Bachelorarbeit

Individualität der Stimmen – Eine Analyse der menschlichen Stimme und der Versuch, Vokale mittels Formantsynthese zu imitieren

Vorgelegt am 18. Januar 2024

Melanie Holzapfel

[redacted]

Erstprüfer: Prof. Dr. Robert Mores

Zweitprüfer: Prof. Thomas Görne

Hochschule für Angewandte Wissenschaften Hamburg
Department Medientechnik
Finkenau 35
20081 Hamburg

Zusammenfassung

In dieser Thesis wird die menschliche Stimme analysiert und teilweise synthetisiert. Phonetische Grundlagen werden erklärt, wie zum Beispiel die Vokale und ihre Relation zu den Formanten. Letztere sind physikalisch messbar und Grundlage für die Formantsynthese, eine Unterart der Sprachsynthese. Formanten sind Frequenzbereiche der menschlichen Stimme, welche einen besonders hohen Schalldruckpegel haben, im Vergleich zu angrenzenden Bereichen. Diese akustischen Parameter, die in Vokalen besonders prägnant vorkommen, werden untersucht und mittels eines Experimentes in die digitale Signalverarbeitung überführt. Die Vokale werden von verschiedenen Personen ausgesprochen und mithilfevon Praat, einem Programm zur Analyseund Synthese von Audio-Dateien, untersucht. Die Stimmen der Personen werden auf die Formantfrequenzen untersucht, denn die sind teilweise sprechendenspezifisch. Anhand der ermittelten Frequenzbereiche werden die Stimmen resynthetisiert. Abschließend werden die aufgenommenen Stimmen mit den synthetisierten Stimmen verglichen und auf personenspezifischen Ähnlichkeiten untersucht.

Abstract

In this thesis, the human voice is analyzed and partially synthesized. It delves into phonetic fundamentals, such as vowels and their relationship with formants. These formants are physically measurable and form the basis for formant synthesis, a type of speech synthesis. Formants are frequency ranges in the human voice that have a particularly high sound pressure level compared to adjacent areas. These acoustic parameters, which are particularly pronounced in vowels, are studied and transferred into digital signal processing through an experiment. The vowels are spoken by different people and analyzed using Praat, a program for the analysis and synthesis of audio files. Thevoices of the individuals areexamined for formant frequencies, as these are partly speaker-specific. Based on the identified frequency ranges, the voices are resynthesized. Finally, the recorded voices are compared with the synthesized voices and examined for person-specific similarities

I. Abbildungsverzeichnis

Inhaltsverzeichnis

1. Einleitung

Auf die Sprachsynthese durch ein Video aufmerksam geworden. In diesem wird die menschliche Stimme als Resultat einer Anreihung von Röhren verschiedenen Durchmessers betrachtet. Die Röhren werden wiederum von verschiedenen Filtern beschrieben, durch die bestimmte Frequenzbereiche verstärkt oder geschwächt werden. Die Stimme wurde nun nicht nur rein physiologisch von mir betrachtet, sondern hat für mich einen technischen Aspekt hinzugewonnen. Wenig später wurde mir bewusst, wie häufig man diesem technischen Ansatz begegnet. Nämlich in Form der Sprachsynthese bei Text-to-Speech (TTS).

Anwendungen mit TTS-Systemen findet man überall. Mit diesen Systemen lässt sich Text in - nicht unbedingt natürlich klingende - Sprache umwandeln. Dies ist zum Beispiel mit dem Google Translator möglich, mit dem man Wörter und Texte übersetzen kann. Hier gibt es zusätzlich die Option, sich das Übersetzte vorlesen zu lassen. Elektronische Stimmen, wie Siri und Alexa, können mithilfe von Sprachsynthese inzwischen auf ziemlich natürliche Art Fragen beantworten. Auch auf verschiedene Sprachen wird Rücksicht genommen, da diese sich unterschiedlich anhören können. Menschen, die aufgrund einer Krankheit nicht selbst sprechen können, können mithilfe von Programmen, die künstliche Sprache entwickeln, kommunizieren. Ein bek anntes Beispiel ist Stephen Hawking.

Schnell ist mir beim Recherchieren bewusst geworden, dass es nicht nur eine Art von Sprachsynthese gibt. Sprache kann auf vielen Wegen nachgestellt werden. Google Translator, Alexa und Siri arbeiten mit der Unit Selection. Bei dieser Art der Synthese spricht ein Mensch zunächst Phoneme, Wörter oder auch ganze Sätze ein. Gegebenenfalls werden sie zusätzlich in einzelne Laute - Phone genannt - geschnitten und in eine Datenbank eingepflegt.

Die Unit Selection weicht somit allerdings von der Vorgehensweise der im ersten Absatz beschriebenen Sprachsynthese ab, bei der ein Eingangssignal mithilfe von Filtern in ein Sprachsignal umgewandelt wird. Im Detail wird das Eingangssignal einer Quelle gefiltert und so entsteht letztendlich die Sprache. Daher wird diese Art der Synthese als sogenanntes Quelle-Filter-Modell bezeichnet. Das Beispiel aus dem ersten Absatz beschreibt das Akustische Modell, welches eine Unterkategorie zum Quelle-Filter-Modell darstellt. Beispiele für Quelle-Filter-Modelle sind:

- Das Akustische Modell, bei welchem der gesamte menschliche Vokaltrakt nachgebildet wird und durch dieses das Eingangssignal geschickt wird.
- Das Artikulatorische Modell, bei dem vor allem Zungen- und Lippenbewegungen betrachtet und nachgeformt werden

- Die Formantsynthese, bei der nur das Eingangs- und Ausgangssignal betrachtet werden und daraus die Übertragungsfunktion des Vokaltraktes ermittelt wird

Vor allem das Akustische Modell, bei welchem - mithilfe von vielen Filtern [1]- der gesamte Vokaltrakt nachgebildet wird, benötigt eine hohe Rechenkapazität. Sie werden meistens in Reihe geschaltet und haben ihre eigenen Parameter. Das Eingangssignal wird Stück für Stück mit jeder Impulsantwort eines Filters gefaltet. So ist es möglich, diesem durch viele Filter nachgebildeten Vokaltrakt eine eigene Stimme zu geben. Aber was sind die Parameter einer eigenen, individuellen Stimme und sind sie objektiv feststellbar? Die Antwort lautet: Teilweise ja und zwar durch die sogenannten Formanten. Und kann man das Wissen über die Formanten nutzen, um eine Stimme zu synthetisieren? Ja. Dafür gibt es die sogenannte Formantsynthese.
Die Formantsynthese ist ebenfalls ein Quelle-Filter-Modell. Im Gegensatz zum Akustischen oder Artikulatorischen Modell wird der Vokaltrakt nicht nachgestellt. Formanten sind Frequenzbereiche - die im Vergleich zu benachbarten Bereichen - einen höheren Schalldruckpegel aufzeigen. Das gewünschte Ausgangssignal wird nur nach Formanten synthetisiert. Das liegt daran, dass die Frequenzen der ersten zwei Formanten auf Vokale schließen lassen. Wie genau wird in Kapitel … erläutert. Ein Formant wird mit je einem Filter nachgebaut. Die Natürlichkeit, für die man die Prosodie, also die Stimmmelodie, benötigt, ist allerdings bei der Formantsynthese nachrangig. Der Fokus in der Praxis liegt bei der Sprachverständlichkeit.
Die ersten zwei Formanten sind vokalspezifisch. Die Vorstellung, dass Vokale durch nur zwei Filter und einer Quelle nachgestellt werden können, hat mein Interesse an der Formantsynthese geweckt. Und wie sieht es mit höheren Formanten aus? In einem großen Teil der Literatur wird beschrieben, dass die Formantsynthese genutzt wird, um Sprache verständlich rüberzubringen, nicht natürlich. Dennoch gibt es in einigen Artikeln Hinweise, dass höhere Formanten auf die individuellen Stimmen der Sprechenden schließen lassen. Dem wird in dieser Thesis nachgegangen.

1.1 Aufbau der Thesis

Zunächst wird auf die theoretischen Grundlagen eingegangen. Zu diesen zählt zu einen die Anatomie. Hier wird kurz auf die Glottis - die Quelle der Stimme - und den Vokaltrakt – der als Filter dient – eingegangen. Abgesehen von der Glottis und dem Vokaltrakt werde ich nicht

[1] Als Beispiel: Es gibt Modelle, die aus 40 Rohrstücken zusammengesetzt sind, siehe Karl Schnell (2003). *Rohrmodelle des Sprechtraktes, S. 86*

großartig auf die umliegende Anatomie, wie zum Beispiel die Lungen oder weitere Muskelpartien, eingehen.
Danach werden die Parameter eines Sprachsignals erläutert. Dazu gehören die Grundfrequenz und die Formanten und deren Einfluss auf die Vokale.
Im Kapitel über die Formantanalyse werden Analyseverfahren aufgeführt, mit denen Formanten grafisch dargestellt werden können. Die Diskrete Fourier-Transformation (DFT) wird erläutert. Sie ist zur Erstellung von Schmal- bzw. Breitbandspektrogrammen notwendig, welche die Frequenzverläufe über die Zeit darstellen. Außerdem wird mit der DFT der Amplituden-Frequenzverlauf dargestellt. Neben ihr gibt es die Lineare Prädiktion (LP) als Analyseverfahren. Für die Entwicklung der Spektrogramme und Spektren wird Praat genutzt. Das Programm analysiert phonetischen Parameter, kann aber auch Signale synthetisieren.
Im Kapitel der digitalen Formantsynthese wird auf die Idee der Quelle-Filter-Theorie eingegangen und welche Methoden es gibt, um Vokale mittels Formanten nachzustellen. Es wird betrachtet, welche Art von Quellsignal nötig ist, wie das digitale Filter aufgebaut ist und wie am Ende das Sprachsignal theoretisch erstellt wird.
Nach der Theoriegeht es in die Praxis. Es wird der Ablauf des Experimentes vorgestellt. Personen sprechen nach abgesprochenen Kriterien und unter bestimmten Voraussetzungen ausgewählte Vokale ein, die daraufhin mit den vorgestellten Analyseverfahren ausgewertet werden. Aus den ermittelten Frequenzen für die Formanten und die Grundfrequenz werden mithilfe von Praat die personenspezifischen Vokale synthetisiert.
Die synthetisierten Vokale werden mit den gleichen Analyseverfahren untersucht wie die aufgenommenen Vokale. Die daraus erstellten Graphen werden miteinander verglichen.

1.2 Ziel dieser Thesis

Zum einen möchte ich mir durch das Experiment einen praktischen Bezug zur Theorie geben, die ich in vielen Büchern und Papern gelesen habe.
Dafür werden die in der Literatur angegebenen Frequenzwerte der ersten zwei Formanten in Vokalen mit den Werten von aufgenommenen Stimmen verglichen. Diese sollen, wie im vorherigen Abschnitt erwähnt, Aufschluss auf die gesprochenen Vokale geben, unabhängig davon, welcher Mensch diese ausgesprochen hat.
In einem Experiment werden die Stimmen von Sprecherinnen und Sprechern aufgenommen. Sie sprechen verschiedene deutsche Vokale ein. Was und wie sie einsprechen sollten, wird in Kapitel 5 näher erläutert. Es werden nur Vokale analysiert, da in diesen die Formanten besonders gut sichtbar sein sollen. Die Konsonanten, bei denen Formanten wenig sichtbar sind, werden dementsprechend weniger betrachtet.

Von diesen Audiodateien werden Spektrogramme entwickelt, in denen die Formantfrequenzen gut sichtbar sein sollen. Im ersten Schritt werden die ersten beiden Formanten (auch F1 und F2 genannt), die für die Vokalunterscheidung wichtig sind, untersucht. Dann werden F3 und höhere Formanten entnommen werden, zu denen man weniger in der Literatur nachlesen kann, weil sie sich von Mensch zu Mensch stärker unterscheiden als F1 und F2.

Aus den Daten, die ich durch die Formantanalyse erlange, sollen Vokale entsprechend den individuellen Stimmen der Sprechenden mithilfe der Formantsynthese nachgebildet werden. Durch die fehlende Prosodie wird keine natürlich klingende, synthetische Stimme erwartet. Wegen der fehlenden Konsonanten ist es kaum möglich, ein komplettes Wort zu synthetisieren. Anhand einer schriftlichen Arbeit kann man nicht direkt beurteilen, ob sich die synthetische Stimme ähnlich wie die aufgenommene Stimme anhört. Daher soll die synthetische Stimme so analysiert werden, wie es bei den originalen Stimmen auch getan wurde. Diese Analysen werden am Ende verglichen.

2. Analyse der menschlichen Stimme

Damit die menschliche Stimme synthetisiert werden kann, muss sie erstmal analysiert werden. Grob gesagt ist die menschliche Stimme das Resultat durch Atemluft, welche die Glottis anregt und in Schwingungen versetzt, und auf dem Weg durch den Vokaltrakt moduliert wird.
Bei der Formantsynthese handelt es sich - wie in Kapitel 1 bereits erwähnt ist - um ein Quelle-Filter-Modell. Das bedeutet, dass für die Synthese das Quellsignal separiert von den Filtern betrachtet wird. Das sollte man bei der Analyse der Glottis und des Vokaltraktes im Hinterkopf behalten. Welche Analogien zur Glottis und dem Vokaltrakt bestehen, wird im folgenden Abschnitt genauer beschrieben.

2.1 Die Quelle: Die Glottis und die Atemluft

Als Quelle der Stimme gilt die Glottis, wo die Stimmlippen durch die Atemluft in Schwingung versetzt werden. In der Anatomie besteht sie aus Schildknorpel, Stellknorpeln und Stimmbändern und ist ein Teil des Kehlkopfes (siehe Abbildung 1). In der Phonetik wird die Glottis allerdings häufig als der Spalt zwischen den paarigen Stimmlippen beschrieben. Sie und die Atemluft stellen zusammen den Oszillator dar.

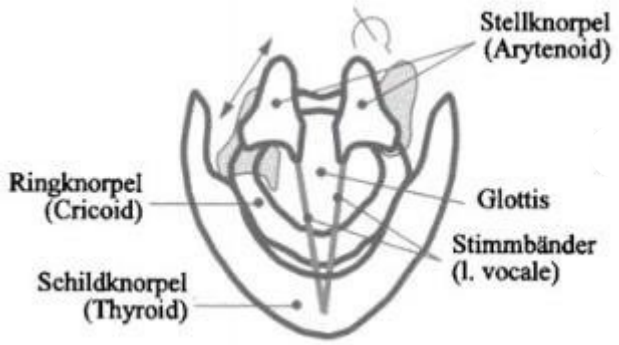


Abbildung 1: Querschnitt des Kehlkopfes mit der Glottis zwischen den Stimmbändern
Quelle: Pompino-Marschall, 2003, S.33

Durch die Glottis wird nicht nur einen Ton erzeugt, der im Sinne der Physik eine Sinusschwingung mit einer Grundfrequenz *f* ist (siehe Gleichung 1).

$$y(t) = sin\ (2\pi f t + \varphi) \tag{1}$$

Sie erzeugt einen Klang, also eine Grundschwingung, die mit ganzzahligen Vielfachen *n* ihrer Grundfrequenz überlappt wird. Die einzelnen Frequenzen werden Harmonische genannt. Die Schwingung mit der Grundfrequenz wird als die erste harmonische Schwingung bezeichnet. Der

n-ten Harmonischen f_n [2]- z.B. zweite, dritte, vierte Harmonische - lässt sich das *n*-fache der Grundfrequenz f_0 mit der Gleichung 2 zuweisen:

$$f_n = n \cdot f_0 \tag{2}$$

Ein Klang, der als Summe aus Vielfachen seiner Grundfrequenz besteht, lässt sich dementsprechend mit der Gleichung 3 beschreiben.

$$y(t) = \sum_{n=1}^{\infty} A_n \cdot sin\,(2\pi f_n\, t + \varphi_n) \tag{3}$$

Mit welchen Grundfrequenzen gesprochen wird, ist unter anderem abhängig von der Form und Größe der Glottis. Ist sie zum Beispiel größer, folgen daraus längere Stimmbänder, welche wiederum tiefere Frequenzen erzeugen, aus denen eine dunklere Stimme resultiert. Aber auch während des Sprechens ändert sie sich. Bei Männern liegt sie im Durchschnitt bei 120 Hz und bei Frauen bei 220 Hz.[3]

Die Ausbreitung des Luftstromes verläuft als eine Longitudinalwelle. Es sind Schallwellen, die durch Druck- und Dichteschwankungen entstehen. Die Stärke des Luftstroms bestimmt die Lautstärke der Stimme. Laute, die durch die Glottis erzeugt werden, werden als stimmhaft bezeichnet.

Wie der Luftstrom durch die Glottis modelliert wird, lässt sich in mehrere Phasen unterteilen. Eine der ersten Ideen kommt von Rosenberg (1971). Er stellte den Glottisimpulswie in Abbildung 2 dar. T_o beschreibt die Phase, in der sich die Glottis öffnet und T_c die Phase, in der sie sich schließt. Die Periode *T* beinhaltet T_o, T_c und die komplett geschlossene Phase, bevor die Glottis sich wieder öffnet. Dieses Modell birgt allerdings ein paar Nachteile. Auf diese wird in Abschnitt 4.1 eingegangen und zusammen mit dem LF-Modell vorgestellt, welchesheutzutage vermehrt für Quelle-Filter-Modelle genutzt wird.

[2] *n* darf nicht mit der diskreten Zeiteinheit *n* aus Abschnitt 2.2 verwechselt werden

[3] In der Literatur gibt es diverse Durchschnittswerte. Als Referenz wurden hier die Werte von Pfister und Kaufmann (2017), *Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung. 2. Auflage, S. 13* herangezogen

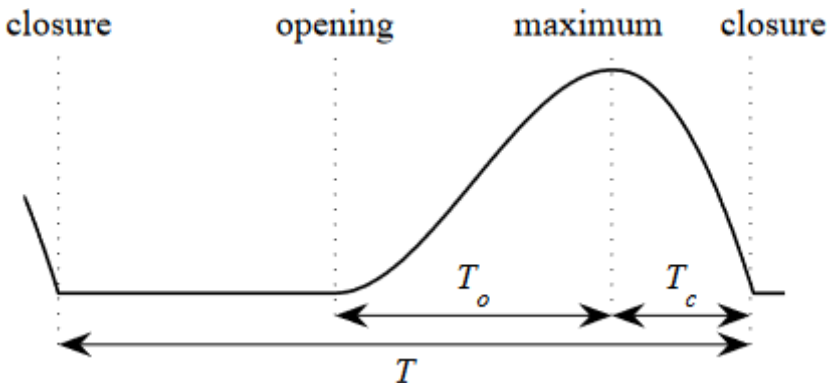


Abbildung 2: Modell eines Glottisimpuls nach Rosenberg (1971)
Quelle: H. Pulakka (2005). Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottograph, S. 22

2.2 Die Filter: Anatomie des Vokaltraktes

Die menschliche Stimme wird durch den Vokaltrakt - auch häufig als Ansatzrohr bezeichnet - moduliert. Dieser beginnt im Kehlkopf über der Glottis und geht durch die Rachen-, Mund- und Nasenhöhlebis zu den Lippen bzw. Nasenlöchern. Die Organe im Vokaltrakt werdenzudem grob in zwei Bereiche geteilt: Artikulationsorte und Artikulationsorgane. Die Artikulationsorte sind dabei der unbewegliche Teil des Vokaltraktes. Die Artikulationsorgane sind die beweglichen Teile. Je nach Lage der Artikulationsorgane und Formungen des Vokaltraktes entstehen unterschiedliche Resonanzen, diewiederumden unterschiedlichen Klang der Stimmeausmachen. Sie werden in Abschnitt 2.2.2 näher betrachtet.
Laut C. Julian Chen (2016) beträgt die durchschnittliche Länge des Vokaltraktes 16,9 cm bei Männern und 14,1 cm bei Frauen.[4]

2.2.1 Artikulationsorgane und -orte

Innerhalb des Vokaltraktes werden einzelne Komponenten in Artikulationsorgane beziehungsweise Artikulationsorte eingeteilt. Beispiele dafür sind die Nasenhöhle, der Rachen oder der Oberkiefer. Die Artikulationsorte sind eher unflexibel und können die Sprache aktiv wenig bis gar nicht modellieren. Durchvon Mensch zu Mensch verschiedene Größen und Formen entstehen allerdings an unterschiedliche Orten Reflektionen, die zum individuellen Klang von Stimmen beitragen.

[4] Chen bezieht sich in seinem Buch (*Elements of human voice, S. 51*) auf die Werte aus *Acoustic Phonetics* von Kenneth N. Stevens (2000)

Die Artikulationsorgane - auch Artikulatoren genannt - sind für die Sprachmodellierung verantwortlich. Zu diesen zählen zum Beispiel die Zunge, die Lippen und der Unterkiefer. Die Lage der Artikulationsorgane bildet verschiedene Laute, auch Phone genannt. Die Aneinanderreihung von Phonen ergibt die Sprache.

2.2.2 Rachen-, Mund und Nasenhöhle

Die Rachen-, Nasen- und Mundhöhle stellen Resonanzräume dar. Entsprechend ändern sich von Mensch zu Mensch die Resonanzen - die den Klang ausmachen - da sich der Aufbau des Vokaltraktes und die Glottis unterscheiden.
Die Rachen- und Nasenhöhle, die sich kaum vom Aufbau und der Größe während des Sprechens ändern, haben kaum einen Einfluss auf die Resonanzfrequenzen.
Vom Rachen kann die Atemluft durch die Mund- und/oder Nasenhöhle entweichen. Die Nasehöhle zweigt sind innerhalb der Mundhöhle ab und ist auch unbeweglich. Wo die Atemluft austritt, ist abhängig von der Lage des Velums und/oder der Zunge, denn diese können die Ausgänge an der Nase und den Lippen versperren.
In der Mundhöhle entstehen die größten Änderungen der Phone. Das liegt zum großen Teil an der Stellung der Zunge. So kann sie eher vorne oder hinten liegen (front/back). Zudem kann der Abstand der Zunge zum Oberkiefer variieren (height). Die Zunge bestimmt am stärksten das Volumen der Mundhöhle, also das Volumen eines Resonanzraumes und die daraus resultierende Verschiebung von Resonanzfrequenzen. Aber auch die Rundung der Lippen (rounding) kann für unterschiedliche Phone sorgen.
„Front/back“, „height“ und „rounding“ sind die Kriterien, mit denen die Entstehung der unterschiedlichen Phone beschrieben werden. Beispiele dazu werden in Abschnitt 2.5 genannt. Das Vokaltrapez in Abbildung 3 zeigt beispielsweise den Zusammenhang der Vokale mit der Lage der Zunge.

2.3 Grundfrequenz F0

Auch wenn die Grundfrequenz, die ein Mensch zum Beispiel durch Singen erreicht, nicht auf die Formanten der menschlichen Stimme zurückführt, so ist sie dennoch wichtig für das Hörempfinden. In der Literatur wird diese häufig als F0 abgekürzt. Trotz des großgeschriebenen F darf sie nicht mit den Formanten verwechselt werden, welchedie Vokale oder die Individualität der menschlichen Stimme aufzeigen können.

Die durch die Glottis erzeugte Grundfrequenz wird zusammen mit ihren Harmonischen, also ganzzahligen Vielfachen, erstellt. Diese Schwingungen gleichen im Idealfall einer Sinusschwingung (erste Harmonische) mit ihren zugehörigen Oberfrequenzen (zweite bis *n*-te Harmonische).[5] Innerhalb des Vokaltraktes können sich diese durch Reflexionen und daraus resultierenden Überlappungen der Schwingungen verstärken, abschwächen oder auslöschen. Die Frequenzbereiche, an denen Harmonische verstärkt sind, werden Formanten genannt.

