EMG-to-Speech: Direct Generation of Speech from Facial Electromyographic Signals

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften
der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Matthias Janke
aus Offenburg

Tag der mündlichen Prüfung:	22. 04. 2016
Erste Gutachterin:	Prof. Dr.-Ing. Tanja Schultz
Zweiter Gutachter:	Prof. Alan W Black, PhD

Summary

For several years, alternative speech communication techniques have been examined, which are based solely on the articulatory muscle signals instead of the acoustic speech signal. Since these approaches also work with completely silent articulated speech, several advantages arise: the signal is not corrupted by background noise, bystanders are not disturbed, as well as assistance to people who have lost their voice, e.g. due to accident or due to disease of the larynx.

The general objective of this work is the design, implementation, improvement and evaluation of a system that uses surface electromyographic (EMG) signals and directly synthesizes an audible speech output: *EMG-to-speech.* The electrical potentials of the articulatory muscles are recorded by small electrodes on the surface of the face and neck. An analysis of these signals allows interpretations on the movements of the articulatory apparatus and in turn on the spoken speech itself.

An approach for creating an acoustic signal from the EMG-signal is the usage of techniques from automatic speech recognition. Here, a textual output is produced, which in turn is further processed by a text-to-speech synthesis component. However, this approach is difficult resulting from challenges in the speech recognition part, such as the restriction to a given vocabulary or recognition errors of the system.

This thesis investigates the possibility to convert the recorded EMG signal directly into a speech signal, without being bound to a limited vocabulary or other limitations from an speech recognition component. Different approaches for the conversion are being pursued, real-time capable systems are implemented, evaluated and compared.

For training a statistical transformation model, the EMG signals and the acoustic speech are captured simultaneously and relevant characteristics in terms of features are extracted. The acoustic speech data is only required as a reference for the training, thus the actual application of the transformation can take place using solely the EMG data. A feature mapping is accomplished by a model that estimates the relationship between muscle activity patterns and speech sound components. From the speech components the final audible voice signal is synthesized. This approach is based on a source-filter model of speech: The fundamental frequency is overlaid with the spectral information (Mel Cepstrum), which reflects the vocal tract, to generate the final speech signal.

To ensure a natural voice output, the usage of the fundamental frequency for prosody generation is of great importance. To bridge the gap between normal speech (with fundamental frequency) and silent speech (no speech signal at all), **whispered speech** recordings are investigated as an intermediate step. In whispered speech no fundamental frequency exists and accordingly the generation of prosody is possible, but difficult.

This thesis examines and evaluates the following three mapping methods for feature conversion:

1. **Gaussian Mapping**: A statistical method that trains a Gaussian mixture model based on the EMG and speech characteristics in order to estimate a mapping from EMG to speech. Originally, this technology stems from the Voice Conversion domain: The voice of a source speaker is transformed so that it sounds like the voice of a target speaker.

2. **Unit Selection**: The recorded EMG and speech data are segmented in paired units and stored in a database. For the conversion of the EMG signal, the database is searched for the best matching units and the resulting speech segments are concatenated to build the final acoustic output.

3. **Deep Neural Networks**: Similar to the Gaussian mapping, a statistical relation of the feature vectors is trained, using deep neural networks. These can also contain recurrences to reproduce additional temporal contextual sequences.

This thesis examines the proposed approaches for a suitable direct EMG-to-speech transformation, followed by an implementation and optimization of each approach. To analyze the efficiency of these methods, a data corpus was recorded, containing normally spoken, as well as silently articulated speech and EMG recordings. While previous EMG and speech recordings were mostly limited to 50 utterances, the recordings performed in the scope of this thesis include sessions with up to 2 hours of parallel EMG and speech data. Based on this data, two of the approaches are applied for the first time to EMG-to-speech transformation. Comparing the obtained results to related EMG-to-speech work, shows a relative improvement of 29 % with our best performing mapping approach.

Zusammenfassung

Seit einigen Jahren werden alternative Sprachkommunikationsformen untersucht, welche anstatt des akustischen Sprachsignals allein auf den, beim Sprechen aufgezeichneten, Muskelsignalen beruhen. Da dieses Verfahren auch funktioniert, wenn komplett lautlos artikuliert wird, ergeben sich mehrere Vorteile: Keine Nutzsignalbeeinträchtigung durch Hintergrundgeräusche, keine Lärmbelästigung Dritter, sowie die Unterstützung von Menschen, die beispielsweise durch Unfall oder Erkrankung ihre Stimme verloren haben. Ziel dieser Arbeit ist die Entwicklung, Evaluation und Verbesserung eines Systems, welches elektromyographische (EMG) Signale der artikulatorischen Muskelaktivität direkt in ein hörbares Sprachsignal synthestisiert: *EMG-to-speech*.

Dazu werden elektrische Potentiale der Artikulationsmuskeln, durch am Gesicht angebrachte, Oberflächenelektroden aufgezeichnet. Eine Analyse dieser Signale ermöglicht Rückschlüsse auf die Bewegungen des Artikulationsapparates und damit wiederum auf die gesprochene Sprache selbst.

Ein möglicher Ansatz zur Erzeugung der akustischen Sprache aus dem EMG-Signal ist die Verwendung von Techniken aus der automatischen Spracherkennung. Dabei wird eine textuelle Ausgabe erzeugt, welche wiederum mit einer textbasierten Synthese verknüpft wird, um das Gesprochene hörbar zu machen. Dieses Verfahren ist jedoch verbunden mit Einschränkungen das Spracherkennungssystems, wie z.B. das limitierte Vokabular sowie Erkennungsfehler des Systems.

Diese Arbeit befasst sich mit der Möglichkeit das aufgezeichnete EMG-Signal direkt in ein Sprachsignal umzuwandeln, ohne dabei an ein beschränktes Vokabular oder weitere Vorgaben gebunden zu sein. Dazu werden unterschiedliche Abbildungsansätze verfolgt, möglichst echtzeitfähige Systeme implementiert, sowie anschließend evaluiert und miteinander verglichen.

Zum Training eines statistischen Transformationsmodells werden die EMG-Signale und die simultan erzeugte akustische Sprache zunächst erfasst und relevante Merkmale extrahiert. Die akustischen Sprachdaten werden dabei

nur als Referenz für das Training benötigt, so dass die eigentliche Anwendung der Transformation ausschließlich auf den EMG-Daten stattfinden kann.

Die Transformation der Merkmale wird erreicht, indem ein Modell erzeugt wird, das den Zusammenhang von Muskelaktivitätsmustern und Lautbestandteilen der Sprache abbildet. Dazu werden drei alternative Abbildungsverfahren untersucht. Aus den Sprachbestandteilen kann wiederum das hörbare Sprachsignal synthetisiert werden. Grundlage dafür ist ein Quelle-Filter-Modell: Die Grundfrequenz wird mit den spektralen Informationen (Mel Cepstrum) überlagert, die den Vokaltrakt wiederspiegeln, um daraus das finale Sprachsignal zu generieren.

Um eine natürliche Sprachausgabe zu gewährleisten, ist die Verwendung der Grundfrequenz zur Generierung der Prosodie von großer Bedeutung. Da sich die sprachlichen Charakteristika von normaler Sprache (enthält Grundfrequenz) und lautloser Sprache (kein Sprachsignal) deutlich unterscheiden, werden als Zwischenschritt **geflüsterte** Sprachaufnahmen betrachtet. Bei geflüsterter Sprache ist keine Grundfrequenz vorhanden und dementsprechend die Prosodiegenerierung notwendig, wobei im Gegensatz zur lautlosen Sprache auf ein akustisches Sprachsignal zurückgegriffen werden kann.

In dieser Dissertation wurden die folgenden drei Abbildungsverfahren zur Merkmalstransformation entwickelt, implementiert, im Detail untersucht und ausgewertet:

1. **Gaussian Mapping**: Ein statistisches Verfahren, das eine Gaußsche Mischverteilung auf Basis der EMG- und Sprach-Merkmale trainiert, um damit eine Abbildung von EMG auf Sprache zu schätzen. Ursprünglich findet diese Technik im Bereich der Sprechertransformation Verwendung: Die Stimme eines Quellsprechers wird so transformiert, dass sie wie die Stimme eines Zielsprechers klingt.

2. **Unit Selection**: Die aufgenommenen EMG- und Sprach-Daten werden in paarweise Einheiten segmentiert und in einer Datenbank abgelegt. Für die Konvertierung des EMG-Signals wird die Datenbank nach möglichst passenden Segmenten durchsucht und die resultierenden Sprachabschnitte konkateniert.

3. **Deep Neural Networks**: Ähnlich dem Gaussian Mapping wird hier eine Abbildung der Merkmalsvektoren trainiert. Als Abbildungsverfahren kommt ein vielschichtiges neuronales Netz zum Einsatz, welches zusätzlich auch Rekurrenzen enthalten kann, um zusätzlich kontextuelle zeitliche Abläufe wiederzugeben.

Ziel dieser Dissertation ist zum einen, jeden dieser Ansätze auf Eignung zur direkten EMG-zu-Sprach-Transformation zu untersuchen und diese jeweils zu optimieren. Zum anderen werden die einzelnen Verfahren sowohl objektiv als auch mittels subjektiver Auswertung miteinander verglichen. Zur Analyse der Leistungsfähigkeit dieser Verfahren wurde ein Datenkorpus aufgezeichnet, welcher sowohl normal gesprochene, als auch lautlos artikulierte EMG- und Sprach-Aufnahmen enthält. Während bisherige EMG und Sprachaufnahmen meist auf 50 Äußerungen limitiert waren, werden die Aufnahmen im Rahmen dieser Dissertation auf bis zu 2 Stunden EMG und Sprachdaten erweitert. Auf Basis dieser Daten werden zwei der genannten Verfahren nach unserer Kenntnis weltweit erstmals zur EMG-basierten Transformation umgesetzt. Ein Vergleich der resultierenden Ergebnisse unseres besten Systems mit verwandten EMG-to-Speech Ansätzen zeigt eine relative Verbesserung von 29 %.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

This chapter serves as a motivation for this dissertation, introduces the field of Silent Speech Interfaces and motivates our approach that directly transforms myographic signals into speech output, entitled EMG-to-speech. The contributions are summarized and highlighted, followed by an outline of the structure of this thesis.

1.1 Speech Communication

Speech communication is the most sophisticated form of human interaction. During the last decades it got high attention for its usage to ease human-computer interfaces, due to its efficiency, naturalness and richness to express information. Today's widely known speech-based systems rely on acoustic data, i.e. the speech is transmitted over air. However, there exist several scenarios where alternative speech signals are desired:

- Adverse noise conditions: High ambient background noise makes the usage of an acoustic speech signal challenging or even impossible.
- Complementing acoustic speech: The addition of a second speech modality can improve a speech processing system.
- Silent speech communication: Transmitting confidential information like passwords or PINs should not be eavesdropped by bystanders.

In general, disturbing the environment with ones voice is not always wanted. Meetings, concerts or libraries are just three of multiple noise-sensitive scenarios. Therefore, a private, confidential and non-obtrusive communication is needed.

- Speech rehabilitation: Help for people who suffer from speech-disabilities or have even lost their voice. According to the National Institute on Deafness and Other Communication Disorders (NIDCD), approximately 7.5 million people in the United States have trouble using their voice [ND16]. Many of these people could obtain assistance and could interact using usual speech communication technologies.

1.2 Silent Speech Processing Technologies

Some of the above mentioned challenges can be alleviated by *Silent Speech Interfaces* (SSI) [DSH+10], i.e. systems that do not rely on an acoustic signal as input. First approaches stem from the 1980s and are either based on visual lip-reading to enhance speech recognition in noisy environments [Pet84], or they rely on facial muscle signals to classify simple vowels [ST85]. Different types of silent speech interfaces have been evolved over the last decades (see Chapter 3 for details) and are actively pursued.

In general, silent speech technologies can be grouped into three categories:

- Capture of very quiet speech signals. e.g. bone-conduction, stethoscopic microphones, etc., which require that at least a small sound is produced. These technologies address the same problems as true silent speech interfaces and are therefore described in the related work in Chapter 3.
- Capture of vocal tract or articulator configurations, e.g. by surface electromyography (EMG), or by visual or ultrasound imaging. This allows to process speech information even when no sound is produced.
- Direct interpretation of brain signals that are related to speech production.

To generate a speech output there exist two different approaches. One way is the usage of a speech recognition system to transform speech-to-text, followed by a text-to-speech synthesis system. The second approach uses a direct feature mapping, which yields some benefits:

- No restriction to a given phone-set, vocabulary or even language.

- Opportunity to retain paralinguistic features, like prosody or speaking rate.

- Preservation of a speaker's voice and other speaker-dependent factors.

- Direct mapping enables faster processing than speech-to-text followed by text-to-speech and thus the ability for nearly real time computation.

1.3 Contributions of this Thesis

In this work we introduce a system that directly converts facial EMG signals into an audible speech output. The capturing of the simultaneous EMG and audio data is followed by the signal processing step to generate relevant features. The next step contains the mapping from EMG features into acoustic representations. The main focus of this thesis is to investigate the direct transformation from EMG to acoustics. This is concluded by the vocoding process, the creation of the final speech wave files from the input acoustic representations.

The main contributions of this thesis are as follows:

- Applying three different state-of-the-art machine learning approaches for *direct speech synthesis based on surface electromyographic signals* (EMG-to-speech): A statistical transformation based on Gaussian mixture models entitled *Gaussian Mapping*, a database-driven approach known as *Unit Selection* and a feature mapping based on *Deep Neural Networks.*

- Generating the fundamental frequency of speech from electromyographic signals (*EMG-to-F0*), enabling a naturally sounding prosody that is based solely on the articulatory muscle movements.

- *Whisper-to-audible speech mapping*: To demonstrate the generation of prosodic information, we investigate the transformation of whispered speech into normal audible speech.

- No need for phone-level time alignments (labels): While comparable electromyography-based approaches need phonetic information in the processing chain, which introduces an additional source for errors, this work establishes a *label-free approach.*

- *Near-real-time processing*: No complex decoding computations are needed and thus we enable fast processing speech output that could be used for acoustic feedback.

1.4 Thesis Structure

This thesis is structured into the following chapters. Section 2 introduces the basic principles that are necessary to understand this work. Chapter 3 gives an overview of the most important related work, followed by an introduction to the experimental setup and the used data corpora in Chapter 4. The challenges and peculiarities of the recorded and analyzed data are discussed in Chapter 5, while the different EMG-to-speech mapping approaches are presented and evaluated in the Chapters 6, 7 and 8. Final comparisons and evaluations are discussed in Chapter 9, while Chapter 10 concludes this thesis.

Chapter 2

Basic Principles

This chapter introduces the basic principles plus scientific and technical foundations that are necessary for the understanding of speech and surface electromyography (EMG) in the context of EMG-to-Speech transformation. The section begins with a brief introduction into speech synthesis that clarifies the basic EMG-to-Speech components that will be described. In addition, the physiological and anatomical basics of speech production and EMG are introduced, concluded by signal processing details. This is complemented by mathematical basics that are relevant for this thesis.

2.1 Introduction and Relation to Speech Synthesis

The artificial generation of acoustic speech is known as speech synthesis. Thus, we start with a short introduction into the history and development of speech synthesis systems that are relevant for the scope of this thesis. More detailed reviews can be found in e.g. [Kla87, Tay09]. Many research has been done in the field of Text-to-Speech (TTS): the production of aural information given from written information, sometimes considered as the counterpart of speech recognition.

There are lots of applications for speech synthesis systems. Screen readers, assisting devices, non-visual human-computer-interfaces, etc. But recreating human speech is a desire that was present long before the upcoming of state-of-the art computer systems. One of the most prominent first speech-like generation systems [Von91] was built by Wolfgang von Kempelen in 1791 and consisted mainly of three components: a pressure chamber (bellows), a vibrating reed and a leather tube. Although this acoustic mechanical speech machine was more like an instrument, it could produce phones and even words that sounded like human speech.

To the beginning and middle of the 20th century researchers started in using electrical methods to synthesize speech. Among the first electrical speech synthesis systems Steward introduced his work in 1922, entitled “Electrical Analogue of the Vocal Organs” [Ste22]. A buzzer acts as excitation and two resonant circuits modeled the vocal tract. Another prominent synthesizer was entitled VODER (Voice operating demonstrator) [DRW39] from Homer Dudley, that was successfully demonstrated at the 1939 World Fair. As computers became widely available for speech synthesis researchers, other approaches were investigated that were not previously possible. e.g. an “analysis-by-synthesis” approach [Bel61] that is inspired by the theory of speech production. In the 1980’s the Klattalk system [Kla82] became popular and first commercial synthesis systems became available.

In general, a speech synthesis approach has two components:

1. a (typically language-specific) frontend that creates a linguistic intermediate representation (similar to a transcription),

2. a backend/waveform generation that creates the speech output based on the linguistic specification.

Regarding a state-of-the-art text-to-speech system the frontend starts with a text preprocessing where non-alphabetic characters (especially numbers) are parsed into a suitable representation. These representations are further processed to get correct pronunciations. This is succeeded by the generation of a meaningful prosody, like duration or pitch. All these steps can be interconnected, e.g. when a question mark has influence on the prosody of a sentence.

Today’s speech synthesis approaches can be grouped into two basic approaches:

1. **Exemplar-based systems:** Segments of target (pre-recorded) speech representations are stored in a database (known as codebook) and are used to concatenate a final output.

2. **Model-based (statistical parametric) systems:** *Parametric* since the built model uses (speech) parameters, *statistical* refers to the fact that distributions are learned from the training data (e.g. by probability density functions).

   The latter model-based approach can be subdivided into two categories:

   (a) **Articulatory approach:** The human speech production system is attempted to be directly modeled.

   (b) **Signal approach:** This tries to mimic the speech signal and its properties itself, using some of the known speech (perceptional) characteristics, e.g. linear predictive coding (see Section 2.6.1).

A term that is closely related to speech synthesis is *voice conversion*, also known as voice transformation. This process covers approaches that transform the speech signal of a source speaker into the voice-characteristics of a target speaker and was introduced in the mid 1980s [CYW85]. A possible application is its usage in a TTS system. Instead of recording multiple of hours of training data (plus the labeling effort) for a synthesis database, one could add a new synthesis voice by applying a voice conversion system, where usually a much smaller amount of training data is required. It is estimated that an amount of 50 speaker utterances can be sufficient for building a voice conversion system [MBB$^+$07]. However, the source and target speaker characteristics can differ in numerous ways, e.g. acoustic differences (male versus female speaker) or prosodic differences (different pitches/volumes).

Training a voice conversion system can be done in a text-dependent or text-independent way. while the first case uses the same text data for source and target speaker and thus only needs a time-alignment, the latter text-independent fashion needs a more sophisticated preparation. One example is the segmentation into a feature space that can be further clustered into similar groups and used for transformation [DEP$^+$06].

The separation into a frontend and backend component can also be done on voice conversion systems. The frontend covers the extraction of appropriate features, that are used for a transformation model that is estimated during a training phase. Different types of transformation techniques exist, e.g. based on Vector Quantization [ANSK88], neural networks [NMRY95] or Gaussian mixture models [SCM98, KM98, TBT07]. The latter approach will be investigated in Chapter 6. The backend represents the vocoder element, that converts the transformed speech characteristics into the final speech sample output.

Figure 2.1 shows the general framework of the EMG-to-speech approach in this thesis, which is inspired by the described techniques and can be divided into a frontend plus a backend component. Our system setup uses facial EMG signals and directly synthesizes an audible speech output. The capturing of simultaneously recorded EMG and audio data is followed by the signal processing step to generate representing features. The next step represents the main part of this thesis, the direct transformation from EMG features into acoustic representations. Finally, the vocoding process creates the final speech wave files from the input acoustic representations. After obtaining the final speech output, there is the question how good the acoustic output actually is and multiple ways for evaluations exist. The theoretic underpinnings of the employed techniques and processes are described in the following sections.

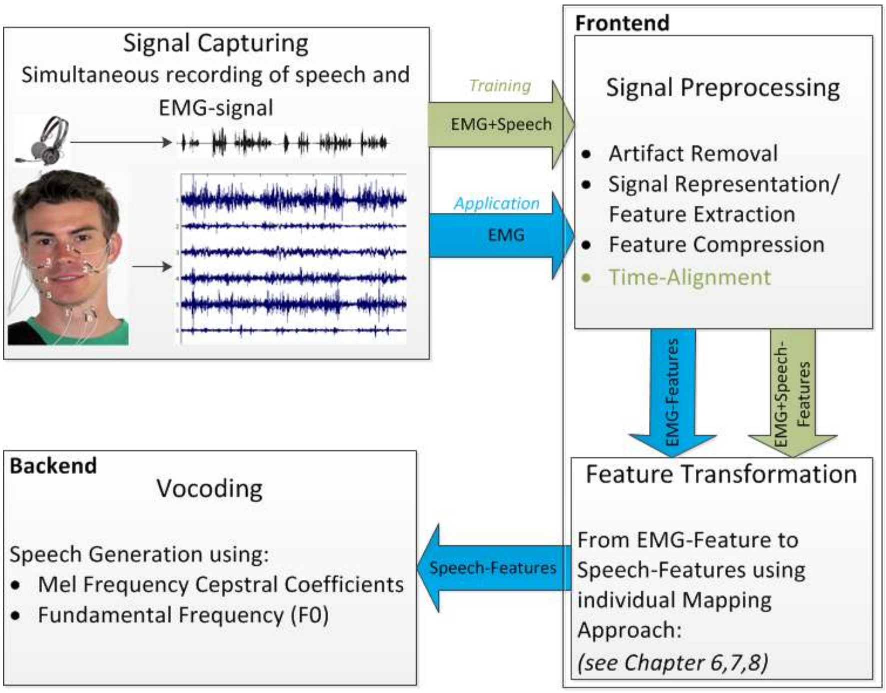


Figure 2.1 – General structure of the proposed EMG-to-speech mapping approach investigated in this thesis. Green arrows represent data flow during training, blue arrows during application of the mapping.

2.2 Physiological and Anatomical Basics

Human motion is the result of muscle contraction, which again is initiated by an electrical stimulus triggered by the brain. Muscles are the executional units in every motoric process. They can be divided into three types: skeletal, cardiac and smooth muscles. The latter two contract only involuntarily and therefore are of no interest to this work. Skeletal muscles are composed of muscle fibers which are surrounded by a plasma membrane called sarcolemma. The muscle fibers are built from a multitude of filamentary structures, called myofibrils. These consist of two types of parallel arranged filaments: myosin filaments and actin filaments repeated in basic functional units entitled *sarcomeres*. Figure 2.2 shows the schematic structure of a skeletal muscle fiber.

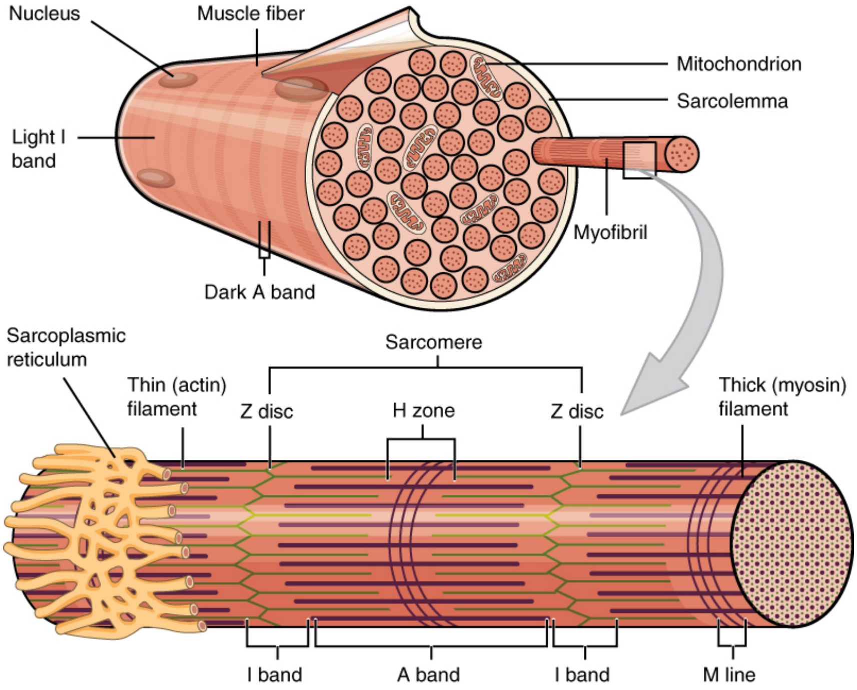


Figure 2.2 – Schematic depiction of the structure of a skeletal muscle fiber, from [Col13].

The actin filaments slide along the myosin filaments, resulting in a muscle contraction. The innervation process is activated by at least one *motor neuron* in the spinal cord. A motor neuron and the muscle fibers it innervates are called *motor unit*. One single motor neuron is able to innervate a multitude of muscle fibers. The motor neuron is connected to the muscle fiber mem-

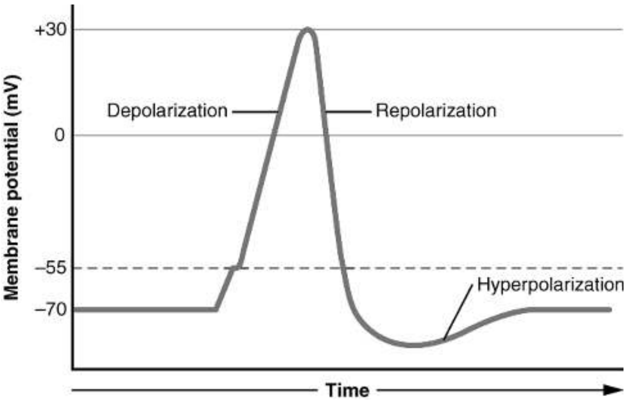


Figure 2.3 – Flow process of a single action potential, from [Col13].

brane by a synapse, the so called motor end-plate. During the activation of the end-plate, transmitter substances are released, which eventually lead to an *action potential* that is propagated along the muscle fiber from the origin and causes muscle contraction.

An motor unit action potential (MUAP) defines the spatial and temporal summation of the electrical activity of those muscle fibers belonging to a motor unit.

The excitability of muscle fibers can be described by a model of the semi-permeable membrane to characterize the electrical properties of the muscle fiber membrane. An ion imbalance between the interior and exterior of the muscle cell generates a resting potential (approximately $-70\,mV$ in uncontracted state). This potential difference is maintained by active physiological processes (ion pump), resulting in a negative charge of the cell interior. The emergence of the electrical pulse is shown in Figure 2.3. When an external stimulus is applied, it results in a depolarization of the muscle fiber membrane (reaching $+30\,mV$), which is restituted immediately by an active compensatory ion backflow (repolarization). The short period of overshooting is defined as hyperpolarization.

According to [Kra07] a muscle contraction can be summarized in three steps:

- Excitable cells react to a stimulus by changing their electrical membrane properties (ion conductivity).
- Resting potential → action potential.
- Action potential leads to muscle contraction.

2.3 Electromyography

While only the basic EMG fundamentals are described, we refer to e.g. [BdL85] for further details. In general, electromyography (EMG) is the study of muscle function through the inquiry of the electrical signal emanating from the muscles [BdL85]. These electrical signals refer to the previously described motor unit action potentials (MUAPs), which stem from several motor units in accordance to muscle contraction. A sequence of MUAPs is defined as a motor unit action potential train (MUAPT). The captured EMG signal consists of a summation of several MUAPTs that can be detected in the vicinity of the recording sensor, thus they can originate from different motor units and even different muscle fibers. Exemplary EMG signals are discussed in Chapter 5.

According to [MP13] the intensity of muscle force depends mainly on two factors:

- The amount of motor units that are recruited, and
- the rate at which they are fired.

This electrical signal is captured by sensors defined as *electrodes*, which transform the ionic potentials from the muscles into electronic currents that can be further precessed with technical devices. In principle, there exist two types of electrodes: 1.) needle or fine-wire electrodes that are directly inserted into the muscle and 2.) surface electrodes that non-invasively measure the electrical activity on the surface of the skin. For the purpose of human-computer interaction in non-medical setups, only surface electrodes are suitable and thus the reported EMG data in this dissertation is captured by surface EMG electrodes. A surface electrode consists of a metal surface (e.g. silver) that is applied to the skin. In this skin-electrode interface, on the moderately conductive tissue side the current is carried by ions, while inside the highly-conductive metal current is carried by electrons. This interface is intrinsically noisy [GB75] and thus highly affected by skin preparation. To improve and stabilize the contact, a conductive gel is applied to the skin-electrode interface. This electrolyte gel contains free chloride ions such that the charge can be carried through the electrolyte. Therefore the electrolyte can be considered as conductive for ion current as the human tissue. “Dry” electrodes (without gel), capacitive electrodes [KHM05] and ceramic electrodes [GSF$^+$95] have also been investigated, but this thesis only reports on standard wet surface Ag/AgCl electrodes.

The recorded EMG signal always represents potential differences between two electrodes, which gives two possible ways of derivation:

Monopolar derivation: Both electrodes are placed on the same muscle fiber.

Bipolar derivation: The neuromuscular activation is derived between one electrode placed on an electrically inactive region and one electrode placed on the active muscle fiber.

The latter configuration is less sensible to artifacts, since disturbances are presented and measured at both electrodes and suppressed by calculating the difference between both sensors. The amplitude of the recorded electromyographic signal is highly dependent on the kind of muscle and varies from μV to the low mV range [BdL85]. The primary energy in the EMG signal lies between 10 and 200 Hz, depending on the skin property and the muscle-electrode-distance [Hay60]. Its measurement depends on various factors, such as the skin properties – conductance of human skin varies with type and thickness of tissue, physiological tissue changes and body temperature – or the types of electrodes and amplifiers.

Different unwanted noise components can be observed in the recorded EMG signal, regarded as *biological artifacts* and *technical artifacts*.

Biological artifacts

Physiological crosstalk: Bioelectric signals that stem from multiple sources inside the human body, e.g. neighboring muscles, can affect the signal and are summarized under the term "crosstalk".

Motion artifacts: During speaking and movement, low-frequent interferences and artifacts can arise, which are the result from the electrodes movement.

Technical artifacts

External interferences: Depending on the environment and the electromagnetic influences of other electronic devices there may exist interferences, e.g. power line noise.

Transducer artifacts: Various influences and artifacts can arise from the recording device, like amplification noise or bad electrode contact.

2.4 Articulatory Muscles

All produced sounds of human speech are the results of muscle contractions. While Section 2.5 gives the details about the production process, this section presents a short overview of the facial muscles that are involved in the articulation process, more precisely the muscles that change the shape of the vocal tract. Figure 2.4 shows the articulatory muscles that are relevant for speech production:

Tongue (not shown in Figure 2.4) is a compound of intrinsic and extrinsic muscles. It occupies most of the oral cavity and has a major influence on vowel articulation. In speech production (see Section 2.5), vowels are characterized by the position of the tongue.

Orbicularis oris a circular muscle that contracts to move the lips.

Levator anguli oris/Levator labii superioris pulls the upper lip upwards.

Zygomaticus major and minor retract the corners of the mouth superiorly and posteriorly, e.g. used for smiling.

Depressor anguli oris pulls the corners of the mouth downwards, e.g. used when frowning.

Depressor labii inferioris lowers the bottom lip.

Risorius retracts the lip corners.

Mentalis protrudes the lower lip and its contraction results in a forward movement of the lower lip.

Buccinator compresses the check and pulls back the angle of the mouth.

Platysma a large muscle that lies under the jaw and down the neck to the upper chest. It depresses the lower jaw and helps to lower the lips.

Masseter a powerful muscle that closes the jaw, e.g. used for mastication.

Since surface electrodes capture signals from multiple muscles, we can not interpret the signals of a particular muscle. In Section 4 we introduce our electrode setup, that uses the information on the positioning of the relevant facial muscles.

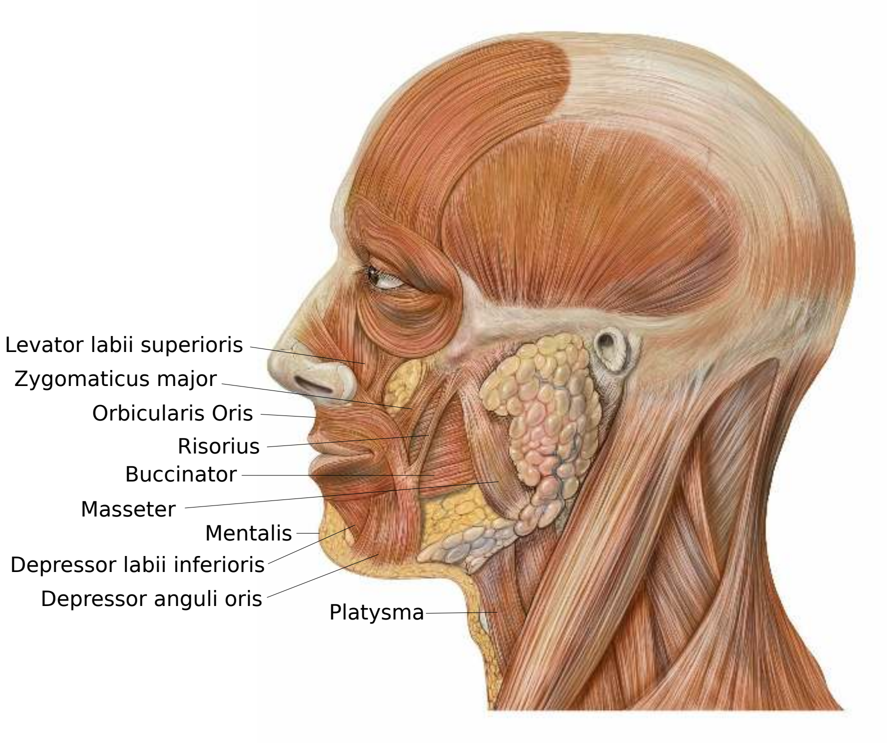


Figure 2.4 – Lateral head anatomy showing the articulatory facial muscles, adapted from [LJ16a]

2.5 Speech Production

Acoustic speech can be regarded as the result of an air-pressure excitation that is modulated by one or more sources and is emitted through a speaker's mouth or nostrils. The main components of the human speech production system are: the lungs, trachea, larynx, throat, nasal cavity and oral cavity. The latter three represent the human vocal tract, which alters the fundamental excitation. As illustrated in Figure 2.5, the anatomy of the human speech production system can be described as follows, in order of the airstream flow [HAH01]:

Lungs: Source of excitating air pressure during speech process.

Larynx: The larynx (or voice box) contains the vocal cords (vocal folds), which are two folds of tissue that can be open or closed to block the air flow from the lungs. They are open during breathing. During speech they can be held close together, which results in a vibration that is excited from the pulmonary airflow. The oscillation of the vocal

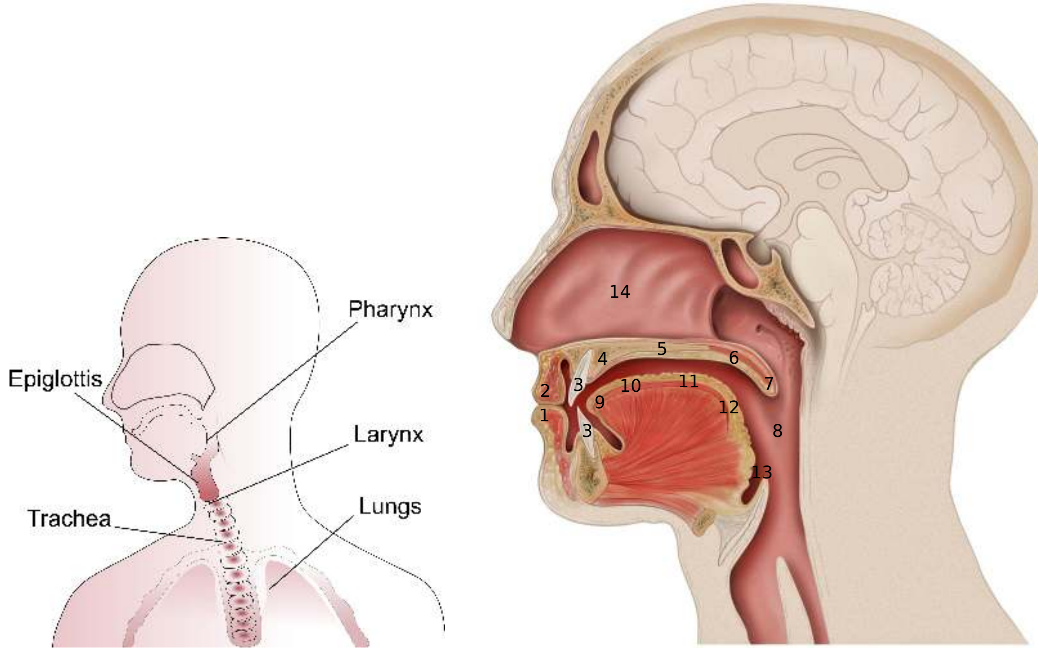


Figure 2.5 – Human throat anatomy (left), from [Wik16] and anatomical diagram of the upper human speech production apparatus (right), from [LJ16b]. 1. Lower lip, 2. Upper lip, 3. Teeth, 4. Alveolar ridge, 5. Hard Palate, 6. Velum, 7. Uvula, 8. Pharynx, 9. Tip of the tongue, 10. Front tongue, 11. Middle tongue, 12. Back tongue, 13. Epiglottis, 14. Nasal cavity.

cords implies a *voiced* sound, thus a quasi-periodic air pressure wave is produced. One important factor in this process is the *fundamental frequency*, i.e. the number of cycles at which the vocal cords vibrate. Influenced by the length, size and tension of the vocal cords, the fundamental frequency can be 60 Hz for men up to 300 Hz for women and children. A lack of vibration of the vocal cords (i.e. when they are too tense to vibrate periodically) implies an *unvoiced* sound. The glottis is defined as the location where the vocal cords come together.

Pharynx: Part of the throat that is placed between the nasal/oral cavity and the larynx (8.) in Figure 2.5).

Oral cavity: Usually this is simply called “mouth”.

Nasal cavity: The nasal tract, above the oral cavity (14.) in Figure 2.5).

The shown positions in the human vocal tract are used to define the point of contact where an obstruction occurs during speech articulation, known as the *place of articulation*. Together with the place of articulation, different sounds in human speech can be characterized by the configuration of the articulators, called the *manner of articulation.* Individual manners are:

Plosive/Stop - an oral occlusive sound, i.e. a sound where the airflow stops completely before resuming again (e.g. the p in “pass”).

Nasal - a sound where the oral cavity is blocked and air flows primarily through the nasal cavity (e.g. the n in “nose”)

Trill - a sound resulting from the repeated opening and closing of the vocal tract, such as a “rolled r”.

Flap/Tap - a momentary closure of the oral cavity (e.g. the t in butter in American English).

Fricative - a sound resulting from turbulent airflow at a place of articulation due to partial obstruction (e.g. the f in “fricative”).

Affricate - a stop changing into a fricative (e.g. the j in “jam”).

Approximant - a sound where there is very little obstruction (e.g. the y in “yes”).

Lateral - an approximant with airflow around the sides of the tongue (e.g. the l in “lateral”).

One fundamental distinction of human speech sounds - called *phones*, is the division into two classes: consonants, that are articulated in the presence of constrictions in the throat or obstructions in the oral cavity, and vowels,that are articulated without significant constrictions and obstructions. However, different placement of the tongue and the oral cavity give each vowel its distinct character by changing the resonance. These major resonances are defined as first (F1) and second formant (F2). The relationship of both formants can be used to characterize different vowels.

