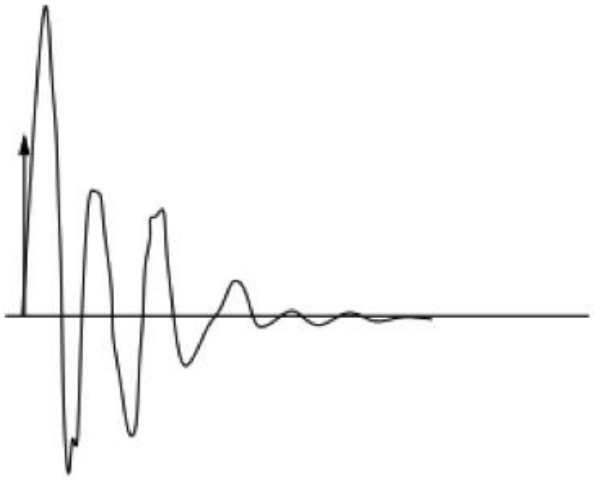

Unit Selection

Unterschiede zwischen Unit Selection und der 2. Generation Sprachsynthese

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
2020 – 08 – 30

Inhaltsverzeichnis

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
30. 08. 2020

Einleitung

Sprachsynthese ist der Bau von Sprach-Apparaturen. Der erste Versuch gelang Ch. G. Kratzenstein 1773, der anfing Vokale zu erzeugen indem er Resonanzröhren auf Orgelpfeifen montierte.

1791 veröffentlichte Wolfgang von Kempelen ein Buch mit dem Titel „Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine“ in dem er jene Maschine genau beschreibt. Mit dieser Maschine lassen sich ganze Worte und kurze Sätze bilden. Sie besteht aus einem Blasebalg mit Gegengewicht, der die Atmung simuliert. Der Luftstrom wird dann durch ein Rohrblatt und ein Rohr in den „Mund“ geleitet, was einen Anstieg des Luftdruckes bewirkt. Stimmlose Laute lassen sich durch das Verschließen des Ansatzrohres erzeugen. Die Stimmlippen werden durch ein Rohrblatt aus Elfenbein dargestellt.

1835 baute Joseph Faber die Maschine „Euphemia“, die normale Sprache, aber auch Gesang und Flüstern synthetisieren kann.

R. R. Riesz baute 1937 eine Maschine, die sehr ähnlich zu der von von Kempelen war, jedoch ein naturgetreueres Ansatzrohr hatte.

Der VODER wurde 1939 von Homer Dudley mit Hilfe von Elektrotechnik aus einem VOCODER gebaut, der eigentlich die Bandbreite von Telefongesprächen reduzieren sollte.

1950 baute Frank Cooper den Pattern Playback, der monotone Sprache erzeugen konnte.

Seit 1950 arbeiten verschiedene Forscher an den Synthesetechniken der 2. Generation, bei der elektrische Impulse erzeugt werden, die dann Lautsprechermagnete in Schwingung

versetzen, und seit 1970 auch an der Unit Selection, bei der vorher aufgenommene Laute wiedergegeben werden.[1]

Für die 1. Generation hatte man ganze Worte einsprechen lassen und die einfach abgerufen, was einen roboterartigen Klang hat. Im Laufe der Zeit hatte man dann die Worte in kleinere Schnipsel aufgeteilt, da sich diese besser zusammenschneiden lassen und das Ergebnis natürlicher klingt. Bei der Unit Selection ist man bei den einzelnen Lauten angekommen.

Das Ziel dieser Arbeit liegt darin die Unterschiede zwischen der Unit Selection, einer TTS[2]-Synthese-Technik, und der 2. Generation der Sprachsynthese herauszuarbeiten und beide miteinander zu vergleichen, um herauszufinden, ob die Unit Selection denn geeigneter ist, als die 2. Generation.