2.4 Formanten

Bei Musikinstrumenten wie Saiteninstrumenten - mit und ohne Hohlkörper - können sich Frequenzen verstärken. Welche das sind ist z.B. abhängig vom Material oder der Form des Instrumentes. Sie werden Eigenfrequenzen oder Moden genannt. Daraus folgen Frequenzbereiche, in denen die Amplitude im Vergleich zu benachbarten Frequenzbereichen besonders hoch ist. Andere Frequenzen hingegen werden gedämpft. Auch der menschliche Körper kann als Resonanzkörper betrachtet werden. Im Fall der Stimm- und Sprachmodellierung spielt der Vokaltrakt eine große Rolle, welcher, wie in vorangegangenen Abschnitt 2.2.2, in die Resonanzräume Rachen-, Mund- und Nasenhöhle aufgeteilt wird. Je nach Form und Größe dieser Resonanzräume werden die von der Glottis und Atemluft erschaffenen Harmonischen unterschiedlich verstärkt oder gedämpft. Frequenzbereiche mit einer höheren Schalldruckpegel (kurz *SPL*), die durch Resonanzen im Vokaltrakt entstehen, werden als Formanten bezeichnet. Den höchsten SPL haben sie, wenn die durch den Vokaltrakt modellierte Atemluft ungehindert austreten kann. Sie wird nicht von Zunge, Zähnen und Lippen blockiert. So gibt es weniger Energieverluste durch Reflektionen innerhalb des Vokaltraktes. Dies ist der Fall, wenn Vokale ausgesprochen werden. Auf diese wird in Abschnitt 2.5 genauer eingegangen.
In der Formantanalyse und -synthese werden in der Regel die ersten zwei Formanten betrachtet, da diese für die Unterscheidung von Vokalen ausreichend sind. Sie sind also vokalspezifisch. Der Formant mit der niedrigsten Frequenz bekommt die Abkürzung F1, die zweitniedrigste heißt F2 und so weiter. Wie man Formanten erkennt, wird in Kapitel 3 mithilfe von Spektrogrammen und Spektren gezeigt. Vor allem durch die Änderung der Zungenlage und Rundung der Lippen ändern sich deren Frequenzen. Ab F3 sind die Formanten sprechendenspezifisch. Sie geben der Stimme, ähnlich wie bei Instrumenten, ihren eigenen Klang.

[5] Siehe Gleichung 1 bis 3 in Abschnitt 2.1

2.5 Vokale

Als Vokale werden in der Phonetik Laute beschrieben, welche durch ein ungehindertes Passieren von Atemluft durch den Vokaltrakt entstehen. Zunge und Lippen blockieren nicht den Weg zwischen der Glottis und dem Ausgang bei den Lippen. So entsteht in der Sprache bei Vokalen ein besonders hoher SPL. Das lässt sich in Spektrogrammen ablesen. Dieses zeigt die frequenzabhängige Schallintensität über den Verlauf der Zeit auf. In den meisten Programmen heißt es: Je dunkler eine Frequenz dargestellt wird, desto höher ist der SPL. An den sehr dunklen Stellen lassen sich die Formanten finden.

Vokale sind dabei nicht gleichzusetzen mit den Buchstaben /a/, /e/, /i/, /o/ und /u/. [6] Letztere werden im Deutschen als Vokalbuchstaben bezeichnet. Sie sind zwar Vokale, sie können dennoch anders klingen, je nachdem in welchem Bereich des Wortes sie stehen und mit welchen anderen Phonen sie kombiniert werden. Um das genauer beschreiben zu können, wird die Lautschrift des IPA (Internationales Phonetisches Alphabet) benutzt. Ein Beispiel ist das Wort „weggehen“. Im IPA wird es mit [ˈvɛkˌgeːən] geschrieben. Die Vokalbuchstaben /e/ werden auf drei verschiedene Arten beschrieben: [ɛ], [eː] und [ə]. [ɛ] ist kurz ausgesprochen und könnte auch dem Vokalbuchstaben „ä“ zugeordnet werden (z.B. [ɛkvivaˈlɛnt] für „äquivalent“), [eː] klingt langgezogen (z.B. [ʃneː] für „Schnee“) und [ə] ist wieder kurz ausgesprochen und unbetont (z.B. [ˈmɪtə] für „Mitte“).

In Kapitel 2.2.2 wurde bereits erwähnt, dass die Lage der Zunge maßgeblich für die unterschiedlichen Vokale ist. Der Zusammenhang zwischen der Lage der Zunge und den Vokalen werden im sogenannten Vokaltrapez, welches man in der Abbildung 3 sehen kann, aufgezeigt.

[6] [*] ist die Notation für Sprachlaute (Phone), die im IPA stehen. Die Notation /*/ ist zur Kenntlichmachung von Phonemen gedacht. Letztere sind nur jene Einheiten, die das Lautsystem einer Sprache bilden (laut B. Kortmann (2020). *English Linguistic: Essentials (S. 27)*)

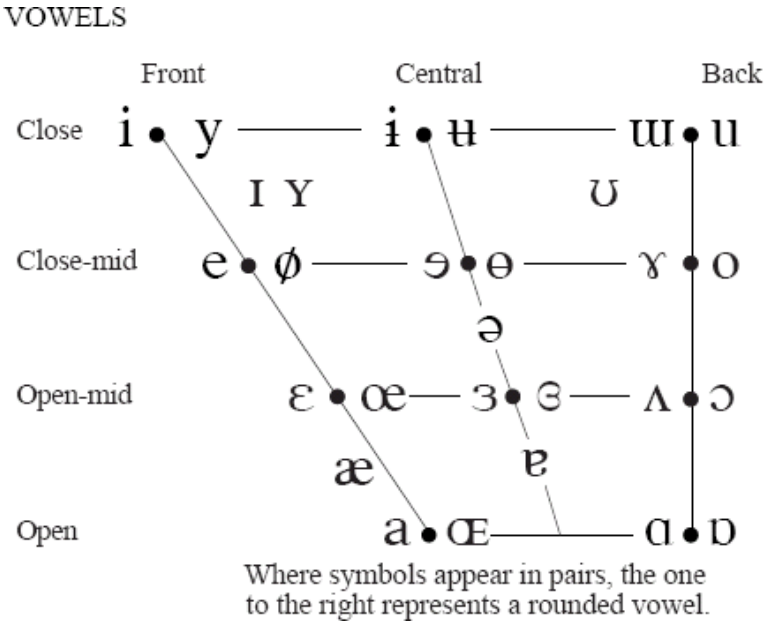


Abbildung 3: Vokaltrapez aus dem International Phonetic Alphabet Chart
Quelle: International Phonetic Association (2015). IPA Chart

Auf der horizontalen Achse wird beschrieben, ob die Zunge in der Mundhöhle eher vorne, mittig oder hinten (front/central/back) liegt. Je nach Lage spricht man dann von einem Vorder-, Zentral- oder Hinterzungenvokal. Auf der Vertikalen sieht man, wie hoch die Zunge liegt (height). Wenn sie hoch liegt, werden die Vokale als geschlossen (close) bezeichnet, da der Luftstrom fast nicht mehr ungehindert durch den Vokaltrakt kommt (z.B. [i], [y] und [u]). Wenn sie tief liegt, kann der Luftstrom entsprechend ungehindert ausströmen (z.B. [a]). Zusätzlich wird zwischen gerundeten und ungerundeten Lippen unterschieden (rounding). So werden Vokale entsprechend der Zungenlage und Lippenformung definiert. Als Beispiel:

[i:]: Ungerundeter geschlossener Vorderzungenvokal
[a:]: Ungerundeter offener Zentralvokal
[o:]: Gerundeter halbgeschlossener Hinterzungenvokal

Das Vokaltrapez lässt zudem auf den Zusammenhang von F1 zu F2 schließen. In Abbildung 4 sieht man ein Diagramm, auf dessen Ordinate die Frequenzwerte von F2 und auf der Abszisse F1 abgebildet sind. Vergleicht man die Position der Vokale in Abbildung 4 mit dem Vokaltrapez in Abbildung 3, dann sind Ähnlichkeiten feststellbar.

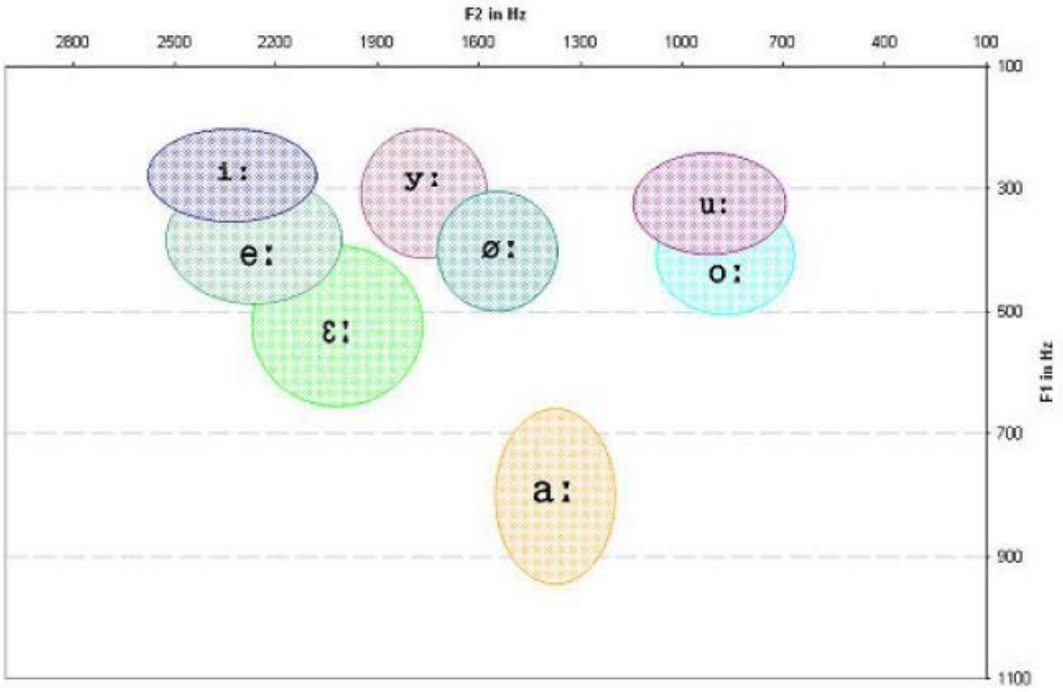


Abbildung 4: Lange Vokale in Abhängigkeit von F1 und F2
Quelle: W. F. Sendlmeier, J. Seebode (2006). Formantkarten des deutsche Vokalsystems

Walter F. Sendlmeier und Julia Seebode haben die ersten beiden Formanten von deutschen Vokalen bei 69 Männern und 58 Frauen ausgewertet und nach Geschlecht unterschieden. Ein Teil der Vokale ist in Tabelle 1 zu sehen.

Tabelle 1: Durchschnittliche Werte für F1 und F2, nach Geschlecht differenziert[7]

M (n=69)	F1 [Hz]	F2 [Hz]
[a:]	737	1275
[e:]	348	2126
[i:]	263	2171
[o:]	383	841
[u:]	310	854
[ɛ:]	482	1902
[ø:]	371	1501
[y:]	302	1722

W (n=58)	F1 [Hz]	F2 [Hz]
[a:]	896	1517
[e:]	434	2461
[i:]	302	2533
[o:]	440	889
[u:]	345	956
[ɛ:]	584	2166
[ø:]	440	1605
[y:]	320	1810

Quelle: W. F. Sendlmeier, J. Seebode (2006)

[7] Es sind hier nicht alle Vokale des deutschen Vokalsystems aufgelistet, sondern die, die im weiteren Verlauf der Thesis im Experiment vorkommen. Weitere Vokale werden in den *Formantkarten des deutschen Vokalsystems* von Walter F. Sendlemeier und Julia Seebode aufgelistet.
Die Abkürzung „M“ steht für die männlichen Sprecher, „W“ für die weiblichen Sprecherinnen , „n“ gibt die Anzahl an Teilnehmenden wieder

3. Formantanalyse

Da die Individualität der menschlichen Stimme in dieser Thesis betrachtet wird, müssen die im Frequenzspektrumhöher liegende Formanten ab F3 beobachtet werden. Bei diesen sind vor allem die Artikulationsorte, die in Kapitel 3.2 beschrieben sind, für den unterschiedlichen Klang der Stimme – auch Tirmbre genannt - verantwortlich.
Um die Formanten einer sprechenden Person zu ermitteln, muss ein Sprachsignal mittels eines Spektrogramms und/oder Spektrums analysiert werden. Im Zeitbereich können sie mithilfe eines Breitbandspektrogramms sichtbar gemacht. Bei Formanten handelt es sich um Frequenzbereiche, weswegen das Sprachsignal neben ihren Resonanzfrequenzen auch auf deren Bandbreiten untersucht werden muss. Das ist durch verschiedene Transformationen möglich. Am geläufigsten ist bei digitalen Signalen die Diskrete Fourier-Transformation (kurz DFT). Aber auch die Lineare Prädiktion (LP) wird bei Sprachsignalen häufig verwendet, um Formanten zu ermitteln.

3.1 Diskrete Fourier-Transformation (DFT)

Um Frequenzen aus dem diskreten Zeitsignal zu ermitteln und sie im Spektrogramm darzustellen, wird die DFT benutzt. Sie wird wie folgt beschrieben:

$$X[k] = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \tag{4}$$

X[k]: Abtastwert
x(n): Eingangssignal; Amplitude am diskreten Zeitpunkt n
N: Anzahl der Abtastwerte

Je höher N ist, desto genauer werden die Frequenzen abgebildet. Die Anzahl ist abhängig von der zeitlichen Länge des Fensters. Je länger das Fenster ist, desto mehr Werte existieren in diesem Fenster. Bei einer geläufigen Abtastfrequenz von 44100 Hz gibt es in einer Sekunde 44100 Werte. In der Regel wird allerdings eine Fenstergröße von wenigen Millisekunden (für Breitbandspektrogramme) bis Zehntelsekunden (für Schmalbandspektrogramm) gewählt.
Die DFT händisch zu berechnen, kann bei den nicht-periodischen Sprachsignalen sehr lange dauern. Programme (z.B. Audacity und Praat) nutzen hierbei die Fast Fourier-Transformation (kurz *FFT*), die durch einen Algorithmus wie eine DFT agiert. Bei der FFT wird das Sprachsignal vorher mit einer Fensterfunktion gewichtet. Die Funktion kann in ihrer Fensterform und -länge variiert werden. Praat hat als voreingestellte Funktion das Gauß-Fenster mit einer Länge von 5

ms. Es sorgt dafür, dass die Amplitude des Zeitsignals am Anfang und Ende des Fensters gleich 0 ist. So werden bei der Transformation unerwünschte Frequenzen vermieden.
In Abbildung 5sind Spektren zu sehen, bei denen die FFT an derselben Stelle des Sprachsignals ausgeführt und mit der gleichen Fensterfunktion - dem Gauß-Fenster - gewichtet wurden. Der einzige Unterschied ist die Fensterlänge. Das schwarze Spektrum wurde mit einer großen Fensterlänge (0,05 s) erstellt. Die Harmonischen sind als die Peaks zu erkennen. Die rote Linie zeigt das Spektrum, welches mit einer kleineren Fenstergröße erstellt wurde (0,005 s). Die Formanten sind als Erhebungen zu sehen und von F1 bis F5 beschriftet. Die erste Erhebung stellt die erste Harmonische dar.

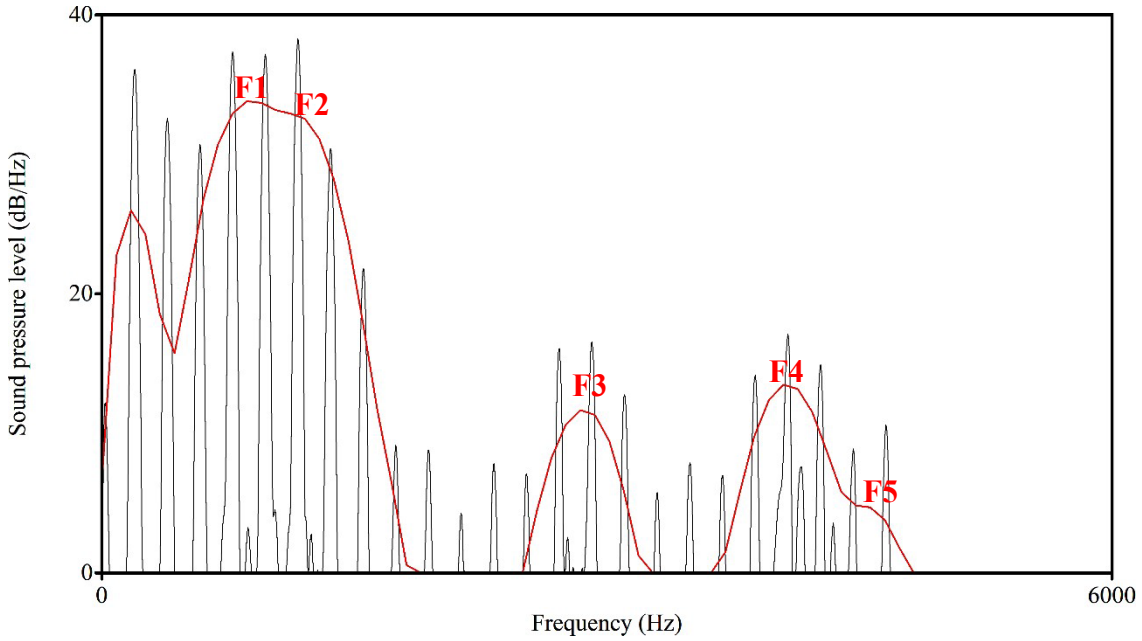


Abbildung: 5 Spektrum von [a:], gesprochen von W1
FFT mit Fensterlänge 0,05 s (rot) und Fensterlänge 0,005 s (schwarz)
Quelle: Eigene Darstellung

3.2 Lineare Prädiktion (LP)

Eine andere Möglichkeit ein Sprachsignal auf Formanten zu untersuchen, ist die lineare Prädiktion. Aufeinanderfolgende Abtastwerte von digitalen Sprachsignalen sind häufig zueinander abhängig. Bei der linearen Prädiktion wird davon ausgegangen, dass ein Abtastwert $\tilde{s}(n)$ durch eine gewichtete Summe aus K vorangegangenen Abtastwerten s(n-1),…, s(n-K) vorausgesagt werden kann. Der prädizierte Abtastwert wird wie folgt berechnet:

$$\tilde{s}(n) = \sum_{k=1}^{K} a_k s(n-k) \tag{5}$$

s(n): Sprachsignal
$\tilde{s}(n)$: prädizierter Abtastwert
K: Ordnung des Prädiktors
a_k: Gewichtungskoeffizient des Prädiktors, auch LPC-Koeffizient genannt

Das LPC-Spektrum ist der Betrag der Übertragungsfunktion H(z) des Synthesefilters.

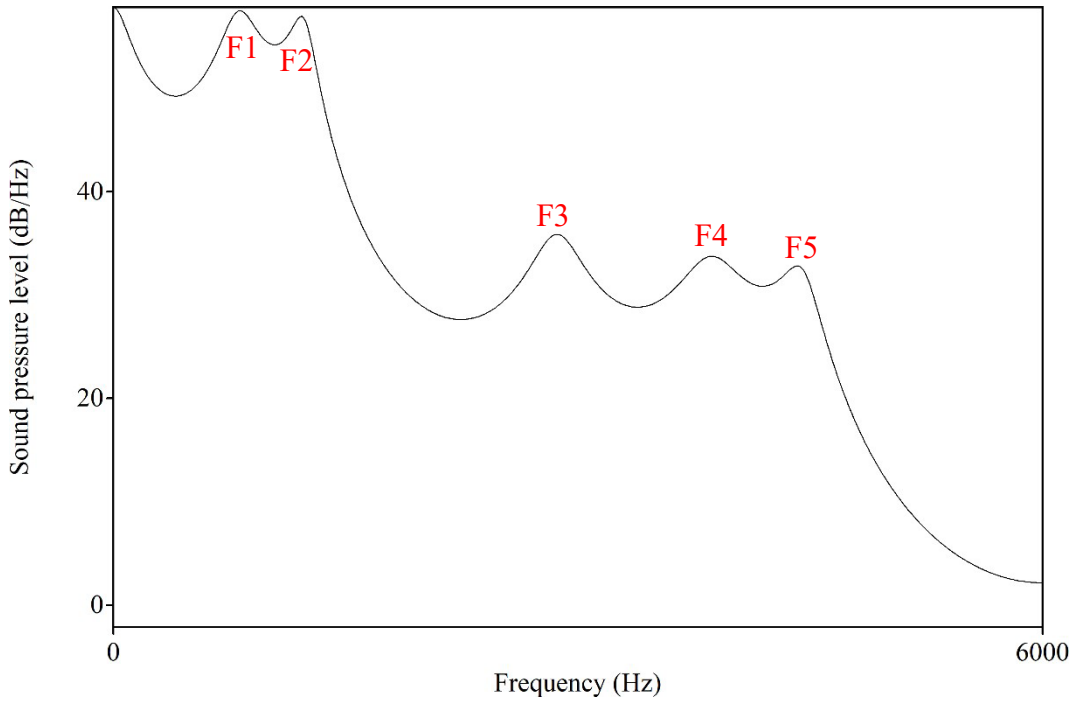


Abbildung 6: LPC-Spektrum vom aufgenommenen [a:], gesprochen von W1, Prediction Order 10
Quelle: Eigene Darstellung

Die ersten fünf Formanten, die bis 6000 Hz zu erwarten waren, sind gut als Peaks sichtbar. Der erste Peak die erste Harmonische. Je höher die Ordnung des Prediktors ist, desto genauer werden die SPL-Werte.

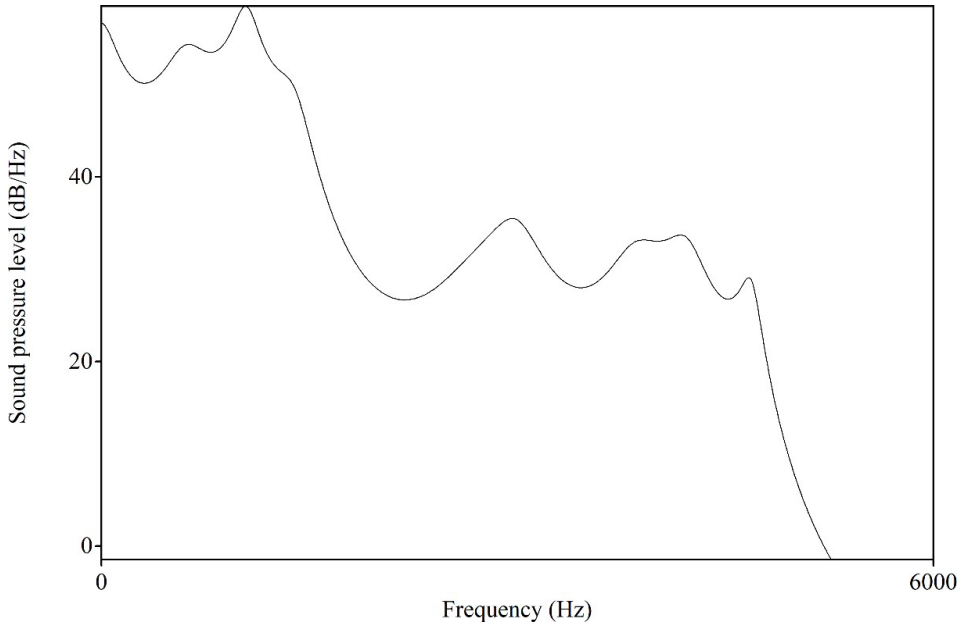


Abbildung 7: LPC-Spektrum vom aufgenommenen [a:], gesprochen von W1, Prediction Order 16
Quelle: Eigene Darstellung

3.3 Schmal- und Breitbandspektrogramm

Ein Spektrogramm wird durch die DFT erstellt. Das Sprachsignal wird zunächst mit einer Fensterfunktion gewichtet. Abhängig von der eingestellten Fenstergröße lassen sich verschiedene phonetische Größen erkennen. Sie wird in Sekunden angegeben. „Höhere“ Werte (lange Fenstergröße) erschaffen ein Schmalbandspektrogramm. Ein Beispiel sieht man in Abbildung 8. Bei diesem sind die Harmonischen gut sichtbar.