This process of human speech generation can be described using a *source-filter model*, depicted in Figure 2.6. The sound excitation signal coming from the lungs can either be voiced or unvoiced, depending on the vocal cords. This reflects the source of the sound, while the vocal tract modulates this excitation source and thereby assumes the role of a filter. The source excitation signal is often modeled as a periodic impulse train, for voiced speech, and white noise for unvoiced speech. Convolution of the excitation signal with the filter $H[z]$ then produces the synthesized speech signal $s(n)$.

The filter component can be estimated using techniques like linear predictive coding or cepstral analysis, which will be discussed in Section 2.6. This

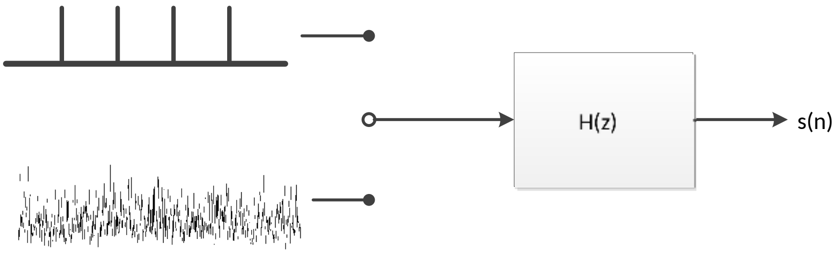


Figure 2.6 – Source-filter model for voiced and unvoiced speech.

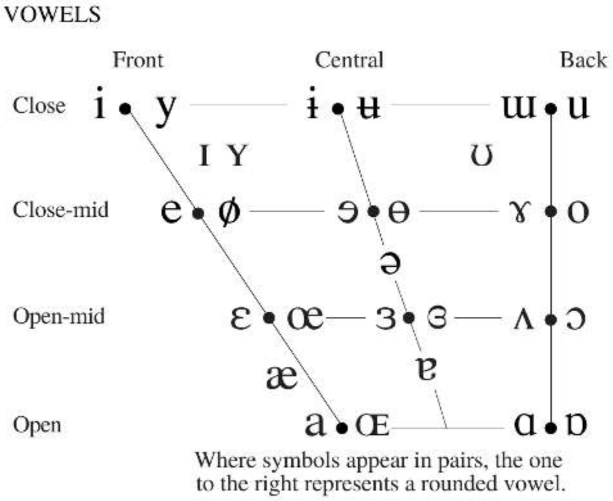
separation of source and filter brings high flexibility in separating the pitch of the source and the filter component.

2.5.1 Phonetical Alphabet

The basic unit to describe speech sounds is defined as a *phone*. A *phoneme* is the smallest unit in speech that serves to distinguish one word from another, resulting in the fact that phonemes are language-dependent. If a single word can be changed into another word by changing one single unit, these units are representations of phonemes (e.g. English words pin vs bin → p and b are phonemes). Different phones that are realizations of the same phoneme are called *allophones*.

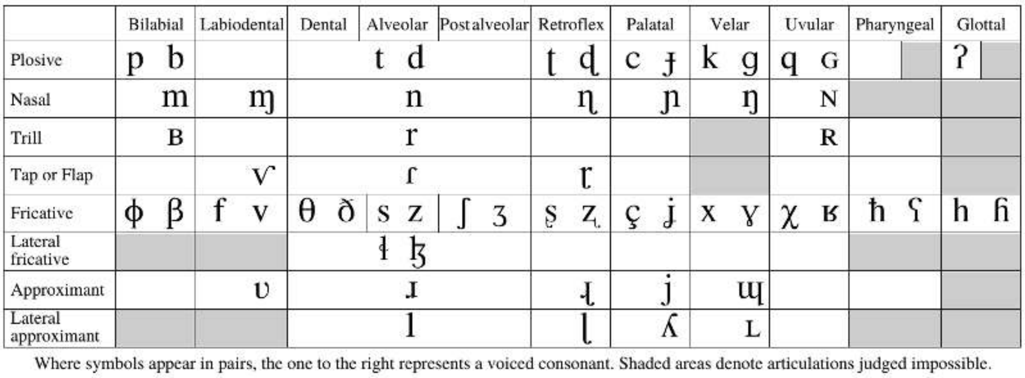
The International Phonetic Association (IPA) [Int99] developed a scheme that groups all pulmonary speech sounds to give a standardized representation of the sounds of oral language. We already introduced one first distinctive separation into vowels and consonants. To further analyze different types of vowels the vowel trapezium is introduced (shown in Figure 2.7). Both dimensions represent the position of the tongue: the horizontal axis discriminates between front and back tongue vowels (vowel backness), e.g. the tongue moves towards the front of the mouth in "bit", while it moves to the back in "but“. The vertical axis discriminates between closed and open vowels (vowel height), related to the vertical position of the tongue, e.g. the tongue is raised to the top of the mouth in 'meet', while lowered to the bottom in "bath“. Additionally, vowels are discriminated between round and unround, depending on the shape of the lips.

For the articulatory description of pulmonary consonants, differentiations are made based on: place of articulation on the horizontal axis (e.g. dental or glottal), manner of articulation on the vertical axis (e.g. nasal or plosive) and



Where symbols appear in pairs, the one to the right represents a rounded vowel.

Figure 2.7 – International Phonetic Association (IPA) vowel trapezium according to [Int99], describing positions of all cardinal vowels and respective IPA symbols.



	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		ⱱ		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 2.8 – List of pulmonary (i.e. obtained by air from the lungs) consonants, grouped by position and manner of articulation, according to [Int99]

voicedness (voiced versus unvoiced). Figure 2.8 gives the detailed consonant classification based on the IPA.

This work uses only English speech data and thus focuses on the characteristics of the English language. Thus, we omit any descriptions and peculiarities of other languages, however we assume that the proposed EMG-to-speech approaches also work with most other Indo-European languages.

2.5.2 Different Speaking Modes

This thesis uses characteristics of speech data from three different speaking modes: normal *audible* speech, *whispered* speech and *silently* articulated speech. Audible speech refers to speech in normal conversational volume and the voiced sounds are produced by modulation of pulmonary air by vibration of the vocal folds. The detailed process of normal speech production is described in Section 2.5.

Whispered speech takes place without major activity of the larynx and is mostly intended for a more private communication. It is thus pronounced more quietly than normal speech [Doe42]. The vocal folds are closed, while the rear part of the glottis is open and builds a kind of triangular shape. The exhaled air produces friction noises in the glottis area. Not only the special vocal fold position distinguishes whisper from normal speech, but also the change of pressure conditions in the sub-glottal space. In whispered speech, voiced sounds are altered, while unvoiced sounds remain unvoiced. For example, if words like “fish” or “six” are whispered, only the vowel is whispered, while the consonants remain voiceless. While it is possible to differentiate between different kinds of whisper, in the scope of this thesis it is sufficient to note that whisper is a strong hissing sound which is produced by a turbulent air flow resulting from a highly constricted glottis.

Silent speech refers to the process when a speaker is not producing audible sounds, while he or she is normally articulating the intended speech. The pulmonary airstream is suppressed, while the conventional speech production process remains unchanged.

The term “silent speech” can be ambiguous, since sometimes different types of speech are referred to as being silent. Three different categories of silent speech are mentioned in the literature:

Imagined speech - also known as thought speech, inner speech or covert speech. The intended speech is only thought of and no articulatory movement is involved. This field is under heavy research in the brain-computer-interface community.

Subvocal speech - Subvocalization, i.e. internal speech used without opening the mouth, is usually observed during reading, when the words are read in mind. Some of the speech articulators - mainly the tongue - slightly move, but the speech process itself can not be noticed by an external bystander.

Mouthed silent speech - sometimes denoted as visible subvocal speech, is regarded as normally articulated speech without emission of air and any underlying sound. Thus, the words are only mouthed. This is the category we are using in this work and that we are regarding as “silent speech”.

For this work we apply mouthed silent speech and normal audible speech on EMG and thus introduce two definitions that are used in the remaining chapters:

Silent EMG: Facial EMG signals that are recorded when mouthed silent speech is produced.

Audible EMG: Facial EMG signals that are recorded when normal speech is produced.

In Chapter 5.2 speaking mode differences are investigated at the EMG signal level.

2.6 Speech and EMG Signal Encoding and Representation

The process of signal processing in this work can be summarized in three stages:

1. Conversion of the input signal (sound pressure waves in air, or ionic myographic current in tissue) to an electrical signal by a standard microphone or EMG electrodes.
2. Sampling of the electrical signal at discrete regularly spaced time intervals, resulting in an analog value per time sample.
3. Quantization of the signal into discrete values (A/D conversion), that can be further processed by technical devices.

Each step needs to be carefully designed in order to avoid erroneous signal recordings. For example, an inappropriate sampling rate can introduce aliasing effects, or a minor quantization resolution can lead to rounding errors.

Speech information, no matter if presented in acoustic or EMG representation, is temporally localized. Thus, it is common to analyze the signal into a sequence of multi-dimensional vectors at regularly spaced intervals called frames. This process is called *windowing*. The windowing is equivalent to multiplying the signal with a function that is non-zero only for a certain amount of time - the *frame length* - starting at the beginning of the signal, shifting it forward a certain amount of time - the *frame shift* - for each successive frame.

2.6.1 Speech Encoding

The speech signal propagates through air as a longitudinal waveform as a series of high-pressure and low-pressure areas, depending on the sound frequency and intensity. The audible frequency bandwidth that is usually used in speech communication ranges up to 8 kHz, resulting in a speech sampling rate of 16 kHz.

One problem in speech processing system is the choice of a compressed representation that still resembles all necessary acoustic information. Illustrating speech sounds in spectrograms gives an important insight for speech analysis, since e.g. formants can be easily spotted. One way to encode the speech signal is the coding of the exact shape of the signal waveform, which is useful when speech and non-speech signals have to be encoded. An example is the Pulse Code Modulation (PCM) which quantizes each sample regardless of their predecessors.

Mel Frequency Cepstral Coefficients Often speech representations are tailored to the human process of speech production and speech perception. The most prominent acoustic features used especially in speech recognition are *Mel Frequency Cepstral Coefficients* (MFCCs). The first step in computing MFCCs is the frame-wise transformation into the frequency domain, e.g. by computing the discrete Fourier transformation.On the short-time spectra a triangular filterbank is used for dimensionality reduction, where adjacent frequency components are bound together. This is based on the mel scale [SVN37] - *mel* referring to the word melody - which is a scale that represents the human auditory perception of pitch, that can be expressed as

$$mel = 2596 \cdot \log_{10}(1 + \frac{f}{700}). \tag{2.1}$$

It mimics the ability of human perception distinguishing lower frequencies at a finer scale compared to high frequencies. Figure 2.9 shows an exemplary mel filterbank, reducing all frequency bands to 12 coefficients. Afterwards each mel spectrum frame is logarithmized and further processed using an inverse

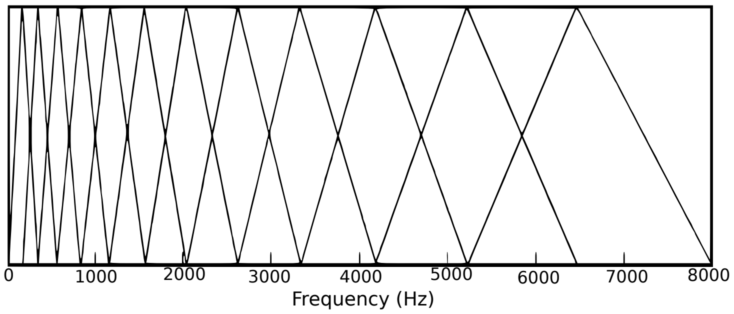


Figure 2.9 – Exemplary mel filterbank reducing to 12 coefficients.

Fourier transform (or alternatively a cosine transform). This transforms the features into the cepstral domain.

This work uses a variation of MFCCs that is implemented for the use of the Mel Log Spectral Approximation filter [Ima83], which generates the acoustic speech signal from input MFCCs. Details on this implementation will follow in Section 2.6.2.

Linear Prediction The coding scheme [AS70] is based on the idea that each observed value $y[n]$ can be predicted as a linear combination of the preceding p values with an error term $e[n]$ called residual signal. The number p of preceding values defines the order of the linear predictive model, $a[i]$ represents the linear prediction coefficients that can be calculated by minimizing the sum of the squared errors per frame. Mathematically this definition can be given as:

$$y[n] = e[n] + \sum_{i=1}^{p} a[i] \cdot y[n-i] \tag{2.2}$$

One drawback using Linear Predictive Coding is its sensitivity to small changes in the coefficients, especially when new data occurs that has not been seen in training. A modification is to use Line Spectral Pairs or Line Spectral Frequencies, which use a transformation of the original LPC data to a space that is more robust [Ita75a].

Fundamental Frequency The fundamental frequency (F0) can be computed using a modification of the auto-correlation method [dCK02]: The framed signal is cross-correlated with itself, and the F0 is extracted as the highest peak in the correlogram within a specified pitch range (manually set, depending on the speaker).

2.6.2 Mel Log Spectral Approximation Filter

The vocoding part of our mapping system generates the final acoustic speech signal using the generated speech features. This work uses the Mel Log Spectral Approximation (MLSA) filter [Ima83]. The implementation uses the fundamental frequency plus a variation of the MFCCs to produce the final acoustic waveform.

Mel Frequency Cepstral Coefficients for MLSA The mel frequency cepstral coefficients (MFCC) implementation that is used in this work is tailored to the MLSA filter. It uses a bilinear warping of frequencies to approximate the mel frequency scale and to obtain a low-dimensional representation of the speech signal. The result can optionally be optimized by applying an iterative method which minimizes the difference between the generated and the original spectrum.

In detail, the MFCC feature extraction algorithm proceeds as follows:

1. Calculation of the power-spectrogram of 70 Hz high pass filtered data using 10 ms frame shift and a 512-point Blackman-Window,
2. Transformation of the spectrogram to its cepstrogram by taking the logarithm and applying the inverse Fourier transform,
3. Application of the frequency warping to transform the frequencies to the mel-scale and reduce the dimension to 25,
4. If desired, iterative optimization steps are performed.

Steps 1 and 2 are straightforward. In step 3, a bilinear, invertible frequency warp is applied in the cepstral domain to approximate the mel frequency scale, in the z-domain it is defined as

$$s = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}} \qquad , \tag{2.3}$$

i.e. mapping the complex unit circle onto itself. The parameter α is chosen as 0.42 in the MLSA implementation. Finally, step 4 consists of an optimization procedure as described by Fukada et al. [FT92]. This (Newton-Raphson-style) optimization is carried out on every frame independently. The cepstral vector is back-transformed into the frequency domain, then the error of this (low-dimensional) representation compared to the original (high-dimensional) spectrum is computed, and in the cepstral domain, the representation is slightly altered to reduce this error. This process is performed iteratively until the iterative improvement falls below a certain threshold, or until the maximum number of iterations is reached. In practice, convergence is usually reached after 4 to 6 iterations. Furthermore, in order to reduce the computation time with respect to real-time applications, the number of iterations can be limited to 1 or 2, as the first step has the largest impact on the final result.

2.6.3 EMG Encoding

For EMG-frames a frame size of 27 ms and a shift of 10 ms was introduced by [WKJ+06] and successfully used for the last decade. The 10 ms frame shift is a factor that is frequently used in speech processing [HAH01]. Since the length of phones is considered at least 30 ms, a 10 ms frame shift corresponds to at least 3 frames per phone. This enables the distinct modeling of begin, middle and end of a phone. Our own preliminary experiments with 5 ms gave only marginally different results.

We use a set of different features for our surface electromyographic data. These time-domain EMG features are based on [JSW+06] and consist of a set of five different features.

In the following description of the features, let $x[n]$ denote the n-th signal value. $w[n]$ is the double application of a nine-point averaging filter:

$$w[n] = \frac{1}{9}\sum_{k=-4}^{4} v[n+k], \quad \text{where} \quad v[n] = \frac{1}{9}\sum_{k=-4}^{4} x[n+k]. \tag{2.4}$$

We define a high-frequency signal:

$$p[n] = x[n] - w[n], \tag{2.5}$$

as well as a rectified high-frequency signal:

$$r[n] = |p[n]|. \tag{2.6}$$

Additionally, for any given feature $\mathbf{f}$, within a frame of size N, $\bar{\mathbf{f}}$ denotes the frame-based mean:

$$\bar{\mathbf{f}} = \frac{1}{N}\sum_{n=0}^{N-1} f[n], \tag{2.7}$$

$\mathbf{P_f}$ denotes the frame-based power:

$$\mathbf{P_f} = \sum_{n=0}^{N-1} f[n]^2 \tag{2.8}$$

and $\mathbf{z_f}$ the frame-based zero-crossing rate (the number of times the signals sign changed). We now define the $TD0$ features as a combination of all of the above:

$$\mathbf{TD0} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_p}, \bar{\mathbf{r}}] \tag{2.9}$$

As time context is crucial for EMG signals that are covering speech information, TD feature vectors are stacked with adjacent vectors to introduce contextual information of the immediate past and future into the final feature.

This stacking can be described as follows: The basic features without any time context are called **TD0**.The 0 represents a stacking context of zero. Consequently, a TD feature vector that has been stacked with its adjacent features N time steps to the past and future is defined **TDN**.

E.g., **TD5** is a **TD0** vector, enhanced with its five respective predecessors and successors. The **TD5** feature vectors of a single EMG channel therefore have $(1+5\cdot2)\cdot5 = 55$ dimensions. Following the same concept, **TD15** feature vectors of a single EMG channel have $(1 + 15 \cdot 2) \cdot 5 = 155$ dimensions.

All direct mapping experiments are done on these TD-features. While they performed best on speech recognition experiments [JSW$^+$06, SW10, WSJS13], we assume that features that capture useful information for speech recognition, also contain information for a direct synthesis approach. Preliminary experiments with different spectral features could not improve our results.

2.7 Mathematical/Technical Background

While the previous sections introduced the details about EMG and speech input and their representation, this section discusses some mathematical principles and algorithms that are used in this thesis.

2.7.1 K-Means algorithm

K-means is an algorithm for grouping a set of N D-dimensional vectors into K clusters, with $N > K$ [Mac05]. The algorithm consists of an initialization step followed by two alternating steps of assignment and update that are repeated until a stopping criterion is met.

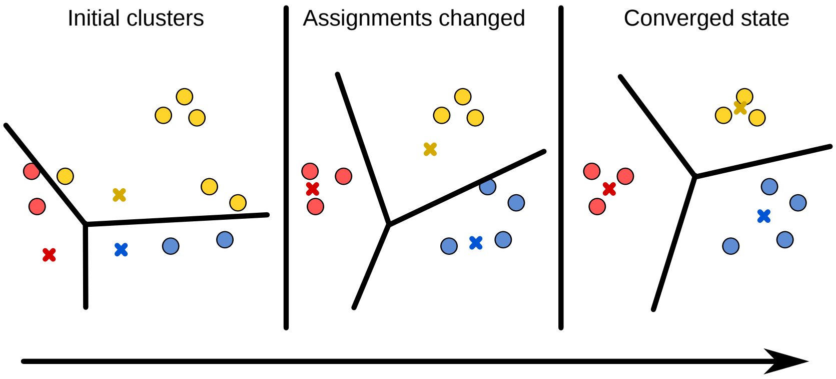


Figure 2.10 – Example for K-means: Clustering 2-dimensional data into 3 clusters using the k-means algorithm, aborting after unit assignments stop changing.

Figure 2.10 demonstrates an exemplary K-means clustering by partitioning a set of 2-dimensional vectors into three clusters. For this purpose, let X_n be the n-th of N input vectors, and $d(x, y)$ a D-dimensional distance metric (e.g. Euclidean distance). Let M_k denote the k-th of K current cluster centroids and $A_n \in \{1 \dots K\}$ the index of the cluster to which the n-th vector is currently assigned to. The procedure is as follows:

1. The set of cluster centroids are initialized. This can be done by e.g. randomly initializing the centroids are simply using the first K input vectors $\{X_1 \dots X_K\}$.
2. Each input vector is assigned to a cluster that has the smallest distance to, thus resulting in
$$A_n = \underset{1..K}{\arg\min}\, d(X_n, M_k)$$
3. Update the centroids by re-computing the cluster means, i.e.
$$M_k = \frac{\sum\limits_{n \in \{m | A_m = k\}} X_n}{\sum\limits_{n \in \{m | A_m = k\}} 1} ,$$
and repeat from step 2 until a stopping criterion is met, e.g. a fixed number of times or until only a fixed fraction of cluster assignments change.

2.7.2 Dimensionality Reduction

After extracting features for our data, the number of features per feature vector (i.e. the dimensionality) may be very high, particularly with EMG features. The "curse of dimensionality" refers to the problem that a high-dimensional input is given in a limited amount of training data. This means that we need to use representations for data that satisfy some properties:

- Generalization must be given, i.e. a good performance on unseen data.
- Stability/robustness: System output is not affected by slight input variations.
- Eliminate high inter-dependent features (e.g. redundancies).
- Discriminability: Extract features that help to distinguish the underlying classes from each other.
- Computational efficiency: Reducing input data to lower dimensions gives noticeable speedup.

In this section two dimensionality reduction techniques are introduced: Linear Discriminant Analysis (LDA) and Canonical Correlation Analysis (CCA). In general, the reduction can also be done using the unsupervised Principal Component Analysis (PCA), which is sensitive to noise and data scaling and thus performs usually worse compared to supervised techniques. Therefore, we only present supervised feature reductions.

Linear Disriminant Analysis (LDA)

(Multi-class) Linear Discriminant Analysis (LDA) is a long-known [Fis36] and popular technique that can be used for dimensionality reduction. It calculates an projection matrix which maximizes the separation between known classes in the training data. Since these class assignments for all input data are needed, LDA is a supervised method.

For k different classes from the training data, the data set of size n is transformed to a $k-1$ dimensional subspace, maximizing the between-class variance, while minimizing the within-class variance.

The between-class scatter matrix $B = \frac{1}{k}\sum_{i=1}^{k}(\mu_i - \mu)(\mu_i - \mu)^T$ is an estimate of the variance between classes, with μ_i being the mean of all data samples that belong to class i and μ represents the mean of all class means.

The within-class scatter matrix $W = \sum_{i=1}^{n}(x_i - \mu_{c(i)})(x_i - \mu_{c(i)})^T$ is an estimate of the variance within the given classes, with $\mu_{c(i)}$ being the mean of the class that sample i belongs to.

LDA finds a projection matrix into a subspace of eigenvectors v, which is arranged in ascending order of discriminability, meaning the first dimensions give the most discriminating features. To ensure class separability, the margin of the class means is maximized, while the in-class-samples are tried to be scattered in a minimal region, resulting in a minimization. Together, following term is maximized:

$$\max_{v} \frac{v^T B v}{v^T W v}$$

The dimensionality of the original input feature can then be reduced by discarding the smallest eigenvalues. After projecting the original feature vectors into linear subspaces, information about which output component is representing which input feature is lost.

Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) is a long known technique that was first published by Harold Hotelling in 1936 [Hot36]. It finds basis vectors for two sets of multidimensional variables such that the correlations between the projections of these variables are maximally correlated. The two sets of variables can also be considered as different views of the same event. Since those views can be any kind of vector representation, there is no restriction to discrete class labels like with the LDA.

One set of variables, the source features, are noted S_x, the other set of target features is defined as S_y. Let w_x and w_y be the basis vectors to which the features are projected upon. CCA finds the w_x and w_y for which the correlation between the projections is maximized:

$$p = \max_{w_x, w_y} corr(S_x w_x, S_y w_y) = \max_{w_x, w_y} \frac{\langle S_x w_x, S_y w_y \rangle}{||S_x w_x|| ||S_y w_y||}.$$

Following Hardoon et al. [HSST04] this can be expressed using the covariance matrices Σ_{xx}, Σ_{xy} and Σ_{yy}:

$$p = \max_{w_x, w_y} \frac{w_x^T \Sigma_{xy} w_y}{\sqrt{w_x^T \Sigma_{xx} w_x w_y^T \Sigma_{yy} w_y}} ,$$

which is equivalent to solving the generalized eigenproblem $\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}w_x = \lambda^2 \Sigma_{xx} w_x$.

Independent Component Analysis

Independent Component Analysis (ICA) [Com92] is a technique used for source separation, i.e. obtaining relevant signals from several observed mixtures of data. In general, it is a particular method for Blind Source Separation (BSS), which can be used to decompose redundant signals into source signal components that satisfy an optimization criterion. The term “blind” refers to the fact that the decomposed source signals, as well as the mixture are unknown. The obtained signals are decorrelated and statistical dependencies are reduced, making the signals as independent as possible from each other.

A common application is its use for high-dimensional electroencephalographic data to separate artifacts (e.g. eye movements or cardiac signals) from electrical signals that are related to a certain brain activity [MBJS96]. For further details on ICA we refer to the detailed descriptions in [HO00].

2.8 Time Alignment

Different acoustic renditions of the same utterance result in different durations based on e.g. different speaking rates. Thus, to compare for example the frames of whispered and normally spoken utterances, there is a need for aligning both speech signals. One solution to this problem is *Dynamic Time Warping* (DTW), which is widely used in speech recognition [Ita75b]. This algorithm efficiently “warps” the time axes of two input sequences $X = x_1, x_2, \dots, x_n$ and $Y = y_1, y_2, \dots, y_m$ to align them to each other. The first step is the construction of an n-by-m distance-matrix, where each matrix element (i, j) contains the distance $d(x_i, y_j)$ between two points x_i and y_j of the input sequences. Typically, a simple distance metric is used (e.g. the Euclidean distance). Based on this distance matrix, a warping path $w = w_1, w_2, \dots, w_l$ is obtained, which defines the mapping from one input sequence to the other. This warping path is subject to three certain constraints [KR05]:

- Boundary condition: $w_1 = (1,1)$ and $w_l = (n,m)$.
- Continuity of step-size: $w_{i+1} - w_i \in \{(0,1),(1,0),(1,1)\}, \forall i \in [1 : l-1]$. This simply restricts the step-sizes of the warping path to its adjacent elements.
- Monotonicity: $n_1 \leq n_2 \leq \ldots \leq n_l$ and $m_1 \leq m_2 \leq \ldots \leq m_l$.

The dynamic warping path can then be found using the path with the minimal total distance among all possible paths, which can be computed using dynamic programming.

2.9 Evaluation Metrics

An appropriate measure for the evaluation of speech quality is a difficult task and under heavy investigation. The main goal is to achieve an *intelligible* output, but to establish a satisfying result, a system needs to also guarantee a *natural* output. Both aspects are based on multivariate influences, including factors like the presence of noise, the kind of noise, breathy or robotic speech, etc. It becomes clear that it is hard to quantify such a quality aspect and that every numeric measurement implies an oversimplification. However, not only the quantification, also the interpretation of a specific term like "intelligibility" is error-prone or at least complicates evaluations. Intelligibility covers the amount of information that is understood. So a first evaluation type could be to simply write down the words that were perceived. However, what about prosodic information? Was the perceived sentence uttered as a statement or a question? Was the speaker emphasizing certain words? Or is this aspect only covered by investigating the "naturalness" of the output?

It is also observed that a concept like "naturalness" may be underspecified [DYK14], regarding that there is no concrete definition for this term. This results in the fact that different studies give different instructions for naturalness evaluation. [DYK14] refers to the example that e.g. the Blizzard evaluation [PVE$^+$13] instructs the listener to "reflect your opinion of how natural or unnatural the sentence sounded. You should not judge the grammar or content of the sentence, just how it sounds." In contrast [AEB12] defines naturalness to rate the probability "that a person would have said it this way?" The second instruction takes into account the grammar and content, standing in contrast to the first definition.

These examples give a first glance at the problems that are related to subjective evaluations of synthesized speech. Despite these problems it is common to estimate both subjective and objective speech quality measures.

2.9.1 Subjective Evaluation

Subjective – also called psychophysical – listening tests are based on the judgment of human listeners. Though it is inconvenient to run a listening test on every incremental improvement of a synthesis system, subjective evaluations represent a solid way to evaluate an output since they are influenced by the most important value for the output: the human. A commonly used metric is the Mean Opinion Score (MOS), where a number of listeners rate the quality of synthesized utterances on a numerical scale. MOS is also defined in ITU-T Rec. P.10 [IT06] (in the telecommunication/transmission domain) as "the value on a predefined scale that a subject assigns to his opinion of the performance of the telephone transmission system used either for conversation or only for listening to spoken material." The commonly used perceptual MOS scale is:

5 (excellent), 4 (good), 3 (fair), 2 (poor), 1 (bad).

Another way to evaluate an output is to explicitly compare two or more utterances, called an *A/B preference test.* The listener hears the same utterance from two different output methods and chooses the one that sounds better. A variation of this technique is to additionally present (if available) the original speech file, so that the subject can compare and choose which of the outputs is closer to the target speech.

2.9.2 Objective Evaluation

While subjective evaluations are an "expensive" way (in terms of time, effort and reproducibility), a different option is the application of computer-derived objective evaluations. A summary of some objective speech quality measures can be found in [HL08]. We list only those that are commonly used in related work and thus are deployed in this work.

Mel-Cepstral Distortion

Mel cepstral distortion, sometimes also called mel cepstral distance, (MCD) [Kub93] is a logarithmic spectral distance measure using the scaled Euclidean distance between MFCC vectors. It is computed between a synthesized utterance (estimate data) and the reference utterance (target data). This can mathematically be formulated as the distance between an estimated MFCC-vector $\mathbf{MFCC_{est}}$ and the reference $\mathbf{MFCC_{ref}}$ at the frame-level as:

$$\mathbf{MCD_f} = 10/\ln 10 \sqrt{2 \cdot \sum_{k=2}^{25} (\mathbf{MFCC}_{est}[k] - \mathbf{MFCC}_{ref}[k])^2} \tag{2.10}$$

In this work, we use 25 MFCCs, resulting in a 25-dimensional MFCC vector. The first coefficient $k = 1$ can be excluded for MCD computation, since it represents the power of the speech signal, which is primarily affected by recording gain rather than signal distortion. Since some other publications include the first coefficient, MCD numbers are not always directly comparable. It is also worth mentioning, that MCD comparability is not always given, due to its dependence on factors like the MFCC implementation, or the frame size.

Since MCD represents a distance measure, lower numbers imply better results. In Voice Conversion research reported MCD results range from 4.5 to 6.0. Raghavendra et al. [RVP10] observe that an MCD difference of 0.2 produces a perceptual difference in the quality of synthetic speech.

The utterance-level MCD for an utterance with n frames is given by:

$$\mathbf{MCD} = 1/n \cdot \sum_{i=1}^{n} \mathbf{MCD_f} \tag{2.11}$$

Since only the similarity of an estimated output to an existing target can be measured, MCD is not fully reliable to rate the general quality of an output.For example, it does not take prosodic effects into account.

F0 Voiced/Unvoiced Error Rate

While the MCD gives a simple and reliable distance measure on MFCCs, when it is compared to a reference signal, we evaluate the estimated F0 excitation with 2 different metrics.

One way is to account for the recognition of the voicing states, in other words to rate the unvoiced/voiced (U/V) decision error on a frame-basis. Regarding silence frames as unvoiced, each estimated (and reference) frame can be either voiced or unvoiced. Thus, we have 4 possibilities: a voiced frame is estimated correctly as voiced ($V \rightarrow V$), an unvoiced frame is estimated correctly as unvoiced ($U \rightarrow U$), a voiced frame is estimated falsely as unvoiced ($V \rightarrow U$), an unvoiced frame is estimated falsely as voiced ($U \rightarrow V$).

The V/U decision error uses the number of voiced and unvoiced errors ($N_{V \rightarrow U}$ and $N_{U \rightarrow V}$) and is defined as:

$$\mathbf{V/U\text{–}Error} = \frac{N_{V\rightarrow U} + N_{U \rightarrow V}}{N} \cdot 100\% \ , \tag{2.12}$$

where N is the total number of frames.

Sometimes the V/U-Accuracy is reported, which is given by:

$$\mathbf{V/U\text{–}Accuracy} = 100\,\% - \mathbf{V/U\text{–}Error} \tag{2.13}$$

This results in a simple metric, accounting for the voicedness of an utterance, yet it gives only little insight to the generated prosody itself.

F0-Correlation-Coefficient

To evaluate the prosodic contour of the generated F0, we also account for the correlation coefficient between all frames of an estimated and a reference speech signal.Correlation coefficients are calculated on an utterance-basis as follows:

$$\mathbf{Corr(est, ref)} = \frac{Cov(est, ref)}{\sigma_{est}\sigma_{ref}}, \tag{2.14}$$

where $Cov(est, ref)$ denotes the covariance between estimated and reference data. σ_{est} and σ_{ref} denote standard deviation of estimated and reference data, respectively. $Corr(est, tar)$ ranges between -1 and 1, where -1, 0, and 1 refer to negative perfect correlation, decorrelation and positive perfect correlation, respectively.

Chapter 3

Related Work

This chapter gives a compact overview of relevant work in the field of biosignal-based speech processing systems, with a focus on myographic input signals. Work related to individual mapping approaches will be discussed in the respective chapters.

3.1 EMG-based Speech Communication

EMG-based analysis of speech production and the articulatory muscles is a topic that is investigated since several decades [TM69, GHSS72, KP77, Hir71]. Initially, instrumentation and recording equipment was far from being user-friendly, e.g. only intramuscular wired electrodes were available. First investigations on silent speech based on this equipment were already published in the 1950s [FAE58].

In the following decades, user-friendly surface electrodes became more and more popular. The first publications describing EMG-based speech recognition date back to the mid 1980s, when Sugie and Tsunoda [ST85] performed a first study on the discrimination of five Japanese vowels based on muscular activity in the oral region recorded using three surface electrodes. Nearly at the same time, Morse et al. [MO86] investigated the suitability of myoelectric signals recorded from four channels (three channels in the neck-area, one channel at the forehead) to attain speech information. Multiple experiments with different sets of words on two different speakers were conducted, achieving an accuracy of 60 % on a 10-digit English vocabulary. Competitive performance was first reported by Chan et al. [CEHL01], who achieved an average accuracy of 93 % on a 10-digit vocabulary of English. A good performance could be achieved Jorgensen et al. [JLA03] even when words were spoken non-audibly, i.e. when no acoustic signal was produced, suggesting this technology could be used to communicate silently. Manabe et al. [MHS03] used a setup where electrodes are attached to the speakers hand, which is moved to the face during recording.

While the former approaches used words as model units, later research [JSW+06, WKJ+06] successfully demonstrated that phonemes can be used as basic model units for EMG-based speech recognition, paving the way for large vocabulary continuous speech recognition. Recent results include advances in acoustic modeling using a clustering scheme on phonetic features, which represent properties of a given phoneme, such as the place or the manner of articulation. Schultz and Wand [SW10] reported that a recognizer based on such *bundled phonetic features* outperforms phoneme-based models by more than 30 %.

Today, a couple of research groups are driving EMG-based speech processing forward [MSH+08, Lee10, JD10, DCHM12, KYHI14], some of them with a particular focus, e.g. investigating Portuguese language specific factors [FTD12] or focusing on disordered speakers [DPH+09]. There also exist multimodal approaches, e.g. combining acoustic signals and EMG signals [KNG05, SHP07, ZJEH09, DPH+09] .

Approaches with a direct transformation of EMG signals to speech will be described in Section 3.3.

3.2 Biosignal-based Speech Interfaces

Systems that are summarized under the keyword “Silent Speech Interface” [DSH+10], define approaches that do not need audible speech signals to process the actual speech. Apart from the already discussed EMG-based processing technologies, other types of input signals can be used:

- Video camera based lip reading approach [HO92].
- Permanent magnetic articulography (PMA) or electromagnetic articulography (EMA) [FEG+08, GRH+10]: Magnets are attached to the

articulators, which enable a capturing of the varying magnetic field by sensors located around the mouth.

- Ultrasound and optical imaging of the tongue and lips [DS04, HAC$^+$07, FCBD$^+$10, HBC$^+$10].
- Non-audible murmur (NAM) microphones [Nak05, NKSC03, HOS$^+$10, TBLT10, HKSS07]: type of stethoscopic microphone that still records acoustic sound wave, but which are conducted via the body and thus are nearly inaudible.
- Glottal activity detection: based on electroglottograph (EGG) [Rot92, TSB$^+$00] or vibration [BT05, PH10], which both records activity across the larynx during speech.
- Brain computer interfaces: based on electroencephalographic [PWCS09, DSLD10], near infrared sensors [VPH$^+$93, FMBG08] or implants in the speech-motor cortex [BNCKG10, HHdP$^+$15]. The first two mentioned approaches are mainly used for imagined speech - due to its high sensitivity to motion artifacts - and only slightly related to this thesis. However, they should be listed since they reveal insights into the processing for silent speech interfaces.

3.3 Direct Generation of Speech from Biosignals

A direct speech generation approach has advantages compared to the recognition-followed-by-synthesis technique, although this may be linked to a possible degradation of the performance, e.g. by omitting the language model. There is no restriction to a given phone-set, vocabulary or even language, while there exists the opportunity to retain paralinguistic features, like prosody or speaking rate, which is essential for a natural speech communication. The speaker's voice and other speaker-dependent factors keep preserved. Additionally, a direct mapping enables faster processing than speech-to-text followed by text-to-speech and thus the ability for nearly real time computation that can also be used for a direct feedback. A couple of research groups investigated these direct speech synthesis techniques.

Lam et al. [LLM06] introduce a first EMG-based speech synthesis approach. Using a two layer neural network on 2 channels (a cheek and a chin channel) of audible EMG, the authors report on a frame-by-frame feature mapping that is trained on seven different phonemes. The neural network is used to generate eight different words. 70 % of these words can be synthesized correctly. It remains unclear how this performance is evaluated.

Tsuji et al. [TBAO08] present a synthesis approach that involves a phone recognition component, rather than a direct transformation of signals. A neural network is used for phoneme classification, which is further processed by hidden Markov models and the final voice generation is obtained using a synthesizer software component. This clearly has the drawback that phone labels are needed for training the system.

Further results on direct EMG-to-speech research are reported by Toth et al. [TWS09], where a GMM-based technique is used without any subword-based recognition steps. Five EMG channels are obtained from single-electrodes recordings with 380 utterances (about 48 minutes) of training data from audible EMG and speech. An average MCD of 6.37 with a standard deviation of 2.34 is obtained. The authors additionally perform a speech recognition experiment using the synthesized speech output. Restricting the testing vocabulary to 108 words, 84.3 % of the words were recognized correctly. First results on silent EMG-to-speech generation are also presented, however this reduces the word recognition accuracy to 20.2 %. Due to the lack of an audible reference for the silent EMG data, no MCD numbers are reported for silent EMG data.

Ultrasound images from the tongue are used to directly generate a speaker's vocal tract parameters in work from Denby et al. [DS04]. The resulting speech is stated to be of poor quality, but shows many of the desired speech properties. Hueber et al. [HCD^{+}07] use ultrasound and additional video images to drive a combination of HMM and Unit Selection approach to synthesize speech. In this promising approach, no objective evaluation scores are presented. The authors state that the synthesized speech is of good quality for correctly predicted sequences; however, the number of errors is still to high to produce a truly usable output signal. The authors [HBDC11] additionally present a GMM-based approach that converts lip and tongue motions, captured by ultrasound and video images into audible speech. Using 1,020 training sentences they achieve an MCD of 6.2.

