

Adaptive audio-visuelle Synthese
Automatische Trainingsverfahren für Unit-Selection-basierte
audio-visuelle Sprachsynthese

Inaugural-Dissertation

zur Erlangung des Doktorwürde

der

Philosophischen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität

zu Bonn

vorgelegt von

Christian Weiss

aus

Heidelberg

Bonn 2007

Gedruckt mit der Genehmigung der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Professor Dr. Caja Thimm (Vorsitzende)
Professor Dr. Wolfgang Hess (Betreuer und Gutachter)
PD Dr. Bernhard Schröder (Gutachter)
PD. Dr. Ulrich Schade (weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 02. November 2006

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn
http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

INHALTSVERZEICHNIS

Liste der Abbildungen	iv
Liste der Tabellen	vi
Anmerkungen	vii
Kapitel 1: Einführung	1
1.1 Einführung und Motivation	1
1.2 Einordnung der Arbeit in den Forschungsbereich	8
1.3 Übersicht über die Arbeit	13
Kapitel 2: Grundlagen der Sprachproduktion, Sprachsynthese und audio-visuellen Synthese	16
2.1 Grundlagen Sprachproduktion und Sprachsynthese	16
2.1.1 Physiologie, Sprachproduktion und Prosodie	17
2.1.2 Sprachsynthese: Überblick und Verfahren	24
2.1.3 Konkatenative Sprachsynthese	26
2.2 Audio-visuelle Synthese	28
2.2.1 Physiologien nonverbaler Kommunikation	31
2.2.2 Modellbasierte Ansätze	32
2.2.3 Videodatenbasierter Ansatz	34
Kapitel 3: Probabilistische Vorbedingungen	37
3.1 Bedingte Wahrscheinlichkeiten	37
3.1.1 Satz von Bayes	39
3.2 Endliche Automaten.....	40
3.3 Hidden-Markov-Modelle.....	42
3.4 Entropie	46
3.4.1 Bedingte Entropie.....	47
Kapitel 4: Korpora und Vorverarbeitung	49
4.1 Definition und Eigenschaften der Sprach- und Video-Korpora	49
4.2 Annotation	52
4.3 Korpora	54
4.3.1 Audiokorpus.....	54
4.3.2 Segmentierung des Audiokorpus.....	55
4.3.3 Videokorpus.....	56
4.3.4 Segmentierung des Videokorpus	57
4.3.5 Viseme	58

4.4 Automatische Vorverarbeitungen	58
4.4.1 Maximum-Entropie-basierte Graphem-Phonem-Transkription,	59
4.4.2 Wortklassen-, Silbengrenzen-, Akzent-Prädiktion	62
4.4.3 Phonem-Viseme Klassifikation	65
Kapitel 5: Automatisches Training für die Sprachsynthese.....	69
5.1 Unit-Selection	70
5.1.2 Quantitative, phonologische und linguistische Merkmale	72
5.2 Spektrale Eigenschaften der Sprachsegmente	74
5.2.2 Mel-Cepstrum-Koeffizienten	75
5.3 HMM-basierte Sprachsynthese	76
5.3.1 Das HTS-Sprachsynthese-System	77
5.3.2 Sprachabhängiges Training der kontext-basierten HMMs	79
5.4 Korpusbasierte Sprachsynthese.....	81
5.4.1 Korpusbasierte Sprachsegmente	81
5.4.2 Datenbasierte Dauer-Modellierung.....	82
5.4.3 Datenbasierte F0-Generierung	85
5.4.4 Karhunen-Loeve Transformation und Eigenspektrum.....	86
5.5 Bedingte Wahrscheinlichkeitsmodelle zur Segmentauswahl	88
5.5.1 Probabilistischer endlicher Automat	89
5.5.2 Bedingte-Entropie-basierte Segmentauswahl	91
5.6 Bestimmung optimaler Sprachbausteinfolge	93
5.6.1 Viterbi-Algorithmus	94
5.6.2 A*-Algorithmus	95
5.6.3 Spektrale Abstandsmaßberechnung	97
5.7 Conditional Random Field basierte Segmentauswahl	98
5.8 Fazit	105
Kapitel 6: Korpusbasierte visuelle Synthese	106
6.1 Visuelle Sprachsynthese	106
6.2 KNN-basierte Auswahl der Video-Frame-Segmente	107
6.3 Parametergewinnung für die Video-Segmentauswahl.....	109
6.4 Audio-Video Konkatenation und Synchronisation.....	111
Kapitel 7: Das audio-visuelle Synthese-System AVISS	113
7.1 AVISS-Software-Implementierung.....	113
7.2 Generierung der audio-visuellen Synthese	116
Kapitel 8: Evaluation.....	119
8.1 Evaluationen der Sprach-Synthese-Ausgabe	119
8.2 Evaluationen der audio-visuellen Synthese-Ausgabe	122
Kapitel 9: Schlussbetrachtung und Ausblick	125

Literaturverzeichnis	129
Abkürzungsverzeichnis:	140
Anhang A:	141

LISTE DER ABBILDUNGEN

<i>Nummer</i>	<i>Seite</i>
1. Schematischer Aufbau audio-visueller Synthese	1
2. Schematischer Querschnitt des menschlichen Sprachtraktes	18
3. Quelle-Filter-Modell	19
4. Sonagramm des Audiosignals: Audio-visuelle Synthese.....	21
5. Blockdiagramm regelbasierte Formantsynthese	25
6. Blockdiagramm der TTS-Komponenten	27
7. Gitternetz-basierte Darstellung eines menschlichen Kopfes.....	33
8. Overlay-Technik im VideoRewrite-System	35
9. Verwendete Gesichtsregionen für das AT&T FaceTalk-System	35
10. Beispielhafte Darstellung einer Markov-Kette.....	41
11. Darstellung elementarer Hidden-Markov Modell Topologien	44
12. Darstellung eines typischen Hidden-Markov-Modells wie es in der HTK Software Verwendung findet.....	45
13. Darstellung der zwei Worte „werden“ in unterschiedlichen Kontexten mit unterschiedlicher prosodischer Realisierung	50
14. LRNE, Verteilungsfunktion von Einheiten	52
15. Darstellung Einheiten-Grenzen (Satz, Wort, Silbe).....	57
16. a) Übersicht der Vokaleinteilung nach Zungenstellung und Grad der Mundöffnung; b) SAMPA-IPA Zuordnung.....	66
17. Auswahlgraph auf Phoneebene zur Konkatenation der Sprachsegmente, hier: Synthese angegeben in phonetischer Transkription /S Y n t e: s @/	72
18. Darstellung des Sprachsignals: (1) Zeitbereich, (2) Frequenzbereich.....	74
19. Blockdiagramm: Training HTS-Sprachsynthese-System	78
20. Blockdiagramm: Erzeugung eines Sprachsignals zur Laufzeit mit dem HTS-Sprachsynthese-System	79
21. Dauerverteilung in Millisekunden für /d a s/ im zugrunde liegenden Sprachdaten-Korpus	83

22. Schemenhafte Darstellung eines Entscheidungsbaumes zur Vorhersage von Silbendauern.....	84
23. Durchschnittlicher log F0 Wert für /d a s/ im zugrunde liegenden Sprachdaten-Korpus	84
24. Hauptkomponentnanalyse der Mel-Cepstrum-Koeffizienten.....	86
25. Hauptkomponentnanalyse von zwei unterschiedlich realisierten /a:/.....	87
26. Probabilistischer endlicher Automat: S0 gibt den Startzustand an	90
27. Darstellung zweier unterschiedlich generierter Sprachsignale. Obere Bild: Segmentauswahl und Konkatenation mittels bedingter Entropie; Unteres Bild: Segmentauswahl und Konkatenation mit 2-dimensionalen Kostenfunktion.....	93
28. Sprachsegment-Graph auf welchem der A*-Algorithmus die Suche vom Startsegment (C11) zum Endsegment (C71) durchführt.	95
29. a) Ungerichteter Graph; b) Beobachtungs- und Zustandssequenz mit zugehöriger Konfiguration	99
30. Schematische Darstellung des Trainings der kontextbasierten CRF Modelle.	102
31. Schematische Darstellung der Sprachsignalgenerierung mittels CRFs.....	103
32. Darstellung zweier unterschiedlich generierter Sprachsignale Oberes Bild: Segmentauswahl und Konkatenation mittels CRF Unteres Bild: Segmentauswahl und Konkatenation mit Kostenfunktion.....	104
33. Schematische Darstellung der KNN in Betracht kommenden Segmente	107
34. Schematische Darstellung der KNN Segmentabfolgenauswahl.....	109
35. Schematische Darstellung Filter zur geometrischen und pixelbasierten Merkmalsextraktion.....	110
36. Darstellung des Eingabe-Vektors für die Hauptkomponentnanalyse und die durchgeführte Hauptachsentransformation.....	111
37. Benutzeroberfläche der AVISS Software	117
38. Diagramm der Evaluationsergebnisse: Markov-Entropie basierte Synthese ...	121
39. Diagramm der Evaluationsergebnisse: CRF basierte Synthese.....	122
40. Diagramm der Evaluationsergebnisse audio-visuelles Synthesesignale	123

LISTE DER TABELLEN

<i>Nummer</i>	<i>Seite</i>
1. Übersicht der visuellen Darstellungsformen in der audio-visuellen Synthese und deren Bezeichnung.....	29
2. Attribute der Sprachsegmente in Sprachdatenkorpus.....	53
3. Übersicht über die Verteilung der Sprachsegmente	54
4. Übersicht über die Lippenstellung bei den einzelnen Visem-Klassen	58
5. Verwendete Phonemeinheiten mit Verteilung im Trainingskorpus.....	60
6. Ergebnisse der korrekten Klassifizierung von Graphem-Phonem.....	62
7. Ergebnisse der korrekten Klassifizierung von Wortklassen (PoS-Tags).....	63
8. Ergebnisse der korrekten Vorhersage von Silbengrenzen	64
9. Ergebnisse der Akzent-Prädiktion.....	64
10. Übersicht der Phonem-Visem-Klassifikation	67
11. Übersicht der quantitativen, phonologischen und linguistischen Merkmale zur Sprachsegmentbeschreibung	73
12. Übersicht der Module zur Vorbereitung der Sprachdaten, Videodaten und statische Modelle	114
13. Übersicht der Module, die zur Laufzeit zur Synthese verwendet werden.....	115

ANMERKUNGEN

Die vorliegende Arbeit entstand während meiner Anstellung bei Professor Wolfgang Hess an dessen Institut für Kommunikationsforschung und Phonetik der Universität Bonn. Dort habe ich die Arbeiten im Rahmen des DFG-Projektes „Automatische Trainingsverfahren für die Unit-Selection-basierte Sprachsynthese“ durchgeführt und daraus resultierend die vorliegende Arbeit erstellt. Herzlichen Dank für die Unterstützung und die Forschungsfreiheit, sowie die fachlichen Anregungen, welche mir Professor Hess während dieser Zeit zukommen ließ. Herzlichen Dank auch an Dr. Karlheinz Stöber, der mich in die Themenstellung eingeführt hat und mir den Start am IKP erleichterte. Dank gebührt auch meinen Kollegen am IKP, mit denen ich nicht nur fachliche Diskussionen führen konnte. Herzlichen Dank an Dr. Bernhard Schröder und Dr. Hans-Christian Schmitz für die abwechslungsreiche Zeit. Herzlichen Dank an meine Kollegen der Phonetik-Gruppe, Dietmar Lance, Stefan Breuer und Dr. Petra Wagner. Besonderer Dank gilt meiner Familie, die mich während des Studiums und während meiner Promotionszeit immer tatkräftig unterstützt hat, sowie Fadja Ehlail, die mir an fröhlichen und weniger fröhlichen Tagen immer mit Rat und Tat beiseite stand. Danke.

EINFÜHRUNG

1.1 Einführung und Motivation

In dieser Arbeit werden Methoden und Verfahren entwickelt, die es ermöglichen, ein Korpus-basiertes audio-visuelles Synthese-System zu erstellen, mit dem natürlich klingende Sprache und synchron zur Sprache eine video-realistische visuelle Ausgabe erzeugt werden kann, welche in einer Videosequenz resultiert. Das Video zeigt einen video-realistischen Talking-Head, der durch Konkatenation von 2D-Bildsequenzen erstellt wurde. Die resultierende Videosequenz ist vergleichbar mit der frontalen Ansicht eines TV-Nachrichtensprechers. Hierbei bestehen die Sprachausgabe und die visuelle Ausgabe aus unabhängig voneinander, automatisch generierten Audio- und Videosegmenten, welche in eine lippensynchrone Audio-Video-Sequenz münden. Abbildung 1 zeigt den schematischen Aufbau des in dieser Arbeit entwickelten Systems.

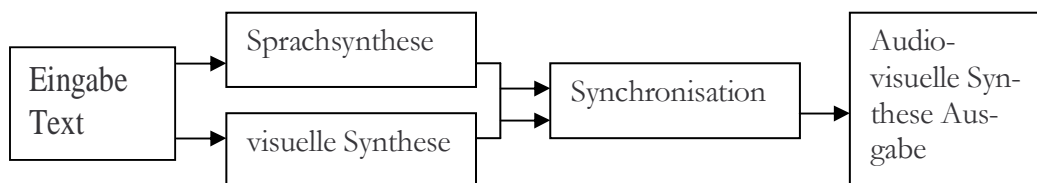


Abbildung 1: Schematischer Aufbau audio-visueller Synthese

Das System kann beliebigen Text audio-visuell wiedergeben, ohne dass dieser von dem jeweiligen Sprecher vorab geäußert wurde. Als Grundlage dienen ein aufgenommenes Sprachdatenkorpus sowie ein Videodatenkorpus, aus diesen werden dann die erwünschten Äußerungen rekonstruiert. Die eingesetzten Algorithmen für die Generierung der auditiven und visuellen Quellen basieren auf einem automatischen Training von Unit-Selection-basierter Sprachsynthese sowie der assoziierten visuellen Segmentauswahl. Die visuelle Segmentauswahl wird anhand einer Distanzmetrik gesteuert, welches es ermöglicht, die entsprechenden Segmente in ihrer zeitlichen Abfolge aus einer Vielzahl von möglichen

Segmenten zu identifizieren und das jeweils passende Segment für die Rekonstruktion zu verwenden. Es wurde auf statistische Lernverfahren aus dem Bereich der Musterklassifikation zurückgegriffen, die für die zu bewältigenden Aufgaben angepasst bzw. für die spezifischen Anforderungen der Sprach- und Bildverarbeitung hinsichtlich der benötigten Funktionen weiterentwickelt wurden. Bei der Sprachdatenverarbeitung zur Generierung der Sprachsynthese wurden statistische Modelle trainiert, die den Sprachproduktionsprozess in seiner zeitlichen Abfolge nachbilden. Auf Grundlage von bedingten Wahrscheinlichkeiten wurde ein endlicher Automat entwickelt, bei welchem die einzelnen Zustände jeweils ein Sprachsegment in variabler Ausprägung repräsentieren. Ausgehend von dem endlichen Automaten wurde ein Modell mit Hilfe von bedingter Entropie erstellt, welches eine statistische Aussage über die Güte des auszuwählenden Sprachsegments für die Unit-Selection-basierte Sprachsynthese gibt, in dem die Entropie über die Gesamtsequenz der potentiellen Sprachsegmente maximiert wird. Als Alternative zu dem auf bedingter Entropie basierenden Modell wurde auf Basis der bedingten Zufallsfelder, „Conditional Random Fields“ (CRF: Lafferty et al. 2001) Modellparameter trainiert, welche ebenfalls statistische Aussagen über die Güte der zu konkatenierenden Sprachbausteine treffen. Mittels graphenbasierter Suche auf den zugrundeliegenden Modellparametern werden die bestmöglichen Sprachbausteinsequenzen ermittelt, indem die Wahrscheinlichkeit über den Gesamtgraphen maximiert wird. Bei der Videosegmentauswahl wurde ein K-Nearest-Neighbor-Verfahren eingesetzt, welches das jeweilige 2D-Videosegment anhand der Minimierung der Merkmalsdistanz auswählt. Eine nähere Betrachtung der Ansätze wird im weiteren Verlauf dieses Abschnitts gegeben sowie detailliert in den jeweiligen Themenbereichen Kapitel 3, 5 und 6.

Die Motivation zu dieser Arbeit resultiert aus der Fragestellung der Mensch-Maschine-Kommunikation, die sich zu einem wichtigen Forschungsgebiet entwickelt hat, in dessen Zentrum die natürliche Interaktion zwischen Menschen und Computern steht. Natürliche Interaktion bedeutet vor allem Mensch-Maschine-Kommunikation auf der Basis von Sprache und visuellen Wahrnehmungen in Verbindung mit der sprachlichen Äußerung. Sprache wird hier gleichgesetzt mit einer akustischen Signalübertragung von Sender A zu Empfänger B mit gleichzeitiger Identifikation, Verifikation und Interpretation des übertragenen akustischen Signals durch den Empfänger. Das übertragene Signal enthält in die-

sem Fall Informationen, die vom Empfänger interpretiert und verarbeitet werden können. Aber nicht nur das Signal in Form eines Sprachsignals enthält Information, sondern auch die Mimik und Gestik, welche bewusst oder unbewusst während der Kommunikation stattfinden und übertragen werden. Diese Mimik in Form von Gesichtsbewegungen liefert dem Empfänger während eines Kommunikationsprozesses zusätzlich zu der sprachlichen Äußerung Informationen des Senders. Im Gegensatz zum Menschen, für den Kommunikation mittels Sprache die natürlichste Form der Verständigung ist, ist diese Kommunikationsform für Computer erstaunlich schwierig und weit davon entfernt gelöst zu sein. Wenn sich zwei Menschen in einer Face-to-Face-Situation unterhalten, ist es für den Empfänger ein Leichtes, die sprachliche und visuelle Information des Senders schnell zu erfassen und zu verarbeiten – vorausgesetzt Sender und Empfänger sprechen die gleiche Sprache und bedienen sich ähnlicher Mimik und Gestik aus dem gleichen Kulturkreis. Überträgt man die gleiche Situation auf eine Mensch-Maschine-Kommunikation, so lässt sich erkennen, dass es trotz aller wissenschaftlich-technischen Fortschritte sehr leicht zu einer Missverständigung zwischen Mensch und Maschine kommen kann. Dies betrifft die Erkennungsleistung der Maschine von Sprache und Mimik wie auch die Syntheseleistung, die als das korrekte und natürliche Sprechen von Computern anzusehen ist, und in Verbindung hierzu die simultane Generierung einer sprachsynchrone visuellen Ausgabe. Dennoch kann die Erkennungsleistung aber ebenso beim Menschen variieren und ist stark abhängig von einer klaren, gut verständlichen Sprachproduktion und einer klaren Mimik und Gestik. Spricht ein Mensch undeutlich oder zu schnell, hat selbst das menschliche Sprachverarbeitungssystem Schwierigkeiten, das Gesagte zu identifizieren bzw. richtig zu interpretieren. Es kommt also auf ein gutes Sprechen an. Die Eigenschaft des verständlichen, gut artikulierten und intonierten Sprechens ist Voraussetzung für eine effektive Kommunikation, die durch die menschliche Mimik und Gestik unterstützt wird. Untersuchungen (Massaro et al. 1998, Beskov 2002) haben gezeigt, dass eine multimodale Mensch-Maschine-Schnittstelle die Interaktion zwischen Mensch und Maschine verbessert, da das menschliche Wahrnehmungssystem bei einer sprachlichen Äußerung zusätzlich die mimischen Informationen wahrnimmt, auf die wir zurückgreifen und die uns so genannte Metainformationen zu der sprachlichen Äußerung liefern. Gestik und Mimik des Sprechers helfen während eines Kommunikationsaktes das Sprachsignal zu verarbeiten, welches inneren und äußeren Störungseinflüssen, wie Umweltgeräuschen, unterliegt.

Auch gesundheitliche Beeinträchtigungen beim Menschen, wie z.B. schlechte Hörleistung beeinträchtigen die Verständlichkeit des Audiosignals. Das Audiosignal kann dadurch vom Hörer falsch wahrgenommen und interpretiert werden. Durch die auditive und visuelle Kombination wird eine Erhöhung der Informationsübertragung erzielt. Diese Eigenschaft macht man sich auch bei der Spracherkennung zu eigen (Potamianos et al. 2004) und erhofft sich so in Umgebungen mit hohen äußeren Störgeräuschen, die Erkennungsleistung von Spracherkennungssystemen zu erhöhen, indem eine audio-visuelle Spracherkennung entwickelt wird, die die visuelle Erkennung in Form von Lippenlesen mit der akustischen Erkennung des Sprachsignals vereint.

Die Störung während eines Kommunikationsprozesses zwischen Mensch und Maschine kann aber auch aus der Anwendung selbst heraus entstehen. So ist bei Unit-Selection-basierten Text-to-Speech-Systemen immer ein LNRE-Phänomen (Möbius 2000), „Large Number of Rare Events“, zu beobachten, das die Synthesequalität eines TTS-Systems beeinträchtigt, da es unmöglich ist, alle in einer Sprache vorkommenden Charakteristika durch das entsprechende Sprachdatenkorpus abzudecken. Dazu zählt z.B. im Deutschen die Möglichkeit der Wortzusammensetzung, die es ermöglicht, immer neue Wörter zu generieren. Auch die Fehler der prosodischen Realisierungen, wie Dauersteuerung und F0 während der Synthese, führen zu Störungen im akustischen Signal und daher zu Verständnisproblemen.

Um Sprachsynthese-Systeme in ihrer Qualität zu verbessern, also ihre Verständlichkeit zu erhöhen, ihre Natürlichkeit der des Menschen anzupassen und das generierte Sprachsignal über längere Zeitabschnitte angenehm zum Zuhören zu gestalten, werden für die in dieser Arbeit entwickelte Sprachsynthese datenbasierte Verfahren eingesetzt. Hierzu wurde auf ein vorhandenes Sprachdatenkorpus zurückgegriffen (Stöber et al. 2000), bei dem eine semi-professionelle Sprecherin aufgenommen wurde. Aus diesem Sprachdatenkorpus werden dann in Form von akustischen Sprachsegmenten Zeitbereichseinheiten für die zu synthetisierenden Äußerungen extrahiert. Die ausgewählten Sprachsegmente werden anschließend konkateniert und als akustisches Sprachsignal wiedergegeben. Dieses Verfahren wird in der Literatur als Unit-Selection-basierte Sprachsynthese (Sagisaka 1988, Hunt & Black 1995, Black & Campbell 1996, Conkie et al. 1999, Stöber 2002) bezeichnet.

Segmentauswahl aus bestehenden Korpora und anschließende Konkatenation bilden ebenso die Grundlagen für die visuelle Synthese. Hierzu wurden Videodaten-Korpora aufgezeichnet, in denen eine Sprecherin in einer frontalen 2D Sitzposition zu sehen ist. Die zu synthetisierenden Äußerungen werden in Teilsegmenten aus diesem Videodatenkorpus entnommen und neu zusammengesetzt. Die Identifikation der visuellen Segmente basiert auf dem phonetisch-visemisch transkribierten Eingabetext (vgl. Kapitel 4, Abschnitt 4.4.1, 4.4.4) und der zeitlichen Entsprechung, die anhand eines Audio-Video-Synchronisation-Algorithmus berechnet wird (vgl. Kapitel 6, Abschnitt 6.4).

Konkatenative Verfahren basieren meist auf zwei-dimensionalen Kostenfunktionen (Campbell & Black 1995, Hunt & Black 1996, Donovan 1999, Stöber 2002), die Abstandsmaße herstellen, um zu berechnen, wie gut einzelne Segmente für die Konkatenation zueinander passen. Die Kostenfunktionen werden in einem zweistufigen Prozess berechnet. Im spezifischen Fall der Unit-Selection-basierten Sprachsynthese richten sich die Abstandsmaße nach spektralen, prosodischen, linguistischen und quantitativen Merkmalen (Hunt & Black 1996). Dieser kostenbasierte Ansatz wird in dieser Arbeit durch eine auf statistischen Lernverfahren beruhende Segmentauswahl ersetzt.

Bei der visuellen Synthese kommt eine metrische Distanzberechnung zum Zuge, die anhand von Minimierung des Abstandes zwischen den Merkmalsvektoren die geeignetsten Segmente auswählt. Die Merkmale, die die Segmente beschreiben, welche während der visuellen Verkettung verwendet werden, beinhalten vor allem Informationen, die mittels Bildverarbeitungsalgorithmen gewonnen werden. Diese Merkmale dienen der Lokalisierung von Kopf- bzw. Gesichtspartien, die sich für die Abstandsmaßberechnung als sinnvoll erwiesen haben, sowie auf pixelbasierten Graustufenwerten.

Mittels der vorgestellten Verfahren kann eine audio-visuelle Synthese generiert werden, welche auf beliebige Sprecher trainierbar und adaptierbar ist, wobei man unabhängig in der Wahl der Eingabe- und Ausgabequelle ist, was impliziert, dass mit ein- und derselben Stimme unterschiedliche visuelle Ausgaben unterlegt werden können sowie eine visuelle Ausgabe für verschiedene Stimmen verwendet werden kann. Die Methoden und Verfahren sind für die datenbasierte Generierung entwickelt und setzen daher nur das Trainings-

daten-Material eines Sprechers bzw. einer Sprecherin voraus. Alle weiteren Schritte werden automatisch vollzogen.

Die Ansätze für die Sprachsynthese, wie auch für die visuelle Synthese, werden durch statistisch motivierte automatische Trainingsverfahren realisiert, um eine möglichst sprachenunabhängige Umsetzung von Text zu audio-visueller Ausgabe zu gewährleisten. Die Entscheidung für die gewählten Lernverfahren resultiert aus den Ergebnissen verwandter Bereiche der Sprachverarbeitung. Der Einsatz von Hidden-Markov-Modellen (Rabiner 1986) ist in der Spracherkennung seit vielen Jahren sehr erfolgreich implementiert. Ebenso wurden Hidden-Markov-Modelle erfolgreich in der Sprachsynthese eingesetzt (Donovan 1996, Acero et al. 1997, Tokuda et al. 2000, 2002). Bei der Betrachtung der Problemstellung, geeignete Sprachsegmente für die Sprachsynthese aus einem großen Sprachdatenkörper auszuwählen, kann der Ansatz der Spracherkennung in umgekehrter Richtung auf diese angewendet werden. Die Sprachproduktion wird als zeitlicher Prozess modelliert wobei, bei dem die Sprachsegmente mittels eines Graphen repräsentiert werden. Die Knoten des Graphen spiegeln jeweils die zur Auswahl stehenden Sprachbausteine für die zu synthetisierende Äußerung wider; die sie verbindenden Kanten sind durch Übergangswahrscheinlichkeiten gewichtet. Daraus ergibt sich folglich zunächst ein probabilistischer endlicher Automat (engl. Probabilistic Finite State Machine). Der optimale Pfad durch den Graphen, der die Zeitbereichseinheiten repräsentiert, wird durch ein Suchverfahren ermittelt, um letztlich das bestmögliche Sprachsegment zu identifizieren, welches die benötigten Charakteristika vereint.

