Simon Stone

A Silent-Speech Interface using Electro-Optical Stomatography

TUDpress

Studientexte zur Sprachkommunikation
Hg. von Rüdiger Hoffmann
ISSN 0940-6832
Bd. 102

Simon Stone

A Silent-Speech Interface using Electro-Optical Stomatography

TUD*press*

2021

Supplemental Materials can be downloaded using the following code

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

ISBN 978-3-95908-457-4

D-01309 Dresden
Tel.: +49 351 4721463
http://www.tudpress.de

TUDpress ist ein Imprint von Thelem

Gesetzt von den Herausgebern.
Printed in Germany.

Technische Universität Dresden

A Silent-Speech Interface using Electro-Optical Stomatography

Dipl.-Ing.

Simon Stone

Von der Fakultät Elektrotechnik und Informationstechnik der Technischen Universität Dresden

zur Erlangung des akademischen Grades

Doktoringenieur

(Dr.-Ing.)

genehmigte Dissertation

Vorsitzender:	Prof. Dr.-Ing. habil. Hagen Malberg (TU Dresden)
1. Gutachter:	Prof. Dr.-Ing. Peter Birkholz (TU Dresden)
2. Gutachter:	Prof. Dr. rer. nat. habil. Gerhard Weber (TU Dresden)
3. Gutachter:	Prof. Pascal Perrier, PhD (Université Grenoble Alpes/Grenoble INP)

Tag der Einreichung: 22.10.2020
Tag der Verteidigung: 27.09.2021

Statement of authorship

I hereby certify that I have authored this document entitled *A Silent-Speech Interface using Electro-Optical Stomatography* independently and without undue assistance from third parties. No other than the resources and references indicated in this document have been used. I have marked both literal and accordingly adopted quotations as such. During the preparation of this document I was only supported by the following persons:

Prof. Dr.-Ing. Peter Birkholz

Additional persons were not involved in the intellectual preparation of the present document. I am aware that violations of this declaration may lead to subsequent withdrawal of the academic degree.

Dresden, 22nd October 2020

Simon Stone

TECHNISCHE
UNIVERSITÄT
DRESDEN

Faculty of Electrical and Computer Engineering Institute of Acoustics and Speech Communication

Chair of Speech Technology and Cognitive Systems

Abstract

Speech technology is a major and growing industry that enriches the lives of technologically-minded people in a number of ways. Many potential users are, however, excluded: Namely, all speakers who cannot easily or even at all produce speech. Silent-Speech Interfaces offer a way to communicate with a machine by a convenient speech recognition interface without the need for acoustic speech. They also can potentially provide a full replacement voice by synthesizing the intended utterances that are only silently articulated by the user. To that end, the speech movements need to be captured and mapped to either text or acoustic speech. This dissertation proposes a new Silent-Speech Interface based on a newly developed measurement technology called Electro-Optical Stomatography and a novel parametric vocal tract model to facilitate real-time speech synthesis based on the measured data. The hardware was used to conduct command word recognition studies reaching state-of-the-art intra- and inter-individual performance. Furthermore, a study on using the hardware to control the vocal tract model in a direct articulation-to-speech synthesis loop was also completed. While the intelligibility of synthesized vowels was high, the intelligibility of consonants and connected speech was quite poor. Promising ways to improve the system are discussed in the outlook.

Zusammenfassung

Sprachtechnologie ist eine große und wachsende Industrie, die das Leben von technologieinteressierten Nutzern auf zahlreichen Wegen bereichert. Viele potenzielle Nutzer werden jedoch ausgeschlossen: Nämlich alle Sprecher, die nur schwer oder sogar gar nicht Sprache produzieren können. Silent-Speech Interfaces bieten einen Weg, mit Maschinen durch ein bequemes sprachgesteuertes Interface zu kommunizieren ohne dafür akustische Sprache zu benötigen. Sie können außerdem prinzipiell eine Ersatzstimme stellen, indem sie die intendierten Äußerungen, die der Nutzer nur still artikuliert, künstlich synthetisieren. Diese Dissertation stellt ein neues Silent-Speech Interface vor, das auf einem neu entwickelten Messsystem namens Elektro-Optischer Stomatografie und einem neuartigen parametrischen Vokaltraktmodell basiert, das die Echtzeitsynthese von Sprache basierend auf den gemessenen Daten ermöglicht. Mit der Hardware wurden Studien zur Einzelworterkennung durchgeführt, die den Stand der Technik in der intra- und inter-individuellen Genauigkeit erreichten und übertrafen. Darüber hinaus wurde eine Studie abgeschlossen, in der die Hardware zur Steuerung des Vokaltraktmodells in einer direkten Artikulation-zu-Sprache-Synthese verwendet wurde. Während die Verständlichkeit der Synthese von Vokalen sehr hoch eingeschätzt wurde, ist die Verständlichkeit von Konsonanten und kontinuierlicher Sprache sehr schlecht. Vielversprechende Möglichkeiten zur Verbesserung des Systems werden im Ausblick diskutiert.

