

Reimplementierung des Formantsynthese- Programms KPE

Studienarbeit Nr. 47 im Fach Computerlinguistik

von
Andreas Madsack

angefertigt am 31. Oktober 2005

Institut für Maschinelle Sprachverarbeitung
Azenbergstraße 12
D-70174 Stuttgart
Germany

Betreuer: PD Dr. Bernd Möbius, IMS Phonetik

Beginn der Arbeit: Juli 2005
Abgabe der Arbeit: Oktober 2005

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Stuttgart, den 31. Oktober 2005

(Andreas Madsack)

Inhaltsverzeichnis

1 Einführung	1
1.1 Überblick über die verschiedenen Sprachsynthesysteme	1
1.2 Korpusbasierte Verfahren	1
1.2.1 Diphonsynthese	1
1.2.2 Unitsselection	2
1.3 Parametrische Syntheseverfahren	2
1.3.1 Artikulatorische Synthese	2
1.3.2 Formantsynthese	2
2 Die Formantsynthese	3
2.1 Quelle-Filter-Modell	3
2.2 Cascade versus parallel	4
2.3 Klatt80-Diagramm	5
2.4 Lautklassen und ihre Realisierung mit <code>imskpe</code>	5
2.4.1 Vokale	5
2.4.2 Sonoranten	6
2.4.3 Frikative / Affrikate	7
2.4.4 Nasale	7
2.4.5 Plosive	8
3 Überblick über <code>imskpe</code>	9
3.1 Allgemeines	9
3.2 Oberfläche	10
3.2.1 Menü	10
3.2.2 Toolbar	10
3.2.3 PAR-Kontrollbereich	11
3.2.4 Preferences	14
3.2.5 Diagrammfläche	15
3.2.6 Statuszeile	16
3.3 Beispielsitzungen	17
3.3.1 Erstellen des Lautes /a/	17
3.3.2 Erstellen der Lautfolge /halo/	18
3.3.3 Erstellen der Lautfolge /jes/	21
4 Die Zukunft von <code>imskpe</code>	25
4.1 Verbesserungsmöglichkeiten	25
4.2 Weiterentwicklung	25
A Anhang	26
A.1 Aufbau des PAR-Formantes	26
A.2 Die Parameterwerte	27
Literatur	28
Abbildungsverzeichnis	29
Tabellenverzeichnis	29

1 Einführung

Die hier verwendete Formantsyntheseimplementierung wurde zur Verwendung in der Psychologie und den Sprachwissenschaften von Dennis Klatt (1980) als flexibles Forschungswerkzeug entwickelt. Die Qualität der erzeugten Sprachsignale ist deutlich schlechter als mit anderen aktuelleren Sprachsynthesystemen. Sie kommt aber ohne Aufnahmen aus und ist somit komplett synthetisch.

Meine Arbeit bestand darin, eine grafische Oberfläche für den Algorithmus von Dennis H. Klatt zu entwickeln, da die bisher aktuelle Implementierung – das Programm KPE80 von Simpson et al. (1995), entwickelt Anfang der 90er – inzwischen in die Jahre gekommen ist und auf vielen aktuellen Systemen nicht lauffähig ist.

Durch Verwendung von GTK2 (Gimp ToolKit) in meiner Arbeit (`imskpe`) wird Plattformunabhängigkeit¹ erreicht und (hoffentlich) eine Verwendung noch in einigen Jahren sichergestellt. GTK bietet ein komplettes Set von Widgets, um grafische Benutzeroberflächen zu entwickeln. Ein gutes Beispiel für die Möglichkeiten von GTK ist die Desktopumgebung Gnome², die zu großen Teilen auf GTK basiert.

1.1 Überblick über die verschiedenen Sprachsynthesysteme

In den letzten Jahrzehnten haben sich verschiedene Ansätze zum Thema Sprachsynthese herausgebildet.

Grundsätzlich unterscheidet man zwei verschiedene Arten von Sprachsynthesystemen. Die erste Art verwendet vorher aufgenommene Sprachsegmente. Vertreter dieser Kategorie sind die häufig eingesetzte Diphonsynthese (z.B. MBROLA³) und die inzwischen allgemein präferierte Unitselection (z.B. Festival⁴).

Die zweite Art von Systemen, die das Sprachsignal komplett synthetisch erzeugt, ist inzwischen fast ausschließlich von akademischem Interesse und ist vor allem für Perzeptionsexperimente interessant. Die Vertreter aus dieser Kategorie sind die artikulatorische Synthese und die hier verwendete Formantsynthese.

Ein guter geschichtlicher Rückblick über die Sprachsynthese findet sich unter anderem in Lemmetty (1999).

1.2 Korpusbasierte Verfahren

1.2.1 Diphonsynthese

Bei der Diphonsynthese wird das Sprachsignal durch Verkettung von Diphonen erzeugt. Diphone sind Kombinationen aus Nachbarlauten, die meist in der Mitte des Lautes geschnitten werden. Die Prosodieanpassung geschieht durch Signalverarbeitung. Die Diphonsynthese wird und wurde häufig verwendet. Die erzeugten Sprachsignale klingen besser als komplett synthetische, wie sie beispielsweise von der Formantsynthese erzeugt werden, wirken aber immer noch künstlich und haben Verkettungsfehler.

¹GTK läuft unter Linux, allen aktuellen Unix-Varianten, MS Windows, ...

²<http://www.gnome.org>

³<http://tcts.fpms.ac.be/synthesis/>

⁴<http://www.cstr.ed.ac.uk/projects/festival/>

1.2.2 Unitselection

Bei der Unitselection werden aus großen Sprachkorpora während der Laufzeit die Sprachsignale zusammengesetzt. Hierbei wird versucht möglichst lange Lautfolgen auszusuchen, um Verkettungen und somit mögliche Signalstörungen zu minimieren. Der große Nachteil der Unitselection ist, dass alle Daten in den Sprachkorpora im Moment meist händisch annotiert werden müssen. Die Unitselection profitiert zunehmend von großen Sprachkorpora und besseren statistischen und signalverarbeitenden Verfahren.

1.3 Parametrische Syntheseverfahren

1.3.1 Artikulatorische Synthese

Bei der artikulatorischen Synthese wird versucht möglichst genau den menschlichen Sprachapparat nachzubilden, durch Simulation der mechanischen Bewegungen der Artikulatoren, der Luftströme, der Lungenfunktion und der Glottis. Auch aus Mangel an Daten und Komplexität des Problems, ist hier noch einiges an Forschung nötig und gute Ergebnisse sind im Moment nicht vorhanden.

1.3.2 Formantsynthese

In der Formantsynthese wird, ohne menschliche Aufnahmen zu verwenden, Sprache erzeugt. Um dies zu erreichen, wird auf ein akustisches Modell zurückgegriffen: Das Quelle-Filter-Modell. Die hierbei erzeugte synthetische Sprache erreicht nicht die Natürlichkeit die andere, auf menschlichen Aufnahmen basierende Systeme erreichen. Der Vorteil gegenüber korpusbasierenden Verfahren ist, dass durch die vielen Parameter sehr viel eingestellt und kontrolliert werden kann. Genau aus diesem Grund wird die Formantsynthese für Perzeptionsexperimente verwendet.

