Universität Zürich[UZH]

Bachelorarbeit der Philosophischen Fakultät der Universität Zürich

Herbstsemester 2019

Aussprachebewertung von Vokalen mittels Formantenanalyse

Institut für Computerlinguistik

Betreuer: Prof. Dr. Volker Dellwo

Verfasser: Alon Cohen

Matrikel-Nr.: 16-730-665

Abgabedatum: 01.12.2019

Abstract

In dieser Arbeit wurde ein Programm entwickelt, das Sprachschülern ermöglicht ihre Aussprache deutscher Vokale zu üben. Dies wurde mit Hilfe von Formantenanalyse erreicht, wobei ein Evaluierungsalgorithmus entwickelt wurde, der die Ähnlichkeit der User-Formanten auf die Target-Formanten des jeweiligen Vokals hin überprüft. Das Programm liefert ziemlich zuverlässige Ergebnisse für die meisten Vokale des Deutschen inklusive Umlaute, hat aber Probleme die Vokale [o] und [u] konsistent zu erkennen.

Danksagung

Ich möchte mich bei Prof. Dr. Volker bedanken, der mich bei dieser Arbeit betreute und mir mit gutem Rat zur Seite stand. Ferner möchte ich mich bei Yannick Jadoul vom Artificial Intelligence Lab an der Vrije Universiteit Brussel bedanken, der mir half, die von ihm geschriebene Python-library *Parselmouth* besser zu verstehen.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation und Hintergrund

Die Entstehungsgeschichte dieser Arbeit nahm ihren Lauf im Frühling des Jahres 2019, als ich als App-Entwickler für die Vox-Sprachschule in Zürich eine App entwickeln sollte, die Sprachschülern beim Erlernen der Aussprachen verschiedener Sprachen unterstützen sollte. Genauer gesagt sollte die App den Schülern einen Satz vorsagen, den diese in das Mikrofon ihres Smartphones wiederholen sollte, woraufhin ihre Aufnahme auf die Aussprache hin überprüft und bewertet werden sollte. Ein - wie ich erst später lernen sollte - sehr ambitioniertes Unterfangen. Im Juni 2019, ein paar Monate nach Beginn der Arbeiten, die sich hauptsächlich mit dem Frontend beschäftigten, musste das Projekt leider eingestellt werden, da die Sprachschule im Sommerloch in finanziell turbulente Zeiten geraten war. Da ich trotzdem das Projekt umsetzen wollte, beschloss ich es im Rahmen meiner Bachelorarbeit fortzuführen. Als Betreuungsperson fand sich Prof. Dr. Volker Dellwo, der als Experte im Bereich der Phonetik an der Universität Zürich eine naheliegende Wahl war. Nach einem ersten Gespräch erschloss sich uns jedoch, dass das ursprüngliche Ziel, beliebige Sätze und Wörter nach ihrer Aussprache zu bewerten, zu ambitioniert für eine Bachelorarbeit war. Stattdessen einigten wir uns darauf, die Ziele etwas kleiner zur stecken: Neu ging es nur noch darum, die Aussprache von Vokalen bzw. Lauten zu bewerten, statt ganze Sätze. Das Grundkonzept blieb aber dasselbe, am Schluss der Arbeit sollte ein Tool bereitstehen, mit dem Sprachschüler ohne weitere technische oder linguistische Kenntnisse, ein Feedback für ihre Aussprache bekommen würden. Nur eben für Vokale, statt ganze Sätze.

Mein persönliches Interesse für dieses Thema dieser Arbeit beruht auf mein Interesse an gesprochener Sprache und am Erlernen dieser. Insbesondere faszinieren mich lokale Dialekte und wie schwierig es sein kann, eine wirklich überzeugende Aussprache in einer Fremdsprache zu erlangen.

1.2 Fragestellung

Diese Arbeit soll die folgende Fragestellung beantworten: Lässt sich mittels Formantenanalyse ein hilfreiches Tool zur Aussprachebewertung von Sprachschülern entwickeln?

1.3 Struktur

Die Arbeit wird sich zunächst mit den theoretischen Grundlagen der Phonetik, insbesondere mit Formanten beschäftigen. Damit wird dem Leser das benötigte Wissen bereitgestellt, um die Funktionsweise des eigentlichen Tools nachvollziehen zu können, die in Kapitel 3 näher erläutert wird. Dieses Kapitel besteht wiederum aus mehreren Unterkapiteln, in welchen auf die Kernidee und die Funktionsweise des Tools, den aufgetretenen Herausforderungen und einer Reflexion zum Evaluierungsalgorithmus eingegangen wird. Abgeschlossen wird die Arbeit von einem Fazit, in welchem die gesamte Arbeit reflektiert wird und einen Ausblick auf mögliche weitere Forschungen an diesem Thema präsentiert.

1.4 Bisherige Erkenntnisse über CAPT

CAPT, also Computer-Assisted Pronunciation Training, ist ein Begriff der Tools zusammenfasst, mit denen Schülern ihre Aussprache am Computer üben können sollen [Pennington, 1999]. Es gab bereits in der Vergangenheit Versuche solche Tools zu entwickeln. Obwohl sich manche Tools auf technischer Ebene durchaus bewähren konnten [Tsai, 2019], liefern sie noch nicht genug gute Ergebnisse, um allein (also ohne Supervision durch einen Sprachlehrer) den Schülern die richtige Aussprache beibringen zu können [Luo, 2016] und [Engwall und Bälter, 2007]. Es besteht also weiterhin Nachfrage nach einem System, dass diese Lücke schliessen könnte, was auch das Ziel des Tools ist, das ich im Rahmen dieser Bachelorarbeit entwickelt habe, zumindest für Vokale.

2 Theoretische Grundlagen

Die Disziplinen der Phonetik und Phonologie werden im Alltag häufig durcheinandergebracht. Historisch betrachtet ist die Phonetik etwas älter und hat ihre wissenschaftlichen, methodischen Ursprünge im 18. Jahrhundert [Mol, 1970], wobei bereits weitaus früher Menschen in verschiedenen Teilen der Welt an ihr forschten, wenngleich nicht als eigenständige Disziplin. Sie befasst sich mit den physikalischen Aspekten des Sprechens und der Sprechorgane und verwendet deshalb naturwissenschaftliche Methoden, während die Phonologie als ein Gebiet innerhalb der Sprachwissenschaft sich geisteswissenschaftlichen Methoden bedient [Pompino-Marschall, 2009]. Diese Arbeit lässt sich an der Schnittstelle zwischen Phonetik und Phonologie ansiedeln, da versucht wird, die phonologische Einteilung verschiedener Laute in Vokale mittels phonetischer Mittel (der Formantenanalyse) zu quantifizieren.