Contents

List of Figures

List of Tables

Acronyms

AC Alternating Current

ADC Analog-to-Digital Converter

ASR Automatic Speech Recognition

ATS Articulation-to-Speech

ATT Articulation-to-Text

BCI Brain-Computer Interface

BLSTM Bidirectional Long Short-Term Memory

C consonant

CNN Convolutional Neural Network

CV consonant-vowel

DAC Digital-to-Analog Converter

DMA Direct Memory Access

DNN Deep Neural Network

DOF Degree Of Freedom

DTW Dynamic Time Warping

ECoG Electrocorticography

EEG Electroencephalography

EMA Electromagnetic Articulography

EMC Electromagnetic Compatibility

EMG Electromyography

EOS Electro-Optical Stomatography

EPG Electropalatography

fMRI functional Magnetic Resonance Imaging

fNIRS functional Near-Infrared Spectroscopy

FSM finite state machine

GPR Gaussian Process Regression

GUI Graphical User Interface

HF High-Frequency Radio Waves

HMM Hidden Markov Model

IC Intergrated Circuit

IPA International Phonetic Association

IPA International Phonetic Alphabet

KRR Kernel Ridge Regression

LED Light Emitting Diode

LPC Linear Predictive Coding

LSTM Long Short-Term Memory

MCU Microcontroller Unit

MEG Magnetoencephalography

MLP Multi-Layer Perceptron

MRI Magnetic Resonance Imaging

OPG Optopalatography

PCB printed circuit board

PDF probability density function

PMA Permanent-Magnetic Articulography

RMSE Root-Mean-Square Error

RNN Recurrent Neural Network

SAMPA Speech Assessment Methods Phonetic Alphabet

SD standard deviation

sEMG Surface Electromyography

SNR Signal-to-Noise Ratio

SPI Serial Peripheral Interface

SSI Silent-Speech Interface

SVM Support Vector Machine

TAM Target Approximation Model

TTS Text-to-Speech

UART universal asynchronous receiver-transmitter

US Ultrasonography

V vowel

VAD Voice Activity Detection

VCSEL vertical-cavity surface-emitting laser

VOT Voice Onset Time

WIG Wearable Intonation Generator

1. Introduction

Listen to the silence. It has so much to say.

(Rumi)

The ability to produce, perceive, and understand speech is arguably the most important human skill. As part of humanity's on-going efforts to create machines ever more similar to itself, attempts to develop a technology to mimic the human speech processing capability were only a matter of time.

In the 20th century, the field of Automatic Speech Recognition (ASR) summarized these attempts and grew into its own scientific discipline. The earliest recognized speech recognition system was "Audrey", introduced in [1], that came out of the legendary Bell Laboratories in 1952 (for more information on that institutions stunning portfolio of inventions and discoveries, see [2]). This groundbreaking, fully analog system was able to recognize the spoken digits from 0 to 9 with a reported accuracy of 97 to 99 %. In the following two decades, some first successes were achieved: William C. Dersch's "Shoebox" system, for example, was presented at the 1962 World's Fair in Seattle [3]. Shoebox extended Audrey's vocabulary by six command words (including "plus", "minus" and "total") to perform simple arithmetic operations entirely based on spoken input. The scientific community, however, also saw some concepts emerge that would stay central to the research efforts in the field of ASR. The "Phonetic Typewriter" [4], a phoneme recognizer developed at the Kyoto University, already tackled the difficult task of continuous speech recognition (as opposed to the isolated command word recognition task other systems of the time focused on). At the University College London, Denes [5] imposed phonotactic constraints by allowing only certain phoneme sequences and thus introduced statistical syntax as another tool to the community. The pace quickened after Vintsyuk [6] proposed dynamic programming to help with the difficult non-linear time alignment of a reference and a sample utterance. This technique was most prominently featured in the Viterbi algorithm [7], which became the de-facto standard for time-alignment (or Dynamic Time Warping (DTW)) more than ten years later, after it crossed over into speech research from the field of information theory, popularized by [8].

Since then, the performance and availability of computer systems rapidly increased and alongside these developments, numerous breakthroughs in ASR research were achieved: Hidden Markov Models and stochastic language models greatly improved the performance of continuous speech recognition systems in the 1980s (e.g., [9, 10]), the vocabularies of the systems grew quickly in the 90s, when statistical learning entered the field, and moved beyond the task of *recognition* towards truly *understanding* speech and even entering a dialog with the user in the 2000s. For a more detailed look at the history of speech recognition, see the review by Juang [11] (which was also the basis of this short introduction) or, for an even more in-depth retrospective, the book by Pieraccini [12].

Today, ASR systems are ubiquitous, used not only as dictation systems on office computers but also in cars, service hotlines, televisions, smart speakers, and many more. We even have voice-enabled personal assistants on smartphones (e.g., Apple's *Siri*, Google's *Google Now*, or Amazon's *Alexa*) that attempt to engage with the user in a way that is supposed to mimic a human interlocutor. The market for speech technology is enormous and still booming (see Figure 1.1).

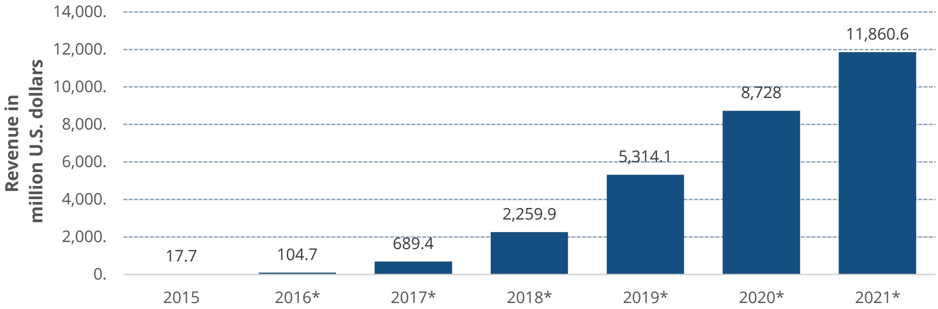


Figure 1.1.: Size of the speech recognition market worldwide from 2015 to 2024. The asterisk (*) denotes projected years. Data according to [13].