Non-audible murmur based transformation into whispered speech has been investigated by Nakagiri et al. [NTKS06]. Although the recorded signals are closely related to normal speech waves, they are nearly silently articulated and therefore can be regarded as silent speech signals. The authors report the challenges in NAM-to-F0 conversation and thus propose a whisper output, using NAM signals in addition with Body Transmitted Whisper signals (BTW). Using the combination of both signals, the best MCD is stated with 4.4 and additionally an F0 correlation coefficient of 0.62 is reported. Converting whisper-like signals to normal speech is proposed by Tran et al. [TBLT08]. A Gaussian mixture model approach and a neural network mapping is used to generate F0, resulting in 93.2 % voiced/unvoiced accuracy using neural networks, and 90.8 % using GMMs. A correlation coefficient of 0.499 is reported on the F0 contour. The authors continue their investigations [TBLT10], reporting an MCD of 5.99, which improves to 5.77 when visual facial information is added.

Electromagnetic articulography (EMA) data is used as a source signal in research by Toda et al. [TBT04]. A GMM-based voice conversion approach is used, a system that is also the basis for the EMG-to-speech research by Toth et al. [TWS09]. The reported average MCD using EMA data is 5.59 with a standard deviation of 2.23, which gets significant improvement when F0 and power features are added to the input, resulting in 4.59 with a standard deviation of 1.61. This approach is improved in [TBT08], achieving an average MCD of 4.5 with the best speaker-dependent MCD of 4.29.

3.3.1 Pitch Generation from EMG

Generating a natural prosody is one of the most challenging problems in speech synthesis. Thus, we list the related work that was done in this field, when instead of text or speech input, other biosignals are used. Most of this research field is intended for people with speaking disabilities who are forced to use an additional prosthetic speaking device: an electrolarynx. This device can be regarded as baseline and used for comparison to the proposed pitch generation approaches. An electrolarynx generates an excitation sound using an electromechanical vibration of a diaphragm when a specific button is pressed. This gives an intelligible speech output, however it sounds very robotic and unnatural.

Considering electromyographic input, Balata et al. [BSM$^+$14] recently reviewed the existing research state regarding the investigation of electrical activity of the larynx during phonation. 27 comparable publications were selected, which shows the interest in this field of research.

One of the first EMG-based investigations was done in 1979, where Shipp et al. [SDM79] use a logistic regression to predict F0, using invasive needle electrodes. [SHRH09, KSZ$^+$09] compare an EMG-controlled electrolarynx to a normal electrolarynx. The spectral envelope from EMG signals of face and neck muscle is used to set the F0 output in a threshold manner, but no detailed comparison to the F0 signal is performed. Similar work is done by DeArmas et al. [DMC14]. A support vector machine is used to predict F0 using two channels of EMG data from the neck muscles, achieving an average voicing state accuracy (voiced/unvoiced accuracy) of 78.05 %.

An approach that is based on Gaussian mixture models and competes with the work from DeArmas [DMC14] is presented by Ahmadi et al. [ARH14]. Pitch generation is performed using data from the neck muscles. Using an improvement on the feature vector they report a correlation coefficient of 0.87 and a voiced/unvoiced accuracy of 87 %. Investigations in tonal languages like Cantonese are reported in [YLN15].

Although surface EMG only records a summation of multiple muscles, many of the proposed EMG-to-F0 approaches focus and motivate on the detection of laryngeal muscles. This contradicts its usage for an application with e.g. laryngectomees, where the voice box has been completely removed.

3.4 Summary

The related work introduced in this chapter shows the wide interest and longstanding research in the field of direct speech generation using a biosignal input. Although these approaches gradually improve and the used recording devices become considerably affordable, many restrictions still exist. Some approaches work only on specific laboratory-like environments (e.g. EMA) or the output is restricted to a limited domain (e.g. [LLM06] train on 8 phonemes only).

Denby et al. [DSH$^+$10] compare EMG to other biosignals and endorse its high potential in terms of cost, non-invasiveness, silent-usage and other factors. While Toth et al. [TWS09] introduced a first EMG-to-speech approach, this thesis uses multi-channel grid-electrodes and advances the results by examining different mapping approaches, paving the way towards an end-user system.

CHAPTER 4

Data Corpora and Experimental Setup

This chapter gives an overview of the EMG and speech data recordings plus the used equipment, software and hardware. We start with an introduction of the different EMG setups and the general recording setup, concluded by details on the used corpora. Additionally, first analyses of the recorded data are presented.

4.1 EMG and Speech Acquisition System

Two different EMG recording setups (see Figure 4.1) were used in the research for this thesis:

A Single-Electrodes Setup using six EMG channels from multiple facial electrode positions.

An Electrode Grid/Array Setup using 35 EMG channels from an electrode grid on the cheek and one electrode array on the chin.

The latter was introduced in the scope of this thesis and involves an easier electrode attachment and new opportunities for signal processing in contrast to the generally used single-electrodes. The single-electrodes setup was first presented in 2005 [MHMSW05] and was only slightly changed over the last years. Details will be given in Section 4.1.2. The array-based setup is a

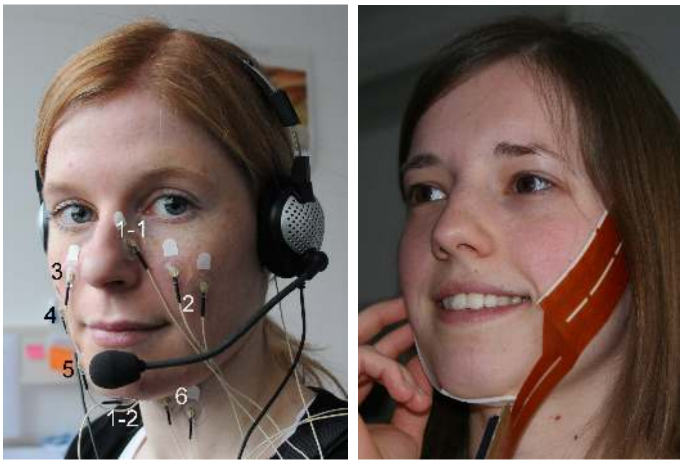


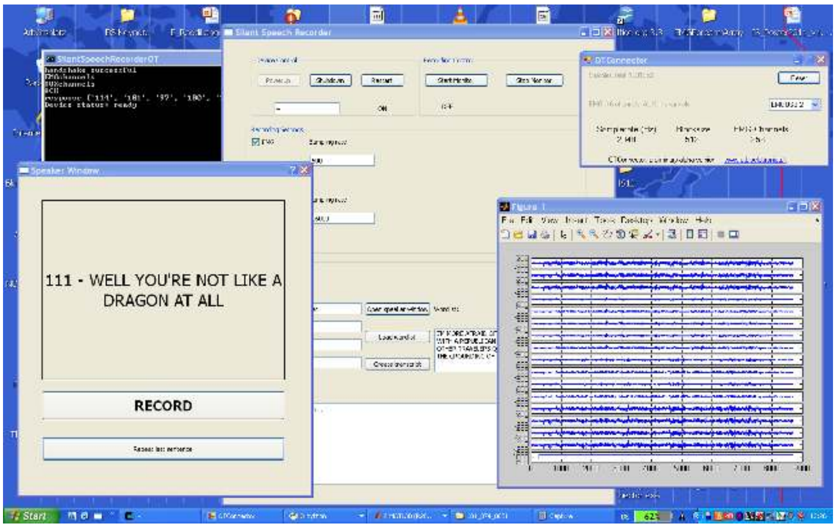
Figure 4.1 – The single-electrodes recording setup (left) and electrode-array recording setup (right).

novel recording system that was first described in [WSJS13] for EMG-based speech recognition, although the use of electrode arrays for EMG was already introduced in the 1980s [MMS85]. See Section 4.1.3 for further descriptions of the array setup. For both setups an audio signal was synchronously recorded with a standard close-talking microphone. The general process of recording kept the same, independent from the type of setup.

4.1.1 General Recording Setup

All recorded data was obtained with a standard close-talking headset and an EMG-recording system in parallel. Details on the recorded utterances follow in Section 4.2. For audio recording we used an ANDREA NC95 headset with inbuilt noise reduction. The headset microphone was connected to a Hercules Gamesurround Muse XL Pocket LT3 USB stereo sound adapter, where the audio signal was saved into the first audio channel (out of two). The other channel contains an analogue marker signal used for synchronization of the acoustic signal with the EMG signal, as described below.

We recorded two different speaking modes: audible speech and silent speech. Recordings were done on a laptop with the in-house developed recording software BiosignalsStudio [HPA$^+$10]. The recording process was as follows: The program presented a sentence of the corpus on the screen to the recording participant. For recording this sentence, the subject had to press the RECORD button and hold it while speaking this sentence. The resulting utterance was considered “valid“ if the pronunciation of each word in the sentence had been articulated properly in English. If the speakers skipped or added words or letters, produced stuttering or pauses that were longer than a regular speech pauses, the utterance was repeatedly recorded until the pronunciation of the complete sentence was valid. Furthermore, the recording participants were allowed to practice the pronunciation if they felt it to be necessary. We define the amount of data that was recorded between attachment and detachment of the electrodes as one recording session. If the electrodes were removed and attached again on the same day, we consider this a new session. Figure 4.2 shows a screen with the existing recording elements. During recording, only the speaker window (bottom left) is presented to the subject, while the recording assistant inspects the EMG signals on the EMG visualization on a second screen. The remaining windows are used for sentence selection and EMG device management.



```
handshake successful
EMGchannels
AUXchannels
ACK
response ['111', '101', '77', '100', '
Device status: ready
```

Figure 4.2 – The Graphical User Interface for the data recording.

All recordings were performed in the same room and with the same equipment and settings and were supervised by a recording assistant. For the recording of audible speech, subjects were told to speak as normally as possible, for the recording of silent speech, subjects were instructed to imagine a meeting situation in which he or she was talking to a person next to him or her without generating any sound. This scenario was intended to prevent hyperarticulation.

4.1.2 Single-Electrodes Setup

For single-electrodes EMG recording, we used a computer-controlled biosignal acquisition system (Varioport, Becker-Meditec, Germany).

The recording device can use a Bluetooth connection, as well as a serial port, to communicate with a connected computer system. The built-in A/D-converter can be operated with different sampling rates, where a rate of 600 Hz was chosen. The single-electrodes are standard Ag/AgCl electrodes with a diameter of 4 mm.

All EMG signals were sampled at 600 Hz (covers the main spectral part of up to 200 Hz [Hay60]) and filtered with an analog high-pass filter with a built-in cut-off frequency at 60 Hz. We adopted the electrode positioning from [MHMSW05] which yielded optimal results. Our electrode setting uses six channels and is intended to capture signals from the levator anguli oris (channels 2 and 3), the zygomaticus major (channels 2 and 3), the platysma (channels 4 and 5), the depressor anguli oris (channel 5) and the tongue (channels 1 and 6). Channels 2 and 6 use bipolar derivation, whereas channels 3, 4, and 5 were derived monopolarly, with two reference electrodes placed on the mastoid portion of the temporal bone (see Figure 4.3). Similarly, channel 1 uses monopolar derivation with the reference electrode attached to the nose.

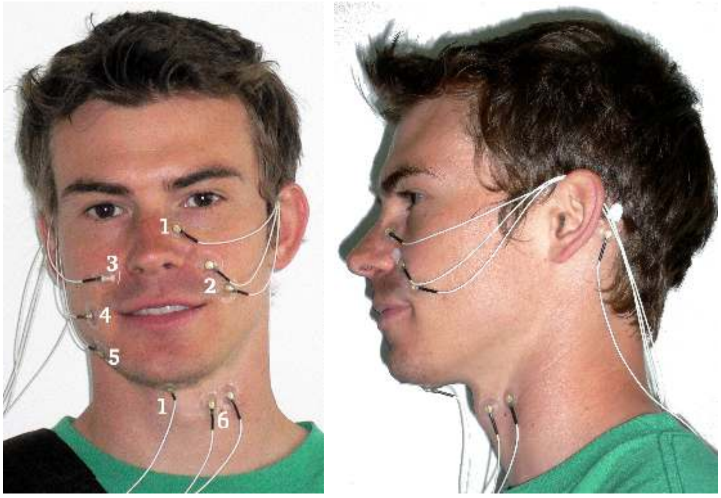


Figure 4.3 – Electrode positioning of the single-electrodes setup resulting in six EMG channels. Reference electrodes for monopolar derivation are placed on the nose (channel 1) and on the mastoid portion of the temporal bone (channels 3,4 and 5).

4.1.3 Electrode-Array Setup

The data acquisition using electrode-arrays is a next step towards practical usage. Compared to the attachment of single-electrodes the setup is much easier and less time consuming. In addition, a higher amount of EMG channels is available due to the large number of electrodes.

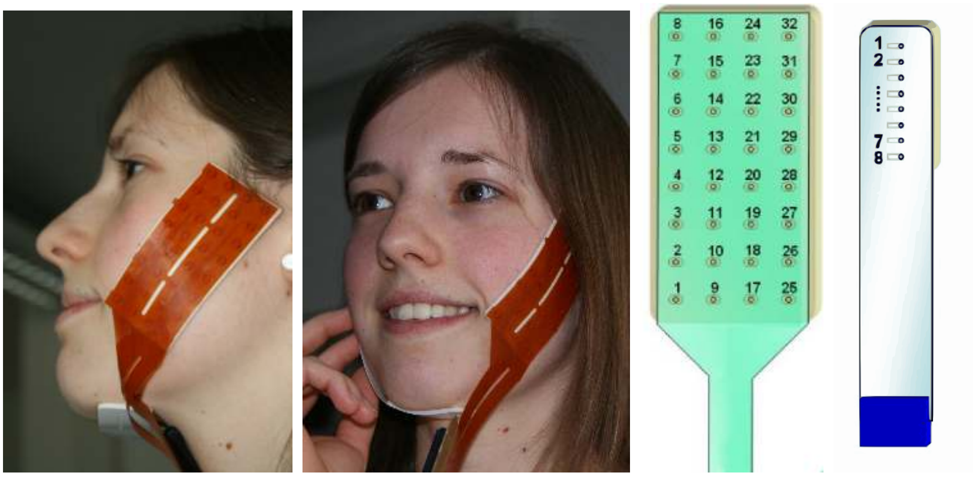


Figure 4.4 – Left side: Electrode array positioning: electrode grid is positioned on the cheek, small array under the chin. Right-hand side: Channel numbering of the 4 × 8 cheek electrode grid, and the 8 electrode chin array.

For array-based EMG recording we use the multi-channel EMG amplifier EMG-USB2 produced and distributed by OT Bioelettronica [Ot2]. The EMG-USB2 amplifier allows to record and process up to 256 EMG channels, supporting a selectable gain of 100 – 10,000 V/V, a recording bandwidth of 3 Hz – 4,400 Hz, and sampling rates from 512 to 10,240 Hz. After some initial experiments, we chose an amplification factor of 1,000 V/V, a sampling frequency of 2,048 Hz, and a high-pass filter with a cutoff frequency of 3 Hz and a low-pass filter with a cutoff frequency of 900 Hz. These numbers follow the suggestions in [FC86]. For power line interference reduction and avoidance of electromagnetic interference, we use the amplifier-integrated Driven Right Leg (DRL) noise reduction circuit [WW83]. To activate the DRL noise reduction circuitry, a ground strip is connected to the recording participant at a position with no bioelectrical activity (wrist or ankle). Using the provided cable, the strip is connected to the DRL IN connector. An additional ground strip must be connected at a point with no bioelectrical activity on the subject and to the DRL OUT connector. To ensure that we have no interference from the muscles of the right hand when pushing the recording buttons, we attach all strips to the left wrist. In our recording setup, we always use the DRL noise reduction circuitry. The electrode arrays and the electrolyte cream were acquired from OT Bioelettronica as well. The cream is applied to the EMG arrays in order to reduce the electrode/skin impedance.

We capture signals with two arrays, as illustrated in Figure 4.4: One cheek array consists of four rows of 8 electrodes with 10 mm inter-electrode distance (IED), and one chin array containing a single row of 8 electrodes with 5 mm IED. The original electrode array includes 8 rows of 8 electrodes, which is too disturbing to the speaker when the array is attached to the face. We therefore cut off the 2 outer rows of the array, so our modified array contained 4 × 8 electrodes. The numbering of the final electrode channels is depicted in Figure 4.4. We use a bipolar derivation mode, thus the signal differences between two adjacent channels in a row are calculated. We therefore obtain a total of 35 signal channels out of the 4 × 8 cheek electrodes and the 8 chin electrodes [WSJS13].

EMG-Array-positioning

Processing speech by means of EMG signals requires signals at least from the cheek area and the throat, where tongue activity can be captured. Also, we rely on the fact that the articulatory activity is rather symmetric on both sides of the face, so that recording one side of the face suffices. It has turned out to be very difficult to capture signals in the close vicinity of the mouth, since neither single-electrodes nor electrode arrays can be attached reliably.

Since the single-electrodes positions were carefully selected and evaluated in [MHMSW05], the array positioning was based on the prior single-electrodes positions. The smaller 8 electrode chin array is used mainly for recording the activity of the tongue muscles and is thus attached to the speakers neck, where there is only little variation on the positioning possible.

We tested several different positions for the cheek electrode grid with respect to both, the signal quality and the interference with the subject's speech production. The cheek array should not be attached too close to the speakers mouth to not interfere with the speech production process. Furthermore, we observed that it was difficult to attach the array firmly to the skin, which resulted in movement artifacts.

The cheek array was always placed approximately 1 cm apart from the corner of the mouth and then rotated around this focal point for further analysis. We started with a nearly vertical position of the array and then rotated by approximately 20 degree clockwise for each new experiment. After 40 degree rotation the subject was negatively affected by the recording cable connection and we stopped. Thus, we focused on these three recording settings: vertical, 20 degree rotation, 40 degree rotation (illustrated in Figure 4.5).

We tested this setup on three speakers. Two of three clearly preferred the middle position and since this position resulted in the least interferences for the speaking process, we used this position for the final attachment of the electrode grid.

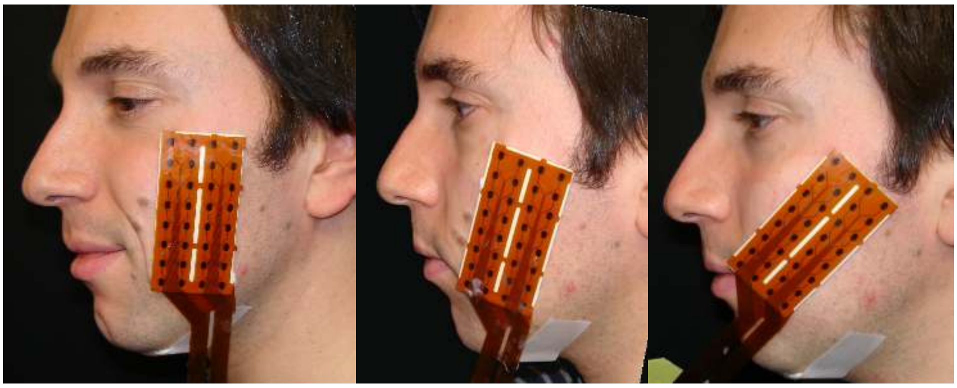


Figure 4.5 – The three different positionings of the cheek electrode array: vertical (left side), 20°(middle), 40° (right-hand side).

4.2 EMG Data Corpora

Using the introduced two EMG recording setups, we grouped the recording sessions based on the amount of recorded data. We additionally recorded a simplified data corpus, which consists of CVCVCVCV nonsense words, where C and V represent selected vowels and consonants. This gives us the opportunity to have a closer look at the signal characteristics and enables a more fine-grained investigation than using whole sentences. Thus, we use three different corpora in this thesis, where only the latter two are used for the EMG-to-speech transformation.

EMG-Array-AS-CV Corpus consisting of *audible and silent* multichannel array-based nonsense words.

EMG-Array-AS-200 Corpus consisting of 200 *audible and silent* multichannel array-based EMG utterances.

EMG-ArraySingle-A-500+ Corpus consisting of at least 500 *audible* utterances from array-based and single-electrodes recordings.

4.2.1 EMG-Array-AS-CV Corpus

The EMG-Array-AS-CV Corpus consists of simplified acoustic and EMG data. The spoken utterances are based on the *mngu0* articulatory-phonetic corpus [RHK11]. One male recording subject, who already participated in multiple EMG recordings, speaks CVCVCVCV nonsense words; C and V representing a selected vowel and consonant, that keeps constant during one utterance. Two sessions with audible and silent data from this speaker have been recorded on different days.

We selected five different vowels (/a/,/e/,/i/,/o/,/u/) in combination with nine different consonants that can be grouped into:

Plosives: /b/, /d/, /k/, /t/

Nasals: /m/, /n/

Vibrants: /r/

Approximants: /l/

Fricatives: /s/

This results in $9 \cdot 5 = 45$ audible and 45 silent speech utterances, giving a total of 90 utterances per session. The second recording session follows the same setup.

4.2.2 EMG-Array-AS-200 Corpus

The EMG-Array-AS-200 corpus consists of normal and silent read speech of five recorded participants. Three of these subjects recorded multiple sessions at different days to investigate intra-speaker-dependent properties. The other subjects recorded one session each. In total, the corpus consists of 12 recording sessions. The participants were four male and one female speakers, with their age ranging from 21 to 31 years.

Each recording session consisted of one en-bloc recording of 200 phonetically balanced English sentences spoken in audible speech mode and one en-bloc recording of the same 200 sentences in *silent speech mode*, i.e. the sentences are identical for both the audible and silent speaking mode. The order of speaking mode recordings was switched between speakers to eliminate any possibilities of ordering effects. We used three different sets of sentences, where each speaker recorded the first sentence set in his or her first recording session. Each sentence set consists of 200 phonetically balanced English sentences originating from the broadcast news domain, with 20 sentences being spoken twice; the three sets of sentences were chosen to have similar properties in terms of phoneme coverage, length, etc.

The total set of recorded data was divided into three sets for training, development (parameter tuning), and evaluation (test) of the forthcoming EMG-to-Speech mapping. Of each session, we set aside 15 % of the corpus as development set and 15 % as evaluation set. Table 4.1 lists the corpus data sizes for the three subsets broken down by speakers.

Table 4.1 – Details of the recorded EMG-Array-AS-200 data corpus, consisting of normally and silently articulated speech data.

Speaker	Sex	Normal speech, in [mm:ss]			Silent speech, in [mm:ss]		
		Train	Dev	Eval	Train	Dev	Eval
S1-1	m	7:00	1:28	1:25	6:51	1:25	1:27
S1-2	m	7:34	1:34	1:32	7:30	1:34	1:29
S1-3	m	7:25	1:31	1:31	7:31	1:34	1:32
S1-4	m	7:45	1:37	1:33	7:48	1:38	1:34
S2-1	m	7:08	1:30	1:29	7:40	1:34	1:36
S2-2	m	7:04	1:29	1:26	7:54	1:38	1:36
S3-1	m	7:10	1:30	1:24	5:50	1:15	1:11
S3-2	m	7:33	1:41	1:43	6:37	1:32	1:30
S3-3	m	7:19	1:31	1:42	5:33	1:09	1:18
S3-4	m	7:43	1:37	1:35	5:31	1:11	1:09
S4-1	f	6:59	1:27	1:27	6:40	1:23	1:22
S5-1	m	9:27	2:00	2:00	8:53	1:58	1:51
Total		90:07	18:55	18:47	84:18	17:51	17:35

4.2.3 EMG-ArraySingle-A-500+ Corpus

While the EMG-Array-AS-200 corpus is designed to foster the investigation of multiple sessions and the different speaking modes silent and audible speech, the EMG-ArraySingle-A-500+ corpus aims at providing large data from few speakers to study high-performance for session-dependent systems. We therefore focus on few sessions with a large amount of recording data.

Since the usage of wet electrodes gives only a limited recording time (due to quality degradation) and since a concentrated/consistent engagement of a speaker has only a limited amount of time, we restrict the speaking mode to normal audible speech.

The corpus consists of six sessions from three speakers with different amount of data. Four sessions incorporate around 500 phonetically balanced English utterances that are based on [SW10], which consist of the same utterances from the EMG-Array-AS-200 corpus. Two larger sessions exist, which additionally incorporate utterances from the Arctic [KB04] and TIMIT [GLF+93] corpora, giving a total of 1,103 utterances for the smaller and 1,978 utterances for the bigger of these two large sessions.

Like before, the corpus data was divided into three sets for training, development, and evaluation. Of each session, we set aside 10 % for development data. A set of 10 sentences (plus repetitions) were selected for evaluation. Those were identical for all speakers and all sessions and also included in the the EMG-Array-AS-200 corpus for comparison reasons.

Table 4.2 gives detailed information about the durations of the recorded utterances from the EMG-ArraySingle-A-500+ corpus.

Table 4.2 – Details of the recorded EMG-ArraySingle-A-500+ data corpus, consisting of normally articulated speech data with at least 500 utterances.

Session	Sex	Length, in [mm:ss]			# of utterances		
		Train	Dev	Eval	Train	Dev	Eval
Spk1-Single	m	24:23	02:47	01:19	450	50	20
Spk1-Array	m	28:01	03:00	00:47	450	50	10
Spk1-Array-Large	m	68:56	07:41	00:48	984	109	10
Spk2-Single	m	24:12	02:42	00:49	447	49	13
Spk2-Array	m	22:14	02:25	01:10	450	50	20
Spk3-Array-Large	f	110:46	11:53	00:46	1,771	196	10
Total		278:32	30:28	05:39	4,552	504	83

4.3 Whisper Corpus

For investigations on prosodic differences between acoustic data, we additionally recorded a database with normal audible and whispered read speech of 9 persons, entitled *WA-200 Corpus*. Each person recorded 200 sentences per speaking mode, i.e. a total of 400 sentences. The corpus consists of 200 phonetically balanced English sentences originating from the broadcast news domain, with 20 sentences being spoken twice. All sentences were recorded using a standard close-talking head mounted microphone (Andrea Electronics NC-8) and a Samsung Galaxy S2 smartphone microphone in parallel. The phone setting for noise cancellation was applied.

The head mounted microphone was positioned about 3 cm to the right of the test person's mouth. The speakers were instructed to hold the smartphone as though they were making a regular phone call with their left hand, so that no interference with the head mounted microphone existed.

The used software and the general process of recording is the same as in the EMG-based recordings. While the head mounted microphone recordings were automatically segmented by the recording software, the smartphone recordings had to be segmented after the recording. This was realized by emitting different sine beep sounds at the start and end of every utterance. These beeps were then recorded by the smartphone microphone and later used to semi-automatically segment the recordings, by searching for start and end beeps. We additionally manually rechecked every recording session in order to correct minor segmentation errors.

For the data collection, six male and three female speakers were recorded, with age ranging from 21 to 33 years. Each recording session consisted of one en-bloc recording of 200 phonetically balanced English sentences spoken in audible speech and one en-bloc recording of 200 phonetically balanced

Table 4.3 – Details of the recorded WA-200 data corpus, consisting of normally articulated and whispered speech data.

Speaker	Sex	Normal speech, in [mm:ss]			Whisp. speech, in [mm:ss]		
		Train	Dev	Eval	Train	Dev	Eval
101	m	08:53	1:49	1:53	08:33	1:44	1:45
102	m	10:23	2:03	2:08	10:40	2:09	2:16
103	m	07:25	1:28	1:33	08:24	1:41	1:46
104	m	09:14	1:50	1:56	09:56	2:01	2:05
105	m	07:46	1:33	1:38	08:42	1:42	1:51
106	m	08:34	1:42	1:50	09:01	1:46	1:56
107	f	08:52	1:48	1:52	08:50	1:47	1:52
108	f	08:17	1:39	1:42	09:17	1:50	1:55
109	f	10:37	2:09	2:14	10:13	2:04	2:09
Total		80:01	16:01	14:46	83:36	16:44	17:35

English sentences spoken in whispered speech. The order of both recording blocks was switched with different speakers to eliminate any possibilities of ordering effects. In order to assure a consistent pronunciation of words, all recording sessions were supervised by a recording assistant.

Finally, the total set of recorded data was divided into three sets for training, development, and evaluation. For both speaking modes, we set aside 15 % of the corpus as development set and 15 % as evaluation set, with identical utterances in both, the audible and the whispered part. In order to allow for speaker-dependent as well as speaker-independent experiments, the three sets are shared across speakers, i.e. each speaker appears in training, development and evaluation set. Table 4.3 lists the corpus data sizes for the three subsets broken down by speakers.

Chapter 5

Challenges and Peculiarities of Whispered and EMG-based Speech

The relation between articulatory muscle contraction and acoustic speech has been investigated since several decades [TM69, GHSS72, KP77, Hir71] and even silent articulation of speech was already investigated in 1957 [FAE58]. It is an important step to have a closer look at the individual peculiarities of the EMG signal and to quantify the relationship between acoustic and EMG signals.

5.1 Relations between speech and EMG

For a first signal analysis we used data from the simplified spoken utterances, described in Section 4.2.1. Investigating complex sentences from the broadcast news domain, makes a signal analysis more complicated than using simpler sentences consisting of basic constant and vowel phrases. Additionally, when using complex sentences, the obtainment of phone labels introduces a source of errors and thus is less reliable.

For an improved inspection of the EMG data, we filter out unwanted noise data. Therefore, we use a *spectral subtraction* technique [Bol79] - an approach that is usually used to suppress acoustic noise data in speech recordings. The first recording part (250 ms) that belongs to a pause before the subject starts to speak is considered to be noise data. This noise data is assumed to be stationary and thus can be subtracted in the frequency domain. Therefore, the power spectrum of the 250 ms noise signal is calculated and subtracted to give the estimated clean signal spectrum. The estimated power spectrum is combined with the original phase information and an application of the inverse Fourier transform gives the final estimated signal.

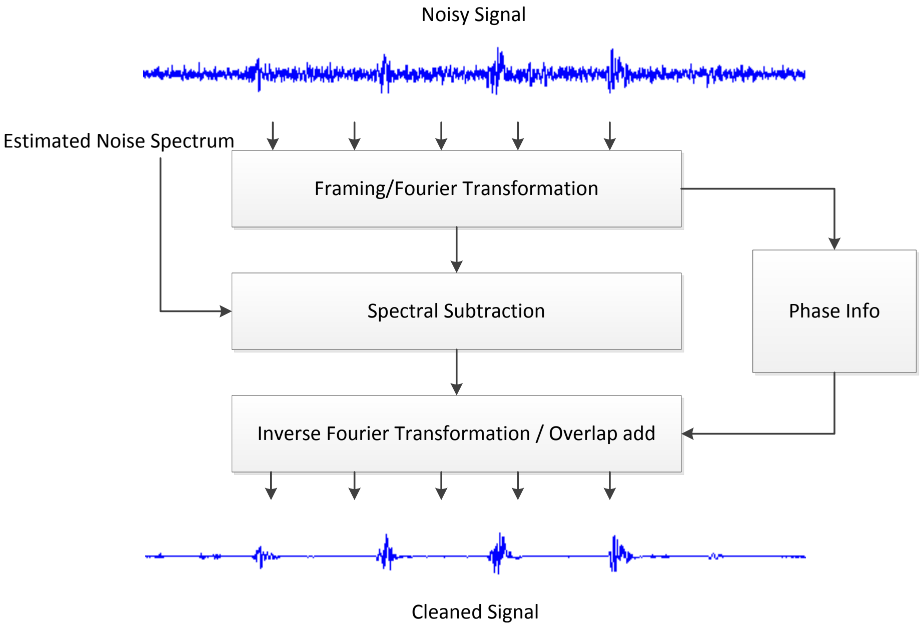


Figure 5.1 – Process flow of spectral subtraction used for EMG signal cleaning.

Figure 5.1 shows the process flow of the spectral subtraction, while Figure 5.2 demonstrates an exemplary application to our EMG data. Note that due to the small inter-electrode distance many channels show redundant signals. For simplicity we show only 6 different channels: channel 2 and 5, which belong to the chin electrode array, plus channels 11, 16, 26 and 32 from the electrode grid on the cheek (see Chapter 4.1.3 for details on the EMG-array setup). It is clearly visible how spectral subtraction helps for the visual signal inspection.

As a first investigation we compared the simultaneously recorded acoustic signal and the EMG signal. Both signals are synchronized using an additional marker signal to ensure a parallel signal. Figure 5.3 gives an example for the utterance “babababa”, showing the spectrogram of the acoustic signal, plus

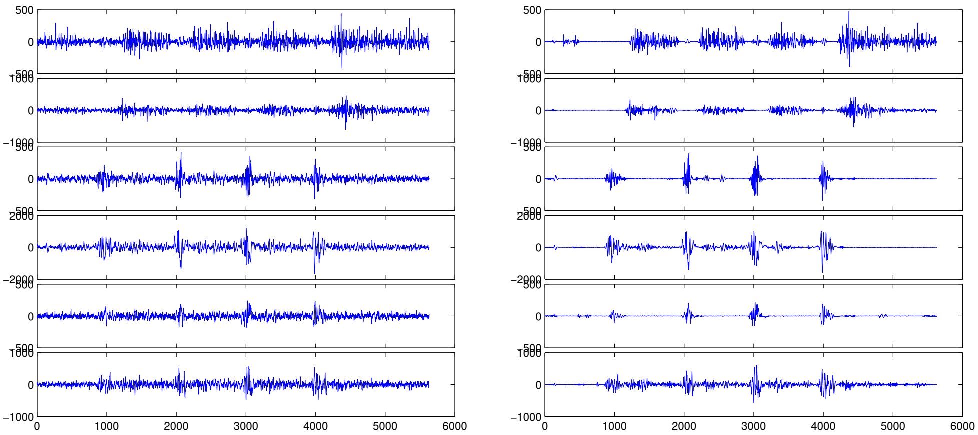


Figure 5.2 – EMG signals from – top to bottom – channels 2, 5 (chin array) + 11, 16, 26 and 32 (cheek electrode grid) of the utterance “babababa”. Raw data before (left) and after using spectral subtraction (right).

one EMG channel from the chin array (channel 2) and one EMG channel from the cheek array (channel 19). A couple of observations can be made:

- Both EMG signals clearly differ from each other, especially regarding onset and offset.
- The opening of the jaw is more represented in the chin array channel 2, which can be seen in the onset/offset similarities of channel 2 and the acoustic channel. During the articulation of “a”, the jaw muscles contract, which is represented by the EMG signal in channel 2.
- The EMG signals precede the acoustic signal, e.g. when considering the beginning of the utterance. While the recording directly starts with an amplitude raise in EMG channel 19, the first b can be acoustically perceived after approximately 200 ms. In general, the delay between detected onset of muscle activity and the concrete realization of force is known as electromechanical delay (EMD) [CK79]. This delay and its effect on the acoustics-to-EMG delay will be discussed in the upcoming Sec. 5.1.1.

5.1.1 Delay between EMG and speech signal

This section aims at the direct investigation of the delay on a signal-basis. In 1956 it was already published [FAB56] that the invasively recorded laryn-

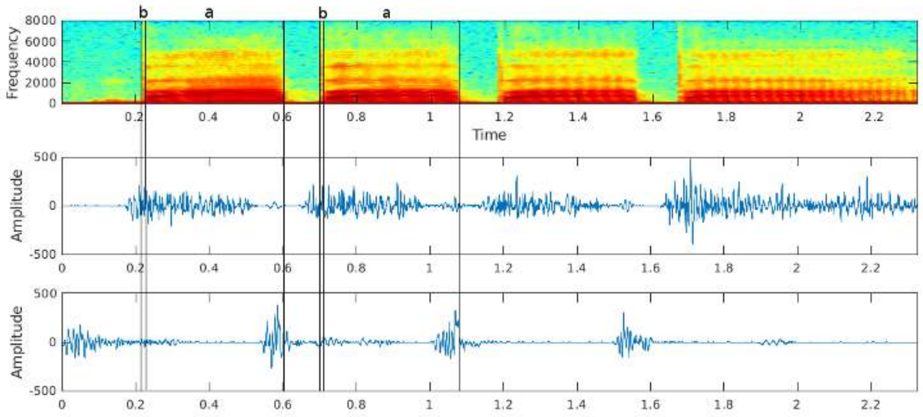


Figure 5.3 – Spectrogram of the audio (top) and the raw EMG signals of cheek-array channel 2 (middle) and chin-array channel 19 (bottom) from the utterance "babababa". The first two occurrences of "b" and "a" are illustrated by vertical lines.

geal EMG signal occurs 0.3 s before the acoustic signal is recorded by the microphone.

Later work [JSW+06] empirically estimated the delay between the articulatory EMG signal and the simultaneously recorded speech signal. They varied the amount of delay and explored which one achieved best classification results on an EMG-based speech recognizer for their further scientific work. The result was an amount of 50 ms.

Scheme et al. [SHP07] also examined this anticipatory EMG effect and stated that "coarticulation effects may even be exaggerated in MES due to anticipatory muscle contractions which occur while prepositioning muscles for articulation." They also varied the offset and obtained the classification accuracy, which gives peak performance on using a delay of 100 ms.

Chan et al. [CEHL02] investigated the delay using a temporal misalignment between 400 ms and 600 ms for the labels, in steps of 25 ms. On a ten-word vocabulary an LDA and a Hidden Markov Model (HMM) classifier were tested on two subjects. Best EMG-based speech recognition results were achieved when a 500 ms delay is used. While the temporal misalignment has only a small effect on the HMM classifier, the LDA results perform significantly worse, when the alignment is differed from the 500 ms delay.

The general effect of the temporal delay between detected onset of muscle activity and the realization of force is known as electromechanical delay (EMD) [CK79]. This is investigated in a more sophisticated analysis, the DIVA model (Directions Into Velocities of Articulators) [GGT06], which gives explanations for the time it takes for an action potential in a motor cortical cell to affect the length of a muscle. The authors divide this amount of time into two components: (1) the delay between motor cortex activation and activation of a muscle as measured by EMG, and (2) the delay between EMG onset and muscle length change. The first delay is measured according to Meyer et al. [MWR+94] to be 12 ms, while the authors measure a 30 ms delay for the second component using electromyographic observations on the tongue. Thus, a delay of 42 ms between movement onset and neuromuscular activation is used.

Having a biologically motivated look at this topic, we can estimate the sound propagation speed with 340 meter per second, and estimate the human vocal tract length with 0.2 m. This would result in a 0.588 ms delay. Hence, the pure acoustic signal propagation time can be considered a minor influence on the EMG-to-acoustic signal delay, even when the begin of innervation is estimated at its pulmonary origin.

Thus, the main impact can be explained by the electromechanical delay (EMD) [CK79]. However, the EMD determination is not trivial [CGL+92]. It is highly muscle-dependent and may be between 30 ms and 100 ms , depending also on the difference of initiation from resting level versus non-resting level [CK79].

Having these delay observations from the literature, we follow [JSW+06] in using a delay of 50 ms between EMG and acoustic signal. This number is in accordance to the described electromechanical delay effect and covers our own investigations for signal analyses.

5.2 Relations between different speaking modes

While we already introduced the definitions of the investigated different speaking modes in Section 2.5.2, we will now investigate the speaking mode properties using the EMG-Array-AS-CV corpus and the whispered speech containing WA-200 corpus. This especially includes a comparison on a EMG signal basis.

5.2.1 Whispered Speech

Comparison to Normal Speech There is a huge amount of research done on acoustic differences between whispered and normal speech (see Section 2.5.2 for details on whispered speech production). According to [ITI05] there is an upward shift of formant frequencies of vowels in whispered speech, compared to normal speech, while voiced consonants have lower energy at frequencies up to 1.5 kHz. Jovicic and Saric [JŠ08] quantify the power differences between whispered and normal speech. While unvoiced consonants have only minor power differences (max. 3.5 dB), the power of voiced consonants is reduced by 25 dB in whispered speech.

