (P24-P25; middle) and EMG6 (P45b; lower). Speaker S3, session 038.

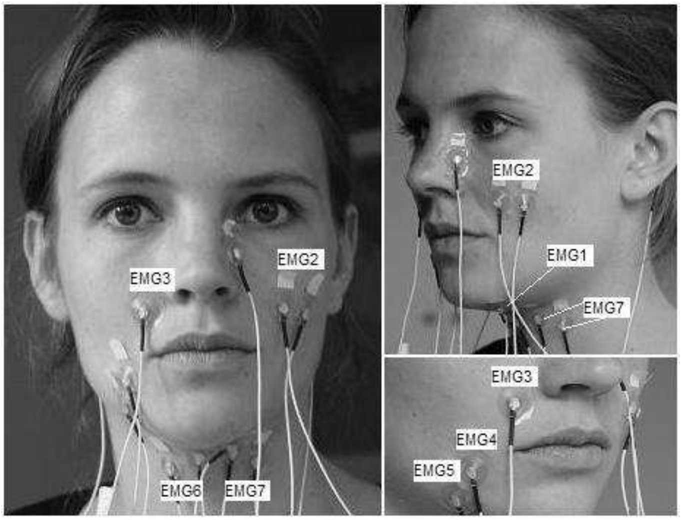


Figure 6.7: Positioning for electrodes EMG1-EMG7. Section 6.2.1 enumerates the associated muscles

Position	Zone	Similar Candidates
P4	1	P46-P47
P46-P47	1	P7
P42	2	-
P17	3	P39, P35-P38
P39	3	P17, P35-P38
P35-P38	3	P17,P39
P22-P23	4	-
P28	4	-

Table 6.8: Set of candidates for finale position set.

Position	S3 audible	S3 non-audible	Signal Quality	Final Position
P35-P38	94.3%	94.0%	good	yes
P46-P47	78.7%	86.3%	good	yes
P39	95.3%	90.7%	good	no
P17	96.7%	95.3%	good	yes
P7	67.3%	79.3%	good	no

Table 6.9: Results for final experiment on electrode positioning

table presents the results for (a) each individual channel, (b) the combination of all channels, and (c) the best combination of $k = 2, 3, 4, 5, 6$ channels. We used a greedy procedure to identify the best combination of k channels: Initially, we simply chose the channel yielding the best individual within-session results. We then added the remaining channels one by one, in the order that gave the best (within-session) performance when combined with the already selected channels.

The results in table 6.10 indicate a significant variation in performance for the individual channels. Channels EMG1 and EMG3 yield the best recognition results for all speakers. These two channels correspond to different muscle groups, and therefore provide orthogonal information. The results from the best channel combination in table 6.10 reveal that it is crucial to apply more than one electrode (highly significant difference between Best 1 and Best 2). Even between two and three electrodes we see a highly significant performance increment on the 9.56E-05 · 100% level, while the performance differences for five, six, or seven electrodes are insignificant.

It should be pointed out that speaker S3 achieved the best recognition results. As already mentioned above this speaker had already recorded several non-audible sessions before recording the sessions for the DIGITS_VIII corpus. He stated that he had developed a particular speaking style for non-audible speech over time. It is worth mentioning here that we noticed for all speakers that an increasing level of experience improved the performance. Recognition results for speaker S3, for example, are better in the final experiment on electrode positioning (table 6.9) than in the previous experiments (tables 6.4 and 6.5) for both, audible and non-audible speech.

Figure 6.8 presents sample signals for the selected positions. The low amplitudes of the bipolarly recorded channels (EMG2,6,7) suggests that the signals of the unipolarly recorded channels (EMG1,3,4,5) provide more information. However, recognition results do not confirm this assumption. EMG2 for example yields better recognition results than EMG4 and EMG5. Moreover, even though several signals look similar they still provide complementary information: The channels EMG6 and EMG7 together, for instance, yield a within-session recognition rate of 75.7% which is significantly higher than the individual channel results (60.1% and 62.0% respectively). It should be pointed out again that these two channels actually represent the same position - just on different sides of the larynx (figure 6.7).

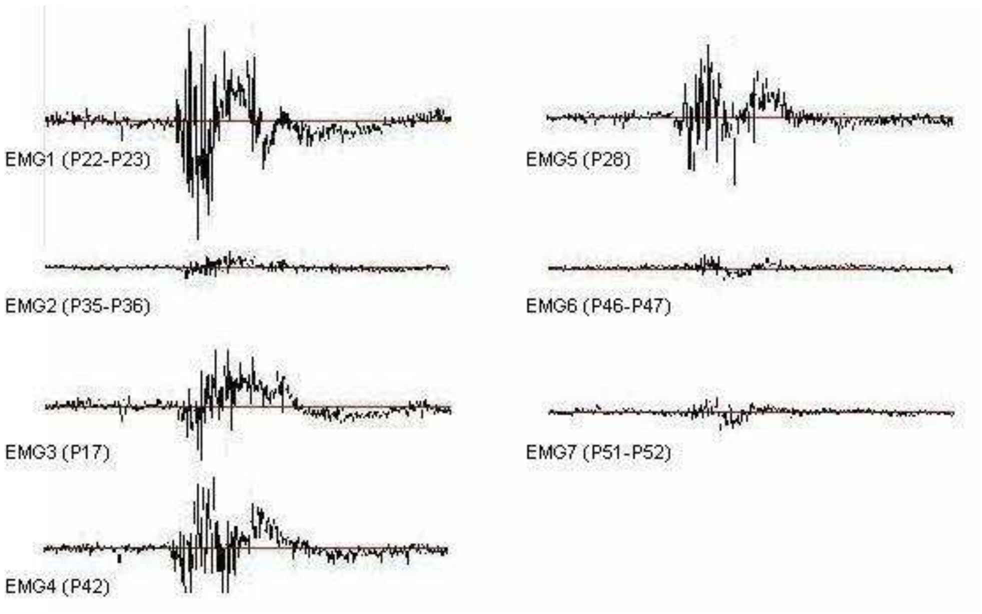


Figure 6.8: Signals for the audibly spoken word zero and electrodes EMG1-EMG7 (speaker S1, session 016).

Suitability of positions for audible and non-audible speech

The purpose of our experiments on electrode placement was to (1) find an appropriate set of electrode positions for non-audible speech recognition and to (2) compare usefulness of positions for audible and non-audible speech. The first aspect was dealt with in the previous section. We want to begin our analysis on suitability of the positions for audible and non-audible speech by mentioning some observations we made in the course of our experiments.

1. In the first experiments (tables 6.4 to 6.7) the results for non-audible speech are significantly worse that those for audible speech *for all positions*. This indicates that audible speech is either more intuitive or contains more information for speech recognition.

2. In the final experiment (table 6.9) some positions (P17, P35-P38, P7, P39) perform better for non-audible speech than for audible speech even though the situation was vice versa in previous experiments (tables 6.5 and 6.6). It is worth mentioning here, that speaker S3 recorded several non-audible sessions for our demo system between the initial experiments and the final experiment. We deduce from this that non-audible speech can be learnt and that certain positions are just as good for non-audible speech as for audible speech (at least for our recognition task).

3. Recognition results for non-audible speech are worse than recognition results for audible speech for all positions in *zone I*. This could be traceable back to the fact that signals from the vocal cords are picked up by the electrodes in *zone I*.

We produced Set II of corpus DIGITS_VIII for further analysis. One audible and one non-audible session was recorded for each speaker. These two “sessions” were in fact recorded as one session with the exact same electrode placement, i.e. the electrodes were not removed between the two parts. The only difference was the speech manner. The recognition results for individual EMG channels are shown in table 6.11 and 6.12 for non-audible and audible session respectively. Table 6.13 presents the differences (Δ) in the recognition results in % (absolute values).

Channels	**S1**	**S3**	**S7**	**Average**
Individual Channels				
EMG1 (P22-P23)	74.2	92.1	**77.4**	**81.2**
EMG2 (P35-P38)	64.1	90.7	69.4	74.7
EMG3 (P17)	**76.1**	**93.8**	72.9	**81.0**
EMG4 (P42)	61.2	83.1	71.6	71.9
EMG5 (P28)	62.4	73.4	63.6	66.5
EMG6 (P46-P47)	63.6	64.4	52.3	60.1
EMG7 (P51-P52)	59.8	66.3	60.0	62.0
Avg EMG1-EMG7	65.9	80.5	66.7	71.1
Channel Combination				
Best 1 (EMG1)	74.2	92.1	77.4	81.2
Best 2 (EMG1,3)	93.5	97.6	90.1	93.7
Best 3 (EMG1,3,6)	97.1	98.1	91.3	95.5
Best 4 (EMG1,3,4,6)	**97.5**	98.3	93.4	96.4
Best 5 (EMG1,2,3,4,6)	97.3	98.6	95.5	97.1
Best 6 (EMG1,2,3,4,5,6)	97.4	**98.8**	**96.2**	**97.4**
All 7 channels	97.2	98.8	96.0	97.3

Table 6.10: Within-session word accuracies (in %) for individual channels and the best combination of $k = 2, 3, 4, 5, 6, 7$ channels on Set I of the DIGITS_VIII corpus using the baseline system (section 6.1).

Channel	**Position**	**S1**	**S3**	**S7**	**Average**
EMG1	P22-P23	69.5	91.3	73.3	78.0
EMG2	P35-P38	65.8	89.5	65.5	73.6
EMG3	P17	75.3	95.0	64.5	78.3
EMG4	P42	59.5	86.3	60.5	68.8
EMG5	P28	56.0	76.5	58.8	63.8
EMG6	P46-P47	57.8	71.0	48.3	59.0
EMG7	P51-P52	57.3	74.3	49.8	60.4

Table 6.11: Within-session word accuracies (in %) for individual channels and the non-audible sessions of Set II of the DIGITS_VIII corpus using the baseline system (section 6.1).

Speakers S1 and S7 have significantly better recognition results for audible than for non-audible speech for all positions ($\Delta > 0$). Again, this is traceable back to the fact that they had no or only little experience in speaking non-audibly. Speaker S3 on the other hand has better recognition results for non-audible speech than for audible speech in some cases (EMG3, EMG5). Once more, this indicates that non-audible speech is indeed learnable. The high Δ values for EMG7 compared to EMG6 for speakers S3 and S7 are surprising because EMG6 and EMG7 are actually placed on the same position - just on different sides of the larynx. We have no explanation for this phenomenon at this point.

It is noticeable that EMG1 and EMG3 yield the best recognition results while EMG6 and EMG7 yield the worst recognition results for both, non-audible and audible speech. This indicates, that positions that are particularly well suited for audible speech are also well suited for non-audible speech and vice versa.

In section 6.7 we compare audible and non-audible speech in general.

As already mentioned above, another challenge for EMG measurements besides locating appropriate electrode positions is to identify (approximately) the same locations in consecutive recording sessions. The following section addresses this issue.

Channel	Position	S1	S3	S7	Average
EMG1	P22-P23	81.5	92.8	79.8	84.7
EMG2	P35-P38	71.5	91.8	82.0	81.8
EMG3	P17	83.5	90.5	77.5	83.8
EMG4	P42	77.3	90.0	71.5	79.6
EMG5	P28	73.8	71.5	66.8	70.7
EMG6	P46-P47	65.8	73.5	51.8	63.7
EMG7	P51-P52	63.8	83.0	63.0	69.9

Table 6.12: Within-session word accuracies (in %) for individual channels and the audible sessions of Set II of the DIGITS_VIII corpus using the baseline system (section 6.1).

Channel	Position	ΔS1	ΔS3	ΔS7	ΔAverage
EMG1	P22-P23	12.0	1.5	6.5	6.7
EMG2	P35-P38	5.8	2.25	16.5	8.17
EMG3	P17	8.3	-4.5	13.0	5.6
EMG4	P42	17.8	3.8	11.0	10.8
EMG5	P28	17.8	-5.0	8.0	6.9
EMG6	P46-P47	8.0	2.5	3.5	4.7
EMG7	P51-P52	6.5	8.8	13.3	9.5
Avg EMG1-EMG7	-	10.8	1.3	10.3	7.5

Table 6.13: Differences of recognition results for audible and non-audible speech. ΔS1 = Word Accuracy S1 audible (table 6.12) - Word Accuracy S1 non-audible (table 6.11).

6.2.2 Repeating Electrode Placement

One of the main challenges in EMG measurements is to ensure repeatability of electrode placement. We investigated the following methods in order to obtain reliable positioning.

Methods

Tape measure: The standard approach for identifying electrode positions for EMG measurements is the use of tape measure. As already mentioned in section 2.2.2, a location should be defined in relation to a line between two anatomical landmarks (e.g. bones). However, this proves to be relatively difficult in areas on the body surface, where no appropriate anatomical landmarks can be found. The region around the larynx that we referred to as *zone I* in section 6.2 for example contains no bones.

Picture: Another approach to obtain stable placement is the use of pictures from previous recordings. This method is quite successful for relatively flat regions like the region around the mouth, where no perspective information is needed. However, finding a specific spot on more curved areas is extremely difficult and unreliable.

Marker: A straightforward idea for ensuring stable positioning is the use of a marker. The exact electrode positions are marked in one recording session prior to electrode removal and re-used in the following recording session. Obviously, this approach is impractical (nobody wants to wear permanent marker outside the recording sessions), yet, it can serve for experimental purposes.

Gypsum masks: Finally, gypsum masks can serve to ensure placement repeatability for one person: once a mask has been produced for a particular speaker, appropriate positions can be marked on that speaker's face (e.g. using lipstick). The mask is then applied to the speaker in order to transfer the marked positions to the inner side of the mask. Finally, holes are drilled into the mask in order to allow position marking in following sessions. Figure 6.9 illustrates the use of a gypsum mask for position identification.

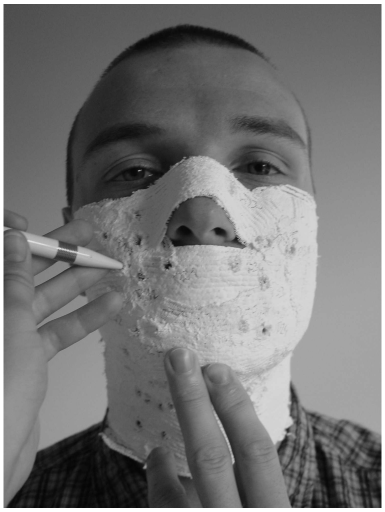


Figure 6.9: Application of gypsum mask for position identification.