However, there is one major problem with the current-day ASR systems: it excludes a significant part of the population. Some people cannot talk to machines, either because of the circumstances (e.g., the loud and noisy environment of a jet plane, the obstructions caused by the breathing masks of fire fighters or divers) or because of physical limitations (e.g., the elderly, laryngectomized cancer patients or intensive care patients with a tracheostoma). Especially the latter demographic is completely shunned by the global innovation drivers in the sector of consumer speech technology (i.e., Google, Amazon, and Apple), despite the fact that they together make up a sizable chunk of the market: According to projections by the United Nations[1], the median age in Germany will be 49 years by the year 2019. While future generations of elderly will be used to the convenience and productivity of speech technology, the physical effort to produce speech makes it inreasingly difficult with age to continue using consumer devices in the same way they used to. But why is it even necessary to talk to the machine? Why does sound need to travel through the air to the machine's microphone, only to be decoded into the actual signal of interest: the speech sound identities (and subsequently the linguistic and semantic content of the speech sound sequence)?

While this is of course merely a matter of convenience and quality-of-life, laryngectomized people have far more pressing concerns regarding speech technology. Given that this demographic is not just of substantial size (five-year prevalence of 488 900 wordwide[2]), but also growing steadily (177 422 new cases worldwide in 2018[3]). In Germany in the year 2018 alone, more than 4800 patients have suffered loss or severe impairment of their voice due to a complete or partial laryngectomy[4].

A few therapies and prostheses are commonly used to rehabilitate the patients' ability in clinical practice, but all of them have their individual drawbacks. There are currently three major kinds of techniques in use [14]: the electrolarynx, esophageal speech (more of a replacement voice than a prostheses), and the so-called tracheoesophageal speech.

The electrolarynx [15] is a hand-held device that is usually pressed against the skin roughly at the height of the (now removed) vocal cords. The device then sends vibrations (usually at a fixed frequency) through the neck into the pharynx, where these vibrations turn into sound pressure

[1] `https://population.un.org/wpp/`

[2] `https://de.statista.com/statistik/daten/studie/1095977/umfrage/zahl-der-weltweiten-krebsfaelle-nach-krebsart/` [In German]

[3] `https://de.statista.com/statistik/daten/studie/286545/umfrage/zahl-der-krebsneuerkrankungen-weltweit/` [In German]

[4] Fallpauschalenbezogene Krankenhausstatistik (DRG-Statistik): Operationen und Prozeduren der vollstationären Patientinnen und und Patienten in Krankenhäusern. Online: `www.gbe-bund.de`[In German]

waves and excite the vocal tract (for more on the speech production process see chapter 2). In the roughly 100 years since the introduction of the first such device in the late 1920s, this basic principle has remained the same and very little improvements of the sound quality and variation of the fundamental frequency have been made [15], with very few notable exceptions (e.g., [16]). Electrolaryngeal speech can be described as robotic, artificial, difficult to understand, and generally unnatural sounding (for an example, please visit `https://www.youtube.com/watch?v=Kmk46U2yjow` [Last visited on September 9, 2020]. Still, it is a widely used technique, probably because it requires very little training (at least in its most basic form).

Esophageal speech avoids any kind of technology, because it re-purposes existing mucosa flaps at the upper end of the esophagus as a *pseudoglottis*: By swallowing air and then expelling it through the esophagus, these flaps can be excited to oscillate, similar to the way that air from the trachea excites the actual vocal folds in a non-laryngectomized speaker (see section 2.1). This manner of speaking is difficult to learn and, even when mastered, usually has a distinct "belching" sound quality to it (visit `https://www.youtube.com/watch?v=UTLg-2N4hyw` for an example of a very capable esophageal speaker [Last visited on September 9, 2020]). Esophageal speech is therefore also sometimes called ructus voice (ructus from Latin *ructare* - belch). Futhermore, many speakers never learn to properly communicate in this way. Exact numbers are unreliable here because these statistics are usually not recorded, but the voice prostheses manufacturer Atos Medical claims that only 20 % of those who try to learn esophageal speech actually succeed [5].

Finally, today's preferred method to rehabilitate laryngectomized patients is tracheoesophageal speech using an artificial valve [14]. These valves are placed into a fistula, a surgically made connection, between the trachea and the esophagus. If not speaking, the valve blocks airflow into the esophagus and air is exhaled from the lungs through the tracheostoma, a hole in the patient's neck[6]. When the patient wants to speak, they can cover the tracheostoma and exhale, thus creating a positive pressure on the valve and forcing it open. The air then escapes into the pharynx, where it is used to excite a pseudoglottis, similar to esophageal speech. In contrast to that, however, the fact that the air does not need to be swallowed and is instead simply exhaled, makes it much more convenient and easier to speak in this way. The resulting high success rate (95 % in long-term users [14]) has helped this technique, which is also called a *voice prostheses*, claim its place as the state-of-the-art in voice rehabilitation after total laryngectomy. But it is not without substantial disadvantages: Laryngectomized patients are often elderly patients and as such have the same difficulties as non-laryngectomized speakers regarding the effort of speech production. The surgery to create the fistula is also not without risks and can result in harmful punctuations of the trachea and/or esophagus. But the main disadvantage of this method is the dependency of the patients on constant clinical and surgical care, because the valves must be regularly checked and replaced to avoid clogging, inflammations, scarring, and other complications. This greatly limits the patients' mobility and self-determined living and may even result in health hazards if patients' miss their checkup appointments.

The state of the art in speech prostheses therefore raises some questions: If it is so difficult to create a new *internal* voice source, why not try to create an *external* voice? So instead of bringing the excitation source *into* the vocal tract, take the articulation *out* of the vocal tract and produce the speech extra-orally?