Nicht alle sprachcharakteristischen Merkmale sind von Beginn an verfügbar; sie werden erst im Laufe der Modellerstellung für das zugrunde liegende Modell dynamisch hinzugefügt. Dies resultiert letztlich in kontextabhängigen Modellen, welche die unterschiedlichen Segmentebenen mit einbeziehen und so eine „non-uniform variable-size unit-selection“ zulassen, was nichts anderes heißt, als dass zu jedem Zeitpunkt des Sprachproduktionsprozesses das bestgeeignete Sprachsegment abhängig von den zuvor realisierten Sprachsegmenten ausgewählt wird. Der Begriff „variable size“ zielt darauf ab, dass die Segmente keine einheitliche Größe haben und „non-uniform“ steht für die unterschiedlichen akustischen und spektralen Eigenschaften der Sprachsegmente. Dies lässt sich leicht

nachvollziehen, wenn man den menschlichen Sprachproduktionsprozess genauer betrachtet, bei dem auf z.B. auf Wortebene jedes Wort abhängig von der zeitlichen Realisierung andere Eigenschaften hat als das gleiche Wort, welches an anderer Stelle in anderem Kontext realisiert wird.

Als visuelle Erweiterung der Sprachausgabe werden vor allem so genannte „Talking Heads“ eingesetzt. Die Basis der visuellen Ausgabe des hier vorgestellten Systems bilden photorealistische Sequenzen, die als 2D-Bilddaten aus einem Videokorpus extrahiert wurden. Die Videosegmente werden in Form von Frame-Sequenzen zur Laufzeit extrahiert. Zur Vermeidung von Störungen an den Verkettungsstellen werden Abstandmaße bzw. statistische Verfahren verwendet, die geeignete Frame-Sequenzen auswählen. Die zeitliche Dimension spielt ebenso wie bei der Sprachproduktion auch bei der Generierung der visuellen Ausgabe eine entscheidende Rolle. In der zeitlichen Abfolge nimmt der Mensch beim Sprechen verschiedene Kopfpositionen in einem Raum ein. Aus diesem Grund muss auch bei der visuellen Realisierung die Auswahl im Kontext betrachtet werden, welches Segment sich nun am besten eignet. Für diese Auswahl wurde auf den bekannten K-Nearest-Neighbor-Algorithmus zurückgegriffen. K-Nearest-Neighbor-Methoden wurden in unterschiedlichen Bereichen der Klassifikation erfolgreich eingesetzt. Sie nutzen die im Trainingsdatensatz vorkommenden Beobachtungen und können so die geeigneten Segmente, welche denen im Modell am nächsten kommen, effizient auswählen.

Als Ergebnis dieser Verfahren wurde die Software AVISS implementiert, welche in der Lage ist, die zuvor beschriebene Aufgabe der audio-visuellen Generierung auf Basis von Texteingaben zu erfüllen und eine natürliche, verständliche audio-visuelle Ausgabe als Videosequenz zu produzieren, bei der der synthetisierte Text synchron zur visuellen Ausgabe ist. Eine Anwendung, die mit Hilfe unserer AVISS Software erstellt wurde, ist das audio-visuelle Wetterinformationssystem, welches vollautomatisch Wetterinformationen in Form von Text audio-visuell synthetisiert und auf einer Internetseite wiedergibt. Diese Art der Anwendung kann als „limited domain lifelike interactive system“ oder „lifelike conversational system“ (Ostermann 1998, Cosatto et al. 2002) gekennzeichnet werden.

1.2 Einordnung der Arbeit in den Forschungsbereich

Mit dieser Arbeit soll ein Beitrag zur Forschung im Bereich der Sprachproduktion, im Besonderen zur konkatenativen Sprachsynthese, und als Erweiterung hierzu, zur audio-visuellen Sprachsynthese geleistet werden. Konkatenative Sprachsynthese ist seit langer Zeit Gegenstand der Forschung und in letzter Zeit wieder sehr in den Mittelpunkt industrieller Anwendungen und universitärer Forschung gerückt. Ebenso ist die audio-visuelle Synthese, die vor allem als „Talking Heads“ in unterschiedlichen Anwendungen zu finden ist, Gegenstand industrieller und universitärer Forschung. Zahlreiche Workshops spiegeln die aktive Forschungslandschaft in diesen Bereichen der audio-visuellen Sprachsynthese wider. Es ist zu erkennen, dass Sprachsynthese und audio-visuelle Synthese sich mehr und mehr in ihren Ansätzen annähern. Nachfolgend werden die Verfahren dieser Arbeit zur adaptiven audio-visuellen Synthese in die aktuelle Forschung eingereiht, und eine Abgrenzung zu den bestehenden Arbeiten wird vorgenommen. Es wird angeführt, wie mit den in dieser Arbeit vorgestellten alternativen Ansätzen bestehende Verfahren vereinfacht bzw. verbessert werden können.

Einen Paradigmenwechsel in der Sprachsynthese stellt der Übergang von den parametrischen Sprachsyntheseverfahren hin zu der nicht-parametrischen Sprachsynthese durch Verkettung von akustischen Zeitbereichseinheiten dar. Die Verkettung von Zeitbereichseinheiten wurde mit dem PSOLA-Algorithmus (Pitch Synchronous Overlap Add, Hamon et al. 1989) realisiert. Mit Hilfe von PSOLA lässt sich das Sprachsegment direkt in seinen prosodischen Eigenschaften manipulieren ohne das Sprachsignal in eine parametrische Darstellung zu überführen. Artikulatorische und Formant-Synthese rückten in den Hintergrund. Konkatenative Sprachsynthese wurde zum State-of-the-Art als Sprachsynthese-Ansatz. Die verwendeten Einheiten für die konkatenative Sprachsynthese waren zumeist Diphone oder vergleichbare Einheiten (Portele 1996). Dies ermöglichte eine geeignete Verkettung der Sprachbausteine und eine kompakte Speicherungsmöglichkeit von wenigen Megabyte, was den zu dieser Zeit entsprechenden Ressourcen hinsichtlich RAM und Speicherplatz Rechnung getragen hat. Der Nachteil dieses Verfahrens ist die benötigte Signalmanipulation, um prosodische Zielvorgaben zu realisieren. Veränderungen der Grundfrequenz mit einem Wert höher oder niedriger als eine halbe Oktave führen zu Qua-

litätseinbußen. Ebenso ist eine spektrale Glättung an den Konkatenationsstellen nicht möglich. Diese Signalstörungen mindern die Qualität des synthetisierten Sprachsignals.

Von Sagisaka (1988), Black & Campbell (1995) und Hunt & Black (1996) wurde ein Wechsel hin zu einem datenbasierten Verfahren mit großen Sprachdaten-Korpora entwickelt, welches als „non-uniform unit-selection“ bezeichnet wird. Das Verfahren nutzt oft mehr als eine Stunde gesprochener Sprache und eine variable Einheitengröße. Die Sprachdaten-Korpora, aus denen die Sprachsegmentbausteine ausgewählt werden, können auf mehrere hundert Megabyte anwachsen, denn die Größe der Segmente, welche für die Konkatenation verwendet werden, kann von ganzen Phrasen (IBM Phrase-Splicing) bis hin zu Halbphonen (AT&T) und in einigen Fällen (Donovan et al. 2001) bis zu Drittelphonen reichen. Dies verringert die Notwendigkeit einer Signalmanipulation durch eine große Anzahl prosodischer Variationen gleicher Einheiten. Das abgeleitete Verfahren wird dann als „non-uniform variable-size unit-selection“ bezeichnet. Neuere Ansätze nehmen noch größere Sprachdaten-Korpora auf, die einen Umfang von bis zu fünfhundert Stunden (N. Campbell, pers. Anmerkung) natürliche Sprache umfassen.

Die Auswahl eines geeigneten Sprachsegments bei Unit-Selection-basierter konkatenativer Sprachsynthese kann als Suchprozess angesehen werden, bei dem jeweils abhängig vom Kontext der passende Sprachbaustein im Suchraum identifiziert werden muss. Bisherige Arbeiten zur Bewältigung der Segmentauswahl basieren auf den Arbeiten von (Black & Campbell 1995; Hunt & Black 1996; Stylianou et al. [AT&T NextGen System] 2000; Klabbers et al. 2001, Donovan, Eide et al. 2001, Stöber et al. 2001). Bei diesen Arbeiten wurde jeweils eine zweidimensionale Kostenfunktion für die Segmentauswahl verwendet. Die Kostenfunktion besteht aus den Einheitenkosten und den Übergangskosten des jeweiligen Sprachsegments und die Minimierung der Kosten bestimmt das geeignete Zeitbereichssegment. In einem ersten Schritt werden die Einheitenkosten berechnet und liefern eine gewichtete Summe der prosodischen, quantitativen und linguistischen Merkmale. Blouin (2001) berechnet die Einheitenkosten als gewichtete Summe unterschiedlicher Unter-Einheitenkosten. Um an den Konkatenationsstellen der Segmente Störungen zu minimieren, werden die Übergangskosten vor allem durch Abstandsmaßberechnung der spektralen Eigenschaften definiert. Klabbers und Veldhuis haben in ihren Arbeiten (Klabbers et

al. 1998) verschiedene spektrale Abstandsmaße untersucht. Hierunter sind Euklidischer Abstand zwischen den Formanten F1 und F2, Euklidischer Abstand der MFCC, das Mittelwert-quadrierte Log-Spektrum sowie Abstand der Amplituden. Als bestes Abstandsmaß wurde der symmetrische Kullback-Leibler-Abstand zwischen der Energie normalisierten LPC Spektra bewertet. Die Ergebnisse stützen sich auf Perzeptionstests, bei denen Zuhörern Stimuli vorgespielt wurden. Der Euklidische Abstand zwischen Mel-skalierten LPC-basierten Cepstral Parametern wurde in einer ähnlichen Untersuchung von Wouters und Macon (Wouters et al. 2000) verwendet. Für die in der vorliegenden Arbeit entwickelten Verfahren (siehe Kapitel 5, Abschnitt 5.5) werden als spektrale Abstandsmaße der Euklidische und der statistisch motivierte Mahalanobis-Abstand zwischen den Mel-Cepstrum-Koeffizienten des letzten Zeitfensters des vorhergehenden Sprachsegments und den Mel-Cepstrum-Koeffizienten des erste Zeitfenster des folgenden Sprachsegments berechnet. Je geringer die Distanz, desto besser - so wird angenommen - passen die Sprachsegmente in ihren spektralen Eigenschaften zusammen. Dies geht auf die Arbeiten von Stylianou und Syrdal (Stylianou et al. 2000) und Donovan (2001) zurück. Stylianou und Syrdal stellten fest, dass das von Klabbers und Veldhuis präferierte Abstandsmaß in ihren Untersuchungen am schlechtesten abschnitt.

Eine theoretische Betrachtung zwischen Spracherkennung und Sprachsynthese wird von Ostendorf und Bulyko (Ostendorf et al. 2004) sowie Eichner und Hoffmann (Eichner et al. 2000) vorgestellt. In ihren Arbeiten stellen sie die Herangehensweise bzw. die Algorithmen der Spracherkennung heraus und stellen die Frage, welche dieser Techniken für datenbasierte Sprachsynthese relevant sind und welche bereits eingesetzt werden. So hat sich aus der Spracherkennung die dynamische Programmierung in Form des Viterbi-Algorithmus für die Sprachsynthese als geeignet erwiesen, um die potentiellen Sprachsegmente in einer schnellen Segmentsuche zu identifizieren. Ebenso kommen die Mel-Cepstrum-Koeffizienten eines Sprachsignals ursprünglich aus der Spracherkennung. Bulyko verfolgt darüber hinaus den Ansatz eines gewichteten Finite State Transducers (WFST) (Bulyko 2002), um den Übergang zwischen den Segmenten zu modellieren. Dieser Ansatz sowie der Ansatz von Tokuda et al. (2002) kommen den in dieser Arbeit entwickelten Algorithmen zur Sprachsegmentauswahl am nächsten.

Tokuda et al. (2000, 2002) stellen einen Ansatz vor, der auf kontextabhängigen Hidden-Markov-Modellen der spektralen Parameter eines Lautes beruht und zur Generierung des akustischen Sprachsignals eine Approximation des logarithmierten Mel-Spektrums vornimmt. Mittels der Mel-Log-Spektrum-Approximation (MLSA, Imai, 1983) wird die lautsprachliche Äußerung letztlich erzeugt. Anders als in den genannten Arbeiten zur non-uniform Unit-Selection (Black & Campbell 1995; Hunt & Black 1996; Stylianou et al. [AT&T NextGen System] 2000; Klabbers et al. 2001, Donovan, Eide et al. 2001, Stöber et al. 2001) werden bei dem HMM-basierenden Ansatz von Tokuda et al. die Sprachsegmente parametrisiert und prosodische Merkmale wie F0 und Lautdauer jeweils mit Hilfe eines Entscheidungsbaumes modelliert. Das Sprachsignal wird aus den konkatenierten HMMs durch den Einsatz des MLSA-Filters (Imai, 1983) generiert. Das von Tokuda et al. entwickelte Sprachsynthesensystem ist aufgeteilt in ein Trainingsmodul, welches die jeweils verwendeten Sprachdaten für die HMM-Synthese aufbereitet und die kontextabhängigen Modelle trainiert, und in ein Laufzeitmodul, welches die Texteingaben in akustische Sprachsignale umwandelt. Für das Training werden spektrale Parameter und Anregung aus dem Sprachsignal extrahiert und kontextabhängige HMMs aus Lauteinheiten erstellt. Zur Laufzeit werden diese wiederum als kontextabhängige HMMs verkettet. Aus diesen werden dann die Parameter für das Sprachsignal generiert, und mittels einer Anregungsfunktion und Filter wird das Sprachsignal synthetisiert. Dieser Ansatz wurde für das Deutsche angepasst, und die sprachabhängigen Module wurden hierfür entwickelt. Siehe hierzu Kapitel 5.1.

Die audio-visuelle Synthese, im Besonderen nachfolgend die Generierung der visuellen Ausgabesequenz, basiert auf 2D-Videobilddaten, die konkateniert werden, um ein Videosignal zu erzeugen. In der aktuellen Forschung haben sich zwei unterschiedliche Richtungen mit jeweils unterschiedlichen Verfahren herauskristallisiert. Die Verknüpfung von Sprachsynthese mit einer visuellen Ausgabe, welches in sog. Talking Heads resultiert, hat sich in einen modellbasierten Ansatz (Bailly 2004, Ostermann et al. 2001, Beskov 2003) und in einen datenbasierten Ansatz, der auf Videodaten (Bregler et al. 1997, Cosatto et al. 2000, Ezzat 1998, 2002) zurückgreift, aufgeteilt.

Eine Vielzahl der modellbasierten „Talking-Heads“ wird mit Hilfe von 3D-Modellen generiert, um den Kopf und die Gesichtsmerkmale zu erfassen, die Parameter zu verarbeiten und letztlich den Kopf und das Gesicht neu zu modellieren. Die erforderlichen Daten werden mittels eines Laser-Scan-Verfahrens gewonnen. Als Ergebnis erhält man Datenpunkte im Raum, bei denen die Knoten miteinander verbunden werden; hieraus entsteht ein polygonales Gitternetzabbild, ein so genanntes „Mesh“ des abgetasteten Kopfes entsteht. Die Flächen zwischen den Knoten und innerhalb der Kanten werden entsprechend farblich realisiert. Während der Animation werden die Knoten anhand vorgegebener Parameter verschoben. Arbeiten hierzu wurden vor allem von Parke (1996), Cohen, Massaro (2002) und Beskov (2003) durchgeführt. Mit der Einführung des MPEG4 Standards¹ wurden so genannte FAP (Facial Animation Parameters) bereitgestellt, die entsprechende Gesichtsparemeter definieren. Es werden High-Level-Parameter und Low-Level-Parameter unterschieden. Zu Ersteren gehören Viseme und Ausdruck, die als Emotionsfunktionen dienen. Zu den Low-Level-Parametern zählen u. a. der Grad der Mundöffnung und die Zungenstellung. Insgesamt umfasst der Parametersatz 68 Einstellungen. Implementierungen der FAP wurden von Ostermann et al. (1998) erfolgreich vorgenommen.

Bei der datenbasierten visuellen Synthese, die wie in der konkatentativen Sprachsynthese Segmentbausteine nutzt, um neue Äußerungen zu synthetisieren, werden aus bestehenden Video-Korpora der jeweiligen Sprecher visuelle Repräsentationen einer Äußerung in Form von 2D-Bilddaten extrahiert und zu einer neuen Sequenz zusammengesetzt. Ein datenbasierter Ansatz wurde von Bregler (1997) in seinem Video-Rewrite-System erfolgreich demonstriert. Bregler extrahierte die Lippenbewegungen, die während einer Äußerung erfolgen, und integrierte diese in einen bestehenden Kopf an die Stelle der ursprünglichen Lippen (vgl. Abschnitt 1.5.2). Cossatto et al. (2000) haben aus ihrem visuellen Korpus nicht nur Lippenbewegungen ausgeschnitten, sondern bedienen sich zusätzlich des Augenbereichs und des Gesichtsbereichs ohne Nase, Mund und Augen. Diese Segmente werden in einer Datenbank gespeichert und bei Bedarf angefordert und miteinander kombiniert (vgl. Abschnitt 1.5.2).

¹ <http://www.m4if.org/>

Ebenfalls auf photorealistischen Sequenzen wurde von Beier (Beier et al. 1992) ein erster Morphing-Ansatz basierend auf Pixelinterpolation präsentiert. Zwischen den Key-Frames werden die entsprechenden Pixel für den zu verändernden Bereich gleichzeitig überblendet und verschoben. Ezzat (1998) hat dieses Verfahren aufgegriffen und in seinem Mike-Talk-System realisiert. Neuere Ansätze von Ezzat (2002) basieren ebenso auf Verschieben von Pixel-Regionen mittels des Optical-Flow-Algorithmus. Dieses Verfahren kann automatisch trainiert werden.

In Abgrenzung zu den genannten Arbeiten von Bregler (1997) Cosatto (2000) und Ezzat (1997, 2002), werden in der vorliegenden Arbeit Verfahren aus der Domäne des statistischen Lernens als Grundlage der auf „variable-size non-uniform unit-selection“ basierten Sprachsynthese herangezogen. Ausgangspunkt der visuellen Synthese sind jeweils Segmentbausteine variabler Größe, die aus dem zugrunde liegenden Videodatenkorpus als Ganzes extrahiert werden und zu einem Videosignal konkateniert werden. Wie in der konkatenativen Unit-Selection-basierten Sprachsynthese wird für die visuelle Synthese keinerlei Signalmanipulation an den Videodaten vorgenommen. Dies hat den Vorteil, dass natürliche Gesichtsmimik, wie Augenzwinkern, in der synthetisierten Ausgabe erhalten bleibt und somit der Gesamteindruck natürlicher erscheint.

1.3 Übersicht über die Arbeit

Die vorliegende Arbeit ist unterteilt in 8 Kapitel. In Kapitel 2 werden die Grundlagen der Sprachproduktion beim Menschen sowie die technische Entwicklung von Sprachsynthese in einem Überblick dargestellt. Die als Grundlage dieser Arbeit verwendeten konkatenativen Sprachsyntheseverfahren, Diphonsynthese und Unit-Selection-basierte Sprachsynthese, werden in Abschnitt 2.1.2 und 2.1.3 dargestellt. Es werden deren grundlegende Funktionsweisen aufgezeigt und ihre verschiedenen Realisierungen dargestellt. Die Grundlagen der audio-visuellen Synthese werden ebenso in Kapitel 2 beschrieben. Hier werden im Hinblick auf die vorliegende Arbeit vor allem die Grundlagen der nonverbalen Kommunikation erläutert und die zwei Hauptansätze der audio-visuellen Synthese beschrieben. Diese zwei Ansätze gliedern sich in einen modellbasierten und in einen datenbasierten Ansatz, welche Gegenstand in Kapitel 2, Abschnitt 2.2.2, 2.2.3 sind.

In Kapitel 3 werden die probabilistischen Grundlagen erläutert, die in dieser Arbeit zur Anwendung kommen. Es wird ein Überblick über bedingte Wahrscheinlichkeit, das Bayes-Theorem, Markov-Modelle, Hidden-Markov-Modelle, Entropie und bedingte Entropie gegeben.

In Kapitel 4 werden die verwendeten Korpora definiert und ihre Eigenschaften dargestellt. Für die Verarbeitung wurden die Korpora aufbereitet und mittels der „Extensionable Markup Language“ XML gespeichert. Es wird eine detaillierte Darstellung der symbolischen Daten-Vorverarbeitung, Segmentierung, Annotation und Speicherung der Korpora dargelegt. Bei der Vorverarbeitung der Sprach- und Video-Korpora werden automatische Trainingsverfahren zur Merkmalsgewinnung eingesetzt. Es wird die automatische Vorhersage der Wortklasse (Engl: Part-of-Speech, POS) beschrieben, sowie die automatische Umsetzung von orthografischem Text in die entsprechende phonetische Lautschrift, desweiteren die Umsetzung von orthografischem Text in die jeweilige visemische Entsprechung. Zur Klassifizierung von Graphemen in Viseme wurde ein visemisches Symbolinventar entwickelt, das als Äquivalent zu den Phonemklassen angesehen werden kann. Ausgehend von den visuell unterscheidbaren Phonemen wurden Visemklassen hergeleitet. Für die Klassifizierungsaufgaben wurde für jeweils ein Maximum-Entropie-Modell trainiert, welches in der Lage ist, mittels überwachten Lernens das klassische Mustererkennungsproblem zu lösen.

Kapitel 5 spiegelt die entwickelten Verfahren für die Unit-Selection-basierte Sprachsynthese wider. Bei der Auswahl von geeigneten Sprachsegmenten für die konkatenative Sprachsynthese kommen unterschiedliche Faktoren zum Tragen. So sind die linguistischen, quantitativen Merkmale wie auch die prosodischen Merkmale Dauer und Grundfrequenz entscheidende Faktoren für die Qualität einer synthetisierten Äußerung. In den Abschnitten 5.1 und 5.2 werden diese Merkmale und die dynamische Merkmalsgewinnung beschrieben. Hier werden vor allem die linguistischen und akustischen Eigenschaften der Sprachsegmente herausgestellt. In Abschnitt 5.2.1 wird die Dimensionsreduktion der spektralen Merkmalsvektoren mittels Karhunen-Loeve-Transformation dargestellt. In Abschnitt 5.3 wird das HMM-basierte Sprachsynthese-Verfahren erläutert. In Abschnitt 5.4 bis 5.7 werden ein statistisch motivierter Selektionsalgorithmus sowie die zugehörigen

Basisbedingungen und Suchalgorithmen entwickelt. Hier bilden die probabilistischen Verfahren „probabilistischer endlicher Automat“, „bedingte Entropie“ und „Conditional-Random-Fields“ die Grundlage.

In Kapitel 6 wird die Erweiterung der Sprachsynthese durch eine datenbasierte video-realistische Synthese erläutert. Ausgangspunkt bilden die artikulatorischen Vorgänge des Sprechens und deren visuelle Repräsentation. Wie in der symbolischen Vorverarbeitung eines TTS-Systems ist die Auswahl der 2D-Bildsequenzen abhängig von der zugrunde liegenden Transkription des Eingabetextes. Hierzu wird die in Kapitel 4 erarbeitete Phonen-Visem-Klassifikation verwendet. Abschnitt 6.2 dient der Beschreibung des verwendeten KNN-Algorithmus. In Abschnitt 6.3 wird die Parametergewinnung, die für die Auswahl der visuellen Segmente verwendet wird, erläutert. Neben der Auswahl der sprachlichen Bausteine und der hierzu analogen visuellen Segmente ist die Synchronisation beider Quellen essentiell und wird in Abschnitt 6.4 wiedergegeben.

Kapitel 7 und 8 dienen der Darstellung der implementierten Software und der aus der Anwendung resultierenden Ergebnisse. Die Software, welche auf Basis der zuvor erarbeiteten Algorithmen und Verfahren erstellt wurde, ist ein vollständiges System zur Erzeugung von video-realistischer audio-visueller Synthese und wurde mittels objektorientierter Programmierung erstellt. Die Architektur wird in Abschnitt 7.1 vorgestellt. Funktionsweisen und Schnittstellen der einzelnen Module werden in Abschnitt 7.2 dargelegt. Kapitel 8 widmet sich der Evaluation der synthetisierten Äußerungen und zeigt die Ergebnisse eines Präferenz-Testes sowie die Evaluation anhand eines MOS-Tests.

In Kapitel 9 schließt sich eine Diskussion über die eingesetzten Verfahren an. Es werden die Fragestellungen betrachtet, die sich während dieser Arbeit ergaben, sowie in einem Ausblick weitere Forschungsarbeiten, welche die Weiterentwicklung auf diesem Gebiet mit sich bringen können, erläutert.

GRUNDLAGEN DER SPRACHPRODUKTION, SPRACHSYNTHESE UND AUDIO-VISUELLEN SYNTHESE

Das nachfolgende Kapitel gibt einen Überblick über die Grundlagen der Sprachproduktion und der Sprachsynthese. Diese Übersicht behandelt die Grundlagen, wie sie für das Verständnis der audio-visuellen Sprachsynthese benötigt werden. Ebenso wird eine kurze Übersicht über bisherige Forschungsarbeiten auf dem Gebiet der audio-visuellen Synthese angeführt.

2.1 Grundlagen

Unter Sprachsynthese versteht man im Allgemeinen eine Anwendung, die anhand maschinell generierter akustischer Signale, oder durch Verkettung von Sprachbausteinen eine sprachliche Äußerung liefert, wobei die gewünschte sprachliche Äußerung zuvor als Text eingegeben werden kann. Eine Spezifizierung von Sprachsynthese ist das als Text-to-Speech (TTS) bekannte Verfahren, einen orthografischen Text, der meist in ASCII-Form zugrunde liegt, als gesprochene Sprache wiederzugeben. In dieser Arbeit wird ausschließlich auf diese Art der maschinellen Sprachgenerierung eingegangen. TTS-Systeme kommen in unterschiedlichsten Anwendungen vor. So werden sie als Vorleseautomaten für sehbeeinträchtigte Personen genutzt oder sie finden sich in Call-Center-Applikationen wieder, um Personal für die Vermittlung zu sparen, wobei der Anrufer durch ein IVR (interactive voice response) in Verbindung mit einem TTS-System an die entsprechenden Stellen weitergeleitet wird. Eine ebenso weit verbreitete Anwendung von TTS ist der Einsatz in Umgebungen, in denen der Nutzer seine Aufmerksamkeit nicht auf geschriebene Textinformationen richten kann, z.B. während des Autofahrens. In Abschnitt 2.1.2 wird auf die ausgehend von der Sprachproduktion beim Menschen entwickelten Sprachsynthese-Verfahren in einem Überblick eingegangen. Eine genauere Betrachtung ist im Rahmen dieser Arbeit aus Platzgründen nicht möglich. Für eine detaillierte Darstellung der jeweiligen Sprachsynthese-Verfahren wird auf die einschlägige Literatur verwiesen.