The advantages and disadvantages of all four methods are summarized in table 6.14. Tape measure and gypsum masks seem to be the most useful utensils for position identification. In order to compare the accuracy of the two methods we conducted the following experiment with speaker S3:

1. Apply each method five times in a row to a set of three different positions (P17, P23, P46). Mark the identified spot each time.

2. For each position determine the distance between each pair of marks and calculate the average distance.

3. Compare the results.

Average and maximal distances for each position and both methods are shown in table 6.15.

Method	inter-session deviation	advantages	disadvantages
Tape Measure	< 10mm	relatively relieable; usable across speakers	requires long preparation time; hard to find anatomical landmarks
Gypsum mask	< 5mm	very reliable method; quick position identification	requires preparation of gypsum masks; only usable for one speaker
Picture	< 15mm	good for transferring positions from one speaker to another speaker	very unreliable
Marker	< 2mm	most reliable method	impractical; only usable for one speaker

Table 6.14: Advantages and disadvantages of different methods for repeating electrode placement.

We deduce from our results that using gypsum masks is the most appropriate method for ensuring reliable placement when a large number of experiments are conducted on a small number of speakers. In this case, producing gypsum masks is worth the effort. In the more general case (few experiments, many speakers) our recommendation is to make use of tape measure.

Method	**Avg/Max P17**	**Avg/Max P23**	**Avg/Max P46**	**Overall Average**
Gypsum mask	2.1mm/4mm	1.8mm/4mm	2.4mm/4mm	2.1mm
Tape measure	2.8mm/5mm	4.2mm/7mm	3.6mm/5mm	3.5mm

Table 6.15: Average and maximal differences between independently marked positions using tape measure and gypsum masks for positions P17, P23 and P46.

Influence of robust positioning

As already mentioned in section 2.2.4 slight changes in electrode position, temperature, amount of applied electrode gel or tissue properties may alter the sEMG signal significantly. In order to measure the impact of some of these factors on across-sessions recognition speaker S0 recorded two sessions (006 and 007) on the phone domain in series. The positions from session 006 were marked and the electrodes were removed and re-applied according to the labels for the following session 007. We then determined within-session and across-sessions recognition results on Set II of the PHONE corpus (sessions 006, 007 and another session 004 from a different day) using our baseline system (section 6.1) and twenty-five training samples per word for each recognizer. The results are shown in table 6.16.

	Session 004	Session 006	Session 007
Session 004	90.4	55.1	62.2
Session 006	52.1	94.8	**77.0**
Session 007	50.9	**78.0**	97.3

Table 6.16: Across-sessions results (in %) on Set II of the PHONE corpus. No segmentation method was applied. Train on (row). Test on (column).

The following conclusions can be inferred from this:

1. Across-sessions results are significantly higher for the marker dependent sessions 006 and 007 (77.5%) than for the independent sessions (004 with 006 and 007; 55.1%). This is most likely traceable back to greater positioning variations as well as to changes in speech patterns, temperature, or skin properties which vary on different days.

2. Within-session results (average: 94.2%) are significantly better than across-sessions results even for the marker related sessions 006 and 007 (average: 77.5%). We infer from this that the properties of the electrode/skin interface (e.g. the amount of applied electrode gel) - which change with removal and re-application of electrodes - have a significant impact on the sEMG signal and thus on across-sessions recognition. Moreover, we assume that even position changes in the range of 1mm - 2mm alter the sEMG signal significantly. Due to the fact that a change in position always requires re-application of electrodes and vice versa, however, these two aspects cannot be examined independently of each other.

We deduce from our observations that session dependence is a major challenge in EMG speech recognition and address this issue in a separate section 6.6.

In conclusion, we derived an appropriate set of electrode positions for EMG based speech recognition taking into account signal quality, recognition results for individual positions, and recognition results for combined positions. Furthermore, we examined several methods for repeating electrode placement and showed that gypsum masks and tape measure are suitable for this task. In the next section we describe the experiments we performed to improve the frontend of our baseline system using the selected electrode positions.

6.3 Optimization of Data Preprocessing

Data Preprocessing is the process of transforming raw input signals into a sequence of *feature vectors*. The aim is to extract information relevant for discriminating speech units (e.g. phonemes or words) while omitting irrelevant information. Over the years a variety of different preprocessing methods for Automatic Speech Recognition have been developed. Section 2.4.2 summarizes the most commonly used ones. However, most approaches incorporate information about human auditory processing and perception. Mel-scale Filterbank coefficients, for example, imitate the frequency dependent spectral resolution of the human ear. We thus investigated various preprocessing methods in order to find an appropriate frontend for *EMG based* speech recognition. Initial experiments were conducted on the PHONE corpus and yielded the baseline system introduced in section 6.1.1. The features that gave promising results in those initial experiments were investigated further on the DIGITS_VIII corpus in order to optimize the baseline system.
Our approach for the development of an appropriate frontend consisted of the following steps:

1. Determining window size and shift: window size and window shift of the observation windows were varied while all other settings of the baseline system were held constant in order to obtain optimal values for these parameters (section 6.3.1)

2. Choosing useful features: Different features were examined while all other setting of the baseline system were held constant (section 6.3.2).

3. Dimensionality reduction: the use of an LDA on the baseline system for dimensionality reduction was explored (section 6.3.3).

The following sections (sections 6.3.1 - 6.3.3) address these aspects. We will show, however, that no significant improvements can be achieved by modifying the frontend. Section 6.3.4 describes our approach for determining appropriate values for the sampling rate and the bandpass filter cut-off frequencies of our system.

6.3.1 Optimizing Window Size and Window Shift

Tables 6.17 and 6.18 show the within-session recognition rates for various window sizes and shifts.

Window size	STFT Coefficients per Channel	S1	S3	S7	Average
54ms (BASE)	17	97.2	98.8	96.0	97.3
27ms	9	97.1	99.1	96.1	97.4
108ms	33	94.6	93.4	58.1	82.0
16ms	9	96.4	99.0	96.2	97.2
32ms	17	97.0	98.6	95.6	97.1
64ms	33	95.6	97.6	95.0	97.1
128ms	65	92.6	93.6	43.6	76.6

Table 6.17: Recognition rates (in %) for different window sizes on Set I of the DIGITS_VIII corpus. The second column shows the corresponding number of STFT coefficients per channel.

It can be seen that no significant improvements can be made by varying window size and window shift; maximum values are obtained for nine or seventeen STFT coefficient per channel and window sizes of up to 8ms. However, it is noticeable, that recognition rates for speaker 007 drop significantly for large window sizes. Interestingly, this holds especially for non-audible speech and could be traceable back to a greater speaking rate which violates the assumption of a quasi-stationary signal (section 2.4.2).

6.3.2 Optimizing Feature Extraction

As already mentioned in section 6.1.1 we examined the use of various features for an EMG speech recognition frontend in initial experiments. Some of them performed so poorly (zero crossing rate, lpc

Window shift	S1	S3	S7	Average
4ms (BASE)	97.2	98.8	96.0	97.3
8ms	97.3	98.8	95.1	97.1
16ms	96.4	98.1	91.4	95.3
27ms	94.8	97.6	89.0	93.8

Table 6.18: Recognition rates (in %) for different window shifts on Set I of the DIGITS_VIII corpus. All other parameters are the same as as in the baseline system.

coefficients, AR coefficients among others) that we did not investigate them any further. The most promising features were

- *STFT*: windowed Short Time Fourier coefficients (JRTk function *FeatureSet:spectrum*)
- *Delta*: Delta coefficients (JRTk function *FeatureSet:delta*; delta=1)
- *Adjacent*: The JRTk adjacent feature (JRTk function *FeatureSet:adjacent*; delta=1)
- *mean*: the windowed mean of the time domain values
- *abs_mean*: the windowed mean of the absolute values of the time domain values

Refer to the JRTk documentation ([Metze, 2004]) for a more detailed description of these functions. Table 6.19 shows the recognition results for individual features and some combinations of features. The following observations deserve mentioning:

1. The time domain features (*mean* and *abs_mean*) by themselves yield high recognition rates (up to 89%).
2. Despite the fact that EMG data processing in the time domain usually involves *absolute* values (section 2.2.4), the *mean* feature performs better than the *abs_mean* feature. We assume, that this relies to the fact that we have an extremely low cut-off frequency for our high-pass filter. Section 6.3.4 addresses this aspect.
3. *Delta* coefficients outperform *STFT* coefficients. The extremely large performance difference for speaker S7 is surprising. It suggests that the signals are dominated by noise which is eliminated by the *delta* coefficients. However, a visual analysis of the signal data did not confirm this assumption.
4. The *adjacent* feature performs rather poorly compared to the remaining features which confirms the results of previously conducted experiments on the PHONE corpus.

Features	S1	S3	S7	Average
STFT	89.6	95.8	56.1	80.5
Delta	95.4	97.8	93.9	95.7
adjacent	69.4	77.8	47.3	64.8
mean	83.9	94.3	88.9	89.0
abs_mean	81.4	92.4	83.3	85.7
delta & mean (BASE)	97.2	98.8	96.0	**97.3**
delta & abs_mean	96.4	98.1	95.3	96.6
STFT & mean	92.3	96.9	59.3	82.8
STFT & abs_mean	81.4	92.4	83.3	85.7

Table 6.19: Recognition rates (in %) for various features on Set I of the DIGITS_VIII corpus.

In conclusion, we were not able to improve performance by choosing different features. The *delta* feature along with the *mean* feature were confirmed to be the optimal choice.

6.3.3 Applying Dimensionality Reduction

Dimensionality reduction is the mapping of a multidimensional space into a space of fewer dimension. The goal is to reduce data complexity without losing information. As already mentioned in section 2.4.2 a Linear Discriminant Analysis (LDA) is commonly applied in conventional speech recognition systems which leads to a reduction of word error rates. We examined the use of an LDA for dimensionality reduction in EMG based speech recognition. Each codebook represented one LDA class. The within-session results for different numbers of LDA coefficients on Set I of the DIGITS_VIII corpus are shown in table 6.20.

Number of coefficients	S1	S3	S7	Average
BASE (no LDA)	97.2	98.8	96.0	97.3
1	21.3	22.3	16.8	20.1
2	51.1	72.5	54.1	59.2
4	77.9	90.6	79.5	82.7
8	85.8	95.3	89.2	90.1
12	87.4	97.5	92.2	92.4
16	89.7	97.5	93.7	93.6
24	**92.1**	97.8	**94.3**	**94.7**
32	91.8	**97.9**	94.1	94.6
50	91.4	97.4	93.8	94.2
64	91.9	97.4	93.1	94.1
126	90.3	96.8	92.8	93.3

Table 6.20: Within-session results (in %) on Set I of the DIGITS_VIII corpus depending on the number of LDA coefficients.

It can be seen that applying an LDA leads to a degradation in performance regardless of the number of coefficients. Results were even worse for across-sessions testing (> 20% (absolute) performance difference between naive baseline system results (no normalization) and results for an additionally applied LDA with 24 coefficients). Jorgensen et al. reported similar results when examining the use of a Principal Component Analysis (PCA) for dimensionality reduction in EMG based speech recognition [Jorgensen et al., 2003]. It should be pointed out that word accuracies decreased even more when different LDA classes (e.g. words) were chosen. Even though, it is noticeable, that relatively small numbers of LDA coefficients (e.g. 24) yield comparable or better recognition results than large numbers of coefficients (e.g. 126).
We suggest to explore feature dimensionality reduction methods for EMG based speech recognition in the future.

6.3.4 Optimizing Sampling Rate and Filter Properties

Just like acoustic signals EMG signals highly depend on the sampling rate and filter properties. Typically, sampling rates above 1000Hz are chosen because of the usable energy of the EMG being in the 1Hz - 500Hz frequency range. Due to the fact that we used the serial port with a maximum transmission rate of 115kBaud (only about 80% of which should be utilized for data transmission) a 1000Hz sampling rate only allowed the simultaneous recording of up to $80\% \cdot \frac{115000 bit/s}{1000 Hz \cdot 16 bit} = 5.75$ channels including the synchronization and the marker channel. Thus, there was a tradeoff between choosing a high sampling rate and recording a large number of channels. The following paragraph addresses this problem. Next, we describe our experiments on determining appropriate high-pass filter settings.

Sampling Rate and Low-Pass Filter

As already mentioned above, the differential electrode configuration acts as a bandpass filter in the spectral region of the EMG signal. For a typical conduction velocity (4.0m/s) and an inter-detection

surface distance of 10mm, the pass frequency is at 200Hz [Luca, 2002]. Moreover, it decreases with increased inter-electrode distance [Soderberg, 1992]. There is no significant loss of information, however, because of the dominant EMG energy lying in the 50Hz - 150Hz frequency range.
We deduced from this that there should be no loss in information when applying a 300Hz low-pass filter along with a 600Hz sampling rate instead of a 500Hz low-pass filter along with a 1000Hz sampling rate for our EMG measurements. The following experiment was performed to confirm this assumption: First, recognition rates on the PHONE corpus and on the DIGITS_II corpus were determined with the standard settings (1010Hz sampling rate, 1Hz-500Hz bandpass filter). Next, we performed downsampling on several test sets, choosing sampling rates ranging from 500Hz to 1000Hz, and determined the corresponding word accuracies. The results are shown in 6.21.

Sampling Rate	**Window**	**S0**	**S1**	**S3 audible**	**S3 non-audible**	**Avg**
1000Hz	32ms	90.6	88.3	81.3	71.1	**82.8**
900Hz	35ms	90.3	86.4	82.2	71.1	82.5
800Hz	40ms	90.7	86.7	83.4	73.3	83.5
700Hz	45ms	90.8	85.8	82.4	73.2	83.0
600Hz	54ms	90.9	86.3	83.6	73.8	**83.6**
500Hz	64ms	87.8	86.0	84.3	72.3	82.6

Table 6.21: Within-session word accuracies (in %) on Set I of the PHONE corpus (speakers S0 and S1) and on Set I (S3 audible) and Set II (S3 non-audible) of the DIGITS_II corpus for various sampling rates. The remaining settings are the same as in the baseline system.

It can be seen that downsampling does not decrease recognition rates - there was even a slight improvement in performance. It is worth mentioning here that the reconstructed downsampled data could theoretically contain aliasing artifacts. Even though, we see our assumption confirmed by the experimental results. As a consequence, a 300Hz low-pass filter was chosen for the VARIOPORT_VIII data recording system.