2 Die Formantsynthese

Vor allem in den 60er, 70er und Anfang der 80er Jahre gab es einige Versuche Sprache künstlich zu erzeugen. Einige versuchten dies mit dem Bau von Hardware Sprachsynthesizierern und zunehmend auch mit Softwarelösungen. Der Algorithmus von Klatt (1980) wird bis heute noch in der Lehre und in manchen Studien verwendet.

2.1 Quelle-Filter-Modell

Die Formantsynthese basiert auf dem Quelle-Filter-Modell aus Fant (1960).

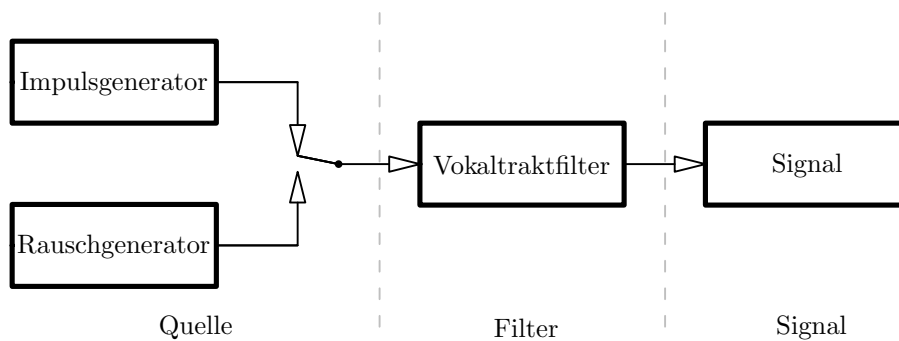


Abbildung 1: Das Quelle-Filter-Modell

Das Quelle-Filter-Modell beruht auf der Idee, dass der menschliche Sprachapparat als Ansatzrohr beschrieben werden kann (siehe Pompino-Marschall, 2003). Dieses Rohr im Modell ist kreiszylindrisch und ca. 17cm lang. Es reicht in der Realität von der Glottis bis zu den Lippen. In diesem Ansatzrohr sind die Wände schallreflektierend, was in der Realität nicht so einfach ist, weil der menschliche Rachenraum nicht absolut schallreflektierend ist.

Es gibt zwei Anregungsmechanismen: Die Phonation und die Friktion. Die Friktion entsteht durch Turbulenzen an einer starken Verengung (z.B. an den Lippen für ein /f/). Bei der Phonation gehen Impulse von den Stimmlippen aus. Wichtig ist hierbei das Öffnen und Schließen der Glottis durch den Bernoulli-Effekt. Die stehenden Wellen, die sich im Ansatzrohr ausbreiten, nennt man Resonanzen. Bei geöffnetem Mund liegen die Resonanzfrequenzen bei 500Hz, 1500Hz und 2500Hz. Berechnet werden diese durch $F_x = \frac{340}{4/N * 0.17}$, wobei für die erste Resonanzfrequenz $N = 1$ ist, für die zweite Resonanzfrequenz $N = 3$ und für die dritte Resonanzfrequenz $N = 5$.

Die Filterung resultiert aus Veränderungen der Resonanzcharakteristik des Ansatzrohres. Dies wird durch die Stellungsänderung der Artikulatoren bewirkt.

Bei Nasalen spielt der Nasalraum eine zusätzliche Rolle. Durch Öffnung des Velums wird das Rohr zwei geteilt. Im Nasenraum werden die höheren Frequenzen stark gedämpft. Der Mundraum ist zu diesem Zeitpunkt geschlossen und erzeugt nun den Antiformant.

2.2 Cascade versus parallel

In Klatt (1980) verweist Dennis Klatt auf viele Synthesizer, die vor seinem entwickelt wurden. Dabei haben sich zwei verschiedene Ansätze als sinnvoll herausgestellt. Zum einen der parallele Formantsynthesizer, bei dem die Formantresonatoren parallel zueinander angeordnet sind und jeder Resonator eine Amplitudensteuerung hat. Mit diesem ist es möglich alle Laute zu erzeugen.

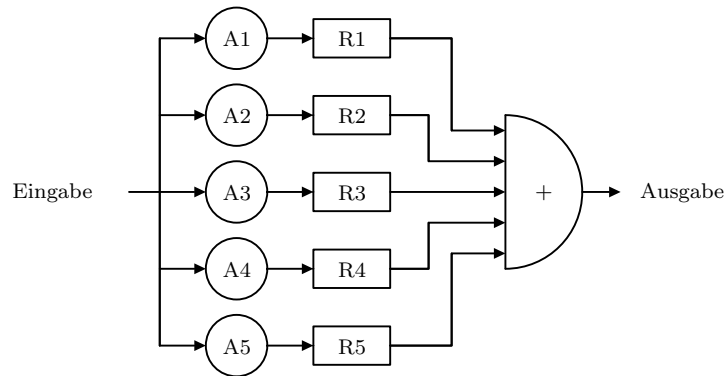


Abbildung 2: Paralleler Zweig

Der zweite Ansatz, der Kaskaden-Formantsynthesizer, ist besser geeignet um Vokale und Sonoranten zu erzeugen. Um Plosive oder Frikative zu erzeugen wird ein paralleler Formantsynthesizer benötigt. Der `klatt80`-Algorithmus, von Dennis Klatt in (siehe Klatt, 1980) beschrieben, verwendet entweder einen parallelen Formantsynthesizer oder einen Synthesizer aus parallelem und kaskadiertem Zweig.

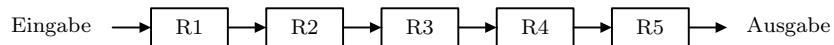


Abbildung 3: Kaskadenzweig

Sonorant	F1	F2	F3	B1	B2	B3
/w/	290	610	2150	50	80	60
/j/	260	2070	3020	40	250	500
/r/	310	1060	1380	70	100	120
/l/	310	1050	2880	50	100	280

Tabelle 2: Wertetabelle für Sonoranten des Englischen aus Klatt (1980)

2.4.3 Frikative / Affrikate

Bei stimmhaften Frikativen muss AF⁵ auf 50dB, AV⁶ auf 47dB und AVP auf 47dB gesetzt werden. Bei stimmlosen Frikativen AF auf 60dB, AV und AB auf 0dB. Die Werte in der Tabelle sind nur für CV-Lautfolgen aufgetragen, wobei der Vokal ein vorderer Vokal ist.

Um ein /h/ zu synthetisieren sollten die Formant- und Bandbreitenwerte des folgenden Vokals genommen werden. Die Frequenz des ersten Formanten sollte nun um ca. 300 Hz erhöht werden. Zusätzlich sollte AV⁷ und AVP⁸ auf 0dB gesetzt werden. Anstelle dessen tritt die Friktion; AF⁹, ASP¹⁰, AB¹¹ sollten auf ca. 50dB gesetzt werden.