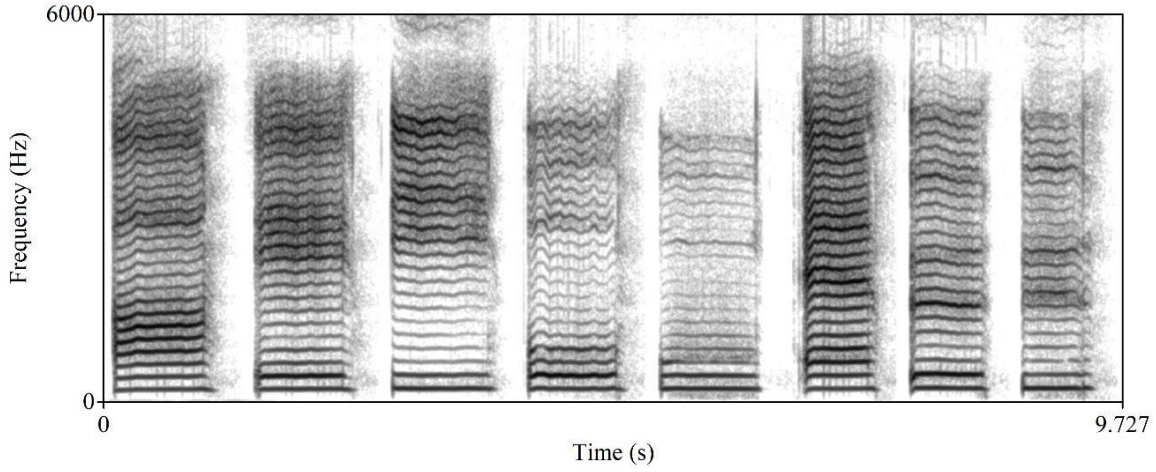


Abbildung 8: Schmalbandspektrogramm von Sprecherin W1 (Fenstergröße: 0,05 s).
Die dunklen Linien zeigen die Harmonischen
Quelle: Eigene Darstellunug

Zur Ermittlung der Formanten sind Schmalbandspektrogramme allerdings nicht geeignet. Dafür bieten sich Breitbandspektrogramme an. In Abbildung 9 wurde eine Fenstergröße von 0,005 s gewählt. Die Formanten sind als dunkel gefärbte Streifen sichtbar. Dort zeichnet sich ein höherer SPL über einen breiteren Frequenzbereich ab im Vergleich zu angrenzenden Bereichen. Wenn man die Werte für F1 und F2 im Kopf hat (oder in der Tabelle 1 nachschaut), dann kann man die Vokale aus dem Breitbandspektrogramm lesen.

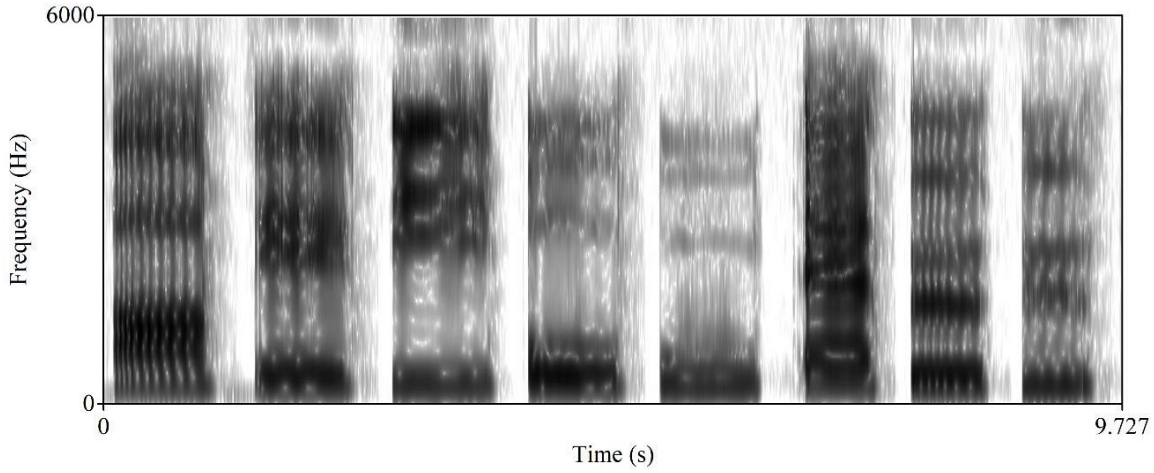


Abbildung 9: Breitbandspektrogramm vom Sprecherin W1 (Fenstergröße: 0,005 s)
Die dunklen Linien zeigen die Formanten
Quelle: Eigene Darstellung

Die beiden Spektrogramme kommen von derselben Person. Nur die Fenstergröße wurde geändert. Von links nach rechts sieht man die folgenden Vokale: [a:], [e:], [i:], [o:], [u:], [ɛ:], [ø:] und [y:].

4. Theorie der digitalen Formantsynthese

In Kapitel 2 wird der Vokaltrakt beschrieben. Es wird erläutert, wie dieser zur Individualität der menschlichen Stimme beiträgt. Aus der Analyse können nun Schlussfolgerungen gezogen werden, die zur Synthese der Stimmen beitragen. Für die digitale Signalverarbeitung kann man folgende Analogien ziehen: Die Glottis (angeregt durch Atemluft) stellt den Oszillator dar, der das Eingangssignal erzeugt. Das Signal wird durch Rachen-, Mund- und gegebenenfalls die Nasenhöhledes Vokaltraktes geschickt und dort moduliert. Den Vokaltrakt einesjeden Menschen kann man sich als Filter vorstellen. Das Ausgangssignal entsteht, sobald die Atemluft die Lippen bzw. Nasenlöcher passiert hat.

Mit dieser getrennten Betrachtung von der Glottis als Quelle und dem Vokaltrakt als Filter befassen sich die Quelle-Filter-Modelle. Die Formantsynthese ist eine Unterkategorie dieses Modells, genauso wie zum Beispiel das Akustische und das Artikulatorische Modell, deren Funktionen grob in Kapitel 1 beschrieben wurden. In der Formantsynthese wird das Wissen genutzt, dass sich bei verschiedenen Vokalen die Formanten F1 und F2 unterscheiden. Untersucht man eine größere Menge an Menschen und vergleicht die Formantfrequenzen in Relation mit den Vokalen, dann ähneln sich deren Frequenzbereiche. Ein Beispiel dafür ist die Tabelle 1. Ziel ist es also, mithilfe von Filtern einen Klang zu entwickeln, der die Sprache eines Menschen verständlich nachahmen kann.

Es gibt drei Parameter, die die Formanten beschreiben: Die Amplitude, Frequenz und die Bandbreite. Manuell können diese Parameter mithilfe eines Spektrums ermittelt werden, welches z.B. mit der DFT/FFT oder LP-Analyse entwickelt wurde. Die so abgelesenen Werte sind häufig fehleranfällig, weswegen in Kapitel 5 Praat bei der Analyse hilft. Die ermittelten Parameter werden für die Resynthese benötigt.

Außerdem darf die Quelle, das Eingangssignal, nicht vernachlässigt werden. Es steht zwar nicht im Fokus, aber ein gewisses Maß an „Natürlichkeit“ sollte die synthetisierte Stimme besitzen. Sonst könnte sie mit einem anderen technischen Gerät verwechselt werden, woraufhin nicht mehr auf Sprache geachtet würde.

4.1 Quelle

Wie in Kapitel 2 bereits erwähnt wurde, erzeugt die Glottis keinereine Sinusschwingung, sondern einen Klang, bei dem die Grundfrequenz mit ihren Oberschwingungen überlappt wird.

In der digitalen Klangsynthesewird oft die Kippschwingung - auch Sägezahnschwingung genannt - genutzt. Sie enthält neben der Grundfrequenz auch alle ihre ganzzahligen Vielfachen, was man im Amplitudenfrequenzgang gut sehen kann. Nimmt man zum Beispiel als Grundfrequenz 150

Hz, dann finden sich bei der Kippschwingung Peaks im Frequenzbereich bei 150 Hz, 300 Hz, 450 Hz usw. (also $f_n = n \cdot 200$ Hz). Diese Schwingung in Verbindung mit Filtern wird bei der subtraktiven Klangsynthese genutzt, um zum Beispiel Musikinstrumente wie elektronische Orgeln oder Streichinstrumente zu erzeugen. Wie der Begriff „subtraktiv“ schon sagt, werden bestimmte Frequenzanteile geschwächt bzw. ausgelöscht werden.

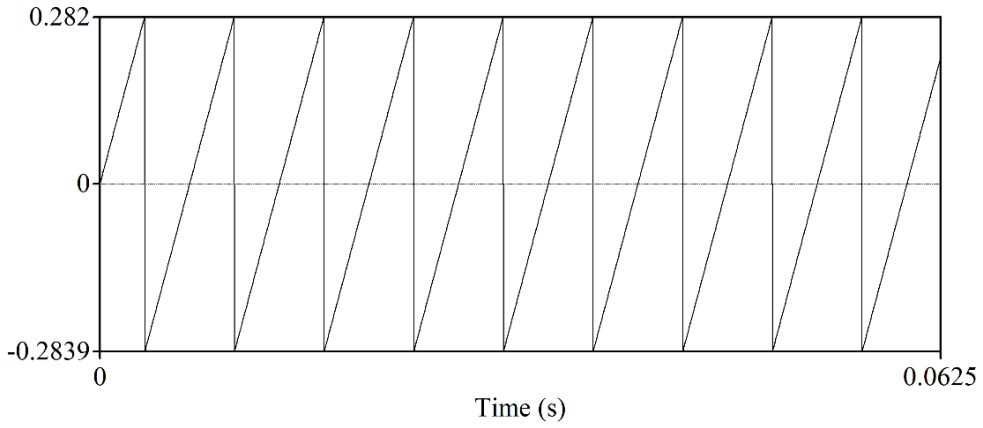


Abbildung 10: Kippschwingung mit 150 Hz
Quelle: Eigene Darstellung

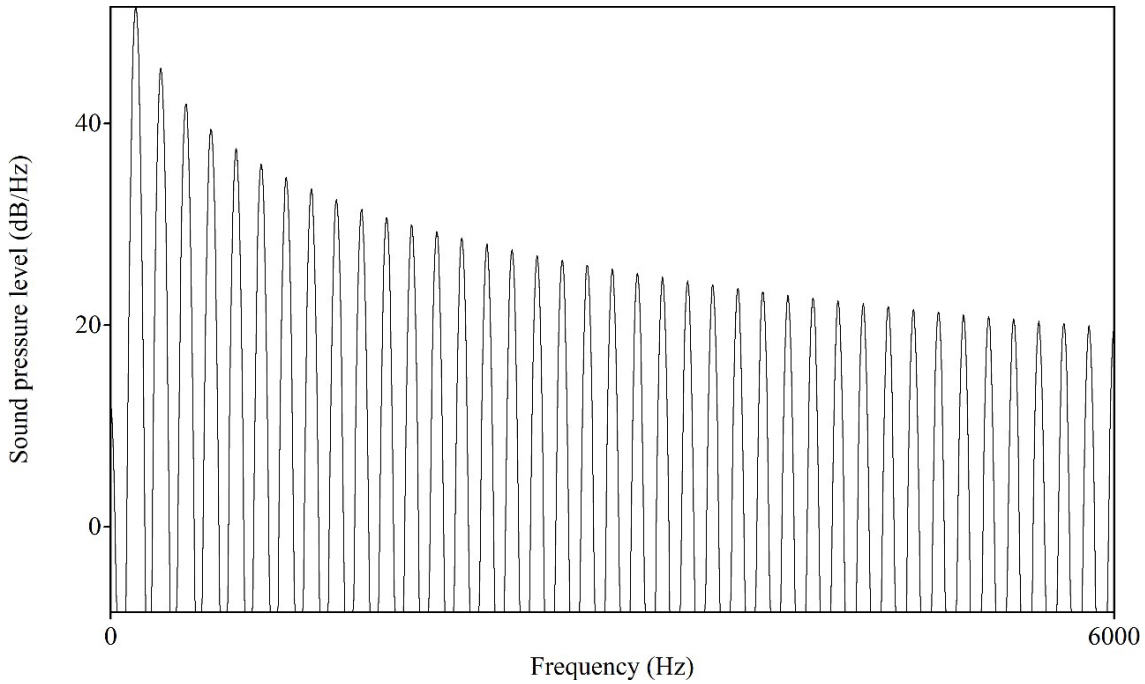


Abbildung 11: Spektrum zur Kippschwingung aus Abbildung 10
Quelle: Eigene Darstellung

Die menschliche Stimme besteht allerdings nicht aus einem einzelnen Rohr gleichen Durchmessers. Der Vokaltrakt ist deutlich komplexer, wodurch der SPL, also die Amplitude, im Vokaltrakt schneller sinkt. Vergleicht man das Spektrum in Abbildung 11 mit den Spektren in Kapitel 3, dann sieht man, dass die Amplituden von der Kippschwingung ab einem Punkt nur

noch langsam sinken. Dadurch zeigt sich, dass die Kippschwingung nicht für die Synthese der menschlichen Stimme geeignet ist.

Es muss eine andere Schwingung her, die den Glottisimpuls darstellt und als Quelle für die Sprachsynthese gewählt werden kann. An dieser haben viele Personen geforscht. Forschende (z.B. Rosenberg und Fant) wollten den SPL über die Zeit abbilden können. Entsprechend wurden einige Modelle entwickelt, die den Glottisimpuls nachstellen wollten. Viele ähneln dem Modell von Rosenberg (1971), das man in Abbildung 2 sehen kann. Allerdings sind bei weiteren Forschungen Schwachstellen aufgetaucht. Das Modell ist gut für stimmhafte Laute. Sobald allerdings eher gehaucht gesprochen wird, dann greift es nicht mehr korrekt. Das sogenannte LF-Modell in Abbildung 12 merzt das aus. Die aus dem Modell hervorgegangene Schwingung wird in Praat als Phonations-Signal benutzt (siehe Abbildung 13), um die menschliche Stimme synthetisieren zu können.

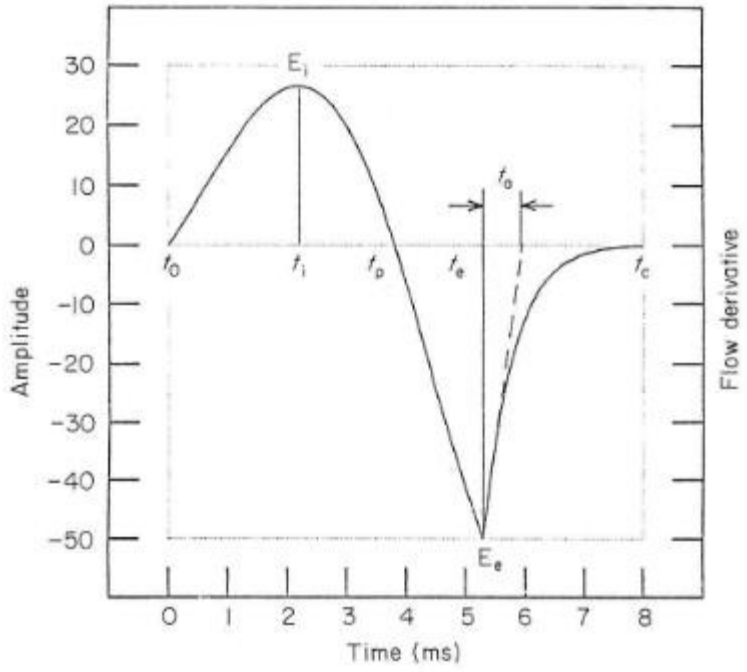


Abbildung 12: LF-Modell
Quelle: Fant, Liljencrants & Lin (1985)

Wie auch das Modell von Rosenberg in Abbildung 2 wird das LF-Modell in mehrere Phasen unterteilt:

$t_0 > t > t_i$: opening phase, E_i ist die maximale Amplitude des Signals bei t_i

$t_i > t > t_e$: closing phase, E_e ist das negative Minimum bei t_e

$t_e > t > t_c$: returning phase

Die Schwingung wiederholt sich periodisch beim Sprechen.

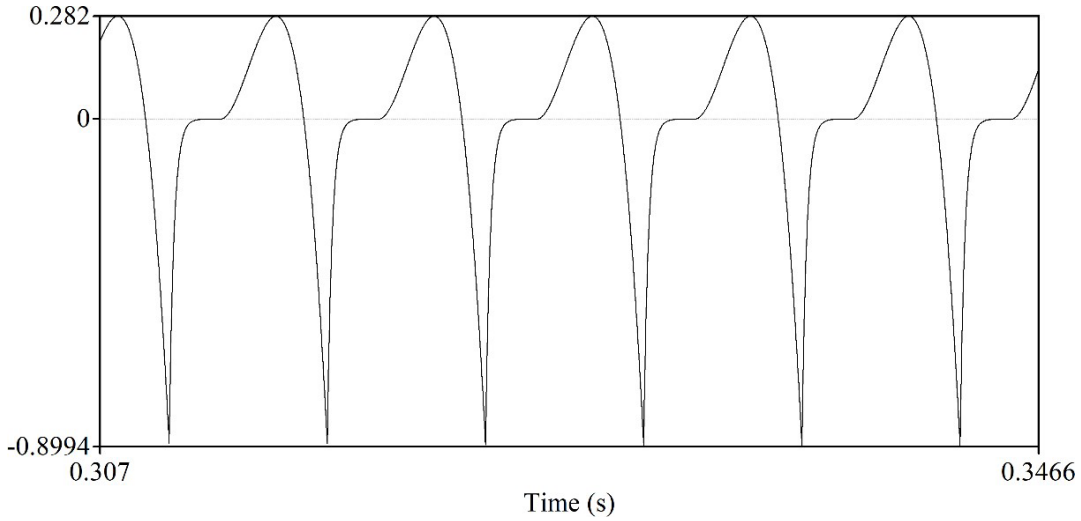


Abbildung 13: Phonations-Signal mit 150 Hz bei Praat
Quelle: Eigene Darstellung

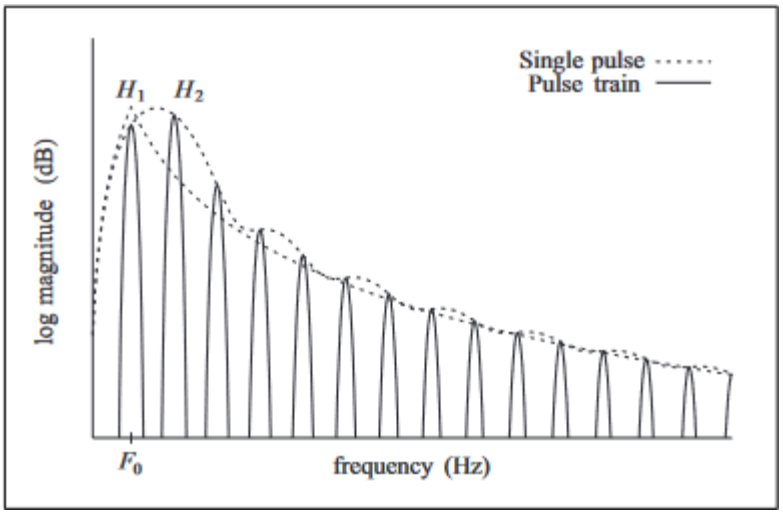


Abbildung 14 Spektrum LF-Modell mit Abbildung der 1. und 2. Harmonischen
Quelle: B. Doval, C. d'Alessandro (2006). The Spectrum of Glottal FlowModels

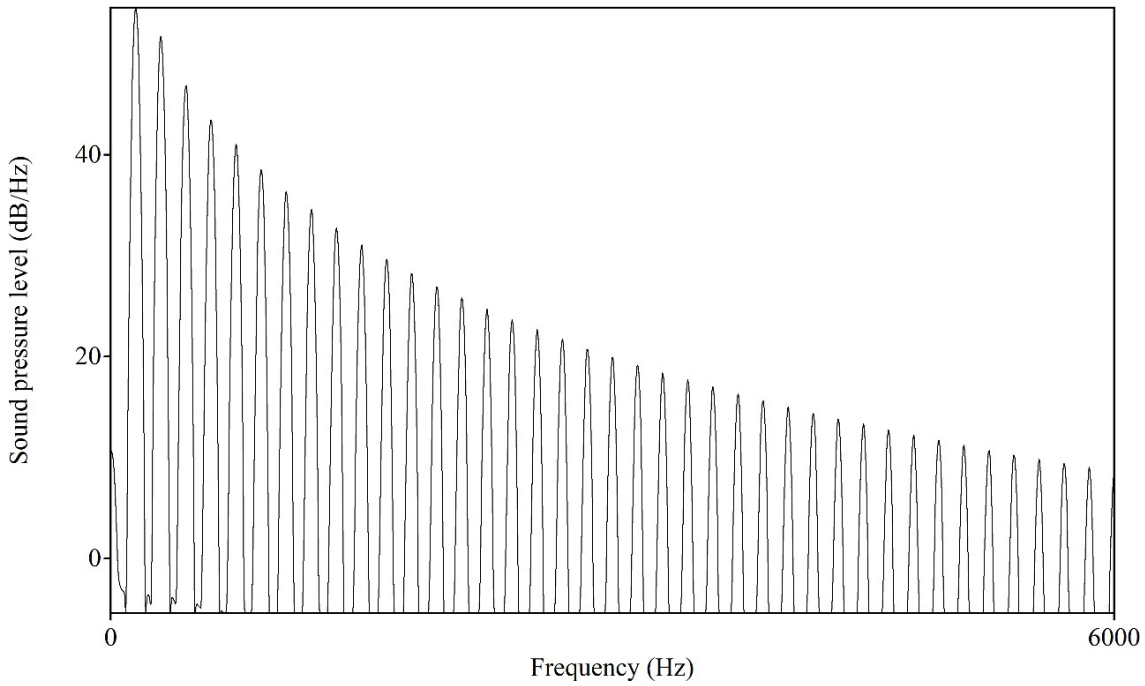


Abbildung 15: Spektrum zum Phonations-Signal in Abbildung 13
Quelle: Eigene Darstellung

4.2 Filter

Das Quellsignal, welches im Analogen von der Glottis kommt, wird durch den Vokaltrakt geformt. Das Signal ist ein Klang, aufbauend auf der Grundfrequenz und ihren harmonischen Schwingungen. Der Vokaltrakt wird im Digitalen durch Filter nachgebildet, welche die Harmonischen des Quellsignals in ihrer Amplitude schwächt oder verstärkt. Formanten entstehen in unmittelbarer Nähe von oder sind lokale Resonanzfrequenzen. Daher rührt die dunkle Färbung im Spektrogramm, die auf einen hohen SPL hinweist.
Für je einen Formanten wird ein digitales Bandpass-Filter benötigt. Bei dem Filter handelt es sich um einen IIR-Filter 2. Ordnung. Die Übertragungsfunktion H(z) des linearen und zeitdiskreten Systems ist in Gleichung… zu sehen. IIR steht für „Infinite Impulse Response“. Dieses Filter nimmt einen Wert vom Ausgangssignal und fügt es beim nächsten Impuls dem Eingangswert zu. Somit ist dieses Filter rekursiv und wird auch Werte ausgeben, wenn das Eingangssignal x(n) = 0 ist.

$$H(z) = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2}} \tag{6}$$

Zur Unterscheidung von Vokalen werden zwei Bandpässe benötigt. Die weiteren Formanten (F3, F4 etc.), die den Klang der sprechenden Person wiedergeben, werden entsprechend mit weiteren Bandpässen nachgebildet. Sie können in Reihe oder parallelgeschaltet werden. Abhängig davon ist, welche Parameter eingestellt werden müssen. Bei der Reihenschaltung müssen nur die Mittenfrequenz (bzw. Formantfrequenz) und die dazugehörige Bandbreite eingestellt werden (siehe Abbildung 16). Bei der Parallelschaltung zusätzlich noch die Amplitude (siehe Abbildung 17).