High-Pass Filter

As already mentioned in section 2.2, EMG data recording systems typically apply 10Hz-20Hz high-pass filters in order to reduce movement artifacts. The EMG channels of our first data recording system, VARIOPORT_II, however, yielded extremely good recognition rates despite the fact that they were configured with 1Hz high-pass filters. In order to examine the MES characteristics depending on the high-pass filter cut-off frequency one channel of our new data recording system VARIOPORT_VIII was configured with a 20Hz high-pass filter whereas the rest of the EMG channels was again configured with a 1Hz high-pass cutoff frequency. We then performed the following experiment to determine which filter yields the better results:

1. Set I of the MEETING corpus was recorded (two sessions from one speaker). In the first session, position P46-P47 (section 6.2) was recorded by channel EMG7 (1Hz high-pass filter). In the second session, the same position was recorded by channel EMG8 (20Hz high-pass filter).

2. Recognition results for position P46-P47 and different features were determined and compared for both sessions as shown in table 6.22. Twenty-five samples per phrase were used for system training and the remaining settings were the same as for the baseline system (section 6.1).

It is noticeable that recognition results for all features except for the abs_mean feature are significantly higher for channel EMG7 (1Hz filter). We deduce from this that a 1Hz high-pass filter is more appropriate for EMG based speech recognition than the standard 20Hz high-pass filter even though the signal contains more movement artifacts. The better recognition results for the abs_mean feature and channel EMG8 are possibly traceable back to the fact that the absolute value of an EMG channel that is configured with a 20Hz high-pass filter yields a more accurate measure of muscular force because no DC is contained in the signal.

Features	**EMG7 (Session 000)**	**EMG8 (Session 004)**	Δ
STFT	58.3	42.9	15.5
Delta	60.0	43.3	16.7
mean	32.9	23.8	9.1
abs_mean	20.8	27.1	-6.3
Delta & mean	70.8	49.5	21.3
Delta & abs_mean	64.6	42.9	21.72

Table 6.22: Recognition accuracies for position P46-P47 for channels EMG7 (1Hz high-pass filter) and EMG8 (20 Hz high-pass filter) on Set II of the MEETING corpus for various features.

6.4 Optimization of Hidden Markov Model Classifier

First order HMMs with Gaussian mixture models are used in most conventional ASR systems as classifiers because they are able to cope with both, variance in the time-scale and variance in the shape of the observed data. In this section we present experiments for optimizing the Hidden Markov Model classifier for isolated word recognition introduced in section 6.1. We will show, however, that no significant improvements in performance can be achieved by varying HMM topologies and the number of gaussians per state.

6.4.1 Optimizing HMM Topologies

When performing isolated word recognition on a relatively small vocabulary (< 20 words) it is feasable to train a Hidden Markov Model for every word in the vocabulary. We examined the following approaches to determine an optimal number of states for each word:

1. Choose the same number of states for every word in the vocabulary.
2. For each word let the corresponding number of states be proportional to the number of phonemes in that word.

The methods were tested on Set I of the DIGITS_VIII corpus as presented in table 6.23. Due to the fact that all digits consist of only one or two syllables we did not choose HMM topologies where the number of states was proportional to the number of syllables in the corresponding word. However, this approach was examined in isolated *phrase* recognition which we used in our demo systems (chapter 7).

Unit	**states per unit**	**S1**	**S3**	**S7**	Average
word	3	95.8	98.4	94.9	96.4
word	4	95.9	98.5	94.5	96.3
word	5	97.2	98.8	96.0	97.3
word	6	96.4	98.9	95.9	97.1
word	7	96.6	98.8	95.9	97.1
phoneme	1	93.9	97.6	94.6	95.4
phoneme	2	96.5	98.9	95.3	96.9
phoneme	3	96.6	98.4	95.6	96.9

Table 6.23: Within-session word accuracies (in %) on Set I of the DIGITS_VIII corpus for different HMM topologies.

The results indicate that choosing at least five states per word and two states per phoneme is crucial. In fact, the χ^2-test confirms that the results for the four-state and five-state HMM are different at a significance level of 0.538%. The two different methods of determining HMM topologies (fixed number of states per word vs. number proportional to number of phonemes in the corresponding word) yield no significant performance differences.

6.4.2 Optimizing Gaussian Mixture Models

The second step for building a HMM speech classifier is to determine an appropriate number of gaussians for each HMM state. The corresponding experiments were performed on Set I of the DIGITS_VIII corpus. Table 6.24 shows the results.

Gaussians per State	S1	S3	S7	Average
8	96.1	98.0	94.7	96.2
12	97.2	98.8	96.3	97.4
16	96.6	98.8	95.4	96.9
20	96.9	99.4	96.4	97.6
24	97.2	99.1	96.7	97.6
32	97.2	98.8	97.0	97.6
50	96.8	98.9	96.7	97.4

Table 6.24: Within-session word accuracies (in %) on Set I of the DIGITS_VIII corpus for different numbers of gaussians per HMM state.

It can be seen that no significant improvements can be made by varying the number of gaussians per state. The experiments suggest, however, to choose more than eight gaussians.
In summary, we did no improve performance by varying HMM topologies or the number of gaussians. The next section addresses the issue of EMG signal segmentation.

6.5 Segmentation

The task of segmentation is to find the boundaries of words within an utterance. In isolated word recognition utterances of the form "*silence [word] silence*" are recorded and classified. Theoretically, the system does not have to "know" the exact beginning and end of the word within the utterance, because viterbi paths for each vocabulary word can be computed using the complete signal. In this case, however, recognition rates may depend on the amount of silence contained in the signal. In order to address this issue, we investigated methods for identifying segments within the EMG signals that yield optimal recognition results when used for training and classification instead of the original signals. Section 6.5.1 introduces the segmentation methods we examined. The results are given in section 6.5.2.

6.5.1 Segmentation Methods

We investigated several different methods for segmenting EMG signals. Some of them rely on additionally recorded audio data.

No Segmentation (with extra silence): the complete signal is used for training and classification. Pure samples of silence are additionally recorded so that the speech recognizer can automatically learn to assign the "speech segment" to the spoken word and recognize the "no speech segment" as silence.

EMG Speech Alignment: An EMG *speech* recognizer is trained on all utterances of the training set. It receives the training signals labelled with "*silence speech silence*" during training so that it learns to distinguish between speech and silence. This recognizer is then used to find the speech segments in all utterances. The isolated *word* recognizer performs training and recognition on the segments identified as speech.

Audio Speech Alignment: The acoustic signal is recorded along with the EMG signal. An audio speech recognizer is then used to find a forced alignment for all utterances using the label "*silence speech silence*" . The EMG segments corresponding to the "speech segments" of the audio signal are used for training and testing the EMG recognizer.

Audio Word Alignment: The acoustic signal is recorded along with the EMG signal. An audio speech recognizer is used to find a forced alignment for all utterances using the label "*silence [word] silence*". The EMG segments corresponding to the audio [word] segments are used for training and testing the EMG recognizer.

Others: We additionally explored several other segmentation methods. Energy threshold based methods, for example, performed poorly. We also examined the segmentation method introduced by [Chan et al., 2002b] who used the beginning of an additionally recorded audio signal as a trigger and cut out an EMG signal block of a fixed size beginning 500ms before the beginning of the audio speech signal. However, we could not find an advantage of this method compared to our "audio speech alignment" approach which performed better.

It should be pointed out that there are three major drawbacks of audio based segmentation methods: firstly, they only work for audible speech. Secondly, they require synchronization of the audio channel and the EMG channels. Finally, muscle activation generally can prior to and after sound generation as shown in figure 6.10. Consequently, it is necessary to examine whether an extension of the speech segment to the left and/or the right side improves performance. We address this issue in the following section 6.5.2.

Table 6.25 summarizes advantages and disadvantages of all segmentations methods.

Figure 6.11 shows three EMG signals of the word "accept" along with the signals resulting from different segmentation methods. It is noticeable that the audio *speech* recognizer yields longer segments than the audio *word* recognizer. Visual inspection of the signals indicated that the audio *word* recognizer generally detects the speech segment in the *acoustic* signal accurately. Yet, it can be seen that EMG activity occurs up to about 200ms prior to sound generation in the sample utterances. Despite the fact that the EMG speech detector seems to cut out the best segment our recognition results presented in the next section suggest that audio segmentation based recognizers perform better.

Segmentation Method	**Advantages**	**Disadvantages**
No Segmentation (with extra silence)	easy	learns silence as part of words; worst recognition results
EMG Speech Alignment	performs better than baseline method; requires less effort than audio based methods; works for non-audible speech	requires the training of two recognizers for each training set
Audio Speech Alignment	yields good recognition results; knowledge on audio data usable for analysis of EMG signals	only works for audible speech; requires synchronization; requires audio recognizer
Audio Word Alignment	yields best recognition results ; knowledge on audio data usable for analysis of EMG signals	requires knowledge of which word was actually spoken; only works for audible speech; requires synchronization; requires audio recognizer

Table 6.25: Advantages and disadvantages of segmentation methods.

6.5.2 Results

The recognition results for the different segmentation methods on Set I of the PHONE corpus are presented in table 6.26. The baseline method BASE differs from the segmentation method *No Segmentation* in that the latter was trained with additional pure silence signals - not merely with the vocabulary words.

The results show that all our segmentation methods lead to a performance improvement compared to the baseline method. The audio based segmentation methods perform particularly well. It is worth mentioning here that the signals in the PHONE corpus contained a relatively large amount of

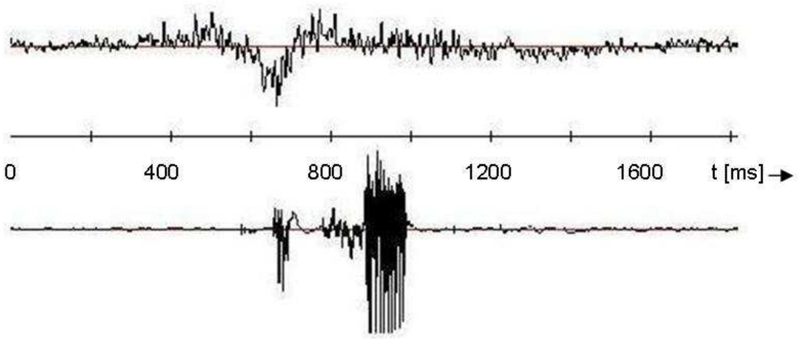


Figure 6.10: EMG and audio signal for the word accept (speaker S0, session 009, EMG1). Muscle activity occurs approximately 200ms prior to sound generation.

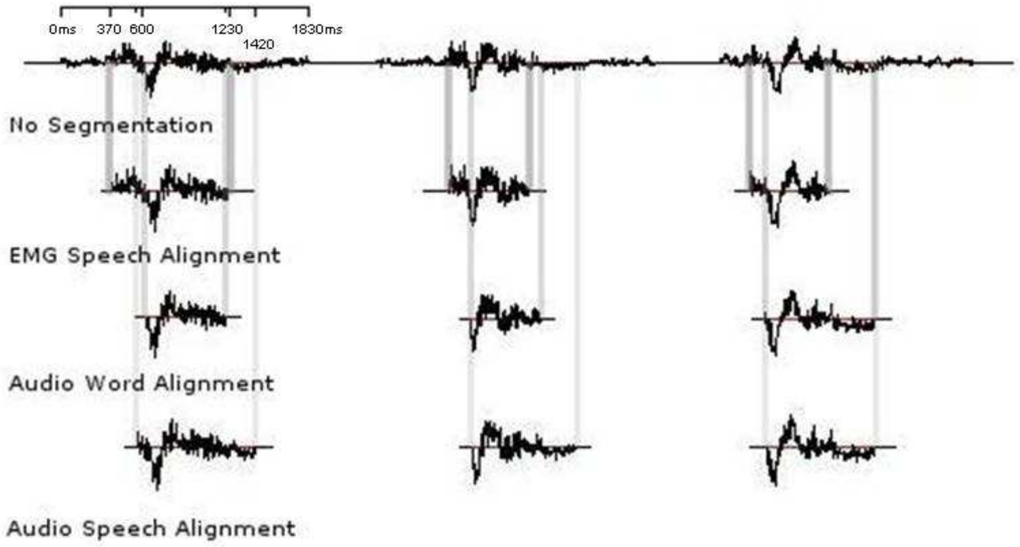


Figure 6.11: Signals resulting from different segmentation methods applied to three sample signals of the word three (speaker S0, session 009, EMG1).

Segmentation Method	S0	S1	Average
BASE	89.3	87.5	88.4
No Segmentation (with extra silence)	91.6	87.9	89.7
EMG speech alignment	92.8	90.5	91.6
Audio Word Alignment	96.5	94.2	95.4
Audio Speech Alignment	96.1	91.7	93.9

Table 6.26: Within-session word accuracies (in %) for diffferent segmentation methods on Set I of the PHONE corpus.

silence (more silence than speech). Experiments on the DIGITS_VIII corpus indicate that the different segmentation methods perform similarly when less silence is contained in the signals.

Table 6.27 shows recognition results for the method *Audio Speech Alignment* for different extensions of the speech segment to the left and right. Maximal recognition rates are achieved for a 200ms extension of the speech segment to the right. It is worth mentioning here, that Chan et al. discovered speech activity up to 500ms prior to sound generation [Chan et al., 2002b].

Extension to left [ms]	Extension to right [ms]	S0	S1	Average
0	0	96.1	91.7	93.9
100	0	96.4	92.3	94.4
200	0	96.3	93.0	**94.7**
300	0	95.9	92.9	94.4
400	0	95.5	89.6	92.6
500	0	95.0	88.6	91.8
600	0	93.0	89.1	91.1
0	100	96.6	93.3	95.0
0	200	97.0	92.8	94.9
0	300	95.7	93.3	94.5
0	400	96.3	92.9	94.6
0	500	95.2	92.0	93.6

Table 6.27: Within-session word accuracies (in %) for different extensions of the audio segmented EMG signal to the right and to the left on Set I of the PHONE corpus.

In conclusion, we have developed a state-of-the-art EMG speech recognition system by investigating various electrode positions, features, HMM topologies and segmentation methods. In the following section, we address the issue of session independence as an attempt to improve practicability of the system.

6.6 Session Independence

As already mentioned above the signal obtained from surface EMG measurements depends on a number of different factors which cannot be held constant over several recording sessions. Exact electrode positioning plays a particularly crucial role [Lamb and Hobart, 1992]. Although gypsum masks were used to increase placement repeatability, the poor across-sessions results indicate existing variation in the positioning. In fact, experiments showed an across-sessions deviation of up to 5mm (6.2.2). Furthermore, other factors like the amount of applied electrode gel may vary from session to session. Moreover, the speakers' speech patterns produced on different days may differ from each other. Speaker S7, for example, stated that he had the impression that he pronounced the non-audibly spoken words differently in different recording sessions.