Frikative	F1	F2	F3	B1	B2	B3	A2	A3	A4	A5	A6	AB
/f/	340	1100	2080	200	120	150	0	0	0	0	0	57
/v/	220	1100	2080	60	90	120	0	0	0	0	0	57
/θ/	320	1290	2540	200	90	200	0	0	0	0	28	48
/ð/	270	1290	2540	60	80	170	0	0	0	0	28	48
/s/	320	1390	2530	200	80	200	0	0	0	0	52	0
/z/	240	1390	2530	70	60	180	0	0	0	0	52	0
/ʃ/	300	1840	2750	200	100	300	57	48	48	46	0	0

Tabelle 3: Wertetabelle für Frikative des Englischen aus Klatt (1980)

Affrikate	F1	F2	F3	B1	B2	B3	A2	A3	A4	A5	A6	AB
/ç/	350	1800	2820	200	90	300	0	44	60	53	53	0
/ʝ/	260	1800	2820	60	80	270	0	44	60	53	53	0

Tabelle 4: Wertetabelle für Affrikate des Englischen aus Klatt (1980)

2.4.4 Nasale

Die Energie bei Nasalen ist eher im unteren Bereich des Spektrums lokalisiert. Sie haben eine zwar erkennbare aber wenig ausgeprägte Formantstruktur

⁵Amplitude of frication

⁶Amplitude of voicing

⁷Amplitude of voicing

⁸Amplitude of sinusoidal voicing

⁹Amplitude of frication

¹⁰Amplitude of aspiration

¹¹Amplitude of bypass path

(Pompino-Marschall, 2003). Um Nasale zu erzeugen muss zudem mit FNP¹² und FNZ¹³ gearbeitet werden.

Nasal	FNP	FNZ	F1	F2	F3	B1	B2	B3
/m/	270	450	480	1270	2130	40	200	200
/n/	270	450	480	1340	2470	40	300	300

Tabelle 5: Wertetabelle für Nasale des Englischen aus Klatt (1980)

2.4.5 Plosive

Die stimmhaften und stimmlosen Plosivpaare unterscheiden sich in F1, B1 und ein wenig Friktion am Ende des Plosivs.

Bei der Erstellung der unterschiedlichen Plosive spielt die Lokustheorie eine wesentliche Rolle. Die Lokustheorie (Delattre et al., 1955) beschreibt die Transitionen der ersten zwei Formanten abhängig vom Plosiv und dem darauffolgenden Vokal.

Plosive	F1	F2	F3	B1	B2	B3	A2	A3	A4	A5	A6	AB
/p/	400	1100	2150	300	150	220	0	0	0	0	0	63
/b/	200	1100	2150	60	110	130	0	0	0	0	0	63
/t/	400	1600	2600	300	120	250	0	30	45	57	63	0
/d/	200	1600	2600	60	100	170	0	47	60	62	60	0
/k/	300	1990	2850	250	160	330	0	53	43	45	45	0
/g/	200	1990	2850	60	150	280	0	53	43	45	45	0

Tabelle 6: Wertetabelle für Plosive des Englischen aus Klatt (1980)

¹²Nasal pole

¹³Zero frequency

3 Überblick über imskpe

3.1 Allgemeines

Die Oberfläche von `imskpe` versucht soweit es möglich ist den Vorschlägen der GNOME Human Interface Guidelines¹⁴ zu folgen, dabei aber nicht komplett die Ideen von `kpe80` (Simpson et al., 1995) zu übersehen.

Alle Screenshots in dieser Dokumentation sind bewusst mit der englischen Version von `imskpe` gemacht worden, um eine spätere Verwendung auf der Webseite¹⁵ als Bestandteil einer (auch) englischsprachigen Dokumentation zu ermöglichen. Zum jetzigen Zeitpunkt ist `imskpe` ins Deutsche und ins Französische übersetzt worden. Eine Übersetzung in weitere Sprachen ist durch den *gettext*-Standard, der auch von KDE, Gnome und vielen anderen verwendet wird, leicht möglich.

Diese Ausarbeitung bezieht sich auf Version 1.0 von `imskpe`.

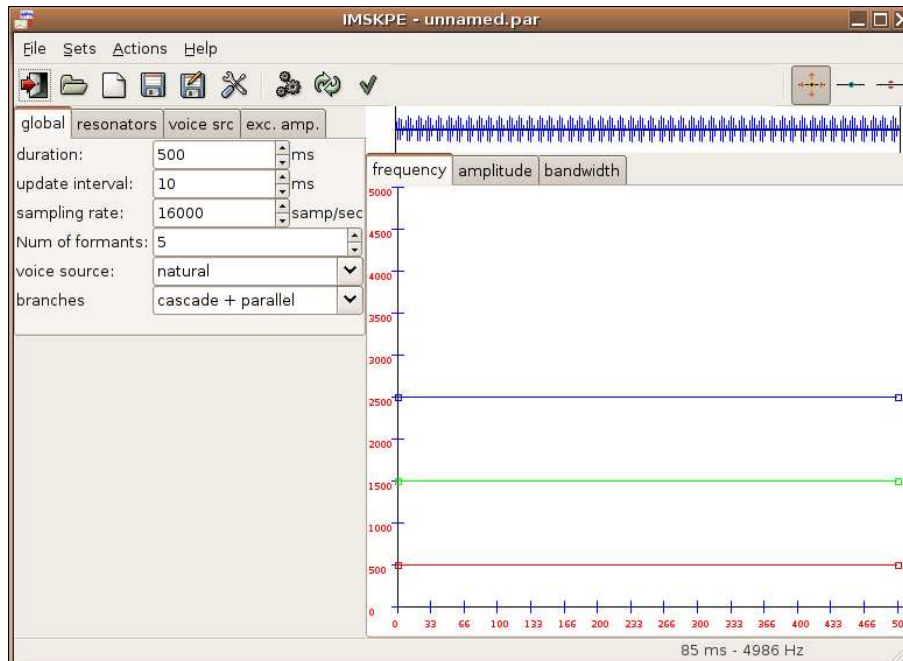


Abbildung 6: `imskpe` – kompletter Screenshot

¹⁴<http://developer.gnome.org/projects/gup/hig/>

¹⁵<http://imskpe.sf.net/>

3.2 Oberfläche

3.2.1 Menü

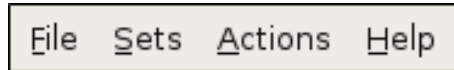


Abbildung 7: Menü

Unter *File* befinden sich die üblichen Operationen (*New*, *Open*, *Save*, *Save As*, *Quit*). *Save As* ermöglicht auch ein Speichern als WAV¹⁶. Das Standard-Dateiformat ist PAR. Der genaue Aufbau von PAR ist im Anhang A.1 beschrieben.

Bei *Sets* handelt es sich um die Möglichkeit Kurvensichtbarkeitszustände abzuspeichern. Das Default-Set bewirkt die Ausblendung aller Kurven. Wichtig hierbei ist, dass der Synthesealgorithmus alle Kurven verarbeitet, auch wenn sie nicht sichtbar sind!

Unter *Actions* kann man mit *Execute* für die aktuelle Kurvenkonstellation eine WAV-Datei erzeugen, diese wird sofort mit dem in den *Preferences* angegebenen Abspielkommando ausgegeben. *Interpolate* bietet die Möglichkeit unnötige Punkte aus den Kurven zu entfernen. Dies ist aber immer mit Vorsicht zu verwenden. Die Interpolation hat Toleranzeinstellungen, die ein Löschen wichtiger Punkte verhindern soll. Dies kann aber nicht immer garantiert werden. Es ist sinnvoll vorher immer zu speichern.

3.2.2 Toolbar



Abbildung 8: Toolbar (links)

Der erste Teil der Toolbar besteht aus folgenden Operationen: *Quit*, *Open*, *New*, *Save*, *Save As*, *Preferences*, *Execute*, *Refresh* der WAV-Anzeige und *Interpolate*.