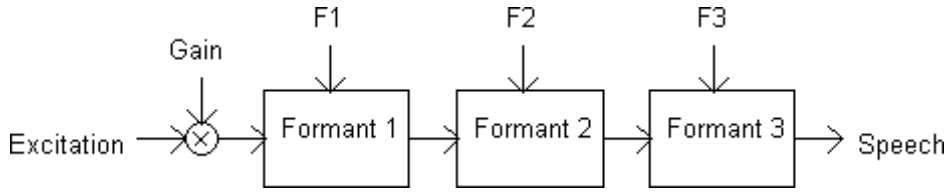


Abbildung: 16 Reihenschaltung von drei Formanten
Quelle: Sami Lemmetty (1999). Review of Speech Synthesis Technology, S. 30

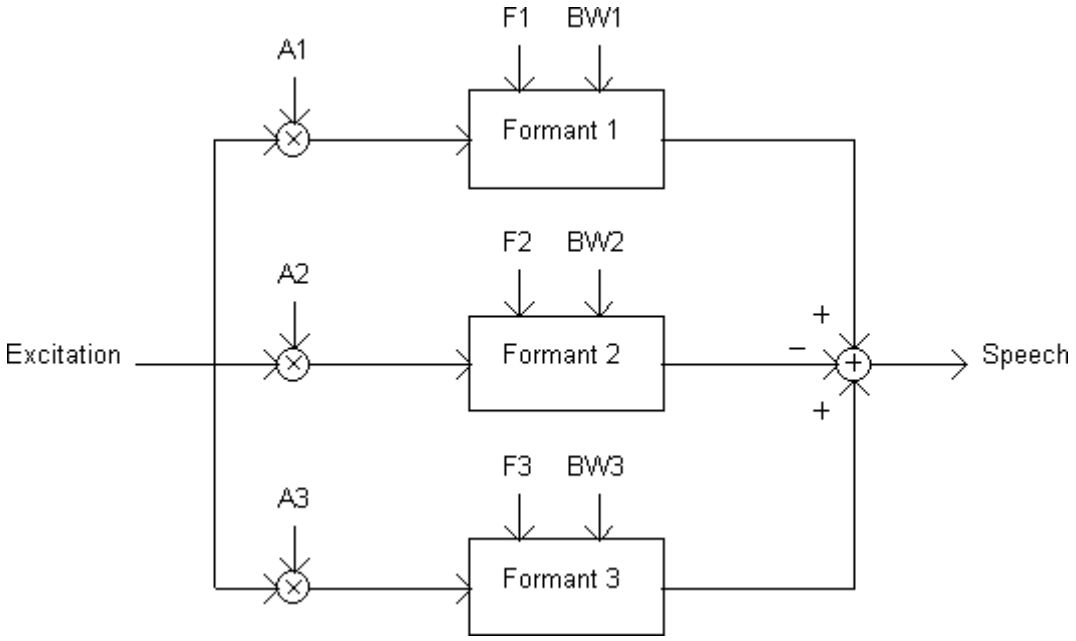


Abbildung 17: Parallelschaltung von drei Formanten
Quelle: Sami Lemmetty (1999). Review of Speech Synthesis Technology, S. 30

5. Experiment: Aufnahme der Stimmen/Synthese

Im folgendem wird ein Experiment beschrieben, um eigene Spektrogramme erstellen zu können, aus welchem dann die Frequenzen der Formanten bestimmt werden sollen. Dies ist keine repräsentative Studie, da bei diesem Versuch inklusive mir nur fünf Personen teilgenommen haben.

In Vokalen sollen die Formanten deutlich zu sehen sein, daher wird eine Auswahl dieser untersucht. Zudem möchte ich auch die Formanten analysieren, die für die Individualität der Stimmen verschiedener Menschen verantwortlich sind und zu denen weniger Werte zu finden sind. Dafür lasse ich Menschen unter gleichen Umständen deutsche Vokale aussprechen, wie in Abschnitt 5.2 erklärt wird. Mit den Stimmsignalen wird ein individuelles Spektrogramm erstellt. Mithilfe von Praat und den Spektrogrammn sollen Formantwerte entnommen werden, die ich für die Stimmensynthese benötige. Es deutet über den zeitlichen Verlauf die Höhe des SPL im Frequenzbereich an. Das menschliche Gehör kann im Idealfall Frequenzen im Bereich von 20 Hz bis 20 kHz wahrnehmen und die menschliche Stimme kann sich über diesen Bereich verteilen, wie in Abbildung 18 zu sehen ist.

Aus den synthetisierten Stimmen werden die Spektrogramme und die Amplituden-Frequenzgänge erstellt. Die aufgenommenen Stimmen werden so auf nicht auditive Weise mit den synthetischen Stimmen verglichen.

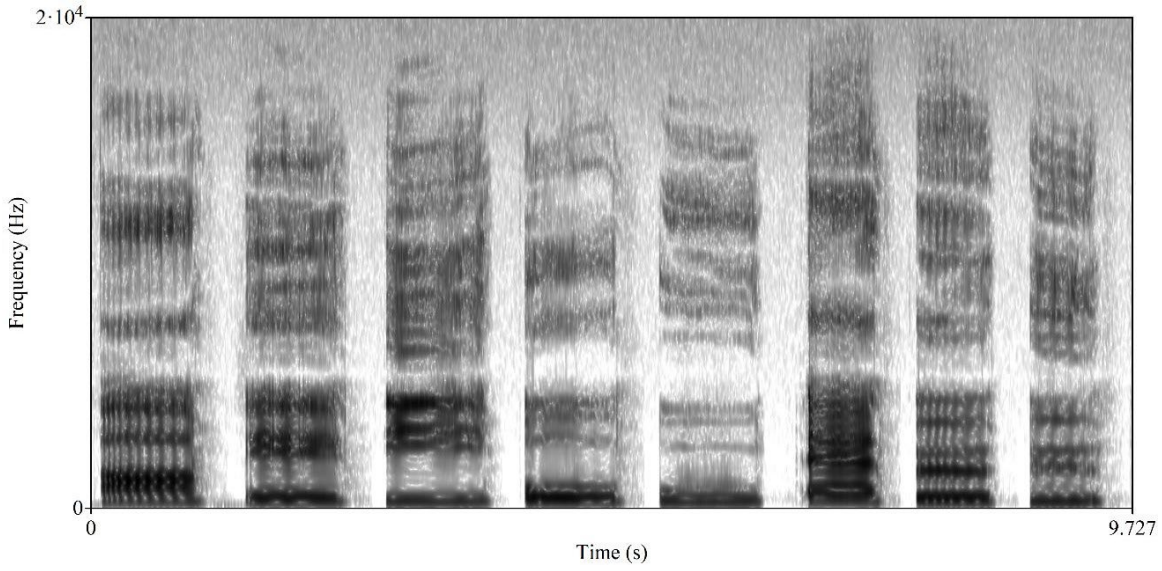


Abbildung 18: Spektrogramm von W1 bis 20 kHz
Von links nach rechts werden folgende Vokale ausgesprochen: [a:], [e:], [i:], [o:], [u:], [ɛ:], [ø:], [y:]
Quelle: Eigene Darstellung

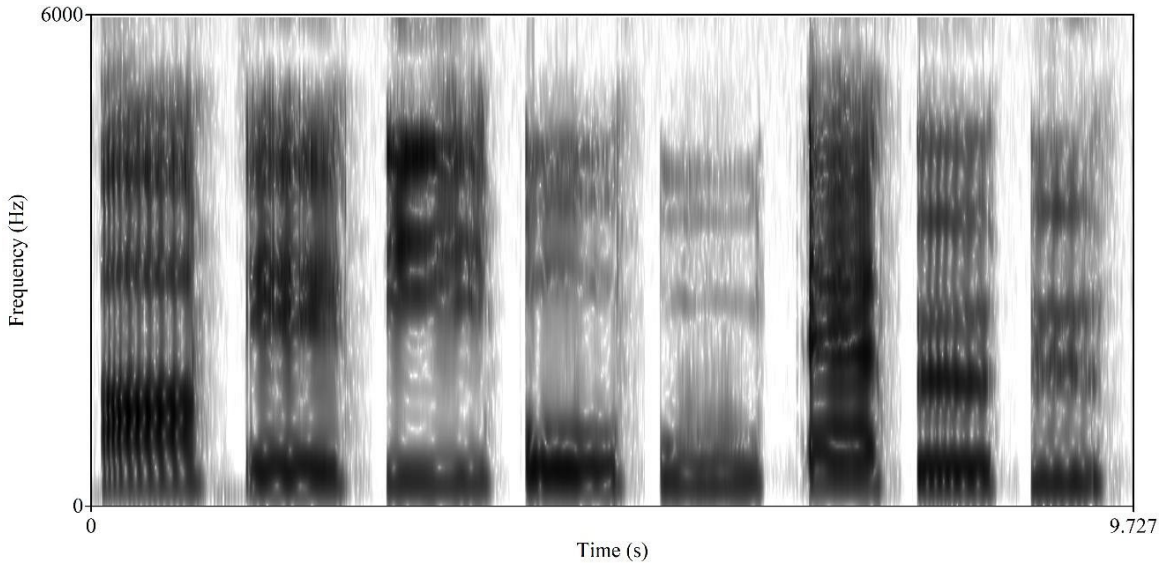


Abbildung 19: Gleiches Spektrogramm wie in Abbildung 18 mit einem Frequenzbereich bis 6 kHz
Quelle: Eigene Darstellung

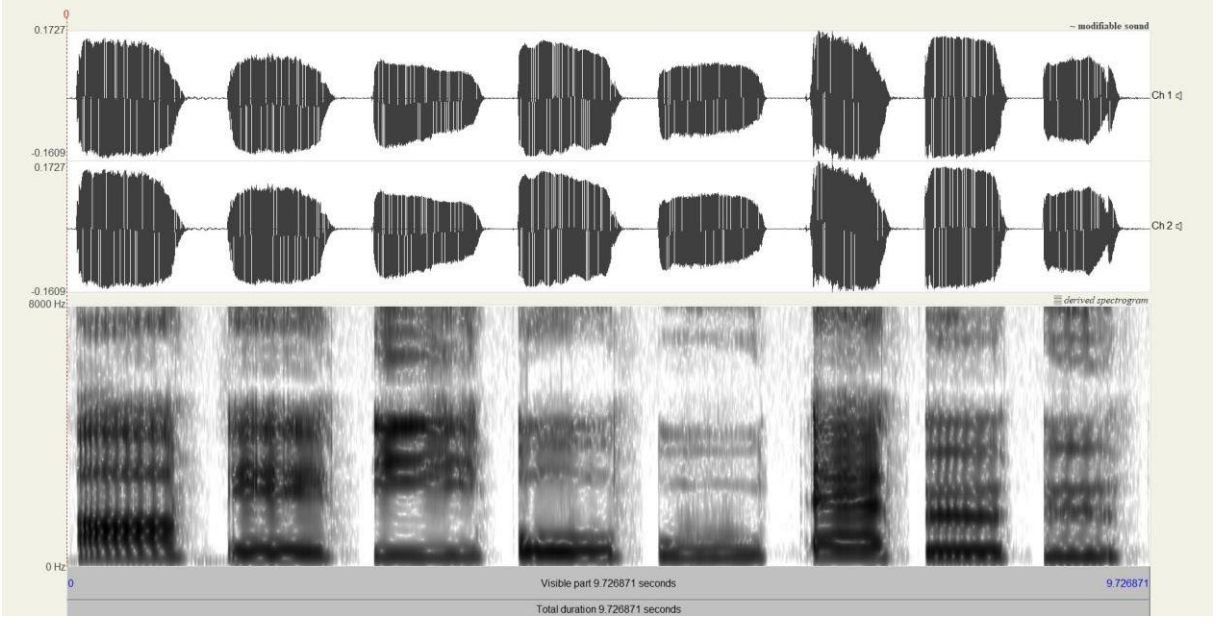


Abbildung 20: Im oberen Bereich ist die Wellenform einer Stereospur zu sehen, im unteren Bereich sieht man das zugehörige Spektrogramm
Quelle: Eigene Darstellung

5.1 Material

5.1.2 Hardware

Dynamisches Mikrofon: Sennheiser MD421
Audio-Mischpult: DHD.audio SX2

5.1.3 Software

Zur Aufnahme: Audacity (Version 3.4.2)

Mit Audacity werden einige Manipulationswerkzeuge geboten. Falls sich eine Person bei der Aufnahme zum Beispiel verspricht, kann es im Nachhinein einfach ausgeschnitten werden. Die Audio-Dateien werden als unkomprimierte WAV-Datei mit einer Abtastfrequenz von 44,1 kHz gespeichert. Audacity besitzt die Möglichkeit Spektro-gramme zu erstellen. Allerdings müssten die Formanten komplett manuell daraus gelesen werden. Es kann hier Personen schwerfallen, die genauen Werte der Formanten zu erkennen. Daher wird zur Formantanalyse das Programm Praat verwendet.

Zur Analyse und Synthese: Praat (Version 6.3.09)

Praat kann wie Audacity kostenlos runtergeladen werden und ist speziell dazu gedacht, Audiodateien auf verschiedene phonetische Parameter zu analysieren, wozu unter anderem die Frequenzen zu den einzelnen Formanten, sowie deren Bandbreiten zählen. Sie sind für die Synthese notwendig.
Ebenso können Audio-Dateien auf nicht weniger vielfältige Art synthetisiert werden.

Für viele Spektrogramme und Spektren wurde Praat verwendet.

5.2 Aufnahme der Stimmen

In der folgenden Tabelle sieht man in der ersten Zeile die Vokale, die die Personen aussprechen sollten, nach der Lautschriftform von IPA. In der zweiten die meistens zugewiesenen deutschen Vokalbuchstaben/Phoneme und in der dritten Zeile deutsche Wortbeispiele.

Tabelle 2: Beispiele von Vokalen an deutschen Wörtern

[a:]	[e:]	[i:]	[o:]	[u:]	[ɛ:]	[øː]	[y:]
/a/	/e/	/i/	/o/	/u/	/ä/	/ö/	/ü/
M**a**len	M**eh**l	M**ie**te	M**oo**r	M**u**ße	M**ä**dchen	M**öh**re	M**üh**le

Quelle: Eigene Darstellung

Dass die Vokale in der Lautschrift aufgeschrieben wurden, liegt an der Varietät, in der die Vokalbuchstaben ausgesprochen werden können. Denn Vokalbuchstaben sind nicht das Gleiche wie Vokale (wie bereits in Abschnitt 2.5 erwähnt).

Damit möglichst gleiche Randbedingungen geschaffen sind, habe ich die Personen an einem Tag in demselben Studio mit demselben Mikrofon einsprechen lassen. Über das Audiomischpult habe ich sie vor ihren Aufnahmen so eingepegelt, dass die Stimmen nicht übersteuern und es dadurch zum Rauschen kommt, welches das Amplituden-Frequenzspektrum verfälschen könnte. Was sie redeten, war beim Einpegeln dabei egal.

Nachdem sie eingepegelt wurden, wurde ihnen kurz vorgezeigt, wie die Vokale ausgesprochen werden sollen: Möglichst monoton, in gleichbleibender Lautstärke und zeitlich etwas gestreckt in die Mitte des Mikrofons. So sieht man die Formanten im Spektrogramm als grade Linien. Bei kleinen Unstetigkeiten können dann an quasistationären Zeitpunkten Werte entnommen werden, die in einem kürzeren Zeitraum innerhalb des Vokals gleichbleibend sind. Die monotone Aussprache soll als Tonlage bei der Synthese genutzt werden.

5.3 Die Formantanalyse mit Praat

Für die Auswertung und Entnahmeder Formantfrequenzen wird das Programm Praat genutzt. Die Formanten und die Grundfrequenz können durch Praat direkt im Spektrogrammangezeigt werden (siehe Abbildung 26) und in einem separaten Fenster. Das Spektrogramm wird bei einem Ausschnitt des Signals von maximal zehn Sekunden angezeigt. Ist das Sprachsignal länger, dann kann man gegebenenfalls in die Datei zoomen.

Die Frequenz- und Zeitskalierungen der Achsen sind linear. Die Ordinate zeigt dabei den Verlauf des Audiosignals über die Zeit und die Abzisse den eingestellten Frequenz-bereich. Die Graustufen deuten auf die Höhe der Amplituden/SPL in Abhängigkeit zur Frequenz. Die Skala in diesem Spektrogramm geht von weiss (SPL = 0 dB) bis schwarz (hohes SPL). Je dunkler die Grautöne werden, desto höher ist der SPL im Frequenzbereich.

Um die richtigen Formanten ermitteln zu können, müssen diverse Einstellungen von Mensch zu Mensch individuell angepasst werden. Dafür wird hier die sogenanntehalbautomatische Methode angewendet. Auf diese wird im Verlaufe des Abschnittes eingegangen.

5.3.1 Spektrogramm-Einstellungen

Das Sprachsignal wird mithilfe der FFT in den Frequenzbereich überführt. Dafür wird es mit einer Fensterfunktion multipliziert. Bei Praat wird in den Standardeinstellungen das Gauß-Fenster mit einer Fensterlänge von 5 ms benutzt. Dieses startet bei einer Amplitude von 0, steigert sich bis zum Mittel (bei dieser Funktion bei 2,5 ms) und sinkt dann wieder auf 0. Je niedriger die Länge ist, desto verschwommener werden die Frequenzbereiche dargestellt, wodurch eine flächenweise Abdunklung sichtbar ist. Es wird als Breitbandspektrogramm bezeichnet und hier sind die Formanten gut zu erkennen. Wenn man die Harmonischen sehen will, muss man höhere Werte (z.B. 20 ms) nehmen. Dann erhält man ein Schmalband-Spektrogramm, welcheszur Bestimmung von Formanten allerdings nicht geeignet ist.
Mit den Standardeinstellungen in Abbildung 21 wird ein Breitbandspektrogramm mit einem Frequenzbereich von 0 Hz bis 7000 Hz erstellt werden.

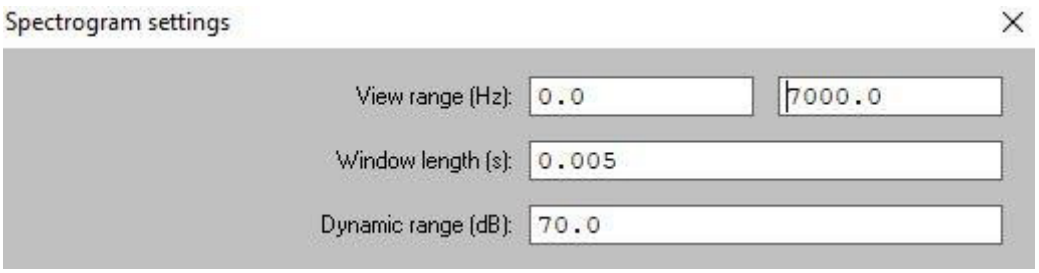


Abbildung 21: Standardeinstellungen für das Spektrogramm bei Praat
Quelle: Eigene Darstellung

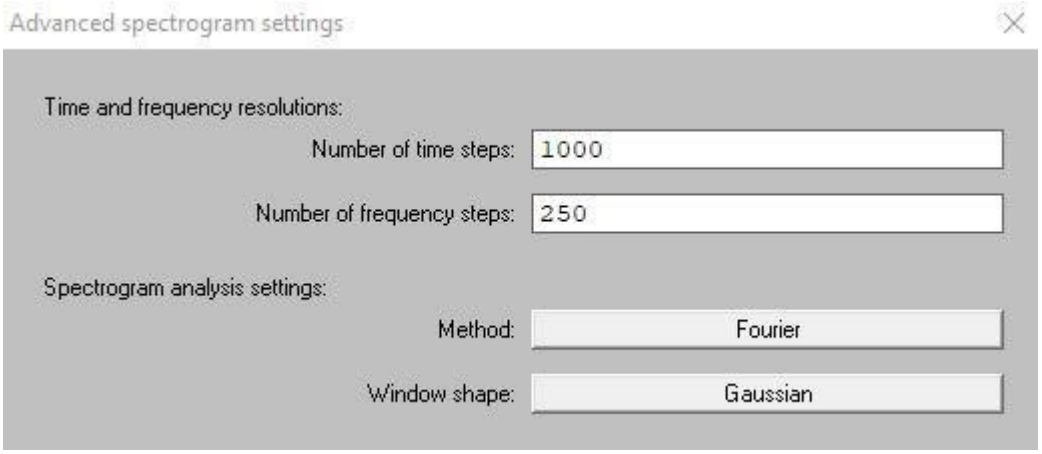


Abbildung 22: Erweiterte Einstellungen für das Spektrogramm bei Praat
Quelle: Eigene Darstellung.