Table 6.28 shows the within-session and naive across-sessions results for speaker S7. Naive across-sessions testing refers to testing without any normalizations and adaptations. The large performance differences between within-session results (values on the diagonal in bold face) and across-sessions results (values in the remaining cells) illustrate the problem of session dependence.

	session 004	session 008	session 010	session 011
session 004	**94.5**	74.3	83.0	58.8
session 008	67.5	**93.5**	80.5	73.8
session 010	48.8	59.5	**97.5**	77.8
session 011	60.5	67.0	91.8	**98.5**

Table 6.28: Word accuracies (in %) for within-session testing and naive (no normalization) across-sessions testing on Set I of the DIGITS_VIII corpus for speaker S7 using all seven channels. Train on (row), test on (column)..

The results for naive across-sessions testing for all speakers are summarized in Tables 6.29 and 6.30 for all channels and for individual channels respectively (method=$BASE$). The numbers represent the average word accuracy when one session is used for training and one session is used for testing. Thus, in Table 6.29 each cell corresponding to method $BASE$ represents the results for $sN \cdot (sN-1) = 4 \cdot 3 = 12$ experiments (sN: number of sessions). In Table 6.30 on the other hand, the entries represent the results for $cN \cdot sN \cdot (sN-1) = 7 \cdot 4 \cdot 3 = 84$ experiments, where cN represents the number of channels. Again, the results for across-sessions testing are significantly worse than those for within-session testing. We address this crucial problem of session dependence in the next section and will show that we achieve significant improvement across sessions by normalizing data and adapting our models.

6.6.1 Normalization and Adaptation Methods

We investigated the following normalization and adaptation procedures to compensate for the described session dependent variations:

1. *Session Combination (SC)*: The data to train the classifiers is shared across three sessions, each contributing the same number of samples (ten samples per vocabulary word)

2. *Session Selection (SS)*: A conventional HMM classifier C_i is trained for every training session i. The incoming unknown signal is then decoded by each classifier C_i, giving a hypothesis W_i and a corresponding Viterbi score v_i. The word with the overall best viterbi score is output as the hypothesis. $W_{hyp} = W_l$; $l = \arg\max_n v_n$

3. *Variance Normalization in combination with SC (SC&VN)*: For each training session two normalization vectors are computed; one containing the mean of each feature vector coefficient for the session's training samples and one containing the variance of each feature vector coefficient. Similarly, two normalization vectors are computed for all test session data. Prior to Viterbi path computation during training or testing, the obtained vectors are applied to normalize the extracted feature vectors o_i.

4. *Variance Normalization with enrollment data and SC (SC&VN_enr)*: Similar to *SC&VN* but the normalization vectors for the test session are computed on enrollment data rather than on the test data itself. The enrollment data set consisted of two examples for each vocabulary word including silence.

5. *Supervised Feature Space Adaptation and SC (SC&FSA_sup)*: Feature Space Adaptation is a constrained Maximum Likelihood (ML) transformation of input features. In analogy to Speaker Adaptive Training (SAT) [Jin et al., 1998] we perform *session adaptive training*. First, an initial classifier is computed on three training sessions. Then, we iteratively (a) adapt each training session to the current classifier (beginning with the initial classifier) and (b) recompute the classifier models using the adapted training data. After four iterations, the final classifier is used for a supervised computation of an adaptation matrix for the test data. During testing, only adapted test data is used.

6. *Unsupervised Feature Space Adaptation and SC (SC&FSA_unsup)*: Like *SC&FSA_sup* but unsupervised adaptation is performed on the test data using hypothesis from the computed classifier.

7. *Feature Space Adaptation with enrollment data and SC (SC&FSA_enr)*: Like *SC&FSA_sup* but the adaptation matrix is computed on an enrollment data set consisting of twenty-two signals (*as in SC&VN_enr*).

8. *Feature Space Adaptation with enrollment data, iterative learning and SC (SC&FSA_enr_it)*: Like *SC&FSA_enr* but the adaptation matrix for the test data is recomputed after each hypothesis computation for a test signal.

9. *Combinations of the above methods*: When both, *VN* and *FSA* are applied, the features are first normalized and then adapted to the model.

Method	**S1**	**S3**	**S7**	**Average**
BASE	74.5	83.7	70.3	76.2
SC	84.6	90.1	77.6	84.1
SS	85.2	88.3	77.3	83.7
SC&VN	83.4	94.3	83.7	87.1
SC&VN_enr	84.3	90.3	79.6	84.7

Table 6.29: Word accuracies (in %) for across-sessions testing on Set I of the DIGITS_VIII corpus using all channels for recognition.

6.6.2 Results

The baseline system (section 6.1) (extended by the respective normalization and adaptation methods) and Set I of the DIGITS_VIII corpus (3 speakers, 4 non-audible sessions each) were chosen for experiments on session independence. It should be pointed out that all recognizers referred to in this section were trained with thirty samples per word in the vocabulary. Refer to section 4.3.3 for a general description on within-session and across-sessions offline testing.

We examined both, across-sessions recognition using all seven channels (table 6.29) and across-sessions recognition using only one channel (table 6.30). In the latter case, the word accuracies for the individual channels were averaged. Due to the fact that FSA computations led to numerical instabilities when high-dimensional data was used (seven channels correspond to 126 dimensions), we did not apply feature space adaptation based methods when using all seven channels for recognition. Initial experiments using an LDA for dimensionality reduction decreased word accuracies.

As shown in Tables 6.29 and 6.30, normalization and adaptation improve performance for all speakers. In fact, the χ^2-test confirms that the results for *BASE* and *SC* are different at a significance level of 2.93E-20%. The additional application of *VN* leads to another increment on a significance level of 2.84E-03%.

As in ASR, combining data from several sessions improves performance considerably (session combination *SC*). Session Selection (*SS*) leads to significant improvements in performance as well. However, this method requires three times as much training material and the training of three times as many parameters. Consequently, *SS* is not directly comparable to the other methods. In fact, we obtained an improvement of 1.9% (1.5% absolute) for all channels and and 4.6% (2.2% absolute) for individual channels when we used the same amount of training material for combination (*SC*) as for selection *SS* (thirty samples per word from each session). We therefore did not combine *SS* with *VN* and *FSA*. Experiments suggest, however, that a similar increase in word accuracy as with *SC* can be achieved.

Both tables show a significant improvement in word accuracy when Variance Normalization (*VN*) is applied. However, the method fails to increase word accuracies for speaker S1. We attribute this to large deviations in recording lengths for speaker S1 which leads to significant deviations in the amount of silence relative to the amount of speech in different recording sessions. This in turn leads to a unreliable estimation of the *VN* normalization vector.

Method	**S1**	**S3**	**S7**	**Average**
BASE	37.0	53.5	41.3	43.9
SC	40.3	59.3	44.2	47.9
SS	43.4	61.4	48.6	51.1
SC&FSA_sup	42.5	62.7	47.7	51.0
SC&FSA_unsup	42.0	62.3	47.0	50.5
SC&FSA_enr	42.3	62.5	47.1	50.6
SC&FSA_enr_it	42.1	62.5	47.2	50.6
SC&VN	40.2	61.6	47.1	49.6
SC&VN_enr	38.8	60.5	45.5	48.3
SC&VN&FSA_sup	42.6	65.0	49.9	52.5
SC&VN&FSA_unsup	42.0	64.6	49.5	52.0
SC&VN_enr&FSA_enr	41.2	63.7	48.2	51.0
SC&VN_enr&FSA_enr_it	41.3	64.1	48.5	51.3

Table 6.30: Word accuracies (in %) for across-sessions testing on Set I of the DIGITS_VIII corpus using one channel for recognition. Each cell represents the average over all seven channels.

Feature Space Adaptation based methods increase the performance for all speakers. Interestingly, supervised adaptation performs equally well as unsupervised adaptation. Combining *FSA* and *VN* leads to further improvements, yet the improvements are not additive, i.e. both methods address similar artifacts. In order to apply FSA based methods when several channels are used for recognition, we suggest to explore feature dimensionality reduction techniques for EMG speech data in the future. Both, *FSA_unsup* and *VN* require the whole set of test data for initial computations. Obviously, this is impractical. We therefore examined the use of enrollment data for the computation of normalization vectors and adaptation matrices. According to Table 6.30 only a small decrease in word accuracy occurs when enrollment data is used. However, *VN_enr* performs significantly worse than *VN* when all channels are used for recognition. Unfortunately, this cannot be explained satisfyingly by the current experiments.

In conclusion, we were able to improve word accuracies for across-sessions testing by 18.5% (8.1% absolute) for individual channels and by 14.3% (10.9% absolute) by sharing training data across sessions and by applying methods based on Variance Normalization and Feature Space Adaptation. This indicates, that conventional speech recognition methods can be transferred to EMG based recognition systems and achieve comparable word error rate reductions.

Remarks

We conducted a large number of additional experiments on session independence that have not yet been mentioned in order to facilitate reading of this thesis. Some observations, however, deserve mentioning at this point:

1. Recognition results improved significantly when signal data from the test session was included in training. 92% word accuracy was achieved when training was performed on four (instead of three) sessions including the test session, each contributing (approximately) the same number of utterances per word (seven or eight). Due to the fact that online training is far too time-consuming when it comes down to large vocabulary tasks this results is irrelevant for practical applications.

2. We also examined training a speech recognizer based on session labels: For each training session a complete speech recognizer was trained which was used for labelling all signals in the corresponding session. We then combined the method *Session Combination (SC)* with training along the session dependent labels. This approach yielded slight improvements compared to SC, yet, we did not investigate it any further because it requires three times as much training material as SC. For a general explanation on training along labels refer to [Metze, 2004].

3. Several normalizations in the time domain have been examined:

   (a) Maximum Voluntary Force: The speaker tried to generate maximum muscular force in a reference contraction . All time domain values were then normalized to the average of the $n = 1, 5, 10, 20$ maximal absolute values in that reference contraction. Yet, production of maximal force with facial muscles proved to be extremely difficult.

   (b) Normalization to Maximum Value: The $n = 1, 5, 10, 20$ maximal absolute values within an utterance were determined and all values were normalized to the average of these values.

   (c) Variance Normalization: Mean Subtraction and Variance Normalization was performed on the time domain values prior to STFT computation

   Due to the fact that these methods performed extremely poorly (across-session recognition rates were significantly worse than naive across-sessions results without normalizations), we did not investigate them further.

4. When segmentation methods were applied (refer to section 6.5) STFT performed better when the normalization vectors were computed on the signal segments rather than on the complete utterances.

5. When the normalization vectors were merely computed on the silence utterances performance dropped significantly.

By presenting methods for improving across-sessions recognition results we have moved one step towards a more practical EMG based speech recognition system. In the following section we investigate differences of non-audible and audible.

6.7 Analysis of Audible and Non-Audible Speech

To investigate the influence of speech manner (audible vs. non-audible) on the performance of EMG based speech recognition, we recorded one audible and one non-audible session for each speaker. These two sessions were in fact recorded as one session with the exact same electrode placement, i.e. the electrodes were not removed between the two parts. The only difference was the speech manner. We investigated the following aspects:

1. do the EMG signals produced by audible speech differ from those produced by non-audible speech?

2. is the recognition performance of audible speech different from that of non-audible speech?

3. is it possible to train a speech recognizer that works accurately on both speech manners?

To investigate the first aspect we determined the recognition results across speech manners, i.e. models trained on audible speech were applied to non-audible speech and vice versa. To examine the second issue we compared the recognition results between the two speech manners in a matching condition, i.e. the models were trained and tested on the same speech manner. In a third experiment, we shared the training data across speech manners from each speaker to determine the performance of a recognizer that works on both, non-audible and audible speech. In the latter case we trained two systems; one with the same number of parameters as our baseline system and one with twice as many parameters.

The results of our experiments are shown in Table 6.31 for all channels and in Table 6.32 for individual channels respectively. It is noticeable that speakers S1 and S7 have much better recognition rates for audible speech than for non-audible speech. By contrast, there is no significant difference in performance for speaker S3. We believe that this relies to the fact that speaker S3 had the most experience in speaking non-audibly. As alluded to in section 6.2 we noticed an improvement in performance with increasing experience for all speakers. We deduce from this, that MES based recognition of non-audible speech can work just as well as MES based recognition of audible speech on our vocabulary provided that the speaker is accustomed to the speaking manner. Note that section 6.2 gives more information on the suitability of *individual* electrode positions for non-audible and audible speech.

The relatively low results in the mismatched condition indicate that muscle movements corresponding to audible speech differ from muscle movements corresponding to non-audible speech. However, the results for the mixed systems indicate that a recognizer can be trained for both, audible and non-audible speech, with reasonable results. The comparison of the 12-Gaussian vs. the 24-Gaussian systems suggests to increase the numbers of parameters for the mixed system.

Speech Manner	**S1**	**S3**	**S7**	**Average**
non-audible	97.0	99.8	93.5	96.8
audible	99.5	98.8	96.0	98.1
audible on non-audible	72.8	84.5	64.3	73.8
non-audible on audible	67.2	92.5	69.3	76.3
mixed; 12 Gaussians	96.1	98.1	91.8	95.3
mixed; 24 Gaussians	96.1	98.4	93.5	96.0

Table 6.31: Word Accuracies (in %) of non-audible and audible speech on Set II of the DIGITS_VIII corpus using all seven channels and the baseline system introduced in section 6.1.

Speech Manner	**S1**	**S3**	**S7**	**Average**
non-audible	63.0	83.4	60.0	68.8
audible	73.9	84.7	70.3	77.5
audible on non-audible	43.3	59.4	39.2	47.3
non-audible on audible	39.0	60.9	32.7	44.2
mixed; 12 Gaussians	62.6	79.3	57.3	66.4
mixed; 24 Gaussians	64.7	81.1	59.7	68.5

Table 6.32: Word Accuracies (in %) for non-audible and audible speech on Set II of the DIGITS_VIII corpus using one channel for recognition and the baseline system introduced in section 6.1. Each entry represents the average over all seven channels.

In order to investigate the suitability of non-audible speech for any given recognition task (not just for discrete word recognition) it is necessary to compare recognition results for non-audible and audible speech for smaller units than words, e.g. for phonemes. In the next section we present initial experiments on connected digits recognition and on phoneme based recognition approaches.