Die meisten dieser Funktionen entsprechen den Funktionen im Menü (siehe 3.2.1). Ausschließlich in der Toolbar befindet sich die Schaltfläche für die Einstellungen (genaueres unter 3.2.4) und die Aktualisierung der WAV-Anzeige.



Abbildung 9: Toolbar – Modus (rechts)

¹⁶16bit, mono, PCM, little-endian

Im rechten Teil der Toolbar kann der Modus gewählt werden. Bewegen, Einfügen und Löschen von Punkten auf den Kurven. Auf dem Anzeigebereich kann mittels eines Popupmenüs, welches mit der rechten Maustaste aktiviert wird, der Modus gewechselt werden. Falls eine Maus mit Scrollrad vorhanden ist, kann diese über dem Anzeigebereich auch zum Moduswechsel genutzt werden.

3.2.3 PAR-Kontrollbereich

Dateieigenschaften

global	resonators	voice src	exc. amp.
duration:	500		ms
update interval:	10		ms
sampling rate:	16000		samp/sec
Num of formants:	5		
voice source:	natural		
branches	cascade + parallel		

Abbildung 10: Tabs – globale Dateieigenschaften

Duration bestimmt die Dauer des zu erzeugenden Signals in Millisekunden.

Das Update-Intervall ist der Abstand zwischen den Parameterwerten. Hier sind Werte zwischen 2ms und 20ms möglich. Sinnvoll sind hier 10ms. Ein Wert von 5ms kann das Ergebnis leicht verbessern, ein niedrigerer Wert ist nicht sinnvoll. Beim Speichern der PAR-Datei werden die Kurven mit genau dem Punktabstand zerteilt, der durch das Update-Intervall festgelegt wird. Dies kann bei erneutem Laden dieser Datei dazu führen, dass mehr Punkte verwendet werden, um die Kurve zu beschreiben. Mit Hilfe der Interpolationsfunktion können einige der unnötigen Punkte entfernt werden.

Die Samplingrate ist im Default auf 16kHz gesetzt. Dennis Klatt verwendet in Klatt (1980) 10kHz. Bei 10kHz klingt das synthetisierte Signal etwas heller und meist besser. Nur ist 10kHz ein sehr ungewöhnliches Format für den WAV-Export und wird möglicherweise nicht von allen Abspielwerkzeugen unterstützt.

Die Anzahl der Formanten darf zwischen drei und sechs liegen. Der Standard liegt bei fünf Formanten.

Die Stimmquelle (*voice source*) kann entweder naturgetreu (*natural*), *impulsed* oder digitalisiert (*sampled*) sein, wobei die letztere Methode (*sampled*) sehr unzuverlässig funktioniert.

Wie in Kapitel 2.2 bereits ausgeführt, kann der `klatt80`-Algorithmus entweder nur einen parallel Zweig (Branch) oder einen kaskadierten und parallelen Zweig umfassen.

Formantwerte

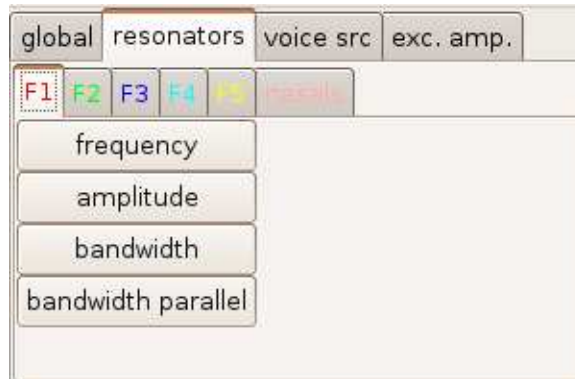


Abbildung 11: Tabs – Formantwerte

Für jeden Formanten können die Kurven zur Frequenz, der Amplitude, der Bandbreite und der Bandbreite im parallelen Zweig angezeigt werden. Hierbei gilt, auch wenn die Kurve nicht sichtbar ist, wird sie im Algorithmus verwendet.

Nasalparameter

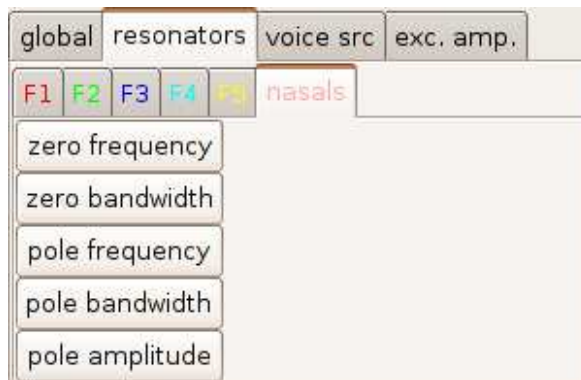


Abbildung 12: Tabs – Nasalparameter

Für die Synthetisierung von Nasalen kann die Frequenz der Auslöschung (*zero frequency*) und deren Bandbreite in Form einer Kurve eingestellt werden. Auch die Kurven für die Frequenz, Bandbreite und Amplitude des Maximums (*pole frequency, bandwidth, amplitude*) werden in diesem Tab aktiviert.

Voice Source Parameter



Abbildung 13: Tabs – Voice Source Parameter

Die *fundamental frequency* entspricht der Grundfrequenz (F_0). *Glottal open quotient* (Parameter: *kopen*) entspricht der Länge oder Dauer der Öffnung der Glottis in Samples. Ein Wertebereich von 10 bis 65 ist möglich. Mit *breathiness* kann die Behauchtheit der Stimmquelle angepasst werden. *Spectral tilt* und *skewness* geben die Neigung der Stimmquelle im Spektrum an.

Erweiterte Amplitudenparameter

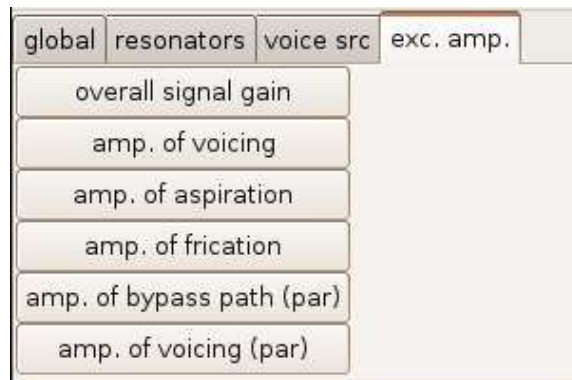


Abbildung 14: Tabs – erweiterte Amplitudenparameter

Overall signal gain ist die Lautstärke des Signals. Mit *amp. of voicing* (AV) kann die Kurve für die Amplitude der Phonation sichtbar gemacht werden. *Amp. of voicing (par)* (AVP) ist das entsprechende im parallelen Zweig. Mit *amp. of frication* (AF) und *amp. of aspiration* (ASP) werden die entsprechenden Amplitudenkurven sichtbar. AB (*amp. of bypass path(par)*) ist im Schaubild in Kapitel 2.3 der Pfad, der unten den parallelen Zweig umgeht.

3.2.4 Preferences

Die Einstellungen werden beim Beenden von `imskpe` automatisch gespeichert. Beim ersten Start wird ein default angenommen, der mit dem `Default`-Button jederzeit wieder herstellbar ist. Die Einstellungen werden in der Datei `.imskpe` im HOME-Verzeichnis des Benutzers gespeichert.