5.3.2 Formantanalyse-Einstellungen

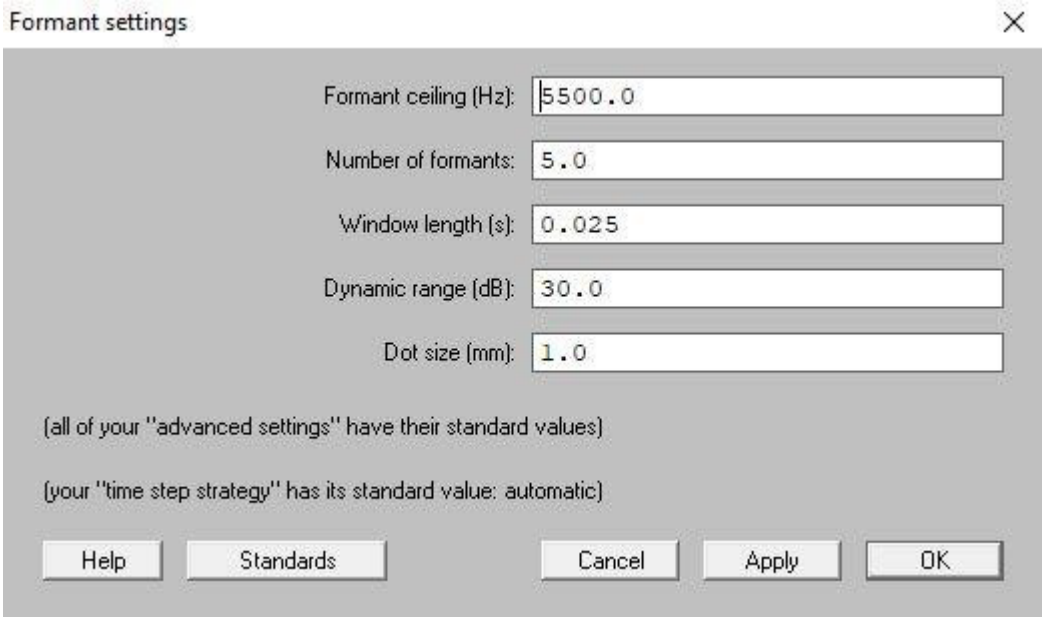


Abbildung 23: Standardeinstellungen zur Formantanalyse bei Praat
Quelle: Eigene Darstellung

„Formant ceiling" ist die maximale Frequenz, bis zu welcher Praat die Formanten bestimmen soll. „Number of formants" ist die Anzahl an Formanten, die man in diesem Bereich erwartet und nach denen daraufhin gesucht werden soll. Die oben gezeigte Obergrenze von 5500 Hz soll sich auf weibliche Stimmen beziehen. In Praat wird zusätzlich eine Grenze von 5000 Hz für Männer empfohlen[8]. Das liegt an der unter-schiedlichen Länge des Vokaltraktes. Dieser ist bei Männern länger als bei Frauen. Je länger der Vokaltrakt ist, desto tiefer liegt der Frequenzbereich, in dem die Formanten zu erwarten sind und desto dunkler klingen die Stimmen.
Wegen dieser Unterschiede ist es wichtig, dass die Werte von Person zu Person angepasst werden. Praat nutzt für die Formantanalyse die LPC-Analyse. Da diese Analyseform mit prädizierten Werten arbeitet, können ungenaue Einstellungen zu Fehlern führen. Aus diesem Grund habe ich mich für die sogenannte halbautomatische Methode entschieden. Sie ist neben der automatischen und der manuellen Methode[9], eine weitere Methode zur Bestimmung von Formanten. Bei dieser werden die vom Programm ermittelten Werte mit dem Spektrogramm verglichen und die Einstellungen gegebenenfalls angepasst. Da ein Formant im Spektrogramm als dunkle Linie erkennbar ist, kann man diese meistens gut abzählen. Idealerweise sollte man laut Jörg Meyer (2022) die Obergrenze zwischen den fünften und sechsten Formanten (F5 und F6) setzen, da sich

[8] Laut „LongSound help" in Praat
[9] Die automatische Methode richtet sich nach den Voreinstellungen, bei der manuellen Methode liest man die Formanten ohne die Formantanalyse aus dem Spektrogramm ab.

da ein „Tal“ wiederfindet[10]. Das ist ein Frequenzbereich, in dem die Amplituden flach sind im Vergleich zu benachbarten Bereichen. Zwischen diesen Formanten gibt es häufig einen größeren Abstand voneinander und durch die geringen Amplituden ist der Bereich sehr hell. So ist es möglich einen Frequenzwert zwischen F5 und F6 zu schätzen und diesen als Obergrenze auszuwählen. In Abbildung 18 kann man bei ca. 5500 Hz so ein Tal sehen.
Da nun bis zu jener Obergrenze fünf Formanten sichtbar sein sollen, kann „Number of formants“ auf 5 eingestellt bleiben. Wenn man noch mehr Formanten entnehmen will, muss sowohl eine höhere „Formant Ceiling“-Grenze gesetzt als auch die Anzahl der Formanten geändert werden. Das Programm wird sonst entweder zu wenig oder, wenn der „Number of Formants“-Wert zu hoch angesetzt ist, zu viele „Formanten“ finden.

Nachdem die Einstellungen gut eingerichtet sind, gibt es die Möglichkeit, sich direkt die Frequenzwerte der ersten vier Formanten auf einmal an einem ausgewählten Zeitpunkt in einem separaten Fenster anzeigen zu lassen. Auch der Amplituden-Frequenzverlauf lässt sich in einem neuen Praat-Objekt darstellen. Es sollte allerdings darauf geachtet werden, dass man sich einen quasistationären Bereich aussucht, um aussagekräftige Formanten und ihre zugehörigen Bandbreiten zu finden.

Nach Entnahme der Formantwerte wird mithilfe von Praat eine Formantsynthese durchgeführt. Dieser Teil wird in Abschnitt 5.5 genauer erläutert.

5.3.3 Zu Formanten zugehörige Bandbreiten

Was für die Synthese nicht vernachlässigt werden darf, ist die Bandbreite der Formanten. Denn, wie es in Abschnitt 2.4 erwähnt ist, handelt es sich bei Formanten um Frequenzbereiche. Die Formantfrequenzen, die von Praat ausgegeben werden, zeigen lediglich den Peak dieses Bereiches an. Wenn das Spektrum um den Formanten stark fällt, dann ist die Bandbreite gering. Umgekehrt gilt das für flach verlaufende Spektren, dass die Bandbreite hoch ist. Ist die Bandbreite sehr hoch, dann sollte geprüft werden, ob an der Stelle wirklich ein Formant ist. Für solche Werte ist die halbautomatische Methode wichtig, denn da kann man abschätzen, ob an jener Stelle im Spektrogramm eine dunklere Linie zu sehen ist.

[10] Jörg Mayer (2022). *Phonetische Analysen mit Praat. Ein Handbuch für Ein- und Umsteiger, S. 64*

5.4 Vergleich von vokalspezifischen Formanten

Die vokalspezifischen Formanten sind F1 und F2, also die zwei niedrigsten Formanten eines Phons. Die Formantfrequenzen sehen bei den Personen wie folgt aus:

Tabelle 3: Formantfrequenzen zu F1 und F2, geordnet nach Vokalen und Sprechenden

		[a:]	[e:]	[i:]	[o:]	[u:]	[ɛ:]	[øː]	[y:]
F1 [Hz]	W1	813,7	412,7	265,5	408,6	297,8	705,2	418,1	233,9
	W2	870,6	473,2	274,3	479,9	312,5	760,4	472,3	316,0
	W3	720,8	354,1	335,9	358,2	357,7	637,2	357,3	307,6
	M1	638,2	289,4	224,6	349,1	221,1	530,6	320,3	256,5
	M2	758,8	297,0	242,8	320,8		557,0	322,2	252,2
F2 [Hz]	W1	1220,4	2285,4	2559,2	799,3	699,9	1867,8	1512,1	1671,6
	W2	1236,8	2392,0	2675,8	752,4	789,9	2093,4	1535,9	1452,2
	W3	1308,7	2504,1	2578,5	746,9	799,5	2100,7	1565,6	1885,4
	M1	1068,2	2091,8	2150,9	671,3	613,1	1653,7	1550,8	1652,9
	M2	1279,2	2107,3	2059,3	589,4		1860,7	1429,0	1585,5

Quelle: Eigene Darstellung

In Tabelle 3 sieht man die ersten zwei Formanten der Vokale [a:], [e:], [i:], [o:], [u:], [ɛ:], [øː] und [y:], eingesprochen von den fünf Personen. Die Ellipsen in Abbildung 24 zeigen den Bereich auf, in dem sich diese zentrieren.

„W“ ist hier die Abkürzung für weiblich, also eine Sprecherin. „M“ steht entsprechend für männlich, also einen Sprecher. Die Zahlen hinter den Buchstaben dienen lediglich zur Unterscheidung der weiblichen und männlichen Personen. Auf diesen Werten wird eine Synthese aufgebaut, die die Stimmen der Personen bei den genannten Vokalen imitiert.

Wie man an der Tabelle 3 sehen kann, konnte bei M2 kein Wert für [u:] festgestellt werden. Darauf wird im Kapitel 6 eingegangen.

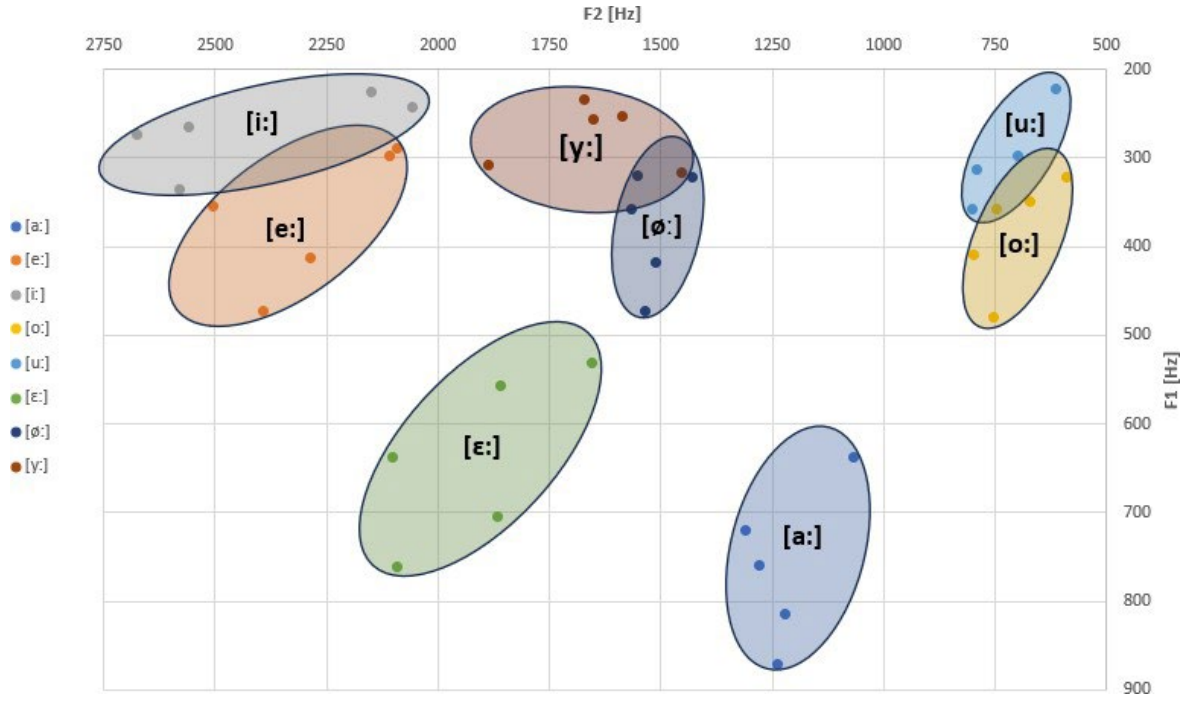


Abbildung 24: Frequenzbereiche, in denen sich die vokalspezifischen Formanten F1 und F2 konzentrieren
Quelle: Eigene Darstellung

5.5 Beispiel: Formanten und zugehörige Bandbreiten einer einzigen Person

Im Abschnitt 5.4.1 wurden nur die ersten zwei Formanten aller Personen betrachtet. Da für eine personenspezifische Formantsynthese allerdings mehr Formanten (ab F3) gebraucht werden, wurden entsprechend mehr Werte aufgenommen. Im Folgenden gibt es ein Beispiel einer Tabelle von einer der sprechenden Personen. F0 ist, wie bereits erwähnt, kein Formant, aber dennoch notwendig für das Hörempfinden und für die spätere Synthese.

Tabelle 4: Formanten von Sprecherin W1, links sieht man die Vokale.

W1	F0 [Hz]	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]	F5 [Hz]
[a:]	193,8	813,692	1220,408	2873,533	4016,861	4531,12
[e:]	201,8	412,666	2285,371	2833,861	4106,675	4479,984
[i:]	207,9	265,496	2559,219	3191,599	3993,711	4339,61
[o:]	204,2	408,643	799,335	2898,648	3816,631	4320,135
[u:]	207,3	297,815	699,941	2468,682	3538,862	4102,591
[ɛ:]	205,7	705,164	1867,784	2720,973	3812,146	4374,86
[ø:]	214,6	418,055	1512,142	2438,378	3523,1	4327,698
[y:]	213,4	233,9	1671,555	2315,925	3629,824	4233,844

Quelle: Eigene Darstellung

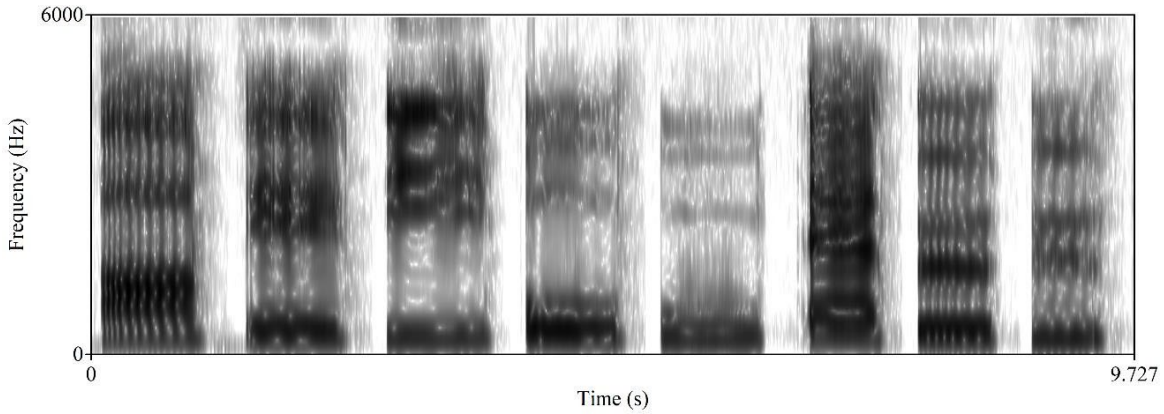


Abbildung 25: Zur Tabelle 4 zugehöriges Spektrogramm ohne Formanten
Quelle: Eigene Darstellung

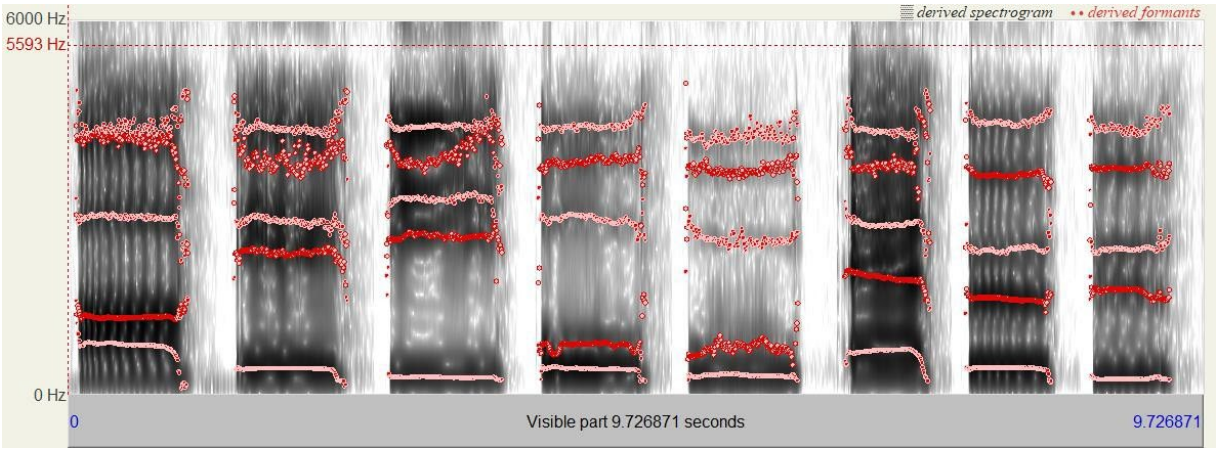


Abbildung 26: Zur Tabelle 4 zugehöriges Spektrogramm mit Formantenanzeige
Quelle: Eigene Darstellung

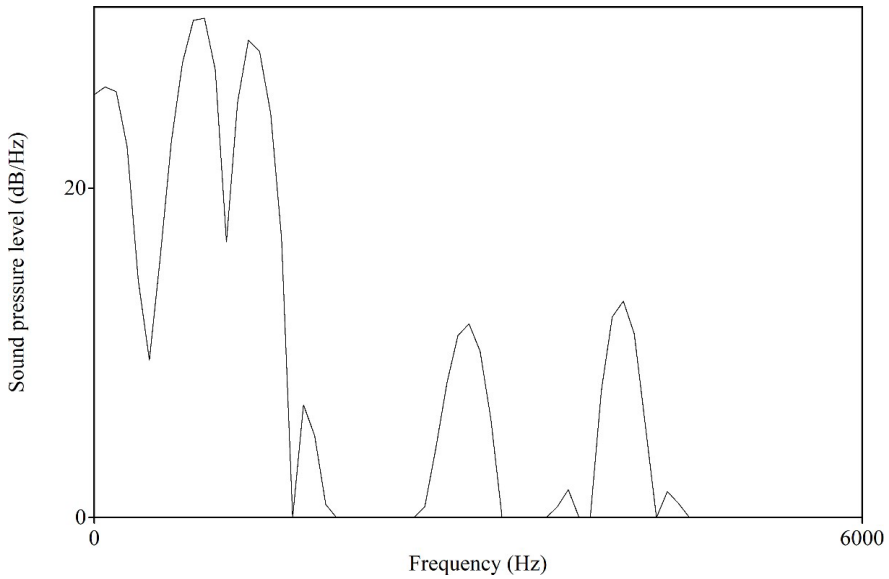


Abbildung 27: FFT-Spektrum von [a:] von der Sprecherin W1 gewichtet mit dem Gauß-Fenster, Fensterlänge 0,005 s
Quelle: Eigene Darstellung

5.6 Digitale Formantsynthese mit Praat

In den vorherigen Abschnitten wurde analysiert, wie man für die Synthese aus Praat bzw. den Spektrogrammen Formanten und Bandbreiten ablesen kann. Die Werte können aus der Tabelle 4[11] entnommen werden.

Es gibt diverse Programme, mit denen Sprache manipuliert und/oder synthetisiert werden kann. Zu diesen zählt auch Praat. Da mit diesem Programm bereits die Analyse der aufgenommenen Stimmen durchgeführt wurde und die synthetisierten Stimmen ebenfalls damit analysiert werden, wird die Synthese nun auch über Praat laufen.

Praat arbeitet mit Objekten, die man zusammenknüpfen muss, um am Ende, wie im Fall dieser Arbeit, eine Sound-Datei zu erstellen. Für synthetische Stimmen benötigt man zum einen ein FormantGrid-Objekt und ein PitchTier-Objekt. In FormantGrid stellt man die Formanten ein. Dieses Objekt wird als Filter dienen. Bei PitchTier erstellt man die Grundfrequenz. Die Änderungen in Objekten werden direkt übernommen. Ein Zwischen-speichern ist nicht nötig. Was genau die Einstellungen tun und wie letzten Endes eine synthetische Stimme entsteht, wird in den folgenden Abschnitten erklärt.

[11] Die Tabellen Spektrogramme der anderen Sprechenden finden sich im Anhang

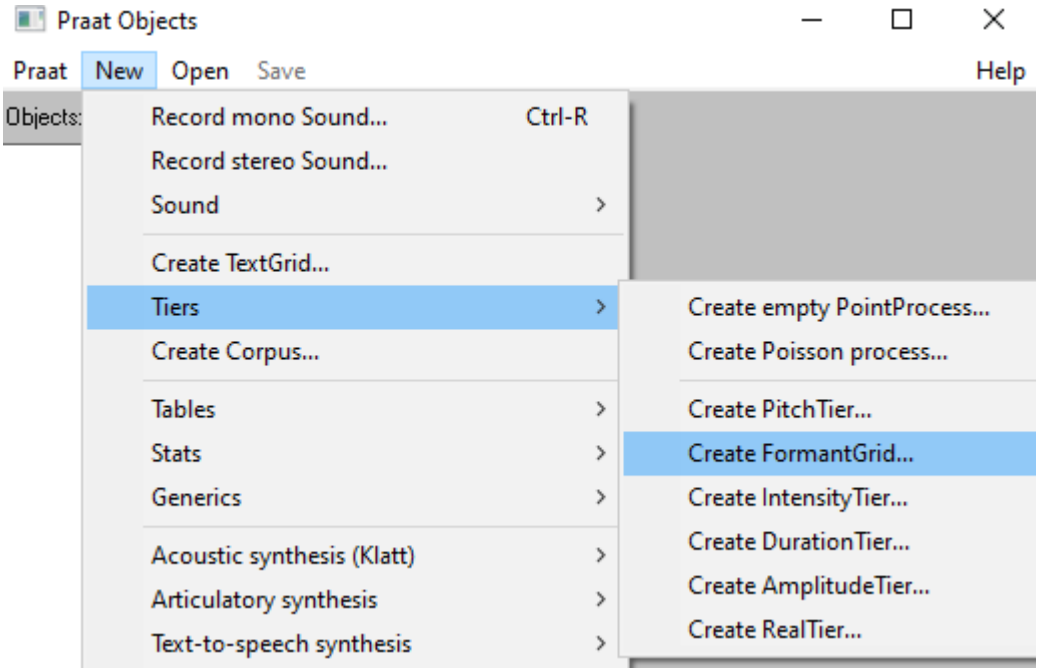


Abbildung 28: Objektauswahl in Praat
Quelle: Eigene Darstellung

5.6.1 Formanteinstellung (FormantGrid)

Hier können die Formanten und die dazugehörigen Bandbreiten eingestellt werden. Allerdings ist durch das FormantGrid allein keine Sound-Datei erstellbar. Es dient als Filter. Erst in Verbindung mit dem PitchTier-Objekt, welches in Abschnitt 5.6.2 genauer erläutert wird, kann eine Sound-Datei erstellt werden.
Ein neues FormantGrid-Objekt wird erstellt, indemman wie in Abbildung 28 gezeigt ist, vorgeht. Also:

New > Tiers > Create FormantGrid…

Daraufhin öffnet sich ein Einstellungsfenster, welches in Abbildung 29 zu sehen ist. In der Standardeinstellung wird ein Zeitfenster von einer Sekunde erstellt. In diesem werden zehn Formanten generiert. Der erste Formant fängt bei 550 Hz an. Jeder weitere wird in einem Abstand von 1100 Hz zum vorherigen Formanten gesetzt, also bei 1650 Hz, 2750 Hz usw. Der erste Formant bekommt zudem die Bandbreite 60 Hz zugeschrieben. Alle weiteren Formanten werden eine um je 50 Hz breitere Bandbreite bekommen. Sobald man das Objekt öffnet, können die Formanten und Bandbreiten nach Belieben verschoben werden.

