TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Elektrotechnik und Informationstechnik, Institut für Akustik und Sprachkommunikation, Lehrstuhl Kognitive Systeme

Hauptseminararbeit zum Thema

Synthese von Sprachsignalen

Karl-Ludwig Besser, Zhongjiu Li, Franz-Marcus Schüffny, Peter Steiner

Betreuer:
PD Dr.-Ing. Ulrich Kordon
Dipl.-Ing. Steffen Kürbis

Hochschullehrer:
Jun.-Prof. Dr.-Ing. Peter Birkholz

Verteidigungsdatum:
14.07.2015

Einleitung

Im Hauptseminar Kommunikationssysteme wurde das Thema „Synthese von Sprachsignalen“ bearbeitet. Der Fokus lag auf der Erzeugung von Sprachsignalen mittels Formantsynthese. Diese wurde in einem Computerprogramm durch ein Quelle-Filter-Modell realisiert. Zur Parameterbestimmung war eine vorherige Analyse realer Sprache notwendig. Mit dieser Thematik beschäftigten sich Forscher schon seit mehreren Jahrhunderten.

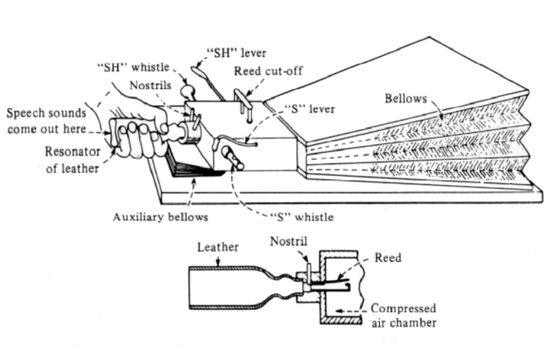

Abbildung 1: Von Kempelen's Sprachmaschine [1, S. 5]

Weitere Meilensteine waren die Entwicklung elektronischer Filter sowie der Einsatz moderner Computertechnik.
Heutzutage kommen Systeme zur Sprachsynthese zum Beispiel in Navigationsgeräten und Smartphones zum Einsatz.

Quelle-Filter-Modell

Das Quelle-Filter-Modell versucht eine Zerlegung von Sprachsignalen in Anregungssignale und Filterstrukturen. Durch geeignete Wahl der Modellparameter soll eine möglichst gute Modellierung des menschlichen Artikulationstrakts erreicht werden.
Die Struktur des Modells, wie es hier verwendet wird, ist in Abbildung 3 gezeigt.

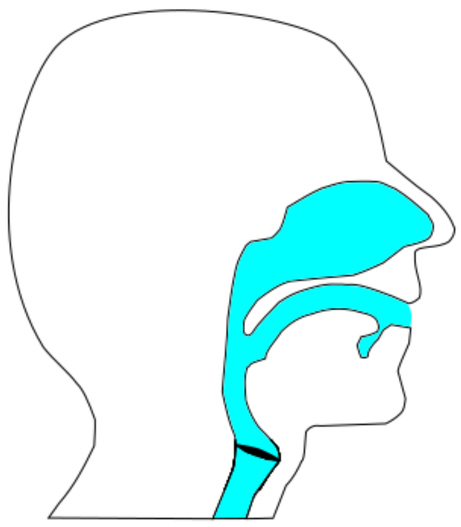

Abbildung 2: Menschlicher Vokaltrakt

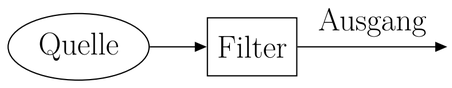

Abbildung 3: Verwendetes Quelle-Filter-Modell

Analyse

Zur Parametrisierung des in Abbildung 3 dargestellten Modells müssen reale Sprachsignale analysiert werden.
Der Schwerpunkt liegt hierbei auf der Bestimmung der Filterparameter. Es werden primär Bandpassfilter 2. Ordnung eingesetzt. Die drei charakteristischen Parameter sind Mittenfrequenz, Bandbreite und Grundverstärkung.

Die Mittenfrequenz wird als Formantfrequenz bezeichnet. Um diese zu bestimmen, wurden verschiedene Analysemethoden verwendet. Von der Software Praat wurde der fertig implementierte Burg-Algorithmus bereitgestellt, selbst nachprogrammiert wurde der Cepstrum-Algorithmus. Das dabei gewonnene geglättete Spektrum ist in Abbildung 4 dargestellt.

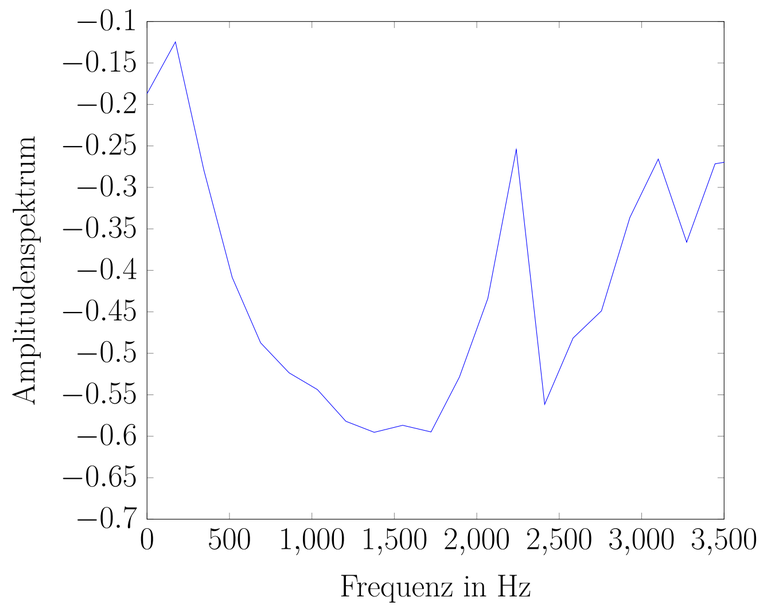

Abbildung 4: Geglättetes Spektrum des Vokals i

Synthese

Zur Synthese von Einzellauten werden verschiedene Quelle-Filter-Modelle verwendet.