Farben der Kurven

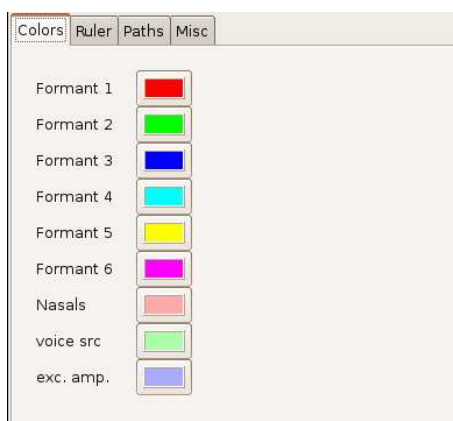


Abbildung 15: Preferences – Farben der Kurven

Dieser Dialog ermöglicht es, jedem Formanten eine Farbe zuzuweisen, wobei Frequenz, Amplitude und Bandbreite des jeweiligen Formanten die gleiche Farbe haben.

Lineal

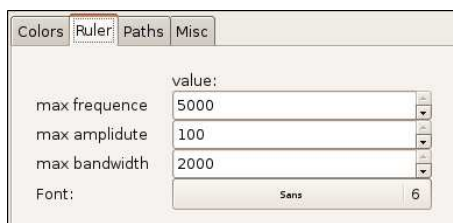


Abbildung 16: Preferences – Linealeinstellungen

Die Lineale der drei Diagramme sind standardmässig auf ihre maximalen Werte eingestellt. Diese sind in vielen Fällen zu groß und können hier geringer eingestellt werden. Zudem kann die Schriftart jedes Lineals verändert werden. Die voreingestellte Schriftart sollte auf allen Plattformen zur Verfügung stehen.

Pfade

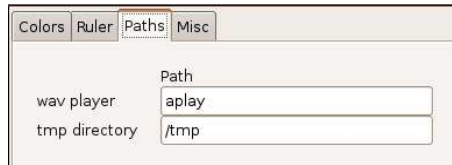


Abbildung 17: Preferences – Pfade

Im Dialog *Paths* kann der Pfad für das Ziel der temporären Dateien angegeben werden. Bei jedem Aufruf von *Execute* wird eine temporäre Datei angelegt. Bei einem erneuten Aufruf von *Execute* innerhalb derselben Session, also ohne Neustart von *imskpe*, wird die Datei überschrieben. Um die temporären Dateien voneinander zu unterscheiden und es mehreren Benutzern zu ermöglichen auf demselben System mit *imskpe* zu arbeiten, steht im Dateinamen der temporären Datei die Prozessnummer der *imskpe*-Instanz. Die erzeugten temporären Dateien werden nicht gelöscht. Auch der Befehl, der zum Abspielen der resultierenden WAV-Datei beim Ausführen verwendet wird, kann hier festgelegt werden. Unter Linux/Unix ist dies meist `play` oder `aplay`. Unter Windows ist der mitgelieferte Soundrecorder die Voreinstellung (`sndrec32 /play /close`).

Verschiedenes

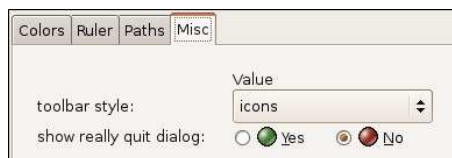


Abbildung 18: Preferences – Verschiedenes

Verschiedenes beinhaltet die Art der Toolbar (nur Icons, nur Text oder beides) und ob ein „Wirklich beenden“-Dialog angezeigt werden soll oder nicht.

3.2.5 Diagrammfläche

Um eine Kurve zu editieren, muss sich der Mauscursor über der Kurve befinden. Zum Editieren gibt es drei verschiedene Modi: Bewegen, Einfügen und Löschen eines Punktes. Der Modus kann durch Klicken der Icons in der Toolbar geändert werden, durch das mit der rechten Maustaste erreichbare Kontextmenü, und als dritte Möglichkeit – falls vorhanden – mit dem Scrollrad der Maus. Im Kontextmenü gibt es zudem die Möglichkeit, den Wert des Punktes exakt einzugeben.

Punkte haben einen Mindestabstand zueinander, der durch das Updateintervall festgelegt wird.

3.2.6 Statuszeile

In der Statuszeile wird, sofern der Cursor sich über der Diagrammfläche befindet, der aktuelle Wert der beiden Koordinaten angezeigt. Wenn der Mauscursor sich über einer Kurve oder einem Punkt befindet, werden der Name der Kurve und die Koordinaten des Punktes angegeben.

3.3 Beispielsitzungen

Die hier beschriebenen Beispiele sind nicht perfekt! An einer Formantsynthese-parameterdatei kann man Stunden verbringen und an dem Resultat Feinarbeit leisten. Meist ist es dann immer noch nicht perfekt. Die Beispiele hier sollen nur aufzeigen, wie mit Hilfe von `imskpe` ein Laut oder eine Lautfolge erstellt werden kann.

3.3.1 Erstellen des Lautes /a/

Nach dem Starten von `imskpe` ist die Anzeige der Frequenzen der ersten 3 Formanten aktivieren. Der Standard ist das Schwa. Bei einem Klick auf die *Execute*-Schaltfläche sollte ein Schwa zu hören sein.

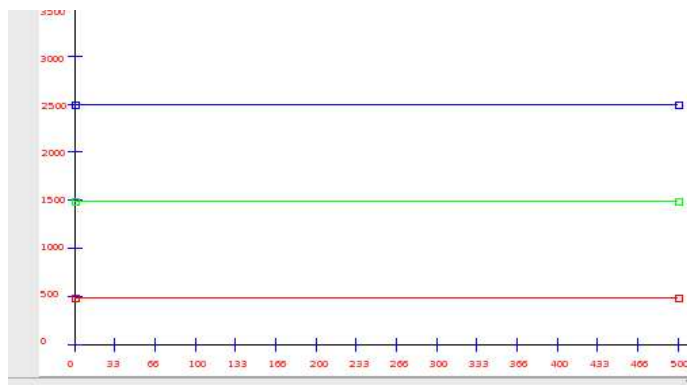


Abbildung 19: Frequenz-Diagrammfläche mit aktivierten F1–F3 für ein Schwa

Um aus einem Schwa ein /a/ zu machen, müssen die Frequenz von F1 bis F3 geändert werden. Um diese herauszufinden, wurden mit Hilfe von `Wavesurfer`¹⁷ die Formantwerte der ersten drei Formanten eines /a/ abgelesen [750Hz, 1400Hz, 3000Hz]. Nun sind die Frequenz von F1 bis F3 in `imskpe` auf diese Werte anzupassen.

¹⁷<http://www.speech.kth.se/wavesurfer/>

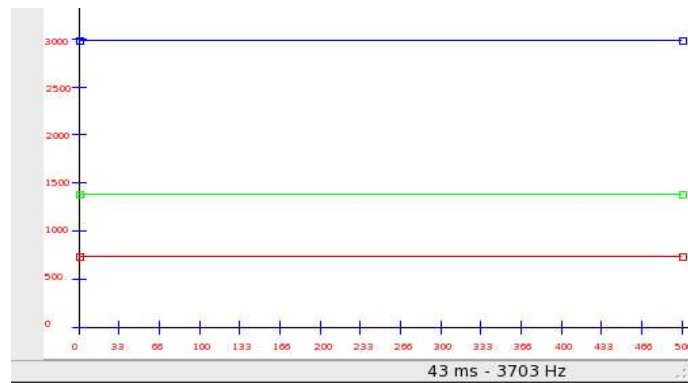


Abbildung 20: Frequenz-Diagrammfläche mit aktivierten F1-F3 für ein /a/

Nach dem Aufruf von *Execute* sollte es nach einem /a/ klingen.