2.1 Stimme und Laute

Die menschliche Stimme entsteht, wenn aus der Lunge strömende Luft die Stimmlippen im Kehlkopf zum Schwingen bringt, welche einen für den Menschen hörbaren Schall erzeugt. Weiter moduliert wird dieser Schall durch die Beschaffenheit des Mundraums, der Nasenhöhle und des Rachens. Tritt dieser Schall weitgehend ungehindert aus dem Mund sprechen wir von *Vokalen* oder genauer gesagt von *Vokallauten*, tritt eine Verengung des Vokaltraktes (z.B. durch die Lippen) auf, sprechen wir von *Konsonanten* [Rosen und Howell, 2011] und [Machelett, 1996]. Es ist beispielsweise nicht möglich ein Phon, in der deutschen Sprache repräsentiert durch den Buchstaben *A*, mit geschlossenen Lippen zu erzeugen und umgekehrt ist es nicht möglich, den durch *P* repräsentierten Phon mit geöffneten Lippen zu erzeugen. Im Rahmen dieser Arbeit interessieren nur die Vokale.

2.2 Vokale

Das Vokalsystem einer Sprache kann mit Hilfe eines sogenannten Vokaltrapez visuell dargestellt werden, das die Zungenposition im Mund bei der Bildung verschiedener Vokale verbildlicht. Es besteht aus zwei Dimensionen, wobei die horizontale Achse die Zungenposition (entweder vorne oder hinten im Mund) darstellt und die vertikale Achse die Höhe der Zunge darstellt, oder in anderen Worten, den Öffnungsgrades des Mundes (nicht der Lippen) [Pompino-Marschall, 2009]. Das Ganze sieht dann folgendermassen aus:

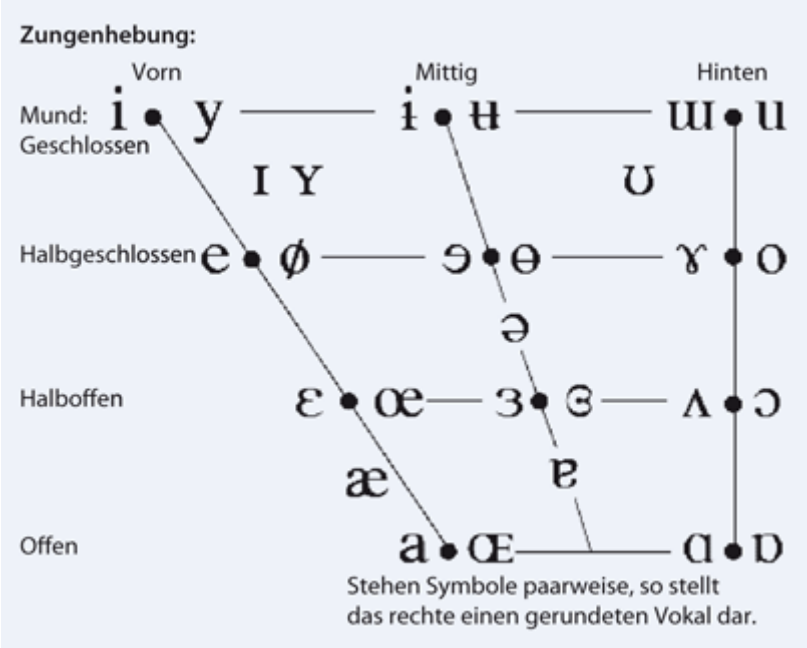

Abb. 1: Vokaltrapez der deutschen Sprache [Ptok, 2009]

Oder noch etwas eindrücklicher dargestellt, im menschlichen Mundraum gezeichnet:

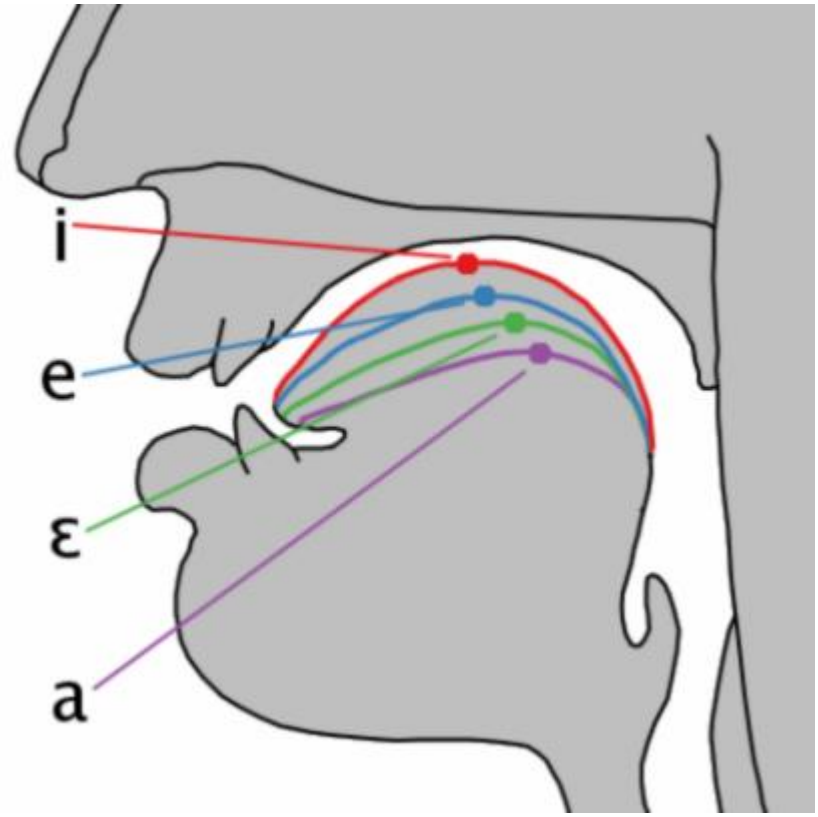

Abb. 2: Zungenpositionen verschiedener Vokale [Jones, 1972]

2.2.1 Quantifizierung durch Formanten

Wie eben gezeigt, lässt sich also jeder Vokal anhand der Zungenposition erkennen. Und da wie in Kapitel 2.2 festgestellt, die Zunge Teil des menschlichen Stimmwerkzeugs und somit auch Teil der Schallmanipulation ist, kann jeder Vokal, von dem man den Obertonbereich misst, mit drei Grössen, den sogenannten Formanten, wiedergegeben werden [Maurer, 2016]:

F1: die Zungenhöhe
F2: die Position der Zunge hinten oder vorne im Mund
F3: die Lippenrundung

Für die Klassifikation von Vokalen sind vor allem die F1 und F2-Werte aussagekräftig und somit für diese Arbeit von besonderem Interesse. Graphisch sieht das Ganze dann folgendermassen aus:

