Technische Universität Berlin
Fakultät I: Geistes- und Bildungswissenschaften
Seminar: Mensch-Maschine-Kommunikation (WS 2016/17)
Dozent: Dr. Felix Burkhardt
Referentin: Cleopatra Christina Moshona
Datum: 15. November 2016

Thema: Formantsynthese

Historie:

- Wendepunkt in der Sprachsynthese → Entwicklung einer **akustischen Theorie** der Sprachproduktion
- Sprache (*speech*) = Anregung eines linearen Filters (hier: simuliert durch die Resonanzfrequenzen des Ansatzrohres) durch eine / mehrere Quellen
    - Primäre Quelle.: Glottis
    - Sekundäre Quelle: Rauschen → Luftdruckveränderungen durch Engebildung
    - → **Quelle-Filter-Modell**

Analoge Systeme:

- Lawrence (1953): PAT (*Parametric Artificial Talker*): Parallelschaltung
- Fant (1953): OVE (*Orator Verbis Electris*): Reihenschaltung, Kaskade

Kommerziell:

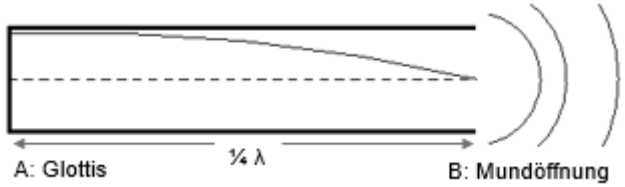
- 1974: Votrax, hardware formant synthesizer
- 1982: Dectalk, basiert auf Klatt's MITTalk, Hybrid-Modell, regelbasierte Formantsynthese
- 1994: Infovox 210/230, KTH Stockholm, Nachfolger des OVE/OVE II von Holmes (1973)
- 1996: TruVoice
- 1998: ETI Eloquence, regelbasierte Formantsynthese, IBM lizensiert
- 2009: Orpheus

Forschung:

- 1994: Multivox 3, TU Budapest
- 1998: Felix, Berkom (Tochterunternehmen der Deutsche Telekom)
- 2006: e-speak, Jonathan Duddington, basiert auf UNIX Programm „speak" von 1995 (open source)

Quelle-Filter-Modell:

- Im Ansatzrohr befindet sich eine Luftsäule → Form der Luftsäule wird durch die Stellung der Artikulationsorgane bestimmt
- Idealisierte Vorstellung anhand des [ə]: einseitig geöffnetes, gerades Rohr mit konstantem Querschnitt (Kundt'sches Rohr)



- Luftsäule wird durch den Primärschall zum Schwingen angeregt → Luftsäule hat Frequenzbereiche, bei denen Resonanzen auftreten (= Eigenfrequenzen)
- Lage der Resonanzfrequenzen ist abhängig von der Länge des Rohres (je kürzer das Rohr, desto höher liegen die Resonanzfrequenzen)
- Resonanzfrequenzen ergeben sich nur für diejenigen Wellenlängen die zwei Bedingungen erfüllen:
    - **1. Druckverteilung an der Glottis maximal**
    - **2. Schalldruck am Mund = 0**

- Bedingungen sind dann erfüllt, wenn die Länge des Rohres ¼ der Wellenlänge (bzw. ¾ λ etc.) entspricht
- **Diese Resonanzbereiche werden Formanten genannt!**
- Lage der Formanten kann mit folgender Formel berechnet werden:

$$F_n = \frac{(2n - 1) \cdot c}{4 \cdot L}$$

- Schallgeschwindigkeit c = 340 m/s und L = 17 cm
- F1: n = 1, F2: n = 2 etc.
- Dämpfung der Formanten durch Schleimhäute des Ansatzrohres oder durch sog. „Antiformanten" (Nasalformanten) → Resonanz weniger intensiv, erstreckt sich aber über einen breiteren Bereich
- Resonanzbereiche wirken wie ein Filter für den Primärschall → nur diejenigen Obertöne des Kehlkopfsignals werden durchgelassen, die im Bereich der Formanten liegen, Rest wird stark gedämpft bzw. rausgefiltert
- 2. Modifikation an den Lippen: höhere Frequenzbereiche werden verstärkt

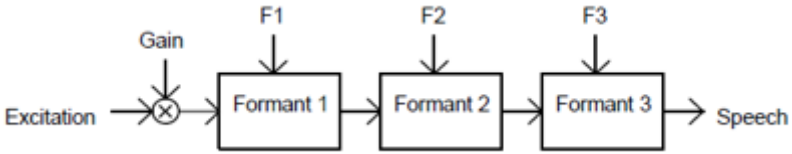
→ Input – Übertragungsfunktion – Output

Warum Formantsynthese?