Nun sind noch einige Optimierungen möglich: Ein genaues Einstellen der Bandbreite für die ersten drei Formanten kann das Ergebnis verbessern.

3.3.2 Erstellen der Lautfolge /ha/

Als Erstes das /a/ aus 3.3.1 laden oder ein /a/ erstellen.

Zu Beginn an den Anfang der Lautfolge ein /h/ bauen, indem AV¹⁸ und AVP¹⁹ von Millisekunde 0ms bis 60ms auf 0dB gesetzt werden und von 60ms bis 100ms auf 60dB ansteigen, was dem schon vorhandenen Wert des /a/ entspricht. Genau entgegengesetzt nun die Werte für AF²⁰, ASP²¹, AB²² von 0ms bis 60ms auf 50dB setzen und von 60ms bis 100ms auf 0dB abfallen lassen.

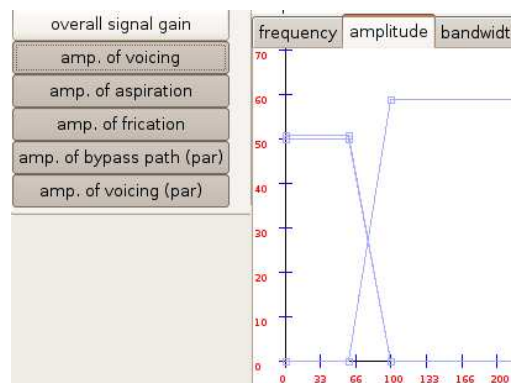


Abbildung 21: AV, AF, ASP, AB, AVP für ein /ha/

¹⁸Amplitude of voicing

¹⁹Amplitude of sinusoidal voicing

²⁰Amplitude of frication

²¹Amplitude of aspiration

²²Amplitude of bypass path

Zusätzlich noch für den Bereich von 0ms bis 60ms die Frequenz von F1 um ca. 300Hz höher setzen als die Frequenz des ersten Formanten des /a/. Von 60ms bis 100ms dann wieder auf die Frequenz von F1 des /a/ abfallen lassen.

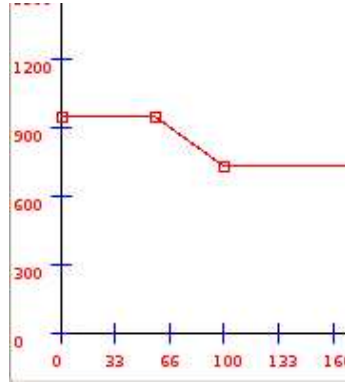


Abbildung 22: Frequenzunterschied von F1 zwischen /h/ und /a/ in /ha/

Für /l/ und /o/ müssen nun noch die Frequenzen und die Bandbreiten der ersten drei Formanten angepasst werden. Da das /a/ und das /o/ ein wenig länger sein sollen als das /l/, habe ich folgende Positionen für die Transitionen gewählt: Das Ende des /a/ bei 250ms; die Transition zum /l/ bis 300ms; die Länge des /l/ 30ms; die Transition zum /o/ bis 330ms. An genau diesen Punkten werden nun Punkte auf den drei Formantkurven (F1–F3) im Frequenzdiagramm eingefügt. Diese Punkte werden nun auf die Werte aus 2.4.2, bzw. für das /o/ aus 2.4.1 angepasst.

Zur besseren Übersicht die Werte in einer Tabelle:

Zeitindex	F1	F2	F3
300	310	1050	2880
330	310	1050	2880
360	450	900	2300
490	450	900	2300

Tabelle 7: Frequenzwerte für /lo/

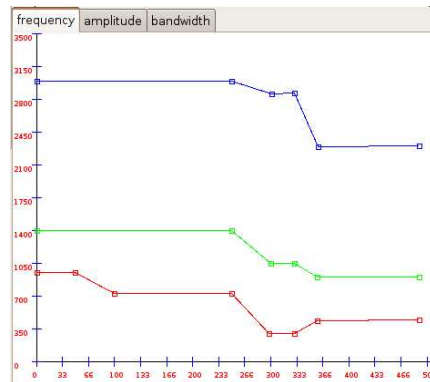


Abbildung 23: Frequenzen von F1–F3 für /haɪo/

An den selben Positionen auf der Zeitachse werden auch auf den Bandbreiten der ersten drei Formanten Punkte eingefügt. Auch hier sind die Werte für /ɪo/ der Tabellen in 2.4 entnommen.

Zeitindex	B1	B2	B3
300	50	100	280
330	50	100	280
360	80	70	70
490	80	70	70

Tabelle 8: Bandbreitenwerte für /ɪo/

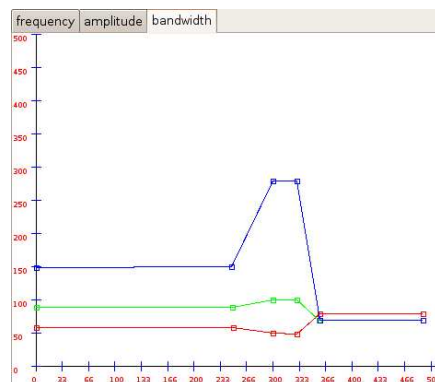


Abbildung 24: Bandbreite von F1–F3 für /haɪo/

3.3.3 Erstellen der Lautfolge /jεs/

Zunächst das ganze Signal zu einem /ε/ einstellen. Dies geschieht durch das Setzen der Frequenzen der Formanten 1–3 auf ca. [530Hz, 2000Hz, 2500Hz].

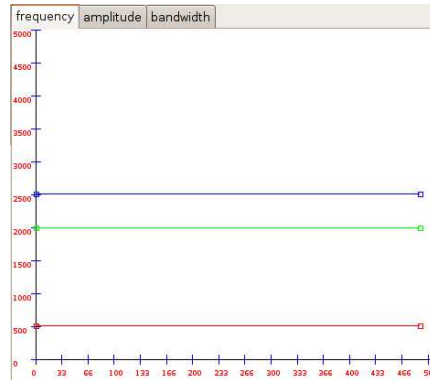


Abbildung 25: Frequenzen von F1–F3 für das /ε/ in /jεs/

Anschließend die Bandbreiten für die ersten drei Formanten auf [60Hz, 90Hz, 200Hz] setzen.

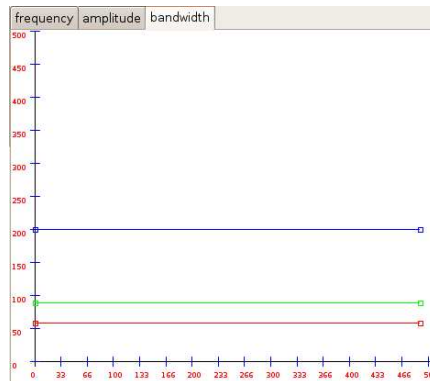


Abbildung 26: Bandbreiten von F1–F3 für das /ε/ in /jεs/

Das /s/ sollte ca. bei 400ms beginnen. Davor ca. 50ms für den Übergang vom /ε/ zum /s/. Wie in der Tabelle in Kapitel 2.4.3 aufgeführt sind die Formanten für /s/ $F1-F3 = [320\text{Hz}, 1390\text{Hz}, 2530\text{Hz}]$ und $B1-B3 = [200\text{Hz}, 80\text{Hz}, 200\text{Hz}]$. Zusätzlich sind A6 auf 52dB, AF auf 60dB, AV und AVS auf 47dB zu setzen. Um A6 setzen zu können muss die Formantanzahl auf 6 gesetzt werden.

