Hochschule der Medien

Ton Seminar 41201

Prof. Oliver Curdt

Semesterarbeit im WS `11/`12

- Sprachsynthese -

von

Katharina Stroh (Matrikelnr. 22763), AMB 4

und Nathaniel Haezeleer (Matrikelnr. 22160), AMB 4

Inhaltsverzeichnis:

Man kann keine Zeile schreiben, ohne einen Sprachfehler zu machen.

Ludwig Tieck, *Der gestiefelte Kater*, 1.Akt 2.Szene (1797)

1. Definitionen

Definition Sprachsynthese:

Sprachsynthese ist die Erzeugung und Ausgabe menschlicher Sprache durch Maschinen zum Beispiel dem Computer.

Zielsetzung ist eine natürlich klingende Aussprache und die Erzeugung einer natürlichen Satzmelodie (Prosodie).

Definition Prosodie:

Prosodie umfasst alle charakteristischen Eigenschaften einer Sprache, also Akzente, Intonation, Rhythmus, Pausen etc.

Sie ist das Hauptmerkmal der menschlichen Sprache. Weil eine natürliche Prosodie extrem schwer zu generieren ist, können wir bisher Computerstimmen von menschlichen Stimmen unterscheiden.

2. Arten der Sprachsynthese:

2.1 Konkatenative Synthese (Signalmodellierung mit Samples, korpusbasiert):

Zwei Möglichkeiten:

Phrase Splicing:

Nur einzelne Worte werden aufgenommen und später zu Sätzen zusammengefügt.

Somit sind die Anwendungsbereiche aufgrund des kleinen Wortschatzes stark begrenzt. (Navi, Bahnhofsdurchsagen)

Unit selection:

Worte werden in Phone, die kleinste sprachliche Einheit, unterteilt. Diese Phone können dann wieder zu Worten und Sätzen zusammengefügt werden. Außer Phonen werden auch Diphone und

Halbsilben zur Sprachsynthese verwendet. Diphone sind die Übergänge an zwei Phonen, Halbsilben sind die Unterteilung von Silben in eine Anfangs- und eine Endsilbe.
Durch die Eingabe einer begrenzten Zahl an Phonemen entsteht auch nur ein begrenzter verfügbarer Wortschatz, dennoch größer als beim Phrase Splicing.
Die eigentliche Synthese findet dann beim Verknüpfen dieser Sprachstücke statt. Im Extremfall entsteht die Synthese nur aus der Verkettung der vorliegenden Segmente. Die Problemstellen sind immer die Übergänge, die nicht hörbar sein sollten, um einer Prosodie nicht im Weg zu stehen. Heikel dabei ist, dass die Betonung bei der Aufnahme so neutral wie möglich sein muss, damit das Sprachstück später im fertigen Satz nicht unnatürlich oder übertrieben wirkt.

Anwendungsmöglichkeiten: Umfangreiche TTS-Systeme

2.2 Artikulatorische Synthese (regelbasiert)**:**

Hierbei wird versucht, die menschliche Sprachphysiologie zu modellieren und so einen künstlichen „Sprachapparat“ zu erzeugen, der über einen unbegrenzten Wortschatz verfügt. Das physiologische Modell muss den realen Umständen weitestgehend entsprechen, damit natürlich klingende Ergebnisse erzeugt werden können. So sind wichtige Faktoren zum Beispiel die Eigenschaften des Vokaltrakts, der Hohlraum der Lungen, die Wölbung des Gaumens oder die Stellung der Artikulatoren selbst. Problematisch ist die Erfassung der Gesamtheit der Faktoren und deren genaue Berechnung, denn umfassende Modelle von Stimmbändern und die Bestimmung des Einflusses der Filterfunktion des Vokaltraktes erfordern viele physiologische Details, die teilweise bis heute nicht im ganzen erfasst werden konnten. Aus diesem Grund wird die Entwicklung artikulatorischer Sprachsynthese-Engines nicht im kommerziellen Bereich gefördert, sondern eher im experimentellen Bereich.