6.8 Towards Continuous Speech Recognition

In the previous sections we have demonstrated the applicability of surface electromyography for discrete speech recognition. Yet, isolated word recognition is only suitable in a limited number of situations. In order to move towards continuous speech recognition on large vocabularies we performed experiments on *connected* digits recognition (section 6.8.1) and examined phoneme-model based approaches for recognizing words and phrases (section 6.8.2).

6.8.1 Connected Digits Recognition

In this section we will show that context dependency and segmentation are among the biggest challenges for connected words recognition.

Context Dependency

In order to investigate the influence of context dependency on connected words recognition we trained a first recognizer on *isolated* words and a second recognizer on *connected* words and compared the performance of the two systems on *connected* words. If context dependency between words was not an issue the system trained on isolated words should not perform worse than the system trained on connected words.
The following experiment was conducted:

1. Two sessions (Set III of the DIGITS_II corpus) were recorded in series without the electrodes being removed between the two sessions. In the first session, we recorded *isolated* words from the extended digits vocabulary {zero, one, ..., nine, call, end}. In the second session, we recorded utterances of the form "call [digit] end".

2. Context independent models for each word of the extended vocabulary were trained on the *first* session, i.e. on the isolated words, using forty samples per word. The resulting recognizer was used to decode the utterances from the *second* session. The set of possible hypothesis was restricted to utterances of the form "call [digit] end" by the language model (a grammar) and the number of correctly recognized digits relative to the total number of utterances was defined as the word accuracy.

3. The second session was split into a set of five disjoint sets and the round robin algorithm was applied to train context independent models for each word in the extended vocabulary (training and testing were thus both performed on connected words). Forty samples of each digit were used for training each recognizer. It is worth mentioning here, that the left and right contexts of each digit were the same in all utterances (call and end respectively). No *labels* were used to present an initial segmentation of the utterances during training, thus, the recognizers had to learn the word boundaries by themselves. Word accuracies were determined as in 2.

Table 6.34 shows the results of our experiments.

	Isolated Digit Session	Connected Digits Session
Isolated Digit Session	82.4	**53.0**
Connected Digits Session	54.6	**78.0**

Table 6.33: Within-session and across-sessions results (in %) on Set III of the DIGITS_II corpus (1 speaker, 2 sessions). Train on (row), test on (column). The recognition results printed in bold-face illustrate the problem of context dependency.

It can be seen that the results for connected words recognition are significantly higher when training is performed on connected words rather than on isolated words, even though training on connected words was performed without labels (i.e. without given word boundaries). We deduce from this that context dependency is an important issue in EMG based speech recognition just like in conventional speech recognition systems.

Segmentation

As already mentioned above the task of segmentation is to find the boundaries of words within an utterance. In order to demonstrate that segmentation is a particularly important issue in EMG based speech recognition we compared connected digits recognition results of (1) a standard system and (2) a system whose models were initialized with the help of an audio recognizer such that an initial segmentation was indirectly presented to the recognizer.
For this purpose we recorded Set I of the DIGITS_VIII_CON corpus, which consists of two sessions from two different speakers. In each session utterances of the form "*silence* [6 · [digit]] *silence*" were recorded such that the complete set of utterances could be divided into two disjoint sets each

containing each digit *tripel* at least once. Context independent models were then trained for each digit in two different ways:

1. Standard: Training was performed as in the baseline system but with the two sets introduced in the previous paragraph used for the round robin algorithm.

2. Audio Label Initialization: Like "Standard" but an audio based initialization of the codebooks was choosen:

   (a) An audio recognizer was used to determine a *forced alignment* (i.e. a segmentation) for all utterances in the training data set. According to the word boundaries identified by the audio recognizer the complete sequence of feature vectors extracted from the myoelectric signals of a given utterance was split into a set of segments each of which was associated with either one of the six spoken digits or with *silence*. It should be pointed out that the audio recognizer was configured to allow *optional silence* in between words.

   (b) For each digit all sequences of feature vectors that were assigned to that digit by the audio recognizer were split into $n_s = 5$ parts of equal length (n_s: number of HMM states for each word) and the ith part was assigned to the ith state of the word model of that digit.

   (c) The sequences of feature vectors corresponding to the *silence* at the beginning and the end of each utterances were assigned to the silence model.

   (d) segments corresponding to *silence* between individual digits were not included in the initialization procedure because experiments suggested to omit them. We attribute this to the fact that the EMG signal and the audio signal are not perfectly synchronized because muscle activity generally occurs prior to sound generation (section 6.5).

   The k-means algorithm was then applied to every state of every word model using the previously assigned feature vectors just like in the baseline system (section 6.1.3).

The recognition results for the two different initialization methods are shown in table 6.34. The space of possible hypothesis (defined by the language model) consisted of all sequences of six connected digits. We used a standard definition of the word error rate (WER):

$$WER = \frac{\#substitutions + \#deletions + \#insertions}{\#words}$$

The word accuracy was defined as $1 - WER$. Refer to [Rogina, 2003] for details.
From the fact that audio label initialization leads to significantly better results than the baseline method we deduce that model initialization and thus segmentation in general are important issues in connected words recognition. Moreover, experiments presented in section 6.9 indicate that segmentation is a greater challenge in EMG based speech recognition systems than in conventional ASR systems.

Method	**Recognition**
Standard	34.7
Audio label initialization	43.1

Table 6.34: Within-session results (in %) on Set I of the DIGITS_VIII_CON corpus for different model initialization methods.

It is worth mentioning here, that significantly better connected words recognition results were obtained on Set I of the COMMANDS corpus (1 speaker, 1 audible session). Training and testing were performed on word *tripels* (each possible tripel was seen in training) and recognition results of 97% were achieved. These results are encouraging, yet, it should be pointed out that only two boundaries between words had to be found instead of five so that the segmentation task was not as difficult as in the experiments described above. Moreover, the commands vocabulary consists of only six words instead of ten.

We assume that automatic segmentation is particularly difficult for *context independent* models. However, it is infeasible to train context *dependent* models for every possible word tripel in a large vocabulary. Consequently the use of smaller units than words for EMG based speech recognition must be explored. The following section introduces initial experiments on phoneme based speech recognition.

6.8.2 Phoneme Models

When performing continuous speech recognition it is infeasible to train a Hidden Markov Model for every possible utterance. In order to minimize the amount of training data phoneme-models are most commonly used in automatic speech recognition (section 2.4). In *context-independent* speech recognition, one HMM is trained for each phoneme. Words are then composed of phonemes and sentences are composed of words.

For experimental purposes we trained phoneme models for isolated digit recognition. The phonemes corresponding to individual digits are presented in table 6.35.

We chose Set II of the DIGITS_VIII corpus (3 speakers, 1 audible and 1 non-audible session each) for these experiments in order to determine differences in performance between non-audible and audible speech. Within-session recognition results for both, word models and phoneme models are shown in table 6.36. Two states per phoneme were used in both cases. The recognizers' only difference to the baseline system (section 6.1) was the changed topology.

Word	**Phonemes**
one	W AH **N**
two	**T** UW
three	TH **R** IY
four	**F** AO **R**
five	**F** **AY** **V**
six	**S** **IH** K **S**
seven	**S** EH **V** AX **N**
eight	EY **T**
nine	**N** **AY** **N**
zero	Z **IH** **R** OW

Table 6.35: Decomposition of digits into phonemes. Phonemes that occur more than once are printed in bold-face.

Model	**S1 a.**	**S1 n-a.**	**S3 a.**	**S3 n-a.**	**S7 a.**	**S7 n-a.**	**Avg a.**	**Avg n-a.**
word	98.5	96.8	98.8	99.8	97.5	94.5	98.3	97.0
phoneme	95.5	93.3	98.3	94.8	87.3	77.3	93.7	88.4
Δ	3.0	3.5	0.5	5.0	10.3	17.3	**4.6**	**8.6**

Table 6.36: Within-session word accuracies (in %) for word models and phoneme models on Set II of the DIGITS_VIII corpus. Results are shown for one audible (a.) and one non-audible (n-a.) session for each speaker. The Δ row presents the difference of row 2 (word models) and row 3 (phoneme models).

It can be seen that performance differences between word models and phoneme models are significantly higher for non-audible speech than for audible speech (significance level of 1.3% according to the χ^2-test). This holds even for speaker S3 whose results for non-audible speech were originally better than those for audible speech (when word models were deployed). We deduce from this that phoneme-models may not be the optimal choice for non-audible speech recognition.

In order to investigate the issue in more depth we examined the word accuracies for *individual* words from the vocabulary for all speakers as shown in table 6.37. Moreover, confusion matrices for the phoneme model recognizers are presented in tables 6.39 and 6.38 for the audible and the non-audible sessions respectively.

Word	**Word models n-a.**	**Word models a.**	**Phoneme models n-a.**	**Phoneme models a.**	**Δ n-a.**	**Δ a.**
one	98.3	100.0	97.5	98.3	0.8	1.7
two	95.8	98.3	95.8	98.3	0.0	0.0
three	98.3	100.0	99.2	97.5	-0.8	2.5
four	100.0	100.0	89.2	99.2	**10.8**	**0.8**
five	97.5	100.0	89.2	95.0	8.3	5.0
six	95.8	96.7	87.5	90.8	8.3	5.8
seven	95.0	98.3	91.7	97.5	3.3	0.8
eight	92.5	95.8	50.0	70.8	**42.5**	**25.0**
nine	97.5	96.7	85.8	92.5	**11.7**	**4.2**
zero	99.2	96.7	98.3	96.7	0.8	0.0

Table 6.37: Within-session word accuracies (in %) for word models on the non-audible sessions (2nd column), word models on the audible sessions (3rd column), phoneme models on the non-audible sessions (4th column) and phoneme models on the audible sessions (5th column) on Set II of the DIGITS_VIII corpus. The last two columns present recognition differences between word and phoneme models for non-audible (Δ n-a.) and audible speech (Δ a.) respectively. All results are averaged over three speakers.

	one	two	three	four	five	six	seven	eight	nine	zero
one	117				1		1		1	
two		115		1				1		3
three			119				1			
four	2		1	107						**10**
five			1	7	107		1		1	3
six			3		1	105	5		1	5
seven			2			2	110			6
eight	5	2	10		4	6	11	60	15	7
nine	6		3		5		1		103	2
zero		1		1						118

Table 6.38: Confusion matrix for all non-audible sessions from Set II of the DIGITS_VIII corpus (i.e. one non-audible session per speaker) and phoneme models. Empty cells represent the value 0.

The following observations deserve mentioning:

1. Huge performance differences between word and phoneme models exist for several words (five, six, eight, nine) for both, audible and non-audible sessions while the performance of several words (two, zero) is almost unaffected by the model change. Yet, there seems to be no relationship between the change in performance of a particular digit and the number of phonemes that digit shares with other digits.

   Moreover, the confusion matrices show that the hypothesis of misclassified digits do not necessarily share phonemes with the actually spoken digits. The misclassified utterances corresponding to the digits "eight", for example are not usually classified as "two" which is the only digits sharing a phoneme with "eight".

   Also, the decrease in performance does not seem to be related to the length (number of phonemes) of a word, because the two shortest words "two" and "eight" show completely different behaviour.

	one	two	three	four	five	six	seven	eight	nine	zero
one	118				1				1	
two	1	118								1
three			117	2						1
four	1			119						
five		1	1	1	114		2		1	
six	1		4			109	3	1		2
seven							117			3
eight	2	6	3	1	2		9	85	8	4
nine	3		2		2		1		111	1
zero		1	2				1			116

Table 6.39: Confusion matrix for all audible sessions from Set II of the DIGITS_VIII corpus (i.e. one audible session per speaker) and phoneme models. Empty cells represent the value 0.

We assume that in many cases misclassification relies to the fact that shared phonemes learn a rather general EMG pattern which can be easily confused with any other phoneme.

2. Performance drops drastically in the non-audible session for several words while there is no significant difference in word accuracy for the audible sessions. The digit "four" is the most obvious example. The confusion matrices (tables 6.39 and 6.38) show that the non-audible utterances corresponding to "four" are often misclassified as "zero" while only one misclassification occurs for the *audibly* spoken "fours". This could possibly be traceable back to the fact that muscle movement corresponding to the phonemes OW and AO is similar when words are spoken non-audibly and that a rather general speech pattern has been learnt for the phoneme R.

We suggest to perform more extensive experiments on phoneme-model based approaches to investigate these issues in more detail. However, our results indicate that phoneme-models are not the optimal choice for EMG based *non-audible* speech recognition.

6.9 Comparison of Conventional Speech Recognition and EMG-Based Speech Recognition

We compared the performance of our EMG based speech recognition systems to conventional speech recognition systems by training audio based speech recognizers on the *acoustic signals* of sessions that had already been used for training *myoelectric signal* based recognizers. Set III of the DIGITS_VIII corpus (3 speakers, 1 session each, individual digits) was chosen for isolated digits recognition and Set I of the DIGITS_VIII_CON corpus (2 speakers, 1 session each, 6 connected digits) was selected for connected digits recognition.

We used the exact same settings as for the EMG systems that had previously been trained on these sets (section 6.7 and section 6.8.1), but replaced the frontend by a standard speech recognition frontend (Melscale-filterbank coefficients).

The word accuracies for both, the MES recognizers and the audio based recognizers are shown in table 6.40.

It can be seen that similar recognition results are achieved for *isolated word* recognition. We assume that this relies to the fact that the recognition task is too easy to make a useful comparison. The difference in recognition would probably be higher for a larger vocabulary.

Connected digits recognition is much more accurate for the acoustic signal than for the EMG signal. However, the EMG speech recognizer yields a significantly higher improvement for optimized model initialization than the audio speech recognizer. This could be traceable back to the fact that small segments of silence can be found in the acoustic signals in between words while we could not (visually) detect silence (i.e. muscle relaxation) between individual digits in the EMG signals. This observation is not surprising considering that X-ray films of the speech organs in action show that they are in

Recognition Task	**Test set**	**EMG Recognizer**	**Audio Recognizer**	**Δ**
isolated digit recognition	DIGITS_VIII, Set VI	98.1	99.9	1.8
six digits recognition, standard	DIGITS_VIII_CON, Set I	34.7	89.0	54.3
six digits recognition with audio label initialization	DIGITS_VIII_CON, Set I	43.1	87.9	44.9

Table 6.40: Comparison of the performance of speech recognizers based on myoelectric and acoustic signals respectively. Δ shows the difference.

continuous motion during the act of speaking [Stüker, 2003]. Consequently, it is more difficult to find an appropriate segmentation within the myoelectric signals than within the acoustic signals.

