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

Einleitung

Im Hauptseminars 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.

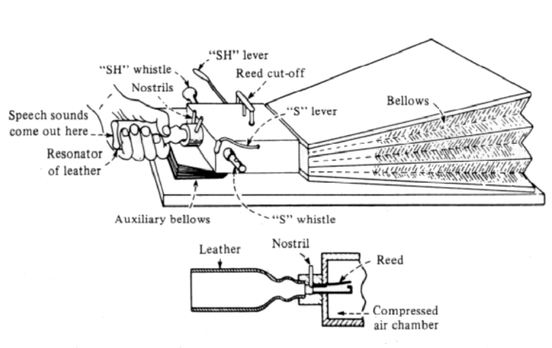

Abbildung 1: Kratzensteins Resonator [1, S. 5]

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 2 gezeigt.

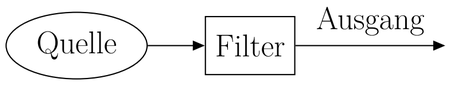

Abbildung 2: Verwendetes Quelle-Filter-Modell

Aenean ac nulla ipsum. Sed nulla dui, consectetur sit amet ultrices eget, semper nec ipsum. Pellentesque lacinia ornare sapien, ac accumsan nulla congue eget. Aliquam gravida nulla id justo egestas accumsan.
Aenean ac nulla ipsum. Sed nulla dui, consectetur sit amet ultrices eget, semper nec ipsum. Pellentesque lacinia ornare sapien, ac accumsan nulla congue eget. Aliquam gravida nulla id justo egestas accumsan.

Analyse

Zur Parametrisierung des in Abbildung 2 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 3 dargestellt.

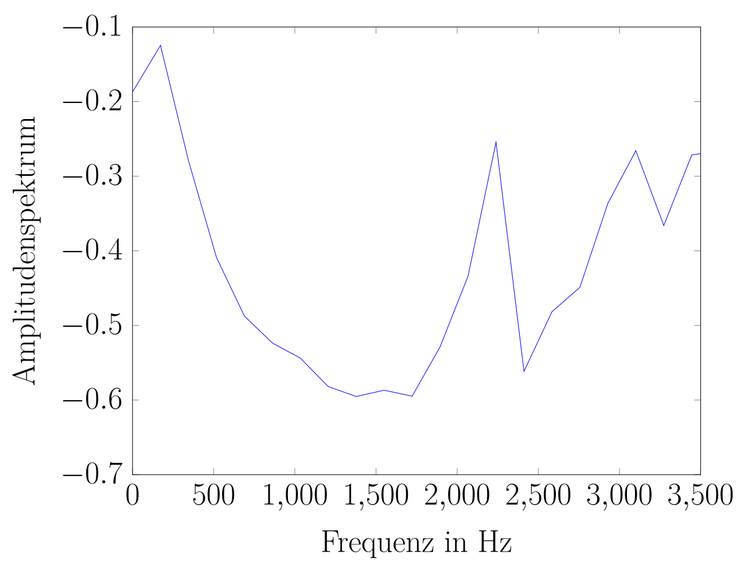

Abbildung 3: Spektrum des Vokals i

Synthese

Zur Synthese von Einzellauten werden verschiedene Quelle-Filter-Modelle verwendet. Nachfolgend sind zwei Beispiele dargestellt. Einmal wird zur Realisierung eine Kaskadenstruktur verwendet, einmal eine Parallelstruktur.

Vokallaute

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 die von Paul Taylor vorgeschlagene Formel benutzt

$$u[n] = \begin{cases} \frac{1}{2}(1 - \cos(\pi n / N_1)) & 0 \le n \le N_1 \\ \cos(\pi (n - N_1)/(2N_2)) & N_1 \le n \le N_2 \\ 0 & \text{sonst} \end{cases} \quad (1)$$

Eine Filterstruktur nach Abbildung 4 wird zur Synthese von Vokallauten genutzt.

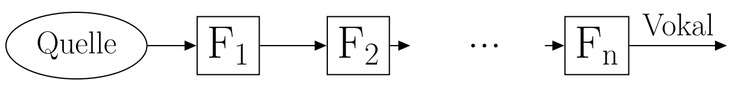

Abbildung 4: Quelle-Filter-Modell für Vokallaute

Beispielhaft ist das Spektrum für den synthetisierten Vokal „i“ in Abbildung 5 dargestellt.