Tran et al. [TMB13] created an audiovisual corpus to analyze whispered speech. They observed that most of the visual differences between modes are observed in lip spreading. Other aspects describe that the orofacial area is preserved during whisper speech.

Quantification of Difference To quantify the acoustic discrepancy between whispered and normally spoken speech, we do a first evaluation on the recorded Whisper Corpus *WA-200* introduced in Section 4.3. We compute the Mel Cepstral Distortion (MCD) between MFCCs of original whispered utterances and the corresponding normally spoken utterances, averaged over all utterances of each speaker. The computation of the MCD is performed on a frame-by-frame basis. To align the frames of the normally spoken to the corresponding frames of the whispered utterances, (Dynamic Time Warping) alignment (see Section 2.8) is used. The results are presented in Table 5.1.

The high MCD values are not surprising since the distortion is calculated between normal audible speech and the whispered recordings without applying any mapping strategy. These numbers are meant to serve as baselines for further investigations. The range between speakers is also quite large, varying between 8.17 and 10.69 with the head mounted microphone and ranging from 6.44 to 8.51 with the smartphone, indicating that the style of whispering may be rather individual.

The MCD values of the recordings with the smartphone are significantly lower. Listening to single whispered utterances indicates a better intelligibility of the whispered smartphone recordings. We assume that the noise suppression technology and similar integrated voice amplifications in the smartphone may help for achieving these good results in whispered speech.

Table 5.1 – Mel Cepstral distortions (MCD) between whispered and normal speech of the WA-200 corpus, per speaker.

Speaker	MCD (Headset)	MCD (Smartphone)
101	8.17	6.44
102	9.23	7.12
103	10.55	8.37
104	10.09	8.46
105	10.39	8.22
106	9.53	8.29
107	10.11	7.75
108	10.69	8.51
109	10.06	6.75
MEAN	9.87	7.77

5.2.2 Silent Speech

First myoelectric-based investigations on silent and audible speech had been published in the 1950s [FAB56, FAE58], investigating the delay between EMG and the acoustic speech signal.

[Jan10] showed that no significant difference between the duration of normal and silently mouthed utterances exist, concluding that the missing acoustic feedback does not result in variations of the speaking rate. The same observation holds with our EMG-Array-AS-200 corpus, described in Section 4.2.2. Analyzing Table 5.1, some speakers tend to speak faster in silent mode, others clearly speak slower. Thus, indicating to be a speaker-dependent phenomena, no significant general speaker-independent difference can be outlined.

The same work presented the existence of significant differences in the amplitude and the spectral properties between silent and audible EMG. These differences were mostly prominent for speakers exhibiting constantly low accuracies, when a silent speech recognition system is used. Figure 5.4 gives an example for the spectral curves of a “good” and a “bad” silent speaker. While a “bad” silent speaker shows high discrepancies between audible and silent EMG in the spectral domain on the left figure, a “good” speaker shows very similar spectral curves.

Figure 5.5 shows two channels (channel 2 from the cheek array, channel 19 from the chin electrode-grid) from the audible EMG and silent EMG of the utterance “babababa”. A visual comparison of the raw EMG signals,

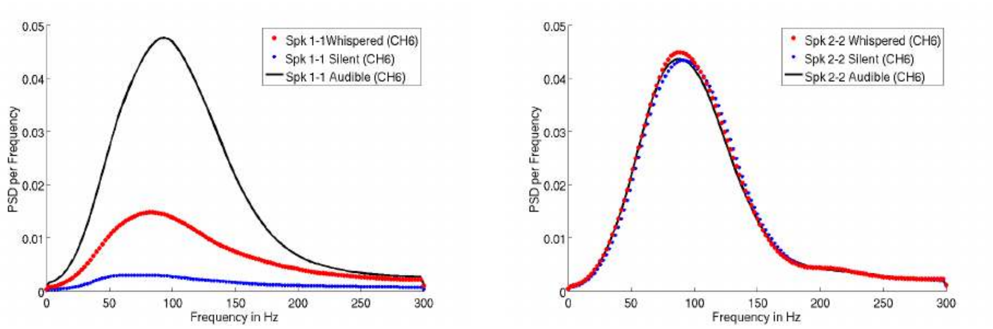


Figure 5.4 – Power spectral densities of whispered/silent/audible EMG from a speaker with constantly high Word Error Rates (left) and from a speaker with low Word Error Rates (right).

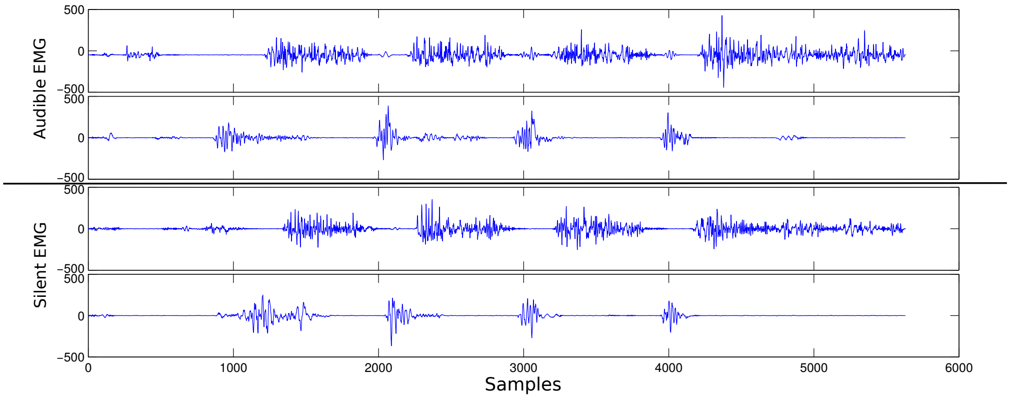


Figure 5.5 – EMG signals of channels 2 and 19 from the normally articulated (top) and the silently mouthed (bottom) utterance "babababa".

reveals no fundamental different properties. The same holds when a power spectral density comparison is investigated for the recordings of the EMG-Array-AS-CV corpus. The similarities between audible and silent EMG can be explained by the fact that the speaker that is recorded in the EMG-Array-AS-CV corpus is very experienced in speaking silently. Following the observations in [Jan10] this justifies the similar signal characteristics. The performance of silent versus audible EMG will be investigated in the following chapters.

Chapter 6

Mapping using Gaussian Mixture Models

This chapter presents a first feature transformation approach to convert EMG-features into acoustic speech based on Gaussian mixture models. This approach was originally introduced for voice conversion [SCM98], i.e. converting acoustic speech of one speaker into acoustic speech of another speaker. This thesis applies this approach to EMG-to-speech mapping. Additionally, we introduce an application where whispered input is transformed into normal audible speech.

Figure 6.1 gives a high level overview of the EMG-to-speech conversion system. We already discussed the types and backgrounds for the input EMG and speech signals and introduced the details of preprocessing steps to build representative features that will be used for the upcoming feature transformation. The first approach for generating MFCCs and F0 from input EMG features is entitled *Gaussian Mapping* (red box in Figure 6.1). Details and evaluations will follow in this chapter.

6.1 Theoretical Background

Gaussian mixtures are based on multi-dimensional Gaussian distributions, i.e. the Gaussian probability density functions, which are defined for d dimensions as follows:

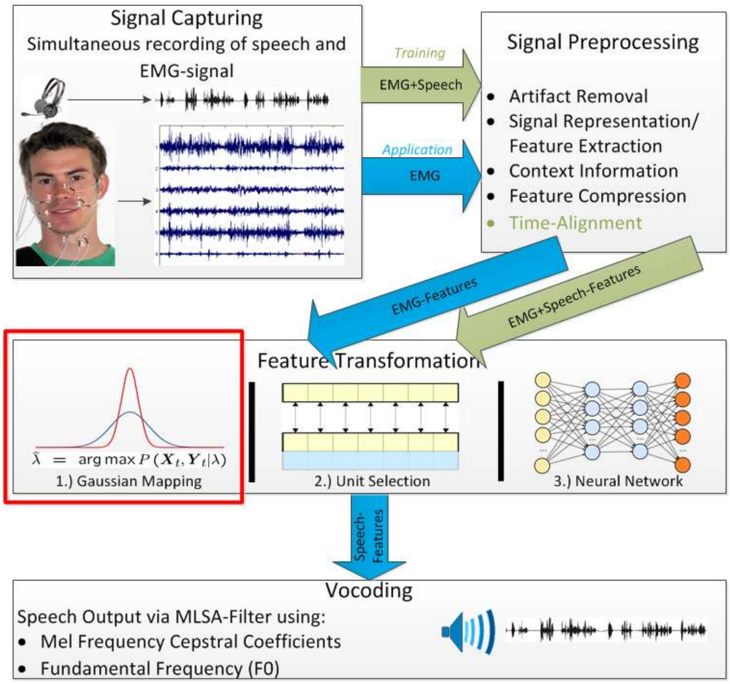


Figure 6.1 – Structure of the proposed EMG-to-speech approach.

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \tag{6.1}$$

A Gaussian distribution is defined by its d-dimensional mean μ and its $d \times d$-dimensional covariance matrix Σ. $|\Sigma|$ is the determinant of the covariance matrix Σ.

To increase the model accuracy a *Gaussian mixture model* λ is used, consisting of a weighted sum of M Gaussians. This defines a probability density function $p(x)$ for a vector x as follows:

$$p(x) = \sum_{m=1}^{M} w_m \mathcal{N}(x;\mu_m,\Sigma_m) \tag{6.2}$$

μ_m and Σ_m are mean and covariance matrix of the m-th Gaussian, and w_m represents the associated mixture weight, which is subject to the following constraints:

$$w_m \geq 0 \quad \text{and} \quad \sum_{m=1}^{M} w_m = 1. \tag{6.3}$$

Gaussian mixtures with a sufficient number of mixtures can approximate any distribution. The parameters of a GMM $\lambda = w, \mu, \Sigma$ are usually learned by applying the Expectation Maximization (EM) algorithm [DLR77].

6.1.1 General Gaussian Mapping Approach

The mapping approach based on Gaussian mixture models (GMMs) was first used by Stylianou et al. [SCM95] and further modified by Kain [KM98] to learn a joint distribution of source and target features. The approach was also later investigated by Toda et al.[TBT04], applying this approach for articulatory-to-acoustic mapping. Its applicability for Silent Speech Interfaces based on Non-Audible Murmur (NAM) has been shown in previous studies [NTSS12]. Thus, we expect that the method is also applicable to EMG-to-speech conversion. The algorithm is frame-based and relatively efficient in terms of computation, so it can be applied continuously in close to real-time [TMB12]. No vocabulary or language constraints exist, making the algorithm flexible for its use with data that was not previously seen in training.

The mapping consists of a training and a conversion phase. The training stage needs parallel data from source and target features to train the GMMs, while the conversion stage only uses source feature data for the conditional probability density function to generate the estimated features.

GMM Training

Source and target feature vectors at frame t are defined as $\boldsymbol{x}_t = [x_t(1), \ldots, x_t(d_x)]^\top$ and $\boldsymbol{y}_t = [y_t(1), \ldots, y_t(d_y)]^\top$, respectively. d_x and d_y denote the dimension of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, respectively.

A GMM is trained to describe the joint probability density of source and target feature vectors as follows:

$$P(\boldsymbol{x}_t, \boldsymbol{y}_t|\lambda) = \sum_{m=1}^{M} w_m \mathcal{N}\left([\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top; \boldsymbol{\mu}_m^{(x,y)}, \boldsymbol{\Sigma}_m^{(x,y)}\right), \tag{6.4}$$

$$\boldsymbol{\mu}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_m^{(x,y} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}, \tag{6.5}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. m denotes the mixture component index, and M denotes the total number of mixture components. The parameter set of the GMM is denoted by λ, which consists of weights w_m, mean vectors $\boldsymbol{\mu}_m^{(x,y)}$ and full covariance matrices $\boldsymbol{\Sigma}_m^{(x,y)}$ for individual mixture components. $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ represent the mean vectors of the mth mixture component for the source and the target features, respectively. $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ represent the covariance matrices and $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ represent the cross-covariance matrices of the mth mixture component for the source and the target features, respectively.

The GMM is trained by first running K-means algorithm (see Chapter 2.7.1) on the training data and then by refining the Gaussian models by application of the Expectation Maximization algorithm [DLR77]. This is based on an iterative method of alternating expectation (E) and maximization (M) steps.

E: The current parameters are used to estimate the likelihood.

M: Based on maximization of the likelihoods the parameters are estimated.

This is either done for a fixed number of iterations or stopped when the difference of likelihoods between two iterations is below a given threshold.

After the training, the GMM describes the joint distribution P of source features x_t and target features y_t.

GMM Conversion based on Mimimum Mean-Square Error

The conversion method is based on work from [SCM98, KM98] and is performed based on the minimization of the mean-square error (mse) of expectation

$$\epsilon_{mse} = \boldsymbol{E}[\|y - \mathcal{F}(x)\|^2], \tag{6.6}$$

where $\boldsymbol{E}[]$ denotes expectation and $\mathcal{F}$ represents the feature mapping function.

The conversion itself is defined by:

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \lambda\right) \boldsymbol{E}_{m,t}^{(y)}, \tag{6.7}$$

where $\hat{\boldsymbol{y}}_t$ is the estimated target feature vector at frame t and

$$\begin{aligned} P\left(m|\boldsymbol{x}_t, \lambda\right) &= \frac{w_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{n=1}^{M} w_n \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)}\right)}, \\ \boldsymbol{E}_{m,t}^{(y)} &= \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\right), \end{aligned}$$

$P\left(m|\boldsymbol{x}_t, \lambda\right)$ is the conditional probability density of the m-th Gaussian component given the feature vector x_t. The given equation can be calculated by applying Bayes theorem.

The remaining unknown elements are $\boldsymbol{\mu}_m^{(y)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$, which can be computed by solving a least squares problem, based on the relations between source and target data. Details can be found e.g. in Stylianou et al. [SCM98].

Note that there exist variations for the implementation of covariance matrices like the usage of diagonal covariances to reduce computational load. Since we are interested in the best possible output and since we perform a mapping from EMG features to acoustic features, we apply full covariance matrices.

6.1.2 Maximum Likelihood Parameter Generation

There exist some modifications to the introduced general Gaussian Mapping approach. One prominent extension is the consideration of time-contextual information on the target acoustic domain. This is realized with an implementation known as *Maximum Likelihood Parameter Generation (MLPG)* [TBT07]. Instead of using only source and target vectors x_t and y_t, additional context-information using $\Delta y_t = \frac{y_{t+1} - y_{t-1}}{2}$ is included to estimate the target parameter trajectory. The general training is only slightly adapted, using $X_t = [x_t^\top, \Delta x_t^\top]^\top$ and $Y_t = [y_t^\top, \Delta y_t^\top]^\top$ and an additional function for EM training, that takes the whole feature sequence into account, not only single frames. The mapping itself is performed using maximum likelihood estimation [TBT07]:

$$\hat{\boldsymbol{y}} = \arg\max_{y} P(Y|X, \lambda)\,, \tag{6.8}$$

where X and Y are source and target feature matrices (consisting of feature vectors in addition to their delta features) and λ represents the set of GMM parameters.

6.2 Mapping on Whispered Data

In a first experiment, we train session-dependent whisper-to-normal conversion systems for all speakers in our Whisper WA-200 data corpus (see Chapter 4.3). We trained a Gaussian mapping for the MFCC features, i.e. a mapping which takes MFCCs of whispered speech as source features and outputs MFCC features of audible speech as target features. In the final vocoding step, the generated MFCC features and the generated F0 contour are converted into an audible sound file using the MLSA filter.

In a first step, we evaluated the systems with the Mel Cepstral Distortion (MCD) measure. To align the frames of the normally spoken to the corresponding frames of the whispered utterances, the DTW (Dynamic Time Warping) alignment algorithm was used on the corresponding MFCCs. Table 6.1 shows the MCDs for this whisper experiment. We calculate the MCD between aligned MFCCs of source whispered utterances and target audible utterances and determine this value the baseline MCD of this experiment, which results in an average MCD of 7.77 of this data corpus.

Table 6.1 – Evaluation results between whispered and normal speech, baseline Mel Cepstral Distortions (MCD) and Gaussian mapping results: MCDs, Unvoiced/Voiced(U/V)-Accuracies and F0 correlation coefficients (r).

Speaker	Baseline	Gaussian Mapping		
	MCD	MCD	U/V Acc.	r
101	6.44	4.79	83.5	0.669
102	7.12	5.16	84.8	0.743
103	8.37	5.03	76.3	0.554
104	8.46	5.21	76.1	0.553
105	8.22	5.25	80.5	0.633
106	8.29	5.71	73.6	0.472
107	7.75	5.35	76.5	0.580
108	8.51	5.66	65.7	0.430
109	6.75	5.02	78.1	0.626
MEAN	7.77	5.24	77.2	0.584

The results in Table 6.1 indicate that large improvements on the MCD criterion can be achieved by Gaussian Mapping on all speakers. A mean MCD improvement of 2.56 can be achieved, which is a relative improvement of 32.9 %. Comparing the average MCD of 5.24 to the literature (e.g. Tran et al. [TBLT10] report 5.99 using whisper-like NAM to speech mapping), this represents decent results, although it should be noted that a direct comparison is not possible due to a multitude of different properties (e.g. speakers, setup, corpus).

6.2.1 Speaker-Independent Evaluation

To investigate the effect of speaker-independence, we train 9 different speaker independent systems with 16 Gaussian Mixture Models (GMMs) in a leaving-one-out cross validation manner which works as follows: For each system, we mark one speaker as test speaker. Training is performed on the training data of the remaining 8 speakers, testing is performed on the development set of the defined test speaker. In this section we report the results of these speaker-independent systems with respect to the impact of adaptation using the Maximum Likelihood Linear Regression (MLLR) approach. The training data of the defined test speaker is used for adapting the mean and covariances of the Gaussian mixture models. The data is additionally normalized, using a simple normalization scheme as follows: the maximum absolute value of the amplitude is determined, followed by normalization using the inverse of this value, such that the amplitude ranges between 0 and 1.

MLLR is an adaptation scheme for Gaussian mixture models which is specifically designed to “modify a large number of parameters with only a small amount of adaptation data” [GW96]. The implementation is based on the details given in [GW96].

MLLR consists of two key parts: In the first step, the Gaussian distributions are clustered by similarity. We perform the clustering by computing a regression tree, where each node consists of a split of a set of nodes. By descending the tree, the set of all Gaussian distributions are split up in increasingly fine-grained partitions. The split for each node is computed by applying a K-Means clustering (see Section 2.7.1 for details). In the second step, the regression tree allows to adapting a set of similar Gaussians simultaneously, thus these Gaussians share their adaptation data, and adaptation is performed even if a Gaussian has not received any training data at all.

The results for the MLLR mean adaptation can be seen in Table 6.2. Normalizing the audio results in a slight MCD improvement. When we apply MLLR to the normalized audio data, we achieve an average MCD reduction of 0.58 compared to the un-normalized, un-adapted baseline system, which is a relative improvement of 11.1 % compared to the session-dependent Whisper-to-Speech mapping and an improvement of 40 % comparing to the baseline.

Table 6.2 – Whispered speech speaker-dependent (SD) versus speaker-independent with adaptation results: Mel Cepstral Distortions (MCD), Unvoiced/Voiced(U/V)-Accuracies and F0 correlation coefficients (r).

Speaker	SD Gaussian Mapping			+ Adaptation		
	MCD	U/V-Acc.	r	MCD	U/V-Acc.	r
101	4.79	83.5	0.669	4.11	83.9	0.687
102	5.16	84.8	0.743	4.43	86.0	0.757
103	5.03	76.3	0.554	4.43	77.7	0.585
104	5.21	76.1	0.553	4.67	79.2	0.620
105	5.25	80.5	0.633	4.59	80.9	0.640
106	5.71	73.6	0.472	5.00	73.2	0.533
107	5.35	76.5	0.580	4.99	76.7	0.632
108	5.66	65.7	0.430	5.12	71.4	0.488
109	5.02	78.1	0.626	4.58	78.0	0.640
MEAN	5.24	77.2	0.584	4.66	78.6	0.620

Increasing the amount of training data, even when using data from different speaker, helps in improving the results.

6.3 Mapping on EMG Data

In the training part we use the simultaneously recorded EMG and audio signals of the same utterance. Since these signals have been synchronously recorded, no alignment is needed for the audible EMG data. However, we delay the EMG signal by 50 ms compared to the speech signal in order to compensate for electromechanical delay. The EMG-TD15 features are reduced to 32 dimensions via an LDA and combined with MFCC features into a single sequence of feature vectors. Note that we use the same frame shift for both EMG and audio signals, so the EMG and audio feature sequences have identical number of frames.

6.3.1 EMG-to-MFCC

Maximum Likelihood Parameter Generation (MLPG) Evaluation

The MLPG-based Gaussian Mapping variation introduced in Section 6.1.2 takes into account temporal dynamics. Thus, Δ-information is added, replacing source frame x_t and target frame y_t with $X_t = [x_t^\top, \Delta x_t^\top]^\top$ and $Y_t = [y_t^\top, \Delta y_t^\top]^\top$.

Figure 6.2 shows the MCDs for *EMG-ArraySingle-A-500+* data sessions and gives a comparison of using MLPG (blue bars) versus the baseline results without MLPG (red bars). There is only a marginal difference between both approaches resulting in an average MCD of 5.95 with MLPG versus 5.91 without MLPG. Only Speaker2-Array gets an improved MCD, while the other five sessions obtain a degradation. Although other reports [TBT07] state promising improvements using contextual information with MLPG, we can not gain a performance boost in terms of MCD. Listening to the resulting output also showed no perceptional difference between both approaches. We assume that the inclusion of stacking adjacent features in the EMG preprocessing part, already covers enough temporal information and thus MLPG can achieve no further improvements. We therefore perform no further investigations in using MLPG and use the previously described frame-based Gaussian mapping approach.

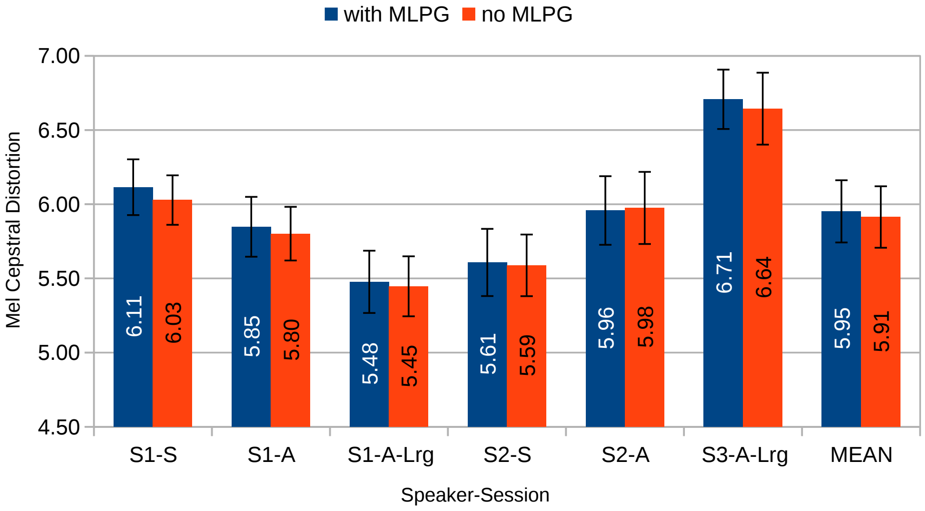


Figure 6.2 – Comparison of Maximum Likelihood Parameter Generation (MLPG) method on the EMG-ArraySingle-A-500+ corpus.

Gaussian mixtures

In the next experiment we evaluate the number of Gaussian mixture components to obtain the best performance in terms of MCD. We vary the number of mixture components between 16 and 128 and train the Gaussian Mapping with the training data. Figure 6.3 shows the Mel Cepstral Distortions using those mixture components. Three out of six sessions get best results with

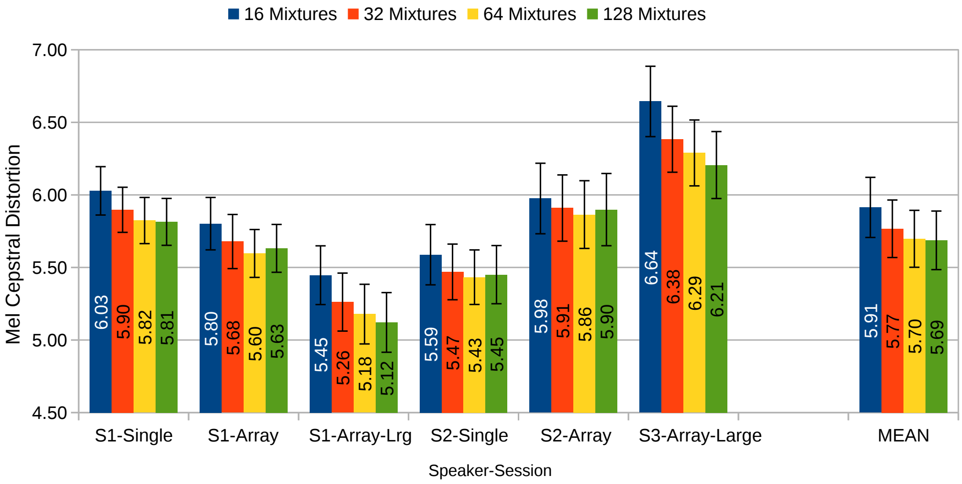


Figure 6.3 – Mel Cepstral Distortions with different numbers of Gaussian mixtures evaluated on the EMG-ArraySingle-A-500+ corpus.

64 Gaussian mixture components, with only small differences to the results with 128 Gaussian mixtures. Since there is only minimal improvement when the number of Gaussians is raised from 64 to 128, we stopped raising the mixtures with 128. Speaker1-Array-Large achieves the best result with an MCD of 5.12, while Speaker3-Array-Large obtains an MCD of 6.21, showing a high MCD variation between speakers. When investigating array-based recordings versus single-electrodes-recordings, no tendency to a further improvement can be seen. While speaker 2 achieves better MCDs with the array setup (5.6 versus 5.81), speaker 2 performs better with the single-electrodes setup (5.43 versus 5.86).

Comparing these results to the related work from the literature, we achieve well performing results. Hueber et al. [HBDC11] report an MCD of 6.2 using about 1,000 training utterances on a similar GMM-based mapping from video/ultrasound data to speech. Toth et al. [TWS09] use a GMM-based EMG-to-speech transformation obtaining an MCD of 6.37 when they use 380 utterances (48 minutes) for training.

Dimensionality Reduction

Canonical Correlation Analysis The training of an LDA for reducing the amount of input dimensions needs phonetic alignments, that are obtained by a speech recognition system on the simultaneously recorded acoustic data. This process is a potential source for errors, since there are no perfect alignments, especially when it comes to myographic data. As a consequence of this severity, we investigate the application of canonical correlation analysis (CCA), a transformation that can use the MFCCs for supervised training instead of discrete phone labels. Similar to the feature reduction with LDA, we compute a subspace projection from the stacked TD15 features that maximizes the correlation between the TD15 features and the MFCCs (instead of labels) of the corresponding speech recording. Additionally, we stack the MFCC feature vectors with their 5 adjacent MFCCs (left and right) to include contextual information and to increase the dimension of the resulting subspace. Since the subspace dimensionality is determined by the dimensionality of the MFCC target data used for the supervised training, including the stacking results in a $11 * 25 = 275$-dimensional subspace. We use only the 32 subspace components that yield the highest correlation. This number is also used for the final LDA feature reduction. Figure 6.4 shows the CCA results and the MCDs obtained with LDA feature reduction, both using Gaussian Mapping with 64 mixtures.

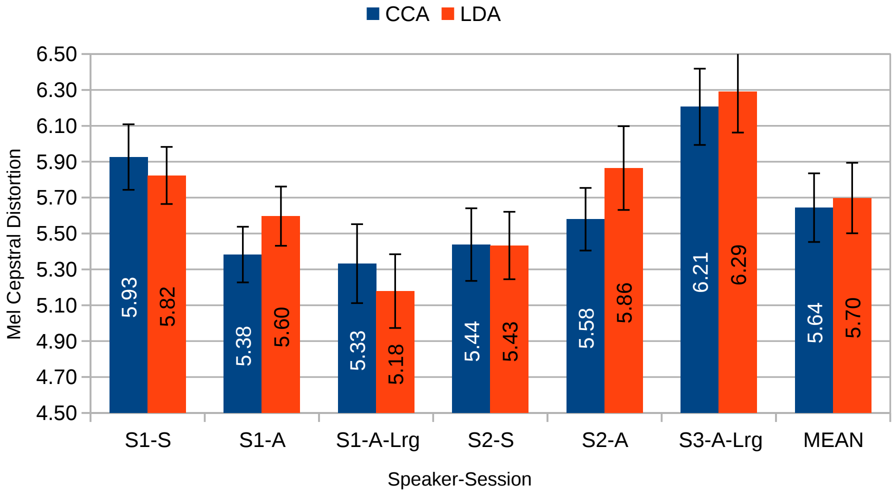


Figure 6.4 – EMG-to-MFCC mapping results using Canonical Correlation Analysis (CCA) versus Linear Discriminant Analysis (LDA) preprocessing.

While the average MCD is improved from 5.70 to 5.64, three out of six sessions perform better with LDA preprocessing. The best session-dependent result from Spk1-Array-Large gains also a significant ($p < 0.05$) MCD loss from 5.18 to 5.33, implying that the use of phone information still helps in producing low MCDs. However, using CCA feature reduction, we not only discard the need for phone alignments, we even slightly improve the mean MCD, making this technique a capable replacement for LDA.

Silent Speech Data

So far all experiments on EMG-to-speech transformation described in this section were done on *audible EMG data*. Mapping on EMG signals that are recorded when the words are spoken non-audibly is a challenging topic. While the main attention of this thesis is dedicated to the feasibility and general improvements on a direct EMG-to-speech approach, we also want to explore real silent speech data.

For a first baseline system we use the Gaussian models that were trained on audible speech and evaluate the mapping on EMG signals from silent test data. Since there is no parallel acoustic data for the silent speech recordings, and thus no acoustic reference data, we can not compute the MCD like before. Therefore, we compare the output data with the audible data from the same session. This data is textually the same, but may of course have differences since it was not simultaneously recorded to the input EMG data. Due to the different signal lengths, we align the synthesized silent data and the reference acoustic data using DTW. This DTW alignment also has effects on the final MCD, since the alignment is improved compared to a simple frame-to-frame MCD comparison. Thus, it would be unfair to compare the MCDs from the converted silent data to the MCDs we obtained in the previous EMG-to-MFCC experiments. We therefore use a slightly modified *DTW-MCD* for the audible EMG data, which precedes a DTW alignment to the input before MCD computation.

Although converted audible EMG data (namely the resulting MFCCs) and the simultaneously recorded acoustic data have the same duration, a DTW with MCD cost function gives better results than a simple frame-to-frame MCD.

Figure 6.6 shows the Gaussian Mapping DTW-MCDs resulting from three different mappings on audible and silent EMG data with 16 mixtures:

1. The baseline results using audible EMG for training and testing (Aud Train - Aud Test),
2. Using audible EMG data for training, and silent EMG data for testing (Aud Train - Sil Test),
3. Using silent EMG data for training and testing (Sil Train - Sil Test).

The latter is not trivial, since there is no synchronously recorded acoustic data that can be taken as target data. Since the EMG-Array-AS-200 corpus contains audible and silent data per session, we can use the acoustic data and align this data to the silent EMG data using DTW. Since a DTW distance metric between acoustic MFCC features and EMG features seems not reasonable, we use the DTW path that is computed between silent and audible EMG features.

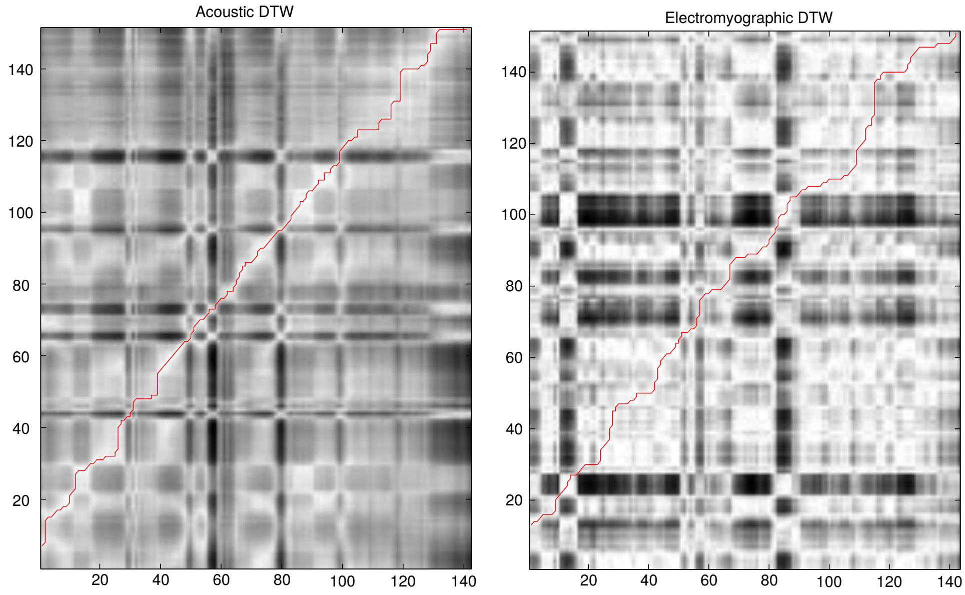


Figure 6.5 – Feature distances and the DTW paths (red line) calculated between acoustic MFCC features stemming from two utterances of the same sentence (left) and the DTW for the corresponding electromyographic power feature (right).

To investigate the DTW alignment and to obtain a correct DTW path, we start by looking at the audible speaking mode. We use two different utterances stemming from the same sentence. This has the advantage that we can use the DTW alignments that are created on the acoustic features, and additionally can compare the DTW path that is created between the EMG features. The "acoustic DTW" is often used in speech processing, and thus we can use it as a valid target for the novel "electromyographic DTW" to achieve well-performing features/parameters/channel combinations. We reduce the 35 EMG channels to selected 5 different channels to decrease redundant information and to exclude artifact-prone channels. Two of these resulting channels are obtained from the inside area of the chin array, while the remaining three channels are taken from the middle of the electrode-grid. We intentionally used no edge channels of the arrays, since they are prone to artifacts. We then use a simple power feature that is also included in our TD feature set (see Section 2.6.3 for feature details). Figure 6.5 shows feature distances and the DTW paths of two acoustic features stemming from two utterances of the same sentence (left) and the DTW path of the corresponding electromyographic features (right). Although the two outlined distance matrices stem from completely different feature domains (acoustic MFCCs versus EMG power), a similar DTW path can be obtained, which encourages a useful electromyographic alignment that can be also applied for aligning silent EMG and audible EMG. We thus use this electromyographic DTW technique to align the source silent EMG data to the target acoustic data, which is necessary for the training in the "Sil Train - Sil Test" experiment.

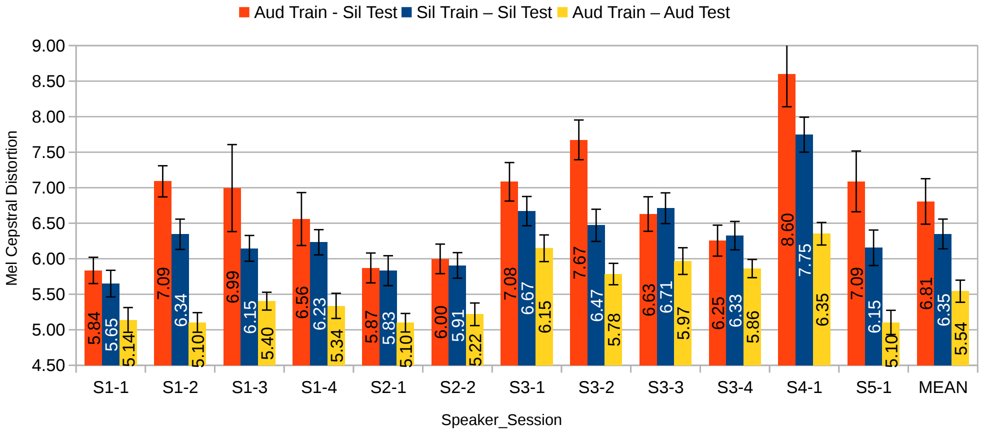


Figure 6.6 – Gaussian Mapping on Silent Speech Data from the EMG-Array-AS-200 corpus. Results are computed using audible EMG train and silent EMG test data (red bar), silent EMG train and test data (blue bar) and audible EMG train and test data (yellow bar). Final Mel Cepstral Distortions are aligned using DTW.

The obtained results from the three different mappings are depicted in Figure 6.6. Looking at the results, it can easily be seen that using silent EMG data on models trained on audible EMG obtains a high MCD degradation from an average MCD of 5.54 to 6.81 compared to the Aud Train - Aud Test scenario. The high discrepancy between silent and audible MCD results can also be observed within a single speaker. While previous work [JWS10] differentiated between "good" and "bad" silent speakers, the Gaussian mapping outputs indicate a more complicated - probably session-dependent - result. Regarding the MCDs from speaker 1, the first session differs in 0.7, while the remaining three sessions differ in at least 1.2. Using silent EMG on models that are also trained on silent EMG, considerably reduces the average MCD from 6.81 to 6.35. However, it should be noted that since the audible reference data was not simultaneously recorded, the obtained MCD results give only a limited validity.

6.3.2 EMG-to-F0

In previous work from Toth et al. [TWS09] the fundamental frequency (F0) of simultaneously recorded audible speech was used for the final synthesis of speech waveforms. For the final application of EMG-to-speech mapping, not only spectral information but also F0 information is essential to be estimated directly from EMG data.

We thus use the introduced Gaussian Mapping approach and replace the target MFCCs with F0 features.

The data for the GMM training consists of EMG feature vectors as the source data and F0 vectors as the target data.

In the conversion part, segmental feature vectors of the source EMG data are constructed in the same manner as in the training part.

Figure 6.7 and 6.8 show the experimental results for different numbers of Gaussian mixtures on the EMG-ArraySingle-A-500+ corpus. We obtain an average correlation coefficient of 0.61 with the best session result of 0.65 on Spk1-Array. The mean voiced/unvoiced accuracy achieves 78.5 %, where Spk1-Array-Large retains an accuracy of 82 %. The results indicate that it is possible to estimate F_0 contours from EMG data. Comparing this to the literature, where e.g. [TBLT08] generate F0 from whispered speech input with 93.2 % voiced/unvoiced accuracy and a correlation coefficient of 0.499, our mapping gets worse voicedness results, but a better correlation coefficient. [ARH14] presented EMG-based pitch generation on neck muscles solely and

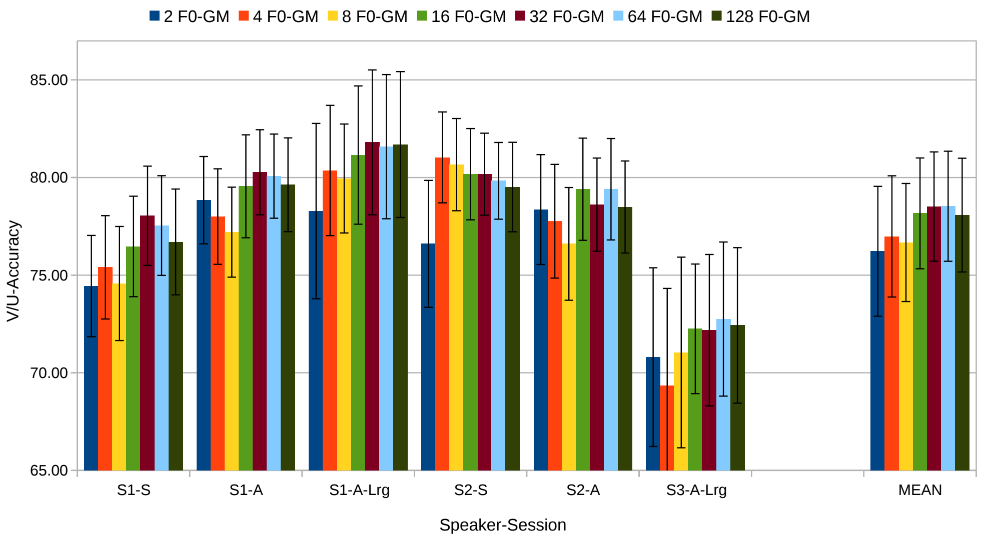


Figure 6.7 – Voiced/Unvoiced accuracies for EMG-to-F0 Gaussian Mapping on the EMG-ArraySingle-A-500+ corpus.