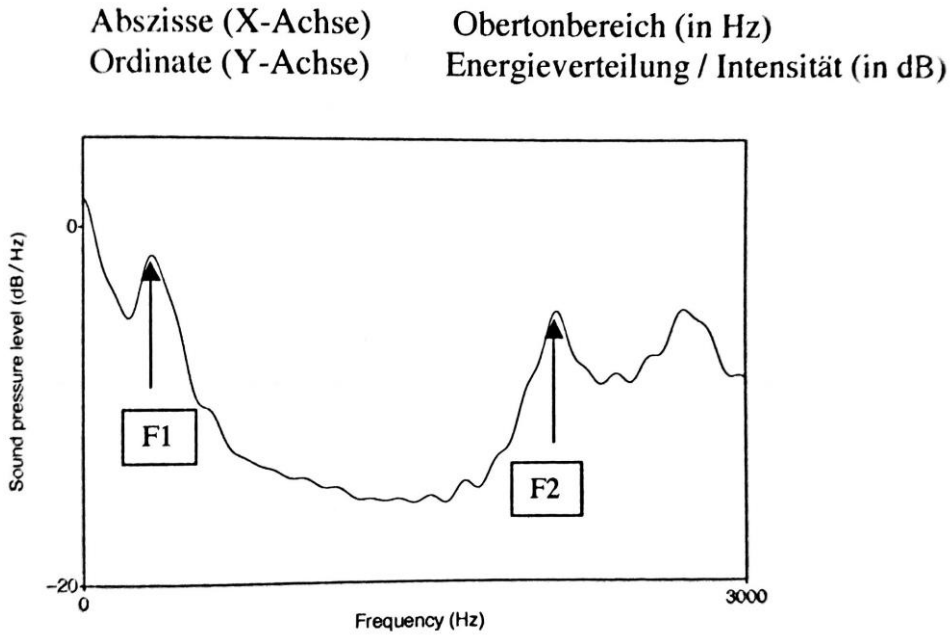

Abb. 3: Spektrum des Vokals [e] [Schmid, 2018]

Wenn man nun auf empirische Weise die Formantwerte verschiedener Vokale von verschiedenen Sprechern ermittelt, kann man eine kleine Datenbank für eine Sprache erstellen, in der die durchschnittlichen Formantwerte einer Gruppe von Muttersprachlern enthalten sind (mit der möglichen Ausnahme von Dialekten):

Vokal	i	y	ɪ	ʏ	e	ø	ɛ	œ
Beispiel	Kiel	fühlen	Fisch	Bitte	Steg	Flöte	Bett	können
2. Formant	2400	2000	2200	1800	2100	1700	1900	1550
1. Formant	275	275	325	325	375	375	500	500
Vokal	ə	a	ɑ	ɔ	o	ʊ	u	
Beispiel	Mitte	Mann	Tat	hoffen	Ton	Mutter	Mut	
2. Formant	1200	1400	1150	900	850	850	750	
1. Formant	500	800	750	500	375	325	275	

Tabelle 1: Formanten der deutschen Sprache [Nawka und Wirth, 2008]

Diese quantifizierte Klassifikation von Vokalen mittels Formanten ist der Schlüssel, mit dem *Pronunciation-Checker* funktioniert.

3 Pronunciation-Checker

Pronunciation-Checker ist der Name der Applikation, die im Rahmen dieser Bachelorarbeit entwickelt wurde. Wie der Name bereits andeutet, soll das Programm, gegeben einen sprachlichen Input, die Aussprache des Gesagten evaluieren, in diesem Fall die Aussprache von verschiedenen Vokalen der deutschen Sprache. Programmiert wurde es in Python.

3.1 Mögliche Use Cases

Das Tool richtet sich an Sprachschüler, die selbstständig ihre Aussprache der deutschen Vokale verbessern möchten. Da das Tool keinerlei linguistische oder programmiertechnische Anwenderkenntnisse erfordert, ist es für eine sehr grosse Userbase geeignet. Insbesondere für Sprachschüler die in einem nicht deutschsprachigen Land wohnen, keinen Zugang zu einem Sprachlehrer haben oder sich leisten können und somit kein Feedback zu ihrer Aussprache erhalten können, könnte *Pronunciation-Checker* hilfreich sein. Ein weiterer denkbarer Use Case sind zum Beispiel Flüchtlinge in Deutschland, Österreich oder in der Schweiz, für die es oftmals an Sprachlehrern mangelt, sich erst recht keinen leisten können und häufig sozial von den Muttersprachlern dieser Länder isoliert sind, also die korrekte Aussprache nicht genug zu hören kriegen.

3.2 Dependencies

Pronunciation-Checker benötigt mehrere, in Python nicht vorinstallierte Libraries. Hier soll nur auf die wichtigsten eingegangen werden, eine vollständige Liste findet sich im Programmcode im Anhang dieser Arbeit. Für die Extraktion der Formanten aus den Aufnahmedateien wurde *Parselmouth*[1] verwendet. *Parselmouth* ist eine open-source Library für Python, die die Funktionalitäten von *Praat*[2] direkt in der Python-Umgebung nutzbar macht [Jadoul et al., 2018]. Für die Berechnungen im Hintergrund wird ausserdem die *SciPy*[3] Library verwendet. Das graphische Interface wurde mit *Tkinter*[4] erstellt, einem GUI-Toolkit für Python.

3.3 Funktionsweise

Im Folgenden soll die Funktionsweise des Programms schematisch erläutert werden:

[1] https://github.com/YannickJadoul/Parselmouth
[2] Praat ist ein open-source Programm für phonetische Analysen
[3] https://github.com/scipy/scipy
[4] https://github.com/python/cpython/tree/master/Lib/tkinter