Beispiel: Martin Riches The Talking Machine (1990)
http://www.youtube.com/watch?v=WClZcQo9l6Q

Definition Artikulatoren:

Lippen, Zunge, Unterkiefer, Kehlkopf etc.
Die Komplexität solcher Modelle ist nur im Ansatz ersichtlich, deshalb beschränken sich die meisten Modelle auf 2 Dimensionen.

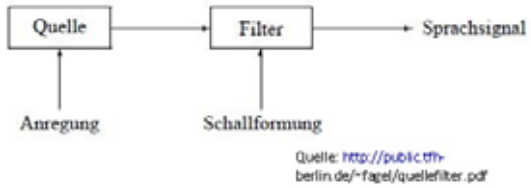
2.3 Formantsynthese (regelbasiert)**:**

Bei der Formantsynthese wird menschliche Stimme ohne Aufnahmen von Sprechern erzeugt.

Defintion Formanten:

Formanten sind Frequenzbereiche im menschlichen Stimmspektrum, die stärker hervortreten, als benachbarte Bereiche. Ihre Lage ist unabhängig von der Grundfrequenz. Wo sich diese Formanten im Frequenzbereich befinden, ist abhängig von mehreren Faktoren. So spielen zum Beispiel die Form des Vokaltrakts, oder die Eigenheiten Artikulatoren eine entscheidende Rolle.

Dies entspricht dem „Quelle-Filter-Modell“ nach Fant (1970). Mit elektronischen Schaltkreisen lässt sich so sehr einfach durch Hinzufügen einer Übertragungsfunktion *f* eine Stimme mit verschiedenen Intonationsoptionen erzeugen. Das hier gezeigte Quelle-Filter-Modell ist sehr vereinfacht dargestellt. Heutige Sprachsynthese-Engines arbeiten mit weitaus komplexeren Modellen, die feinere Einstellungsmöglichkeiten bieten. Allerdings klingen die erzeugten Ergebnisse meistens nicht so natürlich, wie die der korpusbasierten Methoden, da diese auf Segmentierung menschlicher Sprachaufnahmen basieren.



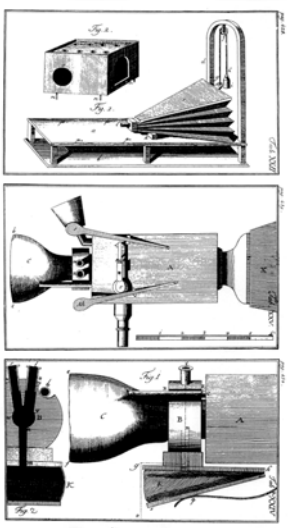
Formanten sind darüber hinaus auch ein wichtiges Unterscheidungsmerkmal zwischen verschiedenen menschlichen Stimmen, ebenso wie für verschiedene Musikinstrumente. Sie sind eine „spektrale Unterschrift“.

Bei folgendem Beispiel nimmt allerdings der Mensch eine interpretierende Aufgabe ein, indem er entscheidet, welche Silben anders betont werden sollen. Ziel ist es, eine Maschine zu schaffen, die dies selbstständig macht.

Beispiel: http://www.youtube.com/watch?v=0rAyrmm7vv0 (VODER nach Dudley 1939)

3. Geschichte

Ende des 18 Jahrhunderts, taucht die ersten Maschinen auf, die versuchten die menschliche Sprache nachzubilden. Dem deutschen Wissenschaftler Christian Kratzenstein gelang es, mit an Orgelpfeifen angeschlossenen Resonanzröhren, die Vokale „A E I O U“ hervorzubringen. Ganze Worte oder Sätze waren allerdings noch nicht möglich.