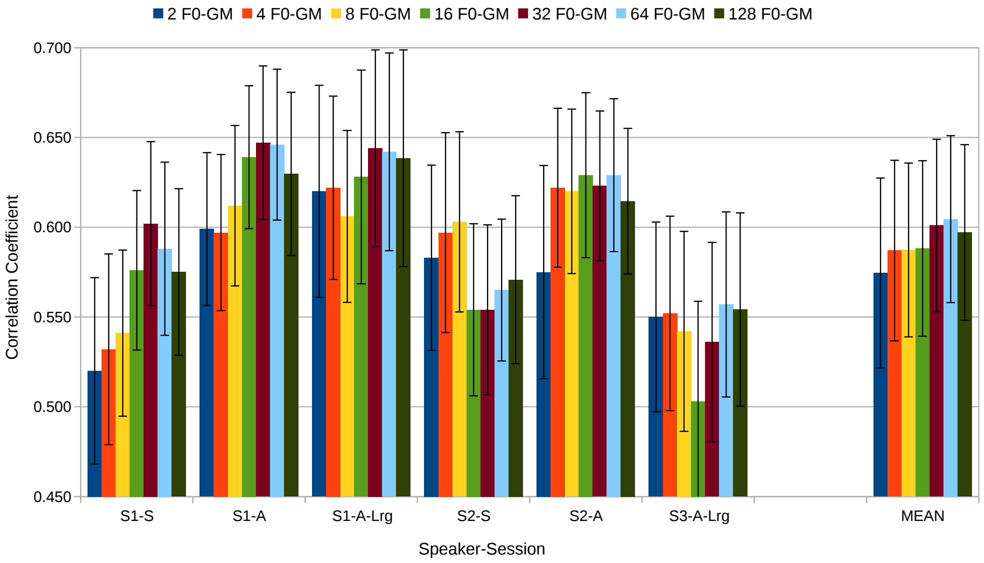


Figure 6.8 – Correlation Coefficients for EMG-to-F0 Gaussian Mapping on the EMG-ArraySingle-A-500+ corpus.

reported a correlation coefficient of 0.87 and a voiced/unvoiced accuracy of 87 %.

While the objective evaluation is a valid technique to investigate the performance of incremental changes in the system, it remains unclear how a final user judges the EMG-to-F0 conversion. Maybe an omission of F0 can also produce natural output? To investigate this hypothesis we conduct the following experiment: We use the MFCCs from the reference audio file, and create three different "artificial" excitations:

1. the converted **EMG-to-F0** output,

2. white Gaussian noise, resulting in unvoiced speech that sounds whisper-like, entitled **0 F0**

3. a constantly **flat F0** contour, resulting in a robotic-like speech output.

We also add the unmodified reference speech recording (reference), plus the re-synthesized reference recording (resynth). The latter one is obtained by using extracted speech features (MFCCs + F0) from the target audio with the MLSA filter to produce a "re-synthesized" reference. This reference contains the quality reductions from feature processing and vocoding steps and thus represents the best output we can achieve with our synthesis setup. A listening test is performed to investigate the user's acceptance on naturalness, where the listening test participant is confronted with the question "How natural does the presented speech recording sound?, please rate between very unnatural and very natural." The formulation of this question follows the naturalness test from the Blizzard challenge [PVE$^+$13]. A continuous slider, which is internally scaled on the interval between 0 and 100, is given and the five output variations EMG-to-F0, 0 F0, flat F0, reference and resynth are presented in randomized order. This setup combines the direct comparison of the five variations and the quantification to a final naturalness score. We randomly selected three different utterances from four different speakers, resulting in twelve utterances, i.e. each of the 12 utterances is synthesized in 5 variations. Thus, a total of 60 utterances are played to the participant. In total 20 participants listened to 60 utterances each and evaluated the naturalness.

Figure 6.9 and 6.10 show the listening test results.While Figure 6.9 gives the mean naturalness scores for each participant, Figure 6.10 shows the direct comparison of the three F0 methods, discarding reference and resynth. Thus, Figure 6.10 shows the number of times the respecting method was preferred.

Some notable observations can be made from the listeners judgments:

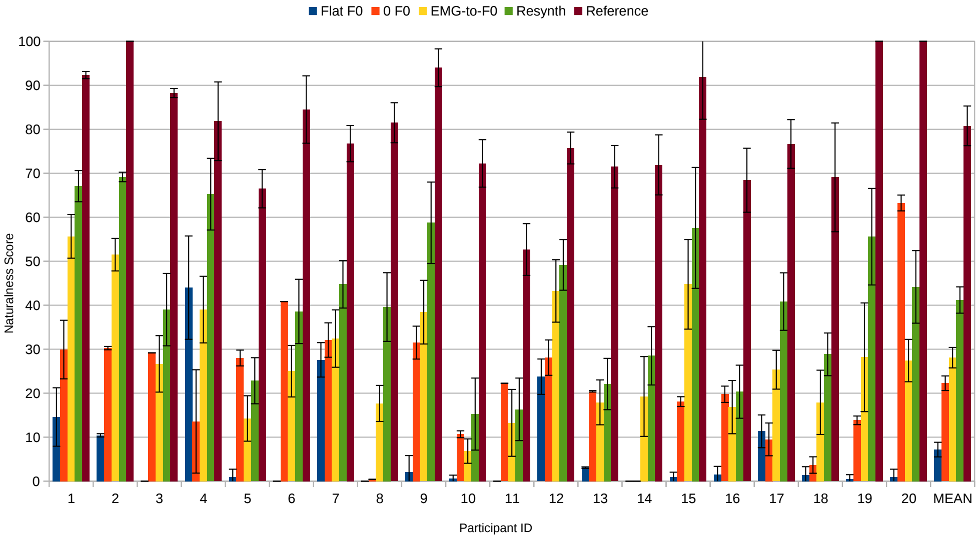


Figure 6.9 – Mean results from the naturalness listening test. 20 participants rated the naturalness from 0 (very unnatural) to 100 (very natural). The error bars represent 95 % confidence intervals.

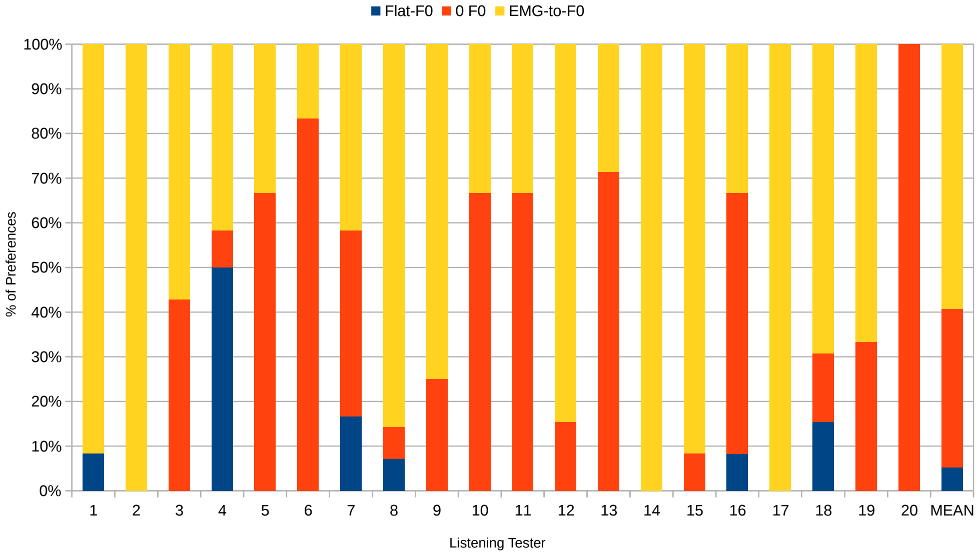


Figure 6.10 – Results from the naturalness listening test. Preference comparison of the three F0-generating methods: Robotic-like flat F0, whisper-like 0 F0 and EMG-to-F0 mapping.

- The highest difference on naturalness is found to be between reference (naturalness score 81 from 100) and resynthesized (score 41 from 100) output.

- The EMG-to-F0 generation significantly ($p < 0.01$) gives the most natural output (naturalness score 28) among the assessed generation methods. **0 F0** obtained the second best naturalness score with 22. Considering a direct comparison (Figure 6.10) averaged over all listened utterances, 59 % preference was given towards the **EMG-to-F0** mapping versus 35 % preference on the **0 F0** approach.
- Comparing the simple F0 generation techniques **0 F0** versus **flat F0**, the unvoiced **0 F0** significantly ($p < 0.001$) outperforms the **flat F0** results (naturalness score 22 versus 7), implying that the whisper-like output is regarded more natural than the robotic-like F0 generation.
- 7 out of 20 participants prefer the whisper-like **0 F0** output over **EMG-to-F0** output. Four of them even regarded it better than the target resynthesized approach.
- 10 out of 20 participants perceived the different approaches in the same order: Reference and resynth clearly give best naturalness, followed by EMG-to-F0, 0 F0 and flat f0 in descending order. Nonetheless, the variances across participants are quite high.

These results show the significant preference of the EMG-to-F0 mapping over simple artificial F0 generation techniques.

Chapter 7

Mapping using Unit Selection Approach

This chapter presents a second approach (see red box in Figure 7.1) for our proposed conversion system that is based on Unit Selection. This corpus-based concatenation technique was introduced in the 1980s [Sag88] and has since become a popular approach for speech synthesis [HB96] known to produce output with a high naturalness.

7.1 Theoretical Background

Unit Selection can be regarded as a corpus-based synthesis approach with the key idea that a speech corpus is used as acoustic inventory, which is utilized to select a best matching sequence of segments during the final Unit Selection application. Appropriate subword units are selected from this inventory - defined as *codebook* - and concatenated to produce a final natural-sounding speech output.

We first build a database consisting of source (e.g. EMG feature) and target (e.g. acoustic feature) segments. We extract segments of a fixed frame number and refer to this amount of frames as *unit width* w_u. To create a larger amount of segments in the database, we shift the segments by one frame at a time, rather than shifting by the whole unit. Together, the (*source*) feature segment and the associated speech segment (*target*) build one *unit*.

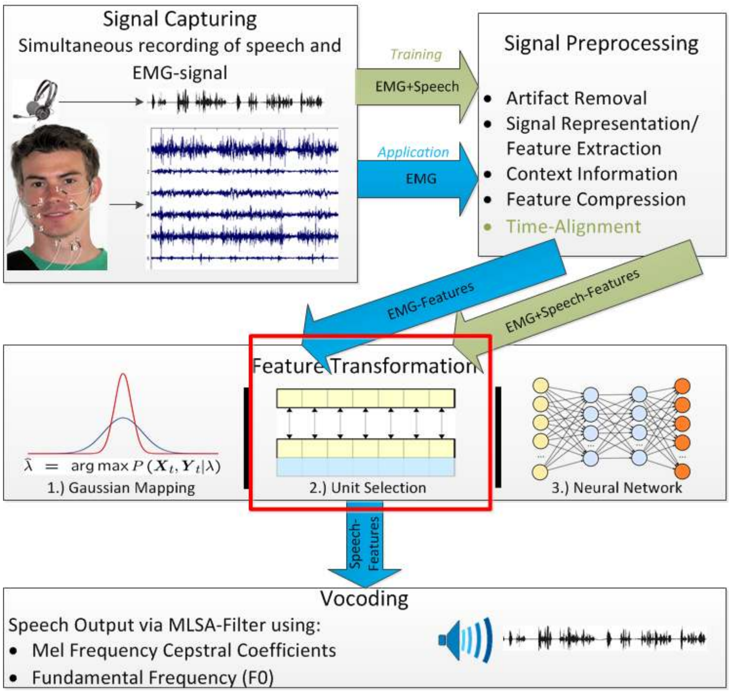


Figure 7.1 – Structure of the proposed EMG-to-speech approach.

In general, the target of the proposed Unit Selection approach is to generate a natural-sounding waveform that resembles the original utterance. The basic entity of processing is called a “unit”. In the EMG-to-speech system a unit consists of simultaneously recorded EMG and audio segments, as illustrated in Figure 7.2.

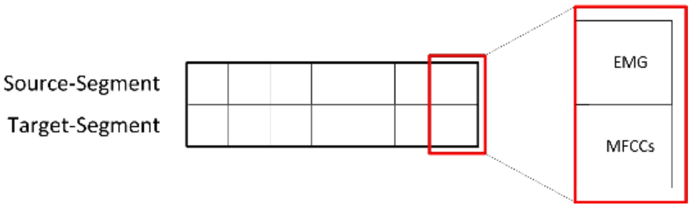


Figure 7.2 – A basic Unit, consisting of source EMG and target MFCC segments used for the proposed Unit Selection approach.

When the codebook is set up, the final application of the conversion can start. During this process the source feature test sequence is also split up into overlapping segments of the same width, as shown in Fig. 7.3. We also vary the shift between two consecutive segments, which we define as *unit shift* s_u.

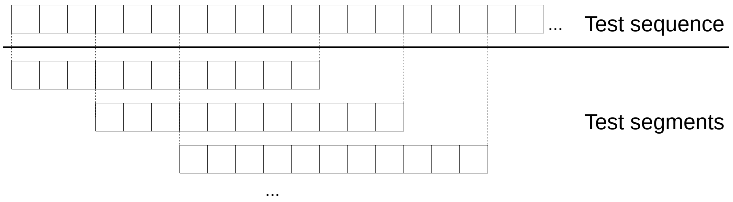


Figure 7.3 – Splitting of the test sequence, here with unit width $w_u = 11$ and unit shift $s_u = 3$.

The desired units are then chosen from the database by means of two cost functions, *target cost* c_t between source EMG segments and the *concatenation cost* c_c between target acoustic segments[HB96]. The target cost measures the similarity between an input test segment and a candidate database unit. The concatenation cost gives the distance between two adjacent units to ensure that the candidate unit matches the previously selected unit. Figure 7.4 illustrates how the cost functions for selecting one unit are evaluated.

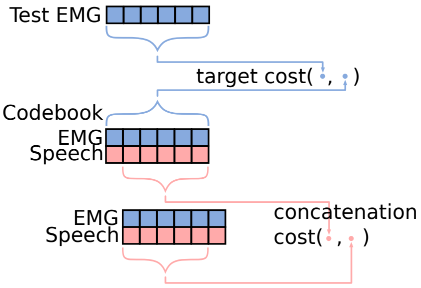


Figure 7.4 – Illustration of combination of target (top) and concatenation (bottom) cost during Unit Selection.

In a simple metric, the target cost can be calculated as the mean Euclidean distance between the respective source EMG segments $s_{test}^{(t)}$ and $s_{db}^{(t)}$:

$$c_t = \frac{1}{w_u} \sum_{k=1}^{w_u} \sqrt{\sum_{d=1}^{D_E} (s_{test}^{(t)}(k,d) - s_{db}^{(t)}(k,d))^2}, \tag{7.1}$$

where D_E denotes the dimensionality of the source EMG features and $s^{(t)}(k,d)$ denotes the d-th dimension of the k-th frame of the segment at time index t. For EMG-to-speech mapping our choice for the target cost is the cosine-distance instead of the Euclidean distance, since we achieved better performances in preliminary experiments. The cosine-distance is defined between two d-dimensional feature vectors $f^{(1)}$ and $f^{(2)}$ as:

$$df_{cos} = 1 - \frac{\sum_{i=1}^{d} f_i^{(1)} \cdot f_i^{(2)}}{\sqrt{\sum_{i=1}^{d} f_i^{(1)2}} \cdot \sqrt{\sum_{i=1}^{d} f_i^{(2)2}}}, \tag{7.2}$$

resulting in values in in the interval $[0, 2]$ (lower is better).

The concatenation cost is based on the mel cepstral distance at the point of concatenation to ensure a proper smoothness of the acoustic output. It can also be interpreted as the scaled mean Euclidean distance between the overlapping frames of the target audio segments of two database units $t_{db}^{(t)}$ and $t_{db}^{(t+1)}$:

$$c_c = \frac{1}{o_u} \sum_{k=1}^{o_u} \sqrt{\sum_{d=1}^{D_A} (t_{db}^{(t+1)}(k, d) - t_{db}^{(t)}(k + s_u, d))^2}, \tag{7.3}$$

where D_A denotes the dimensionality of the target audio features and $o_u = w_u - s_u$ is the number of overlapping frames of two units.

If two units have a natural transition, meaning they originated from the same utterance with starting points i and $i + s_u$, they are favored because the overlapping frames are the same, resulting in a concatenation cost of 0.

Additionally, a weight for the two cost functions is needed to balance naturalness and distinctiveness. If the relative weight of the concatenation cost is too high it results in fluent speech that may differ heavily from the target phone sequence; if it is too low, the synthesized speech sounds chopped and unnatural. Since the ranges are specific for certain features, the optimal weights depend on the feature used for Unit Selection.

The search for the optimal unit sequence is illustrated in Figure 7.5 and can be implemented as a Viterbi search through a fully connected network [Vit67]. The goal is to minimize the overall cost of the chosen sequence, consisting of the combination from target cost and concatenation cost. As an alternative search approach, we use a greedy selection of the lowest costs for each test unit. We implemented both search variations, but decided to evaluate the EMG-to-speech mapping on the greedy approach. This technique may give suboptimal results compared to a full Viterbi search, however we can still achieve decent results more quickly.

After determining the optimal unit sequence, the overlapping audio segments are smoothed using a weight function as proposed by [WVK$^+$13]. We define

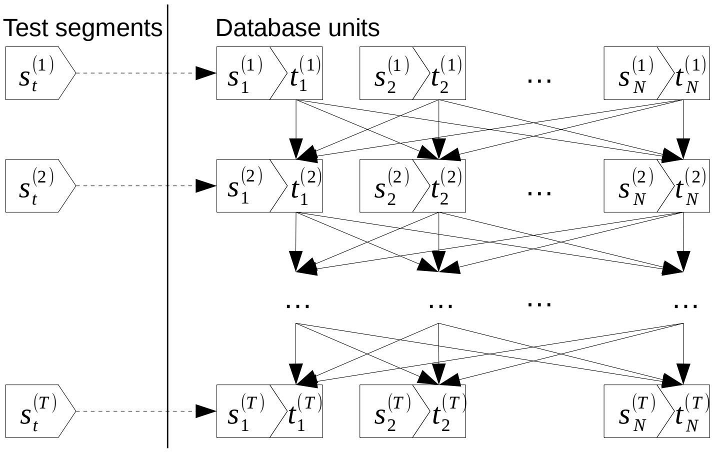


Figure 7.5 – Illustration of the search for the optimal unit sequence. $s_t^{(t)}$ are test source segments, $s_n^{(t)}$ and $t_n^{(t)}$ are database source and target segments, respectively. Dashed lines represent target cost (sketched exemplarily only for one database unit at each time), solid lines represent concatenation cost.

n as the number of units which share a frame, illustrated by the hatched frames of the chosen segments in Fig. 7.6. Since this number of units n varies depending on the unit shift, the weight w for each unit's affected frame is calculated as follows:

$$w[i] = \frac{exp(-0.2 \cdot a[i])}{\hat{w}}, \ i = 1 \dots n, \tag{7.4}$$

with

$$a[i] = \begin{cases} [\frac{n}{2}, \frac{n}{2} - 1, \dots, 1, 1, \dots, \frac{n}{2} - 1, \frac{n}{2}], & n \text{ even} \\ [\lceil \frac{n}{2} \rceil, \lceil \frac{n}{2} \rceil - 1, \dots, 1, \dots, \lceil \frac{n}{2} \rceil - 1, \lceil \frac{n}{2} \rceil], & n \text{ odd}, \end{cases} \tag{7.5}$$

where $\hat{w} = \sum_{i=1}^{n} exp(-0.2 \cdot a[i])$ is used for normalizing the weights to sum to 1. Figure 7.6 shows an example of the smoothing process: The hatched frames of the chosen segments are weighted and added up to create one output frame. This process is repeated at each frame.

Finally, after obtaining the final acoustic feature sequence the speech output can be generated using the MLSA filter on the resulting MFCC and F0 frames.

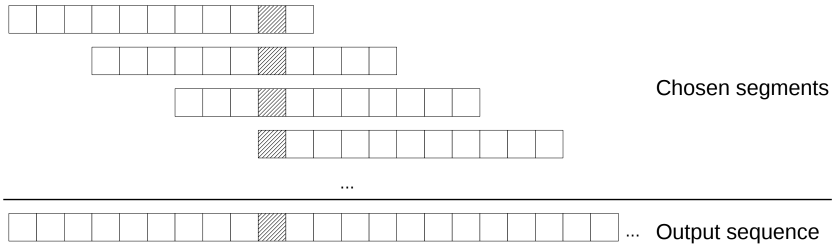


Figure 7.6 – Creating the output sequence from the chosen audio segments.

7.2 Cluster-based Unit Selection

Since a well-performing Unit Selection system relies on sufficient coverage of speech data, there are two obvious improvements on the codebook:

- Adding more units to the codebook.
- Improving the general quality of the codebook.

While the first item adds more confusability and an increase in search time, the second possibility is to be preferred. Another problem by just increasing the amount of codebook data lies in the inter-personal and even inter-session differences of the source feature signal - especially in the EMG signal - caused by variation in electrode positioning and skin properties.

7.2.1 K-Means Codebook Clustering

For the optimization of the unit codebook, we propose a clustering approach that uses the k-means algorithm (see Chapter 2.7.1) to achieve prototypical unit representations and to additionally reduce the general size of the unit codebook.

Initialization

The initialization step can be performed in several different ways. Our implementation chooses the first K input vectors $X_{1..K}$ as initial centroids $M_{1..K}$.

Assignment

In the assignment step, a cluster assignment is computed for each input vector by finding the centroid that the current input vector has the minimal distance to.

Centroid Re-Computation

After all input vectors are assigned, these new assignments are used to recompute the centroid for each cluster as the arithmetic mean of all the vectors that are assigned to it.

This process is iterated until a termination condition is met. One possible way is running the algorithm until only some fraction of cluster assignments change. In this work, clustering is stopped when the assignment of less than 0.1 % of vectors are changed during an iteration. To speed-up the experiments, a parallelized GPU implementation of the K-Means algorithm, employing the NVidia CUDA toolkit, is used [Giu]. The output of the algorithm is the cluster assignments for all vectors.

Like mentioned before, adding more units to the codebook may imply a wider acoustic diversity and thus may improve the synthesis output. The problem of finding suitable units is still problematic, and even hampered since adding redundant or even unnecessary units will not help in improving the system. Additionally, including too much data in the codebook definitely slows down the unit search and the synthesis.

The goal of the proposed *unit clustering* is, to not only reduce the size of the codebook, but also to improve the representability of each single unit. In this way, we hope to be able to use less units to generate a better output.

Two typical reasons, that are responsible for suboptimal Unit Selection results, are alleviated by a codebook clustering approach:

- the non-availability of a proper unit in the codebook,
- the selection of an improper unit, although a correct one is available.

Clustering the available codebook units can have an impact on these problems, since the initial basic units are changed into a more prototypical version of the acoustic and electromyographic segments they represent.

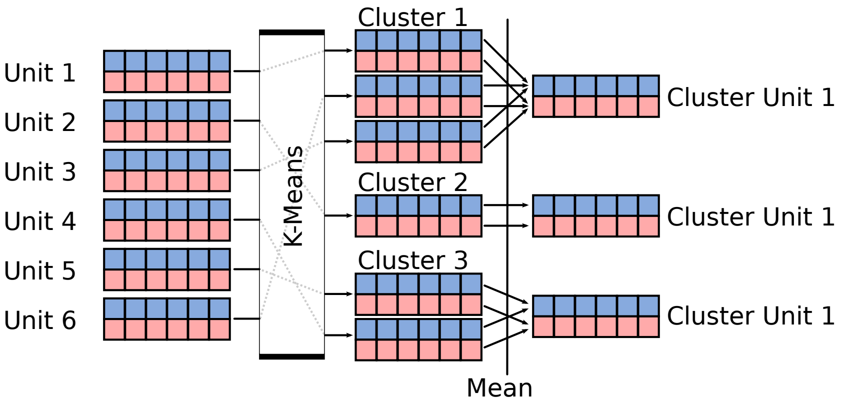


Figure 7.7 – Creation of cluster units from basic units. The units are first assigned to a cluster by performing K-Means clustering, then new *cluster units* are created.

Cluster Unit Creation

The creation of cluster units is depicted in Figure 7.7. At first, the k-means algorithm is applied to build a combination of each unit from the set of basic units. The cluster assignments resulting from this step are then used to build the final cluster units. A cluster of multiple units is turned into one single cluster unit by computing the arithmetic mean of all source and target features belonging to the units that are assigned to that cluster. The resulting cluster units are used for a new codebook and can be utilized to perform unit selection based on the target cost and concatenation cost as described above.

7.3 Oracle: Mapping using Identical Source and Target Speech Data

To get the “golden line” with the best possible result on our recorded data, we conduct a preliminary oracle experiment without the previously described clustering approach. The goal is to evaluate the general feasibility of the proposed Unit Selection approach and to investigate whether the amount of training data is sufficient.

Source and target segments consist of identical acoustic features, which are taken as source **and** target features. The source segment of each unit consists of the audible MFCCs and the target segment contains the MFCCs plus the

F_0 feature. Since only audible data is used in this approach, we refer to this experiment as *Aud-to-Aud* mapping.

The Unit Selection system is trained as described in Section 7.1 and tested on unseen data that is not included in the codebook to investigate if enough units are available and to evaluate the best possible result given the selected parameters. Table 7.1 shows the results, in terms of MCD on the MFCC data and additionally the the F0 Unvoiced/Voiced-accuracies and the F0 correlation coefficients.

Table 7.1 – Aud-to-Aud oracle experiment (using acoustic data for codebook creation) results on the EMG-ArraySingle-A-500+ corpus: MCDs, Unvoiced/Voiced(U/V)-Accuracies and F0 correlation coefficients (r).

Speaker-Session	MCD	U/V-Acc.	r
Spk1-Single	3.62	88.8	0.779
Spk1-Array	3.59	91.4	0.835
Spk1-Array-Large	3.09	86.5	0.721
Spk2-Single	3.50	86.1	0.698
Spk2-Array	3.84	91.0	0.822
Spk3-Array-Large	3.47	83.3	0.682
MEAN	3.52	87.9	0.756

Best MCD results are obtained with the “large” sessions, i.e. the sessions that contain the largest amount of training data. Although Spk3-Array-Large contains the most speech data, Spk1-Array-Large gives a lower MCD suggesting that the latter speaker/session speaks more consistently. The observation that a high amount of codebook data not necessarily gives the best results, additionally holds with the F0 evaluation. Spk1-Array and Spk2-Array perform best on U/V-accuracy and F0 correlation coefficients, although Spk1-Array-Large and Spk-3-Array-Large use a considerably higher amount of training data.

7.3.1 Unit Evaluation

Figure 7.8 depicts the unit size and unit shift evaluation of one exemplary speaker. Reducing the unit shift s_u, results in a reduction of MCD. This is (at least partly) based on the effect that a lower unit shift results in an increased overlap, which means that a higher amount of codebook units are used for the final mapping feature sequence. Thus, bad units can be compensated by a higher amount of other units. The unit width w_u has a minor influence,

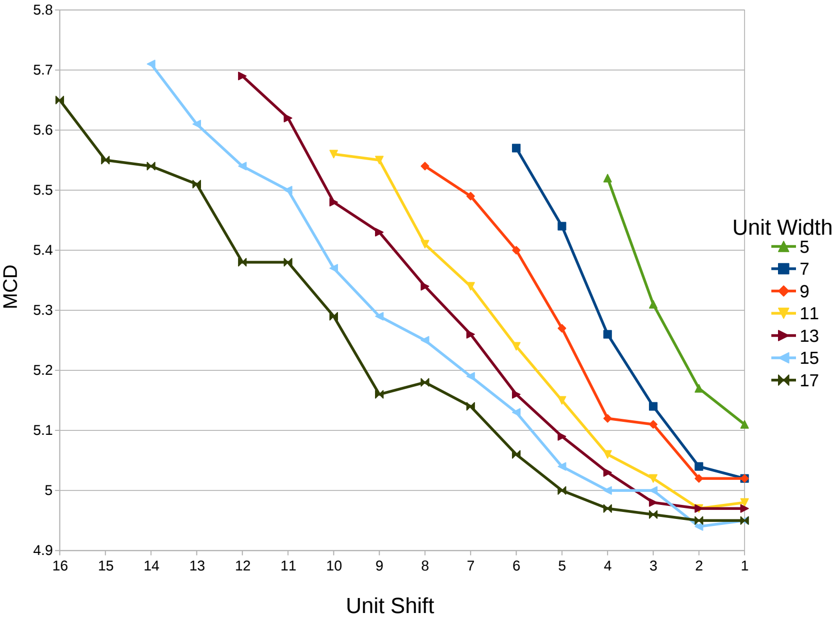


Figure 7.8 – Unit size and unit shift variation and its influence on the mel cepstral distortion.

especially for small unit shifts. There is a tendency of higher unit widths resulting in lower MCDs. Thus, we perform our Unit Selection experiments with a unit size of $w_u = 15$ and a unit shift of $s_u = 2$. Using a frame shift of 10 ms and a frame size of 27 ms, this results in a unit width of 167 ms.

7.4 Mapping using Whispered Speech Data

To the best of our knowledge, there is no (frame-based) Unit Selection approach for the conversion from whispered speech into normal speech described in the literature. Closest related work is proposed by [TBLT08], using Gaussian mixture models and a neural network mapping to generate F0 from whispered speech input. This results in 93.2 % voiced/unvoiced accuracy using neural networks, and 90.8 % using GMMs. A correlation coefficient of 0.499 is reported on the F0 contour. The authors proceed their investigations in [TBLT10], reporting an MCD of 5.99.

We use session-dependent whisper-based conversion systems for each of the nine speakers of our Whisper data corpus, introduced in Section 4.3. Two different evaluations are performed: an MFCC-to-MFCC conversion, that uses whispered MFCC features to estimate MFCCs of audible speech, plus an MFCC-to-F0 transformation, that uses whispered MFCC to compute an F0 contour, which are part of the target segment inside a single unit. All this is done in a frame-by-frame fashion, without considering any additional contextual information. To align the MFCC features of the normally spoken to the corresponding MFCC features of the whispered utterances, DTW (Dynamic Time Warping) is used. We evaluate the basic Unit Selection approach on our Whisper-Corpus *WA-200*. For the issue of misaligned whispered and audible training utterances, we use dynamic time warping (DTW) on the MFCC features.

Table 7.2 – Evaluation results between whispered and normal speech, baseline Mel Cepstral Distortions (MCD) and Unit Selection results: MCDs, Unvoiced/Voiced(U/V)-Accuracies and F0 correlation coefficients (r).

Speaker	Baseline	Unit Selection		
	MCD	MCD	U/V Acc.	r
101	6.44	4.93	85.04	0.664
102	7.12	5.19	86.13	0.710
103	8.37	5.37	78.87	0.566
104	8.46	6.02	77.55	0.541
105	8.22	5.41	84.40	0.673
106	8.29	6.29	76.27	0.426
107	7.75	5.85	78.88	0.564
108	8.51	6.52	74.15	0.436
109	6.75	5.23	82.79	0.605
MEAN	7.77	5.64	80.45	0.576

We calculate the MCD between aligned MFCCs of source whispered utterances and target audible utterances and determine this value the baseline MCD of this experiment, which results in an average MCD of 7.77. The Unit Selection mapping reduces the average distortion to 5.64, representing an MCD reduction of 2.13. An average Unvoiced/Voiced accuracy of 80.45 % with an F0 correlation coefficient of 0.576 can be achieved. Thus, Unit Selection can be regarded a valid approach to generate both F0 and normal MFCCs from whispered input.

7.5 Mapping using EMG Data

Since Unit Selection requires a certain amount of speech data for the codebook, we concentrate the evaluation of this approach on the EMG-ArraySingle-

A-500+ corpus only. The EMG-Array-AS-200 corpus contains only 200 utterances per session with 140 used for training – approximately 7 minutes training data – which gives only unsatisfactory results.

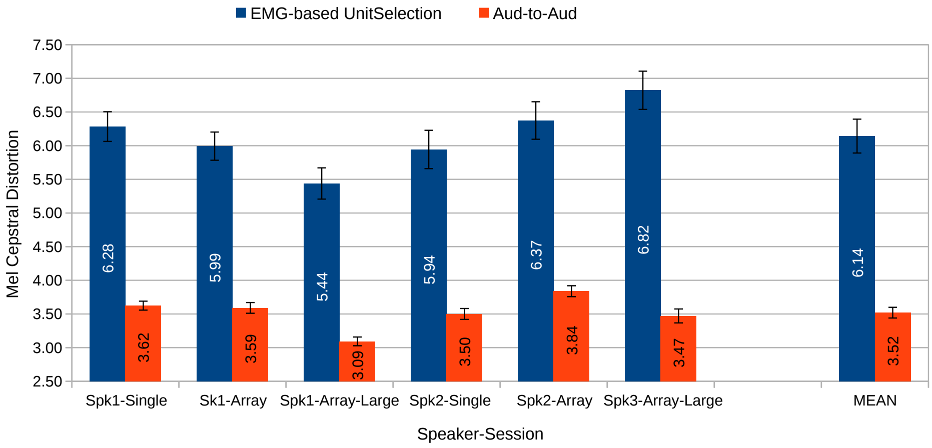


Figure 7.9 – EMG-to-speech Unit Selection MCDs of the EMG-ArraySingle-A-500+ corpus with Aud-to-Aud mapping reference.

The results of the basic Unit Selection system (without Unit Clustering) are depicted in Figure 7.9, in addition to the golden line of the acoustic data based approach Aud-to-Aud. An average MCD of 6.14 is obtained, showing high inter-session dependencies. While Spk1-Array-Large results in a considerably low MCD of 5.44, Spk3-Array-Large performs worst with an MCD of 6.82, although this session contains the most training data. Since Spk3-Array-Large performed on average in the previous Aud-to-Aud experiment, this implies that the session may contain EMG recording problems or articulation inconsistencies.

Figure 7.10 depicts the F0 evaluation for the Unit Selection mapping. An average Voiced/Unvoiced-accuracy of 76.2 % is obtained, with an F0 correlation coefficient of 0.505.

7.5.1 Cluster-based EMG Unit Selection

To investigate the quality of the codebook for unit clustering, we evaluate the effect of reducing the amount of units in the codebook drastically. The experiment is performed on a single session from the EMG-ArraySingle-A-500+ corpus in a cross-evaluation setup: For every training utterance, a

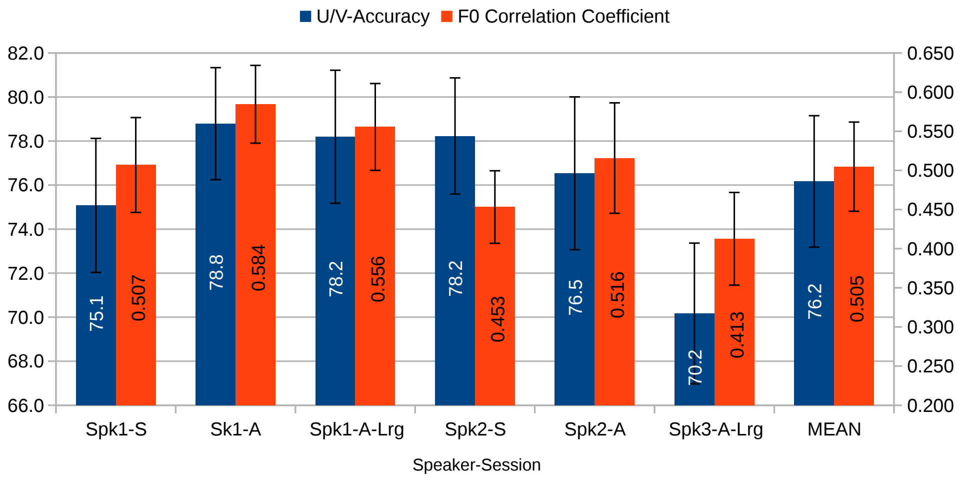


Figure 7.10 – EMG-to-F0 Unit Selection evaluation of the EMG-ArraySingle-A-500+ corpus: Unvoiced/Voiced-accuracy (blue bars) and F0 correlation coefficients (red bars). Error bars denote standard deviation.

unit codebook is created from the training set with that utterance held out. Conversion is then performed for the held out utterance and the list of units that was used to create the output is saved. This process is repeated with acoustic features as both source and target features, like in the previous Aud-to-Aud evaluation. This configuration outputs the best possible audio sequence that can be created using the available units from the codebook.

When we reduce the codebook to units that were used at least once, both in the Aud-to-Aud conversion (leaving only units that provide good output) and EMG-to-Aud conversion, the size of this new codebook is 17,836 units, down from 159,987. This near tenfold reduction leads to a much quicker conversion and has only a minor affect to the mel cepstral distortion at all. The reduced codebook achieves an average MCD of 5.9 on the development set, compared to 5.85 with the original codebook.

From this result, it is obvious that reducing codebook redundancy can be done without considerably reducing the mel cepstral distortion. It also shows that reducing the codebook size without a loss of MCD is feasible and does not require the use of any particularly complicated methods - simply restricting the codebook to units that are useful is already sufficient.

We use the proposed unit clustering to reduce the codebook to a defined amount of cluster units that is varied between 500 and 13,000 cluster units. Figure 7.11 depicts the results of this cluster evaluation for Spk1-Array. The

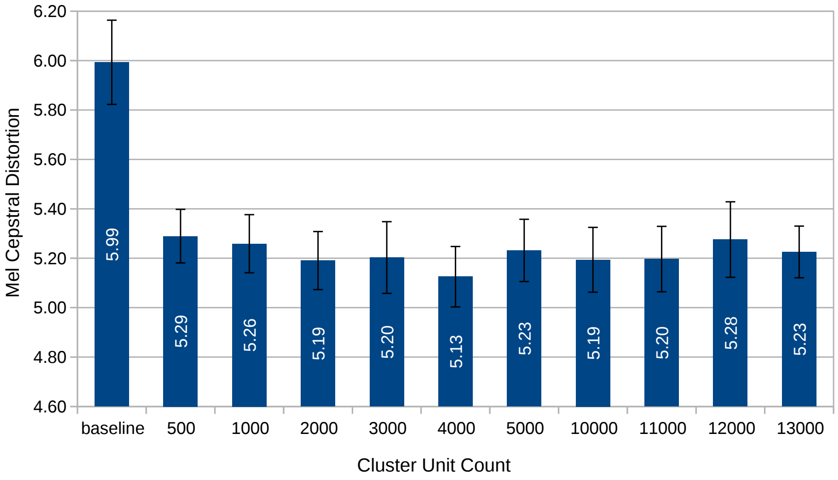


Figure 7.11 – Mel Cepstral Distortions for Unit Selection mapping using Unit Selection baseline and different cluster unit counts for the proposed Unit Clustering approach. The error bars show the standard deviation.

clustering approach clearly helps to reduce the MCD, while the concrete number of cluster units has a minor influence. While there is a notable MCD reduction from the 5.99 baseline, the cluster count variations are between 5.13 and 5.29, a difference that is only slightly perceivable when the output is synthesized.