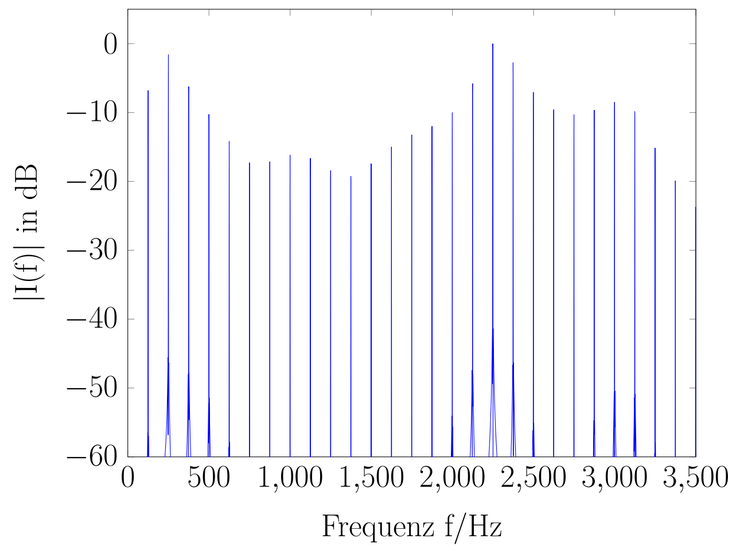

Abbildung 5: Spektrum des synthetisierten Vokals i

Zischlaute

Zur Synthese von Zischlauten wird ein Rauschgenerator als Quelle verwendet. Die Signalformung wird mittels Bandpassfilter in Parallelschaltung realisiert. Eine Filterstruktur nach Abbildung 6 wird zur Synthese von Zischlauten verwendet.

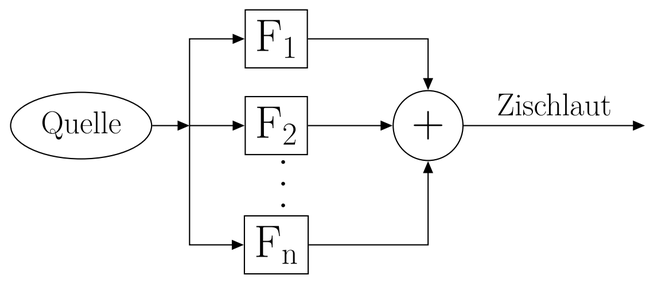

Abbildung 6: Quelle-Filter-Modell für Zischlaute

Kombination

Bei der Lautkombination werden einzelne Laute gefenstert und ineinander verschoben. Als Fensterfunktion wird ein Von-Hann-Fenster verwendet. Die Verschiebung zweier Fenster ist beispielhaft in Abbildung 7 abgebildet.

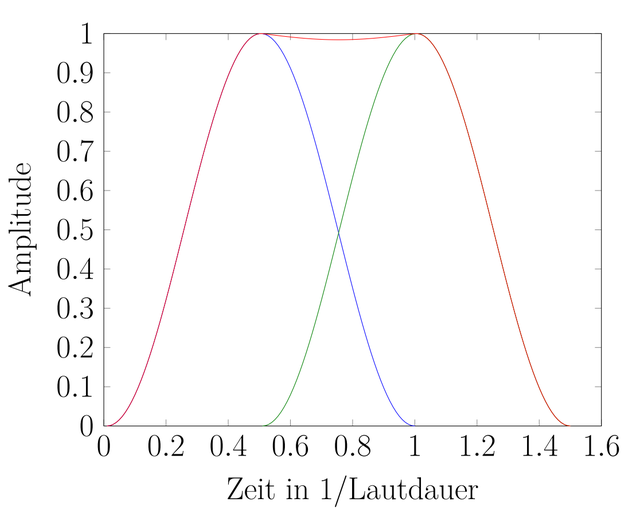

Abbildung 7: Überlappung zweier Von-Hann-Fenster

Zusammenfassung

Aenean ac nulla ipsum. Sed nulla dui, consectetur sit amet ultrices eget, semper nec ipsum. Pellentesque lacinia ornare sapien, ac accumsan nulla congue eget. Aliquam gravida nulla id justo egestas accumsan.
Vestibulum convallis malesuada faucibus. Vestibulum ligula turpis, venenatis vel gravida at, eleifend eget tortor. Phasellus blandit nisi vel leo euismod a vestibulum est vestibulum. Duis convallis dignissim turpis. Nam ullamcorper molestie urna et iaculis.

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.