[1] http://www.cs.columbia.edu/~julia/cs4706/tts-history.htm

[2] TTS ist die Abkürzung für „**T**ext **T**o **S**peech“ und beschreibt die künstliche Übersetzung von geschriebenem Text in gesprochene Sprache

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
30. 08. 2020

Sprachsynthese 2. Generation

Entstehung aus 1. Generation und Linear Prediction

Die 1. Generation dominierte die Sprachsynthese bis in die 1980er hinein.
Verbesserungen erfolgten kaum, da der Speicherplatz noch sehr teuer war und man mit wenig Platz arbeiten musste.[3]

Man hatte bei 512 kB Speicher nicht mehr als 80 kB für den Code und ebenso wenig für die Daten.
Um bei den Signalen Platz zu sparen, werden stimmlose Konsonanten und Transienten[4] so gelassen, wie sie sind und nur die stimmhaften Phoneme synthetisiert.
Die eigentliche Synthese findet dann durch die Konkatenation der Transienten und dem Mittelteil des Phonems statt.
Da man keine Soundcard hatte, nutzte man als Output die Lautsprecher des PCs und parallele Schnittstellen[5], wodurch die synthetisierte Sprache roboterartig und monoton klingt.
Später wurde es dann einfacher die Sprachmelodie und den Sprachrhythmus zu manipulieren. Man kürzte Perioden durch das Löschen der letzten Samples und verlängerte sie, indem man leere Elemente anhing.[6]

---

[3] Taylor, Kapitel 13

[4] „Der Begriff Transient […] bezeichnet den kurzen Zeitabschnitt, in welchem ein plötzlich in Schwingung versetztes schwingendes System […] zunächst noch chaotisch schwingt (Geräusch), bevor sich die Longitudinalwellen in Grund- und harmonische Oberschwingungen ordnen (Klang).“ (Wikipedia, Transient)

[5] „Die parallele Schnittstelle bezeichnet einen digitalen Eingang oder Ausgang eines Computers oder eines Peripheriegerätes. Bei der Datenübertragung über eine parallele Schnittstelle werden mehrere Bits parallel übertragen, im Gegensatz zur seriellen Schnittstelle, bei der die Bits nacheinander übertragen werden.“ (Wikipedia, parallele Schnittstelle)

[6] “Three Generations of Speech Synthesis Systems in Slovakia” Darjaa Sakhia, et al, 2006, Kapitel 2

Die 2. Generation der Sprachsynthese ist eine Weiterentwicklung der Linear Prediction (LP)[7], mit dem Ziel, die Probleme des vereinfachten Impulse/Noise Source Model[8] (z.B. Speicherplatzmangel) zu lösen.[9]

Gründe für die Entwicklung der 2. Generation waren die Möglichkeit eine höhere Sampling Rate[10] zu verwenden und die einfache Wiedergabe der unanalysierten Sprache, was sich in besserer Qualität äußert. Auch kann man hier mit einfachen Rekombinationstechniken Phoneme[11] und Worte erzeugen, die nicht im Originaltext vorhanden sind.
Problematisch ist dabei, dass zwar alle möglichen phonetischen Effekte erzeugt werden können, aber nicht zusätzlich auch noch die prosodischen Merkmale, wie Akzent, Pitch und Druckstärke.[12]

Obwohl die Qualität der einfach zusammengesetzten Sprache sehr hoch sein kann, setzt man sich das Ziel zusätzlich die Prosodie mit einbauen zu können.
Hier sieht man auch, warum die klassische LP Synthese so schlecht klingen kann: Man ersetzt die natürliche Lautquelle (z.B. Stimmbänder) durch einfache Impulse.[8] Phoneme tauchen in charakteristischen Sequenzen auf, z.B. findet man am Anfang von Worten oft /s t r/ aber wahrscheinlich niemals /s b r/.
Diese Phonetischen Grammatiken nennt man Phonotaktiken und sie helfen ein komplettes Diphon-Set zu ermitteln.
Eine der Aufgaben der 2. Generation ist es, ein komplettes Diphon-Set zu ermitteln. Diphone beginnen in der Mitte eines Phonems und enden in der Mitte des folgenden Phonems. [13]