Sprachproduktion beim Menschen ist ein komplexer physischer und kognitiver Vorgang, bei dem viele Einflussfaktoren eine Rolle spielen. Der Mensch hat die Fähigkeit, sprachliche Äußerungen in unterschiedlichsten Formen zu produzieren. Er ist in der Lage, frei zu sprechen, einen Schrifttext laut vorzulesen sowie unterschiedlichste Stimmcharakteristika und Emotionen mit seiner gesprochenen Äußerung auszudrücken. Daraus lässt sich leicht erkennen, dass Sprache ein komplexer kognitiver Prozess ist, der physisch durch die an der Sprachproduktion beteiligten Organe realisiert wird. Die Physiologie der menschlichen Sprachproduktion wird in diesem Kapitel im nachfolgenden Abschnitt 2.1.1, Physiologie, Sprachproduktion und Prosodie, näher betrachtet.

2.1.1 Physiologie, Sprachproduktion und Prosodie

Abbildung 2 zeigt den schematischen Aufbau des menschlichen Sprechapparats. In dieser Arbeit können die Organe, die notwendig sind für die Spracherzeugung bei Menschen, nur in einem Überblick dargestellt werden. Eine detaillierte Übersicht und Erklärung findet sich u. A. in Pompino-Marschall (1996) und Hess (2005).

Das Erzeugen von Schwingungen und Geräuschen, die dann als akustisches Signal abgestrahlt werden, wird allgemein als Anregung bezeichnet (vgl. Pompino-Marschall 1996). Durch Atmung strömt Luft durch Bronchien und Thorax in die oberen Atemwege. Nach Pompino-Marschall (1996) unterscheidet man drei grundlegende Funktionen der Sprachproduktion: Atmung, Phonation sowie Artikulation. Diese werden auch als Initiator, Generator und Modifikator bezeichnet. Für stimmhafte Laute gestaltet sich der Prozess der Klangerzeugung wie folgt: durch die Atmung wird ein subglottaler Luftdruck erzeugt, welcher mittels der Stimmlippen im Kehlkopf in näherungsweise als periodisch anzunehmende Schwingungen versetzt wird. Diese sind in Periode und Amplitude veränderlich und dienen somit als Anregung für verschiedene stimmhafte Laute, wie Sprache oder Schreien. Dieser Vorgang wird als Phonation bezeichnet.

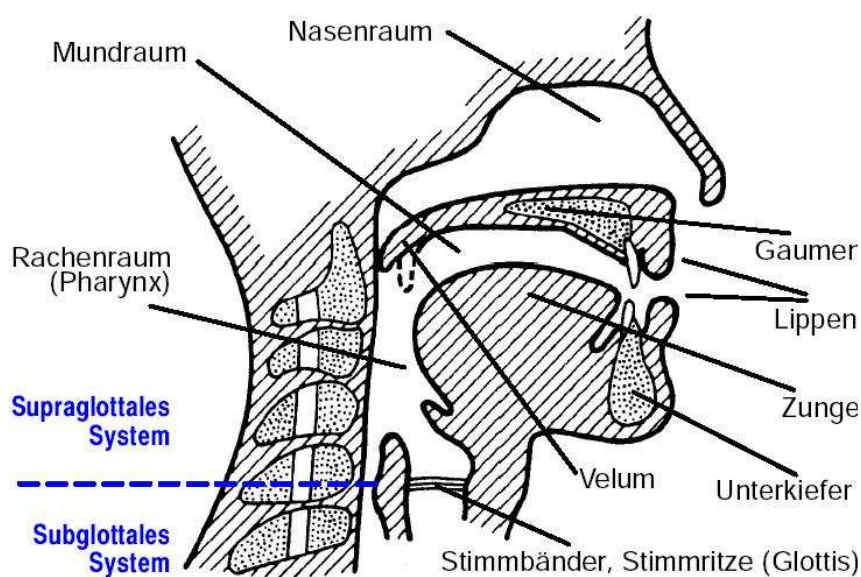


Abbildung 2: Schematischer Querschnitt des menschlichen Sprechtraktes, nach Hess (2005)

Während des Vorgangs der Phonation sind die Stimmbänder leicht gespannt und die Glottis ist geschlossen (vgl.: Abbildung 2). Durch den subglottalen Druck werden die Stimmbänder geöffnet und Luft kann durch die Glottis strömen. Die Öffnung der Glottis ist im Verhältnis zum Larynx kleiner, wodurch die Luft mit hoher Geschwindigkeit durch die Glottis hindurch strömt. Physikalisch wird im Bereich der Stimmbänder ein Unterdruck generiert, der quer zur Strömungsrichtung wirkt (Bernoulli-Kraft). Durch diesen Vorgang wird die Rückstellkraft verstärkt und die Stimmbänder werden bei Wegfall des subglottalen Drucks wieder in die Ausgangsstellung gebracht, da die Bernoulli-, sowie Rückstellkraft zusammen den subglottalen Druck übersteigen und die Stimmbänder aufeinander gepresst werden. Durch kurzzeitiges Schließen der Glottis fällt die Bernoulli-Kraft weg und Stimmbänder nehmen die Ausgangsstellung wieder ein. Der Schwingungszyklus kann von vorne beginnen. Der Schwingungszyklus ist für die Sprachverarbeitung ein wichtiger Prozess, da anhand dieses Schwingungszyklus das Maß für die Sprachgrundfrequenz F_0 extrahiert werden kann. Hier spielt vor allem das abrupte Schließen der Glottis eine wichtige Rolle, da hierdurch eine „Ecke“ im Phonationssignal entsteht, welche es ermöglicht, dass im Sprachsignal Frequenzen bis zu mehreren kHz auftreten können.

Bei der menschlichen Sprachproduktion existieren neben der stimmhaften Anregung noch die stimmlose Anregung und die transiente Anregung. Die stimmlose Anregung dient zur

Generierung von Reibe- oder Friktionsgeräuschen. Hierbei durchströmt Luft die geöffnete Glottis und passiert eine Engstelle im Mundraum oder Rachenraum. An dieser Stelle entsteht durch Reibung eine turbulente Strömung und somit ein Rauschen, dessen Spektrum von der Lage der Engstelle bestimmt wird. Bei der transienten Anregung, die zur Generierung von Plosivgeräuschen dient, staut sich die Luft hinter einer Verschlussstelle im Mund- oder Rachenraum. Durch plötzliches Öffnen des Verschlusses wird der Druck abgebaut.

Um von der Anregung zum eigentlichen Sprachsignal als Informationsträger zu kommen bedarf es noch der Signalformung. Die Signalformung wird im Wesentlichen durch die Artikulation beeinflusst, welche im Vokaltrakt stattfindet (siehe Abbildung 2). Der Vokaltrakt umfasst den Rachenraum über der Glottis (Epiglottis, Pharynx) sowie den Mundraum, welcher das supraglottale System mit Ausnahme des Nasenraums einschließt. Die wesentlich zur Klangfarbe des Sprachsignals beitragenden Komponenten im Vokaltrakt sind die Stellung der Zunge und die Stellung der Lippen, wobei sich die Stellungen im zeitlichen Verlauf ändern. Diese sich in der zeitlichen Abfolge ändernden Stellungen werden als Artikulation bezeichnet.

Nach Fant (1960) wird der Vokaltrakt als akustisches Rohr (Ansatzrohr) modelliert. Abbildung 3 zeigt den Prozess der Sprachsignalproduktion, wobei der Ausgangspunkt die Rohschallquelle ist, deren Modifikation durch ein Filter bewerkstelligt wird.

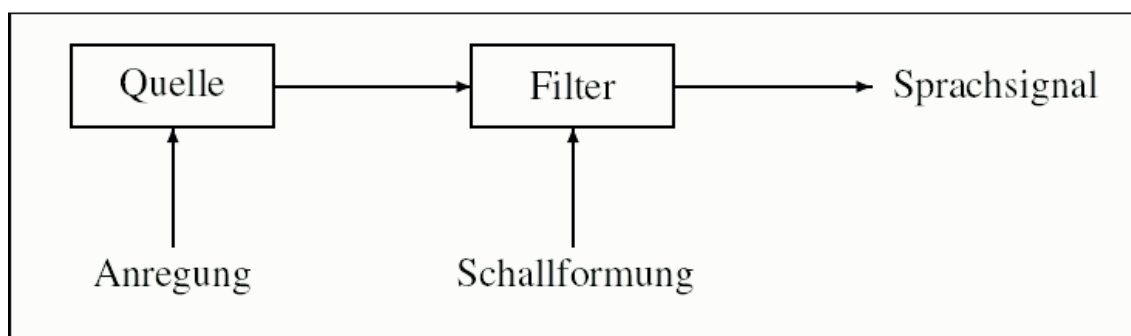


Abbildung 3: Quelle-Filter-Modell: Zusammenhang zwischen Quellsignal und Filtercharakteristik im Zeitbereich (nach B. Pompino-Marshall, 1996)

Die Wände des Vokaltrakts werden bei dieser Modellierung als ideal schallhart angesehen. Ebenso wird die Querschnittsfläche des Ansatzrohres als konstant angenommen, wobei dies in der Praxis am besten durch den neutralen Schwa-Laut repräsentiert ist. Wird

das Ansatzrohr in Schwingung versetzt, ergeben sich Schwingungen mit Wellenlängen, die mit den geometrischen Bedingungen des Ansatzrohres kompatibel sind. Es entsteht bei kompatiblen Schwingungen ein Unterstützungseffekt seitens des Ansatzrohrs, wohingegen Schwingungen anderer Frequenz gedämpft werden. Bei der Konfiguration des Ansatzrohrs, welche ein Ende geschlossen hat und das andere offen, wird die Schallschnelle am geschlossenen Ende Null und erreicht am offenen Ende einen Extremwert, mit Schalldruck Null. Dies erfüllt sich immer dann, wenn ein ungeradzahliges Vielfaches einer Viertelwelle in das Ansatzrohr hineinpasst. Anhand der Formel

$$f_k = \frac{(2k-1)c}{4l}, \quad k = (1, 2, 3, \dots)$$

mit $c = 340\text{m/s}$ und $l = 17\text{cm}$ (Durchschnittswert für Männer) ergeben sich die Resonanzfrequenzen mit

$$f_k = (2k-1) \cdot 500\text{Hz}, \quad k = (1, 2, 3, \dots).$$

Diese Resonanzfrequenzen des Vokaltrakts werden als Formanten bezeichnet (Fant, 1960). Für Sprache sind vor allem die beiden untersten Formanten als F1 bzw. F2 bedeutend. Die Formanten spielen eine entscheidende Rolle bei der Qualität der konkatenativen Sprachsynthese.

Neben den Formanten spielt auch die Koartikulation eine wichtige Rolle. Betrachtet man ein Sprachsignal mittels eines Sonagramms, Abbildung 4, so ist sofort erkennbar, dass eine eindeutige Abgrenzung der einzelnen Laute sich als nicht trivial gestaltet. Dem eindeutigen Abgrenzen, d.h. dem Segmentieren, der einzelnen Sprachlaute kommt bei der konkatenativen Sprachsynthese große Bedeutung zu. Stöber (2002) liefert hierzu eine Betrachtung der automatischen Lautsegmentation. Grundlage für die Lautsegmentation ist die Zuteilung eines Sprachlautes zu einer bestimmten Klasse. Diese Lautklassen sind in einem Phoneminventar zusammengefasst (IPA, SAMPA).

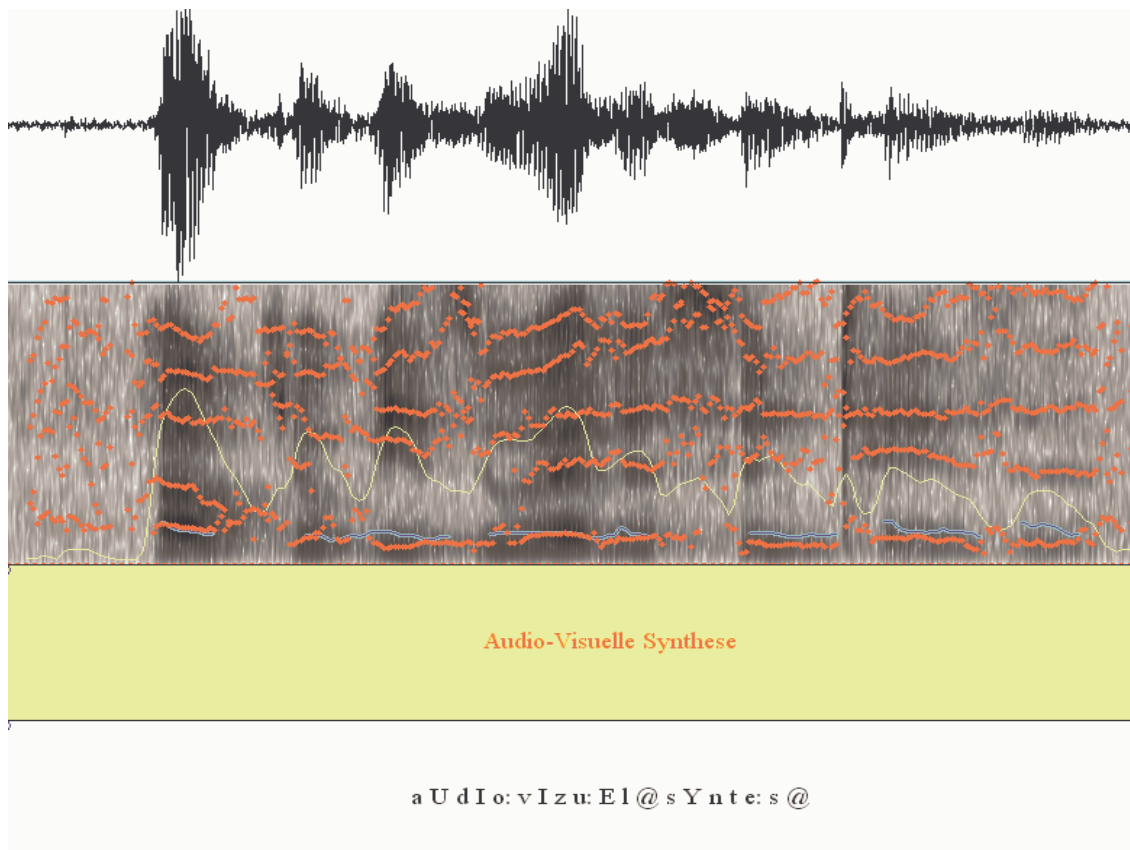


Abbildung 4: Einhüllende, Sonogramm, Intensität, F0-Verlauf und Formanten des Audiosignals Audio-Visuelle Synthese

Wie zuvor erwähnt spiegelt sich vor allem in der zeitlichen Abfolge der Laute und der zugehörigen Stellung der entsprechenden Artikulatoren das sprachliche Phänomen der Koartikulation wider, wobei die Laute je nach Kontext unterschiedlich realisiert werden. Dies bedingt, dass der Position der Artikulatoren eine spezielle spektrale Charakteristik der Laute zugeordnet werden kann. Während des Sprechvorgangs werden die Artikulatoren im zeitlichen Verlauf verändert und nehmen schon vor der eigentlichen Realisierung eines bestimmten Lautes dessen artikulatorische Zielstellung ein. Die Berücksichtigung der Koartikulation ist entscheidend für eine qualitativ hochwertige Sprachsynthese. Eine nähere Betrachtung hierzu findet sich im Abschnitt 2.1.3 „Konkatenative Synthese“. Durch die koartikulationsbedingte Veränderung der artikulatorischen Zielstellung entstehen Lauttransitionen, die sich im zeitlichen spektralen Verlauf des akustischen Signals finden lassen.

Neben den spektralen Eigenschaften, die sich in der zeitlichen Abfolge der Laute ändern, sind die prosodischen Eigenschaften von Sprachlauten von Bedeutung. Die unter dem Begriff „prosodische Merkmale“ zusammengefassten Eigenschaften der lautsprachlichen Kommunikation haben im Hinblick auf eine natürlich klingende Sprachsynthese eine entscheidende Funktion. Nachfolgend werden diese Eigenschaften erläutert. Da über die Prosodie einer Sprache ganze Abhandlungen verfasst wurden, kann auf die prosodischen Eigenschaften einer Sprache in dieser Arbeit nur in einem Überblick eingegangen werden. Für eine detaillierte Betrachtung zum Thema Prosodie wird auf die einschlägige Fachliteratur verwiesen. Siehe hierzu u. a. Pierrehumbert (1991), Uhmann (1991), Möbius (1993), Sonntag (1996), und Günther (2000). Unter Prosodie werden verschiedene Eigenschaften einer lautsprachlichen Äußerung zusammengefasst, die jeweils einzeln sowie im Zusammenspiel eine Funktion in der lautsprachlichen Kommunikation erfüllen und den Suprasegmentalen Merkmalen untergeordnet sind. Die segmentalen Merkmale, Dauer, Intensität und Intonation in sind die Suprasegmentalia eingegliedert. Hierarchisch gesehen kann folgende Gliederung definiert werden: Suprasegmentalia, Prosodie, Intonation. Hier wird zunächst die Prosodie betrachtet.

„Prosodie: aus dem Griechischen , das Hinzugesungene, Beigesang“²

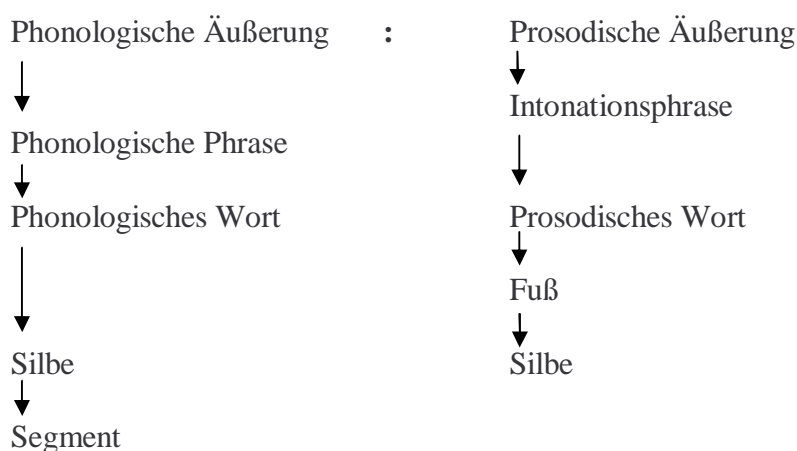
„Prosodie: die Gesamtheit sprachlicher Eigenschaften wie Tonhöhe, Lautheit, zeitliche Strukturierung, Sprechtempo, Stimmlage, Stimmqualität, Klangfarbe, Rhythmus und auch das Fehlen eines sprachlichen Ereignisses. Sprachliche Einheiten, denen diese Eigenschaften zuzuordnen sind, umfassen mehr als einen Laut, weshalb die prosodischen Eigenschaften insbesondere in der angelsächsischen Literatur auch als Suprasegmentale Merkmale bezeichnet werden. Diese sprachlichen Einheiten werden häufig als Prosodeme bezeichnet. Es kann sich dabei um Silben, Wörter, Phrasen, Sätze oder Redebeiträge handeln ...“ (Bussmann 1990, S. 618).

Die Intonation ist nicht an einen sprachlichen Laut gebunden und wird bei Bussmann (1990, S. 352) wie folgt definiert:

² Bußmann 1990

„Gesamtheit der prosodischen Eigenschaften von sprachlichen Äußerungen, die nicht an den Einzellaut gebunden sind.“

Zu den Konstituenten der Intonation zählen die Parameter (Petursson & Neppert 1991) Tonhöhe, Dauer, Intensität, Klangfarbe, Pause, Tempo, Stimmqualität, Musikalität und Emphase. Intonation wird in dieser Arbeit nach Nöth (1990) als funktionaler Aspekt der distinktiven Verwendung aller prosodischen Eigenschaften verstanden. Eine hierarchische phonologische-prosodische Gliederung kann wie folgt definiert werden (Hess 2005):



Die phonologische Gliederung dient den Geltungsbereich von Regeln zu definieren. Die prosodische Gliederung erfüllt den Zweck den Rhythmus einer Äußerung zu definieren und die Betonung und Phrasierung zuzuweisen (vgl. Spencer, 1996)

Auf der Sprachsignalebene werden den prosodischen Eigenschaften physikalische Parameter zugeordnet. Die linguistischen Parameter Quantität, Intensität, Intonation interagieren mit den akustischen Parametern Dauer, Energie, und F0. Die Grundfrequenz (F0) dient ebenso als akustisches Korrelat zur Tonhöhe. Auditiv wird die Tonhöhe allerdings noch durch Lautdauer und Intensität beeinflusst (Lehiste 1970). Die Dauer eines Sprachsegments, z.B. Phon oder Silbe, wird je nach Kontext und Stellung im Satz und Wort unterschiedlich realisiert. Durch das Zusammenspiel der physikalischen Parameter Grundfrequenz, Dauer, Energie und Änderung im Frequenzspektrum wird das Merkmal Akzent im Sprachsignal erkennbar.

Unter Akzent kann die Prominenz von Teiläußerungen verstanden werden wobei Prominenz als linguistisch, perzeptiv relevante Hervorhebung eines Abschnittes des Sprachsignals verstanden werden kann (Möbius 1993). Akzent ist ein relationales Merkmal und wird in den meisten Sprachen hauptsächlich durch die Änderung der Grundfrequenz markiert. Auditiv kann der Akzent als Änderung der Tonhöhe, der Intensität und der Lautdauer wahrgenommen werden. Akzentuierung kann u. A. durch präzisere Artikulationsaktivität realisiert werden. Der Akzent kann als Wortakzent oder auch als Satzakzent dienen und übernimmt als Hervorhebungsfunktion u. A. lexikalische Disambiguierung oder auf semantisch-syntaktischer Ebene die Fokus-Hintergrund-Gliederung einer Äußerung. Akzent und Intonation stehen durch den Parameter Grundfrequenz und deren Änderung im zeitlichen Verlauf in enger Verbindung.

2.1.2 Sprachsynthese: Überblick und Verfahren

Ausgangspunkt für die maschinelle Sprachgenerierung war die menschliche Sprachproduktion. Mitte bis Ende des 18. Jahrhunderts waren es vor allem Kratzenstein und von Kempelen (1791)³, die versuchten, Sprachlaute mechanisch zu erzeugen (vgl. Nikleczy, Olaszy 2003; Olive, J. in: Stork, D. G. 1998). Kratzenstein hatte zum Erzeugen von Vokalen Orgelpfeifen benutzt, an die er Resonanzröhren anbrachte. Von Kempelen hatte zum Produzieren von Vokalen und Wörtern versucht, den Ablauf der Sprachproduktion beim Menschen nachzubilden. Ein Blasebalg simulierte dabei die Lungen. Ein Rohr diente für den Vokaltrakt, wobei die Stimmlippen durch ein aus Elfenbein gefertigtes Rohrblatt nachgebildet wurden.

Artikulatorische Sprachsynthese ist der Teil der Sprachsyntheseforschung, welcher sich das Paradigma der menschlichen Sprachproduktion zu eigen macht und auf Basis von Sprachgenerierungsparametern die menschliche Sprachproduktion mittels Filter- und Übertragungsfunktion nachzubilden versucht. Hierbei werden alle Artikulatoren berücksichtigt. Neben der Artikulatorischen Synthese wurde die Formantsynthese entwickelt (Klatt 1982). Mit Hilfe von Schaltungen und digitalen Filtern wird versucht, verschiedene Formantfrequenzen zu erzeugen.

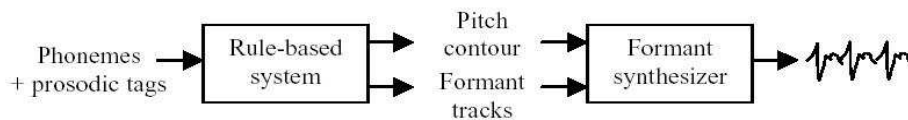


Abbildung 5: Schematisches Blockdiagramm eines regelbasierten Synthesystems, nach: Huang et al. (2001)

Die Formantsynthese unterscheidet zwischen stationären, stimmhaften Klängen und dem Rauschen, also stimmlosen Klängen. Da in realen Sprechsituationen keine stationären Klänge vorkommen, werden über einen Regelsatz Grundfrequenz und Formantfrequenzen abhängig von der Zeit geändert. Abb. 5 zeigt ein Blockdiagramm eines regelbasierten Synthesystems, wobei hier nur die Parameter der Grundfrequenzkontur und Formantverlauf zu sehen sind. In realen Systemen wächst die Anzahl der Parameter auf vierzig an. Die Formantsynthese folgt dem Quelle-Filter-Modell in Abbildung 3 der Sprachproduktion. In der Praxis wird die Vokaltrakt-Antwort des Sprachsignals $S(z)$ als linear angenommen. Die z -Transformierte kann daher synthetisiert werden mittels:

$$S(z) = U(z)H(z)$$

$U(z)$ ist eine Funktion, die das Anregungssignal modelliert. $H(z)$ repräsentiert die Übertragungsfunktion des Filters. Die Übertragungsfunktion $H(z)$ setzt sich zusammen aus der Vokaltrakt-Antwort $V(z)$ und der Funktion $R(z)$, welche die Abstrahlcharakteristiken bei der Lauterzeugung widerspiegelt. Die Anregungsfunktion $U(z)$ umfasst den Impuls bzw. das weiße Rauschen als $P(z)$ und $G(z)$. Daraus ergibt sich für stimmhafte Laute folgende Gleichung:

$$S(z) = P(z)G(z)V(z)R(z)$$

Anwendungen der Forschung zu Artikulatorischer Synthese und Formantsynthese sind u. a. zu finden in CASY (Haskins Laboratories), MITalk (Allen 1987), Klattalk (Klatt 1982). Eine detaillierte Übersicht findet sich in Huang, Acero, Hon (2001).

³ Wolfgang von Kempelen, Ingenieur im Dienste von Maria Theresia in Wien. Er wurde 1734 in Pressburg, der damaligen Hauptstadt von Ungarn, geboren und starb 1804 in Wien.

2.1.3 Konkatenative Sprachsynthese

Konkatenative Sprachsynthese bezeichnet ein Verfahren, das gegebenen Eingabetext lautsprachlich realisiert, indem zuvor aufgenommene und gespeicherte Sprachsegmente zu der entsprechenden Äußerung zusammengefügt wird. Als Sprachbausteine kommen theoretisch alle möglichen linguistischen Einheiten in Frage. Dennoch hat es sich in der Praxis als nicht praktikabel erwiesen, einfach nur jeden einzelnen Laut, der in einem Sprachsystem vorkommt, oder jedes Wort einer Sprache zu verwenden. Wollte man einen Text wiedergeben, welcher durch Wortkonkatenation synthetisiert wird, so müsste man alle möglichen Wörter, die in einer Sprache vorkommen können, abdecken, was zu einer sehr großen Datenmenge führen würde und erheblichen Aufwand benötigt, um damit ein Sprachsynthese-System zu erstellen. Ein ebenso zu lösendes Problem ist hierbei die jeweilige unterschiedliche prosodische Realisierung. Dies führt auch dazu, dass eine Konkatenation auf Phonbasis, wie zuvor erwähnt, sich als nicht sinnvoll erweist, da je nach kontextueller Realisierung der Phone unterschiedliche spektrale Charakteristika zum Tragen kommen, und auch die Übergänge nicht mit realisiert sind.