We deduce from our results that conventional speech recognizers outperform state-of-the-art EMG based speech recognizers provided that there is so ambient noise. On the other hand, Chan et al. have shown that EMG recognizers perform better than audio recognizers (at least for individual words) in noisy environments [Chan et al., 2002a] (chapter 3). Moreover, they can be used for recognizing non-audible speech.

In order to present applications for *EMG* based speech recognition we implemented two demo systems which are introduced in the following chapter.

Chapter 7

Demo Systems

The purpose of this chapter is to introduce the two demo systems we have implemented to show the potential of EMG based speech recognition. Section 7.1 introduces a prototype "silent" mobile phone suitable for conducting phone calls in situtations requiring silence, e.g. in a meeting. Section 7.2 presents a prototype EMG based lecture translation system.

7.1 Silent Mobile Phone

7.1.1 Motivation

One of the major advantages of EMG based speech recognition is the fact that it does not require the transmission of an acoustic signal. Consequently, the resulting speech is non-disturbing and allows confidential conversation with or through a device. One possible application of electromyography in speech is a "silent" mobile phone which can be deployed to conduct phone calls in situations requiring silence, for example in a meeting. The phone consists of a physiological data recording system for capturing EMG signals, an EMG based speech recognizer for translating EMG signals into speech, a speech synthesizer (text-to-speech) that converts the hypothesis into acoustic signals and transmits these signals to the conversational partner, and a receiver that converts incoming acoustic signals to text which appear on a display. Alternatively, acoustic signals could be received using headphones.
The following section introduces the prototype silent phone we have implemented.

7.1.2 Prototype

We have implemented a prototype silent mobile phone consisting of the following components:

- The physiological data recording system VARIOPORT_VIII introduced in section 4.1.1.
- An EMG speech recognizer trained on a set of sentences typically used for answering a phone call during a meeting, for instance "I'm in a meeting", "is it urgent?" and "I'll call back later".
- The speaker interface introduced in section 4.2.1 providing a push-to-talk button for recording an utterance to be transmitted and a display for showing the hypothesis obtained from the EMG speech recognizer.
- A text-to-speech module that displays the hypothesis as acoustic signals through loudspeakers.

Figure 7.1 illustrates the setup.
The EMG speech recognizer deployed by our prototype silent phone is identical to the baseline system introduced in section 6.1. The only difference is the fact that the smallest units in the vocabulary are complete sentences instead of words and that ten states (instead of five) are used per "word" (i.e. sentence) in the vocabulary.

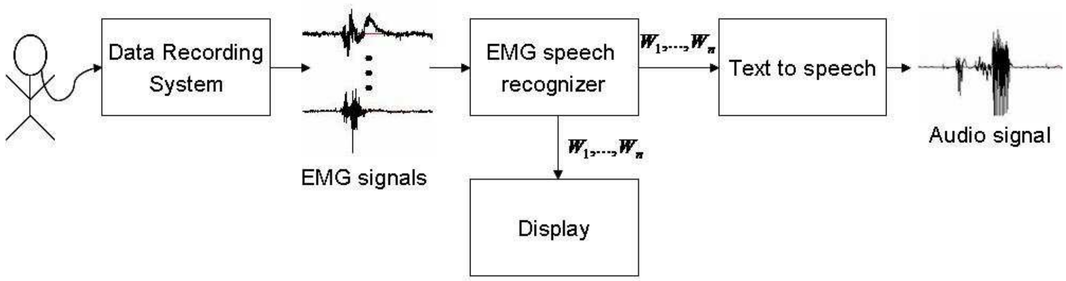


Figure 7.1: Schematic view of silent phone prototype

We created the following scenario to demonstrate our prototype system: P. is sitting in a meeting when he receives a phone call from L. He picks up his silent mobile phone to find out whether or not the call is important and conducts the following short dialog without producing a sound:

P: Hello?
L: Hi Peter, this is Lena. Do you have a minute?
P: I'm in a meeting.
P: Is it urgent?
L: Not really. We can talk about it another time.
P: I'll call back later.
L: Perfect! Talk to you later.
P: Talk to you later.'

As already mentioned above the hypothesis are output by a speech synthesizer in order to simulate what the conversational partner would hear.
The performance of the system is discussed in the following section.

7.1.3 Performance

In order to evaluate our online recognition system we determined online and offline recognition results on Set I of the MEETING corpus. Offline recognition refers to round-robin within-session testing, i.e. the round robin algorithm is applied to the respective recording session as already explained in section 4.3.3. For online recognition the system is trained on a *complete* recording session. Words are then randomly picked, and recorded by the speaker using the online recognition mode of our system (section 4.3.3).
Table 7.1 shows the results for offline and online testing on Set I of the MEETING corpus using six electrodes (EMG1-EMG6).

Recognition	**Session 000**	**Session 004**
offline	97.5	99.5
online	80.0	91.4

Table 7.1: Word Accuracies (in %) for online and offline recognition on Set I of the MEETING corpus (Speaker S5, two non-audible sessions) using six electrodes (EMG1-EMG6). 105 utterances (15 per phrase in the base vocabulary) were used for online testing.

The results show that offline classification generally performs better than online classification. It is worth mentioning here, that a classification error in the online mode increased the probability of the next utterance being misclassified. It is also noticeable that there was a significant improvement in recognition - especially for the online mode - in session 004 compared to 000 which is traceable back to increased experience of speaker S5. We observed increasing online recognition rates with increasing experience for speaker S3 as well.

7.2 Non-audible Speech Translator

7.2.1 Motivation

In the age of globalization, international conventions are part of everyday life for many people. Communication, however, requires a common language. While English can be spoken and understood by the majority of Europeans and thus serves a the common language in the western world, communication between certain nationalities is difficult (e.g. between Russians and Japanese people). In order to overcome this problem, professional interpreters are being employed who translate a conversation or a talk online. This is not only expensive but also quite inconvenient because it requires two people to speak at the same time. EMG based speech recognition suggests a solution to this problem: The idea is to produce a translation system that translates *non-audible* speech into a chosen language. That is, EMG signals resulting from non-audible speech in a certain language (e.g. English) are captured and recognized by an EMG based speech recognizer. The resulting hypothesis is then translated into another language and transformed into an acoustic signal by an appropriate speech synthesizer. The owner of such a translation system could thus give a non-audible talk in English that is translated directly into Chinese.
The following section introduces our prototype non-audible speech translator.

7.2.2 Prototype

We have implemented a prototype non-audible speech translator consisting of the following components:

- The physiological data recording system VARIOPORT_VIII introduced in section 4.1.1
- An EMG speech recognizer trained on a set of English sentences typically used by somebody who is giving a talk, e.g. “good morning ladies and gentlemen”, “my name is ...”, “any questions?”, “thank you for your attention”. The system’s only difference to our baseline system (section 6.1) is the fact that the smallest units in the vocabulary are complete sentences instead of words and that the states of a Hidden Markov Model were chosen to be proportional to the number of syllables in the correponding sentence (two states per syllable).
- two lookup tables for translating the given set of English sentences into German and Spanish respectively.
- Three speech synthesizers (text-to-speech) that can display English, German, and Spanish hypothesis respectively.

Figure 7.2 illustrates the setup.
In order to demonstrate the system the user can choose a language of his or her choice. He or she then conducts the following monologue which is translated directly into the chosen language and output by a speech synthesizer.

S: `Good afternoon, ladies and gentlemen!`
S: `Welcome to the Interact Center.`
S: `My name is Stan Jou.`
S: `Let me introduce our new prototype.`
S: *explains that the body of the talk would follow here*
S: `Any questions?`
S: `Thank you for your attention.`

The performance of the system is discussed in the following section.

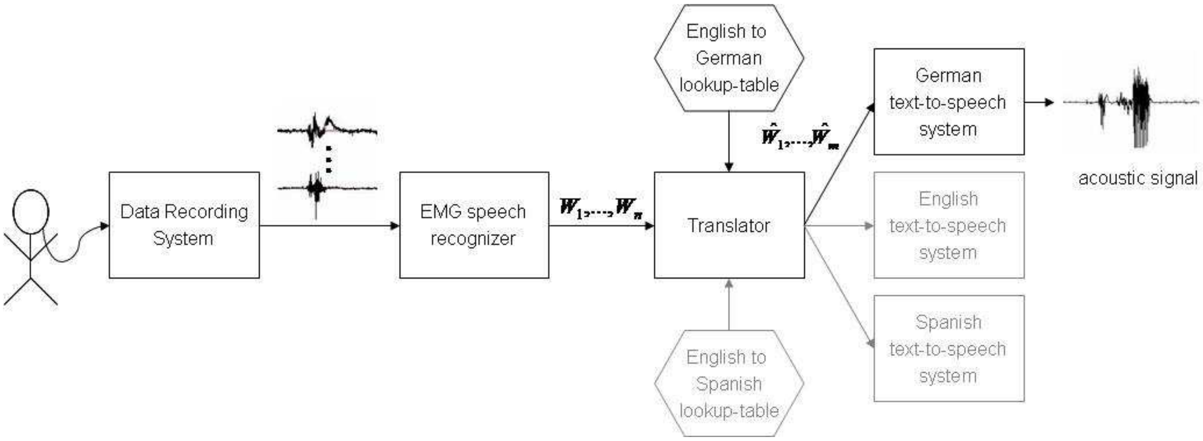

English to German lookup-table
German text-to-speech system
acoustic signal
$\hat{W}_1,\dots,\hat{W}_m$
Data Recording System
EMG speech recognizer
$W_1,\dots,W_n$
Translator
English text-to-speech system
Spanish text-to-speech system
English to Spanish lookup-table

Figure 7.2: Schematic view of our lecture translator prototype. The German translation is activated.

7.2.3 Performance

Offline classification results for the prototype translator ranged from 99.1% to 99.6% for the sessions of Set I of the LECTURE corpus. In an online experiments, all thirty-five randomly spoken sentences (five for each sentence in the base lecture vocabulary) were classified correctly (Set I of the LECTURE corpus). We attribute the better classification results compared to the MEETING demo to the greater lengths of the sentences in the vocabulary.

Chapter 8

Summary, Conclusions and Future Work

8.1 Summary and Conclusions

In this thesis we have presented a state-of-the-art isolated word EMG based speech recognizer that outperforms previously developed systems on the ten English digits vocabulary. Moreover, we dealt with issues that have not yet been addressed in the literature, namely with session dependence, the comparison of non-audible and audible speech, connected digits recognition and phoneme based approaches. In order to demonstrate the practicability of the new technology we have further presented two demo systems showing different applications of speech recognition based on myoelectric signals.

Our baseline isolated word recognition system was developed based on various experiments on electrode positioning, feature extraction and Hidden Markov classification. The system deploys seven surface electrodes placed in various positions on the face and the neck and yields within-session accuracies of 97.3% on the ten English digits vocabulary. Comparative experiments indicate that applying more than two electrodes is crucial, while using more than five electrodes does not lead to significant performance improvements.

To cope with the issue of session dependence, we have investigated a variety of signal normalization and model adaptation methods. Our results suggest that methods used in conventional speech recognition systems for channel and speaker adaptation can be used for session adaptation in EMG based speech recognizers. Sharing training data across sessions and applying methods based on Variance Normalization and Maximum Likelihood Adaptation improve across-sessions performance. We achieved an average word accuracy of 97.3% for within-session testing using seven EMG channels. Naive across-sessions testing - i.e. across-sessions testing without normalizations or adaptations - yielded an average of 76.2%. By applying our normalization and adaptation methods we were able to bring recognition rates back up to 87%. Gains were even higher, when a smaller number of channels were used.

Furthermore, our experiments indicate significant differences between the muscle movement corresponding to non-audible and the muscle movement corresponding to audible speech. Despite this, it is possible to merge training data to obtain a recognizer that deals accurately with both speech manners.

In order to move towards continuous speech recognition we have performed experiments on connected digits recognition and on phoneme based recognition. The results indicate that segmentation and context dependency are major issues in EMG based continuous speech recognition and that classical phoneme models are not the appropriate choice for the recognition of non-audible speech.

Furthermore, we compared the performance of conventional speech recognition systems to the performance of EMG based speech recognition systems in environments with a minimum of ambient noise. While recognition results on isolated words are similar, the conventional speech recognizer performs significantly better on connected digits than the EMG based recognizer. We attibute this to greater

context dependencies between individual words in the myoelectric signals which makes segmentation for the MES recognizer more difficult.
In order to demonstrate the practicability of EMG based speech recognition two demo systems have been implemented showing possible applications of the new technology: a prototype “silent” mobile phone suitable for conducting non-disturbing phone calls in situations requiring silence (e.g. in a meeting) and a prototype lecture translation system that can be applied for an online-translation of a non-audibly spoken talk into a language of one’s choice. Both systems were trained on a set of *sentences* typical for the respective application. Hypothesis are output by speech synthezisers in the selected language.
The presented results are very promising but several limitations still need to be overcome. The following section presents our suggestions for future work.

8.2 Future Work

One of the main challenges of EMG based speech recognition is to move beyond discrete speech recognition and approach continuously spoken large vocabulary tasks. Conventional speech recognition system use phonemes as basic units, yet, our results indicate that classical phonemes are not the optimal basic units for speech recognizers based on myoelectric signals. Several phonemes, for example, are exclusively distinguished by the fact that they are voiced or unvoiced (i.e. if there is a vibration of the vocal cords or not). When non-audible speech is deployed, however, there is never a vibration of the vocal cords. Consequently, we suggest to identify a set of speech units that can be characterized by features detectable in *electromyographic* signals. Whether or not myoelectric signals resulting from non-audible speech are rich enough to disambiguate all words and handle the full richness of a given language remains to be seen.
Another challenge associated with EMG based speech recognition is further the development of more user-friendly data recording systems. The use of sensors that do not require physical contact with the skin and could possibly be integrated in the collar of a shirt is being investigated by Jorgensen et al. [Jorgensen and Binsted, 2005]. The finger electrodes introduced by [Manabe and Z.Zhang, 2004] are also considerably more user-friendly than classical surface EMG electrodes and deserve further research activity.
Finally, the issue of speaker independency must be addressed. This goal requires reliable electrode placement across speakers as well as the investigation of normalization and adaptation methods to compensate for variation in speaking style and speaking rate. It is worth mentioning, however, that possible applications of EMG based speech recognizers focus on personal devices so that speaker independence is not a major issue.

Appendix A

VC++ Control Component

The control component of our VC++ software determines the general workflow. When the PC is connected to a physiological data recording system, the software can communicate with that system, initiate/stop data recording, and receive data.
The software can be in one of three states:

1. *Raw*: the raw mode was designed for experimental purposes. The software can communicate with the recorder and visualize incoming data without storing it.