Producing speech with technology has always fascinated researchers and records of attempts to build speech producing machines go back to the 18th century and the days of Christian Gottlieb Kratzenstein [17], who built a set of acoustic resonators that produced vowel sounds, and Wolfgang Von Kempelen [18], who developed a machine that was even capable to produce short utterances (for more details and examples of historic speech analysis and synthesis systems and devices see [19]). In the second half of the 20th century, three major branches of synthesis systems emerged: articulatory synthesizers (e.g., the Kelly-Lochbaum model [20]) that simulate the

[5] `https://www.atosmedical.us/support/esophageal-speech/`[Last visit: September 9, 2020]

[6] The tracheostoma is not made specifically for this voice prosthesis, but is necessary for all laryngectomized patients because the larynx also protects the trachea from contamination by food or saliva. When the larynx is removed, the connection between the trachea and the pharynx therefore needs to be blocked and the tracheostoma is made to create an airway for breathing.

propagation of sound waves through the human vocal tract for speech production, formant synthesizers (e.g., Klatt's well-known Klattalk system [21]) that use the source-filter model of speech production ([22, 23]), and systems based on concatenation of very short pre-recorded speech segments (e.g., [24]). Today, artificial neural networks working in the cloud directly map written letters to acoustic waveforms in end-to-end systems (e.g, WaveNet [25] and Tacotron [26]) and allow high-quality speech synthesis in portable, miniature devices (as long as they have a fast and stable connection to the internet).

So with a long history of speech synthesis research and a wide range of systems available, connecting a voice-less (or voice-impaired) user to such a system in some way seems like an obvious way of restoring their ability to communicate. Especially since the users described above usually retain their ability to still *articulate* speech, i.e. silently mouthing the intended words, this leads to the fundamental ideas underlying a technology called Silent-Speech Interfaces: What if we could (a) remove the acoustic stage from a speech recognition system and use the speech *movements* as the input, or (b) use the speech movements to control some kind of technological speech generator?

This dissertation presents the development of one incarnation of such a Silent-Speech Interface, using a newly developed measurement technique to capture the speech movements, state-of-the-art algorithms for a silent speech recognition system, and a novel vocal tract model to generate speech based on the measured movements.

1.1. The concept of a Silent-Speech Interface

A Silent-Speech Interface (SSI) is a technologically enabled channel of communication between a human and a machine that uses speech to encode the information but does not require any audible, acoustic speech. There are two basic paradigms for SSIs: Articulation-to-Text (ATT) and Articulation-to-Speech (ATS). An ATS system can also incorporate an ATT frontend, which translates the articulatory data to text as an intermediary representation that is then used with a standard Text-to-Speech (TTS) system to generate speech. These systems can possibly exploit text-based linguistic models to regularize the mapping from articulation to speech, but are limited to the pre-defined vocabulary and thus the language they were trained with. An ATS system without a textual intermediary cannot use text-level linguistic models but can, in theory, generate all speech (and even non-speech) sounds by learning the direct mapping from articulation. Such systems are therefore also called *direct* ATS systems.

The general framework of an SSI consists of three components: an articulatory data acquisition frontend using some kind of sensor technology, a recognition (in ATT systems) or parametric synthesis (in ATS systems) backend, and a mapping between the articulatory data and the vocabulary (ATT) or the parameters of the synthesis (ATS) (see Figure 1.2). Due to the unstandardized interfaces between the components, research around SSIs usually involves the entire pipeline, with each research group setting up their own framework. Some efforts have been made to uncouple research into each component, e.g., by publishing datasets of articulatory data for the specific purpose of allowing other researchers to focus on the mapping. But because of the heterogeneous input modalities across the various technologies, no unified framework or defined interfaces between components have been established in the field, making every SSI a stand-alone solution, which usually needs to be developed "from scratch" every time.

1.2. Structure of this work

To understand the requirements and challenges of SSI development, an at least basic understanding of the speech production processes is necessary. Chapter 2 therefore introduces the fundamentals of phonetics to the reader, limited to and focused on everything directly related to the subjects of this dissertation. As described in section 1.1, developing the synthesis or recognition backend of an SSI usually goes hand in hand with the development of the articulatory data acqui-

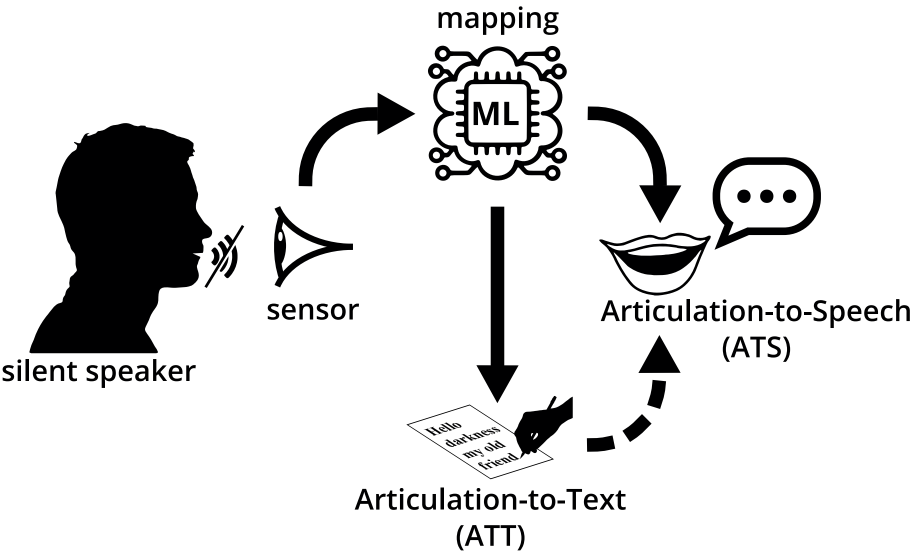


Figure 1.2.: General framework of a Silent-Speech Interface. In the ATT paradigm, the mapping is from the sensor data to a word label (classification). In the ATS paradigm, the mapping is from the sensor data to a set of synthesis parameters. By using a regular TTS synthesizer, an ATT system can be extended to an ATS system.