---

[7] Linear Prediction beschreibt die Basis der Signalverarbeitungsprogramme, die Daten nutzt um das Sprachverhalten zu analysieren. Dies klingt jedoch stark unnatürlich, da die natürliche Lautquelle, z.B. die Stimmbänder durch eine einfache Impulsfolge ersetzt wird. (Taylor, Kapitel 14.1)

[8] Das Impulse/Noise Source Model hält eine Trennung von Quelle und Filter für die LP bereit. Dadurch kann man eine Framesequenz resynthetisieren, indem man die fundamentale Frequenz verändert (Taylor, Kapitel 13.3.2)

[9] Taylor, Paul: Kapitel 14.1

[10] Die Sampling Rate ist „die Häufigkeit, mit der ein Analogsignal […] in einer vorgegebenen Zeit abgetastet […] wird." (Wikipedia, Sampling Rate)

[11] Ein Phonem [...] ist die abstrakte Klasse [...] aller Laute (Phone), die in einer gesprochenen Sprache die gleiche bedeutungsunterscheidende [...] Funktion haben. (Wikipedia, Phonem)

[12] Taylor, Paul: Kapitel 14.1.1

Lässt sich eine Phonemsequenz nicht in die Phonotaktiken einbauen, ist sie kein Teil des gültigen Diphon-Sets. Durch die Übergänge zwischen verbundenen Worten, Namen oder Neologismen[13] gibt es nur sehr wenige ungültige Phonemsequenzen.
Wenn die einzelnen Diphone aus neutralem Kontext stammen, deutlich voneinander zu trennen und von hoher Qualität sind, lassen sie sich konkatenieren[14] ohne verändert werden zu müssen.[15]

Wenn man dann die Diphonsequenz hat, muss man um Prosodie zu erhalten nur noch Pitch und Timing anpassen.[13]

[13] Neologismen sind Wortneuschöpfungen.

[14] Konkatenation ist die Verbindung bzw. das Aneinanderhängen von Elementen

[15] Taylor, Kapitel 14.1.2

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
30. 08. 2020

Die wichtigsten Signalverarbeitungssysteme der 2. Generation

Auflistung der wichtigsten Signalverarbeitungssysteme mit einer kurzen Zusammenfassung.

PSOLA, oder auch **P**itch **S**ynchronous **O**ver**L**ap and **A**dd, gilt als das am meisten genutzte Sprachverarbeitungssystem der 2. Generation.
Es modifiziert Pitch und Timing von gesprochener Sprache ohne die Quelle und den Filter explizit voneinander zu trennen.
Im Prinzip werden die Pitch-Perioden voneinander getrennt modifiziert und dann neu zusammengesetzt.[16]

Residual Excited Linear Prediction ist eine Technik, bei der zuerst die einzelnen Pitch-Perioden voneinander getrennt werden, um dann jeder Pitch-Periode eine asymmetrische Fensterfunktion zugeteilt wird, die die Periode vom Rest trennt. Dadurch kann das Timing geändert werden.[17]

MBROLA übersetzt phonetische Anweisungen in Sprache, kann jedoch nicht direkt mit geschriebener Sprache umgehen und braucht vorher ein anderes Programm, das den Text in gesprochene Sprache umwandelt. [18]

[16] Taylor, Kapitel 14.2
[17] Taylor, Kapitel 14.3.1
[18] Taylor, Kapitel 14.5

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
30. 08. 2020

Unit Selection

Als bestes und bekanntestes Beispiel für Unit Selection kann man die Ansagen in Zügen und Bussen nennen.