2. *Data Collection*: The data collection mode was designed for producing complete *recording sessions*. Section 4.3 describes the corresponding workflow.

3. *Recognition*: In the recognition mode EMG signals are recorded and classified by a selected EMG speech recognizer. Section 4.3 describes the corresponding workflow.

In each mode, various settings can be chosen. They include

- Channel selection: A subset of all channels provided by the physiological data collection system can be selected. Only the values corresponding to the selected channels are transmitted over the serial port.

- Sampling rate: A sampling rate for the physiological signals can be chosen.

- Word list: In the data collection mode a word list can be selected, consisting of the utterances a speaker has to record in the corresponding session. The words can either be presented to the speaker in the same order as in the word list file or be randomized.

- Visualization: The incoming data can optionally be visualized.

- Audio data: An additional acoustic channel can optionally be recorded. Sampling rate and resolution can be selected.

- Synchronization: The marker channel can optionally be deployed for synchronizing the acoustic channel with the remaining channels: The PC is connected to the marker input of the recorder. When an audio recording is started the values of the marker channel are changed from 0 to a value $! = 0$ by the software.

- Translation: In the recognition mode it is possible to select a language to which the hypothesis is translated to.

For a more detailed description refer to [Mayer, 2005].

Appendix B

Data Collection Procedure

All signal data used for our experiments was collected in so-called *recording sessions*. A recording session is defined as a set of utterances collected in series by one particular speaker. All settings (channels, sampling rate, speech mode) remain constant during a session. Data collection consists of the following steps:

1. Application of electrodes: The surface electrodes are positioned on the user's face (e.g. using tape measure or gypsum masks). Individual electrodes must be removed and re-applied until all channels yield good-looking signals.

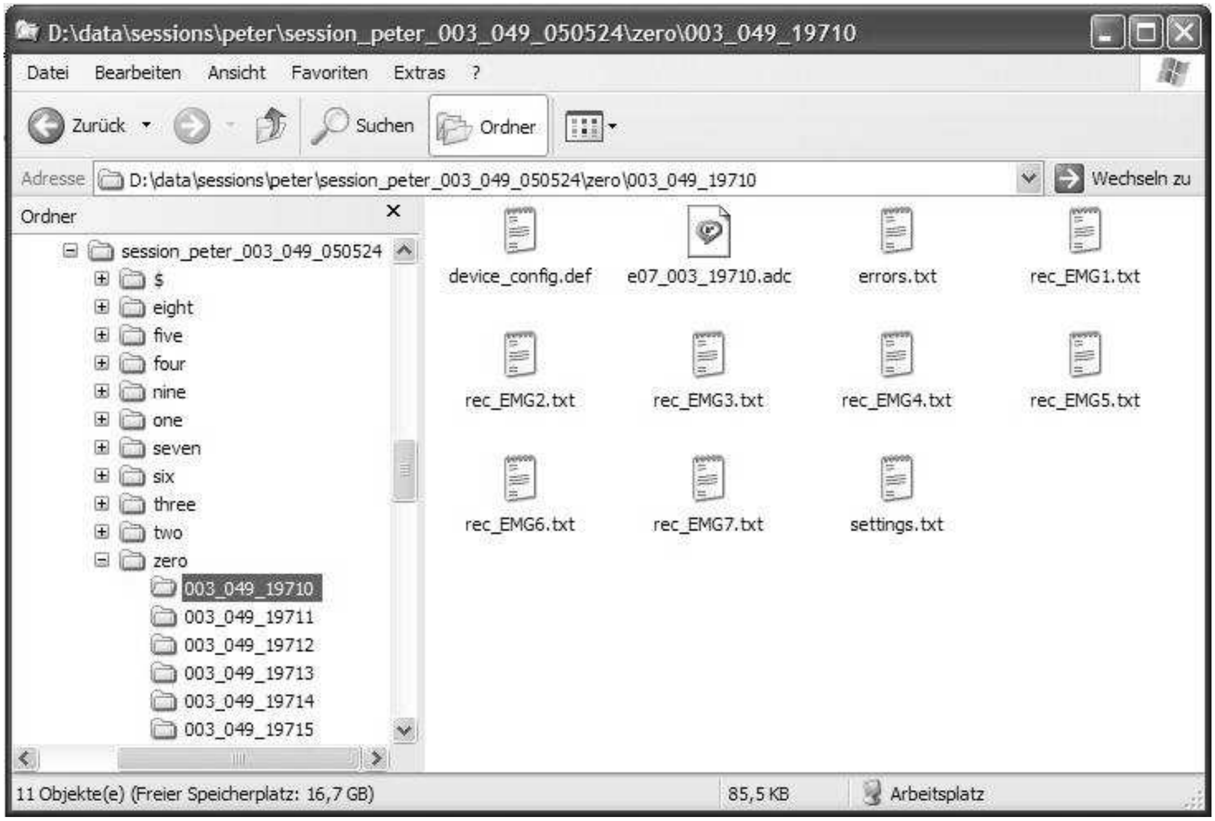
2. Choosing settings: the following settings have to be made prior to starting the recording:

   - a set of channels along with appropriate sampling rates is selected
   - a word list is loaded containing all utterances the speaker is to record during the session.
   - A unique ID (three digits) is entered for the speaker.
   - the session *number* (four digits) is entered. It is equal to the total number of sessions previously recorded by that speaker. Consequently, each tuple (speaker ID, session number) corresponds to exactly one session and vice versa.
   - the *first* utterance number for the session is entered. It is equal to the total number of utterances collected by the corresponding speaker. Thus each tuple (speaker ID, utterance number) corresponds to exactly one recorded utterance and vice versa.

3. Data recording: The speaker records one file set for each word in the word list using the push-to-talk button of the speaker interface. For each channel a .txt file is stored containing the transmitted time domain values in ASCII code. The values from *all* channels are additionally stored in a .adc file which is interpretable by janus. The directory and file structure is shown in figure B.1. After each recording the utterance number is automatically incremented and the next word from the word list is presented to the speaker.

4. Generation of transcript file: When all utterances have been recorded a transcript is (automatically) created for the session (button *create transcript*).

5. Generation of settings file: All settings are stored in a file. A sample settings file is given in appendix C.



D:\data\sessions\peter\session_peter_003_049_050524\zero\003_049_19710

D:\data\sessions\peter\session_peter_003_049_050524\zero\003_049_19710

```
session_peter_003_049_050524
    $
    eight
    five
    four
    nine
    one
    seven
    six
    three
    two
    zero
        003_049_19710
        003_049_19711
        003_049_19712
        003_049_19713
        003_049_19714
        003_049_19715
```

device_config.def
e07_003_19710.adc
errors.txt
rec_EMG1.txt
rec_EMG2.txt
rec_EMG3.txt
rec_EMG4.txt
rec_EMG5.txt
rec_EMG6.txt
rec_EMG7.txt
settings.txt

11 Objekte(e) (Freier Speicherplatz: 16,7 GB)

85,5 KB

Figure B.1: Directory and file structure for data recording session 049 from speaker 003 (S3)

Appendix C

Typical Settings file

```
GENERAL INFORMATION
------------------------------
Speaker name: Peter Osztotics
Speaker ID: 003
Session: 052
Date: 30.05.2005
Time: 19:25 - 19:48
Software version: V2.03


CHANNEL INFORMATION
------------------------------
Channels recorded: EMG1 EMG2 EMG3 EMG4 EMG5 EMG6 EMG7
Number of channels: 7
Sampling rate: 0.60 kHz


Electrode positions:
        EMG1: P22-P23
        EMG2: P35-P38
        EMG3: P17
        EMG4: P42
        EMG5: P28
        EMG6: P46-47
        EMG7: P51-52
        GROUND: LEFT_WRIST
        REF1: LEFT_EAR
        REF2: RIGHT_EAR
Electrode types:
        EMG1: BIPOLAR
        EMG2: BIPOLAR
        EMG3: UNIPOLAR
        EMG4: UNIPOLAR
        EMG5: UNIPOLAR
        EMG6: BIPOLAR
        EMG7: BIPOLAR
        GROUND: ARBO H99LG
        REF1: LE
        REF2: RE
```

Signal quality: GOOD

VOCABULARY

Domain: DIGITS
Connected: NO
Wordlist: digits_40_randomized.wdl

ID INFORMATION

IDs recorded: 00320720 - 00321159 (440 IDs)
Corrupts IDs: -
Bad Signal IDs: -
Comparable sessions: -
Sessions with exactly the same electrode placement: 051

AUDIO INFORMATION

Speech audible: Yes
Audio sampling rate: 16.0 kHz
Synchronization: NONE

COMMENT

Appendix D

Listing of all recording sessions

D.1 Sessions from speaker S0

Table D.1: Listing of all sessions from speaker S0.

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
000	no_and_accept_15.wdl	yes	M_INITa M_INITb	PHONE_INIT	M_INITa: next to larynx; right side; inner electrode. M_INITb: next to larynx; right side; outer electrode.
001	phone_10.wdl	yes	M_INITa M_INITb	PHONE_INIT	-
002	phone_80.wdl	yes	M_INITa M_INITb	PHONE_INIT	-
003	phone_30_ohne_sil.wdl	no	M_INITa M_INITb	PHONE_INIT	-
004	phone_50.wdl	yes	P1 P2	PHONE	-
005	phone_40.wdl	yes	P1 P2	PHONE	Marker was used in order to obtain same placement as in session 004. Bad signal quality.
006	phone_50.wdl	yes	P1 P2	PHONE	-

Continued on next page

Table D.1 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
007	phone_30.wdl	yes	P1 P2	PHONE	Marker was used in order to obtain same placement as in session 006.
008	phone_60.wdl	yes	P1 P2	PHONE	-
009	phone_50.wdl	yes	P1 P2	PHONE	-
010	demo_30.wdl	yes	P1 P2	MEETING	-
011	demo_30.wdl	no	P22-P23 P32 P17 P16 P42 P5	MEETING	Electrode EMG2 (P32) lost contact during session.
012	demo_dialog.wdl	no	P22-P23 P32 P17 P16 P42 P5	MEETING	Electrode EMG2 (P32) lost contact during session 11.

D.2 Sessions from speaker S1

Table D.2: Listing of all sessions from speaker S1.

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
000	phone_10.wdl	yes	L_INITa L_INITb	PHONE_INIT	L_INITa: next to larynx; right side; inner electrode. L_INITb: next to larynx; right side; outer electrode.
001	phone_10.wdl	yes	L_INITa L_INITb	PHONE_INIT	-
002	phone_70.wdl	yes	L_INITa L_INITb	PHONE_INIT	-
003	phone_20.wdl	no	L_INITa L_INITb	PHONE_INIT	-
004	phone_40.wdl	yes	L_INITa L_INITb	PHONE_INIT	-
005	phone_35.wdl	yes	P1 P2	PHONE	-
006	phone_40.wdl	yes	P1 P2	PHONE	-
007	phone_40.wdl	yes	P1 P2	PHONE	Positions from 006 marked and re-used. Bad signal quality. Electrodes possibly lost contact.
008	phone_40.wdl	yes	P1 P2	PHONE	-
009	phone_50.wdl	yes	P1 P2	PHONE	-
010	digits_30.wdl	no	P22-P23 P24-P25 P45_b P29_b P28 P27	DIGITS_VIII_POS	-

Continued on next page

Table D.2 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
011	digits_30.wdl	no	P5 P4 P7 P8 P3 P6 P46-P47	DIGITS_VIII_POS	-
012	digits_30.wdl	no	P17 P16 P36 P43 P37 P15 P39	DIGITS_VIII_POS	-
013	digits_30.wdl	no	P42 P41 P33-P34 P31 P32 P40 P35-P38	DIGITS_VIII_POS	-
014	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
015	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
016	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-

Continued on next page

Table D.2 – continued from previous page

Session Number	**Word List**	**Audible**	**Electrode Positions**	**Corpus**	**Comment**
017	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
018	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-

D.3 Sessions from speaker S2

Table D.3: Listing of all sessions from speaker S2.

Session Number	**Word List**	**Audible**	**Electrode Positions**	**Corpus**	**Comment**
000	phone_2.wdl	yes	MH_INITa MH_INITa	PHONE_INIT	-
001	phone_45.wdl	yes	MH_INITa MH_INITa	PHONE_INIT	-

D.4 Sessions from speaker S3

Table D.4: Listing of all sessions from speaker S3.

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
000	phone_60.wdl	no	P1 P2	PHONE	-
001	phone_40.wdl	yes	P1 P2	PHONE	synchronization not optimal.
002	phone_50.wdl	yes	P1 P2	PHONE	synchronization not optimal.
003	phone_60.wdl	no	P1 P2	PHONE	-
004	phone_50.wdl	yes	P1 P2	PHONE	synchronization not optimal.
005	phone_50.wdl	no	P1 P2	PHONE	-
006	phone_50.wdl	yes	P3 P4	PHONE	synchronization not optimal.
007	phone_50.wdl	no	P3 P4	PHONE	-
008	digits_50.wdl	yes	P3 P4	DIGITS_II	synchronization not optimal.
009	digits_with_call_end_50.wdl	yes	P3 P4	DIGITS_II	synchronization not optimal.
010	digits_with_call_end_50.wdl	no	P3 P4	DIGITS_II	-
011	connected_digits_test_3.wdl	yes	P3 P4	DIGITS_II	synchronization not optimal.
012	digits_with_call_end_50.wdl	yes	P3 P4	DIGITS_II	synchronization not optimal.
013	digits_with_call_end_50.wdl	no	P3 P4	DIGITS_II	-
014	digits_with_call_end_50.wdl	no	P3 P4	DIGITS_II	-
015	call_2digits_end_5.wdl	yes	P3 P4	DIGITS_II	-

Continued on next page

Table D.4 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
016	call_6digits_end_5.wdl	yes	P3 P4	DIGITS_II	-
017	digits_with_call_end_50.wdl	no	P3 P4	DIGITS_II	-
018	digits_with_call_end_50.wdl	yes	P3 P4	DIGITS_II	-
019	call_digit_end_50.wdl	yes	P3 P4	DIGITS_II	-
020	digits_with_call_end_50.wdl	yes	P3 P4	DIGITS_II	-
021	digits_with_call_end_50.wdl	yes	P3 P4	DIGITS_II	-
022	digits_with_call_end_50.wdl	yes	P3 P4	DIGITS_II	Bad signal quality because electrode contact changed during recording.
023	digits_with_call_end_50.wdl	no	P3 P4	DIGITS_II	Bad signal quality because electrode contact changed during recording.
024	digits_30_10sil.wdl	yes	P4 P3 P5_b P6_b P7_b	DIGITS_VIII_INIT	No audio files recorded (corrupt). Amplification too high.
025	digits_30_10sil.wdl	no	P4 P3 P5 P6 P7	DIGITS_VIII_INIT	Amplification too high.
026	digits_30_10sil.wdl	yes	P3 P9 P10 P11 P13-P12 P44	DIGITS_VIII_INIT	Amplification too high.