sition frontend. The literature review in chapter 3 therefore covers the state-of-the-art and the history of both algorithms and measurement technologies in the field of SSI research. After this overview, a newly developed articulometric technology called Electro-Optical Stomatography (EOS) is presented in chapter 4 that aims to overcome the shortcomings and limitations of the previously existing techniques. In chapter 5, EOS is used to develop and evaluate two command word recognition systems in an ATT paradigm. Chapter 6 presents a study on using EOS in an ATS system. To that end, a newly developed vocal tract model well-suited for real-time articulatory speech synthesis is also presented therein. Two additional experiments on the generation of pitch and voicing information in an ATS system close out the chapter. Finally, chapter 7 summarizes the findings and contributions of this dissertation and presents an outlook on future work towards a fully-developed SSI based on EOS.

2. Fundamentals of phonetics

In order to understand the requirements and challenges of articulatory measurements, it is important to understand how humans produce speech and how speech is structured from an articulatory perspective. The field of phonetics, or more specifically articulatory phonetics, concerns itself with the systematic analysis and description of exactly these characteristics of speech, has a long and rich tradition, and is an ongoing, fertile field of research. Within the scope of this dissertation, only the fundamentals of speech production are of immediate interest. To that end, I will discuss the *speech organs* involved in the process (section 2.1), provide a basic breakdown of the various *sounds* making up speech (section 2.2 and section 2.3), take a brief look at the *acoustic properties* of speech sounds (section 2.4), and introduce a somewhat advanced concept called *coarticulation*, which goes beyond the basics of phonetics but has an immediate bearing on Silent-Speech Interface (SSI)-related matters (section 2.5). Finally, the summary of these concepts in section 2.7 further focuses on the presented aspects of speech and articulation most relevant in the context of SSIs. The information presented in this chapter is based on [27], except where stated otherwise. The languages of the world are a very diverse domain and it is not helpful (nor even possible) to describe the entire state of the art in phonetics in the context of this dissertation. Instead, only the sounds most relevant to English and German are the major focus of this chapter because of the global importance of the former and the latter's use in the experiments of this dissertation. Even within these two languages, there are numerous dialectal variants and accents that not only use the same sounds in a different way but also use entirely different sounds. To avoid confusion, the terms English and German are regarded as synonomous with General American English and Standard German, respectively. All schematic articulations in this chapter are reproduced from [28] and slightly modified for clarity.

2.1. Components of the human speech production system

Speech sounds are produced by the time-varying interplay of three functional components (see Figure 2.1): initiation of the airflow from the lungs, modulation of this airflow (phonation) to generate an acoustic excitation, and a "tube" formed by the upper airways and shaped by body parts called articulators that functions as a resonator and/or an aerodynamic tube system (similar to the body of a trumpet or trombone). The airflow from the lungs, funneled through the trachea, passes through the larynx (also known as the "voice box"). Inside the larynx, small pieces of layered soft tissue are stretched across the trachea. When at rest, they form a V-shape pointing towards the front. These *vocal folds* (also sometimes imprecisely called vocal cords) typically have a length of 1.75 cm to 2.5 cm in males and 1.25 cm to 1.75 cm in females [29] and the area between them is called the *glottis*. The vocal folds are kept wide apart (abducted) in a neutral state so that airflow can pass unhindered through the open glottis in both directions during breathing. The muscles that are part of the vocal folds can also completely shut them (keeping them adducted), which happens, e.g., in the initial phase of coughing to build up pressure below the vocal folds. During speech, the vocal folds are slightly less abducted for some sounds and are narrowed for others (see section 2.2 and section 2.3). If the vocal folds are narrowed below a certain critical distance while air is flowing through the gap in between, they start to vibrate and thus produce a complex, wideband sound (similar to a vibrating reed in a woodwind instrument's mouthpiece). This flow-induced oscillation is a complex and multi-faceted process and its analysis and modeling is subject of ongoing research [30, 31]. For the purposes of this overview, it shall suffice to say that due to the airflow from the lungs the pressure below the narrowly constricted (or even closed) glottis builds up until the pressure differential across the vocal folds becomes too large and they are blown open again. The rapid airflow through the glottis resumes and, thanks to the Bernoulli effect, the vocal folds are drawn back together by the created suction, and another cycle of sub-glottal pressure rise, opening burst, and closing suction begins. This oscillation continues as long as the airflow from the lungs is kept up (and sufficiently fast for the Bernoulli effect to occur) and the distance between the vocal folds is small enough. This vibration is called the *voiced excitation* of the vocal tract (i.e., the system of cavities above the glottis consisting of the *pharynx*, the *nasal cavity*, and the *mouth*). Conversely, if the vocal folds do not oscillate but air still flows through the glottis, it is called the *voiceless excitation*. A mixed excitation, where only some part of the vocal folds oscillates and/or the glottis is permanently open to some extent, is also not just possible but actually quite common. However, since only the simplified binary voicing is used to group speech sounds, mixed excitation, the various voice qualities, and other idiosyncracies of the glottal excitation shall not be discussed here and the reader is instead directed to [29] for further research. Similarly, other types of flow than the egressive pulmonic airstream (exhaled air from the lungs), e.g., those occurring in the ejectives or clicks of African languages, are ignored for the purposes of this dissertation.