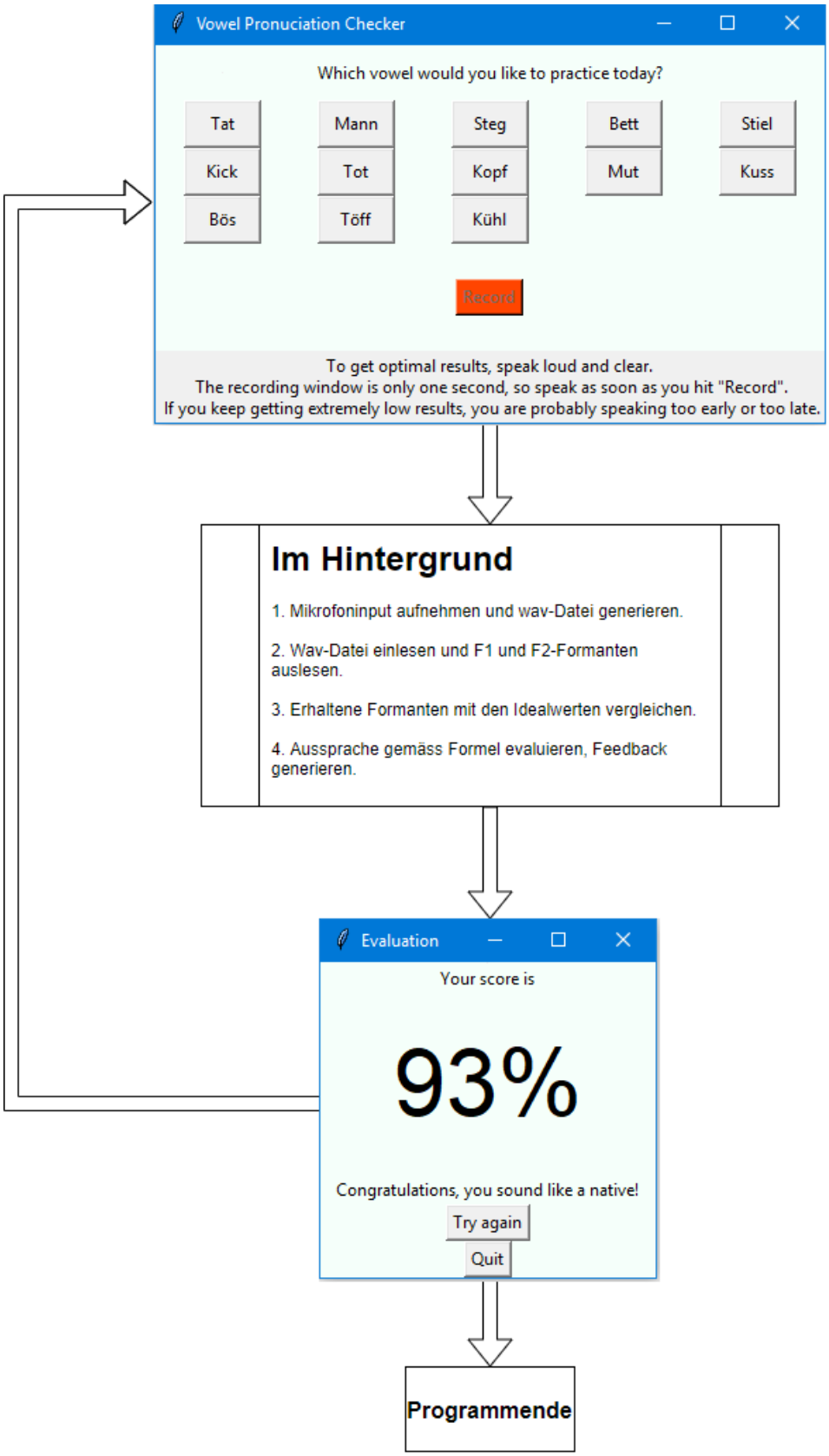

Abb. 4: Schematische Darstellung eines Programmdurchgangs

Nach dem Start des Programms wird der User mit einem ersten Fenster konfrontiert, in welchen er zwischen verschiedenen Vokalen zum Üben wählen kann. Hierbei wird auch unterschieden, ob ein Vokal kurz oder lang ausgesprochen wird, etwa "Mann" vs. "Tat", da die Formanten hier jeweils auch unterschiedlich sind (Vergleiche 2.2.1, Tabelle 1). Jeder Vokal wird von einem Beispielwort repräsentiert. Beim Auswählen eines Wortes, wird das Wort dem User vorgesagt, sodass er weiss, wie sich die korrekte Aussprache anhört. Nach einem Klick auf den Record Button, wird die Stimme des Users aufgenommen. Die erzeugte wav-Datei wird nun eingelesen und mittels *Parselmouth* werden die Formanten ermittelt. Daraufhin wird die Qualität der Aussprache im Hintergrund evaluiert[5]. Ohne Verzögerung wird daraufhin dem User ein Feedback generiert, das seine Aussprache des Vokals auf einer Skala von 0% bis 100% bewertet und in eine von drei möglichen Kategorien einteilt:

80% bis **100%**: Muttersprachliche Aussprache
70% bis **80%**: Befriedigende Aussprache
0% bis **70%**: Unbefriedigende Aussprache

Für den dramatischen Cut-off-Wert bei 70% gibt es einen guten Grund[6]. Je nachdem welche Kategorie die Aussprache erreicht, erhält der Nutzer eine von den drei möglichen, oben erwähnten Bewertungen.
Anschliessend kann der User entscheiden, ob er den Vokal (oder einen anderen Vokal) nochmals üben möchte oder ob er das Programm beenden will.

3.3.1 Evaluation

Gegeben zweier Target-formants[7] eines beliebigen Vokales (entnommen aus 2.2.1, Tabelle 1) und zweier User-formants desselben Vokales (ermittelt mit *Pronunciation-Checker*), habe ich folgenden Evaluierungsalgorithmus für die Aussprache des Vokals entwickelt:

Target-formant F1: T_1
Target-formant F2: T_2
User-formant F1: U_1
User-formant F2: U_2

F1-Score[8]: α
F2-Score: β
Total-Score: γ

[5] mehr dazu in 3.3.1
[6] mehr dazu in 3.4.3.1
[7] auf die Problematik von "idealen" Formanten wird in 3.4.3 und 3.5 eingegangen
[8] Nicht zu verwechseln mit dem im Englischen fast gleichnamigen *F1 score* aus der Statistik

1. Um den F1-Score α zu erhalten, berechne:

$$\alpha = \frac{U_1}{T_1} \cdot 100$$

2. wenn $\alpha > 100$ dann definiere α neu:

$$\alpha = 200 - \frac{U_1}{T_1} \cdot 100$$

3. Um den F2-Score β zu erhalten, berechne:

$$\beta = \frac{U_2}{T_2} \cdot 100$$

4. wenn $\beta > 100$ dann definiere βneu:

$$\beta = 200 - \frac{U_2}{T_2} \cdot 100$$

5. Um den Total-Score γ zu erhalten, berechne:

$$\gamma = \frac{1}{2} \cdot (\alpha + \beta)$$

6. wenn $\gamma < 0$ dann definiere γ neu:

$$\gamma = 0$$

Die Neudefinierung von α und β in Schritt 2 respektive 4 erfolgt aus einem einfachen Grund: Berechnet soll nämlich die Nähe von U_1 zu T_1, also wie nahe der User-formant am Target-formant ist. Je näher, desto besser. Ist nun der Wert eines User-formants höher als der des Target-

formants, schiesst er sozusagen übers Ziel hinaus, würde dies ohne Korrektur einen höheren F1-score ergeben, als wenn User- und Target-formants identisch wären. Ein rechnerisches Beispiel:

$T_1 = 200$
$U_1 = 200$

$$\alpha = \frac{U_1}{T_1} \cdot 100 = 100\%$$

$T_1 = 200$
$U_1 = 250$

$$\alpha = \frac{U_1}{T_1} \cdot 100 = 125\%$$

Wie wir sehen erhalten wir bei der zweiten Gleichung einen höheren F1-score (α) als bei der ersten Gleichung, obwohl bei der zweiten Gleichung die User- und Target-formants viel näher beieinander sind. Mit der Korrektur erhielte die zweite Gleichung folgenden F1-score:

$$\alpha = 200 - \frac{U_1}{T_1} \cdot 100 = 75\%$$

Läge U_1 bei 150, würde der F1-score ebenfalls 75% betragen, denn die Differenz zwischen 150 und 200, und 250 und 200 ist ja beide Male gleich gross. Somit ist eine lineare Korrektur bei U_1 -Werten über dem Target-Wert gewährleistet.
Diese Korrektur wird auch auf identische Art- und Weise auf den F2-Score angewendet.

Die Neudefinierung von γ in Schritt 6 findet statt um die Rückgabe von negativen Total-Scores zu vermeiden. Diese würden den User nur verwirren, so ist ein Score von beispielsweise -23% weniger eindeutig als ein Score von 0%. Dies kann geschehen, wenn entweder einer oder beide der Formant-Scores negativ waren und das arithmetische Mittel beider einen negativen Wert ergibt[9]. In diesem Fall wird das unterste Limit einfach bei 0% angesetzt, was viel verständlicher für den User ist.