Für die konkatenative Synthese wurde mit der Diphonsynthese ein erfolgreicher Ansatz realisiert. Hierbei wurden Sprachaufnahmen eines phonetisch ausbalancierten Textes erstellt, bei dem die jeweiligen Diphone in unterschiedlichen Kontexten realisiert wurden (Portele 1996). Von Mitte bis gegen Ende der 90er Jahre fand in der Sprachsynthese ein Paradigmenwechsel statt. Die klassische Diphonsynthese wurde durch eine korpusbasierte Sprachsynthese in den Hintergrund gedrängt. Diese korpusbasierten Verfahren werden als Unit-Selection-basierte Sprachsynthese beschrieben. Sagisaka (1988) hat eine erste Arbeit hierzu vorgestellt. Das von ihm entwickelte ATR-v-Talk-System benutzte als Grundlage ein Sprachdatenkorpus, aus dem dann die entsprechenden Sprachbausteine extrahiert und in einem zweiten Schritt hinsichtlich der gewünschten Äußerung neu konkateniert wurden. Im Gegensatz zu der Diphonsynthese werden für die Unit-Selection-basierte Sprachsynthese Korpora verwendet, die ein und dasselbe Sprachsegment in mehreren prosodischen Realisierungen enthalten.

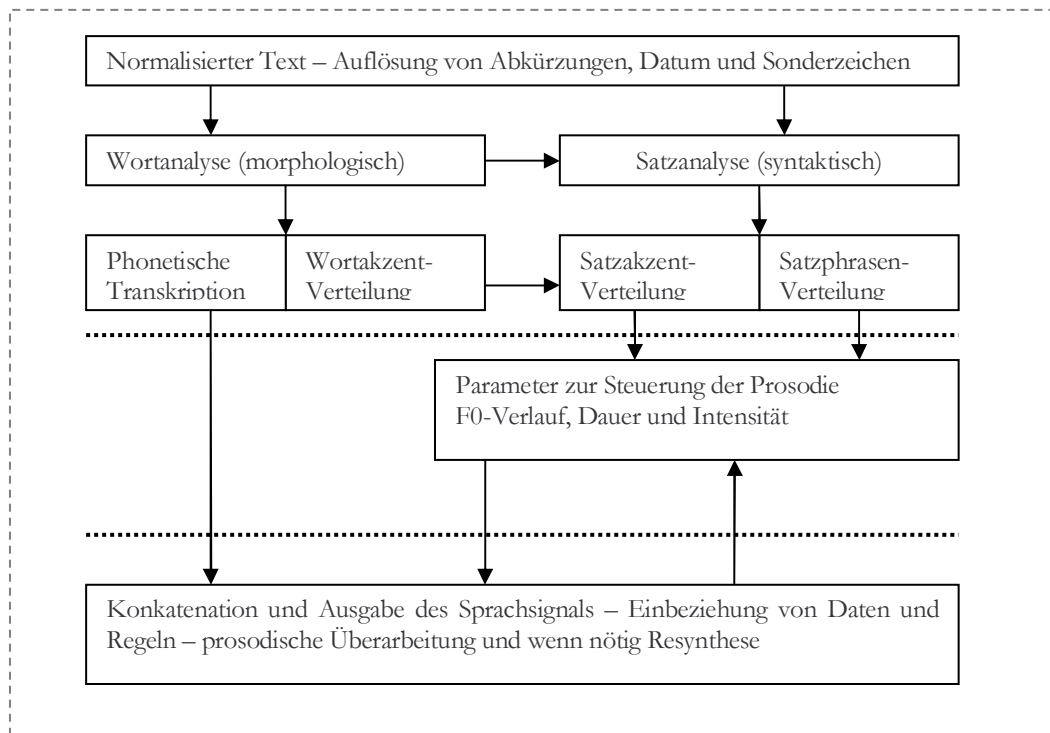


Abbildung 6: Blockdiagramm der TTS-Komponenten (modifiziert nach Kraft, 1997)

Bei der Diphonsynthese hingegen wurden die Diphone möglichst in neutraler Ausprägung im Trägersatz integriert, um bei der späteren Verkettung ein neutrales Sprachsignal zu erhalten, welches durch Signalmanipulation an die gewünschten prosodischen Eigenschaften angepasst wurde.

Bei der Unit-Selection-basierten Sprachsynthese werden alle im Korpus enthaltenen Segmente einer sprachlichen Repräsentation als potentielle Sprachbausteine angesehen, wobei jedes Sprachsegment unterschiedliche prosodische und spektrale Eigenschaften besitzt. Dies hat zum Vorteil, dass schon zu Laufzeit ein Sprachsegment ausgewählt werden kann, das bestimmte Eigenschaften erfüllt. Wird z.B. ein Laut gebraucht, der eine bestimmte Dauer sowie eine bestimmte Grundfrequenz hat, so ist die Wahrscheinlichkeit groß, dass solch ein Laut schon im Korpus vorhanden ist und eine nachträgliche Signalmanipulation vermieden werden kann. Denn jede nachträgliche Signalmanipulation bringt Störungen im Signal mit sich. Die zurzeit eingesetzten konkatenativen Sprachsynthesysteme folgen einem dreistufigen Aufbau. Abbildung 6 zeigt diesen Aufbau schema-

tisch. In Kapitel 4 wird die konkatentative Sprachsynthese anhand der von Hunt, Black und Campbell (1995, 1996) eingeführten Unit-Selection-Sprachsynthese näher betrachtet.

2.2 Audio-visuelle Synthese

TTS-Systeme haben den Nachteil, dass die Wahrnehmung nur eindimensional durch auditive Perzeption von den Kommunikationspartnern verarbeitet werden kann. Dieser Nachteil wirkt sich in einem störungsbedingten Umfeld signifikant auf die Verständlichkeit und somit auch auf die Gesamt-Wahrnehmung der Anwendung aus. Dies führt zu einer negativen Evaluation der Gebrauchsfähigkeit solcher Applikationen und kann einen Akzeptanzverlust bei Nutzern solcher Systeme hervorrufen. Vergleicht man dies mit einem Radiogerät bei dem das Signal gestört ist, ist der Nutzer geneigt, die Störung durch Justieren der Sendereinstellungen zu mindern, einen anderen Sender zu suchen oder das Radio abzuschalten. Nur in Ausnahmefällen neigt der Nutzer dazu, trotz der Störung das Radioprogramm weiter zu verfolgen. Eine systematische sowie logische Weiterentwicklung zur Verbesserung der natürlichen Mensch-Maschine-Interaktion mittels Sprache besteht darin, in den Wahrnehmungs- und Kommunikationsprozess von Anwender und System visuelle Informationen mit einzubinden und TTS-Systeme zu audio-visuellen Systemen auszubauen. Ausgehend von der menschlichen „Face-to-Face-Kommunikation“ wird ersichtlich, dass der Informationsgehalt, wie auch die semantische Auflösung von sprachlichen Äußerungen, durch visuelle Informationen mit gestützt und codiert ist. Der natürliche Kommunikationspartner wird somit in die Lage versetzt, die sprachlich-akustische Äußerung kognitiv besser zu erfassen und einzuordnen. Dies trifft vor allem auf emotionale Äußerungen zu. Hierbei wird ein emotionaler Zustand, wie Verärgerung oder Freude nicht nur sprachlich ausgedrückt, sondern zusätzlich über Gesichtsmimik (Ekman 2004). Ekman und Friesen (2002) haben ein „Facial Action Coding System (FACS)“ erstellt, mittels dessen Mimik, aufgrund von Gesichtsmuskelbewegungen, klassifiziert werden kann.

Audio-visuelle Synthese steht seit einiger Zeit im Interesse von Forschung und Anwendung. Vor allem aus Sicht der Anwendung verspricht man sich eine Steigerung der Akzeptanz der interaktiven Mensch-Maschine-Kommunikation. Beobachtet man die Kommunikation beim Menschen, so lässt sich sehr leicht erkennen, dass während des Kom-

munikationsprozesses nicht nur das auditive Signal in Betracht gezogen wird. Die lautliche Wahrnehmung während der Kommunikation basiert nicht nur auf dem, was man hört, sondern auch auf dem, was man sieht. Bei normalen „Hörern“ (Beobachtern) wird die Fähigkeit, das sprachauditive mit dem visuellen Signal zu integrieren, vor allem klar, wenn das Signal von einem Hintergrundgeräusch überlagert wird, im Besondern von anderen Stimmen. Dieser Effekt der Stimmüberlagerung ist auch als „Cocktail-Party“-Effekt aus der Spracherkennungsforschung bekannt. Der Mensch ist in der Lage, selbst bei einer Kommunikation, die im Umfeld von anderen Konversationen stattfindet, eine Unterhaltung zu führen. Dies ist u. a. auf die Fähigkeit zurückzuführen das Quellsignal eindeutig zu identifizieren und zusätzlich das akustische Signal visuell zu verifizieren. Die lautliche Wahrnehmung basiert nicht nur auf dem, was man hört, sondern auch auf dem, was man sieht, wie es von McGurk und MacDonald demonstriert wurde. Sie spielten Versuchspersonen ein Videoband mit der Audio-Silbe „ba“ vor, synchronisierten das Audiosignal mit einem visuellen „ga“ und stellten fest, dass die Probanden „da“ wahrnahmen (McGurk, MacDonald 1976).

Audio-visuelle Synthese vereint die Generierung von künstlich erzeugter Sprache, meist durch ein Text-to-Speech-System, mit einer visuellen Ausgabe, die den Sprechvorgang des Menschen visuell simuliert und dabei entweder den Kopf und das Gesicht künstlich generiert oder eine Ganzkörperdarstellung verwendet. Allgemein kann nachfolgende Unterscheidung in Tabelle 1 der visuellen Darstellung ausgemacht werden.

Bezeichnung	visuelle Darstellung
Animated Face	ein Gesicht, das artikuliert und eventuell auch Emotionen usw. zeigt
Talking Head	ein kompletter Kopf, der artikuliert
Avatar	Darstellung einer Person, entweder nur eines Gesichts, eines Kopfes oder einer kompletten Gestalt
Conversational Agent	meistens eine komplette Gestalt, so dass auch z.B. Hand- und Armbewegungen genutzt werden können

Tabelle 1: Übersicht der visuellen Darstellungsformen in der audio-visuellen Synthese und deren Bezeichnung, nach Bailly et al. 2003

In der vorliegenden Arbeit wird ausschließlich die Darstellung und Modellierung eines „Talking-Head“ berücksichtigt, bei dem der Audiokanal durch ein TTS-System generiert wird und der visuelle Kanal durch 2D-Bildkonkatenation. Beide Quellen werden in der zeitlichen Abfolge synchronisiert und als audio-visuelles Signal ausgegeben.

Die als „Talking Heads“ bezeichneten visuellen Darstellungsformen unterscheiden sich in der Art ihrer Modellierung, Erscheinung und Implementierung. Es werden in der audio-visuellen Synthese vor allem zwei Hauptrichtungen unterschieden. Dies sind zum einen der modellbasierte Ansatz und zum anderen der auf 2D-Videsequenzen basierende video-realistische Ansatz (Bailly et al 2003). Aktuell gibt es im Gegensatz zur Sprachsynthese keine dominierenden Techniken, die den einen oder anderen Ansatz als state-of-the-art einstufen. Die in dieser Arbeit entwickelte audio-visuelle Synthese stützt sich auf den 2D-Bild-basierten Ansatz und verwendet reale Videosequenzen, welche einen Sprecher in einer Sprechsituation wiedergeben, bei dem der Kopf und die Schultern zu sehen sind. Diese Einstellung wurde einem Nachrichtensprecher nachempfunden. Audio-visuelle Synthese kann in verschiedenen Ausprägungen realisiert werden. So ist die Nachbildung von menschlichen Köpfen mit synchronisierter Sprachausgabe ein aktuelles Forschungsgebiet im Schnittpunkt von Bild- und Sprachverarbeitung. Die Bildverarbeitung liefert hierzu die nötigen Algorithmen zur Entwicklung der physiognomischen Gesichts- und Kopfbestandteile, und die Sprachverarbeitung stellt den Zusammenhang von sprachlicher Äußerung und den Modellen der Lippenbewegung bzw. der visuellen Prosodie, welche Bewegungsabläufe beinhaltet, die während eines Sprechvorgangs Ausdruck in Mimik und Gestik erhalten, her. Audio-visuelle Synthese vereint also im Wesentlichen die sprachliche Äußerung mit einem zugehörigen Kopf, der den Vorgang des Sprechens simuliert.

Diese unterschiedlichen Ansätze werden nachfolgend in Abschnitt 2.2.2 und 2.2.3 aufgeschlüsselt. In dieser Arbeit kann nur übersichtsweise auf die unterschiedlichen audio-visuellen Synthese-Verfahren eingegangen werden. Für eine detaillierte Betrachtung der einzelnen Verfahren muss auf die Arbeiten der einzelnen Autoren verwiesen werden.

Als Motivation der „Talking-Heads“ liegt die Annahme zugrunde, dass der Sprechvorgang beim Menschen durch so genannte nicht-verbalisierte Signale unterstützt wird. Ekman (1999) hat hierzu eine länderübergreifende Studie erhoben und die unterschiedlichen

non-verbale Signale herausgearbeitet. Non-verbale Signale sollen in dieser Arbeit als Begriff verstanden werden, welcher die Gesamtheit der Gesichtsmimik kennzeichnet. So sind die Artikulationsgesten bei Menschen visuell erkennbar und die prosodischen Eigenschaften des akustischen Sprachsignals werden durch visuelle „Cues“ unterstützt, z.B. Augenbrauenbewegungen, Augenblinzeln, Kopfbewegungen, aber auch Hand- und Armbewegungen.

2.2.1 Physiologien nonverbaler Kommunikation

Bei Menschen kann man die Gesichtsbewegungen, welche aus physisch-biologischen Gründen notwendig sind, von denen, die Emotion ausdrücken, abgrenzen. Emotionsbedingte Gesichtsbewegungen sind meist stärker in ihrem Ausdruck als die auf natürlicher Körperreaktion zurückzuführenden Mimik. Diese auf die menschlichen, biologischen Notwendigkeiten abgestimmte Gesichtsmimik ist von Ekman als „manipulators“ kategorisiert. „Manipulators“ lassen sich auf eine mechanische Wirkungsweise zurückführen, die immer wieder vollzogen wird. Zu dieser Kategorie gehören zum Beispiel:

- Augenlidblinzeln zur Befeuchtung der Augen
- Mund-Zungen-Bewegung zur Befeuchtung der Lippen
- Öffnen des Mundes zur Atmung.

Neben der sprachlichen Äußerung spielen während der Kommunikation die nonverbale Signale, die mit einem bestimmten emotionalen Zustand des Sprechers in Verbindung stehen sowie eine visuelle Pragmatik in sich vereinen, eine entscheidende Rolle. Gesichtsmimik und Bewegung der einzelnen Muskelpartien, welche zu Gesten führen, werden entweder vom Menschen unbewusst gesteuert oder gezielt produziert. So verändert sich die Gestik bei starker Erregung nicht durch einen bewussten Akt. Im Gegensatz dazu lässt sich eine bewusste Emotionshaltung durch Anzeigen eines bestimmten Gesichtsausdrucks steuern. Diese Kategorie wird von Ekman als „regulators“ bezeichnet. Regulatorische Gesichtsmimik verleiht einer Konversation die Möglichkeit der Steuerung des Dialogs. So können hierdurch Aufmerksamkeit oder Langeweile während eines Dialogaktes ausgedrückt werden. Ebenso wenn ein Gesprächspartner in eine Unterhaltung einhaken will, kann er dieser durch zuvor begonnenen Gesichtsmimik zum Ausdruck bringen. Bes-

tätigung, Abneigung oder Zweifel, die der Kommunikationspartner während einer Konversation einbringt, wird zum Teil ebenso über Gestik und Mimik vermittelt. Eine weitere Kategorie der non-verbalen Signale bilden die „emblems“. Diese frei als „Zeichen“ übersetzte Kategorie beinhaltet die Anreicherung des Gesagten mit visuellen Unterstützungsgesten. Diese visuellen Gesten „emblems“ verleihen dem Gesagten ein stärkeres Gewicht, als wenn ohne die unterstützende Gesten etwas sprachlich ausgedrückt wird. Hier ist vor allem an die Ja-Nein-Gesten zu denken. Auch Abgrenzungen, welche sprachlich durch Gegensätzliches ausgedrückt werden, können durch diese Art der Gestik unterstützt werden. Hier unterscheidet Ekman nochmals die Unterkategorien „illustrators“ und „punctuators“, wobei die unterstützende Wirkung der Gesichtsgestik zu den „illustrators“ gezählt wird. Zu den „punctuators“ zählen Pausen, welche durch Luftholen vollzogen werden; dabei werden die Augenbrauen hochgezogen und der Kopf bewegt. Die Darstellung dieser nonverbalen Signale kann bei Ekman (Ekman, Friesen 1975) detailliert nachvollzogen werden. Die nonverbalen Signale sind entscheidend für jede Art von „Talking Head“ und tragen zur Natürlichkeit und Akzeptanz einer solchen Anwendung bei.

2.2.2 Modellbasierte Ansätze

Bei der Generierung von Köpfen mittels des modellbasierten Ansatzes wird der menschliche Kopf als Gitternetz erfasst. Abbildung 7 zeigt unterschiedliche Darstellungsformen von menschlichen Köpfen, die mit Hilfe eines Gitternetzes in ihrer Struktur erfasst wurden. Dieses Gitternetz wird als „mesh“ bezeichnet. Typischerweise werden die modellierten Köpfe als polygonale Gitternetze im dreidimensionalen Raum dargestellt. Mit diesem Gitternetz wird die Oberfläche des menschlichen Gesichtes modelliert. Die Haut und Bewegungen werden jeweils mit unterschiedlicher Präzision nachgebildet. Die Netztopologie bleibt bei der Animation konstant, während die Knoten verschoben werden und so unterschiedliche muskuläre Bewegungen simulieren. Die Verschiebung der Knoten wird durch eine Anzahl von Parametern gesteuert. Die Parameter können sich durch Interpolation zwischen Hauptknotenpunkten berechnen. Weitere Verfahren zur Verschiebung der Knoten sind direkte Parametrisierung, pseudo-muskuläre Deformation, physische Simulation oder datenbasierte Verfahren. Interpolation ist die am weitesten verbreitete Methode

und wird in vielen Anwendungen eingesetzt. Es bestehen vordefinierte Schablonen von prototypischen Gesichtsausdrücken oder von Lippen-, Augenstellung.

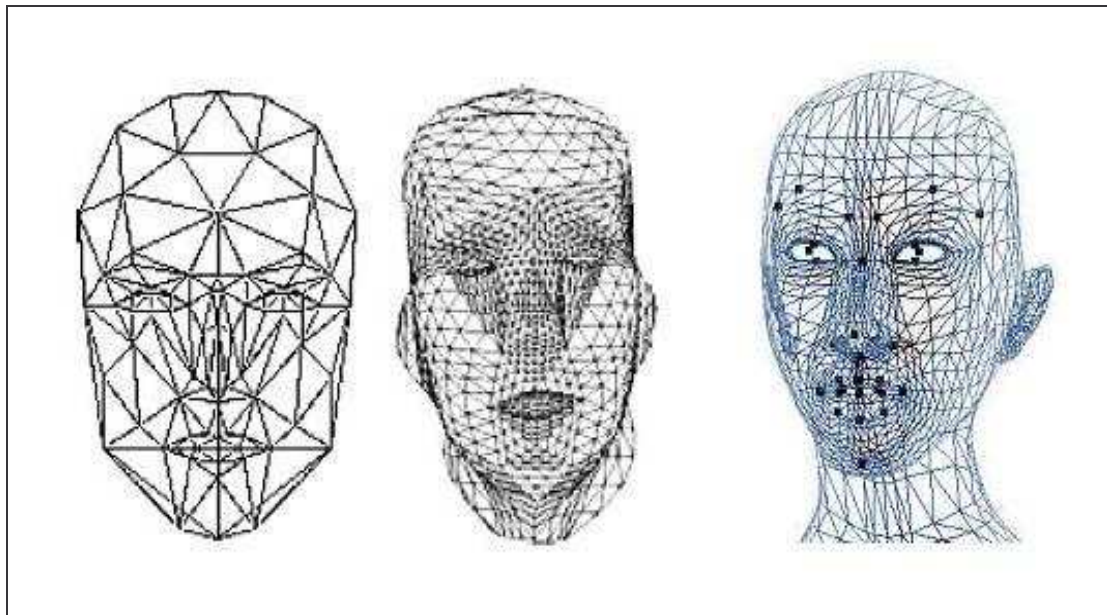


Abbildung 7: Gitternetz-basierte Darstellung eines menschlichen Kopfes nach:
Rydfalk (1987), Tsai et al. (1997)

Während der Animation wird jeweils zwischen der Ausgangsstellung und der Zielstellung interpoliert, wobei die Topologie bestehen bleibt und nur die Knoten verschoben werden. Diese Methode ist einfach zu implementieren und ist daher attraktiv für einfache Anwendungen. Durch die Einfachheit ist man in der Generierung von Gesichtsausdrücken sehr beschränkt. Die Erweiterung erfolgt ausschließlich in der Integration neuer zusätzlicher Schablonen, um den Freiheitsgrad der Gesichtsmodellierung zu vergrößern. Direkte Parametrisierung, welche von Parke (1982) vorgeschlagen und eingeführt wurde, hat diese Einschränkungen nicht und hat eine direkte geometrische Transformation der Knoten entwickelt, die sich je nach Anforderung als Rotation, Skalierung und Interpolation in der zu bearbeitenden Gesichtsregion gestaltet. Parke hat für die Anforderungen Parameter zusammengestellt, die er in „expression“ und „conformation“ Parameter einteilt. Die von Parke entwickelte geometrische Gesichtsmodellierung kann als terminal-analoge Synthese betrachtet werden. Dies bedeutet, es wird nicht versucht, die zugrunde liegenden physiologischen Mechanismen zu modellieren, die das Sprachsignal und die Gesichtsdeformationen verursachen, sondern diese nur in geometrischen Termini zu reproduzieren. Da-

bei bewegen sich wie zuvor erwähnt die Knoten, um Rotation (für den Unterkiefer) und Translation (für Mundöffnung oder Lippenspreizung) zu bewirken. Dabei werden 84 „feature points“ (FP) durch 68 „facial action parameters“ (FAP) kontrolliert, um Gesichtsbewegungen auf allen Niveaus zu beschreiben. Eine Alternative zu Parke wurde von Magnenat-Thalmann et al. (1988) vorgeschlagen, die auf Arbeiten von Platt et al. (1981) zurückzuführen ist. Artikulatorische Parameter stellen ein Set von artikulatorischen Freiheitsgraden dar, die die Bewegung von Knoten im Gitternetz bestimmen. Diese können durch eine statistische Analyse der 3D-Koordinaten hunderter Positionen im Gesicht („facial flesh points“) abgeleitet werden. Diese Art der Modellierung wird auch pseudo-muskulär genannt, weil Muskelbewegungen durch geometrische Deformationsoperatoren simuliert werden. Die Muskeln selbst werden nicht modelliert, sondern als Kräfte implementiert, die die Gesichtsgeometrie ändern. Den Knotenpunkten werden Massen zugeordnet, die mittels Federn miteinander verbunden sind, um die Elastizität der Haut zu imitieren. Bei Waters (1987) und Terzopoulos (1990) werden statt durch geometrische Kontrollparameter die Gesichtsbewegungen direkt durch Muskelaktivierung gesteuert. Die Steuerung wird durch den jeweiligen Grad der kommunikativen Absichten erreicht. Mittels des „facial action coding system“ (FACS) werden Gesichtsausdrücke von 66 Muskelbewegungen beschreibbar.

2.2.3 Videodatenbasierter Ansatz

Visuelle Synthese, die die synchronisierte Lippenbewegung zu dem gewünschten Eingabetext wiedergibt, bedient sich bei 2D-Bild- und videodatenbasierten Ansätzen Gesichtsregionen, die aus den 2D-Einzelbildern gewonnen werden, oder eines Morphing-Ansatzes, der die entsprechenden Pixel im 2D-Bild in die gewünschte Position verschiebt. Unter Morphing ist hier eine Überblendtechnik zu verstehen, wobei die Farbwerte der Pixel interpoliert werden. Diese zwei Ansätze kann man als segmentbasierter „Overlay“- und „Morphing“-Ansatz bezeichnen. Bailly et al. (2003) unterscheiden drei Ansätze, wobei der dritte Ansatz ebenso eine Veränderung der Pixelposition im 2D-Bild beinhaltet und nur zusätzlich die farbliche Erscheinung der Pixel berücksichtigt. Bei dem der Sprachsynthese entlehnten Ansatz mittels Auswahl von Segmenten und der anschließenden Konkatenation von Einheiten besteht die größte Herausforderung darin, die Segmente

so ein- und zusammensetzen, dass dem menschlichen Auge keine erkennbaren Störungen des Bildflusses auffallen.

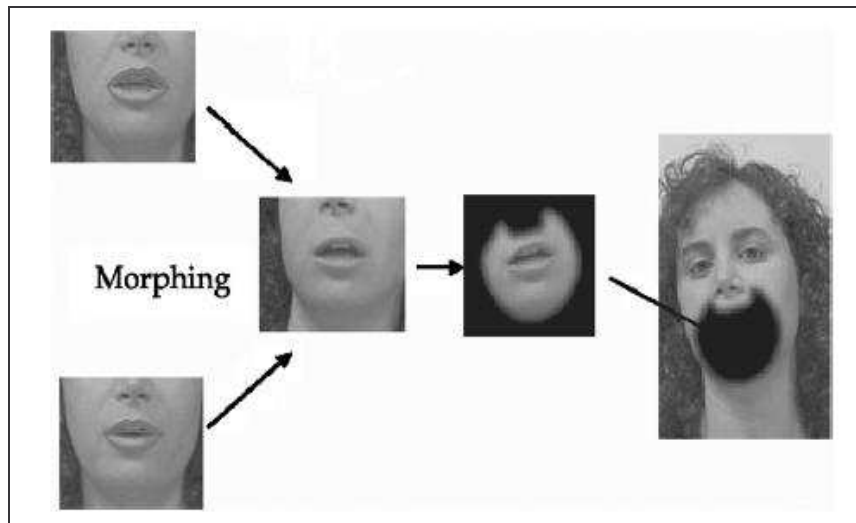


Abbildung 8: Overlay-Technik im VideoRewrite-System (Bregler et al. 1997).

Systeme, die dieser Technik folgen, sind das von Bregler et al. (1997) entwickelte VideoRewrite-System und das von Cosatto et al. (2000) implementierte AT&T FaceTalk-System. Das VideoRewrite-System bedient sich der Mund-Kinn-Gesichtsregion; siehe Abbildung 8, die als Triphon-Visem in einer Datenbank abgelegt ist und bei Bedarf in das zugrunde liegende Video eingefügt wird. Die beste Position für den einzufügenden Bildbereich wird mittels eines Schätzverfahrens errechnet.