- Formantsynthese versucht die Filterfunktion des Ansatzrohres zu imitieren
- Die Filterfunktion erlaubt es uns ein intelligibles Signal wahrzunehmen (Kehlkopfton = schriller Ton von ca. 120-130 dB!)
- Formanten (bzw. Antiformanten, vgl. poles-and-zeros) sind wichtig für:
  - Qualität und Identifizierung von Lauten
  - Natürlichkeit und Stimmklang
- d.h., wenn Sprache synthetisiert werden soll, muss man wissen welche Frequenzbereiche wie stark verstärkt oder gedämpft werden damit die Laute als solches identifiziert werden können

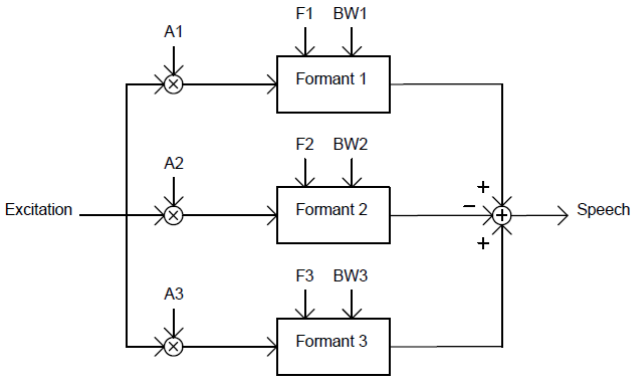
- 1951 - 1967: Cooper, Delattre, Liberman entwickeln Empfehlungen für „Syntheseregeln", basierend auf ihre Beobachtungen bei der Identifizierung englischer Laute (vgl. Locus-Theorie)
- **Akustische Hinweise (*cues*) für die Identifizierung eines Lautes überlappen sich zeitlich mit den *cues* benachbarter Laute und sind kontextbedingt**
- Für die Sprachsynthese sind auch die Übergänge zwischen den Lauten wichtig. Diese Formant-„Übergänge" nennt man Transitionen (= Formantbewegungen, Formantverläufe)

Kaskade:



- Reihenschaltung von Bandpass-Resonatoren, bei der das Output jedes einzelnen Formant-Resonators in das Input des darauffolgenden Formant-Resonators einfließt
- Vorteil: eine Kaskade ist leicht zu bedienen und kommt mit relativ wenig Information aus:
  - Formantfrequenz, Formantbandbreite
  - die relative Amplitude von Vokalen muss nicht einzeln eingestellt werden, das durch die Kaskade produzierte Resultat ist auch so gut
- Geeignet für Vokale + stimmhafte Laute (aber keine Nasale!), ungeeignet für Frikative und Plosive
- Ist vom Prinzip her sehr ähnlich zu der eigentlichen Übertragungsfunktion des Ansatzrohres