Dieses Problem sollte Wolfgang von Kempelen 1791 lösen, als er "die sprechende Maschine" erfand. Sie konnte nicht nur Phone erzeugen, sondern auch in Worten oder sogar in kurzen Sätzen sprechen. Von Kempelen orientierte sich beim Bau seiner Maschine an den menschlichen Sprachorganen.
Ein Blasebalg stellt die Lunge dar, die mit dem rechten Unterarm betätigt wird. Ein Gegengewicht sorgt dafür, dass sich der Blasebalg wieder füllt, so wird das menschliche Atmen nachgeahmt.
Die Windlade regelt die Luftzufuhr, an ihr sind verschiedene Hebel angebracht, die mit der rechten Hand zu bedienen sind. Die Stimmlippen werden von einem aufschlagenden Rohrblatt simuliert, der der Mund von einem leeren Gummitrichter. Durch Verändern der Gummiabdichtung von Hand werden die Vokale beeinflusst und bestimmte Konsonanten erzeugt.
Außerdem gibt es Öffnungen, die die Funktion der Nase übernehmen. Wenn keine Nasale oder Nasalvokale hervorgebracht werden sollen, werden die Nasenlöcher mit den Fingern verschlossen.

http://www2.ling.su.se/staff/hartmut/kempln.htm

Von Kempelen war darin bemüht seine Fortschritte in Büchern nieder zu schreiben, damit Nachfolger sein Werk verbessern konnten.

Demonstration eines Nachbaus:
http://www.youtube.com/watch?v=zYRVqrfY3tQ&feature=related

Joseph Faber entwickelte 1835 auf Basis von von Kempelens Buch die „Euphonia“. Dieses Gerät hatte im Vergleich zu von Kempelens Maschine einen formveränderten Rachenraum und Zunge. „Euphonia“ konnte das Lied "God save the Queen" singen. Der Blasebalg wurde über ein Fuß Pedal betrieben, die restliche Bedienung erfolgte über eine Klaviatur.
Im 19. Jahrhundert wurden einige weitere Maschinen ähnlicher Art konstruiert, jedoch gab es keine grundsätzlichen Neuerungen zu verzeichnen.

(links: http://www2.ling.su.se/staff/hartmut/kempln.htm)

In den 30er Jahren des 20. Jahrhunderts wurde der Vocoder erfunden.
Homer Dudley verbesserte den Vocoder zum Voder, welcher der Öffentlichkeit 1939 bei der Weltausstellung In New York präsentiert wurde. Er ermöglichte es, Sprachschall auf elektrischem Wege hervorzubringen. Die Bedienung erforderte eine sehr lange Übungszeit und diente hauptsächlich der Unterhaltung.
Ein anders Sprachsynthesesystem war der Pattern Playback. Dieses Gerät wurde 1950 von Frank Cooper erstellt. Es diente der Untersuchung der Wahrnehmung der Sprache.

http://www2.ak.tu-berlin.de/Studio/Meyer-Eppler/Meyer-Eppler.html

Martin Riches: The Talking Machine (1990)

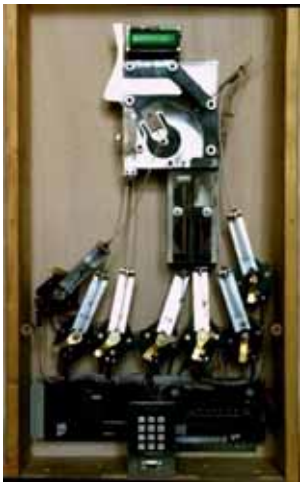
Diese Maschine besteht aus 23 Rohren, die jeweils für einen Sprechton sorgen. Entweder durch eine Pfeife, oder ein Rohrblatt wie bei Instrumenten. Ein Resonanzkörper repräsentiert die Mundhöhle und verändert den bisherigen Klang in ein Stück Sprache. Die Ventile, die die Rohre mit Luft versorgen bzw. den Luftstrom steuern, sind an einen Computer angeschlossen, von dem aus der Benutzer eingaben machen kann. Den eingegeben Silben entsprechend, ist entweder nur ein Rohr oder mehrere Rohre aktiv. So entsteht ein Wortschatz, der aus einigen hundert englischen Wörtern besteht und das zählen bis 100 in verschiedenen Sprachen umfasst.