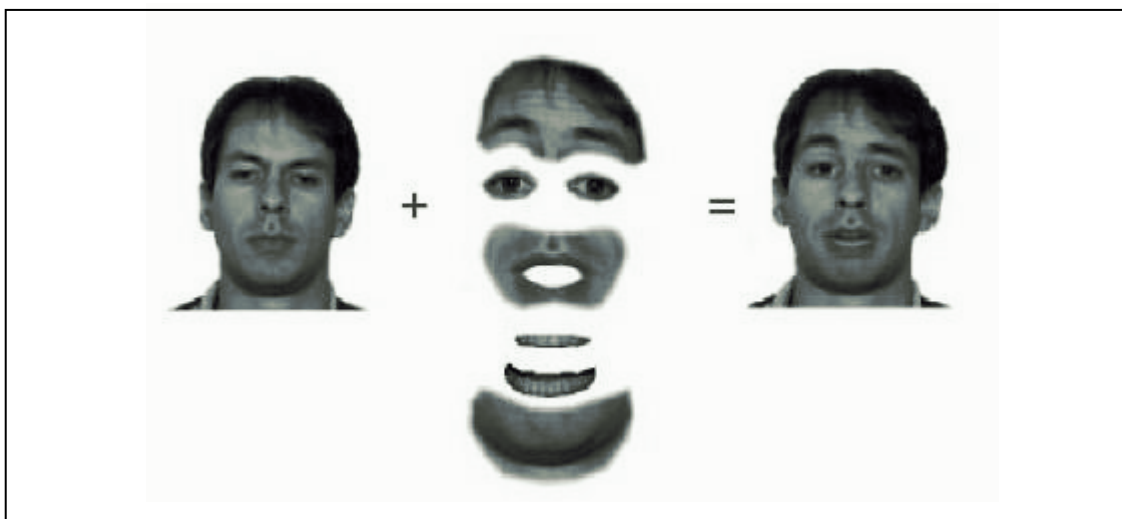


Abbildung 9: Verwendete Gesichtsregionen für das AT&T FaceTalk-System von Cosatto et al. 2002

Das FaceTalk-System von Cosatto et al. verwendet nicht nur die Mund-Kinn-Region, sondern teilt das Gesicht in sechs Teile auf. Abbildung 9 zeigt die Gesichtsregionen, die aus einem Basisgesicht gewonnen und zu einem neuen Gesicht zusammengesetzt werden. Die Gesichtsregionen sind ebenso in einer Datenbank abgelegt und werden bei Bedarf in ein Basisgesicht eingefügt.

Das audio-visuelles Synthese-System Miketalk, das auf der Grundlage von Pixel-Morphing (Beier et al. 1991) besteht, wurde von Ezzat et al. (1998) entwickelt. Grundlage bei diesem System bildet der Optical-Flow-Algorithmus, der die entsprechenden Pixel des Quellbildes hin zu der gewünschten Zielstellung im Zielbild bewegt. Eine Erweiterung dieses Systems wurde von Ezzat et al. (2002) mit der Entwicklung des MMM-Algorithmus (Multidimensional Morphable Model) vorgestellt. Bei diesem Ansatz werden die Pixelbewegungen aus einem Trainingsdatensatz gelernt, und bei der Synthese einer Zieläußerung wird das entsprechende Modell der Pixelbewegung verwendet. Als Alternative zu dem Ansatz von Ezzat wurden von Brook et al. (1998) statistische Modelle als Grundlage verwendet, wobei das Bild in Subregionen unterteilt wird. Aus den Subregionen, Pixelblöcke von 16x16 Pixeln, werden mittels einer Hauptkomponentenanalyse die Hauptkomponenten extrahiert. Es werden 40 bis 50 globale Hauptkomponenten verwendet. Mittels dieser wird mit Hilfe von Texturen das entsprechende Bild erzeugt.

PROBABILISTISCHE VORBEDINGUNGEN

Das vorliegende Kapitel erläutert die probabilistischen Vorbedingungen, die als Grundlage der entwickelten und eingesetzten automatischen Trainingsverfahren für die Sprachsynthese verwendet werden. Aufgrund des mathematischen Umfangs und der mathematischen Vorbedingungen kann nur begrenzt auf die einzelnen Themenbereiche eingegangen werden. Für eine detailliertere Ausführung sei auf mathematische Fachliteratur verwiesen. Nachfolgend werden bedingte Wahrscheinlichkeiten betrachtet, sowie Bayes-Wahrscheinlichkeit, Endliche Automaten, Markov-Ketten, Hidden-Markov-Modelle und Entropie bzw. bedingte Entropie.

3.1 Bedingte Wahrscheinlichkeiten

Der Begriff der Wahrscheinlichkeit wurde von Laplace eingeführt, der die Wahrscheinlichkeit im Zusammenhang mit der Untersuchung von Glücksspielen als mathematisches Gebilde einführte. Der Wahrscheinlichkeitsbegriff basiert auf der Annahme, dass die Anzahl der das vollständige Ereignissystem N bildenden Elementarereignisse, $n \in N$ endlich ist. Weiterhin gilt die Annahme, dass alle Elementarereignisse gleich wahrscheinlich sind, d. h. mit der gleichen Wahrscheinlichkeit eintreten, wie zum Beispiel beim Werfen einer idealen Münze. Dies wird als Laplace-Experiment bezeichnet. In Laplace-Experimenten

lassen sich die Wahrscheinlichkeiten berechnen mit $P(n) = \frac{1}{|N|}$, $n \in N$. Bei den Laplace-

Experimenten ist die relative Häufigkeit eines Ereignisses A der Ausgangspunkt. Wird der Versuch x -mal durchgeführt, heißt die Anzahl, wie häufig das Ereignis A auftritt, ab-

absolute Häufigkeit $a_n(A)$. Der Quotient $H_n(A) = \frac{a_n(A)}{n}$ wird als relative Häufigkeit des

Ereignisses A bezeichnet. Dies ist gleich der Zahl der für dieses Ereignis günstigen Ergebnisse, dividiert durch die Zahl der Ergebnisse insgesamt. Die relative Häufigkeit des sicheren Ereignisses ist 1, die des unmöglichen 0. Anhand des schwachen Gesetzes der

großen Zahlen konvergiert die relative Häufigkeit gegen die a-priori-Wahrscheinlichkeit $P(A) := \lim_{n \rightarrow \infty} H_n(A) = P(A)$. $P(A)$ heißt die Wahrscheinlichkeit von A , die als Grenzwert der Häufigkeit des Auftretens eines Ereignisses A , in Relation zu der Menge aller Punkte im Stichprobenraum. Nach Engeln-Müllges (2001) beruht der streng mathematische Aufbau der Wahrscheinlichkeitsrechnung auf den folgenden Axiomen, mit denen sich alle Regeln für die Wahrscheinlichkeitsrechnung ableiten lassen:

Axiom 1: Jedem zufälligen Ereignis A wird eine reelle Zahl $P(A)$, die Wahrscheinlichkeit für A , zugeordnet, die der Ungleichung $0 \leq P(A) \leq 1$ genügt.

Axiom 2: Die Wahrscheinlichkeit des sicheren Ereignisses A ist 1: $P(A)=1$

Axiom 3: Die Wahrscheinlichkeit einer Summe unverträglicher Ereignisse U ist gleich der Summe der Wahrscheinlichkeiten der Einzelereignisse:

$$A_1 \cap A_2 = U \Rightarrow P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

Weiterhin ist zu beachten, dass die Summe der Wahrscheinlichkeiten eines Ereignisses A und des hierzu komplementären Ereignisses \hat{A} gleich 1 ist.

Im Folgenden wird dargestellt, wie Informationen über das Eintreten gewisser Ereignisse berücksichtigt werden können, um die Wahrscheinlichkeit für das Eintreten anderer Ereignisse dann eventuell neu bewerten zu können. Dies ist das Konzept der bedingten Wahrscheinlichkeiten. Wenn die Informationen keinen Einfluss auf die Wahrscheinlichkeit der eintretenden Ereignisse haben, dann spricht man von unabhängigen Ereignissen in dem anderen Fall von abhängigen Ereignissen⁴. Zwei Ereignisse heißen unabhängig, gerade dann wenn $P(A \cap B) = P(A)P(B)$ gilt. Wenn die Wahrscheinlichkeit des Ereignis-

⁴ Die Ereignisse werden in dieser Arbeit durch die zur Verfügung stehenden Sprachsegmente repräsentiert. Zur Veranschaulichung der bedingten Wahrscheinlichkeit kann der menschlichen Sprachproduktionsprozess herangezogen werden, bei dem Laute in zeitlicher Abfolge produziert werden und die vorhergehenden Laute Einwirkung auf den aktuellen Laut und die folgenden Laute hat. Dieser Einfluss kann u. A. bei der Koartikulation sehr gut beobachtet werden, in dem während der Realisierung eines Lautes sich die Artikulatoren schon auf den nächsten Laut vorbereiten. Folgt, der vorhergehende Laut hat Einfluss auf den nachfolgenden Laut. Aber dies kann nicht nur bei der Koartikulation beobachtet werden, sondern ist auch entscheidend für die prosodische Realisierung der Sprachsegmente. Bezugnehmend auf den oben angeführten Wahrscheinlichkeitsbegriff, ist das eingetretene Ereignis in der zeitlichen Abfolge entweder ein Sprachsegment, oder ein Null-Ereignis, also ein Fehlen eines Sprachsegmentes, welches das Teilergebnis des sicheren Ereignisses ist.

ses A nicht unabhängig von dem Auftreten des Ereignisses B ist, spricht man von bedingten Wahrscheinlichkeiten.

„Unter der bedingten Wahrscheinlichkeit $P(A|B)$ versteht man die Wahrscheinlichkeit für das Eintreten des Ereignisses A unter der Bedingung, dass das Ereignis B bereits eingetreten ist. Sie ist gleich dem Quotienten aus der Wahrscheinlichkeit des Ereignisses $A \cap B$ und der Wahrscheinlichkeit des Ereignisses B, wobei $P(B) > 0$ vorausgesetzt werden muss“ (nach: Engeln-Müller et al. 2001). Für die Wahrscheinlichkeitsrechnung gelten die folgenden Rechenregeln:

Verbundwahrscheinlichkeit berechnet sich nach: $P(A \cap B) := \lim_{n \rightarrow \infty} \frac{H_n(A \wedge B)}{H_n}$

Bedingte Wahrscheinlichkeit berechnet sich nach: $P(A|B) := \lim_{n \rightarrow \infty} \frac{H_n(A \wedge B)}{H_n(B)} = \frac{P(A \cap B)}{P(B)}$

Aus diesen Rechenregeln wird nachfolgend der Satz von Bayes abgeleitet.

3.1.1 Satz von Bayes

Das Bayes-Theorem, oder auch der Satz von Bayes, ist ein Ergebnis der Wahrscheinlichkeitstheorie, benannt nach dem Mathematiker Thomas Bayes. Es gibt an, wie man mit bedingten Wahrscheinlichkeiten rechnet. Für zwei Ereignisse A und B lautet es (3)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (3)$$

Hierbei ist $P(A)$ die A-Priori-Wahrscheinlichkeit für ein Ereignis A und $P(B|A)$ die Wahrscheinlichkeit für ein Ereignis B unter der Bedingung, dass A auftritt. Die Korrektheit des Satzes folgt unmittelbar aus der Definition der bedingten Wahrscheinlichkeit

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(A \cap B) = P(A|B) \cdot P(B) \quad (4),$$

da analog gilt

$$P(B | A) = \frac{P(B \cap A)}{P(A)} \Leftrightarrow P(B \wedge A) = P(B | A) \cdot P(A) \quad (5)$$

folgt wegen

$$P(B \wedge A) = P(A \wedge B) \quad (6),$$

der Zusammenhang

$$P(A | B) \cdot P(B) = P(B | A) \cdot P(A) \quad (7).$$

Durch Auflösen bekommt man obigen Satz von Bayes.

3.2 Endliche Automaten

Ein endlicher Automat ist dadurch charakterisiert, dass er eine endliche Anzahl von Zuständen hat. Endliche Automaten haben eine breite Anwendungsmöglichkeit in unterschiedlichen Gebieten der Sprachverarbeitung, wie Part-of-Speech-Tagging oder Parsing.

Ein Spezialfall eines endlichen Automaten ist die Markov-Kette (oder auch Markov-Prozess). Die Markov-Kette ist ein spezieller stochastischer Prozess. Man unterscheidet eine Markov-Kette im diskreten und eine im stetigen Fall. Sei $T \subseteq \mathbb{N}$ eine Indexmenge, hier z.B. die Position des Worts im Text oder die Position des Phonems, $(X_t)_{t \in T}$ eine Familie von Zufallsvariablen mit Werten in X . Dann heißt X_t stochastischer Prozess. Ein stochastischer Prozess beschreibt eine Sequenz von Zufallsereignissen, die nicht unabhängig voneinander sind und deren Werte von vorangegangenen Ereignissen der Sequenz abhängen. Ein stochastischer Prozess ist also eine Folge von elementaren Zufallsereignissen $X_1, X_2, X_3, \dots, X_i \in \Omega, i = 1, 2, 3, \dots$, wobei die Zufallswerte in einem stochastischen Prozess Zustände des Prozesses heißen (Manning et al. 2001) und Ω die Menge aller Ereignisse beschreibt.

In Abbildung 10 wird eine Markov-Kette schemenhaft dargestellt, bei der jeweils ein Sprachsegment durch ein Phonem repräsentiert ist. Die Markov-Kette wird üblicherweise mit Hilfe einer Matrix dargestellt, wobei die Startwahrscheinlichkeit und die Berechnung der Übergangswahrscheinlichkeiten zu einer Gesamtwahrscheinlichkeit summiert werden.

$$\pi_i = P(X_1 = s_i), \sum_{i=1}^N \pi_i = 1 \quad (8)$$

π_i repräsentiert die Start- oder Einsprungswahrscheinlichkeit. a_{ij} gibt die Gesamtwahrscheinlichkeit der aufsummierten Übergänge zwischen den einzelnen Zuständen wieder, wobei die Summe über j gleich 1 ist und a_{ij} grösser Null sein muss.

$$a_{ij} = P(X_{t+1} = s_j | X_t = s_i), \sum_{j=1}^N a_{ij} = 1, a_{ij} \geq 0 \quad (9)$$

Bei der Markov-Kette ist jeder Zustand sichtbar, und die Übergangswahrscheinlichkeiten, die einen stationären Prozess charakterisieren, erfüllen diese Bedingung.

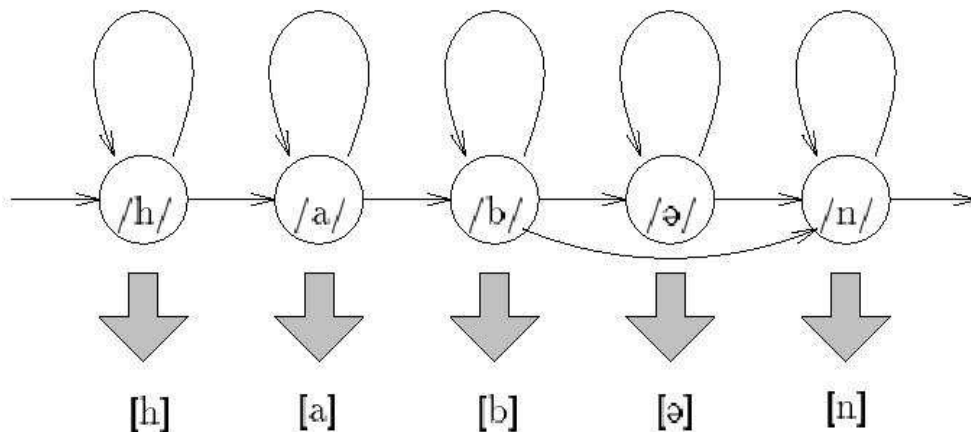


Abbildung 10: Beispielhafte Darstellung einer Markov-Kette, nach Schukat-Talamazzini 1995

Eine Markov-Kette heißt stationär, wenn für alle $i, j \in X$ die Übergangswahrscheinlichkeit $a_{ij} := P(X_{t+1} = s_j | X_t = s_i)$ unabhängig von t ist. Diese Eigenschaft wird auch als Markov-Eigenschaft bezeichnet. So besagt die Markov-Eigenschaft einer Markov-Kette 1. Ordnung, dass ein Zustand nur vom vorangehenden Zustand abhängt. Die Übergangswahr-

scheinlichkeiten a definieren zusammen mit den Anfangsbedingungen diskrete Prozesse, die als Markov-Ketten bezeichnet werden. Die Ausgabeverteilung ist nur durch den aktuellen Zustand bedingt.

Es gibt Markov'sche Ketten

1. Ordnung: das aktuelle Zeichen ist nur vom direkt vorhergehenden abhängig

2. Ordnung: das aktuelle Zeichen ist abhängig von den letzten 2 Zeichen

3. Ordnung: das aktuelle Zeichen ist abhängig von den letzten 3 Zeichen

n-ter Ordnung: das aktuelle Zeichen ist abhängig von den letzten n Zeichen

Die Wahrscheinlichkeit eines Zeichens, das vom vorhergehenden Zeichen abhängig ist, berechnet sich nach der in Abschnitt 3.1 angegebenen Formel für bedingte Wahrscheinlichkeit.

3.3 Hidden-Markov-Modelle

Hidden-Markov-Modelle (HMMs) werden seit langer Zeit erfolgreich in den unterschiedlichsten Anwendungen eingesetzt, wobei die Spracherkennung wohl zu dem populärsten Einsatzgebiet von HMMs zählt (Rabiner 1993). Ein Hidden-Markov-Modell ist ein probabilistischer endlicher Automat ähnlich der Markov-Kette, der zusätzlich in den einzelnen Zuständen Ausgaben produziert und die Zustände im Gegensatz zu den Markov-Ketten nicht beobachtbar sind. Die Zustände sind verborgen (hidden), weshalb von außen betrachtet es sich nicht sagen lässt in welchem Zustand sich das Modell befindet und nur die Beobachtung sichtbar ist. Über die Zustände können nur Annahmen gemacht werden.

Das HMM ist ein generatives Modell, welches die Wahrscheinlichkeiten der Zustände und ihrer Beobachtung gemeinsam modelliert, so dass eine Verbundwahrscheinlichkeit $P(o, s)$ errechnet wird, $o \in O$ bezeichnet die Beobachtung und $s \in S$ den dazugehörigen Zustand. Ein HMM besteht aus der Übergangswahrscheinlichkeit von einem Zustand zum nächsten, $P(s_t | s_{t-1})$ was eine Aussage darüber macht wie beide Zustände zueinander in

Beziehung stehen und die Beobachtungswahrscheinlichkeit $P(o|s)$, die das Verhältnis von Beobachtung O zu den jeweils verborgenen Zuständen S macht. Bei den meisten Anwendungen in der Sprachverarbeitung mittels Hidden-Markov-Modellen repräsentieren die Zustände S die Einheiten und die Beobachtungen O Merkmalsvektoren dieser.

Es gibt verschiedene Modell-Topologien für HMMs: das Links-Rechts-Modell, das Backus-Modell sowie das lineare Modell. In Abbildung 11 sieht man 2 mögliche Modell-Topologien. In der in dieser Arbeit entwickelten Sprachsynthese wird das ergodische Modell verwendet. Ein Hidden-Markov-Modell wird spezifiziert durch die Angabe des Fünftupels (S, O, Π, A, B) (Manning et al. 2001):

- $S = \{s_1, \dots, s_N\}$ endliche Menge von Zuständen
- $O = \{o_1, \dots, o_M\} = \{1, \dots, M\}$ endliche Menge von Ausgabe-Symbolen
- $\Pi = \{\pi(i)\}, i \in S$ Menge der Wahrscheinlichkeiten der Startzustände
- $A = \{a(i, j)\}, i, j \in S$ Wahrscheinlichkeiten der Zustandsübergänge

$$\sum_{j=1}^N a_{ij} = 1$$

- $B = \{b(i, j, k)\}, i, j \in S, \phi \in O$ Wahrscheinlichkeiten der Symbolemissionen

$$\sum_{o=1}^M b_{ijo} = 1$$

In der Spezifikation von Manning et al (2001) wird ein „arc-emission-Modell“ angenommen, d.h. ein Modell, bei dem die Ausgabe beim Übergang von Zustand i zu Zustand j erfolgt. Beim „state-emission-Modell“ ist die Ausgabe jeweils mit dem Zustand j verbunden. Die Spezifikation für die Ausgabewahrscheinlichkeit für das „state-emission-

Modell“ lautet: $B = \{b_{jk}\}, i, j \in S, o \in O, \sum_{k=1}^M b_{jk} = 1$ (10)

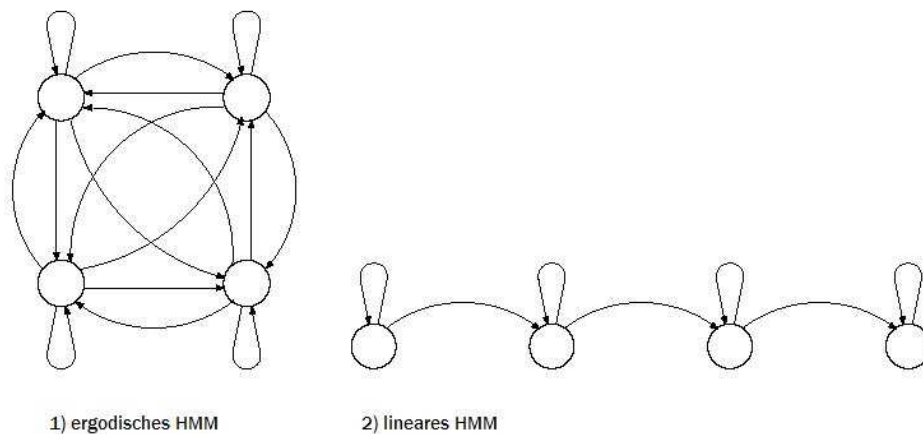


Abbildung 11: Darstellung elementarer Hidden-Markov-Modell Topologien

Praktisch alle Anwendungen von HMMs beruhen auf der Beantwortung von folgenden Fragen:

- Gegeben ist die Beobachtungssequenz $O = \{O_1, O_2, O_3, \dots, O_T\}$ und ein Modell λ . Wie wahrscheinlich ist eine Beobachtung O in einem gegebenen Modell λ ? Diese Fragestellung ist verbunden mit dem Forward-Backward-Algorithmus, der $P(O|\lambda)$ über alle möglichen Pfade berechnet.
- Gegeben ist eine Beobachtungssequenz $O = \{O_1, O_2, O_3, \dots, O_T\}$ und ein Modell λ . Welches ist die beste Zustandsfolge $a^* = \{a_1^*, a_2^*, a_3^*, \dots, a_T^*\}$, die eine optimale Lösung liefert? Die Suche nach dem optimalen Pfad über die gesamte Beobachtungssequenz wird meist mit dem Viterbi-Algorithmus (s. Abschnitt 5.6.1) gelöst.
- Wie werden die Modellparameter $\lambda = (A, B, \pi)$ gewählt, so dass die Wahrscheinlichkeit $P(O|\lambda)$ ein Maximum erreicht? Welcher Algorithmus hier eingesetzt wird, hängt von der gewählten Strategie ab.

Bei der Aufgabenstellung „Automatische Trainingsverfahren für die Unit-Selection-basierte Sprachsynthese“, ist es das Ziel, die am besten geeigneten Sprachsegmentbausteine aus einem Sprachdaten-Korpus auszuwählen, welche die Zielvorgaben hinsichtlich Prosodie und spektraler Eigenschaften maximal approximieren. Im diese Falle ist die Lösung der Fragestellung 2, was ist die wahrscheinlichste Erklärung für die Beobachtung O

$= \{O_1, O_2, O_3, \dots, O_T\}$ unter einem Modell λ , entscheidend. Die Suche der geeigneten Sprachsegmente führt zu dem wahrscheinlichsten Pfad durch die Zustände, welche die gegebene Beobachtung, in diesem Fall die zu synthetisierende Äußerung am besten wiedergeben. Formal ist dies zu erreichen mit:

$$U = \arg \max_s P(O | S) \quad (11)$$

wobei $O = \{O_1, O_2, O_3, \dots, O_T\}$ die Beobachtungen, also die Sprachsegmente, welche zu synthetisieren sind, repräsentiert und $S = \{s_1, s_2, s_3, \dots, s_T\}$ die Zustände, also die Sprachbausteine im Korpus wiedergibt und U die konkatenierten Einheiten.

Abbildung 12 zeigt ein Hidden-Markov-Modell wie es in der HTK Software verwendet wird. Die Beobachtungssequenz $O = \{O_1, O_2, O_3, \dots, O_T\}$ kann beispielsweise, wie in Kapitel 5 näher betrachtet, eine Abfolge spektraler Merkmalsvektoren sein, aus denen dann ein akustisches Signal generiert wird. Die Zustände $S = \{s_1, s_2, s_3, \dots, s_T\}$ sind dann die kontextabhängigen Merkmalsvektoren, welche die Sprachsegmente beschreiben und die eben die entsprechende Ausgabe emittieren und a repräsentiert die Übergangswahrscheinlichkeit der Zustände.

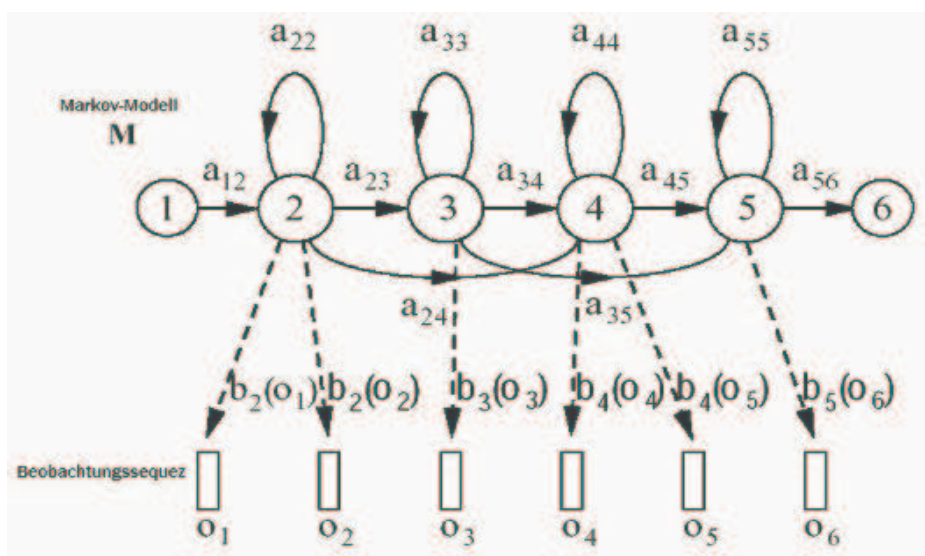


Abbildung 12: Darstellung eines typischen Hidden-Markov-Modells wie es in der HTK Software Verwendung findet (HTK Handbook 2002, S. 13)

3.4 Entropie

Der Ursprung des Wortes Entropie liegt im Griechischen und bedeutet so viel wie „sich zu etwas hinwenden“.