The (voiced or voiceless) excitation signal is modified by the vocal tract before it results in speech sounds. This modification depends on the geometry of the vocal tract, which can be shaped by means of the *articulators*. Articulators are a set of body parts and anatomical landmarks, which in combination can create the speech sounds of all languages from the two basic excitation signals. There are two basic kinds of articulators: active articulators that can be (voluntarily or involuntarily) moved by the speaker (the vocal folds, the larynx, the tongue, the soft palate, the lower jaw, and the lips), and the passive articulators, which usually remain still in Western languages (the pharynx wall, the hard palate, the alveolar ridge, and the upper teeth). Based on the shape created by the articulators, the vocal tract acts in two different but not necessarily mutually exclusive ways: If the vocal tract is mostly open, i.e. there are no narrow constrictions and it is essentially a tube through which air can flow, it functions as an acoustic resonator with a distinct set of resonance frequencies that is defined by the geometrical shape of the complex tube. If there are one or more narrow constrictions (e.g., less than 20 mm^2) anywhere in the vocal tract, they can cause aerodynamic turbulences downstream that create noisy sound sources. Speech sounds, especially in running speech, are usually created through a combination of these two cases, although they are often grouped by the dominant of the two conditions of *open* versus *constricted* or even closed vocal tract. Sounds produced with an open vocal tract are called *vowel* sounds, while sounds produced with a constricted or closed vocal tract are called *consonants*. Besides this distinction, vowels and consonants can also be grouped into two subsets called *sonorants* (which are produced with a non-turbulent airflow in the vocal tract) and *obstruents* (which are produced with some sort of turbulence-causing obstruction of the airflow). Before we can discuss speech sounds, we need an unambiguous way of transcribing them. The orthographic spelling conventions of different languages make it difficult to map letters to sounds in a general, language-independent way. And even within a language, the same letter is used for very different sounds: for example the letter *i* denotes very different sounds in the English pronoun *I* and in the preposition *in*. Sometimes there are also more sounds in a language than are actually used to discriminate words and thus the letters of the alphabet may not be enough. Therefore, the International Phonetic Association (IPA) developed an alphabet that uses unique symbols to denote each sound. This International Phonetic Alphabet (IPA), which is unfortunately going by the same acronym as its inventor, contains not only symbols for the general sounds (a *broad tran-*

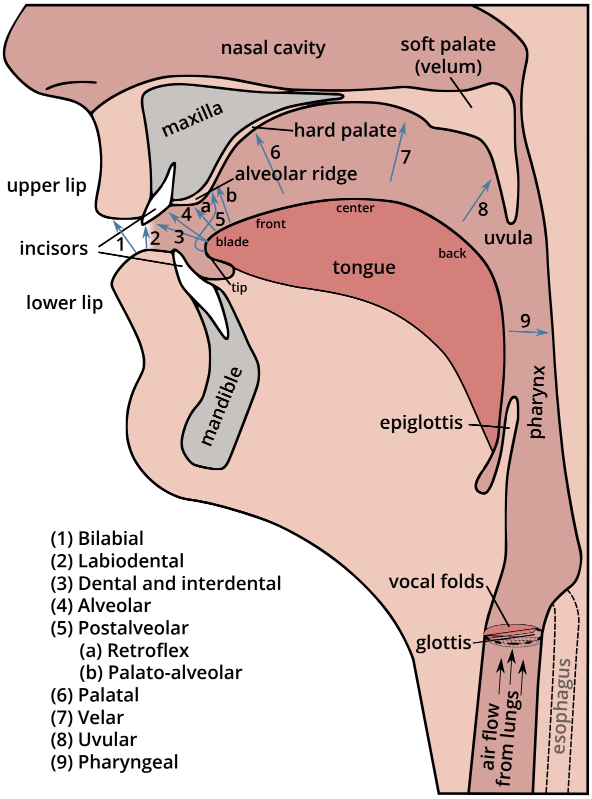


Figure 2.1.: Schematic view of the human vocal tract and places of articulation (adapted and expanded from [27]).