Unit Selection ist eine der dominantesten Synthesetechniken für TTS und die natürliche Weiterentwicklung der 2. Generation konkatenativer Systeme.
Im Prinzip sucht die Unit Selection sogenannten „Units“ (wie Phoneme, Diphone, Halb-Phoneme/-Silben und Disilben) aus einer Datenbank und fügt sie neu zusammen.
Die Units werden so ausgewählt, dass das Ergebnis natürlich klingt. Dieses Kriterium macht den Unterschied zwischen Ziel und Unit besonders wichtig.
Diese Units werden von einem gekürzten Viterbi-Algorithmus[19] ausgewählt.[20]

Wenn davon ausgegangen wird, dass nur eine Unit pro Diphon existiert, wird schnell klar, dass die Qualität stark limitiert ist, da es keine Variationen geben kann.
Die erste Verbesserung dieses Systems war die Erweiterung um mehr als eine Unit pro Diphon.
Dies wurde dann nochmal erweitert, so dass es jetzt eine Unit pro Feature[21] gibt, also pro Diphon eine Unit mit Akzent, eine ohne Akzent, eine mit geändertem Ausdruck und eine ohne geänderten Ausdruck.
Dies ist eine Erweiterung des originalen Diphon-Prinzips, da nicht jedes Diphon einmal aufnimmt und analysiert wird, sondern jede einzelne Kombination aus Features.
Dieses Verfahren lässt sich beliebig erweitern, indem mehr Features hinzufügt werden.
Mehr Features werden jedoch schnell unpraktisch, da mehr Daten gesammelt und verarbeitet werden müssen, genau ein Beispiel pro Feature und Diphon. Auch können die Sprecher bei der Aufnahme nicht einzelne Diphone sprechen, sie brauchen Worte und

[19] „Der **Viterbi-Algorithmus** ist ein Algorithmus der dynamischen Programmierung zur Bestimmung der wahrscheinlichsten Sequenz von verborgenen Zuständen bei einem gegebenen Hidden Markov Model (HMM) und einer beobachteten Sequenz von Symbolen.“ (Wikipedia, Viterbi-Algorithmus)
[20] Hunt, Andrew J, Alan W. Black, 1996 „UNIT SELECTION IN A CONCATENATIVE SPEECH SYNTHESIS SYSTEM USING A LARGE DATABASIS“
[21] Features sind Akzente oder Ausdrucksweisen

Phrasen, wodurch unbenötigte Diphone entstehen. Diese kann man zwar einfach löschen, um Platz zu schaffen, doch wäre dies eine Verschwendung.[22]

Auch eigentlich unbenötigte Diphone werden in einer Datenbank gespeichert.
Der Viterbi-Algorithmus nutzt dann diese Daten, um Sprache zu synthetisieren.
Dann kann es sein, dass Diphone fehlen, hier kann man jedoch mit Konkatenation der bereitstehenden Units nachhelfen.
Je größer die Datenbank, desto kleiner ist die Wahrscheinlichkeit, dass Units fehlen.
Lange, zusammenhängende Sätze, die eingesprochen wurden, bieten die höchste Qualität der Datenbank und der Synthetisierung.

[22] Hunt, Andrew J, Alan W. Black, 1996 „UNIT SELECTION IN A CONCATENATIVE SPEECH SYNTHESIS SYSTEM USING A LARGE DATABASIS“

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
30. 08. 2020

Vergleich zwischen Unit Selection und der 2.Generation

In diesem Kapitel vergleiche ich die 2. Generation der Sprachsynthese mit der Unit Selection.

Einer der Hauptvorteile der Unit Selection ist, dass man aus den einzelnen Units jeden beliebigen Satz synthetisieren kann, während man für die 2. Generation jeden einzelnen Satz neu einsprechen muss.
Prosodie ist kein Problem bei der Unit Selecion, jedoch bei der 2. Generation, wenn überhaupt, kaum möglich.
Auch braucht man für die Unit Selection nur wenige Units um viele verschiedene Satze zu synthetisieren, wobei es bei der 2. Generation nur wenige Möglichkeiten gibt einen Satz neu zu erstellen.
Die großen Datenansammlungen der Unit Selection sorgen für eine hohe Genauigkeit der Synthetisierungen, verbrauchen aber auch viel Speicherplatz und es dauert länger die richtigen Units zu finden. Die 2. Generation hingegen hat deutlich weniger Daten, die aufgenommen werden müssen, und verbraucht dadurch auch deutlich weniger Speicher.