3.4 Herausforderungen

Bei der Entwicklung von *Pronunciation-Checker* traten zahlreiche Herausforderungen und Stolpersteine auf. Manche von ihnen konnten nach viel Kopfzerbrechen durch wenige Zeilen Code überwunden werden, aber manche sind fundamentale Schwierigkeiten, wenn es um Formanten

[9] Der User kriegt den F1- und F2-Score nie zu sehen, da sie allein wenig oder gar keine Aussagekraft über die Aussprache des Vokals haben

und Aussprache im Allgemeinen gilt. Im folgenden Unterkapitel sollen ein paar von ihnen näher beschrieben werden.

3.4.1 Technische Limitation von Parselmouth bzw. Praat

Eine bisher ungelöste Schwäche von *Pronunciation-Checker* ist die korrekte Identifikation und Bewertung der Vokalen [u] und [o]. In etwa der Hälfte der Fälle werden die Vokale korrekt identifiziert, in unregelmässigen Abständen wird aber konstant ein Score von 0% zurückgegeben, obwohl die Aussprache des Vokals korrekt war. Nach vielen vergeblichen Schrauben am Evaluierungsalgorithmus und Herumexperimentierens mit dem Cut-off-Wert, erwähnte meine Betreuungsperson Prof. Dr. Volker Dellwo, dass Praat notorisch dafür ist, Mühe bei der korrekten Erkennung von Formanten hat, wenn diese nahe beieinander liegen. Und genau dies ist bei den Vokalen [u] und [o] der Fall, siehe 2.3.1, Tabelle 1. So kommt es vor, dass Praat die Erhebungen des ersten und zweiten Formanten, als die des ersten interpretiert und diejenige des dritten als die des zweiten. Das natürlich führt dazu, dass der Evaluierungsalgorithmus mit unbrauchbaren Parametern gefüttert wird, der dann wiederum einen Score von 0% zurückgibt.

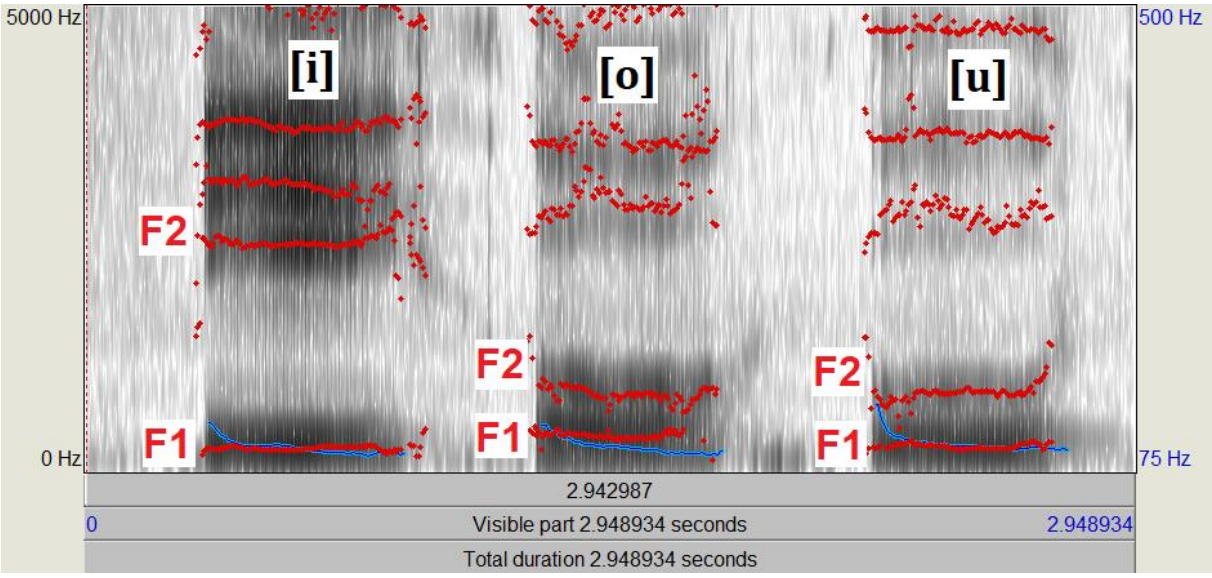

Abb. 5: Mit Praat generierte Spektrogramme dreier Vokale, inklusive Formanten

Gegen diese Vermutung spricht allerdings, dass die Formanten des Vokals [a] im Schnitt noch näher beieinander liegen, als diejenigen von [o] und [u], das Programm aber meist kein Problem hat [a] richtig zu identifizieren. Im Gespräch mit Volker Dellwo mutmassten wir auch die tiefere Energiekonzentration der Formanten von beispielsweise [u] im Vergleich zu [a], was aber bei [i] ebenfalls der Fall ist. Gesichertes liess sich nicht feststellen

3.4.2 Physiologische Unterschiede im Vokaltrakt zwischen Männern, Frauen und Kindern

Es gibt physiologische Unterschiede zwischen den Vokaltrakten von Männern und Frauen, welche Einfluss auf die Formanten hat. Genauer gesagt haben Frauen im Schnitt kürzere Stimmlippen[10] als Männer und Kinder nochmal kürzere als Frauen [Schmid, 2018]. Dies erzeugt wiederum höhere Stimmen. Die durchschnittliche Stimmsprechlagen sind [Schmid, 2018]:

Männer	100-150 Hz
Frauen	200-300 Hz
Kinder	300-450 Hz

Was auch für im Schnitt bei Frauen höhere Formant-Werte bedeutet als bei Männern [Peterson und Barney, 1952]:

		i	ɪ	ɛ	æ	ɑ	ɔ	ʊ	u	ʌ	ɝ
Fundamental frequencies (cps)	*M*	136	135	130	127	124	129	137	141	130	133
	W	235	232	223	210	212	216	232	231	221	218
	Ch	272	269	260	251	256	263	276	274	261	261
Formant frequencies (cps)											
F_1	*M*	270	390	530	660	730	570	440	300	640	490
	W	310	430	610	860	850	590	470	370	760	500
	Ch	370	530	690	1010	1030	680	560	430	850	560
F_2	*M*	2290	1990	1840	1720	1090	840	1020	870	1190	1350
	W	2790	2480	2330	2050	1220	920	1160	950	1400	1640
	Ch	3200	2730	2610	2320	1370	1060	1410	1170	1590	1820
F_3	*M*	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690
	W	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960
	Ch	3730	3600	3570	3320	3170	3180	3310	3260	3360	2160
Formant amplitudes (db)	L_1	−4	−3	−2	−1	−1	0	−1	−3	−1	−5
	L_2	−24	−23	−17	−12	−5	−7	−12	−19	−10	−15
	L_3	−28	−27	−24	−22	−28	−34	−34	−43	−27	−20

Tabelle 2: Durchschnittliche Formantwerte von Männern (M), Frauen (W) und Kindern (CH) verschiedener Vokale des Englischen [Peterson und Barney, 1952]

In der Praxis kann dieses Problem gelöst werden indem man für die Target-Werte auf Datensammlungen zurückgreift in denen Männer und Frauen berücksichtigt wurden (oder einfach den Schnitt zweier Sammlungen nimmt in denen jeweils nur Frauen respektive nur Männer berücksichtigt wurden). User mit überdurchschnittlich hohen oder tiefen Stimmen könnten aber Probleme mit dem Programm bekommen, wenn das Vokal fälschlicherweise als nicht korrekt ausgesprochen identifiziert werden würde, obwohl ein Muttersprachler zum gegenteiligen Schluss käme. In meinen nicht repräsentativen Versuchen mit etwa einem Dutzend

[10] bei Frauen ca.1.7cm - 2cm, bei Männern ca. 2cm - 2.4cm

Usern, davon etwa die Hälfte männlich und die andere Hälfte weiblich, konnte ich niemanden finden bei dem das Programm systematisch Mühe hatte.