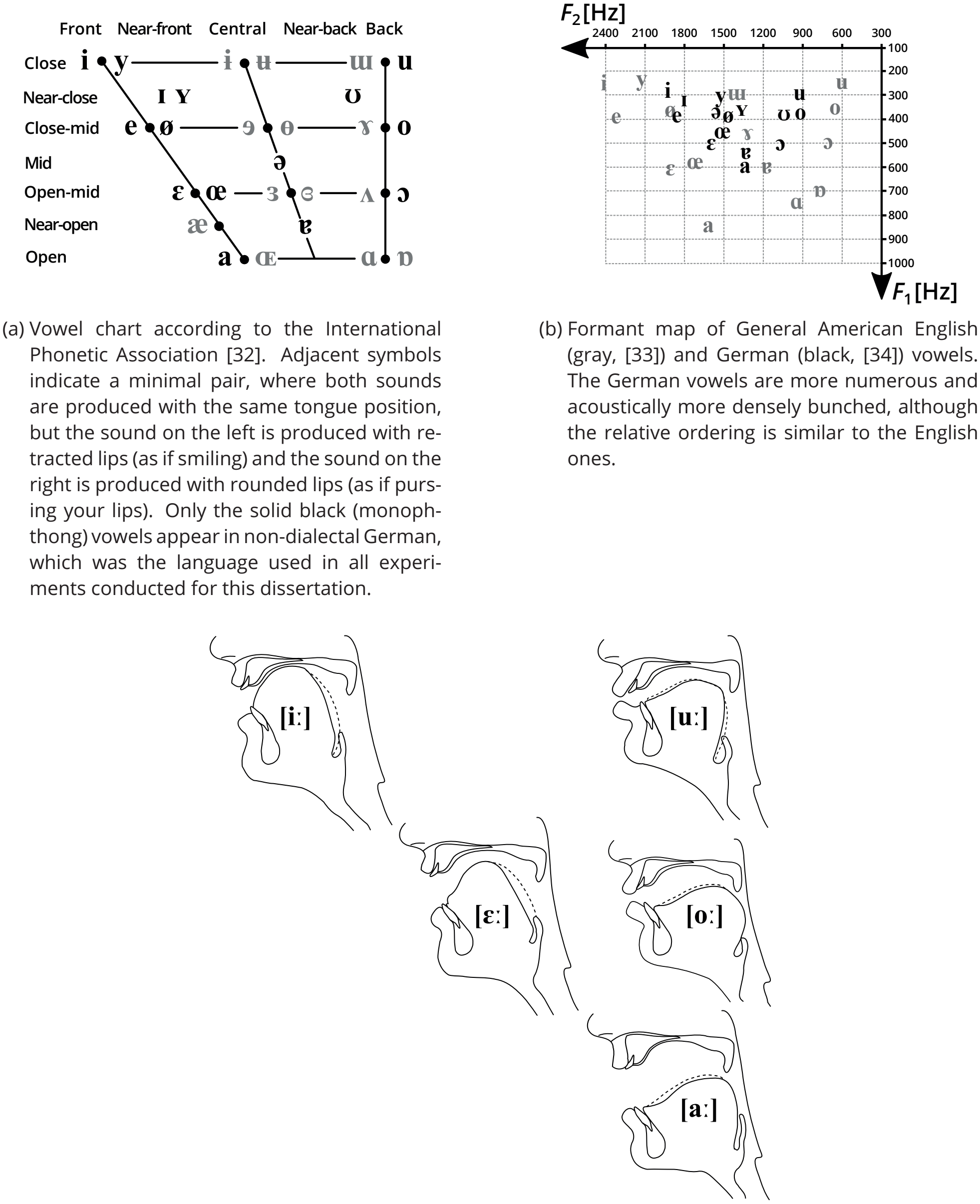
scription), but also provides diacritics to mark the exact pronunciation variations and minute details of phonation and articulation (a *narrow transcription*). The broad transcription is also called *phonemic*, because it only identifies the *phonemes* used to make up a word. Phonemes are the sounds used to discriminate meaning in a language and are the smallest units that cannot be swapped for a different one in a word without changing its meaning. Phonemic transcriptions are usually enclosed in forward slashes /·/. Narrow transcriptions are also called *phonetic* because they transcribe words at the level of the *phones*. Phones are any and all discernible speech sounds, regardless of their importance regarding the meaning of words. Phonetic transcriptions use square brackets [·]. The slashes-versus-brackets convention is not generally adhered to, however, especially in more technical-leaning works. A comprehensive chart with all symbols in the alphabet can be found in Appendix A.

2.2. Vowel sounds

Vowel sounds (Latin vocalis meaning “voiced”) are produced with a mostly open vocal tract and a voiced excitation (except when whispering, when they may be produced with a voiceless excitation). They are usually entirely characterized by only three articulatory parameters: the degree of lip rounding or spreading, and the tongue position along the high/low and front/back axes. From the perspective of the airflow, a high tongue position means a more closed vocal tract, while a low tongue position means a more open vocal tract. Therefore, the high/low dimension is also often called closed/open. Another way of grouping the vowels is by a quality called *tension*, where tense vowels are those produced with larger muscular effort and generally longer durations than lax vowels. This is, however, a rather vague distinction, since it is not well-defined what constitutes large effort, and the duration criterion is often fluid. The vowel system is very distinctive for any given language and the subsets of vowels occurring in English and German are shown in the summary in Table 2.2a at the end of this chapter. The IPA has compiled a vowel chart (see Figure 2.2a) based on the tongue and lip characteristics, which is language-independent and should in theory be able to assign a location for any vowel from any language. While this is certainly true with regards to the relative configurations of the articulators, the acoustic realizations of these "canonical" vowels can vary drastically across languages and not all vowels exist in every language. In some cases, the same sound is transcribed with a different symbol due to historic conventions (e.g., the sound /ɐ/ is often transcribed as /ʌ/ in English). Articulations of some of the most common vowels are shown in Figure 2.2c. Some languages also contain additional vowel sounds, e.g., the nasalized vowels in French, which are produced with a lowered velum and thus require another articulatory dimension. Lastly, all the vowels discussed up to here consist of only a single, quasi-static articulatory configuration and are therefore called monophthongs (Greek *monóphthongos* from mónos "single" and phthóngos "sound"). There are, however, also diphthongs (Greek *diphthongos* from di "double" and phthóngos "sound"), which are produced by a non-stationary articulation. where the beginning vowel glides towards an end vowel: the phrase *no highway cowboys*, for example, contains five of these gliding vowels, or diphthongs. The diphthongs occurring in both English and German are [a͜ɪ] and [a͜ʊ], while German additionally contains [ɔ͜ʏ] and English instead uses [ɔ͜ɪ] and additionally [eɪ] and [oʊ]. All vowel sounds are sonorants.