Continued on next page

Table D.4 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
027	digits_30.wdl	no	P17 P16 P21-P20 P18-P19 P14 P15	DIGITS_VIII_INIT	-
028	digits_30.wdl	yes	P17 P16 P21-P20 P18-P19 P14 P15	DIGITS_VIII_INIT	Electrode contact changed during session (all channels bad)
029	digits_30.wdl	no	P22-P23 P24-P25 P26 P27 P28 P29 P30	DIGITS_VIII_INIT	Electrodes EMG4 and EMG6 lost contact during session. EMG3 and EMG5 not so good either.
030	digits_30.wdl	yes	P22-P23 P24-P25 P26 P27 P28 P29 P30	DIGITS_VIII_INIT	Electrodes EMG4 and EMG6 lost contact during session 029. EMG3 and EMG5 not so good either.
031	digits_30.wdl	no	P5 P4 P7 P8 P3 P6	DIGITS_VIII_POS	-
032	digits_30.wdl	yes	P5 P4 P7 P8 P3 P6	DIGITS_VIII_POS	-

Continued on next page

Table D.4 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
033	digits_30.wdl	no	P42 P41 P33-P34 P31 P32 P40	DIGITS_VIII_POS	-
034	digits_30.wdl	yes	P42 P41 P33-P34 P31 P32 P40	DIGITS_VIII_POS	-
035	digits_30.wdl	no	P17 P16 P36 P34 P37 P15	DIGITS_VIII_POS	-
036	digits_30.wdl	yes	P17 P16 P36 P34 P37 P15	DIGITS_VIII_POS	-
037	digits_30.wdl	no	P22-P23 P24-P25 P45_b P29_b P28 P27	DIGITS_VIII_POS	-
038	digits_30.wdl	yes	P22-P23 P24-P25 P45_b P29_b P28 P27	DIGITS_VIII_POS	-
039	demo_10wds_20sil_35.wdl	no	P24-P25 P22-P23 P28 P42 P32 P17 P5	MEETING	-

Continued on next page

Table D.4 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
040	digits_40_20sil.wdl	no	P24-P25 P22-P23 P28 P42 P32 P17 P5	DIGITS_VIII_INIT	-
041	digits_40_20sil.wdl	yes	P24-P25 P22-P23 P28 P42 P32 P17 P5	DIGITS_VIII_INIT	-
042	demo_10wds_20sil_35.wdl	no	P22-P23 P24-P25 P28 P42 P32 P17 P5	MEETING	-
043	digits_40_20sil.wdl	no	P22-P23 P24-P25 P28 P42 P32 P17 P5	DIGITS_VIII_INIT	-
044	digits_40_20sil.wdl	yes	P22-P23 P24-P25 P28 P42 P32 P17 P5	DIGITS_VIII_INIT	-
045	demo_dialog_7wds_35.wdl	no	P22-P23 P24-P25 P28 P42 P32 P17 P5	MEETING	All ADC files wrong because of change in the software. Session not usable!

Continued on next page

Table D.4 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
046	demo_dialog_7wds_45.wdl	no	P22-P23 P24-P25 P28 P42 P32 P17 P5	MEETING	-
047	demo_dialog_7wds_40.wdl	no	P22-P23 P24-P25 P28 P42 P32 P17 P4	MEETING	-
048	demo_dialog_7wds_45.wdl	no	P22-P23 P24-P25 P28 P42 P32 P17 P4	MEETING	Marker from session 047 (previous day!).
049	digits_30.wdl	no	P35-P38 P46-P47 P39 P17 P4 P50 P48-P49	DIGITS_VIII_POS	-
050	digits_30.wdl	yes	P35-P38 P46-P47 P39 P17 P4 P50 P48-P49	DIGITS_VIII_POS	-
051	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-

Continued on next page

Table D.4 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
052	digits_40.wdl	yes	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
053	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
054	commands_tripels.wdl	yes	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	COMMANDS	-
055	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
056	6connected_digits.wdl	yes	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII_CON	-
057	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	signals for EMG6 were bad.

Continued on next page

Table D.4 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
058	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-

D.5 Sessions from speaker S4

Table D.5: Listing of all sessions from speaker S4.

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
000	demo_dialog_7wds_35.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	MEETING	EMG1 and EMG5 were placed on beard.
001	demo_dialog_7wds_35.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	MEETING	Speech was whispered. EMG1 and EMG5 were placed on beard.

D.6 Sessions from speaker S5

Table D.6: Listing of all sessions from speaker S5.

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
000	demo_7wds_30.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	MEETING	-
001	lecture_14wds_30.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	LECTURE	-
002	lecture_7wds_30.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	LECTURE	-
003	lecture_7wds_30.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	LECTURE	-
004	demo_7wds_35.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	MEETING	-

Continued on next page

Table D.6 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
005	demo_dialog_7wds_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	MEETING	Speaking rate was different (faster) than in related sessions 000 and 004.
006	6connected_digits.wdl	yes	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII_CON	-
007	demo_dialog_7wds_45.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	MEETING	Speaking rate was different (faster) than in sessions 000 and 004.

D.7 Sessions from speaker S7

Table D.7: Listing of all sessions from speaker S7.

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
001	phone_40_sil_40.wdl	yes	P1 P2	PHONE_INIT	-
002	digits_30.wdl	yes	P3 P9 P10 P11 P13-P12 P16	DIGITS_VIII_INIT	Amplification too high. Device crashed many times.
003	digits_30.wdl	no	P3 P9 P10 P11 P13-P12 P16	DIGITS_VIII_INIT	Amplification too high. Device crashed many times. Electrode contact got worse during session.
004	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	Amplification too high.
005	digits_40.wdl	yes	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	some audio files corrupt because of wrong microphone settings (up to approximately utterance 1800).
006	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	EMG2 was supposed to be 35-38 but was 38-35; subject felt tired from about utt_id 0072040 to utt_id 0072180

Continued on next page

Table D.7 – continued from previous page

Session Number	Word List	Audible	Electrode Positions	Corpus	Comment
007	digits_40.wdl	yes	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	EMG2 was supposed to be 35-38 but was 38-35; subject felt exhausted during whole session.
008	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
009	digits_40.wdl	yes	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
010	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-
011	digits_40.wdl	no	P22-P23 P35-P38 P17 P42 P28 P46-P47 P51-P52	DIGITS_VIII	-

Bibliography

[Ankrum, 2000] Ankrum, D. (2000). Questions to ask when interpreting surface electromyography (semg) research. In *Proceedings of the IEA 2000/HFES 2000 Congress*.

[Becker, 2003a] Becker, K. (2003a). *Erweiterbarer Achtkanal-Recorder für physiologische Daten*.

[Becker, 2003b] Becker, K. (2003b). Varioport ^{TM}. http://www.becker-meditec.de.

[Becker, 2004] Becker, K. (2004). *Varioport ^{TM}- Gebrauchsanweisung*.

[Bowden and HAJEK, 1999] Bowden, J. and HAJEK, J. (1999). Handbook of the international phonetic association.

[Chan et al., 2002a] Chan, A., Englehart, K., Hudgins, B., and Lovely, D. (2002a). A multi-expert speech recognition system using acoustic and myoelectric signals. In *Proceedings of the Second Joint EMBS/BMES Conference, Houston, TX, USA*.

[Chan et al., 2001] Chan, A., K.Englehart, Hudgins, B., and Lovely, D. (2001). Myoelectric signals to augment speech recognition. *Medical and Biological Engineering and Computing*, 39:500–506.

[Chan et al., 2002b] Chan, A., K.Englehart, Hudgins, B., and Lovely, D. (2002b). Hidden markov model classification of myolectric signals in speech. *Engineering in Medicine and Biology Magazine, IEEE*, 21:143–146.

[Coleman et al., 2002] Coleman, J., Grabe, E., and Braun, B. (2002). Larynx movements and intonation in whispered speech. Summary of research supported by British Academy grant SG-36269.

[Englehart et al., 1999] Englehart, K., Hudgins, B., Parker, P., and Stevenson, M. (1999). Classification of the myoelectric signal using time-frequency based representations. *Medical Engineering and Physics*, 21:431–438.

[Finke et al., 1997] Finke, M., Geutner, P., Hild, H., emp, T. K., Ries, K., and Westphal, M. (1997). The Karlsruhe Verbmobil Speech Recognition Engine. In *Proc. ICASSP 97*, München; Germany. IEEE.

[Gales, 1997] Gales, M. J. F. (1997). Maximum likelihood linear transformations for HMM-based speech recognition. Technical report, Cambridge University, Cambridge, UK. CUED/F-INFENG/TR 291.

[Hermens and Freriks, 2005] Hermens, D. H. and Freriks, B. (2005). Surface electromyography for the non-invasive assessment of muscles (seniam). http://www.seniam.org/.

[Jin et al., 1998] Jin, H., Matsoukas, S., Schwartz, R., and Kubala, F. (1998). Fast Robust Inverse Transform SAT and Multi-stage Adaptation. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA; USA.

[Jorgensen and Binsted, 2005] Jorgensen, C. and Binsted, K. (2005). Web browser control using emg based sub vocal speech recognition. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*.

[Jorgensen et al., 2003] Jorgensen, C., Lee, D., and Agabon, S. (2003). Sub auditory speech recognition based on emg/epg signals. In *Proceedings of the International Joint Conference on Neural Networks*.

[Kumar et al., 2004] Kumar, S., Kumar, D., Alemu, M., and Burry, M. (2004). Emg based voice recognition. In *Proceedings of the 2004 ISSNIP conference*.

[Laboratory, 2002] Laboratory, U. P. (2002). Dissection of the speech production mechanism.

[Lamb and Hobart, 1992] Lamb, R. and Hobart, D. (1992). Anatomic and physiologic basis for surface electromyography. In *Selected Topics in Surface Electromyography for Use in the Occupational Setting: Expert Perspective*. U.S. Department of Health and Human Services. DHHS(NIOSH) Publication No 91-100.

[Lehtonen, 2002] Lehtonen, J. (2002). Eeg-based brain computer interfaces. Master's thesis, Helsinki University of Technology.

[Lemmetty, 1999] Lemmetty, S. (1999). Review of speech synthesis technology. Master's thesis, Helsinki University of Technology.

[Leveau and Andersson, 1992] Leveau, B. and Andersson, G. (1992). output forms: Data analysis and applications. In *Selected Topics in Surface Electromyography for Use in the Occupational Setting: Expert Perspective*. U.S. Department of Health and Human Services. DHHS(NIOSH) Publication No 91-100.

[Luca, 2002] Luca, C. D. (2002). Surface electromyography: Detection and recording. http://www.delsys.com/library/papers/SEMGintro.pdf.

[Maier-Hein, 2005] Maier-Hein, L. (2005). *Speech Recognition using Surface Electromyography - software documentation.*

[Manabe, 2004] Manabe, H. (2004). Evaluations of the ring-shaped emg measurement system. In *Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, CA, USA*.

[Manabe et al., 2003a] Manabe, H., Hiraiwa, A., and Sugimura, T. (2003a). A ring-shaped emg measurement system for applying to user interface. In *Proceedings of the IEEE EMBS2003*.

[Manabe et al., 2003b] Manabe, H., Hiraiwa, A., and Sugimura, T. (2003b). Unvoiced speech recognition using emg - mime speech recognition -. In *Proceedings of the 2003 Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA*.

[Manabe and Z.Zhang, 2004] Manabe, H. and Z.Zhang (2004). Multi-stream hmm for emg-based speech recognition. In *Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, CA, USA*.

[Matthews, 1991] Matthews, G. (1991). Cellular physiology of nerve an muscle. http://artsci-ccwin.concordia.ca/psychology/psyc358/Lectures/APfigure.htm.

[Mayer, 2005] Mayer, C. (2005). *UKA EMG/EEG Studio v2.0.*

[Metze, 2004] Metze, F. (2004). Jrtk online documentation. http://isl.ira.uka.de/ jrtk/janus-doku.html.

[Meyer-Waarden, 1985] Meyer-Waarden, K. (1985). *Bioelektrische Signale und ihre Ableitverfahren.* Schattauer Verlag.

[Morse et al., 1991] Morse, M., Gopalan, Y., and Wright, M. (1991). Speech recognition using myoelectric signals with neural networks. In *Engineering in Medicine and Biology Society. Proceedings of the Annual International Conference of the IEEE*, volume 13, pages 1877–1878.

[Morse M., 1986] Morse M., O. E. (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. *Comput Biol Med*.

[Noffke, 2005] Noffke, W. (2005). Lecture notes on anatomy and physiology. http://www.cptc.ctc.edu/library/Bio

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

[Rash, 1999] Rash, G. (1999). Electromyography fundamentals. www.ac.wwu.edu/ chalmers/EMGfundamentals.pdf.

[Ritchison, 2005] Ritchison, G. (2005). Lecture notes on human physiology. Department of Biological Sciences Eastern Kentucky University.

[Rogina, 2003] Rogina, I. (2003). Sprachliche mensch-maschine-kommunikation. Entwurf der Habitilationsschrift.

[Scott, 2003] Scott, D. (2003). Important factors in surface emg measurement. www.bortec.ca/Images/pdf/ EMG

[Soderberg, 1992] Soderberg, G. (1992). Recording techniques. In *Selected Topics in Surface Electromyography for Use in the Occupational Setting: Expert Perspective*. U.S. Department of Health and Human Services. DHHS(NIOSH) Publication No 91-100.

[S.Silbernagel and Despopoulos, 2003] S.Silbernagel and Despopoulos, A. (2003). *Taschenatlas der Physiologie*. Thieme.

[Stüker, 2003] Stüker, S. (2003). Multilingual articulatory features. Master's thesis, Universität Karlsruhe, Carnegie Mellon University.

[Wikipedia, 2005] Wikipedia (2005). Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Main_Page.

[Wölfel, 2003] Wölfel, M. (2003). Minimum variance distortionless response spectral estimation and subtraction for robust speech recognition. Master's thesis, Universität Karlsruhe, Carnegie Mellon University.