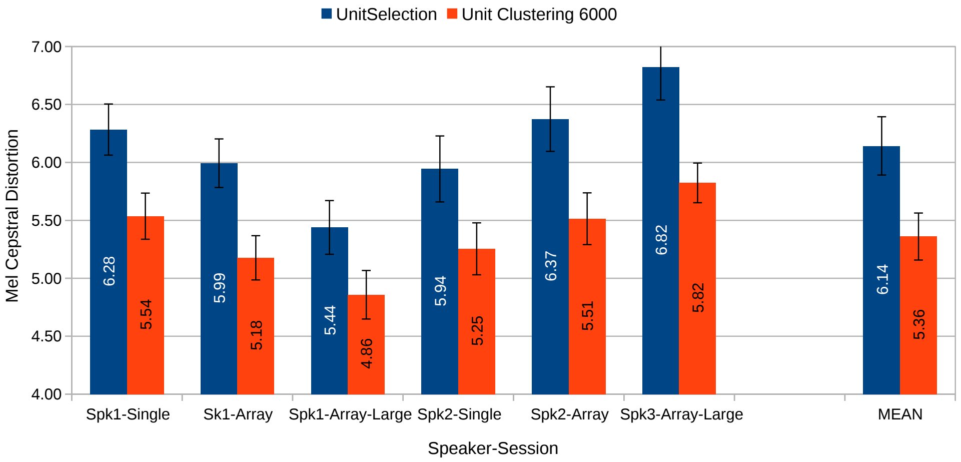


Figure 7.12 – Mel Cepstral Distortions on the EMG-ArraySingle-A-500+ corpus for Unit Selection mapping using Unit Selection baseline and the proposed Clustering approach. The error bars show the standard deviation.

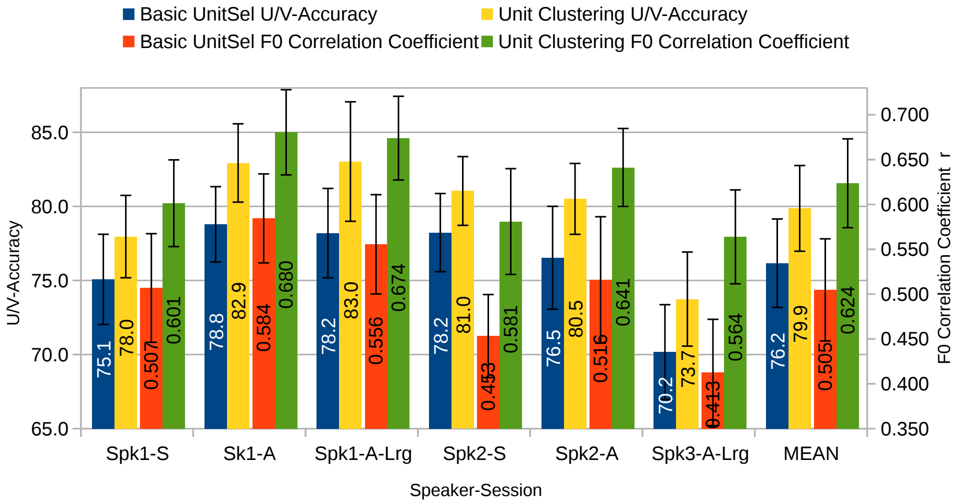


Figure 7.13 – EMG-to-F0 Unit Selection mapping results, basic Unit Selection (without Clustering) versus Unit Clustering: Voiced/Unvoiced Accuracy and correlation coefficient r between output and target F0 contour. The error bars show the standard deviation.

Finally, we apply the Unit Clustering to the full EMG-ArraySingle-A-500+ corpus (see Figure 7.12), which results in an average MCD of 5.36 using 6000 cluster units compared to 6.14 in the baseline Unit Selection. While all sessions obtain a significant ($p < 0.001$) MCD reduction, the best result achieves Spk1-Array-Large with an MCD of 4.86, while the highest MCD reduction is observed with Spk3-Array-Large: The baseline MCD of 6.82 is reduced to 5.82. Comparisons to the other proposed mapping approaches are investigated in Chapter 9.

Chapter 8

Mapping using Deep Neural Networks

This chapter introduces the third approach (red box in Figure 8.1) for a direct conversion system: Artificial Neural Networks. Although they are known for decades, neural networks gained much popularity in recent years. Generation of speech using neural networks has been published by e.g. [KP04, AGO15] using a multilayer perceptron on electromagnetic articulography, electropalatography and laryngography data [KP04], or a deep architecture for articulatory inversion (generating articulatory configurations from speech data) [UMRR12]. Two different artificial neural networks are investigated in this chapter: a classic feed-forward network and a novel recurrent neural network.

8.1 Theoretical Background

Artificial neural networks (ANNs) are computational models that are inspired by biological networks, consisting of interconnected simple processing units entitled *neurons*, that exchange information. The connections between neurons are adapted by different weights. This enables the ANN to be tuned and to learn different characteristics. ANNs can be specified by different parameters, e.g. the type of interconnection between the neurons (recurrent, or feed-forward), the type of activation function that passes the input to the

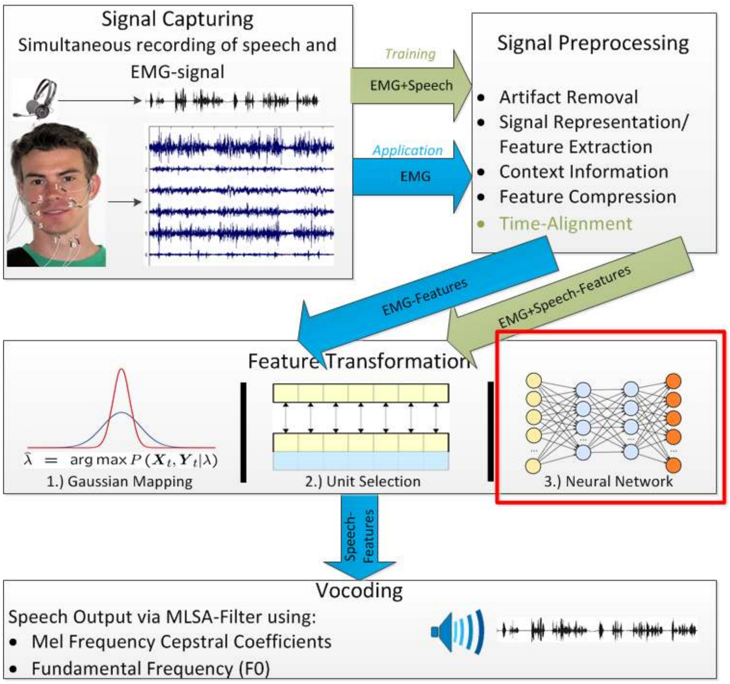


Figure 8.1 – Structure of the proposed EMG-to-speech approach.

output (some logistic function or a rectified linear function) and the type of learning to update the different weights.

We evaluate the direct feature mapping approach on two different kinds of neural networks:

1. Classical feed-forward networks
2. Recurrent Neural Networks (RNNs), i.e. *Long Short-Term Memory (LSTM)* networks [HS97]

8.1.1 Feed-Forward Network

The simplest kind of ANN is a feed-forward type of architecture that consists of three consecutive layer types:

1. Input layer: n-dimensional input data is directly associated to n inputs.
2. Hidden layer(s): the central component of the neural network.
3. Output layer: m-dimensional output is obtained from m output neurons in this layer.

To model more complicated representations, multiple levels of hidden layers are introduced, which is known as *Deep Neural Networks.* In theory every layer gets more details onto the to-be-learned output, like building up different layers of abstraction. For a visual classification task, the first hidden layer would learn to recognize edges, while the second layer would learn more complex structures like circles, etc [PMB13].

The activation function of the last layer acts as an output function, e.g. a soft-max function that is used for classification problems, or a linear output function that may be applied for regression tasks, like the generation of speech from electromyographic features.

In general, neural networks can model complex, non-linear functions, but are bound to many parameters like the number/type of hidden layers and the number of neurons per layer. Given these topological parameters, the ANN needs to be trained, such that a target function represented by a set of observations is learned in an optimal sense. The most prominent way of training neural networks, is the *backpropagation* of errors, in conjunction with an optimization like *stochastic gradient descent (SGD).* The weights of the network are updated based on the error between the network's output and the training data. This delta error is propagated beginning from the output to the input. This results in the weight update Δw:

$$\Delta w = -\alpha \frac{\partial \epsilon}{\partial w},$$

where α is known as the *learning rate* and the "$-$" enables the direction to a minimum. The choice of α has a strong influence on the convergence behavior of the training.

Figure 8.2 shows the architecture of the employed five-layer neural network which we use for mapping the input features to the acoustic space of audible speech. The topology is chosen on empirical basis from prior experiments and confirmed by related work that uses similar architectures [BHG$^+$15, AGO15]. The neural network is trained to map an n-dimensional input feature vector to the target features of audible speech, i.e. if $G(\boldsymbol{x}_t)$ denotes the mapping of $\boldsymbol{x}_t$, then the error of the mapping is given by $\epsilon = \sum_t \|\boldsymbol{y}_t - G(\boldsymbol{x}_t)\|^2$. $G(\boldsymbol{x}_t)$ is defined as

$$\begin{aligned}G(\boldsymbol{x}_t) =&\widetilde{g}(\boldsymbol{w}^{(4)}, \boldsymbol{b}^{(4)},\\ &g(\boldsymbol{w}^{(3)}, \boldsymbol{b}^{(3)},\\ &\quad g(\boldsymbol{w}^{(2)}, \boldsymbol{b}^{(2)},\\ &\quad\quad g(\boldsymbol{w}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{x}_t))))\end{aligned}$$

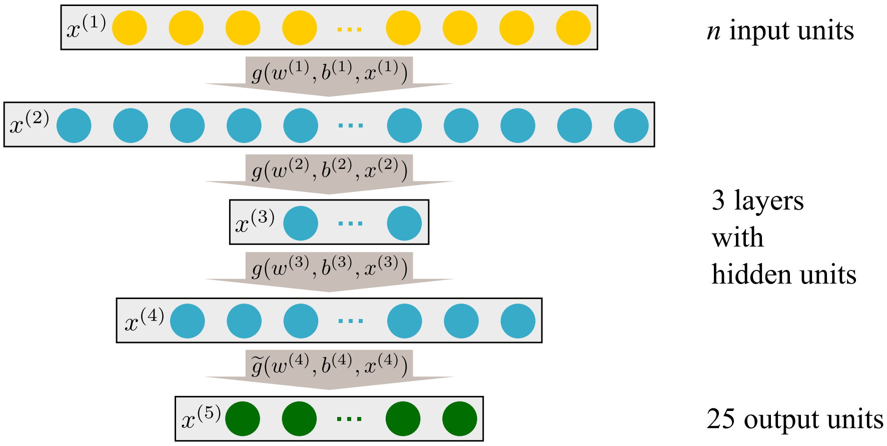

$x^{(1)}$
n input units
$g(w^{(1)}, b^{(1)}, x^{(1)})$
$x^{(2)}$
$g(w^{(2)}, b^{(2)}, x^{(2)})$
$x^{(3)}$
3 layers
with
hidden units
$g(w^{(3)}, b^{(3)}, x^{(3)})$
$x^{(4)}$
$\widetilde{g}(w^{(4)}, b^{(4)}, x^{(4)})$
$x^{(5)}$
25 output units

Figure 8.2 – Structure of the neural network used to convert input features to target Mel Frequency Cepstral Coefficients on a frame-by-frame basis.

where

$$\widetilde{g}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b}$$

and

$$g(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}) = \mathrm{ReL}(\widetilde{g}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}))$$

Here, $\boldsymbol{w}^{(n)}$ and $\boldsymbol{b}^{(n)}$ represent the weight and bias matrices of the hidden and output layers and ReL denotes the rectified linear activation function $\mathrm{ReL}(x) = max(0, x)$. Our own preliminary experiments gave good results using this setup and also other researchers state that substituting logistic units with ReL units results in improved accuracy and fast convergence [ZRM+13].

For the feed-forward neural networks, we ran our first experiments using the Computational Network Toolkit [YES+14], but later sticked to the *brainstorm* implementation [GS15].

8.1.2 Long Short-Term Memory Network

A Long Short Term Memory (LSTM) network is a specific type of recurrent neural networks (RNNs), introduced in 1997 [HS97] and enable a long-range temporal context by using memory cell units that store information over a longer period of time. LSTMs are state of the art for a couple of problems, e.g. recognition of speech [GMH13, SSB14] or hand-writing [DKN14]. The addition of RNNs is motivated in their ability to cope with temporal dependencies directly by the network model, instead of stacking feature frames at the preprocessing level. This is realized by introducing recurrencies into the network model which allow to use the activation patterns from the previous time steps as an input to the current one.

The main element of an LSTM is a memory cell that can maintain information over time. This is accompanied by non-linear gating units that regulate the data flow into and out of the memory cell. The schematic structure can be seen in Figure 8.3. Since its introduction in 1997 the LSTM architecture experienced some modifications (e.g. the first version did not have forget gates). The implementation that is used in this work, matches to the bidirectional LSTM type that is described in [GS05]. For a description of the historic development of LSTMs, we refer to [GSK$^+$15].

There are four units that are connected to the input: a block input, an input gate, output gate and a forget gate. The latter three multiplicative gates determine whether the connected values are propagated into or out of the inner memory cell. The input gate can stop the gate input values, the forget gate resets the cell memory and the output gate triggers the output values. To enable the recurrence, the block output is connected back to the block input and to all three gates. The self connected units in the middle represent the *Constant Error Carousel* (CEC). When no new input or errors are presented, the CEC's inner values remain constant and are stored by the self-connection. There also exist *peephole* connections (not shown in Figure 8.3), which directly link the memory cell to the three gates and help to learn precise timings.

The training of LSTMs is performed using backpropagation through time [Wer90], where the training sequences are processed in temporal order. The bidirectional LSTMs that are used in this work, also use the future context. In the training stage all layers are duplicated, thus one network processes the data in forward order, while the duplicate is processing the data in backward order. It has been shown in speech recognition applications [GS05, WZW$^+$13], that this bidirectional approach outperforms classical LSTMs. The usage of LSTMs for text-to-speech synthesis has been previously been investigated in [FQXS14]. However, it should be noted that the bidirectional characteristic limit the online capability of a feature mapping, due to its usage of future time context.

We optimized most of the LSTM parameters session-dependently and in accordance with the underlying application. The momentum was chosen fixed with 0.9, as well as the number of 2 hidden layers. LSTM research [GSK$^+$15] and our own preliminary experiments showed that these factors have only minor influence on the system performance. To determine the number of

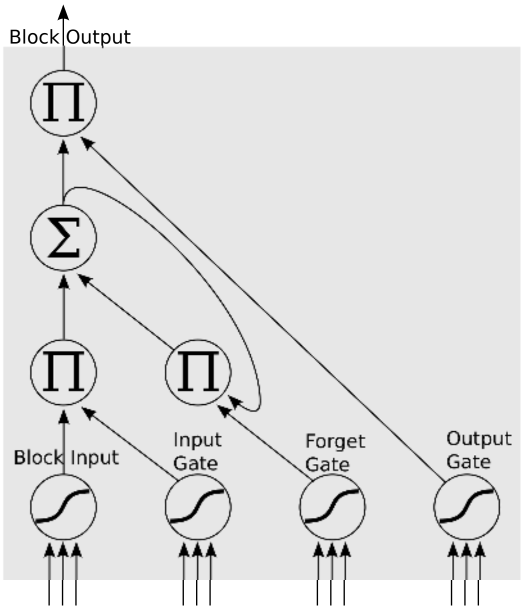


Figure 8.3 – Structure of a long short-term memory network.

memory blocks per layer, we varied the numbers from the literature using our own input data. The speech feature enhancement in [WZW+13] used three hidden layers, consisting of 78, 128 and 78 memory blocks. We use this as orientation and vary different combinations of hidden layer numbers (from 1 to 4) with memory blocks per layer (60, 80 and 100). We obtained best results using two hidden layers, consisting of 100 and 80 memory blocks. To avoid overfitting, we stopped training after 20 epochs without improvement of the validation sum of squared errors. Input and output data was normalized to zero mean and unit variance.

For the evaluation of Long Short-Term Memory (LSTM) networks, we used the CUda RecurREnt Neural Network Toolkit (*currennt*) implementation [WB15]. All LSTM experiments were done using an Intel Core i7 950 CPU with an Nvidia GeForce GTX 970 graphic device.

8.2 Mapping on Whispered Speech

Applying DNNs on whispered speech data is no novel approach, but can give us insights into the feature transformations that are hard to investigate on a complicated signal like the EMG signal. Since the source and target features are speech features, we have a the same preprocessing steps for source and target features and can directly investigate the general feasibility of the feature transformation. Research on whispered Mandarin speech recognition using DNNs was recently presented [LML+15], reporting significant improvements to a baseline GMM approach. The direct generation of speech from whispered input, with a high focus on F0 generation, was proposed by Li et al. [LMDL14]. They use multiple restricted Boltzmann machines and compare their results to a GMM-based approach (see Section 6 for details to this Gaussian mapping). A voiced/unvoiced accuracy of 90.53 % is reported, with an F0 correlation coefficient of 0.61.

We train session-dependent whisper-based conversion systems for each of the nine speakers of our Whisper data corpus, introduced in Section 4.3. Two different DNNs are trained: an MFCC-to-MFCC conversion, that uses whispered MFCC features to estimate MFCCs of audible speech, plus an MFCC-to-F0 transformation, that uses whispered MFCC to compute an F0 contour. All this is done in a frame-by-frame fashion, without considering any additional contextual information. We use DTW (Dynamic Time Warping) to align the MFCC features of the normally spoken utterances to the corresponding MFCC features of the whispered utterances.

8.2.1 Feed Forward DNN

We train three hidden layers for 100 epochs, using a decreasing learning rate. Dropout [SHK+14], which means randomly dropping units from the neural network training, is used to suppress overfitting.

Table 8.1 gives the evaluation results of the proposed DNN system, using the Mel Cepstral Distortion (MCD) measure on the output and target MFCC features plus the voiced/unvoiced accuracy and correlation coefficient r between output and target F0 contour. Baseline MCD refers to the distance between unaltered whispered and target normal speech MFCCs. The DNN mapping gives a significant improvement from an average MCD of 7.77 to 5.46. The F0 results look also encouraging, resulting in an average voicedness accuracy of 79 % and a correlation coefficient of 0.65. Comparing this to the literature [TBLT08] (93.2 % voiced/unvoiced accuracy using neural networks, correlation coefficient of 0.499), we get worse voicedness results, but a better correlation coefficient. Detailed results of the voicedness evaluation are depicted in Figure 8.4. Most of the errors are unvoiced frames that are falsely mapped as voiced frames (17 % $U \rightarrow V$ rate), compared to only 4 % $V \rightarrow U$ rate.

Table 8.1 – Whisper-to-Speech feed-forward DNN mapping results, per speaker: Mel Cepstral Distortions (MCD) between MFCCs, Unvoiced/Voiced (U/V) Accuracy and correlation coefficient r between output and target F0 contour.

Speaker	Baseline	Deep Neural Network		
	MCD	MCD	U/V Acc.	r
101	6.44	4.71	84.14	0.705
102	7.12	5.01	85.90	0.772
103	8.37	5.32	78.98	0.625
104	8.46	5.68	78.69	0.643
105	8.22	5.34	82.78	0.688
106	8.29	6.06	76.28	0.535
107	7.75	5.65	76.92	0.658
108	8.51	6.29	67.93	0.555
109	6.75	5.09	79.39	0.689
MEAN	7.77	5.46	79.00	0.652

8.2.2 Long Short-Term Memory Networks

As already explained in Section 8.1.2, we train two hidden LSTM layers based on related work [GSK+15]. Input and output features are z-normalized.

Table 8.2 – Whisper-to-Speech LSTM mapping results, per speaker: Mel cepstral distortions (MCD) between MFCCs, Unvoiced/Voiced(U/V) Accuracy and correlation coefficient r between output and target F0 contour.

Speaker	Baseline	Long Short Term Memory Network		
	MCD	MCD	U/V Acc.	r
101	6.44	4.52	86.85	0.754
102	7.12	4.77	87.37	0.798
103	8.37	5.03	81.68	0.679
104	8.46	5.46	79.63	0.656
105	8.22	5.09	83.55	0.734
106	8.29	5.72	78.30	0.580
107	7.75	5.44	75.54	0.714
108	8.51	5.99	66.95	0.598
109	6.75	4.85	79.73	0.729
MEAN	7.77	5.21	79.96	0.693

Table 8.2 gives the LSTM evaluation results, using the Mel Cepstral Distortion (MCD) measure on the output and target MFCC features plus the

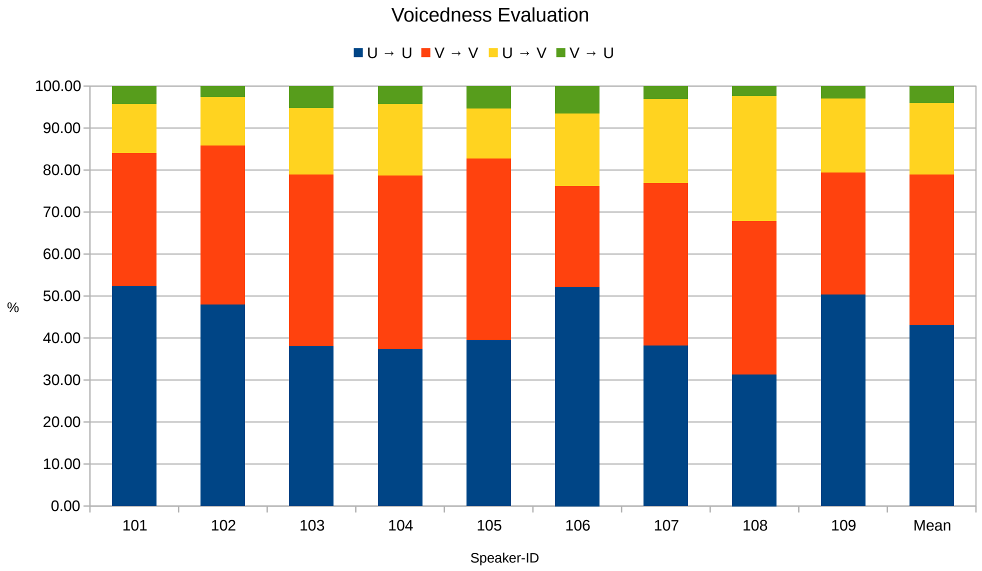


Figure 8.4 – Voicedness results of the DNN-based Whisper-to-Speech mapping. X → Y representing X = reference, Y = output, e.g. U → V are unvoiced frames recognized as voiced.

voiced/unvoiced accuracy and correlation coefficient r between output and target F0 contour. Compared to the feed-forward DNN-based whisper mapping, we get an improvement from an average MCD of 5.46 to 5.21. The F0 results achieve also an accuracy gain, resulting in an average voicedness accuracy of 79.96 % instead of 79.0 % and a correlation coefficient of 0.69 from 0.65. Detailed results of the voicedness evaluation are depicted in Figure 8.5.

8.3 Mapping on EMG data

Tsuji et al. [TBAO08] introduced an EMG-based neural network approach for phone classification and parts of this thesis have also been publicated introducing a direct speech feature regression [DJS15]. Kello and Paul [KP04] presented a DNN-based mapping approach from articulatory positions (electromagnetic articulography (EMA), electropalatograph (EPG) and laryngograph data) to acoustic outputs, obtaining a word identification as high as 84 % conducted in a listening test. A similar articulatory-to-acoustic mapping approach based on Deep Neural Networks (DNN) was presented by

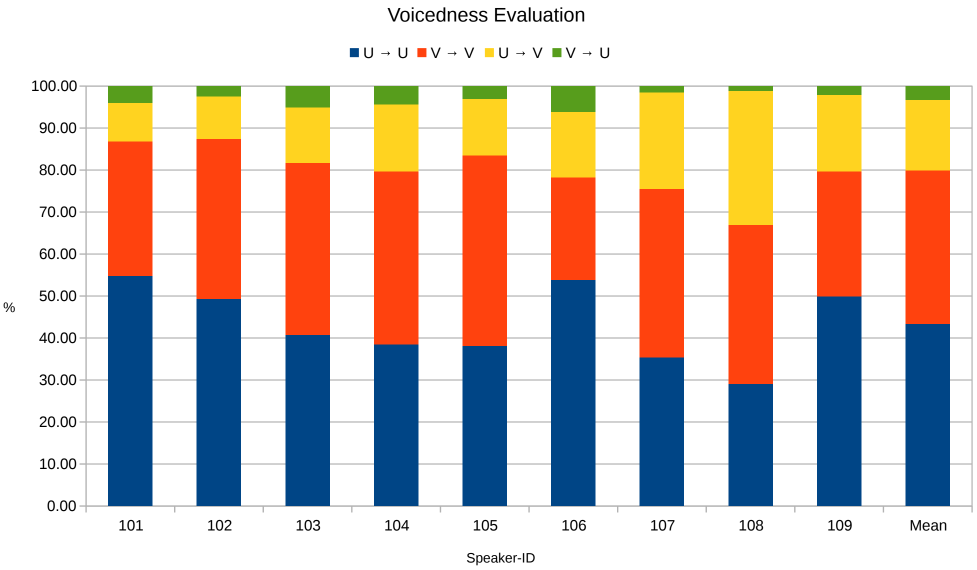


Figure 8.5 – Voicedness results of the LSTM-based Whisper-to-Speech mapping. X → Y representing X = reference, Y = output, e.g. U → V are unvoiced frames recognized as voiced.

[BHG+14] and further investigated in [BHG+15]. They trained on electromagnetic articulography (EMA) data which was recorded simultaneously with the articulated speech sounds. Objective and subjective evaluations on a phone-basis instead of sentences reached a phone recognition accuracy of around 70 %.

We investigate the direct EMG-to-speech mapping based on the proposed artificial neural network types: feed-forward DNN and Long Short Term Memory networks. We estimate the neural network parameters during training using corresponding EMG and speech data. In the final conversion stage EMG signals from the development set are converted to acoustic speech features, i.e. MFCCs and fundamental frequency.

To avoid bias towards numerically larger EMG- or audio features, the signal is normalized to zero mean and unit variance. Since some sessions contain channels with noise and artifacts and since we ran into memory issues with the two sessions with high amount of data, we reduced the number of input channels for these sessions by manually discarding noisy and artifact-prone channels. Table 8.3 shows the selected number of channels and source feature dimensionality the evaluated sessions.

Session	Used Channels	TD0 Dim.	TD15 Dim.
Spk1-S	All 6	30	930
Spk1-A	All 35	175	5425
Spk1-A-Large	18 of 35	90	2790
Spk2-S	All 6	30	930
Spk2-A	All 35	175	5425
Spk3-A-Large	15 of 35	75	2325

Table 8.3 – Number of used channels and source feature dimensionality per session.

8.3.1 Feed Forward DNN

We use a five layer feed forward neural network with the DNN setup described in Section 8.1.1 and perform the following experiments with different combinations of epochs, mini-batch sizes and learning rates per session and report the best results.

As depicted in Figure 8.2, we use the EMG feature vector input, followed by three hidden layers g with different sizes concluded by a final regression layer $\widetilde{g}$ having as many nodes as the number of acoustic output parameters, resulting in an "hourglass" configuration with a $c_2 = 512$ node bottleneck in the center surrounded by a $c_1 = 2500$ node computation layer between input and bottleneck and a $c_3 = 1024$ node computation layer between bottleneck and output. The DNN structure is chosen on empirical basis from prior experiments [JWH+14, DJS15] and our own experience showed only marginal differences using other topologies.

Once the training process has converged, we get a set of weight and bias matrices which represent the mapping function from source EMG features to target acoustic speech features. These matrices can be used in the conversion stage to transform an EMG feature vector to a feature vector of the audible speech.

Figure 8.6 depicts the results for the EMG-ArraySingle-A-500+ corpus. All DNN parameters were optimized independently per session on an additional development set, that consisted of 5 % of the training set. This 5 % development set was separated from the training set and used for the error function during training. The neural network parameters with the lowest error on this 5 % validation set were retained.

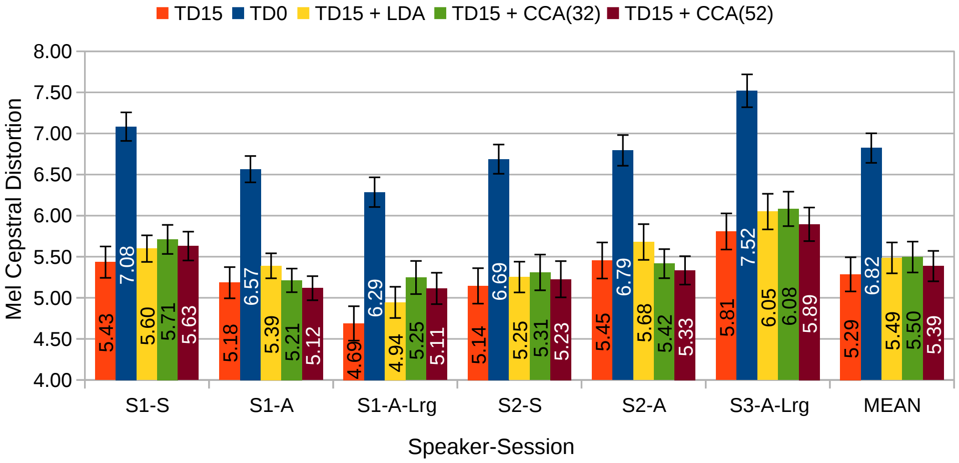


Figure 8.6 – Mel Cepstral Distortions on the EMG-ArraySingle-A-500+ corpus for DNN-based mapping using different EMG feature sets: TD0 vs TD15 vs TD15-LDA vs TD15-CCA-32 vs TD15-CCA-52 features. The error bars show the standard deviation.

While comparable mapping approaches use a feature reduction technique to make the transformation of high dimensional EMG data feasible, neural networks can handle high-dimensional data per se. While e.g. an LDA computation needs phone label information, the usage of direct features need no label alignment and is thus even less error-prone. Having those advantages it shows that using high-dimensional TD15 data gives best MCD results compared to Canonical Correlation Analysis (CCA) and Linear Discriminant Analysis (LDA) feature reduction, resulting in an average MCD of 5.29 using TD15 with the best session giving an MCD as low as 4.69. However, it should be mentioned that not all sessions achieve best MCD results with TD15 features. Although the differences are not significant, Spk1-Array ($p = 0.22$) and Spk2-Array ($p = 0.1$) perform best results with the CCA preprocessed TD15 feature. Spk1-Single ($p = 0.02$), Spk1-Array-Large ($p < 0.001$) and Spk3-Array-Large ($p = 0.04$) perform significantly best with the TD15 feature set.

A considerable MCD reduction can be achieved by using TD15 feature stacking. While unstacked TD0 features obtain an average MCD of 6.82, using adjacent feature information in TD15 boosts the MCD to 5.29. This implies that DNNs have difficulties in modeling time-context information by themselves and Section 8.3.2 will investigate how LTSMs will handle this information.

Independent Component Analysis

To exploit the nature of the array-based recordings and to eliminate redundancies, we use Independent Component Analysis (ICA) (see Chapter 2.7.2 for details) for a decomposition of the raw EMG signal into statistically independent single components, which are used for the further preprocessing steps. Since the multi-channel array electrodes are positioned in close proximity to each other, adjacent electrodes record similar signals, which justifies the ICA application. We modify the logistic infomax ICA algorithm [BS95], as implemented in the Matlab EEGLAB toolbox [DM04]. We separate the EMG channels into two sets: one set representing the chin array, one set for the cheek array. The ICA transformation is then computed for both array sets independently per session, using only the training data sets. We interpret the obtained components as new input signals and select which of these new components are deemed as artifacts and discard the selected noise components. This is followed by the standard TD15 feature computation.

Figure 8.7 shows the MCD results on the EMG-ArraySingle-A-500+ corpus.

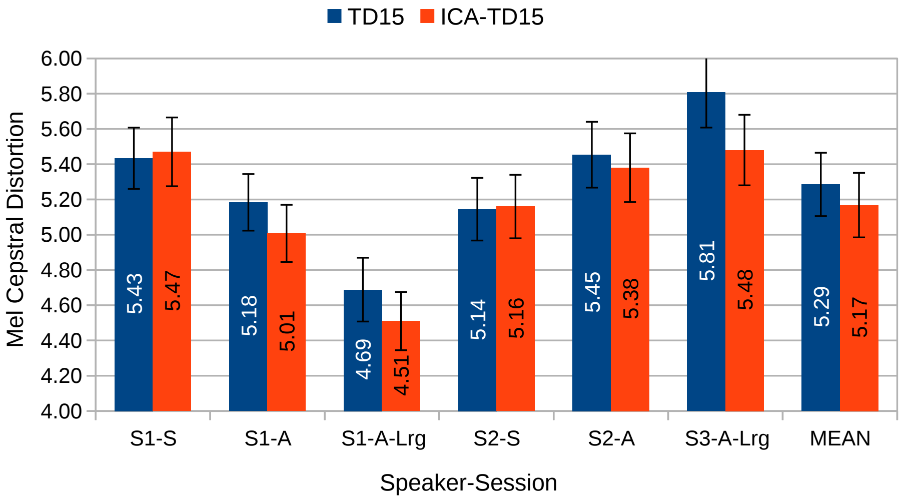


Figure 8.7 – Mel Cepstral Distortions on the EMG-ArraySingle-A-500+ corpus for EMG-to-MFCC DNN mapping results with Independent Component Analysis preprocessing (ICA-TD15) and without ICA (TD15). The error bars show the standard deviation.

An average MCD of 5.17 marks the best MFCC output that was obtained so far, with the best session (Spk1-Array-Large) achieving an MCD of 4.51. Since ICA utilizes the multi-channel array setup, the MCDs on the single-electrodes setup with only six channels get a small performance degradation. The improvement on Spk2-Array is not significant ($p > 0.05$), while the others are (Spk1-Array $p < 0.01$, Spk1-Array-Large and Spk3-Array-Large $p < 0.001$). In general, this experiment shows that the separation into independent components from multi-channel array input helps the neural network to improve the mapping results.

EMG-to-F0

An additional neural network is trained using the best performing TD15 feature-set with target F0 features instead of MFCCs. Results are shown in Figure 8.8.

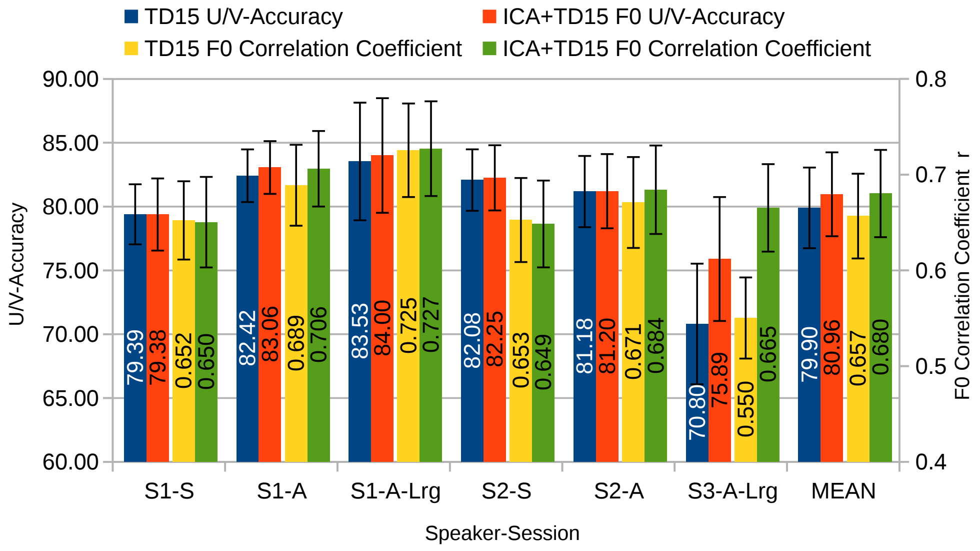


Figure 8.8 – EMG-to-F0 DNN mapping results, without ICA (blue/yellow bars) and with ICA preprocessing (red/green bars): Voiced/Unvoiced Accuracy and correlation coefficient r between output and target F0 contour. The error bars show the standard deviation.

Adding the proposed ICA preprocessing gives generally a little U/V-accuracy improvement from 79.9 % to 80.96 % , with a raised correlation coefficient from 0.657 to 0.68. The results per session are additionally depicted in Figure 8.8. Since ICA performs well on high-dimensional input data, we see only slight differences on the single-electrodes data. Comparing these results to the Whisper-to-F0 mapping, where we achieved 79 % V/U-accuracy with a correlation coefficient of 0.65, the EMG-to-F0 output is encouraging.

Silent Data

After obtaining and improving our results on the EMG-ArraySingle-A-500+ corpus, we investigate the application of neural networks on silent EMG data. Therefore, we use the best performing TD15 setup. As already discussed in Chapter 6, we lack matching target data in the silent EMG domain. Thus, we evaluate three different experimental setups:

- The baseline EMG-to-speech mapping using audible EMG for train and test data (Aud Train - Aud Test).
- Using audible EMG data for training the neural network, and silent EMG data for testing (Aud Train - Sil Test).
- Using silent EMG data for training and testing (Sil Train - Sil Test).

The latter is difficult since we need to use the target acoustic data in training and testing, which is not exactly matching the EMG signals. To compensate for the temporal differences, we apply dynamic time warping (DTW) to align the silent EMG signal with the acoustic data. We still face the problem that the silent EMG signal and the acoustic signal are not simultaneously recorded and thus may include inconsistencies that may not be fully alleviated by DTW. Figure 8.9 depicts the Mel Cepstral Distortions for the silent and audible EMG data of the EMG-Array-AS-200 corpus.

While the training on silent EMG data improved the results in the Gaussian mapping approach (see Section 6.3.1), only three sessions (S1-2, S2-4, S4-1) benefit from the usage of silent EMG data in DNN-based training. When a feature reduction technique (i.e. LDA) is applied, a similar result is obtained. Although the Sil Train - Sil Test MCDs get further reduced to an average of 6.81, the Aud Train - Sil Test setup performs still better with a mean MCD of 6.6.