Parallelschaltung

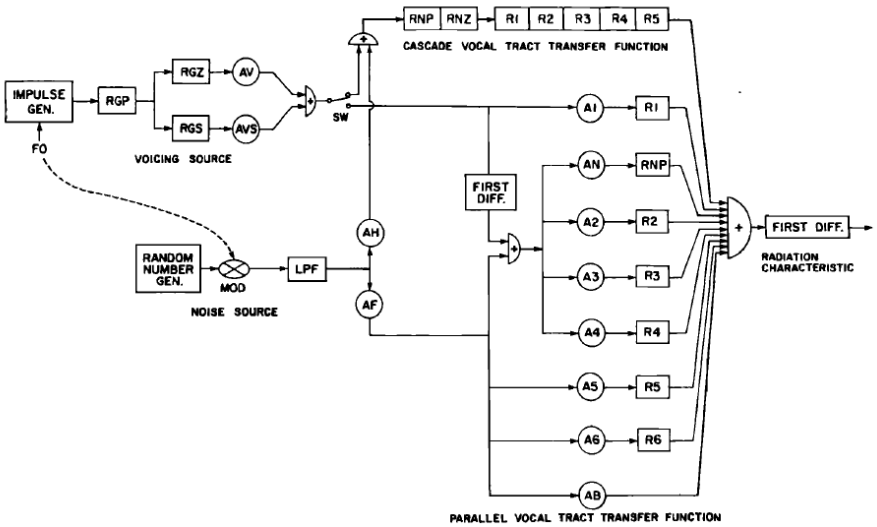


- Die Resonatoren sind parallelgeschaltet, für Nasale werden zusätzliche Resonatoren verwendet
- Der Impuls (*excitation*) wird an alle Resonatoren gleichzeitig weitergeleitet und deren Output wird aufsummiert
- Vorteil: Die Bandbreite und Amplitude (*gain*) jedes einzelnen Formanten kann separat eingestellt werden
- Nachteil: benötigt mehr Information
- Besser geeignet für Nasale, Plosive, Frikative
- Ungeeignet für einige Vokale

Hybrid-Modell von Klatt:

- Zusätzliche Resonatoren und Anti-Resonatoren für Nasale bzw. nasalierte Laute
- 6. Formant für hochfrequentes Rauschen
- Bypass-Filter für flache Übertragungsfunktionen und Abstrahlcharakteristik
- 39 Kontrollparameter, die jede 5 ms aktualisiert werden

Technische Umsetzung: Der Klatt-Synthesizer



Funktionsweise:

Zunächst wird zwischen periodischer und rauschhafter Quelle differenziert

Periodische Quelle:

1. Es wird eine Impulsfolge generiert → entspricht den einzelnen Glottisschlägen, zeitlicher Impulsabstand ist frequenzabhängig
2. Impuls wird tief-pass-gefiltert und dabei geglättet um ein Glottis-ähnliches Signal zu simulieren
3. Das geglättete Signal wird dupliziert und a) erneut geglättet um Sinus-Charakter zu erhalten (RGS) und b) durch einen Anti-Resonator gedämpft (RGZ), der das Signal weiter verfeinert
4. Die Amplituden von RGS und RGZ werden eingestellt
5. Diese werden mit dem aspirativen Teil (AH) aufsummiert und an die Kaskade weitergeleitet

6. In der Kaskade kann Nasalität berücksichtig werden (RNP, RNZ), ansonsten wird das Signal durch fünf Resonatoren (= Formanten) geleitet

Rauschhafte Quelle:

1. Der geräuschhafte Charakter wird durch die Verwendung von Zufallszahlen generiert
2. Die Lautstärke kann analog zur periodischen Quelle moduliert werden (MOD)
3. Das Signal wird tief-pass-gefiltert und dabei geglättet
4. Der AH-Teil wird an die Kaskade, der AF-Teil an die Parallelschaltung weitergleitet
5. AF wird für die Simulierung von Frikativen verwendet. Da diese im hochfrequenten Bereich viel Energie aufweisen, kann ein 6. Resonator verwendet werden

Abschließend werden die einzelnen Signalkomponenten aufsummiert und die Abstrahlungs-Charakteristik durch eine Funktion berechnet

Vorteile-Nachteile

Vorteile:

- Verbraucht wenig Ressourcen, hat geringe Anforderungen
- f0 gut kontrollierbar
- Ist recht flexibel, weil theoretisch unendlich viele Sounds produziert werden können
- Gut verständlich
- Qualität bleibt konstant

Nachteile:

- Klingt insgesamt ziemlich unnatürlich
- Zu simpel um komplexe artikulatorische Synergie wiederzugeben
- Zu „symmetrischer“ spektraler Abfall
- Prosodische Aspekte fließen nicht mit ein (insbesondere Mikroprosodie)
- Artikulatorische Übergänge zu abrupt
- Verhältnis periodischer-aperiodischer Anteile unnatürlich
- Frauenstimme sind schwieriger zu simulieren

Literatur:

Burkhardt, F. (2016): *Deutsche Sprachsynthese*. Abgerufen am 14. November 2016 von: http://ttssamples.syntheticspeech.de/deutsch/

Klatt, D. (1980): *Software for a Cascade/Parallel Formant Synthesizer*. Journal of the Acoustical Society of America. JASA, Vol. 67: 971-995.

Klatt, D. (1987): *Review of Text-to-Speech Conversion for English*. Journal of the Acoustical Society of America. JASA Vol. 82 (3), pp.737-793.

Lemmetty, S. (1999): *Review of Speech Synthesis Technology*. MSc. Thesis. Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland.

O‘Shaughnessy, D. (2000): *Speech Communications: Human and Machine*. Wiley-IEEE Press.

Taylor, P. (2009): *Text-to-Speech Synthesis*. Cambridge University Press