Der Begriff der Entropie kommt ursprünglich aus der Physik und wurde speziell in der Thermodynamik entwickelt und bedeutete dort etwas wie den Grad der Unordnung. Natürliche Prozesse streben in der Regel einem Gleichgewichtszustand zu. Ist die Gleichverteilung erreicht, ist die Entropie 1. In sich selbst überlassenen geschlossenen Systemen nimmt die Entropie ständig zu. Die Entropie wächst mit der Wahrscheinlichkeit des Zustands. Die Entropie ist dementsprechend ein Maß für die Menge an Zufallsinformationen, die in einem oder mehreren Zufallsereignissen oder einer Informationsfolge steckt. Auf Sprachäußerungen bezogen wird die Entropie ein Maß für den Informationsgehalt.

Die geschichtliche Entwicklung der Informationstheorie begann mit den Arbeiten von Hartley 1928. Als Begründer der Informationstheorie gilt jedoch Claude Shannon (1948). Die Informationstheorie gibt Antworten auf die Fragen: Wie viel Information steckt in einer Nachricht und wie misst man Information? Wie kann ein System optimiert werden, damit es möglichst viel Information pro Zeiteinheit überträgt bzw. verarbeitet und wie viel Information kann maximal übertragen bzw. verarbeitet werden? Der Begriff der Information wird hier wahrscheinlichkeitstheoretisch definiert. Wird Information gesendet, heißt das, dass Unsicherheit beseitigt wird, wobei Unsicherheit durch die Wahrscheinlichkeit wiedergegeben ist. Ist ein Ereignis „sicher“, kann dieses Ereignis mit „kein Zufall“ gleichgesetzt werden, woraus folgt, dass das Ereignis mit Sicherheit bestimmt werden kann. Die Wahrscheinlichkeit eines „sicheren Ereignisses“ ist $P(x) = 1$.

Etwas vorher nicht Bekanntes ist nach Zuführung der Information bekannt. Vor dem Empfang der Information gibt es eine Menge von Ereignissen, die eintreten könnten (jedes mit einer gewissen Wahrscheinlichkeit). Man weiß aber nicht, welches dieser Ereignisse eintreten wird. Erst nach Empfang der Information (=Eintreten eines Ereignisses) weiß man, welches Ereignis eingetreten ist. Die Entropie ist ein Maß für die Wahrschein-

lichkeit, mit dem einen System im Zustand i „im nächsten Augenblick“ in den Zustand j übergehen wird.

Der mittlere Informationsgehalt, die Entropie H , gibt die Anzahl von Binärentscheidungen an, die man im Mittel zur Kennzeichnung eines Zustands braucht, wenn die einzelnen möglichen Zustände verschiedene Wahrscheinlichkeiten $P(x)$ haben.

$$H(x) = - \sum_{x \in X} P(x) \log_2 P(x) \quad (12)$$

$H(x)$ kennzeichnet die Entropie und $P(x)$ die Wahrscheinlichkeit des Zustandes. Der Logarithmus Dualis \log_2 kennzeichnet die binäre Ausdrucksweise der Entropie, welche sich in *bits* ausdrückt.

3.4.1 Bedingte Entropie

Bei der Entropie von Symbolfolgen, die als Markov'sche Ketten beschreibbar sind, wird berücksichtigt, dass die Zeichen, die von einer Quelle nacheinander gesendet werden, nicht zusammenhanglos, bunt durcheinander gewürfelt daherkommen, sondern in einer Beziehung zueinander stehen. Als Beispiel kann eine Nachrichtenquelle dienen, die deutschen Text übermittelt. Hierbei sind die Zeichenfolgen nicht beliebig, sondern folgen den Gesetzmäßigkeiten (Phonotaktik und Grammatik) der deutschen Sprache. Daher sind die aufeinanderfolgenden Buchstaben voneinander abhängig und die Entropie sinkt (= höherer Ordnungsgrad). Im Grenzfall vollständiger Abhängigkeit, wenn nach dem ersten Zeichen bereits der gesamte folgende Text bestimmt wäre, wird keine weitere Information übermittelt und die Gesamtentropie entspricht der Entropie des ersten Zeichens. Die Auftretenshäufigkeit eines Zeichens ist nicht zufällig, sondern davon abhängig, welches bzw. welche Zeichen vorher gesendet wurden, z.B. folgt im Deutschen nach einem „q“ immer ein „u“ und nach „sc“ höchstwahrscheinlich ein „h“.

Die Entropie Markov'scher Ketten n -ter Ordnung wird bedingte Entropie oder Markov-Entropie (15) n -ter Näherung bezeichnet und leitet sich ab aus der Entropie (13) und der Anwendung der Kettenregel (14).

$$H(Y|x) = -\sum_{y \in Y} P(y|x) \log P(y|x) \quad (13)$$

$$H(X,Y) = H(X) + H(Y|X) \quad (14)$$

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(y|x) \quad (15)$$

Die bedingte Entropie einer Zufallsvariable X bei einem gegebenen Ereignis Y , bzw. einer Zufallsvariable Y , ist die Unsicherheit über X , die verbleibt, wenn Y bereits bekannt ist. Sind X und Y stochastisch unabhängig, dann bleibt die Entropie von X vollständig erhalten. Die Entropie fällt umso stärker, je mehr Zeichen berücksichtigt werden. Wenn die Anzahl der berücksichtigten Zeichen hinreichend groß gewählt wird, so konvergiert die Entropie gegen Null.

KORORPA UND AUTOMATISCHE VORVERARBEITUNG

Das nachfolgende Kapitel, Korpora und automatische Vorverarbeitung, befasst sich mit den Korpora, die für die Sprachsynthese und audio-visuelle Synthese verwendet werden. Es werden die Eigenschaften der Korpora erläutert, sowie die automatische Vorverarbeitung der Korpora. Die automatische Vorverarbeitung dient der Annotation der Sprach- und Videodaten, wie sie für die Erfordernisse der entwickelten Verfahren der audio-visuellen Sprachsynthese benötigt werden. Für die Sprachsynthese wird das benötigte Sprachdatenkorpus definiert sowie die Einheiten, welche für die konkatentative Synthese verwendet werden. Es wird die Segmentierung der aufgenommenen Sprachdaten beschrieben, sowie auf die Annotation des Sprachdatenkorpus eingegangen. Desweiteren werden die Verfahren vorgestellt, die für die symbolische Vorverarbeitung verwendet wurden. Die symbolische Vorverarbeitung für die Sprachsynthese umfasst die Graphem-Phonem Umsetzung, Vorhersage von Wortklassen, Akzent und Silbengrenzen. Für die audio-visuelle Synthese wird ein Überblick über das verwendete Videokorpus gegeben und die verwendeten visuellen Einheiten werden definiert, wie sie bei der Rekonstruktion des Videosignals verwendet werden. Es wird die Segmentierung der Videodaten in die benötigten Einheiten beschrieben, sowie eine Definition der Visemklassen erarbeitet, die bei der Graphem-Visem Umsetzung eingesetzt werden. Die automatischen Trainingsverfahren beruhen alle auf dem statistisch motivierten Maximum-Entropie-basierten Lernen.

4.1 Definition und Eigenschaften der Korpora

Bei dem in dieser Arbeit entwickelten Algorithmus für die Sprachsynthese werden Sprachsegmente aus dem zur Verfügung stehenden Korpus ausgeschnitten und entsprechend der Vorgaben neu zusammengesetzt. Dieser Ansatz lässt sich darauf zurückführen, dass Sprache im Allgemeinen und Sprache aus der Sicht der Signalgenerierung aus lautlichen Einheiten besteht, die entsprechend der gewünschten Zieläußerung im zeitlichen Ab-

lauf aneinandergereiht werden und so eine Äußerung ergeben. Es ist also möglich, aus den lautlichen Äquivalenten der diskreten Kategorien, den Phonemen einer Sprache, eine sprachliche Äußerung zu erzeugen. In der Praxis jedoch erweist sich dies als nicht realistisch, da jeder Laut abhängig von seiner Umgebung und der lautlichen Realisierung unterschiedliche Eigenschaften und Charakteristika hat und so ein nicht gut verständliches Sprachsignal erzeugen würde. Die lautlichen Einheiten einer Kategorie unterscheiden sich je nach Kontext in ihren prosodischen und spektralen Eigenschaften erheblich. Abbildung 13 zeigt die unterschiedliche prosodische Realisierung von „werden“. Das gleiche Wort hat eine unterschiedliche Dauer sowie eine unterschiedliche Formantstruktur, obwohl es an gleicher Position im Satz realisiert wurde. Die spektralen Unterscheide lassen sich leicht im Spektrum nachvollziehen. Auch die Intensität variiert. Nur der F0-Verlauf ist nahezu gleich. Je nach Kontext passt das eine „werden“ besser als das andere in die zeitliche Abfolge des Sprachsignals.

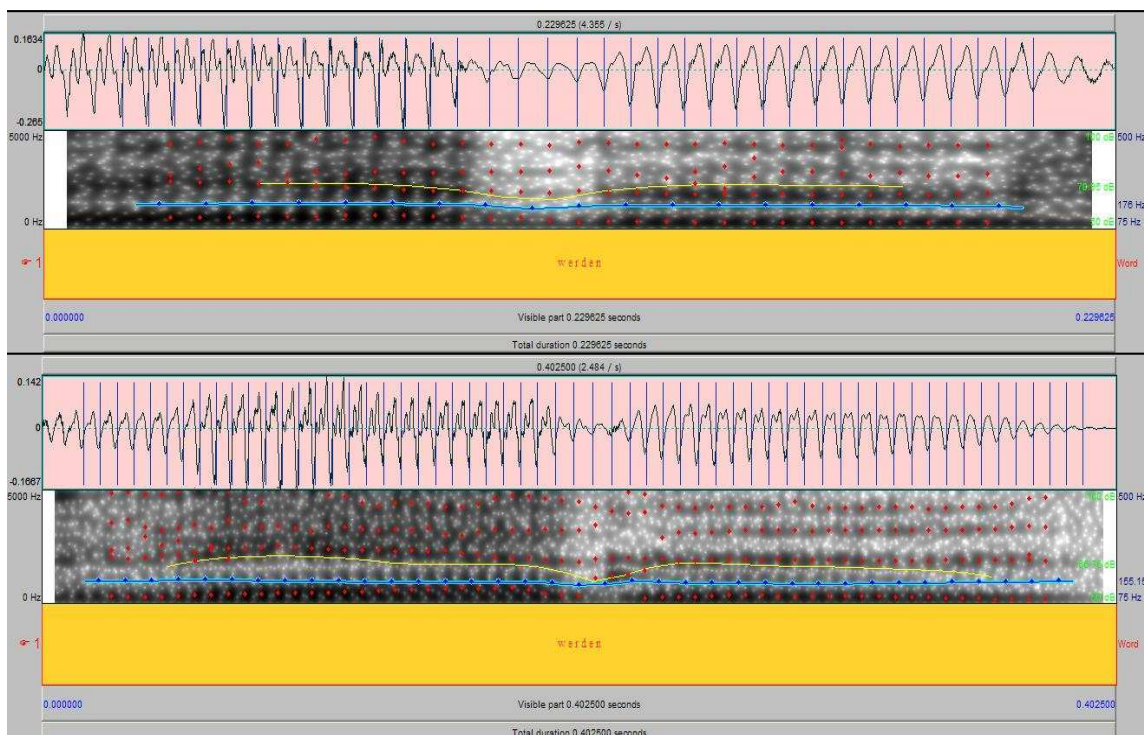


Abbildung 13: Darstellung der zwei Wörter „werden“ in unterschiedlichen Kontexten. Das obere Signal hat eine Dauer von 222 Millisekunden, das Untere eine Dauer von 402 Millisekunden. Weiterhin liegt die durchschnittliche Grundfrequenz oben bei 176Hz, unten bei 155Hz.

Zur Generierung von künstlicher Sprache wurden bei der Diphon-Synthese alle möglichen Diphon-Kombinationen, die durch das bestehende Phonemsystem generiert werden können, in Trägersätze integriert. Diese Trägersätze wurden dann von einem Sprecher mit möglichst neutraler Stimme und flacher Prosodie gesprochen und digital aufgenommen. Einschränkungen konnten durch die phonotaktischen Regeln der jeweiligen Sprache gemacht werden und so kann die Größe des aufgenommenen Korpus reduziert werden, da nicht jedes Diphon-Paar sich in der Sprache wiederfindet. Nachtäglich wurde mit Hilfe eines Signalmanipulationsalgorithmus der jeweilige Baustein an seine prosodischen Erfordernisse angepasst.

Bei der entwickelten Sprachsynthese Software „AVISS“ wird auf eine nachträgliche Signalmanipulation verzichtet. Die grundlegende Idee der auf der „variable-size non-uniform Unit-Selection“ basierten Sprachsynthese ist, Einheiten variabler Größe aus einem vorgegebenen Sprachdaten-Korpus auszuwählen, wobei die ausgewählten Sprachbausteine schon die erforderlichen Vorgaben hinsichtlich phonetischer und akustisch-spektraler Merkmale möglichst gut erfüllen. Variable Einheitengröße bedeutet hier, dass das Sprachsegment aus einem oder mehreren Wörtern, einer Silbe, einem Triphon oder Diphon bestehen kann. Die Auswahl folgt einer hierarchischen Struktur, wobei die Wortebene als oberste Auswahlebene angesehen wird und die Phonemebene als unterste Auswahlebene. Daraus ergibt sich, dass bei der Erstellung von Korpora prosodische Varianten der jeweiligen Kategorie vorhanden sein müssen. Jedoch ist bei der „variable-size non-uniform Unit-Selection“ mit variabler Bausteingröße nicht intuitiv abschätzbar, wie ein bestgeeignetes Korpus definiert werden soll, um alle möglichen potentiellen lautlichen Vorkommnisse einer Sprache abzudecken. Eine Technik, mit der eine fehlende prosodische Realisierung eines Sprachsegments ausgeglichen werden kann, basiert auf dem bekannten PSOLA-Algorithmus (Moulines et al. 1990). Mittels des PSOLA-Algorithmus werden die Dauer und der F0-Wert der lautlichen Einheit nachbearbeitet und an die entsprechenden Zielvorgaben angepasst. Der Leser sei hier auf die einschlägige Literatur zur Funktionsweise des PSOLA-Algorithmus verwiesen.

Zur Verdeutlichung sei hier noch einmal das in Abschnitt 1.1 erwähnte LNRE-Problem (large number of rare events) aufgegriffen. Die Verteilungskurve in Abbildung 14 zeigt,

dass wenige Einheiten häufig auftreten und die überwiegende Mehrheit der Einheiten geringe Auftrittswahrscheinlichkeit hat. Es hat sich gezeigt (Möbius 2000), dass paradoxerweise die Gesamtheit der seltensten Einheiten der natürlichen Sprache in Unit-Selection-basierten TTS-Systemen besonders häufig ausgewählt wird. So ist die Wahrscheinlichkeit des Auftretens einer bestimmten Einheit gering, doch in der kumulativen Wahrscheinlichkeit ist mindestens ein seltenes Ereignis in jedem zu generierenden Satz enthalten.

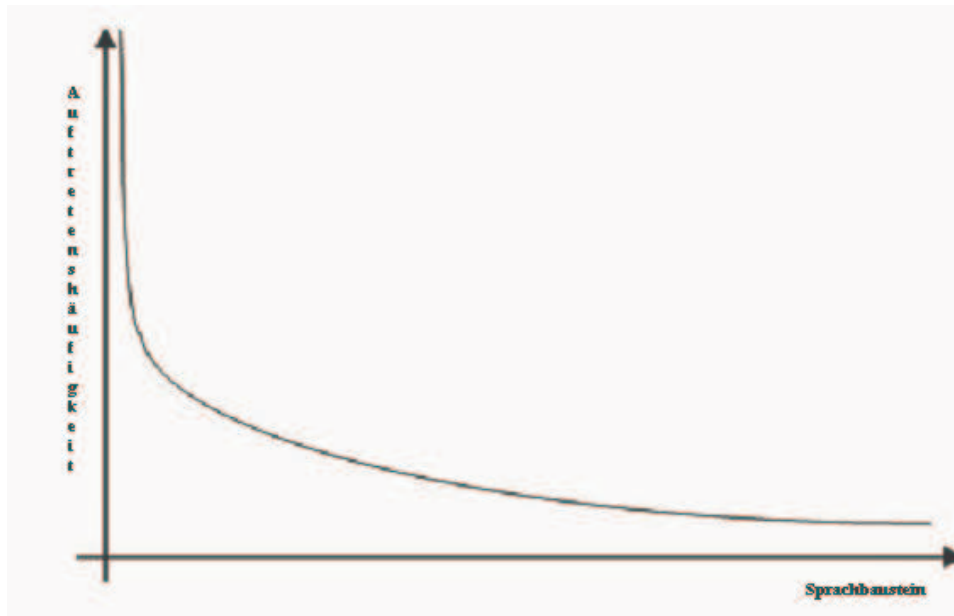


Abbildung 14: LNRE; Verteilungsfunktion von Einheiten: wenige Sprachbausteine treten sehr häufig auf, wobei die meisten Bausteine durch eine geringe Anzahl repräsentiert sind.

Dies ist oft der Grund für Diskontinuitäten im konkatenierten Sprachsignal. Beim Entwurf eines geeigneten Korpus für die Sprachsynthese ist eine bestmögliche Abdeckung aller sprachlichen Ereignisse das Ziel. Will man bei der Unit-Selection-basierten Sprachsynthese auf eine nachträgliche Manipulation der Sprachsignal-Einheiten verzichten, muss auf die prosodischen und spektralen Varianten der Sprachbausteine eingegangen werden und das Korpus entsprechend gestaltet und auf den Anwendungsbereich abgestimmt werden.

4.2 Annotation

Die Annotation der Sprachdaten beruht auf automatischen Verfahren. Die Annotation der

Sprachdaten wurde in ein XML-Format übertragen und folgt einer Attribut-Wert-Struktur. Tabelle 2 gibt einen Überblick der aufgeführten Werte und ihrer zugehörigen Attribute. X kennzeichnet, ob für das jeweilige Segment das entsprechende Attribut Verwendung findet.

	Satz	Wort	Silbe	Phonem
Satztyp	x	---	---	---
Orthografische Realisation	x	x	x	<i>x</i>
phon. Transkription	x	x	x	<i>x</i>
PoS	---	x	x	<i>x</i>
Akzent	---	---	x	---
Stimmhaft/-los	---	---	---	<i>x</i>
Dauer	---	x	x	<i>x</i>
F0	---	x (Mittelwert)	x (Mittelwert)	<i>x (Mittelwert)</i>
MFCC	---	<i>x</i>	<i>x</i>	<i>x</i>

Tabelle 2: Attribute der Sprachsegmente im Sprachdaten-Korpus

Ein Nicht-Vorhandensein des Attributes für das Segment wird durch „---“, gekennzeichnet. Die Attribute sind horizontal angeführt, die Segmente vertikal. Die Struktur folgt einer hierarchischen Gliederung, welche vom Satz ausgeht und die Einheiten Wort, Silbe und Phonem einbezieht. Durch diese Struktur können aus dem annotierten Korpus auch die Einheiten generiert werden, welche sich aus mehreren Einheiten zusammensetzen. Diese sind Wort-Bigramme, Triphon- und Diphon-Einheiten.

4.3 Korpora

4.3.1 Audiokorpus

Das für diese Arbeit verwendete Audio-Korpus wurde im Rahmen des Verbmobil-Projektes am Institut für Kommunikationsforschung und Phonetik aufgenommen. Das Korpus wurde von einer semi-professionellen Sprecherin gesprochen und digital mit 32kHz, 16bit aufgenommen. Als Vorgabe für das Korpusdesign diente ein Anwendungsszenario realer Geschäfts-Kommunikation. Es wurden Terminvereinbarungen und Reiseverbindungen sowie Szenarien der Hotelreservierung integriert. Das Korpus enthält 5000 Sätze gesprochener Sprache. Für die Entwicklung der Sprachsynthese, wie sie in der Applikation verwendet wird, wurden 3000 Sätze nach statistischer Vorauswahl benutzt. Tabelle 3 gibt einen Überblick über die Verteilung der Sprachsegmente im benutzten Korpus. In dem originalen Ausgangskorpus wurden Phonemeinheiten neu definiert, die im BOSS Sprachsynthese-System verwendet werden. Die Sprachsegmente auf Phonebene haben sich auf 63 Daten-Typen verringert. Das in dieser Arbeit entwickelten Sprachsynthese-System verwendet nicht alle Phonemeinheiten welche im Rahmen der Entwicklung des Sprachsynthesystems BOSS II definiert wurden. Die verwendeten Phonemeinheiten folgen dem Standard-SAMPA-Inventar sowie zusätzlichen Phoneinheiten, die aus BOSS II übernommen wurden, wie z.B. /@n/. Dies ist darauf zurückzuführen, dass das Graphem-Phonem-Transkriptionsmodul eine Standard-SAMPA-Graphem-Phonem-Konvertierung plus die zusätzlich verwendeten Einheiten durchführt.

	Bigramm	Wort	Silben	Triphon	Diphon	Phon
Daten-Typ	16068	6123	1497	15511	3271	63
Daten-Token	31958	61338	51570	37003	55505	111011

Tabelle 3: Übersicht über die Verteilung der Sprachsegmente

In Abschnitt 4.4.1, Tabelle 5 wird das verwendete Phonem-Inventar beschrieben sowie das angewendete Verfahren zur Graphem-Phonem-Umsetzung und die erzielten Ergebnisse der statistisch motivierten Transkription.

Entscheidend für die Qualität und Verständlichkeit eines Sprachsynthesystems sind die Aussprachefehler, die bei den Aufnahmen vom Sprecher gemacht werden. Diese führen zu einer falschen Umsetzung des orthografischen Textes in phonetisch transkribierten Text. Aus diesem Grund ist während der Aufnahmen sorgfältig darauf zu achten, dass der Sprecher die Aussprachefehler minimiert und bei einer falschen Aussprache die Aufnahme wiederholt wird. Die Aufnahmen sollten immer von einer geschulten Aufsichtsperson überwacht werden. Auch lässt die Konzentration eines Sprechers nach einiger Zeit nach. Daher sollten Pausen während der Aufnahmen eingehalten werden, um Fehler beim Ablesevorgang zu vermeiden.

4.3.2 Segmentierung des Audiokorpus

Segmentierung von Sprachsignalen in lautliche Einheiten ist ein zentrales Thema für die Qualität der Sprachsynthese. Da bei der konkatenativen Sprachsynthese lautliche Einheiten aneinandergereiht die gewünschte Zieläußerung als Sprachsignal generieren, ist das Einhalten von Lautgrenzen entscheidend. Die Segmentierung von einem Sprachsignal in Sprachsegmente ist ein zeit- und kostenintensiver Prozess, der Experten voraussetzt, die phonetische Kenntnisse haben, und auch dann gibt es bei menschlichen Experten keine 100%ige Konsistenz in der Sprachsignalsegmentierung. Aus diesem Grund wird auf automatische Segmentierung zurückgegriffen. Automatische Segmentierung ist in der Forschung immer wieder Thema. Für die Sprachsynthese bedeutet eine automatische Segmentierung eine enorme Zeitersparnis, auch wenn nach der automatischen Segmentierung noch eine manuelle Korrektur erfolgen muss.

Man unterscheidet zwei Arten von Ansätzen zur automatischen Lautsegmentierung. Dies sind zum einen der „klassifikationsbasierte Ansatz“ und zum anderen das „alignmentbasierte Verfahren“. Unter den klassifikationsbasierten Ansätzen ist ein weit verbreitetes Verfahren zur Lautsegmentierung die Verwendung von Hidden-Markov-Modellen. Das Audio-Korpus für die Sprachsynthese wurde mittels des klassifikationsbasierten Ansatzes

in Zeitbereichseinheiten segmentiert. Zur Anwendung kam die an der Universität Cambridge entwickelte Software HTK (Woodland, Young 1992). Nach Stöber (2002) wurde ein solches Verfahren zuerst von Talkin et al. (1994) angewendet. Eine Übersicht über das Thema Lautsegmentierung findet sich in Hosom (2000).

Lautsegmentierung kann als Nebenprodukt der Spracherkennung angesehen werden. Während die Spracherkennung gesprochene Sprache in Text umsetzt, wird bei der Lautsegmentierung die zeitliche Dimension der Laute erfasst. Für die Lautsegmentierung muss für die Sprachdaten deren orthografische Realisation bekannt sein. Bei Lautsegmentierungsverfahren müssen gegebenenfalls auch die Aussprachevarianten berücksichtigt werden. Dies trifft auf Spontansprache zu, da hier oft Variationen der Aussprache bzw. Reduktionen verwendet werden. Wie in der Spracherkennung wird für die Lautsegmentierung ein Lauthypothesengraf verwendet (Stöber 2002), der die möglichen Lautfolgen repräsentiert. Zwischen dem Start- und Endknoten wird derjenige Pfad gewählt, der die maximale Wahrscheinlichkeit aufweist. Diese Wahrscheinlichkeiten können mittels Hidden-Markov-Modellen modelliert werden und ergeben somit den Pfad mit der maximalen Beobachtungswahrscheinlichkeit hinsichtlich Signal und Lautfolge sowie deren Segmentgrenzen. Wie in der Spracherkennung wird das Lauthypothesengraf-Verfahren als „forced alignment“ bezeichnet. Die Beurteilung der Ergebnisse der Lautsegmentierung kann durch den Betragsmittelwert der Abweichungen von manuellen und automatisch gesetzten Lautgrenzen errechnet werden. Ein anderes Maß ist der AKZM-Wert (Abweichung Kleiner Zwanzig Millisekunden) (Stöber 2002). Dieses Verfahren misst die Anzahl der Lautgrenzen, die weniger als 20 Millisekunden von den manuell gesetzten Lautgrenzen abweichen. Hier sind im Mittel Werte von 84% und darüber ein vernünftiges Ergebnis (vgl.: Stöber 2002, Hosom 2000). Eine nähere Betrachtung der Lautsegmentierung liegt außerhalb der Themenstellung dieser Arbeit. Es wird auf die ausführliche Beschreibung von Lautsegmentierung in Stöber (2002) verwiesen.

4.3.3 Videokorpus

Das für das entwickelte audio-visuelle System aufgenommene Video-Korpus folgt der Vorgabe, einen Sprecher aufzunehmen, bei dem der Sprecher von Kopf bis zu den Schul-

tern zu sehen ist. Diese Aufnahmeeinstellung ist der eines Nachrichtensprechers angenähert. Die Videoaufnahmen wurden mit einer Bildwiederholrate von 25 Einzelbildern pro Sekunde aufgenommen und mit einer Auflösung von 768 x 512 Bildpunkten abgespeichert. Die Daten wurden mit dem Indeo Video 5.1 komprimiert. Das Gesamtvideodatenmaterial beträgt ca. 3 Gigabyte. Es werden ca. 35 Minuten an Videoaufnahmen für diese Arbeit verwendet. Es wurden Trägersätze aus dem für die Sprachsynthese verwendeten Sprachdatenkorpus verwendet. Die Auswahl der geeigneten Sätze folgte einer statistisch motivierten Rangfolge, bei der die Sätze am besten bewertet wurden, die die meisten Viseme beinhalten. Die statistisch motivierte Auswahl der Trägersätze berechnet sich folgendermaßen: die relative Häufigkeit des Visems im Korpus multipliziert mit der Auftretensanzahl im Satz. Dies ergibt eine Kennzahl für ein spezielles Visem in einem Trägersatz. Pro Trägersatz wird diese Kennzahl für jedes Visem errechnet und aufaddiert. Dies ergibt eine Kennzahl für den Satz. Anhand der Satz Kennzahl werden dann die Sätze mit der größten Satz Kennzahl ausgewählt. Durch diesen Ansatz wurde ermöglicht, dass alle zur Laufzeit auftretenden Visemöglichkeiten abgedeckt werden können.