Synthetisierbare Einzellaute

Implementiert wurden folgende Laute: Vokale, Diphtonge (/au/, /ei/, /eu/), stimmlose Frikative (/ch/, /f/, /s/, /sch/), stimmhafte Frikative (/w/), stimmhafte Plosive (/b/, /d/, /g/), stimmlose Plosive (/k/, /t/, /p/), Liquide (/l/, /r/) sowie Nasale (/m/, /n/).
Allerdings können stimmhafte Plosive nur in Verbindung mit einem darauffolgenden Vokal synthetisiert werden.

Vokale

Zur Synthese von Vokallauten wird ein periodisches breitbandiges Anregungssignal als Quelle verwendet. Die Signalformung wird mittels Bandpassfilter in Reihenschaltung realisiert.
Das Anregungssignal für die stimmhaften Laute soll den glottalen Luftstrom nachbilden. Dazu wurde eine von Paul Taylor vorgeschlagene Formel benutzt.

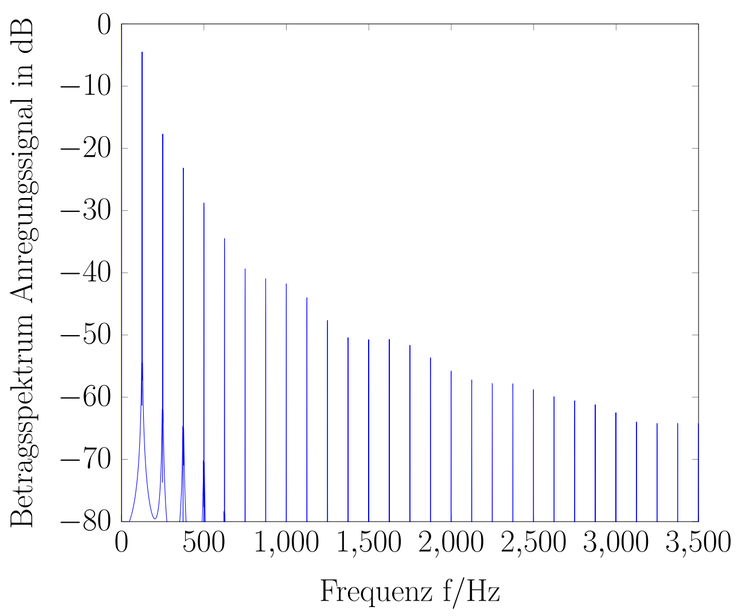

Abbildung 5: Spektrum des Anregungssignals

Beispielhaft ist das Spektrum für den synthetisierten Vokal /i/ in Abbildung 6 dargestellt.

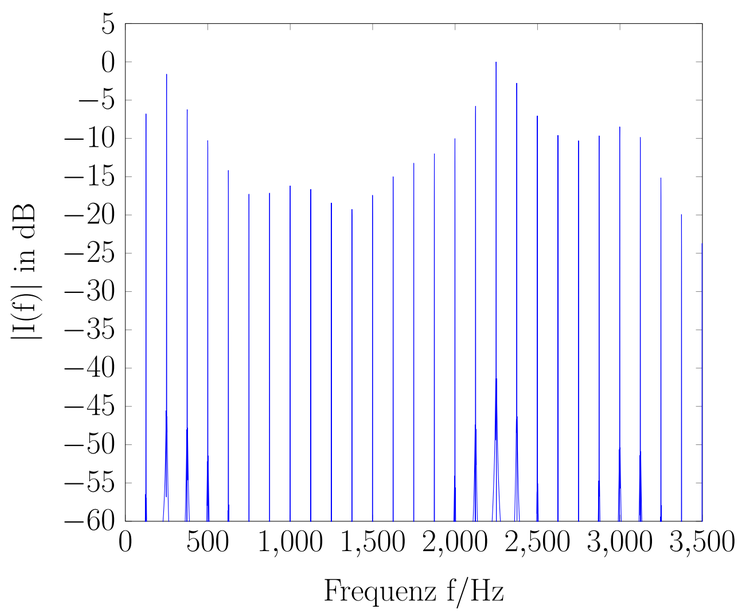

Abbildung 6: Spektrum des synthetisierten Vokals i

Frikative

Zur Synthese von Zischlauten wird ein Rauschgenerator als Quelle verwendet. Die Signalformung wird mittels Bandpassfilter in Parallelschaltung realisiert.

Kombination

Bei der Lautkombination werden einzelne Laute mit einem Von-Hann-Fenster gefenstert und ineinander verschoben. Die Ineinanderverschiebung ist in Abbildung 5 abgebildet.

Zusammenfassung

Im Rahmen des Hauptseminars wurden Einzellaute und Lautkombinationen analysiert und deren Parameter bestimmt. Dies bildete die Grundlage für eine anschließende Synthese von Sprachsignalen.
Der Sprachsynthesizer wurde in GNU Octave implementiert. Der Benutzer hat die Möglichkeit eine Lautkombination zur Synthese einzugeben. Diese wird anschließend als Audiodatei ausgegeben.

Literatur

[1] Sami Lemmetty. „Review of speech synthesis technology“. In: *Helsinki University of Technology* (1999).

[2] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.