3.4.3 Variabilität der Formanten

Ein etwas grundlegenderes Problem kam mit der 1952 von Peterson und Barney beschriebenen Variabilität der Formanten auf. Nach der Analyse 76 englischer Muttersprachler in Amerika fanden sie heraus, dass es innerhalb der gleichen Sprache (und Dialektes) zu zum Teil erheblichen Differenzen innerhalb der Formanten kommen kann:

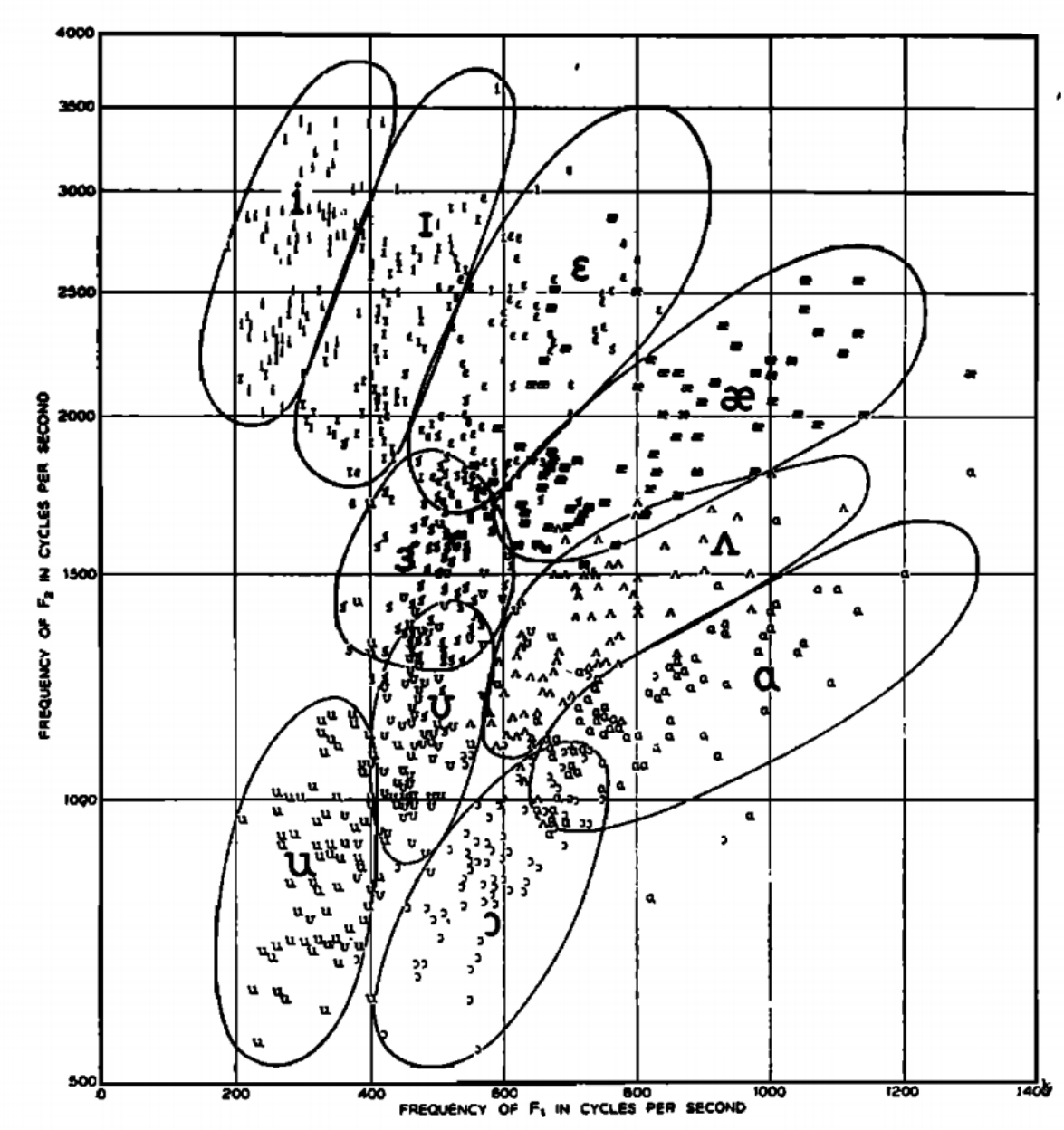

Abb. 6: F1 und F2-Formaten von 76 Sprechern verschiedener englischer Vokale [Peterson und Barney, 1952]

Trotz dieser Unterschiede ist die Aussprache dieser Vokale in der subjektiven Hörwahrnehmung auf muttersprachlichem Niveau - es waren ja alles Muttersprachler. Erwartet worden war, dass die Formanten der Muttersprachler wesentlich näher beieinander liegen würden als sie dies in der Realität tatsächlich taten. Gleichzeitig gibt es viele Überlappungen (siehe in Abb. 6 die sich überlappenden Kreise), also unterschiedliche Vokale, bei welchen dieselben oder sehr ähnliche Formanten gemessen wurden. Bedeutet dies nun, dass der *Pronunciation-Checker* und diese ganze Arbeit auf einer falschen Prämisse beruhen? Nicht wirklich.

3.4.3.1 False Negatives, False Positives und Cut-off-Werte

Denn im Grunde genommen bedeutet die Variabilität der Formanten lediglich, dass wir den Bewertungsmassstab anpassen müssen. In anderen Worten, wie bereits in Kapitel 3.3 erwähnt, muss der Cut-off-Wert für eine nicht akzeptable Aussprache, also eine nicht mehr akzeptable Differenz zwischen Target- und User-formant nach unten gesetzt werden. Das Feedback muss also mehr vergeben, um *False Negatives* zu vermeiden. Gleichzeitig darf der Cut-off-Wert nicht zu weit nach unten gesetzt werden, da sonst viele *False Positives* in Kauf genommen werden würden. Zur Erinnerung:

True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
Korrekt ausgesprochene Vokale und vom Programm als korrekt identifiziert	Falsch ausgesprochene Vokale, aber vom Programm als korrekt ausgesprochen identifiziert	Falsch ausgesprochene Vokale und vom Programm als falsch ausgesprochen identifiziert	Korrekt ausgesprochene Vokale, aber vom Programm als falsch ausgesprochen identifiziert