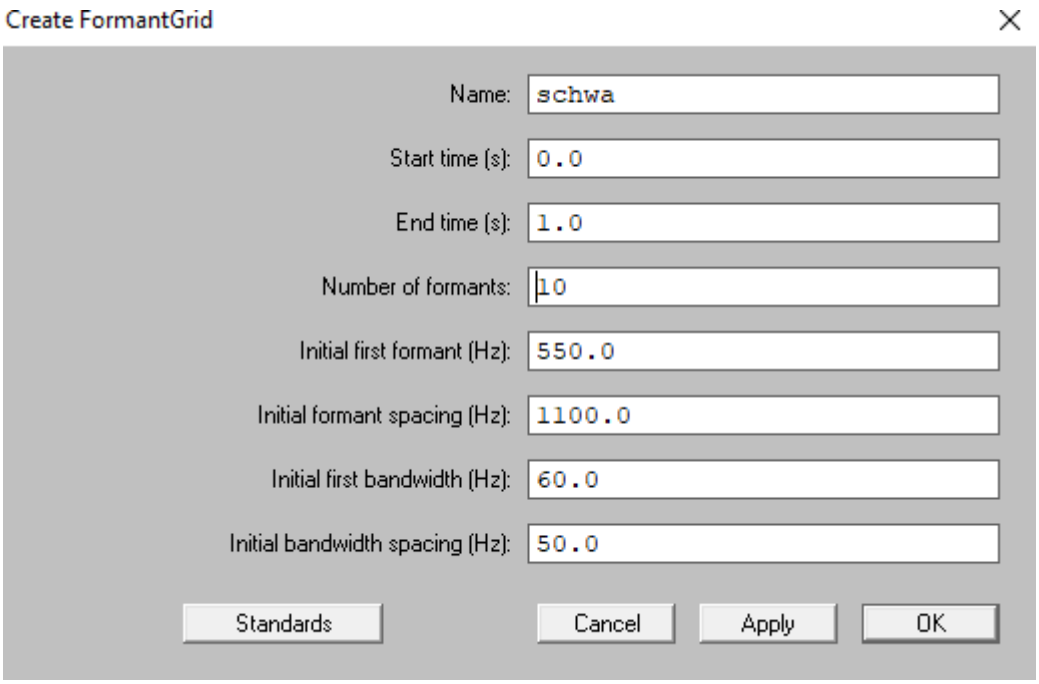


Abbildung 29: Standardeinstellungen für FormantGrid
Quelle: Eigene Darstellung

Die Daten aus der Formantanalyse werden hier nun genutzt. Die gemessenen Formanten und Bandbreiten von jeder Person und jedem Vokal werden eingestellt und im weiteren Verlauf mit dem PitchTier-Objekt zu einer Audiodatei verbunden.

5.6.2 Tonhöheneinstellung (PitchTier)

In diesem Objekt wird allein die Tonhöhe eingestellt. Bei den Personen, die für mich die Vokale eingesprochen haben, lag diese zwischen 101,3 Hz (bei M2) und 267,3 Hz (bei W3). Hier geht man wie folgt vor:

New > Tiers > Creat PitchTier…

Die Tonhöhe kann über einen vorher festgelegten Zeitbereich variiert werden. In der Voreinstellung beträgt dieser eine Sekunde. Wenn sich die Tonhöhe innerhalb dieses Bereiches ändern soll - weil man z.B. eine singende Person nachstellen möchte -, dann kann man das durch Hinzufügen von Punkten erreichen. In Abbildung 30 ist ein Beispiel aufgezeigt, welches nicht zur Synthese genutzt wird. Da die Sprecherinnen und Sprecher die Vokale monoton aussprechen sollten, wird auch nur eine den Personen und Vokalen entsprechende Grundfrequenz F0 eingestellt.

PitchTier > Add point at…

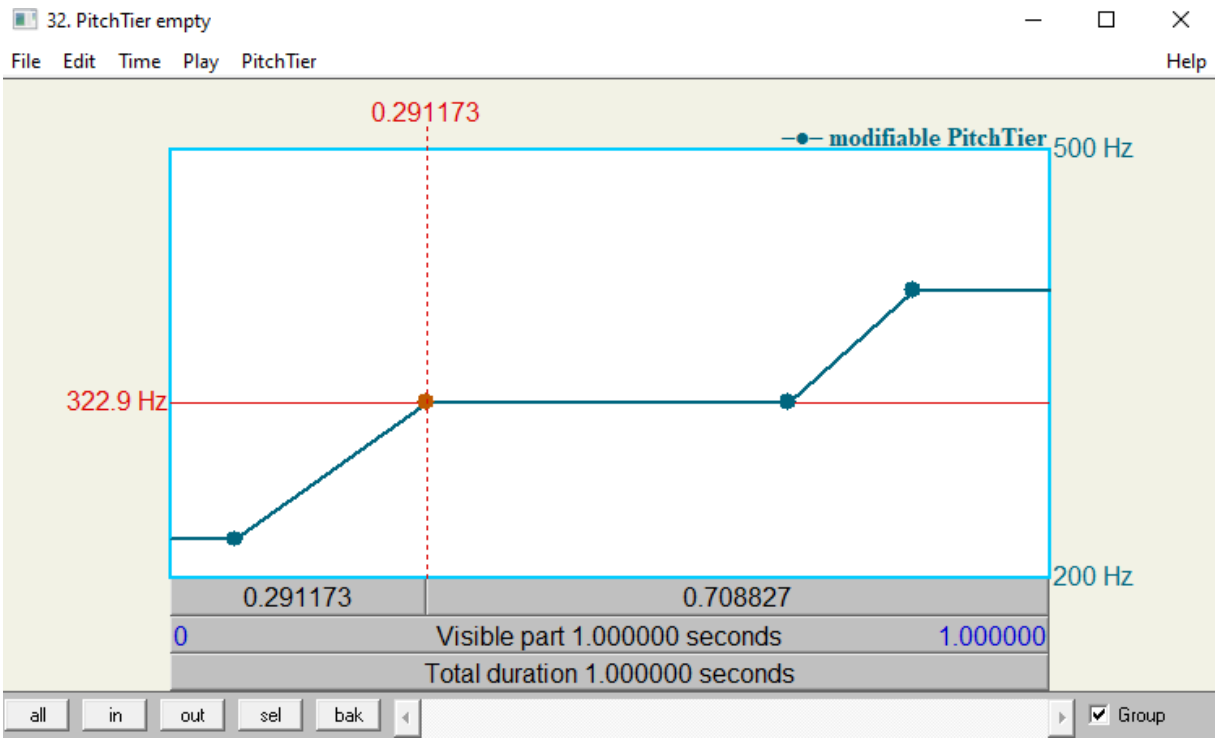


Abbildung 30 PitchTier-Ansicht
Quelle: Eigene Darstellung

Sobald die Tonhöhen im PitchTier eingestellt wurden, muss ein Sound-Objekt daraus erstellt werden. Dafür markiert man diese im „Praat Objects"-Fenster und geht auf „Synthesize". Nun gibt es drei „To Sound"-Möglichkeiten, die dem PitchTier eine Waveform (Form der Schwingung) zuweist: Pulse Train, Phonation und Sine.
„Pulse Train" erstellt eine Reihe von Impulsen auf Basis der Tonhöheneinstellung. Mit „Sine" wird eine Sinusschwingung entsprechend der eingestellten Grundfrequenz erzeugt. „Phonation" erstellt ein Signal, das dem LF-Modell in Abbildung 12 ähnelt. Es zeigt ein Modell der Glottisöffnung und -schließung. Daher wird zum Synthetisieren dieses Signal genommen.

PitchTier > Synthesize > To Sound (phonation)…

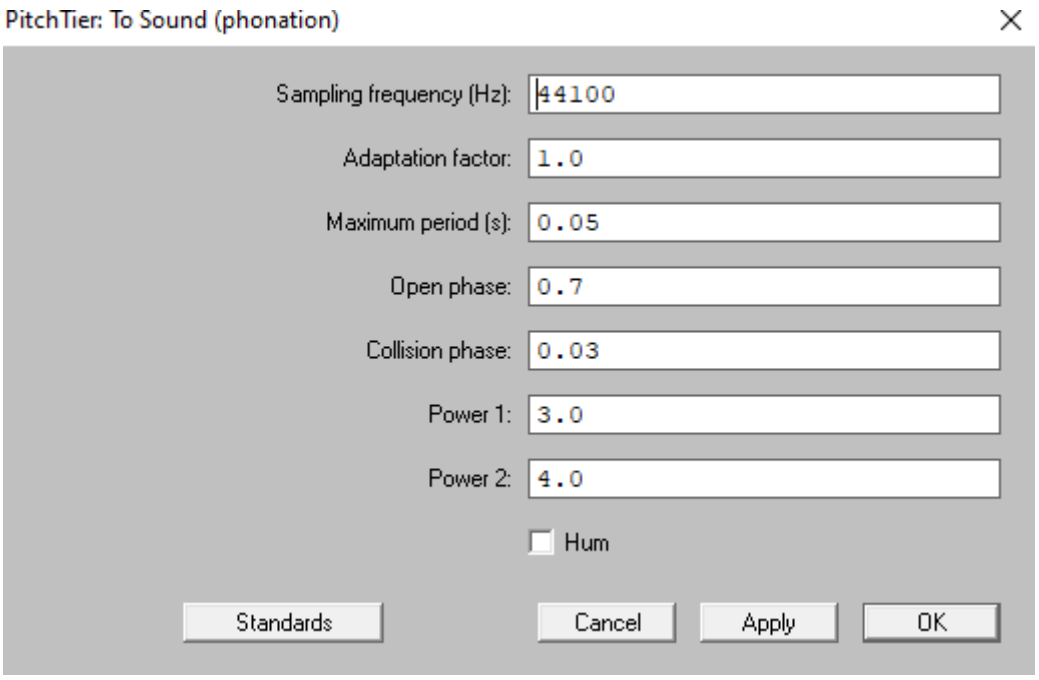


Abbildung 31: Standardeinstellungen, um die Tonhöhen (PitchTier) mittels Phonation in eine Sound-Datei zu überführen
Quelle: Eigene Darstellung

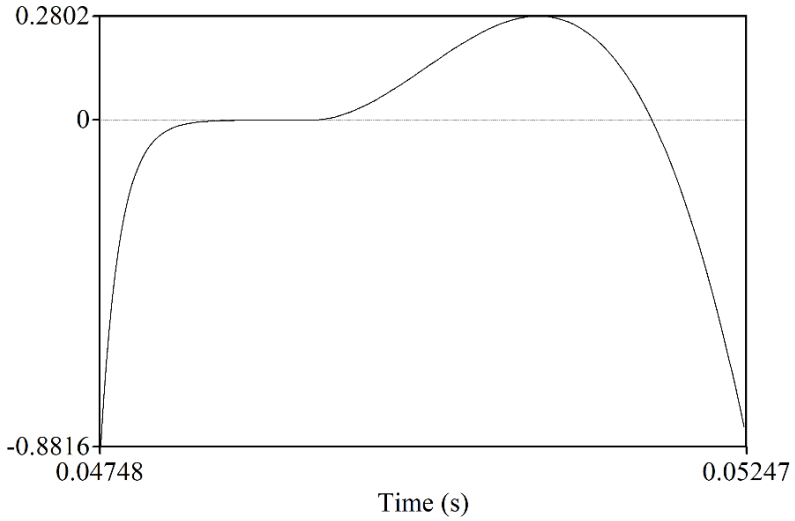


Abbildung 32: Phonations-Signal von Praat
Quelle: Eigene Darstellung

Bei der eingestellten Grundfrequenz von 200 Hz (in Abbildung 32) gibt es bei jeder Harmonischen einen Peak, also bei 200 Hz, 400 Hz, 600 Hz usw. Das dazugehörige Spektrum sieht man in Abbildung 33. Hier kann eine Analogie zur Definition von Klang (Gleichung 3) hergestellt werden, die man im Abschnitt 2.1 sehen kann. Es existiert eine Grundschwingung, die mit ganzzahligen Vielfachen ihrer selbst überlappt wird.

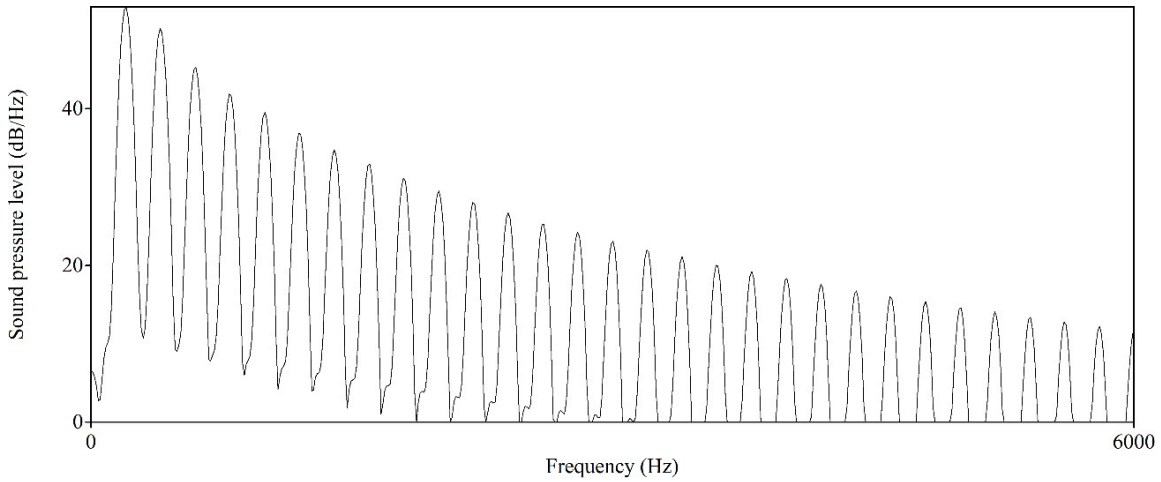


Abbildung 33: Spektrum zum Phonations-Signal bei 200 Hz Grundfrequenz
Quelle: Eigene Darstellung

Da im Vokaltrakt allerdings einige Harmonische verstärkt bzw. geschwächt werden, wird ein Filter für die Synthese benötigt. Nun kommt das FormantGrid-Objekt ins Spiel.

5.6.3 Sprachsynthese

Das Sound-Objekt, in der die Informationen zur Tonhöhe liegen, und das FormantGrid-Objekt werden beide markiert. Daraufhin wird im Objekt-Fenster die Aktion „Filter“ erscheinen, wie man in Abbildung 34 sieht. Es wird ein neues Sound-Objekt erstellt, bei welcher das Phonations-Signal durch das FormantGrid gefiltert wird. Nun sind die Harmonischen, die sich außerhalb der Formantbereiche befinden, stark gedämpft. Es wird eine subtraktive Synthese angewendet. Das entstandene Sound-Objekt ist die synthetisierte Stimme eines Vokals und kann als diverse Audio-Formate (z.B. .wav oder .mp3) exportiert werden. Für weitere Vokale wird der Vorgang ab Abschnitt 5.6.1 weiter wiederholt.

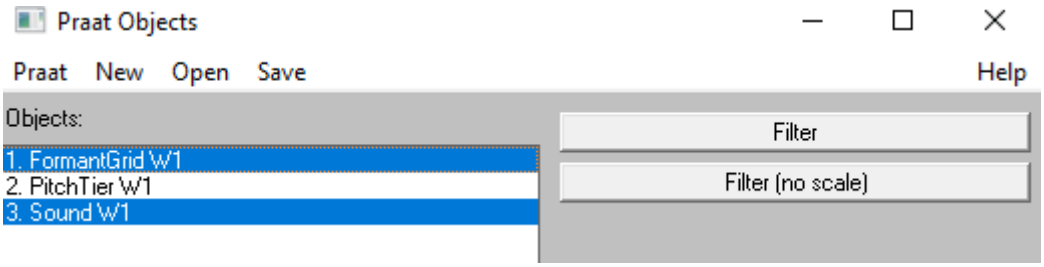


Abbildung 34: Kombination vom Formant-Grid mit dem Phonations-Signal
Quelle: Eigene Darstellung

6. Diskussion

Die Idee, die Stimme von meinen Freundinnen und Freunden, sowie meine selbst, nachbilden zu können, hat mich sehr gereizt. Auch als ich schon zu Anfang wusste, dass die Stimme nicht vollkommen natürlich klingen würde, da dieser Negativpunkt bei der Formantsynthese häufig aufgeführt ist. Auf diese (Un-)Natürlichkeit wird in Abschnitt 6.1 eingegangen.
Je mehr ich recherchiert habe, desto sicherer wurde ich, dass mir eine einfache Synthese gelingen wird. Nur der Klang der aufgenommenen und der synthetischen Stimmen sollten sich ähneln.

6.1 Zur Natürlichkeit der synthetisierten Stimmen

In der Thesis wird betrachtet, wie man eine Stimme vokalspezifisch synthetisieren kann und einen Vokal so klingen lassen kann, dasser einer speziellen Person zugeordnet werden kann. Ein großer Negativpunkt der Formantsynthese ist allerdings die unnatürlich klingende Sprache.

Laut Kaufmann und Pfister (2008) gibt es mehrere wichtige Eigenschaften, damit ein Sprachsignal natürlich klingt. Zum einen müssen Signalabschnitte quasiperiodisch und rauschartig sein, wobei auch Übergangs- und Mischbereicheauftreten können. Sie sind an einigen Stellen quasistationär. Wären sie exakt stationär und somit periodisch, dann werden die Sprachsignale nicht als solche wahrgenommen, sondern als technische Geräusche. Zudem ist die relative Bandbreite von Sprachsignalen groß und setzt sich somit aus Frequenzen zusammen, die über den gesamten Hörbereich verteilt sind. Die Resonanzen des Vokaltraktes zeigen sich im Spektrum als Formanten. Zum Schluss erwähnen Kaufmann und Pfister, dass es für die Natürlichkeit des Sprachsignal – bis auf ein paar Ausnahmen - keine abrupten Änderungen in ihrer spektralen Zusammensetzung gibt. [12]

In Bezug auf den vorherigen Absatz wird nun erläutert, warum sich die synthetisierten Stimmen nicht natürlich anhören werden.
„Rauschartige Signalabschnitte“, „Übergangs- und Mischbereiche“ treten z.B. vor und nach Lauten auf. Diese Übergänge werden nicht nachgebildet. Zudem werden die Formanten in den synthetisierten Vokalen eine Sekunde lang absolut periodisch – und somit stationär – sein, da keine Geräusche von außen oder Schwankungen in der Stimme vorkommen. Außerdem geht die Bandbreite von natürlichen Sprachsignalen über den gesamten Hörbereich (20 Hz bis 20 kHz).

[12] Beat Pfister und Tobias Kaufmann (2017). *Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung. S. 241 f.*

Die ersten vier bis fünf Formanten, die für die Synthese genutzt werden, befinden sich allerdings im Bereich bis 5000 Hz. Hohe Frequenzen ab 5000 Hz werden nicht erzeugt.

6.2 Vergleich von aufgenommenen und synthetisierten Stimmen

6.2.1 Spektrogramme

Die Spektrogramme dienen als Vergleich, um objektiv auf nicht-auditive Weise die (Un-)Ähnlichkeit der aufgenommenen und synthetischen Vokale zu verweisen. Wie in Kapitel 5 wurde Praat genutzt. Es sind Breitbandspektrogramme, die durch ein Gauß-Fenster mit der Länge von 0,005 s erstellt wurden.

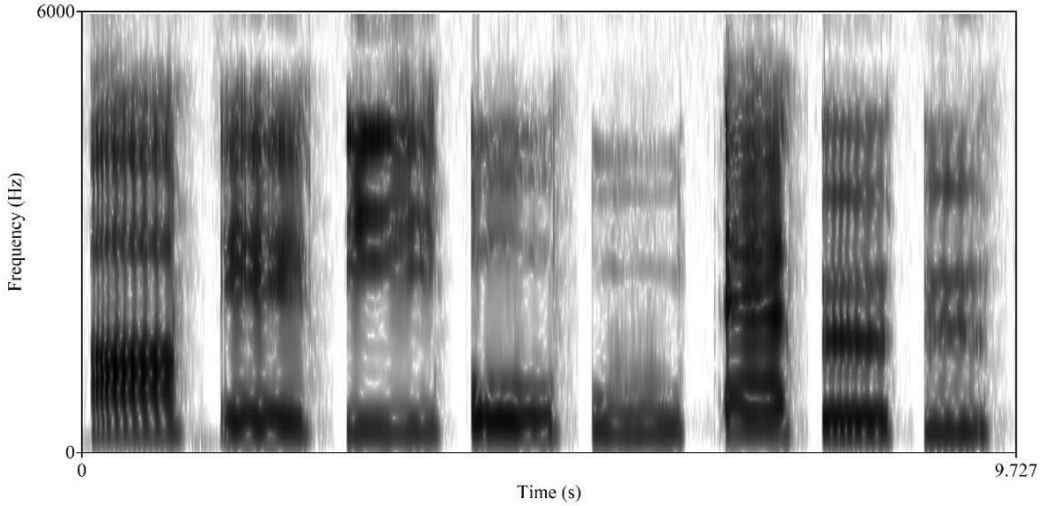


Abbildung 35: Spektrogramm der aufgenommenen Stimme von W1
Quelle: Eigene Darstellung

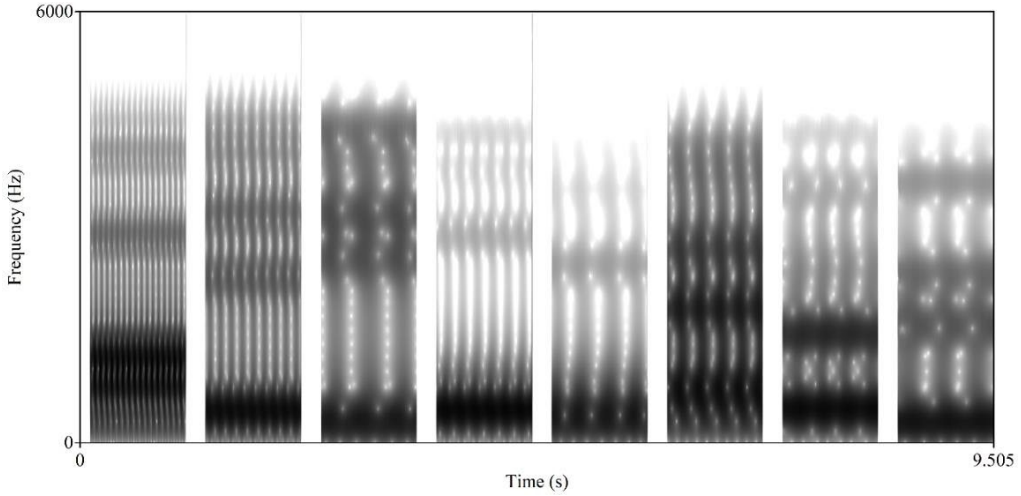


Abbildung 36: Spektrogramm der synthetischen Stimme von W1
Quelle: Eigene Darstellung

Vergleicht man die Spektrogramme der aufgenommenen und synthetischen Stimme, so scheinen die Formanten den gleichen Frequenzbereich zu umfassen. Wie erwartet sind bei der synthetischen Stimme die Formanten allerdings gradliniger.
Das Rauschen fehlt, welches entweder von der Aufnahme mit dem Mikrofon kommt und/oder von der Stimme selbst. Es wurden fünf Formantfrequenzen für die Synthese eingestellt. Nach dem vierten Formanten ist der SPL-Wert sehr gering und der Bereich ist weiss. Der fünfte scheint nicht mehr vorhanden zu sein, obwohl er eingestellt wurde. Besonders sichtbar ist das bei [o:] und [u:], bei denen die ersten fünf Formanten allgemein niedrig verlaufen.