2.3. Consonantal sounds

Consonants (Latin consonans from con meaning "with" and sonare meaning "to sound") are sounds that are produced with an obstruction somewhere in the vocal tract (although confusingly, not all consonants are also obstruents). They are classified by the place and manner of this obstruction. The obstruction is formed when an articulator moves towards a *place of articulation*. A consonantal sound can therefore be specified by naming these two components. A "labio-dental" sound, for example, is a sound where the lower lip (Latin *labium*) as the articulator moves towards the teeth (Latin *dens*) to form the obstruction (see the numbered arrows in Figure 2.1). To differentiate further between sounds formed at the same place of articulation with the same articulators, the so-called *manner of articulation* describes the way the sound is articulated at that place in categorical terms. There are a total of seven manners of articulation [35], although some sources use six (e.g., [27], see section 2.3). Three of these categorize the degree of the obstruction: A *stop* is a sound including a complete closure as the obstruction (e.g., [p] in *peace*), a *fricative* has a narrow constriction instead (e.g., [f] in *fleece*), and an *approximant* has a slightly wider constriction (making the sound vowel-like, e.g., [w] in *wheeze*). In addition to these, other manners of articulation used to describe consonantal sounds are *trill* (caused by an airflow-induced vibration of the articulator, e.g., [r] in Spanish *perro*), *tap* (which is essentially a very brief stop, e.g., [ɾ] in *latter*), *lateral* (in which the airflow is directed through a lateral canal formed by the tongue, e.g., [l] in *fall*), and *nasal* (which is produced with a lowered velum, e.g., [m] in *home*). Since consonants can be produced using either a voiced or voiceless excitation (see section 2.1), a fully qualified consonant name consists of three components: (1) excitation mode, (2) articulator and place of articulation, and (3) manner of articulation. The consonant chart in Appendix A uses this terminology to describe the consonantal sounds of most languages. The subsets occuring in English and German, which are most relevant for this dissertation, are summerized with examples at the end of this chapter in Table 2.2b. The following subsections discuss the manners of articulation mentioned above in greater detail.



(a) Vowel chart according to the International Phonetic Association [32]. Adjacent symbols indicate a minimal pair, where both sounds are produced with the same tongue position, but the sound on the left is produced with retracted lips (as if smiling) and the sound on the right is produced with rounded lips (as if pursing your lips). Only the solid black (monophthong) vowels appear in non-dialectal German, which was the language used in all experiments conducted for this dissertation.

(b) Formant map of General American English (gray, [33]) and German (black, [34]) vowels. The German vowels are more numerous and acoustically more densely bunched, although the relative ordering is similar to the English ones.

(c) Example vowel articulations. The dashed lines are the contours of the side of the tongue.

Figure 2.2.: Articulatory and acoustic vowel spaces. While the relative order of the vowels is very similar in both spaces, the distances between the sounds are very different.

Nasals

A nasal sound is articulated with a lowered velum and thus an open velo-pharyngeal port, which is the opening between the nasal cavity and the pharynx. Theoretically, many sounds can be nasalized in this way (e.g., the French nasalized vowels [ɔ̃] in *bonjour*), but in English and German, only the three nasal consonants [m], [n], and [ŋ] exist as the nasalized versions of [b], [d] and [g], respectively (see Figure 2.3). Nasalized sounds in English and German are always voiced. They are also counted as sonorants because there are no major turbulences in the airflow through the vocal tract.

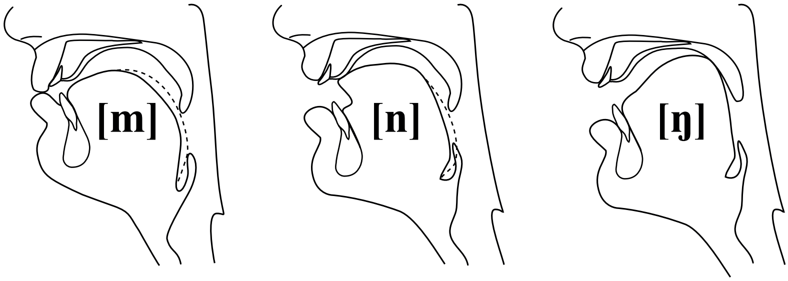


Figure 2.3.: Example articulations of the three English and German nasal consonants [m], [n], and [ŋ]. Note the lowered velum and thus open velo-pharyngeal port, which causes the airflow to continue despite the closed oral cavity. The dashed lines are the contours of the side of the tongue.

Stops

During the articulation of stops, a complete closure is formed in the vocal tract that stops the airflow (hence the name). With the airflow stopped and the velum raised, the pressure in the oral cavity rises. After typically 50 ms to 150 ms of complete closure (and thus a short period of silence in the speech signal), the closure is rapidly released and the built-up pressure discharges in a sudden burst sound (see subsection 2.4.3), e.g., in *pie* or *buy*. The vocal folds can be either abducted (glottis is open), which results in voiceless stop sounds ([p, t, k]), or adducted so that they start vibrating once the airflow resumes (i.e., the closure is released), which results in voiced stops ([b, d, g]. Both English and German use all six of these stops in addition to the glottal stop [ʔ], which is a sudden, deliberate closure of the glottis causing the flow-induced vocal fold vibration to stop.

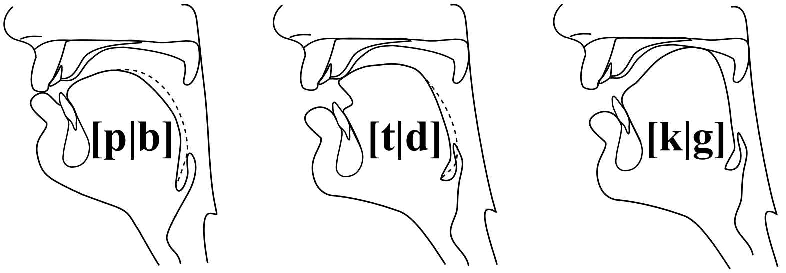


Figure 2.4.: Example articulations of the three English and German stops [p|b], [t|d], and [g|k]. The supra-glottal articulatory configuration is the same for each voiced-voiceless pair. The dashed lines are the contours of the side of the tongue.

Some sources (e.g., [27]) include nasals such as [m] and [n] in this category, since they are very similarly articulated and also include a closure in the oral cavity (see Figure 2.3). However, these sounds do not exhibit the closure release dynamics that are characteristic for stop sounds and they belong to the subset of sonorant sounds, whereas stop sounds are obstruents. Therefore, for the purposes of this dissertation, I will adhere to the commonly used system that includes nasals as their own manner of articulation.