Tabelle 3: TP, FP, TN und FN samt Erklärung

In der Praxis bewährte sich 70% als guter Kompromiss, mit dem sich die meisten *False Negatives* vermeiden lassen, ohne zu viele *False Positives* in Kauf zu nehmen. Fehlerhafte Klassifikation von Vokalen ist übrigens nicht nur ein Problem das *Pronunciation-Checker* betrifft, sondern auch zwischen Muttersprachlern selbst:

	i	ɪ	ɛ	æ	ɑ	ɔ	ʊ	u	ʌ	ɝ
i	10267	4	6	...	...	3	...	...	...	...
ɪ	6	9549	694	2	1	1	...	...	...	26
ɛ	...	257	9014	949	1	3	...	...	2	51
æ	...	1	300	9919	2	2	...	...	15	39
ɑ	...	1	...	19	8936	1013	69	...	228	7
ɔ	...	...	1	2	590	9534	71	5	62	14
u	...	...	1	1	16	51	9924	96	171	19
u	...	...	1	...	2	...	78	10196	...	2
ʌ	...	1	1	8	540	127	103	...	9476	21
ɝ	...	...	23	6	2	3	...	...	2	10243

Tabelle 4: Wahrnehmungsunterschiede verschiedener englischer Vokale. Senkrechte Achse: Vom Sprecher beabsichtigter Vokal. Horizontale Achse: Vom Hörer wahrgenommener Vokal [Peterson und Barney, 1952]

Hören sich Muttersprachler isolierte Vokale anderer Muttersprachler an, kommt es zum Teil zu erheblichen Missverständnissen, also Vokale die vom Zuhörer als andere Vokale identifiziert werden, als der Sprecher beabsichtigte. In der Praxis fällt das aber kaum auf, da Kontext, gegeben durch den Sprachinhalt, die Ambiguitäten auflöst.

3.5 Reflektion zur Aussagekraft des Evaluierungsalgorithmus

Wie in den vorangehenden Kapiteln beschrieben, gehört ein gewisses Fingerspitzengefühl und Trial & Error um ein aussagekräftiges, hilfreiches Feedback zurückgeben zu können. Der in 3.3.1 beschriebene Evaluierungsalgorithmus vergleicht ja absolute Werte zueinander, unter der Annahme, dass ein User-formant möglichst nahe an einem absoluten, “idealen” Target-formant liegen sollte. Wie in 3.4.2 aber jedoch dargelegt, gibt es eigentlich den idealen Wert für Target-formants gar nicht. Möglicherweise ist der Vergleich zu absoluten Werten gar nicht der geeignetste Weg, um die Aussprachequalität eines Vokales zu beurteilen. Hiermit wird auch die Brücke zum Fazit inklusive Ausblickes dieser Arbeit geschlagen.

4 Fazit

In dieser Arbeit wurde der Frage auf den Grund gegangen, ob man mittels Formantenanalyse ein Programm entwickeln kann, dass Sprachschülern beim Erlernen der richtigen Aussprache deutscher Vokale unterstützen kann. Nach einer Auseinandersetzung mit der Theorie aus der Phonetik, im Speziellen Vokalbildung und Formanten, wurde festgestellt, dass zumindest in der Theorie jeder Vokal anhand zwei charakteristischer Formantwerte identifiziert werden kann. In der Praxis bewährte sich diese Idee, ich entwickelte das Programm *Pronunciation-Checker*, das die Stimme eines Users über ein Mikrofon aufnimmt, die Formanten daraus ausliest und mit den Target-Formanten vergleicht. Mit meinem Evaluierungsalgorithmus wird dem User daraufhin ein Feedback zu seiner Aussprache zurückgegeben. Hierbei vergleicht der Algorithmus zuerst die Ratio, also die Ähnlichkeit, des ersten und des zweiten User-Formanten mit denen der jeweiligen Target-Formanten für den jeweiligen Vokal. Der Durchschnitt der somit ermittelten Scores ergibt dann einen Gesamtscore, der dann entscheidet, wie gut die Aussprache des Vokals war und welches Feedback dem User zurückgegeben wird. Dies funktioniert für die Vokale des Deutschen sehr gut, nur das [o] und das [u] werden gelegentlich nicht richtig erkannt. Die Forschungsfrage lässt sich also mit "Ja" beantworten, mittels Formantenanalyse lässt sich Software entwickeln, die Sprachschülern beim Erlernen deutscher Vokale helfen kann.

Als nächstes gilt es empirisch zu ermitteln, wie gut das Programm tatsächlich die Vokale verschiedener Sprecher erkennt. Dafür müsste eine Studie mit einem genügend grossen Stichprobenumfang durchgeführt werden, denn wie in 3.4.3 dargelegt wurde, gibt es ein ideales Paar von Formantwerten für ein bestimmtes Vokal eigentlich gar nicht. Ebenfalls interessant wäre es einen Algorithmus zu entwickeln, der statt absolute Formantwerte als ideal zu erachten, die Differenz zwischen den Formanten als entscheidendes Merkmal hinzuzieht (bspw. würden die User-Formanten 500 und 1000 einen Volltreffer erzielen, wenn die Target-Formanten 1500 und 2000 betragen, da die Differenz beide Male 500 beträgt). Auch hier müsste empirisch untersucht werden, welcher Algorithmus besser abschneidet.

Darüber hinaus muss ebenfalls anerkannt werden, dass beim momentanen state-of-the-art-approach, wenn es um Vokale und Spracherkennung im Allgemeinen geht, nicht reine Formantenanalyse, sondern *Mel Frequency Cepstral Coefficients* gebraucht werden (kurz MFCC) [Kathiresan et al., 2017]. Ein genaues Auseinandersetzen mit den MFCC geht über den Rahmen dieser Arbeit hinaus, für allfällige Arbeiten, die auf den Erkenntnissen dieser Arbeit aufbauen wollen, ist ein Blick auf die Möglichkeiten die MFCC bieten definitiv wert. Zum Beispiel könnte mit

diesen bessere Ergebnisse bei der Identifikation der [u] und [o] Vokale erzielt werden, als mit der Formantenanalyse die in *Pronunciation-Checker* verwendet wird.

Glossar

CAPT: Computer-Assisted Pronunciation Training.

Formant: Akustische Energiekonzentration bei der Lautbildung.

Phon: Linguistische Lauteinheit.

Phonetik: Naturwissenschaftliche Erforschung der gesprochenen Sprache.

Phonologie: Geisteswissenschaftliche Erforschung der gesprochenen Sprache.

Stimmlippen: Schwingender Teil des Kehlkopfes, Teil der eigentlichen Stimmbildung.

Vokaltrakt: Über dem Kehlkopf gelegener Teil des Sprechvorganges (Mundraum, Nasenraum, Zunge, Lippen, etc.

Vokaltrapez: Vokalsystems einer Sprache, visuell dargestellt.

Abbildungsverzeichnis

Tabellenverzeichnis

Bibliografie

[Engwall und Bälter, 2007] Engwall, O. und Bälter, O. (2007). *Pronunciation feedback from real and virtual language teachers.* In: *Computer Assisted Language Learning,* Volume 20, Issue 3.