http://www.martinriches.de/

Demonstration: Zählen auf Deutsch:

http://www.youtube.com/watch?v=WClZcQo9l6Q

Im Laufe der Jahre wurde viele verschiedene Modelle entwickelt, erneuert und verbessert, alle mit einem Ziel - das Erzeugen menschlicher Sprache durch Maschinen. Doch auch ein *praktischer* Nutzen war den Forschern ein Anliegen: Bei der Übertragung von menschlicher Sprache in einem Telefongespräch, konnte man durch Codierung der Sprache Bandbreite sparen und somit eine größere Anzahl an Gesprächen über eine einzelne Leitung schicken. Hierzu wurde der *Vocoder* benutzt. Der *MotorMouth* (1996-1999) war ein eher futuristisch aussehende Nachempfindung des menschlichen Vokaltraktes, mit 86 cm Höhe allerdings um einiges kleiner als sein großer Vorgänger *The Talking Machine*. Mit 8 kleinen Motoren, Rohren und Bowdenzügen konnte er bereits Nasale und Halbsilben erzeugen. Auch die Intonation eines Satzes war auf Wunsch variabel. Parallel zu Sprachmaschinen wurden auch Maschinen zum Spielen von Musik entwickelt, welche unter ähnlichen Funktionsschemata funktionierten.



(Rechts: MoMo“(1996-1999) von Martin Riches www.martinriches.de)

Weitere Beispiele im Internet:

Voder (1993) http://www.youtube.com/watch?v=0rAyrmm7vv0&feature=related

4. Probleme bei der Sprachsynthese:

Qualitätsmerkmale für Sprachsynthese:

- Silbenübergänge: Wie natürlich werden die Übergänge gesprochen?
- Prosodie: Ergibt die Satzmelodie Sinn oder ist sie ein Wirrwarr?
- Sprechtempo: Zu schnell oder zu langsam?
- Sprechrhythmus: Gliederung der Information oder monotones „Ablesen“?
- Pausen: Gibt es Pausen an den richtigen Stellen?

Unterscheiden von Syntax und Semantik:

Die Maschinen bekommen nur ein Textstück, welches dann natürlich klingend ausgesprochen werden soll. Um eine echte Prosodie zu erlangen, müsste die Maschine wissen, was vor dem Satz war, was die Intention des Senders ist und wer der Empfänger des Satzes ist.
Ohne diese Informationen können, wenn überhaupt, nur Texte ohne größeren Informationsgehalt übertragen werden, ohne maschinell zu klingen. Im Prinzip also Sätze, die auch ein Mensch monoton aussprechen würde.

Anglizismen/Abkürzungen:

Anglizismen können, genau wie Abkürzungen, einfach in das Wörterbuch übernommen werden. Allerdings mangelt es hier meist an der korrekten Aussprache. Mit Eigennamen bestehen dieselben Probleme. Beispiele sind auf folgender Seite zu finden:
http://ttssamples.syntheticspeech.de/deutsch/

Zahlen/Jahres-/Mengenangaben:

Für die richtige Aussprache von Zahlenreihungen ist die Kenntnis des Kontextes unabdingbar.

Richtiges Einsetzen von Emotionen:

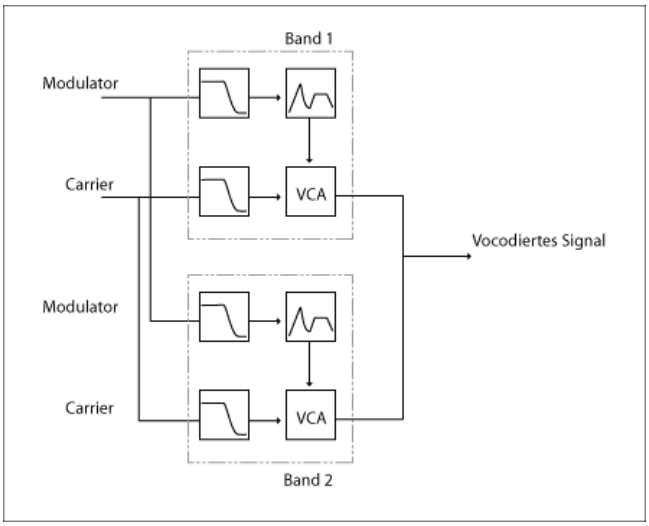
Für eine menschliche klingende Aussprache sind die dem Kontext mitschwingenden Emotionen ausschlaggebend. Bisher fehlt in der Sprachsynthese der Ausdruck von Stimmungen, den die Maschine automatisiert aus dem Kontext erkennt. Forscher analysieren deshalb das Klingen der Sprache von Menschen die traurig, fröhlich, müde u.v.m. sind. In einem zweiten Schritt werden die

so gesetzten Regeln für die Synthese verwendet. Doch schon beim Erstellen von Verhaltensregeln stellen sich große Definitionsprobleme, denn woran wird z.B. Sarkasmus oder Ironie erkannt?

5. Vocoder

Geschichte und Entstehung:

Der Vocoder taucht in der Geschichte das erste mal 1936 auf. Der Name ist ein englischer Neologismus aus "voice" und "coder" zu Deutsch: "Stimmcodierer". Homer Dudley entwickelte und verbesserte den Vocoder zum Voder oder auch „Voice Operation Demonstrator“, der 1939 der Öffentlichkeit bei der Weltausstellung in New York vorgestellt wurde.
Anfangs wurde er beim Militär genutzt, zum Frequenzmultiplexing durch Reduzierung der Bandbreite und zur Verschlüsselung von Gesprächen über elektrische Leitungen. Dadurch wird die menschliche Sprache elektrisch kodiert und dann beim Empfänger wieder reproduziert. Heute wird er häufig in der Computertechnik zur verlustfreien Reduktion des Datenumfangs, in der Sprachforschung und in der Musik als Effektgerät verwendet.



Links: Vocoder mit 2 Bändern

http://www.dma.ufg.ac.at/app/link/Grundlagen%3AAudio/module/8086?step=1

Funktionsweise:

Der Vocoder besteht aus einer Aufnahmeeinheit die für die Klanganalyse zuständig ist und einer Wiedergabeeinheit zur Klangsynthese.

Analyse:

Über die Aufnahmeeinheit (Mikrofon) gelangt das Analysesignal, die menschliche Stimme, in das Gerät. Das Signal durchläuft mehrere Bandfilter (die Filterbank) und wird in Frequenzbändert zerlegt. Jeder Filter analysiert dabei einen kleinen Bereich des Frequenzspektrums des Gesprochenen.

Die Zahl der Filter von Vocodern reicht von 10 bis zu 30. Mit der Anzahl der Kanäle erhöht sich die Genauigkeit der Analyse und damit die Wiedergabequalität aber auch der entsprechende Schaltaufwand.
Beim Menschen stellen die Mund-, Nasen- und Rachenhohlräume Filter dar, die bestimmte Frequenzbereiche selektieren und den typischen Spektralverlauf der einzelnen stimmhaften Laute formen.
Diese Analysesignale geben dem Vocoder vor, wie das Trägersignal gefiltert werden soll. Es ist also mehr als Steuerungssignal zu verstehen, welches bestimmt, wann (Rhythmus) und wie viel (Lautstärke) die verschiedenen Filter des Vocoders geöffnet werden. Somit werden die Stellungen von Mund, Nase und Rachenraum nachempfunden.
Auf das Trägersignal, auch Carrier genannt, wird die Auswertung des Analysesignals übertragen, daraus entsteht die elektronische Nachbildung der menschlichen Stimme. Dieses Trägersignal bestimmt den Sound, Timbre und Tonhöhe, die den Ausgang des Vocoders verlassen.
Als Trägersignal eignen sich obertonreiche Synthesizerklänge wie Sägezahn- oder Pulswellen besonders gut, da die Schwingung des Kehlkopfes neben der Grundschwingung noch viele harmonische Oberschwingungen bis über 4kHz enthält.