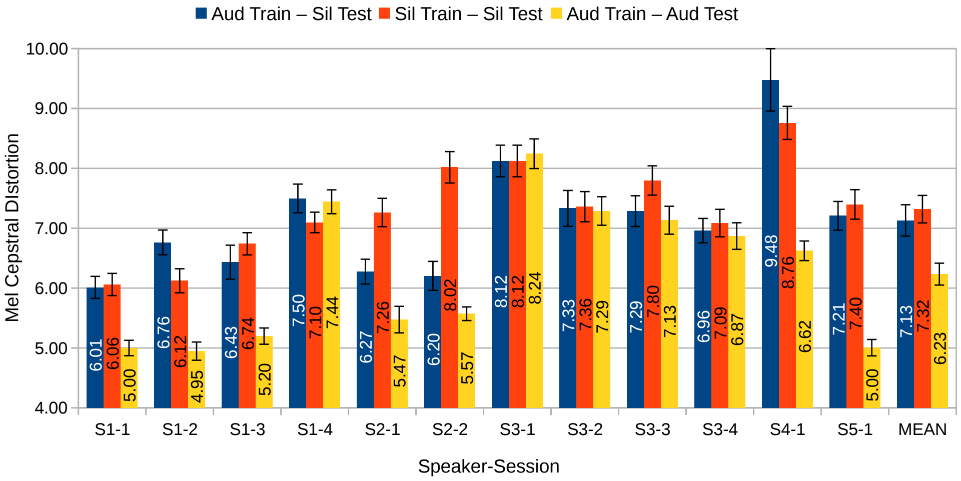


Figure 8.9 – DTW-aligned Mel Cepstral Distortions using feed-forward DNNs for audible EMG training with silent EMG testing (blue bars), silent EMG training with silent EMG testing (red bars) and audible EMG training with audible EMG testing (yellow bars), evaluated on the EMG-Array-AS-200 corpus. Error bars denote standard deviation.

8.3.2 Long Short-Term Memory Networks

Context Feature Evaluation

One of the promising benefits of LSTMs is the intrinsic modeling of time-contextual information, thus including the context-information from the feature domain directly into the mapping/modeling part. An interesting investigation is to look at the performance of an LSTM, when different context-based features are presented. Therefore, we train LSTM networks with different kinds of EMG feature sets:

1. Context-independent *TD0* features,
2. high dimensional context-dependent (15 preceding and succeeding frames) *TD15* features,
3. *TD15+LDA* features: TD15 features, that are post-processed with an LDA to reduce the feature vector to 32 dimensions,
4. *TD15+CCA (32)* features: TD15 features, that are post-processed with a CCA to reduce the feature vector to 32 dimensions,

5. *TD15+CCA (52)* features: TD15 features, that are post-processed with a CCA to reduce the feature vector to 52 dimensions.

Details on the dimension reduction techniques can be found in Section 2.7. The LDA is optimized to maximize discriminability of the phone classes, while the CCA is trained to optimize the correlation to stacked MFCC features.

Figure 8.10 shows the MCD results for all six sessions of the EMG-ArraySingle-A-500+ corpus. The results show that for all of the examined sessions, the high dimensional TD15 setup performs worst, while TD0 features without context information are notably better. This implies that the LSTM network indeed is capable of learning the temporal context by itself. The results can be further improved when stacking *and* feature reduction is used, achieving an average MCD of 5.43 with LDA preprocessing and 5.51 using CCA feature reduction, which can be improved to 5.38, when a 52 dimensional post-CCA feature is retained. Thus, using CCA shows two advantages compared to LDA: It does not require phone alignments for training, and it results on average in a better performance in terms of MCD.

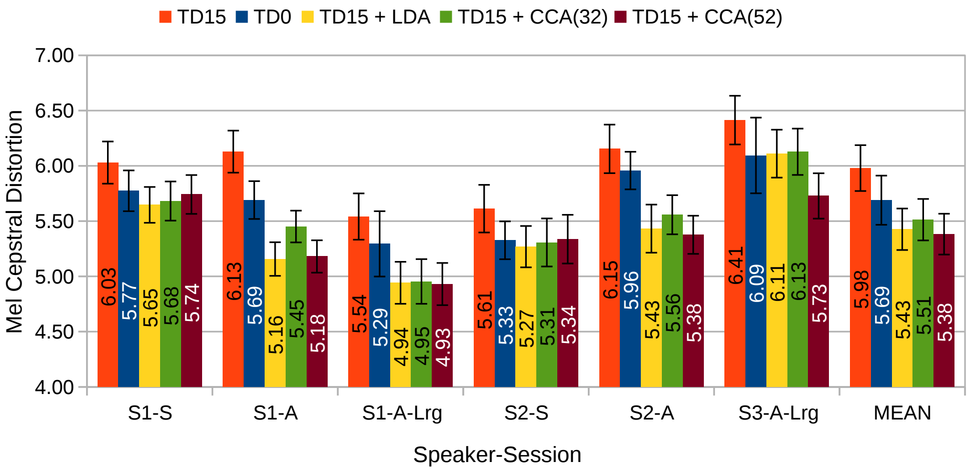


Figure 8.10 – Mel Cepstral Distortions on the EMG-ArraySingle-A-500+ corpus for LSTM-based mapping from different EMG feature sets: TD0 vs TD15 vs TD15-LDA vs TD15-CCA-32 vs TD15-CCA-52 features. The error bars show the standard deviation.

Multitask Learning

Multitask Learning [Car97] is a machine learning approach where a secondary related task is added to a primary main task with the intention to improve the model by learning commonalities of both tasks. Both tasks are learned jointly and in theory generalization is obtained by adding an additional information source. In the context of neural network based EMG-to-speech mapping, a secondary target feature is presented for training, e.g. an LSTM is trained to predict two different features: MFCCs plus phone class labels. Since these primary and secondary tasks are related, the hypothesis is that the network weights will be shared among tasks, which leads to a better representation of the feature relationship and thus an improved performance of the mapping.

When the multitask neural network parameters are learned, the final feature mapping of the primary task can be obtained as before: simply mapping EMG features to MFCCs. Thus, there is no need to include phone labels into the vocoder component, they are only required during training. To apply the multitask learning approach, we use our previously best LSTM setup and change the output layer to 71 dimensions. The first 25 dimensions present the MFCC output while the remaining 46 dimensions correspond to the 45 English phone classes plus a silence class. For each frame the feature vector entry of the current phone is set to 1, while the others are set to 0. A second variation is the usage of "subphones", which further divides each phone p into three position-dependent phone classes: p-begin, p-middle p-end, representing the begin, middle and end of a single phone. Thus, we use an 161-dimensional vector instead of the previous 71-dimensional feature. In a third experiment setup we use a simple delta feature for the secondary task target, i.e. the derivative of the MFCC feature sequence.

Figure 8.11 shows the evaluation results for the EMG-ArraySingle-A-500+ corpus.

Multitask learning with additional phone class information gives no notable improvement over the TD15+CCA-to-MFCC baseline system. While two sessions give a slight MCD improvement, 4 sessions perform even worse, possibly based on the increased amount of training dimensions or due to inaccuracies in the phone alignment. Using delta features for the secondary task reveals a slight and constant (however not significant with $p \approx 0.3$) improvement on all six sessions and thus obtains the best results of a mean MCD of 5.35 with the LSTM-based mapping approach.

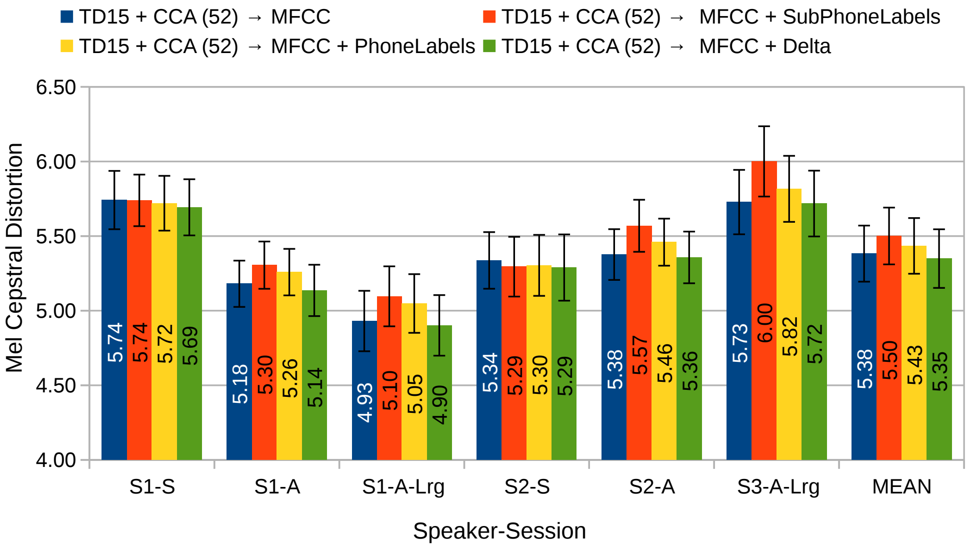


Figure 8.11 – Mel Cepstral Distortions on the EMG-ArraySingle-A-500+ corpus for LSTM-based mapping using Multi Task Learning. The error bars show the standard deviation.

EMG-to-F0

For a final evaluation of audible EMG-to-speech mapping, we estimate F0 information directly from EMG data of th EMG-ArraySingle-A-500+ corpus. We thus use the introduced LSTM approach and replace the target MFCCs with F0 features. Figure 8.12 shows the evaluation results for the EMG-ArraySingle-A-500+ corpus.

Silent Data

As already evaluated with DNNs in Section 8.3.1, we investigate the application of LSTMs on silent EMG data. Due to the lack of simultaneously recorded target data in the silent EMG domain, we use DTW alignment and evaluate three different experimental setups:

- The baseline results using audible EMG for train and test data (Aud Train - Aud Test).
- Using audible EMG data for training the LSTM, and silent EMG data for testing (Aud Train - Sil Test).

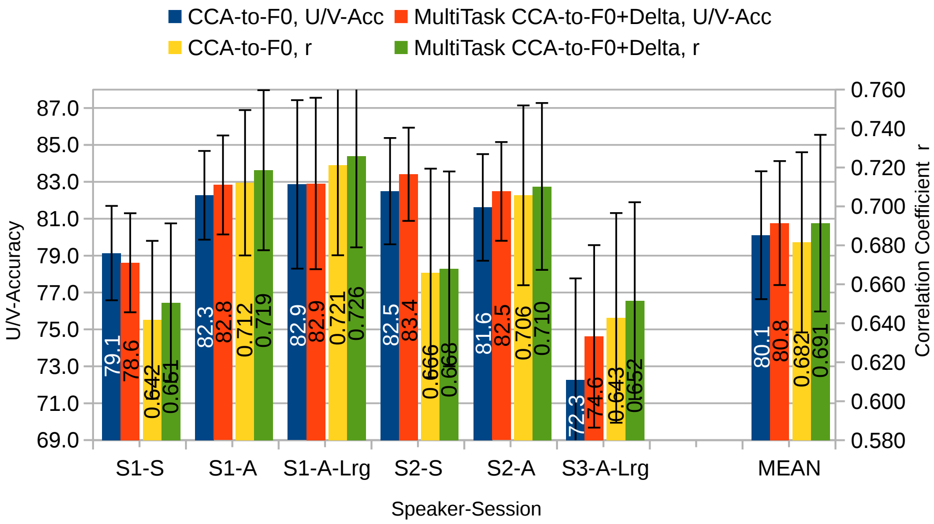


Figure 8.12 – EMG-to-F0 LSTM mapping results: Voiced/Unvoiced Accuracy and correlation coefficient r between output and target F0 contour. The error bars show the standard deviation.

- Using silent EMG data for training and testing (Sil Train - Sil Test).

Figure 8.13 shows the Mel Cepstral Distortions for the three setups on the EMG-Array-AS-200 corpus. It can be observed that generally the MCD for silent test data is notably higher (average MCD of 7.14) than for audible test data (average MCD of 5.79), revealing the differences between silent EMG and audible EMG. Using the DTW-aligned silent EMG data for DNN training *and* testing, improves the MCD to 6.65.

8.3.3 Comparative Summary

Table 8.4 summarizes the evaluation results obtained in this chapter to provide a final comparison of DNN and LSTM results.

Table 8.4 – Comparative summary of LSTM and DNN experiment results.

	Whisper-to-Speech			EMG-to-Speech			
	MCD	U/V	r	MCD	U/V	r	Sil. MCD
LSTM	5.21	80.0	0.693	5.35	80.8	0.691	6.65
DNN	5.46	79.0	0.652	5.17	81.0	0.680	6.60

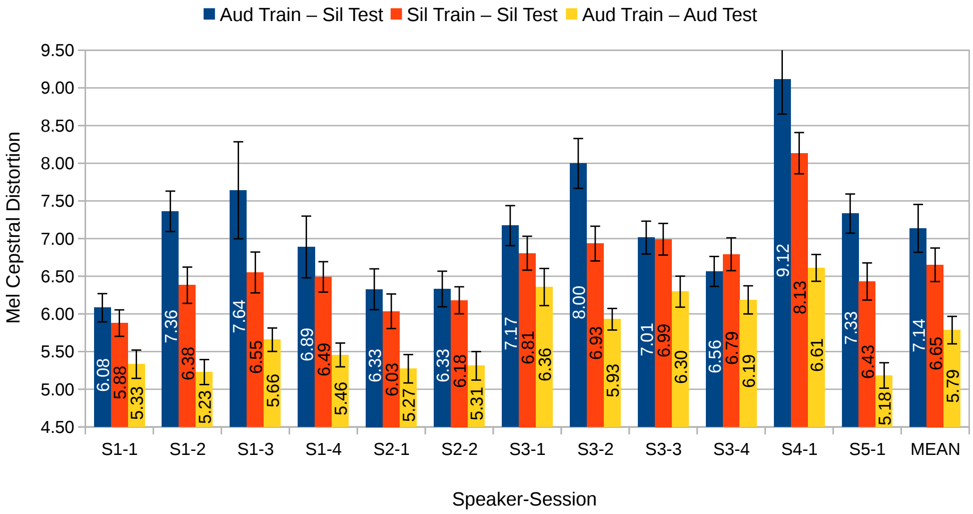


Figure 8.13 – DTW-aligned Mel Cepstral Distortions using Long Short Term Memory networks for audible EMG training with silent EMG testing (blue bars), silent EMG training with silent EMG testing (red bars) and audible EMG training with audible EMG testing (yellow bars), evaluated on the EMG-Array-AS-200 corpus. Error bars denote standard deviation.

Comparing both approaches, the LSTM network performs better using input whispered speech data, while the DNN performs best MCD results using EMG data. Since the DNN achieves best results using TD15 features, the coupling of feature reduction and feature mapping outperforms the LSTM, which is not capable of dealing with high-dimensional feature input. This may also explain, why the whisper-to-speech LSTM outperforms the DNN when 25-dimensional input MFCCs are used and thus no feature reduction is necessary. On the contrary, the best session-dependent MCD result could be obtained with the feed-forward network, resulting in an MCD of 4.51. While the DNN can also handle high-dimensional input with remarkably good results, the LSTM-based mapping performs best with low-dimensional data that must be obtained by a feature reduction technique.

Although the current mapping setup works on a per-utterance basis, it is advisable to prefer the DNN approach to the LSTM setup. Especially considering the property that the bidirectional LSTM has a clear disadvantage when it comes to real-time applications, an aspect (among others) that will be further investigated in the comparative evaluations in Section 9.

Chapter 9

EMG-to-Speech: Comparative Evaluation

While the last chapters introduced three different techniques for the EMG-to-speech mapping, this chapter gives a comparison of the proposed approaches. Objective output evaluations of the generated acoustic output as well as quantitative run-time analyses are presented. This is concluded by a subjective qualitative evaluation represented by a listening test conducted on ten participants, which rated the output quality.

9.1 Objective Comparison between Mapping Approaches

In the following sections we will perform two different objective output quality evaluations on the proposed EMG-to-speech mapping approaches: mapping performance based on mel cepstral distortions (MCD), and the performance of the synthesized output on an automatic speech recognition system. This is concluded by a third objective evaluation: a quantitative comparison of the conversion times of the proposed EMG-to-speech approaches.

9.1.1 Mel Cepstral Distortion Comparison

In the previous chapters, we optimized and evaluated three different approaches for the direct transformation of EMG features to speech: Gaussian mapping, neural network based techniques (LSTMs and feed-forward DNNs) and Unit Selection. Figure 9.1 shows the mel cepstral distortions on the evaluation set of the EMG-ArraySingle-A-500+ corpus.

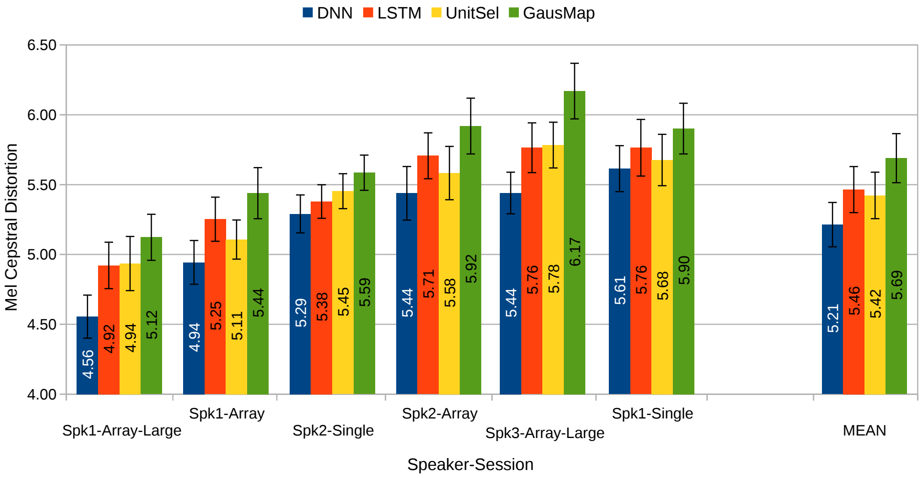


Figure 9.1 – MCD Comparison on the evaluation set of all investigated EMG-to-speech mapping approaches: Deep Neural Networks (DNN), Long Short Term Memory Networks (LSTM), Unit Selection (UnitSel) and Gaussian Mapping (GausMap). The error bars display standard deviation.

While there is only little difference between LSTM and Unit Selection (mean MCD of 5.46 vs 5.42), DNNs significantly ($p < 0.01$) give the best results with an average MCD of 5.21. Gaussian Mapping obtains the highest mean MCD with 5.69. The best session-dependent result is achieved on the Spk1-Array-Large session with an MCD of 4.56. Although Spk3-Array-Large contains the largest amount of training data (111 minutes versus e.g. 28 minutes with Spk1-Array), the MCD is above the average MCD of 5.21, indicating that the speaker encountered inconsistencies during recording the data. When a comparison of array-based recordings versus single-electrodes recordings is made, no superiority in terms of MCD can be stated. While speaker 1 achieves better results on EMG-array data, speaker 2 obtains lowest MCDs with the single-electrodes recording.

9.1.2 Speech Recognizer Evaluation

To interpret the EMG-to-speech synthesis results, we additionally use an automatic speech recognition (ASR) system [SW10] to decode the synthesized evaluation sentences. The goal of this evaluation approach is to objectively evaluate the general intelligibility of the converted speech output rather than a spectral similarity with a reference utterance. The acoustic model of the recognizer consists of Gaussian mixture models. The number of Gaussians is determined by a merge-and-split algorithm [UNGH00]. The recognizer is based on three-state left-to-right fully continuous Hidden-Markov-Models using bundled phonetic features (BDPFs) for training and decoding. Phonetic features represent properties of a given phoneme, such as the place or the manner of articulation, which can improve an acoustic speech recognizer [KFS02]. Since a phonetic feature is generally shared by multiple phonemes, the usage of combined training data can make use of these phonemes in order to train a phonetic feature model more reliably than a single phone model. Thus, a BDPF-based recognition system performs well on small data sets, like they exist in the scope of this thesis, compared to general speech domains where many hours of training data exist. Following the work from [SW10] we use a set of nine phonetic features: Voiced, Consonant, Vowel, Alveolar, Unround, Fricative, Unvoiced, Front, Plosive. These phonetic features additionally can be grouped into bundles, e.g. “voiced fricative”, giving the term bundled phonetic features.

We train the acoustic speech recognizer on the synthesized training set that was generated from the EMG input and test on the synthesized evaluation set. This is done on each of the proposed EMG-to-speech mapping techniques: Gaussian Mapping, DNN-based mapping, LSTM-based mapping and cluster-based Unit Selection. We use the array-based EMG data from Speaker 1 and Speaker 2 (contained in our EMG-ArraySingle-A-500+ corpus), which was also used in related work [WSJS13] and gives us the option for a comparison to the EMG-based speech recognition results reported by Wand et al. [WSJS13].

For decoding, we use the evaluation data corpus together with a trigram language model estimated on broadcast news. To ensure comparability to [WSJS13], we restrict the decoding vocabulary to three different sizes: 108, 905 and 2,111 words, including word variants.

We use the *Word Error Rate (WER)* to measure the performance. The hypothesis and reference text are aligned and the number of insertions, deletions and substitutions (n_i, n_d, n_s) are counted. Thus, for n words in the reference, the WER can be calculated as:

$$WER = \frac{n_i + n_d + n_s}{n} \cdot 100\,\%.$$

The final ASR results are given in Figure 9.2. In general, the DNN-based mapping obtains the best word error rates, confirming the good results that were previously reported in the MCD evaluation. The Gaussian Mapping output achieves second best results in the ASR evaluation, e.g. with a WER of down to 3.5 % (Spk2) using a small vocabulary of 108 words, which also marks the best session-dependent ASR result. Since the Gaussian Mapping uses GMMs that are trained on MFCCs, while the acoustic model GMMs of the recognition system are trained on phone classes, we are not expecting a systematic bias of the ASR system. The listening test in Section 9.2 will additionally investigate the output quality of the mapping approaches. Increasing the vocabulary size raises the WER notably. However, this WER growth has a stronger effect on Spk2 than on Spk1. While a dictionary extension from 108 words to 2,111 words approximately doubles the WER on Spk1, Spk2 has at least a threefold WER increase. This indicates that the output from Spk2 is more confusable. While a small vocabulary can compensate this effect, a large amount of words reduces the discriminability.

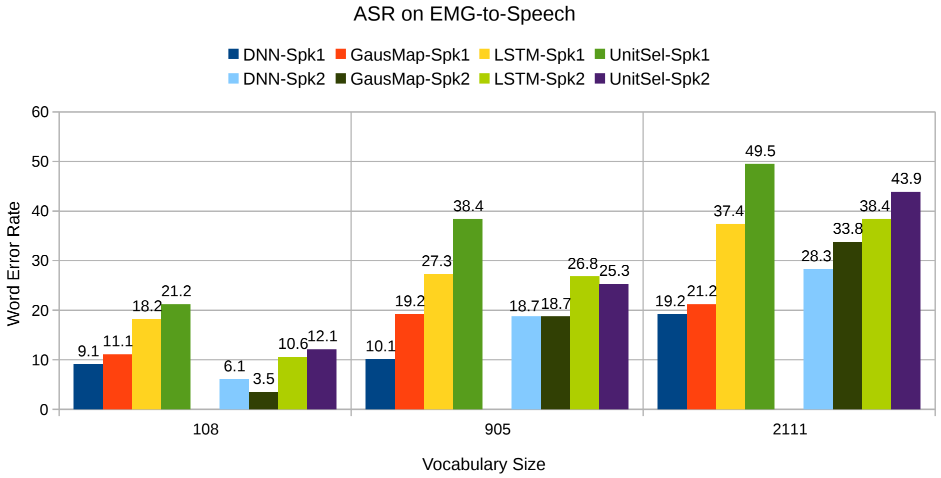


Figure 9.2 – Word error rates of a speech recognition system trained and tested on two different speakers with Deep Neural Network (DNN) versus Gaussian Mapping versus Long Short Term Memory (LSTM) versus Unit Selection (with Clustering) output. Three different decoding vocabulary sizes (including variants) were used: 108, 905 and 2,111 words.

Comparing the obtained word error rates to related work that uses the same setup and a similar corpus on an EMG-based ASR system [WSJS13], our results are encouraging. Wand et al. [WSJS13] achieve an average word error rate of 10.9 % on a vocabulary of 108 words with 160 training utterances, while we can report a mean WER of 7.3 % using our Gaussian Mapping approach. Toth et al. [TWS09] observed the effect that "WER results for EMG-to-speech are actually better than results from training the ASR system directly on the EMG data". Our higher amount of training data may result in the fact that we get even better results, thus a direct comparison is complicated, but shows that we achieve very good results when the synthesized output is used on an ASR system.

9.1.3 Run-Time Evaluation

Since one of the motivations for the direct transformation from EMG to speech is its fast processing time, we will evaluate the real-time capability of the proposed conversion approaches. Thus, we evaluate the conversion time of the validation set on the EMG-ArraySingle-A-500+ corpus using the proposed mapping techniques. A few statements should be kept in mind for this quantitative run-time comparison:

- Only the *conversion time* for mapping EMG features to MFCC features is stated. The additional model-loading, synthesis time and file-I/O is assumed to be constant between different mapping approaches and thus omitted in the comparison evaluation.
- Mapping is done on complete utterances (although in an online situation this frame-to-frame mapping could in theory be started in an incremental fashion – except for LSTMs, which need complete utterances).

Since we want to evaluate the best performing parameter setups per mapping approach, the input EMG feature set depends on the mapping approach. Thus, we use two different input EMG feature sets: 32-dimensional TD15+LDA-features and the high-dimensional TD15 features. The TD15 dimensionality depends on the session, ranging from 930 dimensions on the single-electrode recordings to 5,425 dimensions on the 35-channels array recordings. The Gaussian mapping is set up with 64 Gaussian mixtures. Increasing the number of Gaussians is linearly related to the conversion time, i.e. using 32 Gaussians approximately halves the conversion time.

The result of the quantitative run-time evaluation for the corpus-based approaches (cluster-based unit selection versus baseline unit selection) can be found in Table 9.1, while Table 9.2 presents the conversion times of the model-based approaches: Feed-forward Neural network (DNN) versus Long Short Term Memory (LSTM) network versus Gaussian Mapping (GMM). All times were obtained on an Intel Core i7-2700 CPU running at 3.5 GHz. The numbers in brackets state the real-time factors: the ratio of conversion time to the duration of the utterances, i.e. the sum of the development set (see Chapter 4.2.3). Thus, a real-time factor of 1 means that 1 second of input requires 1 second pure conversion time. The Unit Selection conversion times are heavily related to the amount of training/codebook data, since the unit search is calculated on the whole codebook. Thus, the "Large"-sessions obtain considerably high conversion times. Therefore, the proposed baseline Unit Selection is less useful for real-time setup and even with the proposed unit clustering approach, which considerably reduces computation time, the real-time factor stays above 1.

Table 9.1 – Computation time and real-time factor on the development set for the baseline unit selection system as well as cluster-based unit selection using 6,000 cluster units.

Session	Time taken for conversion in [hh:mm:ss] (RT-Factor)	
	Baseline	Clustering
S1-Single	1:02:39 (22.5)	0:05:02 (1.8)
S2-Single	1:03:05 (23.4)	0:05:57 (2.2)
S1-Array	1:23:11 (27.7)	0:05:29 (1.8)
S2-Array	0:46:24 (19.2)	0:04:22 (1.8)
S1-Arr-Lrg	10:27:06 (81.6)	0:27:35 (2.3)
S3-Arr-Lrg	27:02:22 (136.5)	0:17:34 (2.3)

Compared to the Unit Selection results, the model-based neural network and Gaussian Mapping approaches are independent from the amount of training data, since they only need to load the previously trained models. However, they are heavily influenced by input dimensionality and length of recordings (see Chapter 4.2.3). The neural network conversion time is faster than all the other compared approaches. Even with a high-dimensional feature input (S1-Array and S2-Array having a input feature 5,425 dimensions), the mapping is performed below 0.1 times real-time, making this method the preferred choice for an online EMG-to-speech system. Using the reduced 32-

Table 9.2 – Computation time and real-time factor on the development set for DNN, LSTM and Gaussian Mapping (GMM) approaches using high-dimensional TD15 features and 32-dimensional LDA processed features.

Session	Time taken for conversion in [sec] (RT-Factor)			
	DNN(TD15)	LSTM(LDA)	LSTM(TD15)	GMM(LDA)
S1-Single	2.9 (0.02)	5.6 (0.03)	26.6 (0.16)	42.7 (0.26)
S2-Single	2.9 (0.02)	5.4 (0.03)	23.4 (0.14)	41.4 (0.26)
S1-Array	12.3 (0.07)	6.1 (0.03)	161.4 (0.90)	45.7 (0.25)
S2-Array	10.2 (0.07)	4.7 (0.03)	144.9 (1.00)	35.5 (0.24)
S1-Arr-Lrg	14.6 (0.03)	15.5 (0.03)	139.9 (0.30)	118.7 (0.26)
S3-Arr-Lrg	16.2 (0.02)	23.4 (0.03)	155.4 (0.22)	181.5 (0.25)

dimensional feature input, Gaussian mapping still achieves 0.25 times real time, not excluding it from a possible real-time application.

9.2 Subjective Quality Comparison: Listening Test

The MCD score is a perceptually motivated distance measure and thus correlates with human perception of speech quality [Kub93]. However, this quite simple Euclidean distance measure inherits some drawbacks: e.g. there is no judgment of the naturalness or prosodic information. For this reason, we conduct a set of subjective listening test evaluations, where participants were asked to listen to the outputs of the proposed EMG-to-speech systems and compare them to each other.

Listening Test Setup

All listening tests were performed as a combination of comparison test and opinion-score test with an additionally given reference. The participating subject listens to five synthesized output variations that were obtained from the same sentence using the proposed EMG-to-speech mapping approaches: Mapping using Deep Neural Network (DNN), Long Short Term Memory (LSTM) network, Unit Selection and Gaussian Mapping. Additionally, a re-synthesized reference recording (resynth) of the utterance was presented. This resynth variant is obtained by using extracted speech features (MFCCs + F0) from the target audio with the MLSA filter to produce a “re-synthesized” reference. This reference contains the quality degradations from feature processing and vocoding steps and thus represents the best output we can achieve with our synthesis setup.

The best performing parameters for each mapping approach were used (see details in the previous three chapters) to generate MFCC and F0 features from the input EMG signals. The participant has to score the speech quality of an utterance on a continuous slider, which is internally scaled to the interval $[0, 100]$. The listening test was implemented as a browser-based system[KZ14]. Subjects can always go back to previous sentences to change their ratings. Following instructions were presented before the listening test can be started:

- Please listen to the audio files using quality headphones.
- Please rate the quality of the given utterance, ranging from bad to excellent on a continuous scale using the slider.

Additionally, we instructed the participants to arrange the sliders of the presented variations from one utterance to directly compare the five different variations to each other.

We randomly selected ten different evaluation-set utterances from the three different speakers of the EMG-ArraySingle-A-500+ corpus, i.e. each of the ten utterances is synthesized in 5 variations. Thus, a total of 50 utterances are played to the listening test participant. For speakers with multiple sessions in the corpus, we discarded the session with the worst MCD to reduce the amount of utterances the participant has to listen to. We also wanted to cover the inclusion of array-based *and* single-electrode sessions into the listening test files. Thus, utterances from Spk1-Array-Large, Spk1-Array, Spk2-Single and Sp3-Array-Large were used.

In total, 10 German listeners (aged between 19 and 30 years) participated in the test. The participants had no known hearing impairments and no particular speech synthesis expertise. Figure 9.3 and 9.4 depict the results. While Figure 9.3 shows the mean opinion scores of each participant, Figure 9.4 outlines the direct comparison of the four mapping approaches, discarding the resynth-reference. Thus, Figure 9.4 shows the number of times the respecting method was preferred.

The resynthesized speech from the acoustic target data obtains a mean opinion score (MOS) of 61 (max 100), implying a considerable quality loss due to the acoustic feature processing. Investigating the EMG-to-speech preferences, most of the participants perceived the different approaches in the

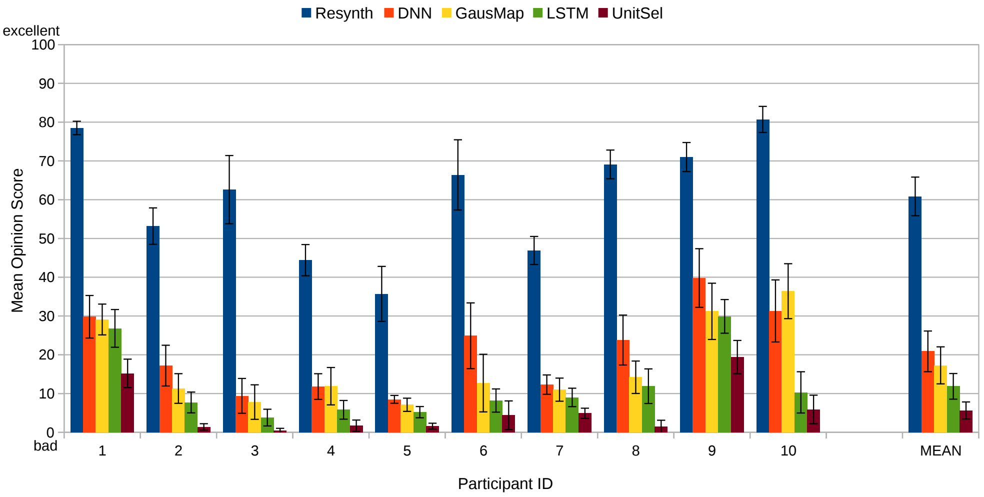


Figure 9.3 – Results from the evaluation listening test comparing the proposed EMG-to-speech mapping approaches: Deep Neural Networks (DNN), Gaussian Mapping (GausMap), Long Short Term Memory networks (LSTM) and Unit Selection (UnitSel). An additional reference utterance (Resynth) is given. 10 participants rated the speech quality from 0 (bad) to 100 (excellent). Error bars represent standard deviation.

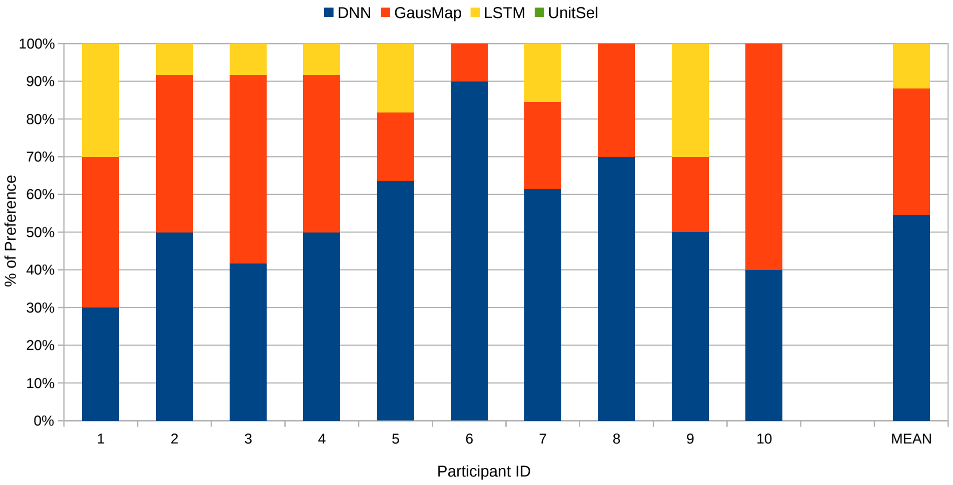


Figure 9.4 – Results from the evaluation listening test. Preference comparison of the proposed EMG-to-speech mapping approaches.

same way: best output quality is obtained with DNNs, followed by Gaussian Mapping, which is succeeded by LSTMs. Unit Selection achieves the worst opinion scores and was never preferred in a single utterance. On average, the participants preferred the DNN-based output in more than 50 % of the cases. A comparison of these results to the ASR evaluation from the previous section reveals a similar outcome: best performance on DNN output, while Unit Selection results are ranked last.

However, there are noticeable variations between listeners judgments. E.g. the resynth-reference output is evaluated with opinion scores between 36 and 81, and also the DNN-based EMG-to-speech mapping obtains opinion scores in the range between 9 and 40. This implies that the acceptance of a distorted speech output is in a large part very user-dependent.

Since the MOS difference between reference and EMG-to-speech is still improvable, we investigate the possible errors on the synthesized signal. Figure 9.5 shows the spectograms from the DNN-based EMG-to-speech output and additionally the resynthesized reference signal from an exemplary utterance taken from Spk-1-Array-Large. Both spectrograms show similar outputs, although two potentials for improvement on the EMG-to-speech output can be spotted: 1.) there are minor artifacts (marked by a black star) that are probably erroneously generated by single frames and 2.) some detailed spectral contours are blurred (e.g. the marked area in the black box). It seems that these kinds of errors are not properly modeled by our proposed EMG-to-speech approach.

To additionally explore the variations between the four investigated sessions, we separate the listening test results into the obtained scores per session. The left part of Figure 9.6 depicts the opinion scores per session. Similar to the previous evaluations, Spk1-Array-Large obtains best results with a mean opinion score of 28 (from 100), while Spk3-Array-Large tends towards lower opinion scores. We additionally compare these findings with the objective Mel Cepstral Distortion evaluation from Section 9.1.1. Therefore, the right part of Figure 9.6 displays a scatter plot of the four sessions and the four EMG-to-speech approaches comparing MCDs with MOS. While lower MCDs imply better performance, mean opinion scores act vice versa. Thus, a high MOS should be associated with a low MCD.

The scatter plot party confirms this hypothesis. For all Unit Selection, as well as all DNN-based outputs, an increased MOS implies a reduced MCD. However, the worst MCD of 6.17 achieves a “mid-range” MOS of 14. Thus, we can state that an objective measure like MCD is adequate for a first quality indication, but being close (or far) to a reference thus not necessarily mean that the output is perceived well (or bad).

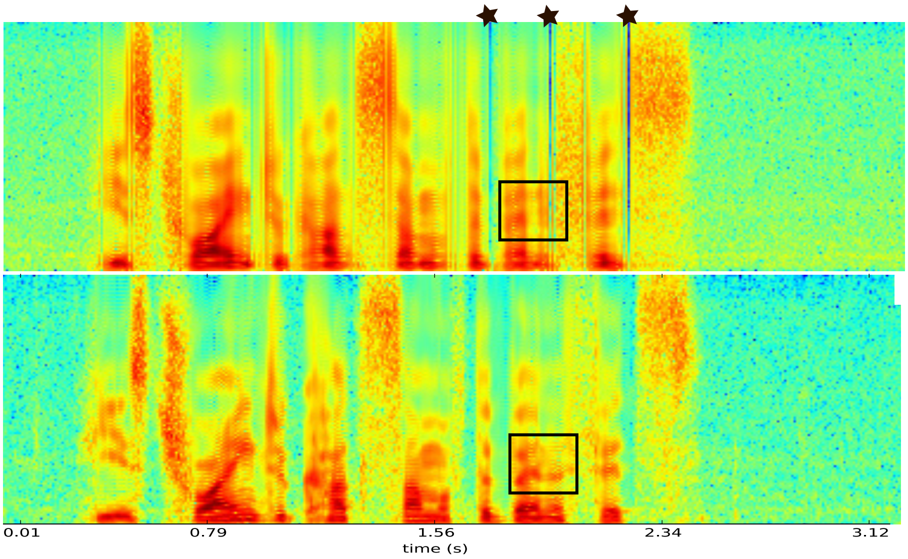


Figure 9.5 – Example spectrograms of DNN-based EMG-to-speech output (top) and resynthesized reference (bottom) of the utterance "He is trying to cut some of the benefits."