[Jadoul et al., 2018] Jadoul, Y., Thompson, B. und de Boer, B. (2018). *Introducing Parselmouth: A Python Interface to Praat.* In: *Journal of Phonetics*, Volume 71, Issue 11.

[Jones, 1972] Jones, D. (1972). *An outline of English phonetics.* Cambridge University Press.

[Kathiresan et al., 2017] Kathiresan, T., Maurer, D., Suter, H. und Dellwo, V. (2017). *Enhancing the objectivity of interactive formant estimation: Introducing euclidean distance measure and numerical conditions for numbers and frequency ranges of formants.* In*: Elektronische Sprachsignalverarbeitung 2017.* TUDpress.

[Luo, 2016] Luo, B. (2016). *Evaluating a computer-assisted pronunciation training (CAPT) technique for efficient classroom instruction.* In: *Computer Assisted Language Learning,* Volume 29, Issue 3.

[Machelett, 1996] Machelett, K. (1996). *Das Lesen von Sonagrammen.* Institut für Phonetik und Sprachliche Kommunikation der Universität München.

[Maurer, 2016] Maurer, D. (2016). *Acoustics of the Vowel.* Peter Lang Verlag.

[Mol, 1970] Mol, H. (1970). *Fundamentals of Phonetics II: Acoustical Models Generating the Formants of the Vowel Phonemes*. Mouton.

[Nawka und Wirth, 2008] Nawka, T., Wirth, G. (2008). *Stimmstörungen: Lehrbuch für Ärzte, Logopäden, Sprachheilpädagogen und Sprechwissenschaftler.* Deutscher Ärzteverlag.

[Pennington, 1999] Pennington, M. (1999). *Computer-Aided Pronunciation Pedagogy: Promise, Limitations, Directions.* In: *Computer Assisted Language Learning,* Volume 12, Issue 5.

[Peterson und Barney, 1952] Peterson, G. Und Barney, H. (1952). *Control Methods Used in a Study of the Vowels*. In: *The Journal of the Acoustic Society of America*, Volume 24, Issue 2.

[Pompino-Marschall, 2009] Pompino-Marschall, B. (2009). *Einführung in die Phonetik.* Walter de Gruyter.

[Ptok, 2009] Ptok, M. (2009). *Sprachlaute und ihre Darstellung für die Diagnostik.* In: *HNO*, Volume 57, Issue 10.

[Rosen und Howell, 2011], Rosen, S. und Howell, P. (2011). *Signals and Systems for Speech and Hearing.* Emerald Publishing.

[Schmid, 2018] Schmid, S. (2018). *Einführung in die allgemeine Phonetik und Phonologie für Studierende der Romanist*ik. Phonetisches Laboratorium der Universität Zürich.

[Tsai, 2019] Tsai, P. (2019). *Beyond self-directed computer-assisted pronunciation learning: a qualitative investigation of a collaborative approach.* In: *Computer Assisted Language Learning,* Volume 32, Issue 7.

Python Skripte

`app.py`

Dieses Skript führt *Pronunciation-Checker* aus. Es beinhaltet die Aufnahme der Stimme, die Bewertung der Aussprache sowie ein grafisches Interface.

Input: Keinen, generiert eigenständig ein wav-File mit der Stimme des Users.

Output: wav-File sowie Aussprachebewertung des Vokals.

`app_simple.py`

Dieses Skript führt ebenfalls *Pronunciation-Checker* aus, aber in einem CLI statt graphischen Interface, falls **`app.py`** nicht richtig funktionieren sollte. Manche Funktionen von **`app.py`** wie die gesprochenen Beispielwörter sind nicht enthalten.

Input: Keinen, generiert eigenständig ein wav-File mit der Stimme des Users.

Output: wav-File sowie Aussprachebewertung des Vokals.

Lebenslauf

Persönliche Angaben

Alon Cohen
16-730-665
8600 Dübendorf
alon.cohen@uzh.ch

Akademische Bildung

seit 2016	Bachelorstudium Computerlinguistik & Sprachtechnologie, Geschichte der Neuzeit und Soziologie an der Universität Zürich

Berufliche und nebenberufliche Tätigkeiten

seit November 2019	Studentische Assistenz beim Phonogrammarchiv der Universität Zürich
März - Juni 2019	App Entwickler bei der Vox-Sprachschule
Februar - Juni 2019	Internes CL-Praktikum: Tutoratsleitung für die Veranstaltung *Einführung in die Computerlinguistik II* von Dr. Manfred Klenner

Selbstständigkeitserklärung

Universität Zürich[UZH]

Institut für Computerlinguistik

Selbstständigkeitserklärung

Originalarbeit

Ich erkläre ausdrücklich, dass es sich bei der von mir im Frühjahrs-/Herbst-Semester 2019 an der Universität Zürich eingereichten schriftlichen Arbeit mit dem Titel

Aussprachebewertung von Vokalen mittels Formantenanalyse

um eine von mir selbst und ohne unerlaubte Beihilfe sowie in eigenen Worten verfasste Originalarbeit handelt. Sofern es sich dabei um eine Arbeit von mehreren Verfasserinnen oder Verfassern handelt, bestätige ich, dass die entsprechenden Teile der Arbeit korrekt und klar gekennzeichnet und der jeweiligen Autorin oder dem jeweiligen Autor eindeutig zuzuordnen sind.

Ich bestätige überdies, dass die Arbeit als Ganzes oder in Teilen weder bereits einmal zur Abgeltung anderer Studienleistungen an der Universität Zürich oder an einer anderen Universität oder Ausbildungseinrichtung eingereicht worden ist noch inskünftig durch mein Zutun als Abgeltung einer weiteren Studienleistung eingereicht werden wird.

Verwendung von Quellen

Ich erkläre ausdrücklich, dass ich sämtliche in der oben genannten Arbeit enthaltenen Bezüge auf fremde Quellen (einschliesslich Tabellen, Grafiken u. €.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos und nach bestem Wissen sowohl bei wärtlich übernommenen Aussagen (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen anderer Autorinnen oder Autoren (Paraphrasen) die Urheberschaft angegeben habe.

Sanktionen

Ich nehme zur Kenntnis, dass eine Arbeit, welche zum Erwerb eines Leistungsnachweises verwendet wird und sich als Plagiat im Sinne des Dokuments *Erläuterung des Begriffs „Plagiat"*erweist, in leichten Fällen zu Notenabzug führt, in schweren Fällen mit Note 1 (eins) ohne Möglichkeit einer Überarbeitung bewertet werden kann und in ganz gravierenden Fällen die entsprechenden rechtlichen und disziplinarischen Konsequenzen nach sich ziehen kann (gemäss §§ 7ff der Disziplinarordnung der Universität Zürich sowie § 36 der Rahmenordnung für das Studium in den Bachelor- und Master-Studiengängen der Philosophischen Fakultät der Universität Zürich).

Ich bestätige mit meiner Unterschrift die Richtigkeit dieser Angaben.

Name: Cohen Vorname: Alon

Matrikelnummer: 16-730-665

Datum: 01.12.2019 Unterschrift:

Seite 1/1