4.3.4 Segmentierung des Videokorpus

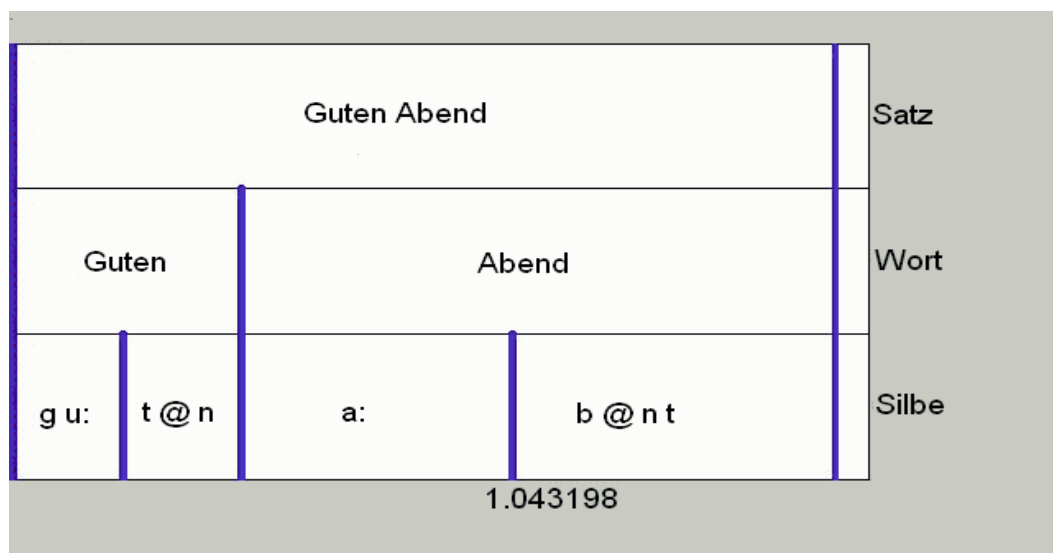


Abbildung 15: Darstellung Einheiten-Grenzen (Satz, Wort, Silbe)

Die Segmentierung der Videodaten beruht auf den Zeitmarken der Sprachdaten. Die Segmentierung der Sprachdaten wurde mit Hilfe des „Software-Toolkits“ Praat⁵ manuell durchgeführt und ist in Abbildung 15 beispielhaft dargestellt. Aus den Videoaufnahmen wurden die Audiospur und die Videospur extrahiert. Dies wurde mit der für diese Arbeit entwickelten audio-visuellen Synthese-Software AVISS (siehe Abschnitt 7.1) vollzogen. Die Audiospur wurde dann in die Einheiten Wort, Silbe und Phonem unterteilt. Die Start-Zeitmarke und die End-Zeitmarke der Sprachdaten-Einheiten liefern die Schnittstelle für die Videosegmente. Hierzu wurde folgende Umrechnung verwendet:

$$V = (E(\text{Seg}_i) \cdot 25) - (S(\text{Seg}_j) \cdot 25) ,$$

wobei V die Anzahl der Videoframes ist und $E(\text{Seg}_i)$ die Endzeit des Sprachsegments in Sekunden wiedergibt und $S(\text{Seg}_j)$, die Startzeit des Sprachsegments ebenfalls in Sekunden. Mit dem Wert der Startzeit werden beginnend die Startnummer des Start-Video-Frames errechnet und die Anzahl der Video-Frames entsprechend der lautlichen Äußerung. Hier wird vorausgesetzt, dass das Video mit 25 Bildern pro Sekunde codiert wurde und dadurch bei nicht eindeutigen Videobildgrenzen der Wert abgerundet bzw. aufgerundet wird.

4.3.5 Viseme

Viseme bezeichnen die äußerlich sichtbaren Artikulatoren-Einstellungen, die zum Realisieren eines bestimmten Lautes nötig sind. Tabelle 4 gibt einen Überblick über die unterscheidbaren Artikulatoren-Stellungen mit ihrer spezifischen Visem-Klasse. Die Visem-Klassen werden nach dem in Abschnitt 4.4.3 beschriebenen Verfahren erstellt. Ausgehend von dem in dieser Arbeit eingesetzten Phoneminventar wurde ein Inventar von Visem-Klassen entwickelt, welches die Laute in ihre visuellen Klassen einteilt und somit nutzbar macht für eine Umsetzung der Grapheme in ihre entsprechenden Viseme (Weiss 2005).

4.4 Automatische Vorverarbeitung

Die Annotation des Sprachdatenkorpus und des Videodatenkorpus enthält neben der or-

⁵ <http://www.praat.org>

thografischen Repräsentation, den Wortklassen, Akzent, den Zeitmarken und akustischen Parametern zusätzlich die phonetisch transkribierte Entsprechung des Textes.
















Viseme	Visuelle Repräsentation	Viseme	Visuelle Repräsentation
<i>P</i>		<i>C</i>	
<i>T</i>		<i>E</i>	
<i>N</i>		<i>A</i>	
<i>M</i>		<i>O</i>	
<i>F</i>		<i>U</i>	
<i>S</i>		<i>Q</i>	
<i>Z</i>		<i>Y</i>	
<i>R</i>			

Tabelle 4: Übersicht über die Lippenstellung bei den einzelnen Visem-Klassen. Siehe hierzu auch Tabelle 10: Phonem-Visem-Zuordnung

Damit eine hierarchische Segmentauswahl während der Synthese erfolgen kann, werden die Segmente in der annotierten Form zusätzlich in ihre silbische Struktur zerlegt. Nachfolgend werden die Verfahren beschrieben, mittels derer die automatische Umsetzung von orthografischem Text in phonetisch transkribierten erfolgt. Weiterhin werden die automatische Wortklassenvorhersage, die Akzentprädiktion und die Silbengrenzen-Erkennungsverfahren beschrieben.

4.4.1 Maximum-Entropie-basierte Graphem-Phonem-Transkription

Die Umsetzung von Graphemen in die entsprechenden phonetischen Klassen ist eine der

entscheidenden Aufgaben, die das Textvorverarbeitungsmodul eines Sprachsynthese-Systems lösen muss. Mit der richtigen Umsetzung von Graphemen in Phoneme steigt und fällt die Qualität eines solchen Systems. Wird ein Eingabetext nicht in die entsprechende phonetische Lautschrift transkribiert, werden die Einheiten nicht im Sprachdaten-Korpus gefunden oder falsch ausgewählt, was zu einer Minderung der Verständlichkeit der Sprachsynthese führt.. Die Umsetzung von Graphemen in Phoneme wurde mit Hilfe des Standard-SAMPA-Inventars für Deutsch (Anhang C) durchgeführt, welches um Phonemeinheiten aus der Einheitendefinition von BOSS II erweitert wurde. Tabelle 5 gibt einen Überblick über die Phonemeinheiten zur Graphem-Phonem-Umsetzung.

2:	2:6	6	9	96	@	@n	C	E	E6
(71)	(11)	(746)	(27)	(3)	(1518)	(1369)	(483)	(653)	(157)
E:	E:6	I	I6	I:6	N	O	O6	OY	S
(135)	(6)	(866)	(31)	(1)	(353)	(295)	(66)	(94)	(669)
U	U6	U:	Y	Y6	Z	a6	a:	a:6	aI
(435)	(54)	(1)	(141)	(24)	(10)	(50)	(560)	(89)	(624)
aU	b	d	e:	e:6	f	g	h	i	i:
(162)	(694)	(591)	(420)	(51)	(1056)	(914)	(341)	(3)	(488)
i:6	j	k	ks	l	m	n	o:	o:6	p
(39)	(121)	(822)	(1)	(1285)	(787)	(1856)	(306)	(97)	(497)
r	s	t	ts	u:	u:6	v	x	y:	y:6
(1165)	(1614)	(3144)	(13)	(311)	(18)	(505)	(246)	(83)	(12)
z	&								
(473)	(6853)								

Tabelle 5: Verwendete Phonemeinheiten mit Verteilung im Trainingskorpus

Es gibt eine Vielzahl von Ansätzen, die diese Aufgabe zufrieden stellend lösen. Neben regelbasierten Ansätzen existieren datenbasierte Ansätze, die auf bekannte maschinelle

Lernverfahren wie Entscheidungsbaumlernen, Neuronale Netzwerke oder N-Gramm-Modelle zurückgreifen. In dieser Arbeit wurde eine Graphem-Phonem-Transkription entwickelt, die auf dem bekannten Maximum-Entropie-Ansatz beruht. Maximum-Entropie-basiertes maschinelles Lernen wurde erfolgreich für unterschiedliche Aufgaben im Bereich der natürlichen Sprachverarbeitung eingesetzt. Berger et al. (1996) benutzten diesen Ansatz für die maschinelle Übersetzung. Ratnarparkhi (1998) entwickelte auf Basis von Maximum-Entropie u.a. einen Part-of-Speech Tagger und trainierte ein Modell zur Erkennung von Satzgrenzen. Der Ansatz von Ratnarparkhi diente in dieser Arbeit dazu, ein Modell für die Graphem-Phonem-Transkription zu entwickeln.

Die Konvertierung von Graphemen in Phoneme kann als klassisches überwachtes Lernen angesehen werden, bei dem jedes Phonem einer Klasse entspricht und jedes Graphem einer bestimmten Klasse zugeordnet werden muss. Graphem-Phonem-Transkription kann also als Klassifikationsproblem angesehen werden mit dem Ziel eine Funktion $C : X \rightarrow Y$ zu haben, die jedem Element $x \in X$ seine korrekte Klasse $y \in Y$ zuweist, wobei x die Grapheme eines Alphabets X und y die Phoneme eines Phoneminventars Y repräsentiert, $x \in \{a, b, c, \dots, z\}$ und $y \in \{a, a:, ?a, b, x, X, \dots, ts\}$. Um eine geeignete Klassifikation zu erreichen muss der Kontext $q \in Q$ des umgebenden Graphems mit einbezogen werden, um die richtige Klasse $k \in K$ zu präzisieren. Der Kontext Q mit $q \in \{a, b, c, \dots, z\}$ bezeichnet eine rechtsseitige Folge von Graphemen sowie eine linksseitige Folge von Graphemen, die das eigentliche Graphem, welches in ein Phonem umgesetzt werden soll, umgeben.

Der Kontext Q ist frei wählbar. Hier wird ein Kontext von zwei Graphemen links und zwei Graphemen rechts von dem aktuellen Graphem verwendet. Dies impliziert, eine bedingte Wahrscheinlichkeit p mit einzubeziehen, so dass die Wahrscheinlichkeit $p(k/q)$ ermittelt werden kann. Das Modell wurde mit einem Trainingskorpus trainiert, das 6500 Wörter mit ihrer phonetischen Entsprechung beinhaltet. Das Trainingskorpus wurde manuell überprüft. Insgesamt enthält das Trainingskorpus 36384 Ereignisse, wobei als Ereignis hier ein Graphem-Phonem-Paar angesehen wird. Zum Überprüfen der korrekten Transkription wurde das aus der Spracherkennung bekannte WER (Word Error Rate)

Maß auf PER (Phoneme Error Rate) übertragen. Die Fehlerrate der falsch zugeordneten Phoneme kann man nun berechnen mit:

$$PER \text{ in } \% = 100 \cdot \left(1 - \frac{\# \text{ richtig erkannte Phoneme}}{\# \text{ Phoneme gesamt}}\right)$$

Hierzu wurden 100 Wörter aus dem Trainingskorpus entnommen und mittels der trainierten Modelle neu transkribiert, sowie 100 Wörter, die nicht im Korpus enthalten sind. Die Transkription wurde mit den manuell korrekt transkribierten Wörtern automatisch abgeglichen. Tabelle 6 zeigt die Ergebnisse.

G-P Trainingskorpus:	82500 Zeichen
PER in %	8.72 %

Tabelle 6: Ergebnisse der korrekten Klassifizierung von Graphem-Phonem

4.4.2 Wortklassen-Bestimmung, Silbengrenzen-Erkennung, Akzent-Prädiktion

Neben der Graphem-Phonem-Transkription ist die Merkmalsgewinnung zur Beschreibung der Eigenschaften der Sprachsegmente eine wichtige symbolische Vorverarbeitung der Trainingsdaten, aus denen im weiteren Verlauf statistische Modelle entwickelt werden, mittels derer eine möglichst genaue Annäherung der ausgewählten Sprachsegmente an die Zielvorgaben zur Auswahl erreicht wird. Diese Merkmale müssen auch zur Laufzeit zur Verfügung stehen. Da diese Merkmale nicht wie die quantitativen Merkmale Position im Satz, Position im Wort etc. ermittelt werden können, sondern statistisch geschätzt werden müssen, werden für diese Merkmale nachfolgend Modelle trainiert, die die Prädiktion dieser Merkmale zur Sprachsegmentbeschreibung im Sprachdaten-Korpus wie auch zur Laufzeit vornehmen. Auch bei der Bestimmung von Wortklassen, Silbengrenzen-Erkennung und Akzent-Prädiktion wurde auf das Maximum-Entropie-Framework zurückgegriffen und entsprechende Modelle trainiert. Wortklassen (Engl: Part-of-Speech Tags, POS) teilen Wörter ihren linguistischen Klassen zu. Beispiel: Die Wörter des Satzes

„Die Weltmeisterschaft findet 2006 in Deutschland statt“ werden in folgende Wortklassen (angegeben in [...]) eingeteilt:

Die [ART] Weltmeisterschaft [NN] findet [VVFİN] 2006 [CARD] in [APPR] Deutschland [NE] statt [VVPP]

In dieser Arbeit wurde das Stuttgart-Tübinger Wortklassen-Inventar (siehe Anhang A) verwendet, um die Wortklassen im Sprachdaten-Korpus, wie auch zur Laufzeit des audiovisuellen Synthese-Systems zu bestimmen. Für das PoS-Tagging wurde die Korrektheit ermittelt durch:

$$\text{Korrekte Klassifikation in \%} = \frac{\# \text{ richtig erkannte PoS Tags}}{\# \text{ PoS Tags gesamt}} \cdot 100$$

Hier wurden 50 Sätze aus dem Trainingsdaten extrahiert und 50 Sätze, die nicht in den Trainingsdaten erscheinen, ausgewählt und die Korrektheit ermittelt. Die Ergebnisse des Maximum-Entropie-basierten PoS-Taggings sind in Tabelle 7 angeführt.

PoS-Tag Trainingskorpus:	720000 Wörter
Korrekte PoS-Tag Klassifikation:	96.26 %

Tabelle 7: Ergebnisse der korrekten Klassifizierung von Wortklassen (PoS-Tags)

Die Erkennung von Silbengrenzen spielt eine wichtige Rolle, da die Silbe in dem in dieser Arbeit entwickelten System als Sprachbaustein dient und gleichzeitig als Container für die Merkmalsgewinnung von Phonen fungiert. So werden die quantitativen Merkmale für Phone auch hinsichtlich ihrer Position in der Silbe bestimmt. Silbentrennung ist eine nicht-triviale Aufgabe, und es kann selbst bei Experten unterschiedliche Meinungen zur Silbentrennung geben. Die Silbentrennung in dieser Arbeit beruht auf den Silbentrennungsregeln der alten deutschen Rechtschreibung. Abgeleitet von den Maßen zur korrekten Graphem-Phonem-Klassifikation und der richtigen Zuweisung von Wortklassen, wurde mittels nachfolgender Berechnung die Erkennungsrate der richtigen Erkennung von Silbengrenzen errechnet.

$$\text{Korrekte Silbenerkennung in \%} = \frac{\# \text{ richtig erkannte Silbengrenzen}}{\# \text{ korrekte Silbengrenzen}} \cdot 100$$

Die Ergebnisse werden in Tabelle 8 dargestellt. Es wurden 10 Sätze aus den Trainingsdaten verwendet und 10 Sätze, die nicht in den Trainingsdaten vorkommen.

Letztes Merkmal, welches mit Hilfe des Maximum-Entropie-Frameworks vorhergesagt wird, ist der Wortakzent.

Silbentrenn Trainingskorpus:	69037 Silben
Korrekte Silbenerkennung:	92.72 %

Tabelle 8: Ergebnisse der korrekten Vorhersage von Silbengrenzen

Unter Wortakzent ist die Prominenz von Wortteilen gemeint, in diesem Falle das Festlegen des Akzents auf eine Silbe des Wortes. Der Akzent wird akustisch meist durch eine Anhebung der Grundfrequenz und der Energie gekennzeichnet, was ihm als relevantes Merkmal in der Sprachsynthese Bedeutung zukommen lässt. Das Maß für die richtige Wortakzentzuweisung errechnet sich durch

$$\text{Korrekte Wortakzentzuweisung in \%} = \frac{\# \text{ richtig erkannter Wortakzent}}{\# \text{ korrekter Wortakzent}} \cdot 100$$

Wortakzent Trainingskorpus:	69037 Akzenttragende Einheiten
Korrekte Wortakzentzuweisung:	95.58%

Tabelle 9: Ergebnisse der Wortakzent-Prädiktion

Tabelle 9 gibt die Ergebnisse wieder, welche durch die zugrunde liegenden Trainingsdaten erreicht wurden.

Die Ergebnisse aus den Tabellen 6, 7, 8, 9 sind im weitesten vergleichbar mit spezialisierten und optimierten Werkzeugen, die augenblicklich von der Forschungsgemeinschaft verwendet werden. Siehe hierzu Demberg (2006, Seite 92ff).

4.4.3 Phonem-Visem Klassifikation

Wie in der Motivation in Kapitel 1 bereits eingeführt, benutzt der Mensch zum Verifizieren des akustischen Signals während des Sprechvorgangs die Bewegung der Lippen und konzentriert sich nicht nur auf den auditiven Kanal. Diese Multimodalität bei der menschlichen Sprachproduktion und Sprachperzeption wird u. A. durch die Artikulatoren gestützt. Die Artikulatoren sind ein wichtiges Instrument der Lautformung. Die Artikulatoren, die einem Betrachter von außen sichtbar sind, sind vor allem die Lippen. Bei genauerer Betrachtung kann man beim Sprechvorgang des Menschen weitere Artikulatoren beobachten, die an der Sprachproduktion beteiligt sind. Das sind die Zunge, die Zähne und der Unterkiefer. Uns interessieren hier nur die drei Artikulatoren Lippen, Unterkiefer und Zunge. Die Nase ist natürlich auch als Artikulator zur Lautformung von außen zu betrachten, genauso wie die Zähne, diese sind aber ein statischer Artikulator, der sich beim Sprechvorgang nicht verändert, sondern wegen der menschlichen Morphologie fest an seinem Platz verankert ist. Mehr als eine Dekade befasste sich die Forschung mit der Relation von auditiver und visueller Sprachproduktion und Perzeption. McGurk und MacDonald (1976) haben in ihren Untersuchungen herausgestellt, dass die akustische Sprachproduktion kombiniert mit der visuellen Sprachproduktion die Perzeption beim Menschen beeinflusst. Der bekannte McGurk-Effekt wurde nach diesen Untersuchungen fester Bestandteil der audio-visuellen Sprachforschung.

Die erste von zwei Gruppen bilden die Konsonanten. Die beiden Plosive /p/ und /b/ werden zu einer Klasse zusammengefasst. In englischsprachigen Phonem-Visem-Klassifikationen wurde noch das /m/ in diese Klasse mit aufgenommen. Hier wurde das /m/ einer eigenen Klasse zugeordnet. Der Artikulationsort ist zwar derselbe, doch wegen der stationären bilabialen Lippenstellung wird das /m/ visuell anders realisiert als die Plosive /p/ und /b/, bei denen der Mund geöffnet wird.

Die vier Plosive /t/, /d/, /k/ und /g/ bilden eine Klasse. /t/ und /d/ werden zwar als alveolare Laute produziert und /k/ und /g/ als velare Laute, doch dieser Unterschied ist äußerlich nicht sichtbar und führt aus diesem Grund zu einer Zusammenfassung der Laute zu einer Visem-Klasse.

Die Phonem-Kombinationen /n/, /@n/, /l/ und /@l/ wurden als Einheiten zu einer Visem-Klasse zusammengefasst. /n/ und /l/ unterscheiden sich zwar in der Zungenstellung und das Velum ist bei /n/ abgesenkt, doch ist dies nicht äußerlich unterscheidbar.

Die labiodentalen Frikative /f/ und /v/ wurden zusammengefasst, wie auch die zwei alveolaren Frikative /s/ und /z/.

Eine weitere Klasse bilden die postalveolaren Frikative und die Affrikaten /S/, /Z/, /tS/ und /dZ/, wobei auch hier keine Unterscheidung der visuellen Artikulation getroffen werden kann.

Die drei Frikative /h/, /r/ und /x/ sowie der Nasal /N/ bilden eine Klasse, obwohl diese unterschiedliche Artikulationsorte nach sich ziehen. Die Artikulation findet im hinteren Abschnitt der Mundhöhle statt, doch bleibt bei allen die Mundstellung geöffnet.

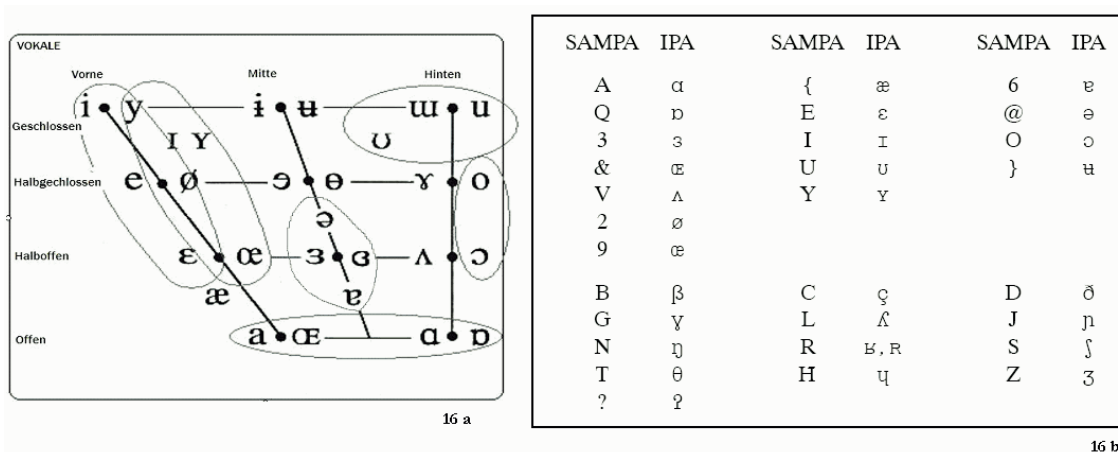


Abbildung 16: (16a) Übersicht der Vokaleinteilung nach Zungenstellung und Grad der Mundöffnung.

Die Visem-Klassen der Vokale sind eingezeichnet.

(16b) SAMPA – IPA Zuordnung. Modifiziert nach Hess

Die letzte Visem-Klasse der Konsonanten bilden die beiden Phoneme /j/ und /C/, die denselben Artikulationsort besitzen und sich nur bei der Approximation unterscheiden. Aber auch hier gilt, dass sie von außen betrachtet nicht unterscheidbar sind.

Die Einteilung der Gruppe der Vokale in einzelne Visem-Klassen beruht auf dem in Abbildung 16a zu sehenden Vokaltrapez, mittels dessen die Zungenstellung bei der Vokalarikulation und der Öffnungsgrad des Mundes nachvollzogen werden kann. Abbildung 16b zeigt die SAMPA-IPA Zuordnung zum besseren Verständnis, da in dieser Arbeit SAMPA verwendet wird. Die Zungenstellung teilt sich ein in Vorne, Mitte und Hinten und der Grad der Mundöffnung in Geschlossen, Halbgeschlossen, Halboffen und Offen. Es wurde keine Unterscheidung zwischen gespannten und ungespannten Vokalen durchgeführt, da dies nicht visuell unterscheidbar ist. Die erste vokalische Visem-Klasse bilden die Vokale /i:/, /I/, /e:/, /E:/ und /E/.

	Phoneme	Viseme	Beispiel
1	p, b	P	Pause, Bitte
2	t, d, k, g	T	Tonne, Dach, König, Gier
3	n, @n, l, @l	N	Nadel, raten, Liebe, Igel
4	M	M	Mutter
5	f, v	F	Finder, Vase
6	s, z	S	Fass, Sein
7	S, Z, tS, dZ	Z	Schar, Rage, Tscheche, Dschungel
8	h, r, x, N	R	Hase, Reden, Dach, Wange
9	j, C	C	Junge, Wicht
10	i:, I, e:, E:, E	E	Bier, Tisch, Weg, Räte, Menge
11	a:, a	A	Wagen, Watte
12	o:, O	O	Wolle, Wogen
13	u:, U	U	Buch, Runde
14	@, 6	Q	Bitte, Weiher
15	y:, Y, 2:, 9	Y	Tür, Mütter, Goethe, Götter

Tabelle 10: Übersicht der Phonem-Visem-Klassifikation

Als zweite Visem-Klasse wurden /a/ und /a:/ zusammengefasst. Die Phoneme /o:/ und /O/ bilden eine Klasse sowie /u:/ und /U/. Die neutralen Phoneme /@/ und /6/ gehören derselben Visem-Klasse an und als Letztes wurden /y:/, /Y/, /2:/ und /9/ einer eigenen Visem-Klasse zugewiesen.

Tabelle 10 gibt einen Überblick über die Visem-Klassen mit entsprechender Bezeichnung. Die Bezeichnung der einzelnen Visem-Klassen wurde durchgängig mit Großbuchstaben vollzogen, wobei die Bezeichnungen an die Phoneme angelehnt wurden, mit Ausnahme der Visem-Klasse 14. Ausschlaggebend für die Bezeichnung war auch eine leichte maschinelle Verarbeitung.

AUTOMATISCHES TRAINING FÜR DIE SPRACHSYNTHESE

Unit-Selection-basierter Sprachsynthese liegt die Annahme zugrunde, dass ein Sprachsignal durch Konkatenation von Sprachsegmenten synthetisiert werden kann, die aus einem zuvor aufgenommenen Sprachdatenkorpus extrahiert werden. Hierbei ist nicht so sehr entscheidend, welche Einheitengröße die extrahierten Sprachsegmente haben, sondern vielmehr, dass die Sprachsegmente in einem Korpus alle sprachlich realisierbaren Möglichkeiten abdecken. Bei einer begrenzten Domäne, wie z.B. einer Sprachsynthese für die Wetterauskunft, ist es nicht notwendig, dass mit dem zugrunde liegenden Sprachdatenkorpus Fachbegriffe aus der Wirtschaft synthetisiert werden können. Will man jedoch ein Synthesystem entwickeln, das jegliche Art von Text synthetisieren kann, so muss sichergestellt sein, dass mit dem verwendeten Sprachdatenkorpus jede mögliche textliche Repräsentation akustisch realisierbar ist.