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
30. 08. 2020

Fazit

Die Unit Selection ist eine Weiterbildung der 2. Generation der Sprachsynthese. Sie erleichtert die Erstellung von neuen Sätzen und lässt prosodische Merkmale zu. Die einzigen Nachteile sind der höhere Speicherplatzbedarf und die umfangreichere Datenerhebung. Die Preise für Speicherplatz sind in der Vergangenheit kontinuierlich gesunken, wodurch der Nachteil des höheren Speicherbedarfs der Unit Selection nicht mehr stark ins Gewicht fällt.

1970, als die Unit Selection erfunden wurde, waren die Kosten für Speicher so teuer, dass sie nur sehr selten tatsächlich verwendet wurde[23], weswegen man damals weiterhin mit der 2. Generation gearbeitet hatte. Erst als der Speicher günstiger wurde, lohnte es sich dann die Unit Selection tatsächlich einzuführen, wodurch die Synthetisierungen noch mal stark an Natürlichkeit gewannen. Mittlerweile ist es möglich mit wenig Platz und wenig Geld mehrere Terrabyte Speicher zur erwerben. Selbst kleine Geräte, wie Smartphones können problemlos die benötigten 1,278,998 Bytes[24] in ihrem Speicher unterbringen.

Bestehen bleibt allerdings der Aufwand für die Erhebung der Daten. Einige Datenbanken für Unit Selection sind frei verfügbar, andere können käuflich erworben werden. Auch die Datenerhebung wird immer einfacher, da man sich die Datenbanken für die Unit Selection im Internet[25] herunterladen kann.

Heutzutage bietet es sich an die Unit Selection zu nutzen, da der benötigte Speicher weder viel Platz benötigt, noch übermäßig teuer ist. Auch die Prosodie, die sich leicht einbauen lässt, sowie die Natürlichkeit sprechen dafür die Unit Selection der 2. Generation vorzuziehen.

[23] https://de.wikipedia.org/wiki/Solid-State-Drive

[24] Boston University Radio Speech Corpus: https://catalog.ldc.upenn.edu/LDC96S36

[25] z.B. hier: https://catalog.ldc.upenn.edu/LDC96S36 oder hier: http://mary.dfki.de/

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
30. 08. 2020

Literaturverzeichnis

Boston University Radio Speech Corpus (2020-08-13):
https://catalog.ldc.upenn.edu/LDC96S36

Hunt, Andrew J., Alan W. Black (2020-08-13):
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.9132&rep=rep1&type=pdf

Taylor, Paul "Text-to-Speech Synthesis" (2020-08-13):
https://pdfs.semanticscholar.org/41ed/fc6c719fd1ba746f772a5f8e8225a1270c20.pdf

Three Generations of Speech Synthesis Systems in Slovakia, Darjaa Sakhia, et al (2020-08-13): https://www.eurasip.org/Proceedings/Ext/SPECOM2006/papers/052.pdf

Wikipedia, Parallele Schnittstellen (2020-08-13):
https://de.wikipedia.org/wiki/Parallele_Schnittstelle

Wikipedia, Phoneme (2020-08-17): https://de.wikipedia.org/wiki/Phonem

Wikipedia, Sampling Rate (2020-08-13):
https://de.wikipedia.org/wiki/Abtastrate#:~:text=Die%20Abtastrate%20oder%20Abtastfrequenz%2C%20auch,ein%20zeitdiskretes%20Signal%20umgewandelt)%20wird.

Wikipedia, Transient (2020-08-13): https://de.wikipedia.org/wiki/Transiente

Wikipedia, Viterbi Algorithmus (2020-08-13): https://de.wikipedia.org/wiki/Viterbi-Algorithmus

Carla Oppermann

Verarbeitung gesprochener Sprache 2020
Timo Baumann
30. 08. 2020

Bildquellenverzeichnis

Titel: Taylor, Figure 14.5:
https://pdfs.semanticscholar.org/41ed/fc6c719fd1ba746f772a5f8e8225a1270c20.pdf