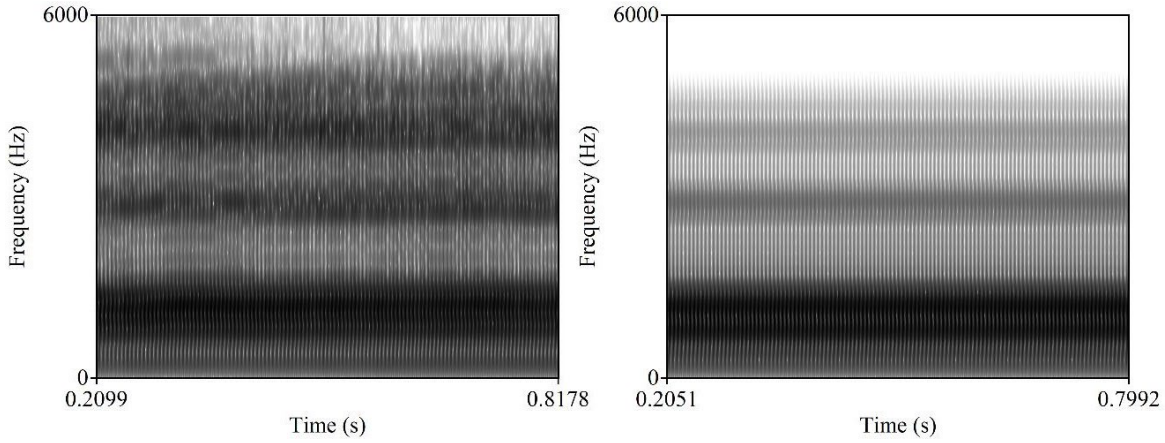


Abbildung 37:Vergleich vom aufgenommenen und synthetischen [a:] von W1
Links: Aufgenommenes [a:] von W1, rechts: Synthetisiertes [a:] von W1
Quelle: Eigene Darstellung

Wie in Abschnitt 5.4 bereitsangedeutet ist bei einigen Vokalen schwierig gewesen, die Formanten auszuwerten. Besonders bei den Vokalen [o:] und [u:] bei den Stimmen der männlichen Sprecher hatten Praat und ich Probleme, die Formanten zu bestimmen. Vermutlich liegt das an der Lage der Formanten, da diese teilweise sehr dicht beieinanderliegen. Hier kommt die halbautomatische Methode ins Spiel. Die im Spektrogramm mit roten Punkten angezeigten Formanten sollen mit den dunklen Linien übereinstimmen (siehe Abbildung 40), die man ohne die Formantanzeige sehen kann. Es war also ein ständiges Abwiegen und Ändern der Werte in den Formant-einstellungen. Bei den tiefen Formanten, bei denen die erste Harmonische noch viel Einfluss hat, wurden die Werte immer wieder mit denen aus Tabelle 1 verglichen.
Besonders bei den männlichen Sprechern war es teilweise schwierig, die Formanten aus dem Spektrogramm zu entnehmen. Wie in Abschnitt 5.4 in der Tabelle 3 zu sehen ist, konnte ich für M2 bei [u:] keine genauen Werte entnehmen. In Abbildung 38 ist sein Spektrogramm zu sehen.

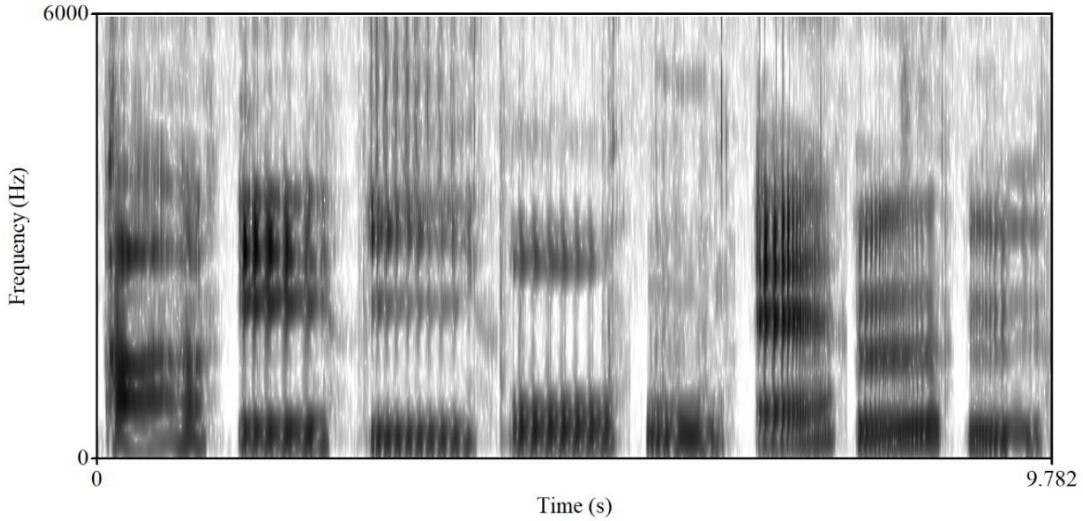


Abbildung 38: Spektrogramm der aufgenommenen Stimme von M2
Quelle: Eigene Darstellung

Bei [a:] bis [i:], [ɛ:], [øː] und [y:] kann man noch vier Formanten erkennen. Bei [o:] sind es noch maximal drei für mich sichtbare Formanten und bei [u:] maximal zwei. Vermutlich liegt das daran, dass die Formanten sehr dicht beieinander liegen. Deswegen hat die LPC-Analyse- die zur Formanatanalyse verwendet wird - möglicherweise Probleme die Formanten zu finden. Da ich beim [u:] keine guten Werte entnehmen konnte, habe ich von M2 diesen Vokal nicht synthetisiert.

6.2.2 Spektren

Bei dem Amplituden-Frequenzverlauf kann man auch einige Unterschiede feststellen, welche sich schon im Spektrogramm abzeichneten. Die Formanten in der aufgenommenen und der synthetischen Stimme sind sichtbar. Allerdings nimmt der SPL beim synthetischen Sprachsignal deutlich schneller ab als beim aufgenommenen. Bereits der vierte Formant ist bei knapp über 0 dB. Der fünfte bei ca. -20 dB.
Das kann zumeinenan der Aufnahme liegen. Bei der aufgenommenen Stimme ist noch zusätzlich ein Rauschen im Spektrogramm sichtbar. Zum anderen könnte es am Phonations-Signal liegen. In Abbildung 39 ist der Amplituden-Frequenzverlauf vom aufgenommenen (schwarze Linie) und synthetisierten (rote Linie) [a:] von W1 zu sehen.

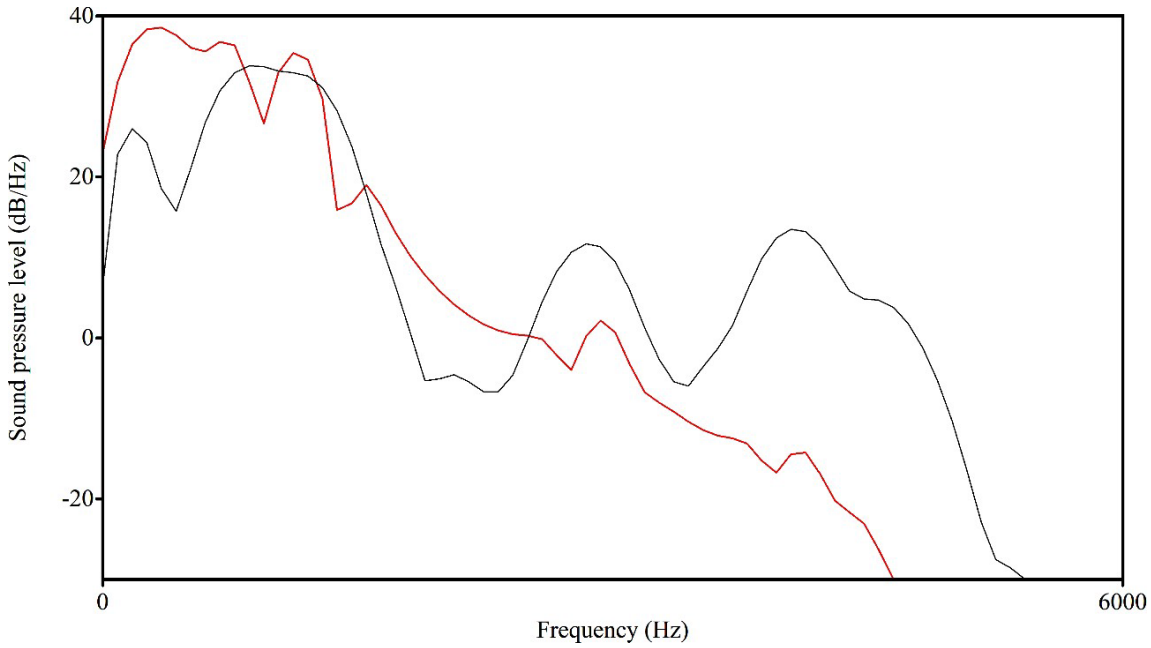


Abbildung 39: Amplituden-Frequenzverlauf vom aufgenommenen und synthetischen [a:] von W1
Quelle: Eigene Darstellung

6.2.3 Formanten

Die Einstellungen, mit denen ich in den aufgenommenen Stimmen die Formanten analysiert habe, haben bei den synthetischen Stimmen nicht mehr gegriffen. Mit den gleichen Einstellungen bei der Formantanalyse sind Formanten an Stellen aufgetaucht, die es in der originalen Aufnahme nicht gibt.

Zum Beispiel bei Sprecherin W1: Bei [a:] wird direkt bei F2 ein dritter Formant hinzugedichtet. Bei [e:] und [i:] ist F2 zu niedrig beziehungsweise taucht an einer Stelle auf, wo man im Spektrogramm keine außergewöhnlich dunkle Graustufe sehen kann. F3 hat in der Synthese allerdings die gleichen/ähnlichen Werte wie F2 bei der aufgenommenen Stimme. Bei [o:] ist F3 zu niedrig und auch hier ist keine besondere Abdunklung im Spektrogramm festzustellen.

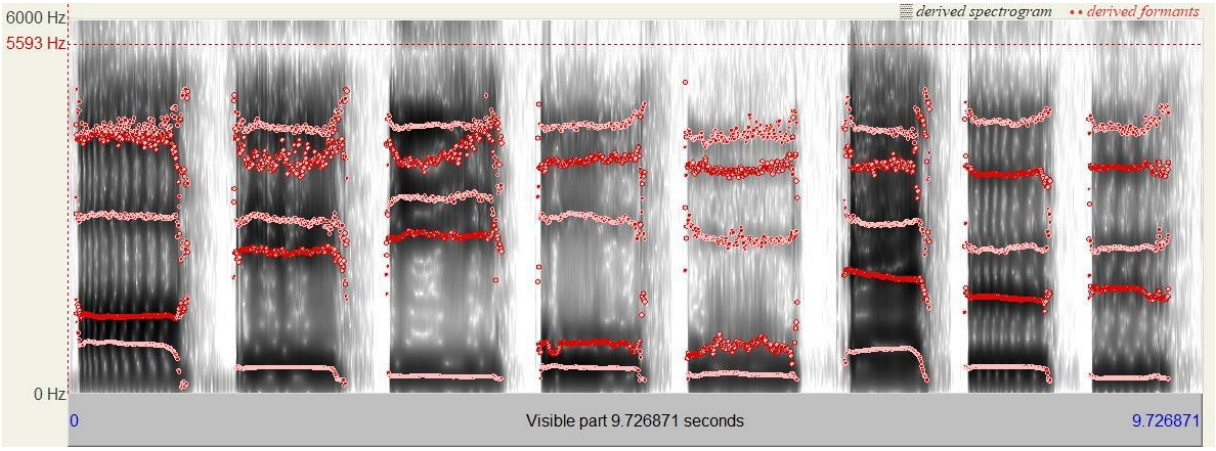


Abbildung 40: Spektrogramm von W1 (aufgenommene Stimme) mit Formanten
Quelle: Eigene Darstellung

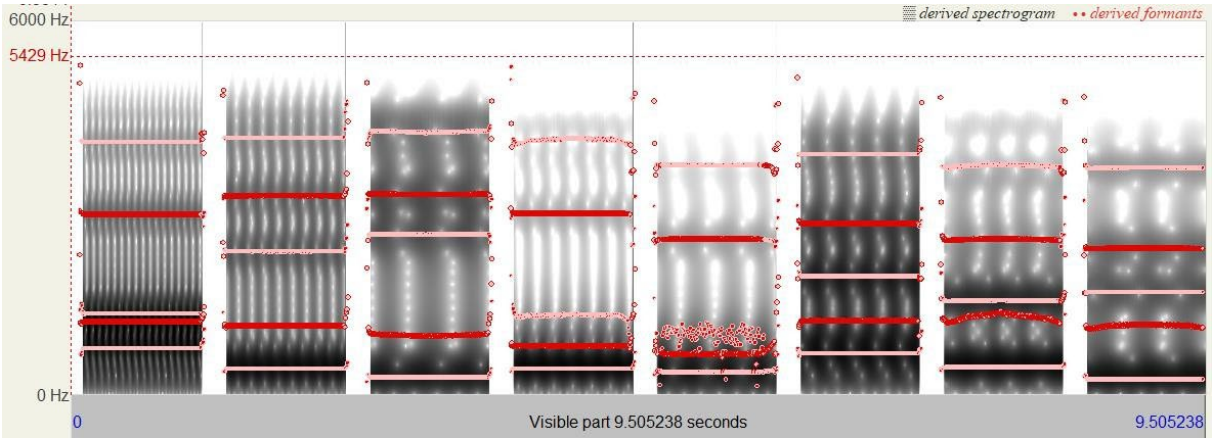


Abbildung 41: Spektrogramm von W1 (synthetisierte Stimme) mit Formanten
Quelle: Eigene Darstellung

Das liegt daran, dass Praat den fünften Formanten im Spektrogramm „verschluckt“. Das Programm sollte im Bereich bis 5500 Hz nach fünf Formanten schauen. Wegen diesen Formanteinstellungen, die ich für die Analyse der aufgenommenen Stimmen genommen habe, hat Praat gezwungenermaßen versucht, einen fünften Formanten zu finden. Als die Formantanzahl auf vier runtergesetzt wurde, wurden die ersten vier gemessenen und für die Synthese genutzten Werte problemlos angezeigt.

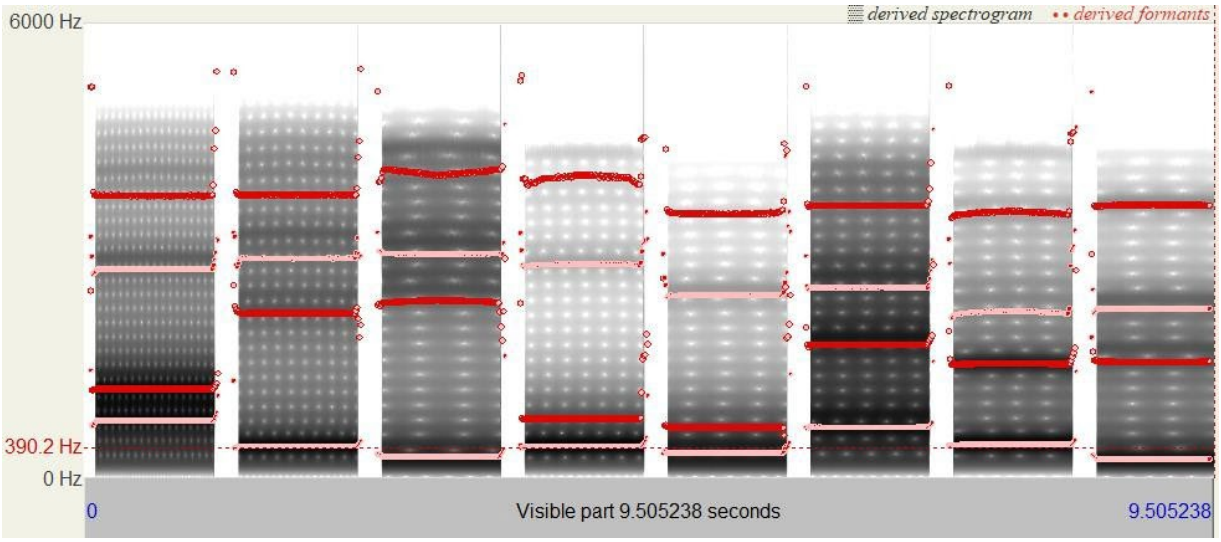


Abbildung 42: Spektrogramm von der synthetisierten Stimme W1 mit vier Formanten
Quelle: Eigene Darstellung

In Abschnitt 6.2.2 wird ein möglicher Grund erwähnt. Im Spektrum sieht man, dass F4 in Abbildung 39 einen SPL von knapp über 0 dB hat. F5 ist bei ca. -20 dB. Vermutlich fehlt bei den synthetisierten Stimmen das Rauschen, auch wenn man es bei den aufgenommenen Stimmen kaum bemerkt.

7. Fazit

Der Bereich der Sprachsyntheseist breit gefächert. Viele Konzerne wie Google oder Applenutzen Text-to-Speech-Systeme, um Geschriebenes in Sprache umwandeln zu können. Am Verbreitesten ist heutzutage die Unit Selection. Diese greift auf vorher eingesprochene Laute (Phone), Wörter oder komplette Sätze in einer Datenbank zurück, setzt diese wie gewollt zusammen und gibt sie als neues Sprachsignal wieder.

Quelle-Filter-Modelle bieten ein Gegenstück zur Unit Selection. Bei diesen Modellen wird ein Quellsignal durch mehrere Filter geschickt und kommt - durch diese ummodelliert - als Sprachsignal aus dem System raus. Eines dieser Quelle-Filter-Modelle ist die Formantsynthese. Sie baut darauf auf, dass verschiedene Vokale mit zwei Frequenzbereichen unterschieden werden können. Diese Frequenzbereiche werden Formanten genannt und zeichnen sich durch einen höheren Schalldruckpegel (SPL) ab als angrenzende Bereiche. Vokale sind Laute, in denen der Luftstrom ungehindert den Vokaltrakt passieren kann. Dadurch entsteht der höhere SPL, welcher mithilfe eines Spektrogramms als dunkler Bereich sichtbar gemacht wird. Das Spektrogramm stellt den SPL von Frequenzen über den zeitlichen Verlauf dar. In Vokalen sind die Formanten als dunkle Linien sichtbar. Die zwei niedrigsten Frequenzbereiche werden mit F1 und F2 gekennzeichnet.

Mit diesem Wissen können Vokale - und somit ein Teil von Sprache - synthetisiert werden. Aber sie klingen nicht so, als würden sie von einer Person gesprochen werden.

Es gibt weitere Formanten. Diese tragen weniger zur Sprache bei. Sie sorgen allerdings für den individuellen Klang der Stimme eines Menschen.

In einem Experiment werden sowohl F1 und F2 als auch die weiteren Formanten (mit den Kürzeln F3, F4, etc.) synthetisiert. Diese sind sprechendenspezifisch. Wenn man sie mit in die Synthese einbezieht, kann man die Vokale so klingen lassen, als seien sie von verschiedenen Personen gesprochen.

Um die Formanten zu ermitteln, sprechen fünf Personen diverse deutsche Vokale - unter gleichen Randbedingungen in einer monotonen Stimmlage - aus. Diese werden daraufhin mit Praat analysiert. Das Programm ist spezialisiert darauf phonetische Parameter zur ermitteln, wie z.B. Formanten. Es ist ein sehr umfang- und hilfreich, wenn man sich in seine Funktionen eingearbeitet hat. Bei ungenauen Einstellungen kann es allerdings zu fehlerhaften Messwerten kommen. Diese wären dann für eine Synthese nicht geeignet. Daher müssen von Person zu Person die Einstellungen für Analysen neu angepasst werden. Untersucht werden ihre Stimmen auf Formanten, deren Bandbreiten und Grundfrequenz. Die Grundfrequenz trägt maßgeblich zum Klang bei, da die Glottis - die mit dem Luftstrom angeregt wird - Harmonische erzeugt, welche durch die Resonanzen im Vokaltrakt verstärkt werden können. Die Resonanzen sind die Formanten.

All diese Parameter werden in der Synthese gebraucht. Für die Sprachsynthese wird wieder Praat genutzt. Es bietet die Möglichkeit als Quellsignal ein Phonations-Signal auszuwählen. Es hat im Zeitbereich eine ähnliche Form wie Modelle, die von Forschenden entwickelt wurden, um den Glottisimpuls nachzustellen. Bei dem Phonations-Signal wird vorher die Grundfrequenz eingestellt. Die Formanten und Bandbreiten werden in einem anderen Objekt eingestellt, welches als digitales Filter dient. Am Ende werden das Phonations-Signal und das Filter zusammengefügt und ergeben das synthetisierte Sprachsignal.
Wie erwartet klingen die synthetisierten Stimmen nicht sonderlich natürlich. Dennoch kann man akustisch die Ähnlichkeiten zwischen den aufgenommenen und synthetisierten Stimmen erkennen. Für die nicht-auditive Belegung dienen die Spektrogramme. In diesen lassen sich die Ähnlichkeitenebenfalls aufzeigen, auch wenn die synthetisierten Stimmen im wahrsten Sinne des Wortes gradliniger sind. Deutlich sichtbar sind allerdings nur noch die ersten vier Formanten. Zudem macht es offenbar einen großen Unterschied, ob Rauschen in der Audiospur enthalten ist. Während die Fläche bei den aufgenommenen Stimmen über den gesamten Frequenzbereich leicht gräulich ist – was vermutlich vom Rauschen kommt – sieht man im Spektrogramm der synthetisierten Stimme sehr viel weiss. Das lässt auf einen geringen bis nicht vorhandenen SPL schließen.
Im Spektrum ist dies auch klar erkennbar. Der SPL fällt nach dem vierten - von fünf eingestellten – Formanten rapide ab. Das macht sich bei der Formantanalyse bermerkbar. Die Einstellungen, die zur Analyse der Formanten in den aufgenommenen Stimmen genutzt wurden, führen bei den synthetisierten Sprachsignalen zu falschen Formanten, die nicht im Originalsignal vorkommen. Praat erkennt nur noch vier Formanten, weshalb die Anzahl an erwarteten Formanten in den Einstellungen auf vier statt fünf gestellt wird. Mit dieser Einstellung werden nun die ersten vier Formanten der Vokale korrekt angezeigt. Scheinbar wurde der fünfte Formant „verschluckt“ und trägt nicht mehr zum Klang bei.

Trotzdessen ist das Endprodukt, das synthetisierte Sprachsignal, für den ersten Versuch gut gelungen. Dieses Experiment hat mir die Komplexität der Stimmen und Stimmensynthese vorgeführt.

8. Quellen

Chen, C. J. (2016). *Elements of human voice*. World Scientific Publishing Co. Pte. Ltd.