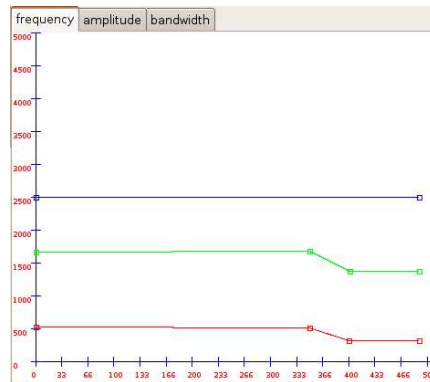


Abbildung 27: Frequenzen von F1–F3 für /εs/ in /jεs/

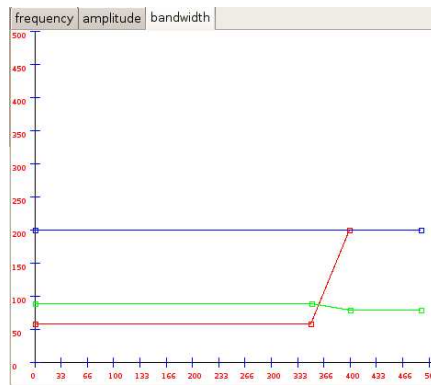


Abbildung 28: Bandbreiten von F1–F3 für /εs/ in /jεs/

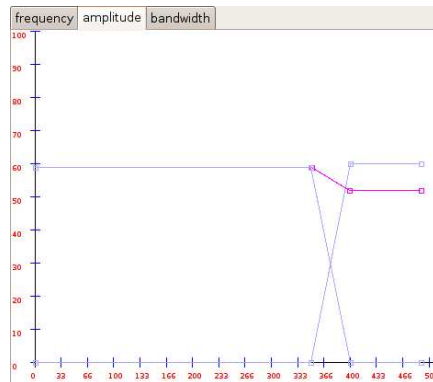


Abbildung 29: Amplitudenwerte (A6, AF, AV, AB) für /εs/ in /jεs/

Zu Abschluss das /j/ am Beginn der Lautfolge. Die Länge des /j/ wird auf 100ms und die Dauer der Transition auf 50ms gesetzt. Die Werte sind in Kapitel 2.4.2 nachzusehen. Für $F1-F3 = [260\text{Hz}, 2070\text{Hz}, 3020\text{Hz}]$ und für $B1-B3 = [40\text{Hz}, 250\text{Hz}, 500\text{Hz}]$.

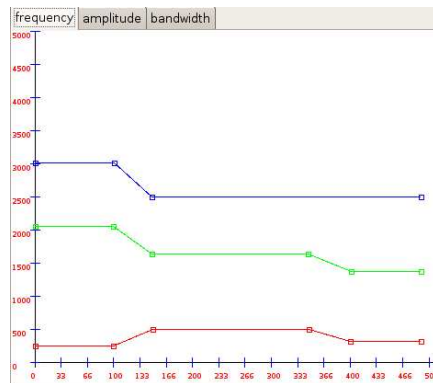
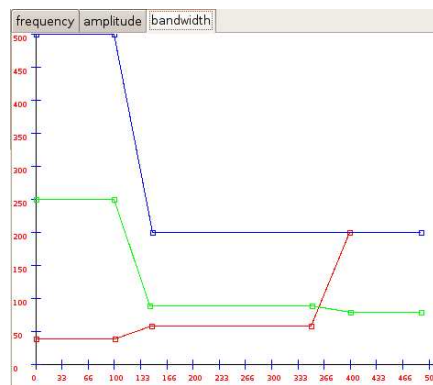


Abbildung 30: Frequenzen von F1-F3 für /jεs/

Abbildung 31: Bandbreiten von F1–F3 für $/j\epsilon s/$

4 Die Zukunft von imskpe

4.1 Verbesserungsmöglichkeiten

- Die Einblendung eines Spektrogramms einer anderen zugeladenen WAV-Datei, um die Formantwerte an dieser zu orientieren, ermöglichen.
- Ein besseres Fileformat würde die interne Repräsentation besser abbilden.
Das Format hat sich seit `klatt80` nicht wesentlich verändert. Das Abspeichern aller Werte zu jedem Zeitindex entspricht nicht der internen Repräsentation von `imskpe`. Intern wird alles in Listen gespeichert, die nur die nötigen Punkte beinhalten. Ein Format, das ebenfalls nur die notwendigen Punkte speichert wäre sinnvoller.
- Die Möglichkeit veränderte Formantsynthesealgorithmen einzusetzen fehlt. Eine Ausgliederung des `klatt80`-Teils und eine API zur Anbindung anderer Formantsynthesysteme mit möglicherweise anderen Parametern wäre anzustreben.
- Ein Installer für die Windowsversion würde den Windowsnutzern die Benutzung erleichtern.
- Shortcuts fehlen bisher völlig. Vorschläge werden hier gerne angenommen!²³
- Weitere Übersetzungen der Sprachdatei sind, wie in Kapitel 3.1 erwähnt, einfach zu erstellen.
Bisher wurde `imskpe` neben Englisch nur in Deutsch und Französisch übersetzt.
- Ein schöneres Icon würde `imskpe` nicht schaden.

4.2 Weiterentwicklung

Die weitere Entwicklung von `imskpe` findet auf den Sourceforge-Projektseiten unter <http://sf.net/projects/imskpe> statt. Hier ist auch der beste Ort um Bugreports oder Featurerequests einzureichen. Auch der Sourcecode liegt hier in einem offenen Versionskontrollsystem. Wichtig hierbei ist allerdings, dass es sich bei `imskpe` immer um ein Werkzeug handeln soll, mit dem einzelne Laute oder Lautfolgen erstellt werden können. Es ist und will kein komplettes Text-To-Speech-System sein.

²³Meine E-Mailadresse: bolsog@users.sf.net

A Anhang

A.1 Aufbau des PAR-Formantes

Das PAR-Format ist ein Textformat.