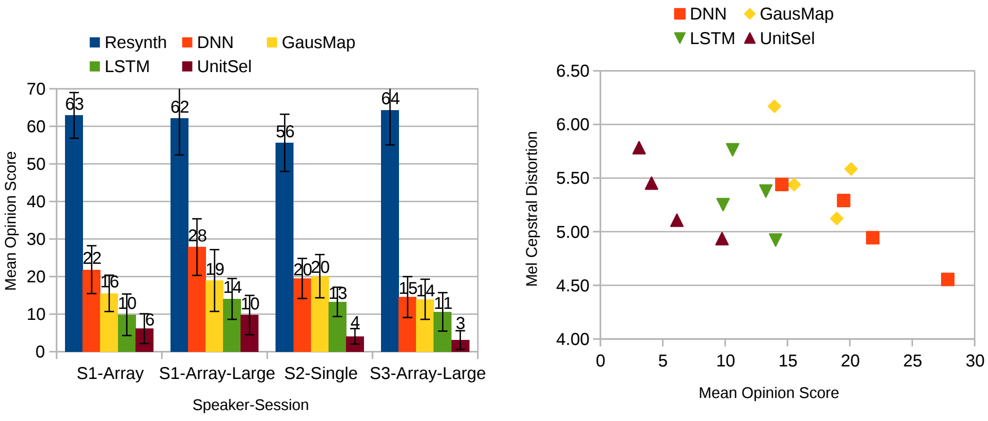


Figure 9.6 – Left side: Results from the evaluation listening test per session. Right-hand side: Comparing opinion scores from the listening test to Mel Cepstral Distortion.

9.3 Summary

In this chapter we directly compared the investigated EMG-to-speech approaches against each other. The feed-forward DNN mapping achieves the best results in multiple evaluations: the best mapping performance in terms of MCD (best session with 4.56), speech recognizer performance and opinion scores in the listening test. Additionally, DNNs are computationally very fast for EMG-to-speech mapping and thus show clear superiority over the other investigated approaches.

Chapter 10

Conclusion and Future Work

This chapter concludes the thesis giving a short summary of the results and the main aspects that were tackled in this thesis. This is concluded by a list of arising research questions and suggestions for potential future work

10.1 Conclusion

This thesis presented a direct transformation from surface electromyographic signals of the facial muscles into acoustic speech. While many silent speech approaches today use a speech recognition component, the proposed direct speech generation has advantages over a recognition-followed-by-synthesis technique: There is no restriction to a given phone-set, vocabulary or even language, while there exists the scope to retain paralinguistic information, like prosody or speaking rate, which is crucial for natural speech communication. The speaker's voice and other speaker-dependent factors keep preserved. Additionally, a direct mapping enables faster processing than speech-to-text followed by text-to-speech and thus the option for nearly real-time computation is given.

The main aspects that were tackled in this thesis are as follows:

- Data collection of simultaneous facial electromyographic and acoustic speech signals. While previous EMG/speech recordings were often limited to 50 utterances, the recordings performed in the scope of this thesis use sessions with up to 2 hours of parallel EMG and speech data.

- Application and comparison of three different state-of-the-art machine learning approaches for the purpose of directly transforming EMG signal to speech: A statistical transformation based on Gaussian mixture models entitled *Gaussian Mapping*, a database-driven approach known as *Unit Selection* and a feature mapping based on Deep Neural Networks. Several sets of parameters for these methods were evaluated on a development data set, achieving MCD scores of up to 4.51. A final round of comparative evaluations, objective as well as subjective, were performed and revealed that the Deep Neural Network approach performs best in several categories. Using a speech recognizer on the synthesized output obtains a word error rate (WER) of up to 19.2 % with a vocabulary size of 2,111 words, which improves down to 3.5 % WER when the vocabulary is reduced to 108 words.
- Transformation of EMG signals recorded during silent articulation into acoustic speech output. While many silent speech researchers are performing on data that is generated during normal articulation, we introduce experiments on silent speech, even for the training of the mapping system where no parallel acoustic data exists. Therefore, we established a dynamic time warping approach that works on silent EMG data and enables an average MCD reduction from 6.81 to 6.35.
- Abolition of the need for label transcriptions (phone-level time alignments). The adoption of Canonical Correlation Analysis performs a feature dimensionality reduction using target acoustic features instead of label information, which would need a recognizer component that is prone for potential decoding errors.
- A realization of Whisper-to-Audible speech mapping with focus on prosody preservation could be achieved, resulting in an F0 generation that performs with a mean voiced/unvoiced accuracy of 80 % and a correlation coefficient of 0.69.
- The feasibility of real-time processing is investigated and evaluated with a real-time factor lower than 0.1 using the proposed feed-forward Deep Neural Network approach.

The performance of EMG-based and other biosignal-based speech interfaces has slowly, but steadily been improved over the last years. While the approach presented in 2009 [TWS09] introduced a mel cepstral distortion of 6.37, the results of this thesis raise the best performance in terms of MCD to 4.51. To compare these MCD numbers to related work, we list the other approaches that are most relevant and report MCD scores. Hueber et al [HBDC11] achieve an MCD of 6.2 on transforming lip and tongue motions, captured by ultrasound and video images into audible speech. The reported average MCD by Toda et al. [TBT04] using Electromagnetic Articulography (EMA) data is 5.59. The same approach is improved by Toda et al. [TBT08], achieving an average MCD of 4.5 with the best speaker-dependent MCD of 4.29. While EMA recordings need specific laboratory-like environments, the EMG-to-speech mapping is clearly more user-friendly.

10.2 Suggestions for Future Work

Although the obtained results outline the ongoing success in EMG-based speech processing, several avenues of research still arise. Among them are:

- Further improvements of the EMG recording hardware. Although the current array and grid-based electrode technology is more user-friendly compared to multiple single-electrodes, the attachment is still cumbersome. Additionally, the recording device is clumsy and immobile and prevents the application of EMG-based speech processing from general usage. Recently introduced epidermal electronics [KLM+11] and small-sized recording devices will definitely achieve user acceptance and applicability. Wei et al. [WRH+14] have already introduced such a device for the application with EMG, which is a next step towards a handy facial EMG-device which again pushes forward EMG-to-speech end-user research and acceptance.

- Further investigations on acoustic feedback realized by real-time/online usage. Since we showed the theoretical applicability of near-real time usage, a logical next step is the investigation of acoustic feedback to the silently articulating speaker. A direct acoustic feedback enables the user to learn, improve and gain a finer control for his or her silent articulation, leading to a system that can be realized in a user-in-the-loop scenario. The user adapts according to the acoustic feedback and additionally the system can directly adapt to the user's utterances.

- Further investigations on speaking mode and speaker differences. The previously mentioned acoustic feedback can help to compensate for speaking-mode differences and may reveal speaker-dependent peculiarities. Additionally, direct acoustic phone-level information can be per-

> mitted from the generated speech output and not from the error-prone phonetic alignments that a recognition system would be involved with.

Yet, no EMG-based speech processing system has been used at an end-user or clinical setting, although one of the most promising applications is the usage on speech-disabled persons. We hope that the findings presented in this thesis have contributed towards achieving this.

Bibliography

[AEB12] Jordi Adell, David Escudero, and Antonio Bonafonte. Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, 54(3):459–476, 2012.

[AGO15] Sandesh Aryal and Ricardo Gutierrez-Osuna. Data driven articulatory synthesis with deep neural networks. *Computer Speech & Language*, pages 1–14, 2015.

[ANSK88] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. Voice Conversion through Vector Quantization. In *Proc. ICASSP*, pages 655–658, 1988.

[ARH14] Farzaneh Ahmadi, Matheus Araújo Ribeiro, and Mark Halaki. Surface Electromyography of Neck Strap Muscles for Estimating the Intended Pitch of a Bionic Voice Source. In *Proc. Biomedical Circuits and Systems Conference (BioCAS)*, pages 37–40, 2014.

[AS70] Bishnu S. Atal and Manfred R. Schroeder. Adaptive Predictive Coding of Speech Signals. *Bell System Technical Journal*, 49(8):1973–1986, 1970.

[BdL85] John V Basmajian and Carlo J de Luca. *Muscles Alive.* Baltimore, MD, 5th editio edition, 1985.

[Bel61] C. G. Bell. Reduction of Speech Spectra by Analysis-by-Synthesis Techniques. *The Journal of the Acoustical Society of America*, 33(12):1725, 1961.

[BHG$^+$14] Florent Bocquelet, Thomas Hueber, Laurent Girin, Pierre Badin, and Blaise Yvert. Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications. *Proc. Interspeech*, pages 2288–2292, 2014.

[BHG$^+$15] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux, and Blaise Yvert. Real-time Control of a DNN-based Articulatory Synthesizer for Silent Speech Conversion: a pilot study. In *Proc. Interspeech*, pages 2405–2409, 2015.

[BNCKG10] Jonathan S Brumberg, Alfonso Nieto-Castanon, Philip R Kennedy, and Frank H Guenther. Brain-computer Interfaces for Speech Communication. *Speech Communication*, 52:367–379, 2010.

[Bol79] Steven F Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, 1979.

[BS95] Anthony J Bell and Terrence I Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[BSM$^+$14] Patricia Balata, Hilton Silva, Kyvia Moraes, Leandro Pernambuco, and Sílvia Moraes. Use of surface electromyography in phonation studies: an integrative review. *International Archives of Otorhinolaryngology*, 17(03):329–339, 2014.

[BT05] Jeffrey C Bos and David W Tack. Speech Input Hardware Investigation for Future Dismounted Soldier Computer Systems. DRCD Toronto CR 2005-064, 2005.

[Car97] Rich Caruana. *Multitask Learning*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1997.

[CEHL01] Adrian D C Chan, Kevin Englehart, Bernard Hudgins, and Dennis Lovely. Myoelectric Signals to Augment Speech Recognition. *Medical and Biological Engineering and Computing*, 39:500–506, 2001.

[CEHL02] Adrian D C Chan, Kevin Englehart, Bernard S Hudgins, and Dennis Lovely. Hidden Markov Model Classification of Myolectric Signals in Speech. *Engineering in Medicine and Biology Magazine, IEEE*, 21(9):143–146, 2002.

[CGL$^+$92] Daniel M Corcos, Gerald L Gottlieb, Mark L Latash, Gil L Almeida, and Gyan C Agarwal. Electromechanical delay: An experimental artifact. *Journal of Electromyography and Kinesiology: Official Journal of the International Society of Electrophysiological Kinesiology*, 2(2):59–68, 1992.

[CK79] P. R. Cavanagh and Paavo V. Komi. Electromechanical delay in human skeletal muscle under concentric and eccentric contractions. *European Journal of Applied Physiology and Occupational Physiology*, 42(3):159–163, 1979.

[Col13] OpenStax College. *Anatomy & Physiology.* Rice University, 2013.

[Com92] Pierre Comon. Independent Component Analysis. *Higher-Order Statistics*, pages 29–38, 1992.

[CYW85] D Childers, B Yegnanarayana, and K Wu. Voice Conversion: Factors Responsible for Quality. In *Proc. ICASSP*, pages 748–751, 1985.

[DCHM12] Yunbin Deng, Glen Colby, James T Heaton, and Geoffrey S Meltzner. Signal processing advances for the MUTE sEMG-based silent speech recognition system. *Military Communications Conference - MILCOM*, pages 1–6, 2012.

[dCK02] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[DEP$^+$06] Helenca Duxans, Daniel Erro, Javier Pérez, Ferran Diego, Antonio Bonafonte, and Asuncion Moreno. Voice Conversion of Non-Aligned Data using Unit Selection. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 237–242, 2006.

[DJS15] Lorenz Diener, Matthias Janke, and Tanja Schultz. Direct Conversion from Facial Myoelectric Signals to Speech using Deep Neural Networks. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2015.

[DKN14] Patrick Doetsch, Michal Kozielski, and Hermann Ney. Fast and robust training of recurrent neural networks for offline handwriting recognition. In *Proc. International Conference on Frontiers in Handwriting Recognition*, pages 279–284, 2014.

[DLR77] Arthur P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the royal statistical society. Series B (methodological)*, (1):1–38, 1977.

[DM04] Arnaud Delorme and Scott Makeig. EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics including Independent Component Analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[DMC14] Winston De Armas, Khondaker a. Mamun, and Tom Chau. Vocal frequency estimation and voicing state prediction with surface EMG pattern recognition. *Speech Communication*, 63-64:15–26, 2014.

[Doe42] E Doehne. Bedeutet Flüstern Stimmruhe oder Stimmschonung? Technical report, 1942.

[DPH+09] Yunbin Deng, Rupal Patel, J.T. Heaton, Glen Colby, L.D. Gilmore, Joao Cabrera, S.H. Roy, C.J.D. Luca, and G.S. Meltzner. Disordered Speech Recognition Using Acoustic and sEMG Signals. In *Proc. Interspeech*, pages 644–647, 2009.

[DRW39] Homer Dudley, R R Riesz, and S S A Watkins. A synthetic speaker. *Journal of the Franklin Institute*, 227(6):739–764, 1939.

[DS04] Bruce Denby and Maureen Stone. Speech Synthesis from Real Time Ultrasound Images of the Tongue. In *Proc. ICASSP*, pages 685–688, 2004.

[DSH+10] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, and James Gilbert. Silent Speech Interfaces. *Speech Communication*, 52(4):270–287, 2010.

[DSLD10] Siyi Deng, Ramesh Srinivasan, Tom Lappas, and Michael D'Zmura. EEG classification of imagined syllable rhythm using Hilbert spectrum methods. *Journal of Neural Engineering*, 7(4):046006, 2010.

[DYK14] Rasmus Dall, Junichi Yamagishi, and Simon King. Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation. In *Proceedings of the 7th International Conference on Speech Prosody (SP2014)*, pages 1012–1016, 2014.

[FAB56] Knud Faaborg-Andersen and Fritz Buchthal. Action potentials from internal laryngeal muscles during phonation. *Nature*, 178:1168–1169, 1956.

[FAE58] Knud Faaborg-Andersen and A Edfeldt. Electromyography of Intrinsic and Extrinsic Laryngeal Muscles During Silent Speech: Correlation with Reading Activity: Preliminary Report. *Acta oto-laryngologica*, 49(1):478–482, 1958.

[FC86] Alan J Fridlund and John T Cacioppo. Guidelines for Human Electromyographic Research. *Psychophysiology*, 23:567–589, 1986.

[FCBD$^+$10] Victoria-M. Florescu, Lise Crevier-Buchman, Bruce Denby, Thomas Hueber, Antonia Colazo-Simon, Claire Pillot-Loiseau, Pierre Roussel, Cédric Gendrot, and Sophie Quattrocchi. Silent vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface. In *Proc. Interspeech*, pages 450–453, 2010.

[FEG$^+$08] Michael J Fagan, Stephen R Ell, James M Gilbert, E. Sarrazin, and P. M. Chapman. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical Engineering and Physics*, 30(4):419–425, 2008.

[Fis36] RA Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[FMBG08] Elia Formisano, Federico De Martino, Milene Bonte, and Rainer Goebel. Who Is Saying What? Brain-Based Decoding of Human Voice and Speech. *Science*, 322(5903):970–973, 2008.

[FQXS14] Yuchen Fan, Yao Qian, Fenglong Xie, and Frank K. Soong. TTS synthesis with bidirectional LSTM based Recurrent Neural Networks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (September):1964–1968, 2014.

[FT92] Toshiaki Fukada and Keiichi Tokuda. An adaptive algorithm for mel-cepstral analysis of speech. *Proc. ICASSP*, pages 137–140, 1992.

[FTD12] Joao Freitas, Antonio Teixeira, and Miguel Sales Dias. Towards a Silent Speech Interface for Portuguese. In *Proc. Biosignals*, pages 91–100, 2012.

[GB75] Leslie Alexander Geddes and Lee Edward Baker. *Principles of Applied Biomedical Instrumentation.* John Wiley & Sons, 1975.

[GGT06] Frank H Guenther, Satrajit S Ghosh, and Jason A Tourville. Neural Modeling and Imaging of the Cortical Interactions underlying Syllable Production. *Brain and Language*, 96:280–301, 2006.

[GHSS72] Thomas Gay, Hajime Hirose, Marshall Strome, and Masayuki Sawashima. Electromyography of the Intrinsic Laryngeal Muscles during Phonation. *Annals of Otology, Rhinology and Laryngology*, 81(3):401–409, 1972.

[Giu] Serban Giuroiu. Cuda K-Means Clustering. *https://github.com/serban/kmeans*.

[GLF$^+$93] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403, 1993.

[GMH13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. *Proc. ICASSP*, (6):6645–6649, 2013.

[GRH$^+$10] James M Gilbert, Sergey I Rybchenko, Robin Hofe, Stephen R Ell, Michael J Fagan, Roger K Moore, and Phil D Green. Isolated Word Recognition of Silent Speech using Magnetic Implants and Sensors. *Medical Engineering and Physics*, 32:1189–1197, 2010.

[GS05] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM networks. *Proceedings of the International Joint Conference on Neural Networks*, 4:2047–2052, 2005.

[GS15] Klaus Greff and Rupesh Srivastava. Brainstorm. *https://github.com/IDSIA/brainstorm*, 2015.

[GSF$^+$95] Ch Gondran, E. Siebert, P. Fabry, E. Novakov, and P. Y. Gumery. Non-polarisable dry electrode based on NASICON ceramic. *Medical & Biological Engineering & Computing*, 33(3):452–457, 1995.

[GSK$^+$15] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutnik, Bas Steunebrink, and Jürgen Schmidhuber. LSTM: A Search Space Odyssey. *arXiv preprint arXiv:1503.04069*, 2015.

[GW96] Mark JF Gales and Philip C Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech & Language*, 10(4):249–264, 1996.

[HAC$^+$07] Thomas Hueber, Guido Aversano, Gerard Chollet, Bruce Denby, Gerard Dreyfus, Yacine Oussar, Philippe Roussel, and Maureen Stone. Eigentongue feature extraction for an ultrasound-based silent speech interface. In *Proc. ICASSP*, pages 1245–1248, 2007.

[HAH01] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken language processing: a guide to theory, algorithm, and system development*. Prentice Hall, 2001.

[Hay60] Keith J. Hayes. Wave Analyses of Tissue Noise and Muscle Action Potentials. *Journal of Applied Physiology*, 15:749–752, 1960.

[HB96] Andrew J Hunt and Alan W Black. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In *Proc. ICASSP*, pages 373–376, 1996.

[HBC$^+$10] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips. *Speech Communication*, 52:288–300, 2010.

[HBDC11] Thomas Hueber, Elie-Laurent Benaroya, Bruce Denby, and Gérard Chollet. Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface. In *Proc. Interspeech*, 2011.

[HCD$^+$07] Thomas Hueber, Gérard Chollet, Bruce Denby, Maureen Stone, and Leila Zouari. Ouisper: Corpus Based Synthesis Driven by Articulatory Data. In *Proceedings of 16th International Congress of Phonetic Sciences*, pages 2193–2196, 2007.

[HHdP$^+$15] Christian Herff, Dominic Heger, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Brain-to-Text: Decoding Spoken Phrases from Phone Representations in the Brain. *Frontiers in Neuroscience*, 9(June):1–11, 2015.

[Hir71] Hajime Hirose. Electromyography of the Articulatory Muscles: Current Instrumentation and Techniques. *Haskins Laboratories Status Report on Speech Research*, SR-25/26:73 – 86, 1971.

[HKSS07] Panikos Heracleous, Tomomi Kaino, Hiroshi Saruwatari, and Kiyohiro Shikano. Unvoiced Speech Recognition Using Tissue-Conductive Acoustic Sensor. *EURASIP Journal on Advances in Signal Processing*, 2007:1–11, 2007.

[HL08] Yi Hu and Philipos C. Loizou. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

[HO92] T. Hasegawa and K. Ohtani. Oral image to voice converter-image input microphone. In *Singapore ICCS/ISITA'92.'Communications on the Move*, pages 617–620, 1992.

[HO00] Aapo Hyvärinen and Erkki Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13:411–430, 2000.

[HOS+10] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. Silent-speech Enhancement using Body-conducted Vocal-tract Resonance Signals. *Speech Communication*, 52:301–313, 2010.

[Hot36] Harold Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377, 1936.

[HPA+10] Dominic Heger, Felix Putze, Christoph Amma, Michael Wand, Igor Plotkin, Thomas Wielatt, and Tanja Schultz. BiosignalsStudio: A flexible framework for biosignal capturing and processing. In *KI 2010: Advances in Artificial Intelligence*, pages 33–39, 2010.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1–32, 1997.

[HSST04] David Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical Correlation Analysis: An Overview with Application to Learning Methods. Technical report, 2004.

[Ima83] Satoshi Imai. Cepstral analysis synthesis on the mel frequency scale. *Acoustics, Speech, and Signal Processing, IEEE*, pages 93–96, 1983.

[Int99] International Phonetic Association. Handbook of the International Phonetic Association. Cambridge University Press, 1999.

[IT06] ITU-T. Itu-T Recommendation P.10. Technical report, 2006.

[Ita75a] Fumitada Itakura. Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57(1):35, 1975.

[Ita75b] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67 – 72, 1975.

[ITI05] Taisuke Ito, Kazuya Takeda, and Fumitada Itakura. Analysis and recognition of whispered speech. *Speech Communication*, 45(2):139–152, 2005.

[Jan10] Matthias Janke. *Spektrale Methoden zur EMG-basierten Erkennung lautloser Sprache*. PhD thesis, Karlsruhe Institute of Technology, 2010.

[JD10] Charles Jorgensen and Sorin Dusan. Speech Interfaces based upon Surface Electromyography. *Speech Communication*, 52(4):354–366, 2010.

[JLA03] Charles Jorgensen, D.D. Lee, and S. Agabont. Sub auditory speech recognition based on EMG signals. In *Proceedings of the International Joint Conference on Neural Networks, 2003*, volume 4, pages 3128–3133, Portland, Oregon, 2003.

[JŠ08] Slobodan T. Jovičić and Zoran Šarić. Acoustic Analysis of Consonants in Whispered Speech. *Journal of Voice*, 22(3):263–274, 2008.

[JSW$^+$06] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel. Towards Continuous Speech Recognition using Surface Electromyography. In *Proc. Interspeech*, pages 573–576, 2006.

[JWH$^+$14] Matthias Janke, Michael Wand, Till Heistermann, Kishore Prahallad, and Tanja Schultz. Fundamental frequency generation for whisper-to-audible speech conversion. In *Proc. ICASSP*, pages 2579–2583, 2014.

[JWS10] Matthias Janke, Michael Wand, and Tanja Schultz. A Spectral Mapping Method for EMG-based Recognition of Silent Speech. *Proc. B-INTERFACE*, pages 2686–2689, 2010.

[KB04] John Kominek and Alan W Black. The CMU Arctic speech databases. *Fifth ISCA Workshop on Speech Synthesis*, pages 223–224, 2004.

[KFS02] Katrin Kirchhoff, Gernot A Fink, and Gerhard Sagerer. Combining Acoustic and Articulatory Feature Information for Robust Speech Recognition. *Speech Communication*, 37(3-4):303–319, 2002.

[KHM05] Anna Karilainen, Stefan Hansen, and Jörg Müller. Dry and Capacitive Electrodes for Long-Term ECG-Monitoring. In *8th Annual Workshop on Semiconductor Advances*, volume November, pages 155–161, 2005.

[Kla82] Dennis H Klatt. The Klattalk text-to-speech conversion system. In *Proc. ICASSP*, pages 1589–1592, 1982.

[Kla87] Dennis H Klatt. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America*, 82(3):737–793, 1987.

[KLM$^+$11] Dae-Hyeong Kim, Nanshu Lu, Rui Ma, Yun-Soung Kim, Rak-Hwan Kim, Shuodao Wang, Jian Wu, Sang Min Won, Hu Tao, Ahmad Islam, Ki Jun Yu, Tae-il Kim, Raeed Chowdhury, Ming Ying, Lizhi Xu, Ming Li, Hyun-Joong Chung, Hohyun Keum, Martin McCormick, Ping Liu, Yong-wei Zhang, Fiorenzo G Omenetto, Yonggang Huang, Todd Coleman, and John A Rogers. Epidermal electronics. *Science*, 333:838–843, aug 2011.

[KM98] Alexander Kain and Michael W Macon. Spectral Voice Conversion for Text-to-speech Synthesis. In *Proc. ICASSP*, pages 285–288, 1998.

[KNG05] Raghunandan S. Kumaran, Karthik Narayanan, and John N. Gowdy. Myoelectric Signals for Multimodal Speech Recognition. In *Proc. Interspeech*, pages 1189–1192, 2005.

[KP77] Diane Kewley-Port. EMG signal processing for speech research. *Haskins Laboratories Status Report on Speech Research*, SR-50:123–146, 1977.

[KP04] Christopher T Kello and David C Plaut. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America*, 116(4 Pt 1):2354–2364, 2004.

[KR05] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, 2005.

[Kra07] Rüdiger Kramme. *Medizintechnik*. Springer-Verlag, 2007.

[KSZ$^+$09] Heather L. Kubert, Cara E. Stepp, Steven M. Zeitels, John E. Gooey, Michael J. Walsh, S. R. Prakash, Robert E. Hillman, and James T. Heaton. Electromyographic control of a hands-free electrolarynx using neck strap muscles. *Journal of Communication Disorders*, 42(3):211–225, 2009.

[Kub93] Robert F Kubichek. Mel-cepstral Distance Measure for Objective Speech Quality Assessment. In *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 125–128, 1993.

[KYHI14] Takatomi Kubo, Masaki Yoshida, Takumu Hattori, and Kazushi Ikeda. Towards excluding redundancy in electrode grid for automatic speech recognition based on surface EMG. *Neurocomputing*, 2014.

[KZ14] Sebastian Kraft and Udo Zölzer. BeaqleJS : HTML5 and JavaScript based Framework for the Subjective Evaluation of Audio Quality. In *Linux Audio Conference (LAC-2014)*, 2014.

[Lee10] Ki-Seung Lee. Prediction of acoustic feature parameters using myoelectric signals. *IEEE transactions on bio-medical engineering*, 57(7):1587–1595, 2010.

[LJ16a] Patrick J. Lynch and C. Carl Jaffe. Lateral Head Anatomy. *https://en.wikipedia.org/wiki/File:Lateral_head_anatomy.jpg*, 2016.

[LJ16b] Patrick J. Lynch and C. Carl Jaffe. Upper Speech Production Apparatus. *https://commons.wikimedia.org/wiki/File:Head_lateral_mouth_anatomy.jpg*, 2016.

[LLM06] Yuet-Ming Lam, Philip Heng-Wai Leong, and Man-Wai Mak. Frame-Based SEMG-to-Speech Conversion. In *IEEE Interna-*

tional Midwest Symposium on Circuits and Systems (MWS-CAS), pages 240–244, 2006.

[LMDL14] Jingjie Li, Ian Vince McLoughlin, Li-Rong Dai, and Zhenhua Ling. Whisper-to-speech conversion using restricted Boltzmann machine arrays. *Electronics Letters*, 50(24):1781–1782, 2014.

[LML+15] Jingjie Li, Ian Vince McLoughlin, Cong Liu, Shaofei Xue, and Si Wei. Multi-task deep neural network acoustic models with model adaptation using discriminative speaker identity for whisper recognition. In *Proc. ICASSP*, pages 4969–4973, 2015.

[Mac05] David J C MacKay. *Information Theory, Inference, and Learning Algorithms.* 2005.

[MBB+07] Larbi Mesbahi, Vincent Barreaud, Olivier Boeffard, De Kerampont, and F-Lannion Cedex. GMM-Based Speech Transformation Systems under Data Reduction. In *ISCA Workshop on Speech Synthesis*, pages 119–124, 2007.

[MBJS96] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. Independent Component Analysis of Electroencephalographic Data. In *Advances in Neural Information Processing Systems 8*, pages 145–151, 1996.

[MHMSW05] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel. Session Independent Non-Audible Speech Recognition Using Surface Electromyography. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 331–336, San Juan, Puerto Rico, 2005.

[MHS03] Hiroyuki Manabe, Akira Hiraiwa, and Toshiaki Sugimura. Unvoiced speech recognition using EMG-mime speech recognition. In *CHI'03 extended abstracts on human factors in computing systems*, pages 794–795, 2003.

[MMS85] Tadashi Masuda, Hisao Miyano, and Tsugutake Sadoyama. A Surface Electrode Array for Detecting Action Potential Trains of Single Motor Units. *Electroencephalography and Clinical Neurophysiology*, 60:435–443, 1985.

[MO86] Michael S. Morse and Edward M. O'Brien. Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes. *Computers in Biology and Medicine*, 16(6):399–410, 1986.

[MP13] Roberto Merletti and Philip Parker. *Electromyography: Physiology, Engineering, and Noninvasive Applications.* John Wiley and Sons, Inc., 2013.

[MSH$^+$08] Geoffrey S Meltzner, Jason Sroka, James T Heaton, L Donald Gilmore, Glen Colby, Serge Roy, Nancy Chen, and Carlo J De Luca. Speech Recognition for Vocalized and Subvocal Modes of Production Using Surface EMG Signals from the Neck and Face. In *Proc. Interspeech*, 2008.

[MWR$^+$94] B. U. Meyer, K. Werhahn, J. C. Rothwell, S. Roericht, and C. Fauth. Functional Organisation of Corticonuclear Pathways to Motoneurones of Lower Facial Muscles in Man. *Experimental Brain Research*, 101(3):465–472, 1994.

[Nak05] Yoshitaka Nakajima. Development and Evaluation of Soft Silicone NAM Microphone. Technical report, IEICE, 2005.

[ND16] National Institute on Deafness (NIDCD) and Other Communication Disorders. Statistics on Voice, Speech, and Language, 2016.

[NKSC03] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. Non-Audible Murmur Recognition. In *Proc. Eurospeech*, 2003.

[NMRY95] M. Narendranath, Hema A Murthy, S. Rajendran, and B. Yegnanarayana. Transformation of Formants for Voice Conversion using Artificial Neural Networks. *Speech Communication*, 16(2):207–216, 1995.

[NTKS06] Mikihiro Nakagiri, Tomoki Toda, Hideki Kashioka, and Kiyohiro Shikano. Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion. In *Proc. Interspeech*, pages 2270–2273, 2006.

[NTSS12] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. *Speech Communication*, 54(1):134–146, 2012.

[Ot2] OT Bioelettronica. *http://www.otbioelettronica.it.*

[Pet84] Eric David Petajan. Automatic Lipreading to Enhance Speech Recognition. In *Proceedings of the IEEE Communication Society Global Telecommunications Conference*, 1984.

[PH10] Sanjay A Patil and John H L Hansen. The Physiological Microphone (PMIC): A Competitive Alternative for Speaker Assessment in Stress Detection and Speaker Verification. *Speech Communication*, 52:327–340, 2010.

[PMB13] Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the Number of Response Regions of Deep Feed Forward Networks with Piece-wise Linear Activations. pages 1–17, 2013.

[PVE$^+$13] Kishore Prahallad, Anandaswarup Vadapalli, Naresh Elluru, Gautam Mantena, Bhargav Pulugundla, Peri Bhaskararao, Hema A Murthy, Simon King, Vasilis Karaiskos, and Alan W Black. The Blizzard Challenge 2013 - Indian Language Tasks. *Blizzard Challenge Workshop*, (1), 2013.

[PWCS09] Anne Porbadnigk, Marek Wester, Jan-P. Calliess, and Tanja Schultz. EEG-based Speech Recognition - Impact of Temporal Effects. In *Proc. Biosignals*, 2009.

[RHK11] Korin Richmond, Phil Hoole, and Simon King. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Proc. Interspeech*, pages 1505–1508, 2011.

[Rot92] Martin Rothenberg. A multichannel electroglottograph. *Journal of Voice*, 6(1):36–43, 1992.

[RVP10] E.V. Raghavendra, P Vijayaditya, and Kishore Prahallad. Speech Synthesis using Artificial Neural Networks. *Communications (NCC), 2010 National Conference on*, pages 5–9, 2010.

[Sag88] Yoshinori Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. *Proc. ICASSP*, pages 679 – 682, 1988.

[SCM95] Yannis Stylianou, Olivier Cappe, and Eric Moulines. Statistical Methods for Voice Quality Transformation. In *Proc. Fourth European Conference on Speech Communication and Technology*, pages 447–450, 1995.

[SCM98] Yannis Stylianou, Olivier Cappe, and Eric Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142, 1998.

[SDM79] Thomas Shipp, E. Thomas Doherty, and Philip Morrissey. Predicting vocal frequency from selected physiologic measures. *The Journal of the Acoustical Society of America 66*, 3:678–684, 1979.

[SHK$^+$14] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958, 2014.

[SHP07] Erik J. Scheme, Bernard Hudgins, and Phillip A. Parker. Myoelectric signal classification for phoneme-based speech recognition. *IEEE transactions on bio-medical engineering*, 54(4):694–9, 2007.

[SHRH09] Cara E Stepp, James T Heaton, Rebecca G Rolland, and Robert E Hillman. Neck and Face Surface Electromyography for Prosthetic Voice Control After Total Laryngectomy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17:146–155, 2009.

[SSB14] Hasim Sak, Andrew Senior, and Franc Beaufays. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. In *Proc. Interspeech*, pages 338–342, 2014.

[ST85] Noboru Sugie and Koichi Tsunoda. A Speech Prosthesis Employing a Speech Synthesizer – Vowel Discrimination from Perioral Muscle Activities and Vowel Production. *IEEE Transactions on Biomedical Engineering*, 32(7):485–490, 1985.

[Ste22] John Q Steward. Electrical Analogue of the Vocal Organs. *Nature*, (110):311–312, 1922.

[SVN37] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[SW10] Tanja Schultz and Michael Wand. Modeling Coarticulation in EMG-based Continuous Speech Recognition. *Speech Communication*, 52(4):341–353, 2010.

[Tay09] Paul Taylor. *Text-to-Speech Synthesis.* Cambridge University Press, Cambridge, 2009.

[TBAO08] Toshio Tsuji, Nan Bu, Jun Arita, and Makoto Ohga. A speech synthesizer using facial EMG signals. *International Journal of Computational Intelligence and Applications*, 7(1):1–15, 2008.

[TBLT08] Viet Anh Tran, Gérard Bailly, Hélène Loevenbruck, and Tomoki Toda. Predicting F0 and voicing from NAM-captured whispered speech. *Proceedings of the 4th International Conference on Speech Prosody*, pages 107–110, 2008.

[TBLT10] Viet Anh Tran, Gerard Bailly, Hélène Loevenbruck, and Tomoki Toda. Improvement to a NAM-captured Whisper-to-Speech System. *Speech Communication*, 52:314–326, 2010.

[TBT04] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Mapping from Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Model for Articulatory Speech Synthesis. In *Proc. of the 5th ISCA Speech Synthesis Workshop*, 2004.

[TBT07] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2222–2235, 2007.

[TBT08] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 2008.

[TM69] Mark Tatham and Katherine Morton. Some electromyography data towards a model of speech production. *Language and speech*, 12:39–53, 1969.

[TMB12] Tomoki Toda, Takashi Muramatsu, and Hideki Banno. Implementation of Computationally Efficient Real-Time Voice Conversion. *Proc. Interspeech*, pages 94–97, 2012.

[TMB13] Tam Tran, Soroosh Mariooryad, and Carlos Busso. Audiovisual Corpus To Analyze Whisper Speech. In *Proc. ICASSP*, pages 8101–8105, 2013.

[TSB$^+$00] Ingo R. Titze, Brad H. Story, Gregory C. Burnett, John F. Holzrichter, Lawrence C. Ng, and Wayne a. Lea. Comparison between electroglottography and electromagnetic glottography. *The Journal of the Acoustical Society of America*, 107(1):581, 2000.

[TWS09] Arthur Toth, Michael Wand, and Tanja Schultz. Synthesizing Speech from Electromyography using Voice Transformation Techniques. In *Proc. Interspeech*, pages 652–655, 2009.

[UMRR12] Benigno Uria, Iain Murray, Steve Renals, and Korin Richmond. Deep Architectures for Articulatory Inversion. In *Proc. Interspeech*, pages 867–870, 2012.

[UNGH00] Naonori Ueda, Ryohei Nakano, Zoubin Ghahramani, and Geoffrey E. Hinton. Split and merge EM algorithm for improving Gaussian mixture density estimates. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, (1):133–140, 2000.

[Vit67] Andrew J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[Von91] Wolfgang Von Kempelen. *Mechanismus der menschlichen Sprache*. 1791.

[VPH$^+$93] A Villringer, J Planck, C Hock, L Schleinkofer, and U Dirnagl. Near infrared spectroscopy (NIRS): A new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience Letters*, 154:101–104, 1993.

[WB15] Felix Weninger and Johannes Bergmann. Introducing CURRENNT - the Munich Open-Source CUDA RecurREnt Neural Network Toolkit. *Journal of Machine Learning Research*, 16:547–551, 2015.

[Wer90] Paul J. Werbos. Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[Wik16] Wikipedia. Throat anatomy diagram. *https://en.wikipedia.org/wiki/File:Throat_anatomy_diagram.svg*, 2016.

[WKJ$^+$06] Matthias Walliczek, Florian Kraft, Szu-Chen Jou, Tanja Schultz, and Alex Waibel. Sub-Word Unit Based Non-Audible Speech Recognition Using Surface Electromyography. In *Proc. Interspeech*, pages 1487–1490, 2006.

[WRH$^+$14] Pinghung Wei, Milan Raj, Yung-yu Hsu, Briana Morey, Paolo Depetrillo, Bryan Mcgrane, Monica Lin, Bryan Keen, Cole Papakyrikos, Jared Lowe, and Roozbeh Ghaffari. A Stretchable and Flexible System for Skin - Mounted Measurement of Motion Tracking and Physiological Signals. In *36th Annual International Conference of the IEEE Engineering in Medicince and Biology Society*, pages 5772–5775, 2014.

[WSJS13] Michael Wand, Christopher Schulte, Matthias Janke, and Tanja Schultz. Array-based Electromyographic Silent Speech Interface. In *Proc. Biosignals*, 2013.

[WVK$^+$13] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li. Exemplar-Based Unit Selection for Voice Conversion Utilizing Temporal Information. In *Proc. Interspeech*, pages 950–954, 2013.

[WW83] Bruce B Winter and John G Webster. Driven-Right-Leg Circuit design. *IEEE Transactions on Bio-Medical Engineering*, 30(1):62–66, 1983.

[WZW$^+$13] Martin Wöllmer, Zixing Zhang, Felix Weninger, Björn Schuller, and Gerhard Rigoll. Feature Enhancement by Bidirectional LSTM Networks for Conversational Speech Recognition in Highly Non-Stationary Noise. *Proc. ICASSP*, pages 6822–6826, 2013.

[YES$^+$14] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Huaming Wang, Jasha Droppo, Geoffrey Zweig, Chris Rossbach, Jon Currey, Jie Gao, Avner May, Baolin Peng, Andreas Stolcke, and Malcolm Slaney. An Introduction to Computational Networks and the Computational Network Toolkit. Technical Report MSR-TR-2014-112, 2014.

[YLN15] Shing Yu, Tan Lee, and Manwa L. Ng. Surface Electromyographic Activity of Extrinsic Laryngeal Muscles in Cantonese Tone Production. *Journal of Signal Processing Systems*, 2015.

[ZJEH09] Quan Zhou, Ning Jiang, Kevin Englehart, and Bernard Hudgins. Improved Phoneme-based Myoelectric Speech Recognition. *IEEE Transactions on Biomedical Engineering*, 56:8, 2009.

[ZRM$^+$13] Matthew D. Zeiler, Marc'Aurelio Ranzato, Rajat Monga, Min Mao, Kun Yang, Quoc Viet Le, Patrick Nguyen, Andrew Senior, V. Vanhoucke, J. Dean, and Geoffrey E. Hinton. On Rectified Linear Units for Speech Processing. In *Proc. ICASSP*, pages 3517–3521, 2013.