Doval, B. und d'Alessandro, C. (2006). *The Spectrum of Glottal Flow Models. Acta Acustica united with Acustica Vol. 92*. S. Hirzel Verlag

Kortmann, B. (2020). *English Linguistics: Essentials (2nd revised, updated and enlarged edition.)*. J.B. Metzler Verlag

Lemmetty, S. (1999). *Review of Speech Synthesis Technology*. (Master Thesis, Department of Electrical and Communications Engineering). Helsinki University of Technology in Helsinki, Finnland

Mayer, J. (2022). *Phonetische Analysen mit Praat. Ein Handbuch für Ein- und Umsteiger.* [online] https://praatpfanne.lingphon.net/das-praat-handbuch/ [zuletzt abgerufen am 10.01.2024]

Pfister, B. und Kaufmann, T. (2017). *Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung. 2. Auflage.* Springer-Verlag GmbH Deutschland

Pompino-Marschall, B. (2003). *Einführung in die Phonetik*. Walter de Gruyter GmbH & Co. KG

Pulakka, H. (2005). *Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottography* (Master's Thesis, Language Technology). Helsinki University of Technology in Helsinki, Finnland

Sendlemeier, W.F. und Seebode, J. (2006). *Formantkarten des deutschen Vokalsystems.* TU Berlin, Institut für Sprache und Kommunikation, Berlin

9. Anhang

9.1 Tabellen mit Formant- und Grundfrequenzwerten der Sprechenden

Tabelle: Sprecherin W1

W1	F0 [Hz]	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]	F5 [Hz]
[a:]	193,8	813,692	1220,408	2873,533	4016,861	4531,12
[e:]	201,8	412,666	2285,371	2833,861	4106,675	4479,984
[i:]	207,9	265,496	2559,219	3191,599	3993,711	4339,61
[o:]	204,2	408,643	799,335	2898,648	3816,631	4320,135
[u:]	207,3	297,815	699,941	2468,682	3538,862	4102,591
[ɛ:]	205,7	705,164	1867,784	2720,973	3812,146	4374,86
[ø:]	214,6	418,055	1512,142	2438,378	3523,1	4327,698
[y:]	213,4	233,9	1671,555	2315,925	3629,824	4233,844

Tabelle: Sprecherin W2

W2	F0 [Hz]	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]	F5 [Hz]
[a:]	244,2	870,6	1236,79	3013,893	4229,01	
[e:]	254,8	473,236	2392,043	3122,493	4005,429	
[i:]	267,2	274,251	2675,847	3425,549	4264,116	
[o:]	261,7	479,917	752,395	2883,777	3791,393	4140,753
[u:]	267,3	312,502	789,942	2723,591	3757,513	4761,961
[ɛ:]	260,5	760,407	2093,446	2928,175	3923,764	4271,239
[ø:]	256,8	472,284	1535,948	2351,405	3850,299	4274,252
[y:]	264,3	316,006	1452,19	2517,78	3753,988	4627,927

Tabelle: Sprecherin W3

W3	F0 [Hz]	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]	F5 [Hz]
[a:]	186,8	720,816	1308,664	3004,276	4045,361	4876,137
[e:]	180	354,112	2504,091	3024,424	4142,28	4737,045
[i:]	179,8	335,851	2578,534	3407,109	4144,409	4818,102
[o:]	179,8	358,196	746,937	2873,137	3767,851	4499,157
[u:]	184	357,686	799,549	2581,849	3862,472	4232,608
[ɛ:]	174,3	637,153	2100,68	2862,988	3847,855	4449,801
[ø:]	177,5	357,311	1565,646	2356,254	3544,175	4535,308
[y:]	179,1	307,566	1885,428	2522,517	3696,236	4388,902

Tabelle: Sprecher M1

M1	F0 [Hz]	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]	F5 [Hz]
[a:]	122,6	638,246	1068,232	2425,683	3903,759	5083,232
[e:]	120,4	321,84	2091,775	2625,665	3391,299	4611,598
[i:]	121,4	224,643	2150,862	3043,604	3956,697	4581,116
[o:]	121,5	349,099	671,269	2445,166	3369,761	4878,274
[u:]	128,2	221,129	613,116	2144,302	3224,086	4347,257
[ɛ:]	126,1	530,63	1653,745	2451,714	3665,915	4687,977
[ø:]	120,4	320,26	1550,787	2163,771	3130,165	4388,777
[y:]	124	256,504	1652,922	1992,648	3090,354	4493,751

Tabelle: Sprecher M2

M2	F0 [Hz]	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]	F5 [Hz]
[a:]	102,8	758,848	1279,219	2745,862	3623,745	
[e:]	107,3	296,974	2107,313	2811,595	3289,516	
[i:]	109,8	242,752	2059,314	2953,725	3411,204	
[o:]	109,4	320,829	589,432	2629,836	2960,01	
[u:]						
[ɛ:]	112,1	557,005	1860,73	2676,013	3220,538	
[ø:]	124	322,172	1429,031	2235,756	3112,416	
[y:]	101,3	252,166	1585,547	2068,702	3073,285	

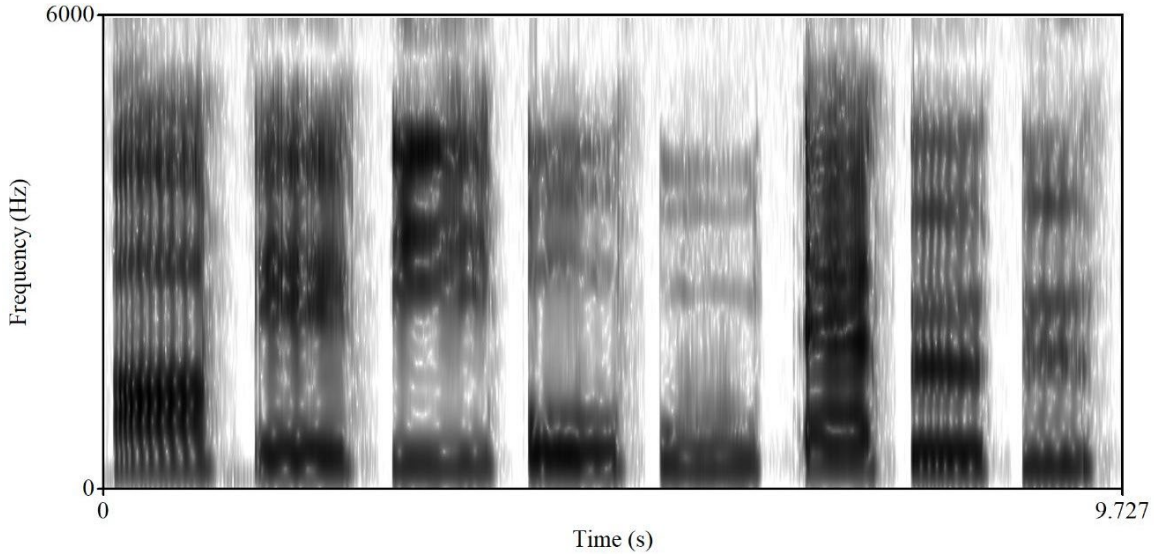
9.2 Spektrogramme der aufgenommenen und synthetisierten Vokale

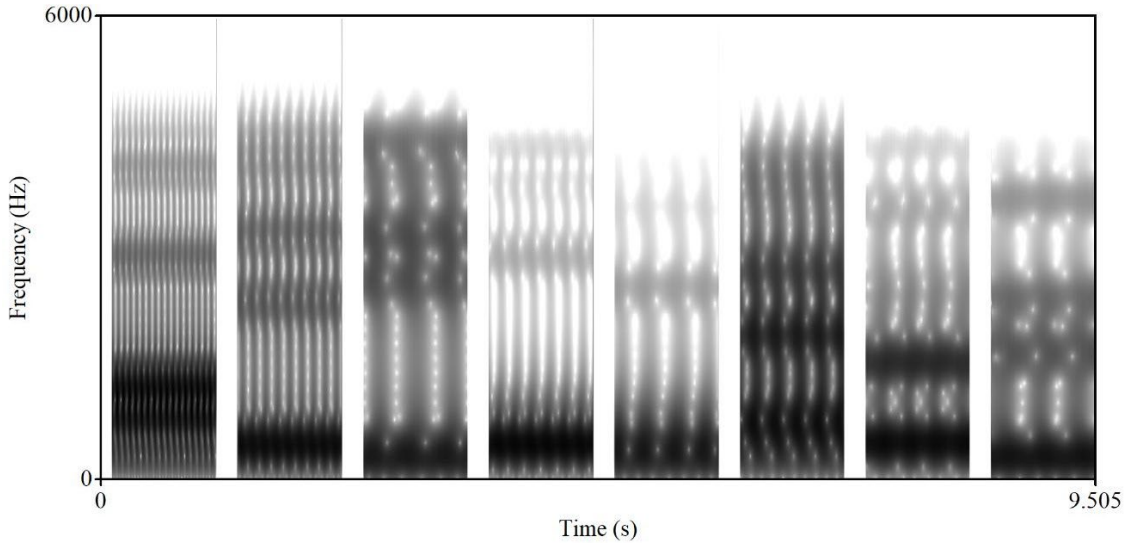
Von links nach rechts werden die folgenden Vokale aufgelistet:

[a:], [e:], [i:], [o:], [u:], [ɛ:], [ø:] und [y:]

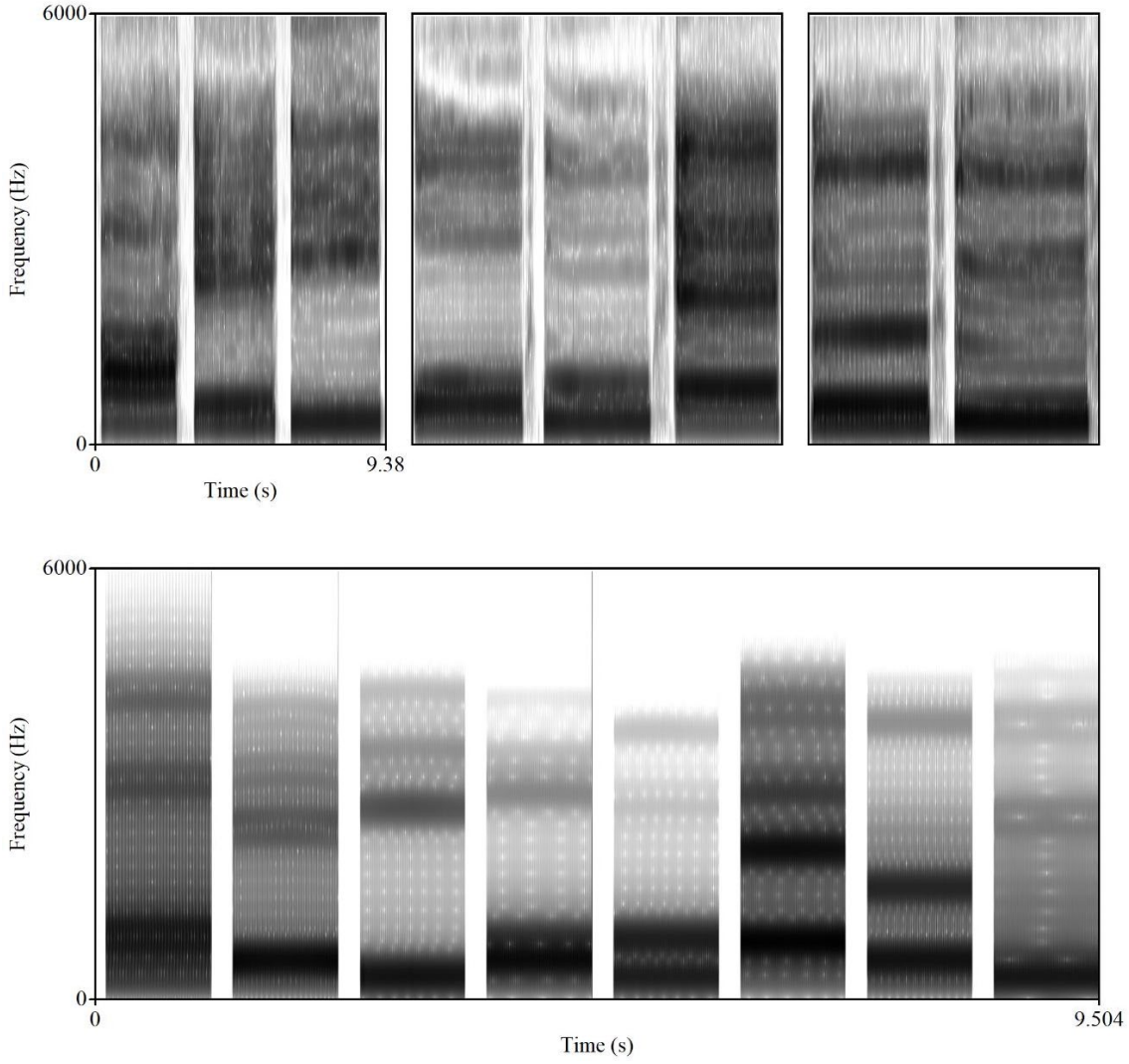
Das erste Spektrogramm eines Sprechenden zeigt die aufgenommenen Vokale, das zweite zeigt die synthetisierten Vokale

9.2.1 Sprecherin W1

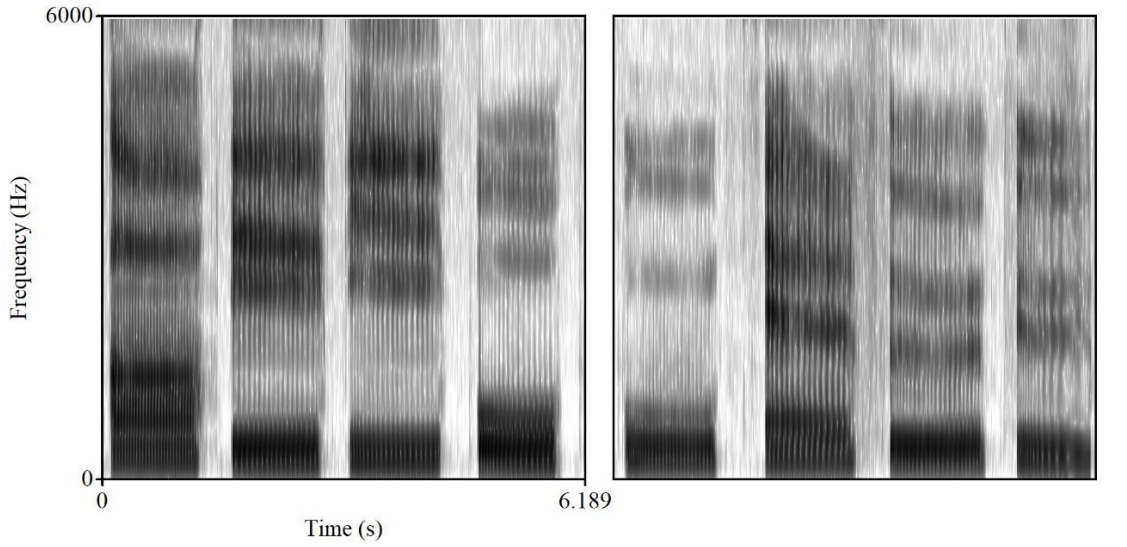


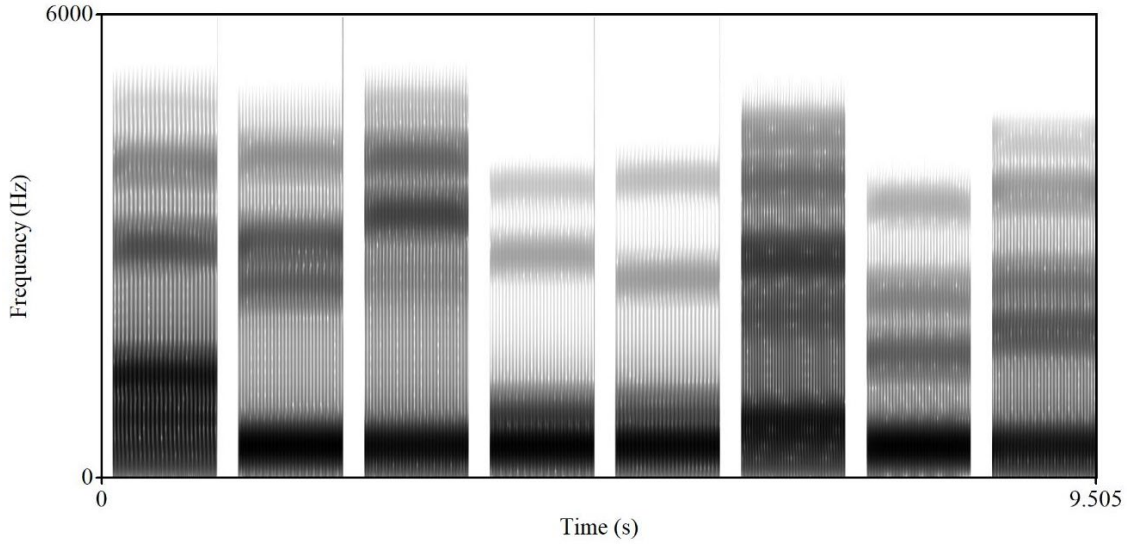


9.2.2 Sprecherin W2

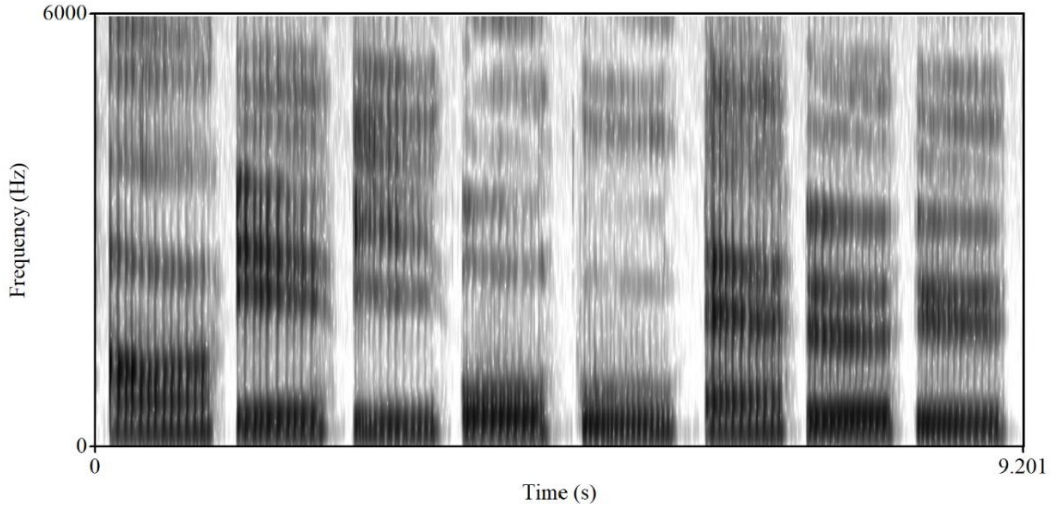


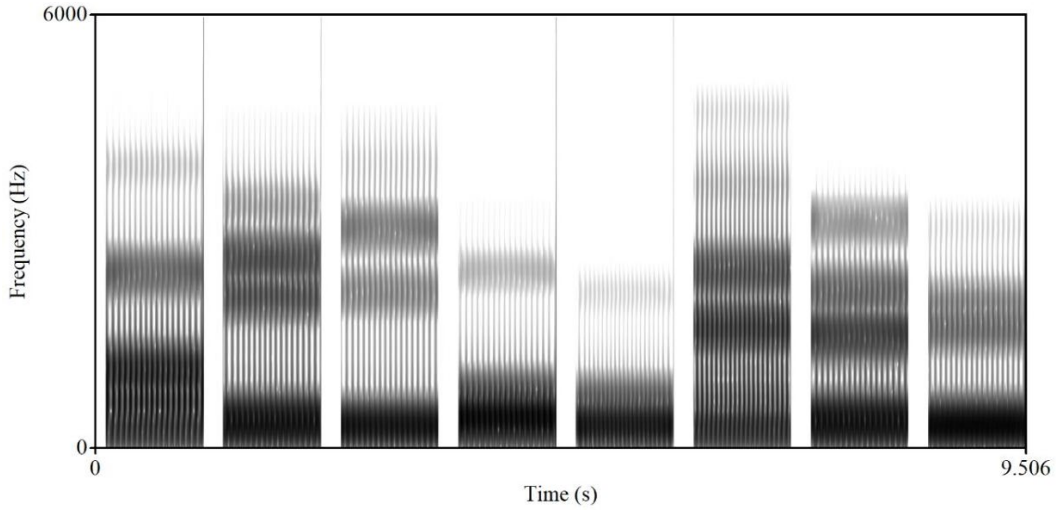
9.2.3 Sprecherin W3



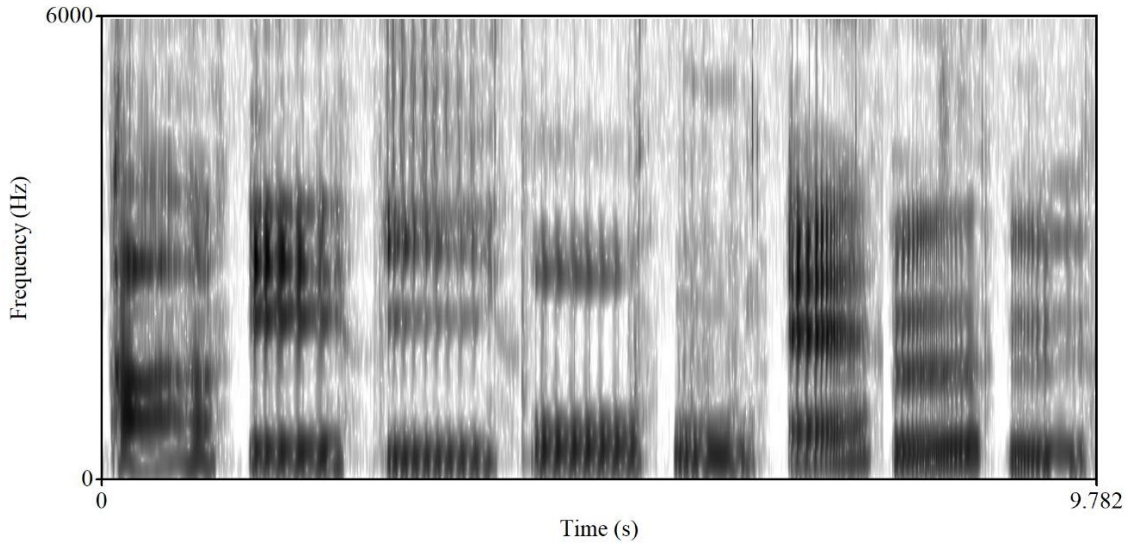


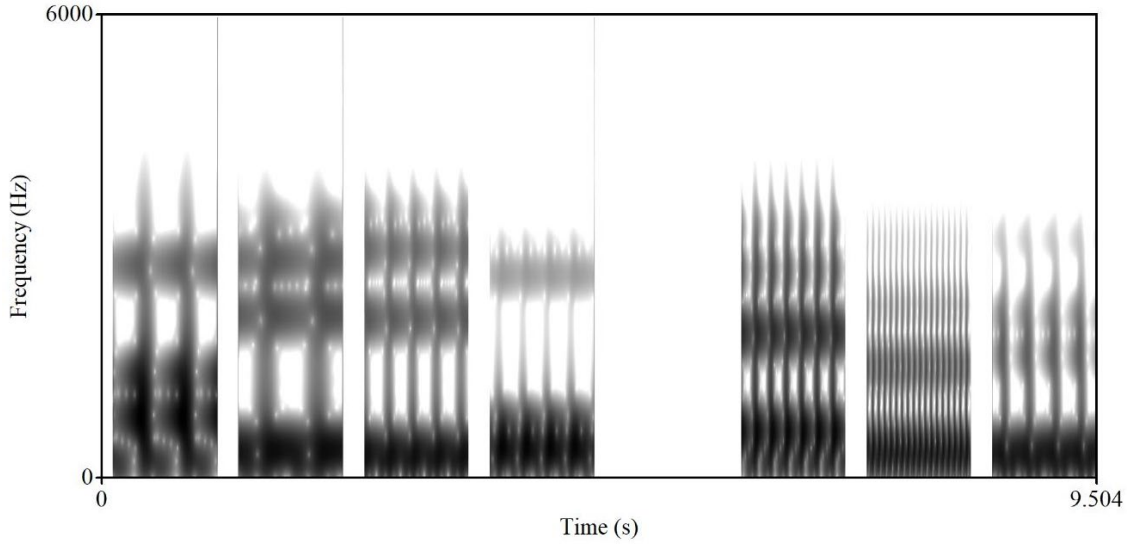
9.2.4 Sprecher M1





9.2.5 Sprecher M2





Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit mit dem Titel:

_____

selbständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

Datum	Unterschrift