Der Header

```
/* Produced by imskpe */
/* DU : 500 */
/* UI : 10 */
/* SR : 16000 */
/* NF : 5 */
/* SS : 2 */
/* CP : 1 */
```

Symbol	Name	Min.	Max.	Default
DU	Duration		5000	500
UI	Update interval	2	20	10
SR	Sampling rate	11025	22050	16000
NF	Number of formants	4	6	5
SS	VoiceSource: impulse [1], natural [2], sampled [3]	1	3	2
CP	Cascade & parallel [1] OR only parallel [2]	1	2	1

Tabelle 9: Headerwerte des PAR-Formates aus Simpson et al. (1995)

A.2 Die Parameterwerte

Jede Zeile besteht aus einem Zeitindex. Die erste Zeile der Daten beginnt beim Zeitindex 0 und die zweite beim Zeitindex $0+UI^{24}$ usw. Nach dem Zeitindex folgt ein Doppelpunkt und 40 Werte in der Reihenfolge der folgenden Tabelle:

N	Symbol	Name	Min.	Max.	Default	Diagramm
1	f0	Fundamental freq. of voicing (Hz)	0	500	0	freq
2	av	Amplitude of voicing (dB)	0	80	0	ampl
3	f1	First formant frequency (Hz)	150	900	500	freq
4	b1	First formant bandwidth (Hz)	40	500	50	band
5	f2	Second formant frequency (Hz)	500	2500	1500	freq
6	b2	Second formant bandwidth (Hz)	40	500	70	band
7	f3	Third formant frequency (Hz)	1300	3500	2500	freq
8	b3	Third formant bandwidth (Hz)	40	500	110	band
9	f4	Fourth formant frequency (Hz)	2500	4500	3300	freq
10	b4	Fourth formant bandwidth (Hz)	100	500	250	band
11	f5	Fifth formant frequency (Hz)	3500	4900	3850	freq
12	b5	Fifth formant bandwidth (Hz)	150	700	200	band
13	f6	Sixth formant frequency (Hz)	4000	4999	4900	freq
14	b6	Sixth formant bandwidth (Hz)	200	2000	1000	band
15	fnz	Nasal zero frequency (Hz)	200	700	250	freq
16	bnz	Nasal zero bandwidth (Hz)	50	500	100	band
17	fnp	Nasal pole freq (Hz)	248	528	-	freq
18	bnp	Nasal pole bandwidth (Hz)	50	500	100	band
19	asp	Amplitude of aspiration (dB)	0	80	0	ampl
20	kopen	Number of samples in open period	10	65	-	ampl
21	aturb	Breathiness in voicing	0	80	-	ampl
22	tilt	Voicing spectral tilt (dB)	0	24	-	ampl
23	af	Amplitude of frication (dB)	0	80	0	ampl
24	skew	Skewness of alternate periods	0	40	-	ampl
25	a1	First formant amplitude (dB)	0	80	0	ampl
26	b1p	Par. 1st formant bw (Hz)	40	1000	-	band
27	a2	Second formant amplitude (dB)	0	0	0	ampl
28	b2p	Par. 2nd formant bw (Hz)	40	1000	-	band
29	a3	Third formant amplitude (dB)	0	80	0	ampl
30	b3p	Par. 3rd formant bw (Hz)	40	1000	-	band
31	a4	Fourth formant amplitude (dB)	0	80	0	ampl
32	b4p	Par. 4th formant bw (Hz)	40	1000	-	band
33	a5	Fifth formant amplitude (dB)	0	80	0	ampl
34	b5p	Par. 5th formant bw (Hz)	40	1000	-	band
35	a6	Sixth formant amplitude (dB)	0	80	0	ampl
36	b6p	Par. 6th formant bw (Hz)	40	2000	-	band
37	anp	Amp of par nasal pole	0	80	-	ampl
38	ab	Bypass path amplitude (dB)	0	80	0	ampl
39	avp	Amplitude of sinusoidal voicing (dB)	0	80	0	ampl
40	gain	Overall gain control (dB)	0	80	48	ampl

Tabelle 10: Parameterwerte des `klatt80`-Algorithmus aus dem Code von Simpson et al. (1995). Analog auch in Klatt (1980) und Allen et al. (1987).

Bei manchen Parameterwerten wären höhere Werte wünschenswert. Der Algorithmus ist aber nur für Werte in diesen Bereichen definiert. Zum Beispiel liegt die Frequenz des sechsten Formanten in einem natürlichen Spektrum über 5000 Hz.

²⁴Update interval

Literatur

- Allen, J., Hunnicutt, M. S., Klatt, D., 1987. *From text to speech: The MITalk system*. Cambridge University Press.
- Delattre, P. C., Liberman, A. M., Cooper, F. S., 1955. Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America* **27** (4), 769–773.
URL <http://link.aip.org/link/?JAS/27/769/1>
- Fant, G., 1960. *Acoustic theory of speech production*. Mouton, The Hague.
- Klatt, D. H., 1980. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America* **67** (3), 971–995.
URL <http://link.aip.org/link/?JAS/67/971/1>
- Klatt, D. H., 1987. Review of text-to-speech conversion for english. *The Journal of the Acoustical Society of America* **82** (3), 737–793.
URL <http://link.aip.org/link/?JAS/82/737/1>
- Lemmetty, S., 1999. *Review of speech synthesis technology*. Master's thesis, Helsinki University of Technology.
URL <http://www.acoustics.hut.fi/~slemmet/dippa/contents.html>
- Pompino-Marschall, B., 2003. *Einführung in die Phonetik*, 2nd Edition. de Gruyter, Berlin, 1st edition 1995.
- Simpson, A., Iles, J., Ing-Simmons, N., 1995. *KPE80 – A Klatt synthesiser and parameter editor*.
URL <http://www.speech.cs.cmu.edu/comp.speech/Section5/Synth/klatt.kpe80.html>

Abbildungsverzeichnis

1	Das Quelle-Filter-Modell	3
2	Paralleler Zweig	4
3	Kaskadenzweig	4
4	Blockdiagramm nach Klatt (1980)	5
5	IPA – Vokalviereck	6
6	imske – kompletter Screenshot	9
7	Menü	10
8	Toolbar (links)	10
9	Toolbar – Modus (rechts)	10
10	Tabs – globale Dateieigenschaften	11
11	Tabs – Formantwerte	12
12	Tabs – Nasalparameter	12
13	Tabs – Voice Source Parameter	13
14	Tabs – erweiterte Amplitudenparameter	13
15	Preferences – Farben der Kurven	14
16	Preferences – Linealeinstellungen	14
17	Preferences – Pfade	15
18	Preferences – Verschiedenes	15
19	Frequenz-Diagrammfläche mit aktivierten F1–F3 für ein Schwa	17
20	Frequenz-Diagrammfläche mit aktivierten F1–F3 für ein /a/	18
21	AV, AF, ASP, AB, AVP für ein /ha/	18
22	Frequenzunterschied von F1 zwischen /h/ und /a/ in /ha/	19
23	Frequenzen von F1–F3 für /halo/	20
24	Bandbreite von F1–F3 für /halo/	20
25	Frequenzen von F1–F3 für das /ε/ in /jes/	21
26	Bandbreiten von F1–F3 für das /ε/ in /jes/	21
27	Frequenzen von F1–F3 für /εs/ in /jes/	22
28	Bandbreiten von F1–F3 für /εs/ in /jes/	22
29	Amplitudenwerte (A6, AF, AV, AB) für /εs/ in /jes/	23
30	Frequenzen von F1–F3 für /jes/	23
31	Bandbreiten von F1–F3 für /jes/	24

Tabellenverzeichnis

1	Wertetabelle für Vokale des Englischen aus Klatt (1980)	6
2	Wertetabelle für Sonoranten des Englischen aus Klatt (1980)	7
3	Wertetabelle für Frikative des Englischen aus Klatt (1980)	7
4	Wertetabelle für Affrikate des Englischen aus Klatt (1980)	7
5	Wertetabelle für Nasale des Englischen aus Klatt (1980)	8
6	Wertetabelle für Plosive des Englischen aus Klatt (1980)	8
7	Frequenzwerte für /lo/	19
8	Bandbreitenwerte für /lo/	20
9	Headerwerte des PAR-Formates aus Simpson et al. (1995)	26
10	Parameterwerte des Klatt-Algorithmus aus Simpson et al. (1995).	27