Synthese:
Bei der Synthese wird, wie oben beschrieben, aus dem Trägersignal und dem Analysesignal ein drittes Signal erzeugt. Dieses Signal hat einen metallischen Roboter Sound.

http://www.metacafe.com/watch/sy-1568171509/lenka_vocoder_official_music_video/

Was hier zu hören ist, ist das Trägersignal, welches vom Analysesignal unter Einsatz von Filterbank und spannungsgesteuerten Verstärkern moduliert wird.
Oder anders: die Filterbank des Vocoders öffnet ihre Tore für das Trägersignal immer nur für den Frequenzbereich, den das Analysesignal vorgibt. Gleiches gilt für den VCA, der die Lautstärke dynamisch regelt.

Zusammenfassend untersucht der Vocoder das eingehende Modulationssignal in Bezug auf seine Lautstärkeamplitude und sein Frequenzspektrum und übersetzt diese Werte auf das Trägersignal.

6. Anwendungsbereiche

Die synthetische Sprache hat sich in vielen Bereichen als sehr nützlich erwiesen. Mit Text-To-Speech Systemen wird Sehbehinderten Menschen das Leben einfacher gemacht, indem ihnen Emails, SMS oder Texte auf Websites vorgelesen werden. Navigationssysteme, Bahnhof- oder Zugdurchsagen arbeiten ebenfalls mit einem synthetischen Wortschatz. Es gibt kaum eine Telefongesellschaft, die Anrufe ohne eine künstliche Stimme empfängt, bevor sie dann, nach Auswahl verschiedener vorgeschlagener Optionen, an einen menschliches Gegenüber verbindet. Die Vorteile liegen auf der Hand: Einsparung von menschlichen Mitarbeitern, die ständige Verfügbarkeit und Beständigkeit der Stimme. Dem gegenüber stehen die noch zu künstlich klingende Aussprache, das Fehlen von echter Prosodie und die Trennung von Syntax und Semantik.
Die meisten Sprachsyntheseunternehmen bieten männliche und weibliche, erwachsene und jugendliche Sprecher an, um ihren Kunden größere Auswahl zu bieten. Die Abdeckung vieler Sprachen versteht sich von selbst. Auch verschiedene Akzente sind oftmals verfügbar, viele davon sind jedoch nur zum „Show-off“ geeignet.
Auch im Bereich der Unterhaltung findet die Sprachsynthese bereits Anwendung. So kann man auf der Seite www.xtranormal.de aus verschiedenen Umgebungen und Räumen, Personen und Charakteren, Atmosounds und Hintergrundmusik wählen. Die Personen, die sich im Comic-Style auf Befehl auch bewegen und artikulieren, können dann mit Text gefüttert werden, den sie dann, nach einer kurzen Renderphase, in Sprache umwandeln. So hat der Nutzer die Möglichkeit, einen eigenen „Film“ im Web zu erstellen und diesen mit anderen zu „teilen“. Den ersten Film gibt es in der Regel kostenlos, alle nachfolgenden und Extrafeatures werden dann in der websiteeigenen Währung bezahlt, welche durch echte Währung gekauft werden kann.

Quellen:

http://www.csounds.com/man/appendix/formants.htm

http://ttssamples.syntheticspeech.de/deutsch/index.html#eloquent

http://public.tfh-berlin.de/~fagel/quellefilter.pdf

http://www2.ak.tu-berlin.de/Studio/Meyer-Eppler/Meyer-Eppler.html

http://www.bonedo.de/artikel/einzelansicht/waldorf-lector.html

http://www.bonedo.de/artikel/einzelansicht/tal-vocoder.html

http://www.ths-nation.de/recall/vocoder.htm

http://www.dma.ufg.ac.at/app/link/Grundlagen%3AAudio/module/8086?step=1

http://de.wikipedia.org/wiki/Vocoder

http://www2.ling.su.se/staff/hartmut/kempln.htm