Die Problemstellung bei Unit-Selection-basierter Sprachsynthese besteht in der geeigneten Auswahl der Sprachsegmente. Dies ist unabdingbar, um ein natürlich klingendes und verständliches Sprachsignal erzeugen zu können. In diesem Kapitel werden die Verfahren beschrieben, die für die Unit-Selection-basierte Sprachsynthese entwickelt und angewendet wurden. Ausgehend von der Problemstellung, welche in Abschnitt 5.1 wiedergegeben ist, wird die Vorverarbeitung erläutert, die Voraussetzung ist für die automatischen Trainingsalgorithmen für Unit-Selection-basierte Sprachsynthese. Die automatischen Trainingsverfahren sind statistisch motiviert. Diese sind zum Einen die HMM-basierte Sprachsynthese, bei der das Sprachsignal aus den Sprachparametern selbst durch den Einsatz eines Filters erzeugt wird und zum Anderen Graphen-basierte Modelle, bei denen das Sprachsignal durch Konkatenation akustischer Sprachsegmente generiert wird, wobei die Sprachsegmente durch einen Graphen repräsentiert werden und die Kanten eine statistische Kennzahl als Kantengewichte haben. Bei den graphischen Modellen wird nochmals die Unterteilung in gerichtete Graphen und ungerichtete Graphen vorgenommen. Beiden Ansätzen liegt jedoch die Annahme der bedingten Wahrscheinlichkeit zugrunde, welche

die Ausgangsüberlegung zur Segmentauswahl bildet. Mit den vorgestellten Verfahren kann natürlich klingende Sprache synthetisiert werden, und es ist darüber hinaus möglich, Sprachsynthesysteme für andere Sprachen als Deutsch zu entwickeln, da die Verfahren leicht auf andere Sprachen adaptierbar sind und nur die Erstellung der Sprachdaten-Korpora, die Textvorverarbeitung und Prosodie sprachabhängig ist.

5.1 Unit Selection

Sprachsynthese-Systeme, die mit Hilfe des Unit-Selection-Algorithmus Einheiten aus einem zuvor aufgenommenen Sprachdaten-Korpus auswählen, haben das Grundproblem, dass zu einer gegebenen linguistisch motivierten Beschreibung der zu synthetisierenden Äußerung die Sprachbausteine ermittelt werden müssen, so dass deren Verknüpfung die äquivalente lautsprachliche Äußerung am besten approximiert. Unit-Selection-basierte Sprachsynthese kann also als Suchproblem artikuliert werden, bei dem die Einheiten aus den zugrunde liegenden Sprachdaten gesucht werden, mit denen die natürliche Sprache rekonstruiert werden kann. Das Suchproblem kann unterschiedlich gelöst werden. Das zurzeit am häufigsten eingesetzte Verfahren ist das von Black et al. entwickelte Auswahlverfahren, welches zweidimensionale Kostenterme verwendet, um die geeigneten Sprachbausteine auszuwählen. Black et al. (1996) verwenden zwei Kostenfunktionen, welche zum einen die Einheitenkosten (orig.: unit distortion) und zum anderen die Übergangskosten (orig.: continuity distortion) berechnen. Die Kostenterme bezeichnen also Abstandsmaße, die die Sprachbausteine zueinander in Beziehung setzen. Die Abweichung der Merkmale des Sprachbausteins zu den aus der Eingabe berechneten statischen Merkmalen sowie den vorhergesagten dynamischen Merkmalen der Zieleinstellung, d.h., welche Eigenschaften der Sprachbaustein erfüllen muss, um die zu synthetisierende Äußerung bestmöglich wiederzugeben, wird mit diesen Einheitenkosten dargestellt. In den Einheitenkosten wird die phonetische Transkription als Hauptmerkmal verwendet. Nur wenn eine solche Einheit im Sprachdaten-Korpus gefunden wird, werden die phonologischen, quantitativen und prosodischen Merkmale miteinbezogen und deren Abstand zueinander bemessen. Die Merkmale sind nicht alle zur Laufzeit vorhanden und müssen vorhergesagt werden. Die Vorhersage von Merkmalen wird über statistische Lernverfahren gelöst. Hier spielt die Prädiktion der prosodischen Merkmale eine entscheidende Rolle. Dauer und F0-

Prädiktion haben unmittelbaren Einfluss auf die Qualität der erzeugten Sprachausgabe. Wird Dauer und F0 schlecht geschätzt, werden schlechte Bausteine bzw. ganz und gar falsche ausgewählt. Die Kostenterme berücksichtigen eben diese Einschränkungen und gewichten die prosodischen Merkmale Dauer und F0 höher als die Prädiktion von Part-of-Speech Tags.

Die Übergangskosten spiegeln den Abstand zwischen zwei Sprachbausteinen wider, die die zu konkatenierenden Sprachbausteine in ihren spektralen Eigenschaften aufweisen. Die spektralen Eigenschaften eines Sprachsignals können durch unterschiedliche Parametrisierung dargestellt werden. In der Spracherkennung wie auch in der Sprachsynthese sind diese Parameter meist durch Mel-Cepstrum-Koeffizienten (MFCC) wiedergegeben. Abschnitt 5.2 geht auf diese parametrisierte Darstellung ein. Um nun die beste Folge der Sprachsegmente anhand der Einheitenkosten und Übergangskosten zu berechnen werden die jeweiligen gewichteten Kosten summiert. Die Sprachsegmentabfolge mit den jeweils geringsten Kosten repräsentiert die optimalen Sprachbausteine, die durch Konkatenation das erwünschte Sprachsignal ergeben. Nachfolgend wird die Kostenberechnung formal dargestellt.

$$s = \arg \min_{\{x_1, \dots, x_N | \forall n \in \{1, \dots, N\}\}} \sum_{i=1}^N \left(\sum_{j=1}^J cd_j(u_{x_i}, u_{x_{i-1}}) \cdot cw_j + \sum_{k=1}^K ud_k(u_{x_i}, t_i) \cdot uw_k \right)$$

Die Funktion gibt die Sprachbausteinfolge s wieder, welche durch den Vektor u_i , der das Sprachsegment im Korpus beschreibt, und den Vektor t_i , der die gewünschten Eigenschaften hinsichtlich der Eingabe darstellt, bestimmt wird. cw (continuity weights) und uw (unit weights) sind die Gewichte der Kanten. N repräsentiert die Anzahl der Syntheseeinheiten und M ist die Anzahl der Sprachsegmente, wie sie im Korpus auftreten. $\sum cd_j$ sind die Übergangskosten (continuity distortion) und $\sum ud_i$ die Einheitenkosten (unit distortion). Die Auswahl der Sprachbausteine kann als gerichteter Auswahlgraph dargestellt werden, bei dem jeder Knoten einen Sprachbaustein repräsentiert. Ein Graph $G = (V, E)$ besteht aus einer Menge V von Knoten (engl.: vertices), einer Menge E von Kanten (engl.: edges), wobei die Kanten Paare von Knoten $e = (v1, v2)$ sind bzw. zwei Knoten miteinander verbinden. Bei einem ungerichteten Graphen ist jede Kante $e = (v1, v2)$ ein

ungeordnetes Paar, wohingegen bei einem gerichteten Graphen (engl.: directed graph) jede Kante $e = (v1, v2)$ ein geordnetes Paar ist. Die Kanten des Auswahlgraphen werden jeweils mit den Kosten gewichtet und gestatten somit die Verwendung von Kürzest-Pfade-Algorithmen, wie den bekannten A*-Algorithmus⁶. Die Berechnung des optimalen Pfades vom Startknoten zum Endknoten des Auswahlgraphen gibt die Sprachbausteinfolge wieder, wie sie zur Konkatenation verwendet werden soll. Der optimale Pfad ist derjenige, dessen Kantengewichte eine minimale Summe bilden.

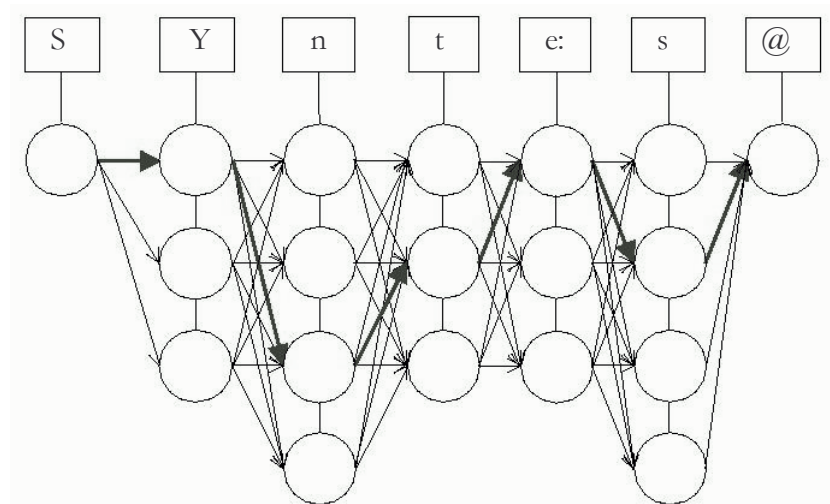


Abbildung 17: Auswahlgraph auf Phonebene zur Konkatenation der Sprachsegmente, hier: Synthese angegeben in phonetischer Transkription /S Y n t e: s @/

Da eine sprachliche Äußerung eine zeitliche Abfolge von Lauten ist, wurde abweichend von den Kostentermen ein Verfahren auf probabilistischen Grundlagen entwickelt. Diesem Ansatz liegt die Annahme zugrunde, dass eine lautliche Äußerung phonotaktischen Bedingungen unterliegt, so dass nicht jedes Sprachsegment mit beliebigen Eigenschaften aneinandergereiht werden kann. Die sprachliche Äußerung kann vielmehr als ein Übergangsnetzwerk angesehen werden, welches mittels der Auftretenswahrscheinlichkeiten der Sprachbausteine im Kontext modelliert werden kann. Nachfolgend werden die Verfahren beschrieben, die die lautliche Sprachproduktion als Übergangsgraphen ansehen. Das HMM-basierte Verfahren wurde hierfür für das Deutsche angepasst und ausgehend davon, dass HMMs nur die Verbundwahrscheinlichkeit des aktuellen Zustands und der

⁶ Abschnitt 5.6 beschreibt diesen Algorithmus detailliert, da er auch für die Auswahl der Sprachbausteine mittels probabilistischer endlicher Automaten verwendet wird.

Beobachtung berücksichtigen, wurden zwei Modelle entwickelt, welche zurückliegende Zustände mit in Betracht ziehen. Dies ist zum einen ein Bedingte-Entropie-basiertes Modell, welches als gerichteter Graph implementiert wurde, und zum anderen das Verfahren der Conditional-Random-Fields, welches einen ungerichteten Graphen als Grundlage verwendet.

5.1.2 Quantitative, phonologische und linguistische Merkmale

Als Merkmale zur Beschreibung der Sprachbausteine dienen quantitative, phonologische und linguistische Parameter. Tabelle 11 gibt einen Überblick über die verwendeten Merkmale.

Phoneme	Aktuelles Phonem in Phon, Silbe, Wort, Bigramm
	Vor-vorhergehendes, vorhergehendes, nachfolgendes und nach-nachfolgendes Phonem
	Phoneme in vorhergehender/nachfolgender Silbe, Wort, Bigramm
Silbe	Anzahl der Phoneme in der vorhergehenden, aktuellen und nachfolgenden Silbe
	Akzent der vorhergehenden, aktuellen und nachfolgenden Silbe
	Position der aktuellen Silbe im aktuellen Wort; rechtszählend/linkszählend
	Position der Silbe im Satz; rechtszählend/linkszählend
	Nukleus der aktuellen Silbe
Wort	PoS von aktuellem, vorangehendem und nachfolgendem Wort
	Position des Wortes in der Äußerung; rechtszählend/linkszählend
Äußerung	Anzahl der Silben in der Äußerung
	Anzahl der Wörter in der Äußerung
	Satzmodus der Äußerung

Tabelle 11: Übersicht der quantitativen, phonologischen und linguistischen Merkmale zur Sprachsegmentbeschreibung

Jeder Sprachbaustein kann im Trägersatz, wie auch in der zu synthetisierenden Äußerung, anhand seiner Position quantitativ beschrieben werden. Hierzu wird die Position des Sprachbausteins links und rechts zählend bestimmt. Es wird die Position im Satz, im Wort und in der Silbe beschrieben.

5.2 Spektrale Eigenschaften der Sprachsegmente

Die Analyse von Sprachsignalen wird zumeist im Frequenzbereich durchgeführt. In der zeitlichen Darstellung können die Sprachsignale nicht so einfach und konsistent analysiert werden wie im Frequenzbereich (O'Shaughnessy 2000). Spektrale Eigenschaften von Sprachsignalen sind parametrisierte Darstellungsformen, die durch geeignete Verfahren gewonnen werden. Die einzelnen Sprachbausteine haben je nach Auftreten im Korpus unterschiedliche spektrale Eigenschaften. Diese Eigenschaft ist darauf zurückzuführen, dass der Sprecher die Sprachlaute niemals gleich realisieren kann. Das Sprachsignal selbst ist ein nichtstationärer Vorgang, der nur kurzzeitig stationär ist: Dies bedeutet das Sprachsignal ändert sich mit der Zeit. Die spektralen Eigenschaften des Sprachsignals, also der Frequenzanteil im Sprachsignal, ändern sich somit ebenfalls im Verlauf der Zeit. Diese Änderungen im Frequenzbereich spielen sich zwischen 100 und 8000 Hz ab. Abbildung 18 zeigt die Darstellung eines Sprachsignals im Zeitbereich und im Zeit-Frequenz-Spektrogramm, woraus die Änderungen im Frequenzbereich ersichtlich werden.

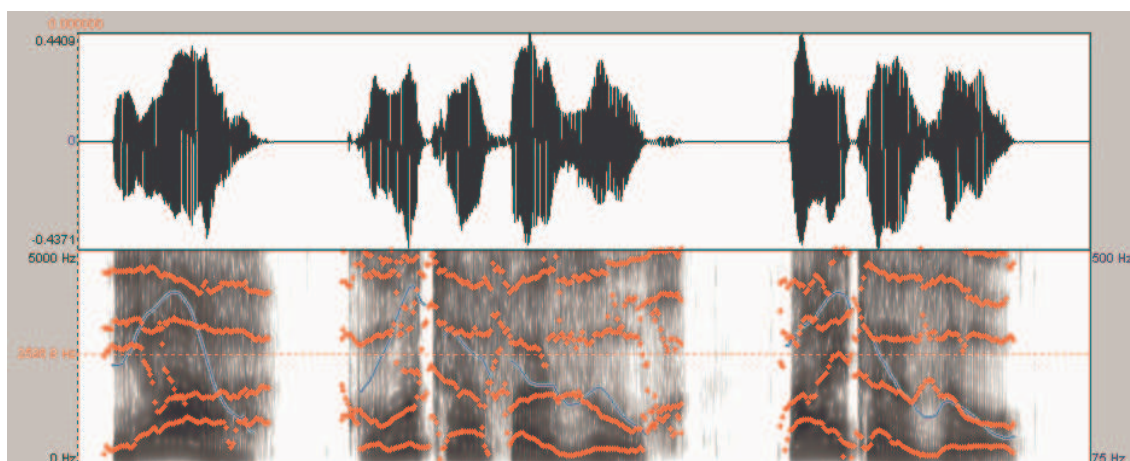


Abbildung 18: Darstellung des Sprachsignals „Ja, guter Versuch. Weiter so.“:
Sichtbar: Zeitbereich, Spektrum, Formanten und F0-Verlauf.

In der oberen Darstellung in Abbildung 18 wird das Sprachsignal mit seiner Einhüllenden im Zeitbereich dargestellt. Darunter wird das zugehörige Spektrogramm mit einem Formantverlauf dargestellt. Anhand der unterschiedlichen Farbstärke werden die Änderungen im Spektrogramm ersichtlich. Um die spektralen Eigenschaften eines Sprachsignals zu untersuchen, hat sich das Prinzip der Kurzzeit-Spektralanalyse bewährt. Bei der Kurzzeit-Spektralanalyse von Sprachsignalen wird nur ein kurzer Block bzw. ein Ausschnitt des Sprachsignals verarbeitet. Hierbei wird durch Fensterung ein Block aus dem Signal ausgeschnitten. Bei der Fensterung wird zumeist auf das Hamming-Fenster zurückgegriffen, wobei sich eine Größe von 20 bis 50 Millisekunden für das Fenster als sinnvoll erwiesen hat. Zumeist wird die Fenstergröße mit 25 Millisekunden gewählt. Der kurze Signalabschnitt wird einer diskreten Fourier-Transformation (DFT) unterzogen, wobei die Blocklänge ein Vielfaches von 2 ist. Die Fourier-Transformierte wird als Spektrum bezeichnet.

Zur parametrisierten Darstellung der spektralen Eigenschaften von Sprachsignalen werden verschiedene Ansätze verfolgt. Zum einen wird die Lineare Prädiktion (LP) verwendet um Koeffizienten aus dem Sprachsignal zu extrahieren. Das LP-Spektrum wird ebenso verwendet wie die Line Spectrum Frequencies. Eine etwas andere Darstellung wird mit den an der menschlichen Perzeption entwickelten Maßen erreicht, die Frequenzen zusammenfassen und so die menschliche Perzeption simulieren. Hierzu zählen die Perzeptiven Linearen Prädiktor-Koeffizienten (PLP) und die Mel-Cepstrum-Koeffizienten. In der Spracherkennung wie auch in der Sprachsynthese haben sich die Mel-Cepstrum-Koeffizienten als Quasi-Standard-Parametrisierung durchgesetzt. Wie die Mel-Cepstrum-Koeffizienten gewonnen werden, wird im nachfolgenden Abschnitt erläutert.

5.2.2 Mel-Cepstrum-Koeffizienten

Mel-Cepstrum-Koeffizienten (mel-frequency cepstrum coefficients, MFCC) gelten als zuverlässige Parameter und sind in der Spracherkennung wie auch in der Sprachsynthese die meistgenutzten Parameter zur Repräsentation spektraler Eigenschaften von Sprachsignalen. Sie werden in der Sprachsynthese zur Berechnung des spektralen Abstandes zwischen zwei Sprachsegmenten verwendet. Dies bescheinigt zwei aufeinander folgenden

Sprachbausteinen, dass sie in spektralen Merkmalen zueinander passen und somit die Störungen an den Konkatenationsstellen minimieren.

Die Mel-Cepstrum-Koeffizienten werden formal als c_n angegeben. Die MFCC stellen eine psychoakustisch motivierte Erweiterung der Cepstralanalyse dar. Mel ist die Maßeinheit für eine psychoakustische Größe und beschreibt die wahrgenommene Tonhöhe. Die Mel-Skala ist eine Nachbildung der Frequenz-Orts-Transformation im menschlichen Gehör. Insofern bietet sie eine optimale Darstellung der Frequenzskala (Zwicker 1982). Die MFCC bieten eine Alternative zu Linearen Prädiktor Koeffizienten (LPC).

Die Berechnung der Mel-Cepstrum-Koeffizienten läuft wie folgt ab:

- Berechnung des (Amplituden-)Spektrums via DFT
- Abbildung der Frequenzskala auf die Mel-Skala; manchmal auch
- Berücksichtigung der Maskierung im Frequenzbereich (durch gewichtete Mittelung)
- Logarithmierung
- Rücktransformation in den Zeitbereich und damit Berechnung des Cepstrum.

Es werden zumeist acht bis vierzehn Koeffizienten extrahiert. Die Koeffizienten geben unterschiedliche Werte wieder. So repräsentiert der Null-Koeffizient c_0 die durchschnittliche Energie in den Sprachsignalabschnitten. Der erste Koeffizient c_1 steht für das Verhältnis von niedrigen und hohen Frequenzwerten. Für die parametrische Darstellung der spektralen Eigenschaften der im Korpus annotierten Sprachsegmente wurden jeweils zwölf Koeffizienten des ersten Sprachsignalabschnitts und des letzten Sprachsignalabschnitts verwendet.

5.3 HMM-basierte Sprachsynthese

HMM-basierte Sprachsynthese wurde u. a. von Donovan (1996) und Acero et. al (1997) vorgestellt. HMM-basierte Sprachsynthese ist in vielerlei Hinsicht ein Erfolg versprechender Ansatz für die Sprachsynthese, der, wie von Tokuda et al. (2002) gezeigt wurde,

gute Synthesequalität liefert. In vielen Bereichen der Sprachverarbeitung wurden HMM-basierte Verfahren angewendet. Im Bereich der Spracherkennung sind HMMs das meist genutzte Vorhersagemodell. Während bei der automatischen Spracherkennung die gesprochenen Äußerungen, d.h. das akustische Sprachsignal, unter Zuhilfenahme eines geeigneten Inventars auf eine möglichst exakte orthographische Repräsentation der Äußerung abgebildet wird, wird bei der Sprachsynthese diese Einsatzweise umgekehrt und mittels HMMs eine gegebene orthographische Textrepräsentation als akustisches Sprachsignal wiedergegeben.

HMM-basierte Sprachsynthese bietet den Vorteil, dass anhand von aufgenommenen Sprachdaten in der Größenordnung ab 15 Minuten Modelle trainiert werden können, die eine verständliche Sprachsynthese liefern (Weiss 2005). Bei diesem Ansatz wird das Sprachsignal direkt aus den Sprachsignalparametern erzeugt, die durch die kontextuellen HMMs generiert wurden. Die Merkmals-Vektoren, welche die Eigenschaften der Sprachbausteine beschreiben, werden als Zustände betrachtet und die Sprachgenerierungsparameter im Falle des nachfolgend beschriebenen Sprachsynthese-Systems, also die MFCCs der Sprachbausteine, als Beobachtung. Die Ausgaben der einzelnen Zustände werden aneinandergereiht und ergeben somit die spektrale Repräsentation des zu synthetisierenden Sprachsignals. Mit dem Einsatz eines Filters wird aus den spektralen Parametern dann das akustische Signal erzeugt. Hierin liegt auch der Nachteil eines solchen Systems. Da das Sprachsignal aus den MFC-Koeffizienten und zusätzlich den Anregungsparametern erzeugt wird, folgt dieser Ansatz einem traditionellen Vocoder. Das ausgegebene Sprachsignal klingt daher etwas künstlich. Dies ist auf den Quelle-Filter-Ansatz zurückzuführen, der zur Erzeugung des Sprachsignals verwendet wird. Das nachfolgende HTS-System wurde von Tokuda et al. (2000, 2002) eingeführt und für das Deutsche angepasst (Weiss 2005). Es wurde ein sprachabhängiges Modell trainiert, welches es ermöglicht, Sprachsignale in deutscher Sprache zu erzeugen.

5.3.1 Das HTS-Sprachsynthese-System

Das HTS-Sprachsynthese-System (**HMM Text-to-Speech**) wurde am Nagoya Institute of Technology, Japan (Tokuda et al. 2000, 2002) entwickelt und erfolgreich für unterschied-

liche Sprachen verwendet. Das System wurde für Englisch (Tokuda et al. 2002), wie auch für die portugiesische Sprache angepasst und trainiert, so dass Sprachsignale in der jeweiligen Sprache generiert werden können. Nachfolgend wird das HTS-System eingeführt und die Arbeiten werden erläutert, die für die Anpassung des Systems erforderlich waren, um deutsche Sprachsynthese durchführen zu können. Die Arbeiten hierzu bedurften eines annotierten Sprachdatenkorpus von mindestens 15 Minuten aufgenommener Sprache. Hier wurde auf das Sprachdatenkorpus, welches am Institut für Kommunikationswissenschaften aufgenommen und annotiert wurde, zurückgegriffen und eine Untermenge der aufgenommenen Sprachdaten verwendet. Eine Beschreibung der Trägersätze findet sich in Abschnitt 5.3.2 wieder.

Das HTS-System ist aufgeteilt in einen Trainingsabschnitt und in einen Syntheseabschnitt. Das Training des Systems wird auf den zugrundeliegenden, annotierten Sprachdaten durchgeführt. Hierbei werden die kontextuellen HMMs trainiert, welche zur Laufzeit die Sprachparameter generieren. Die eigentliche Synthese basiert auf einem eigenständigen Modul. Der Syntheseprozess erlaubt es zur Laufzeit die Sprachsignale zu erzeugen.

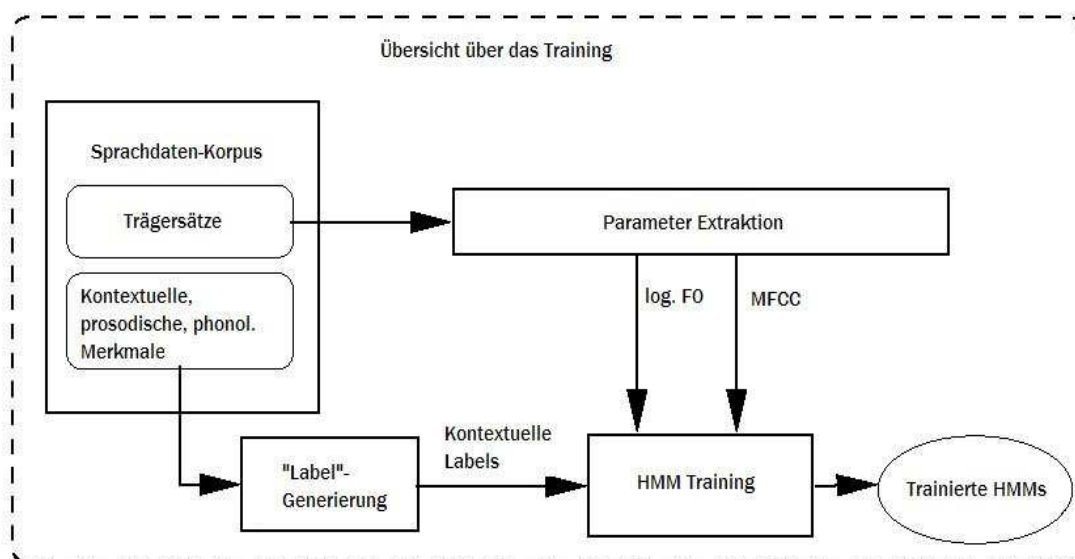


Abbildung 19: Blockdiagramm: Training HTS-Sprachsynthese-System

Das Training der HMM wird nach Erstellung der kontextuellen Labels und der Sprachparameter-Merkmalvektoren durchgeführt. Abbildung 19 gibt einen Überblick über die notwendigen Schritte, die zu einem erfolgreichen Training des Systems führen. Das Sys-

tem arbeitet auf Basis von Phon-Einheiten. Aus den Trägersätzen der Sprachdaten werden die Sprachsignalparameter extrahiert. Dies sind zum einen die logarithmierten F0-Werte, die Dauern sowie die MFC-Koeffizienten. Neben den Signalparametern werden zusätzlich aus den Trägersätzen die quantitativen, prosodischen, phonologischen und linguistischen Merkmale generiert, die die Einheit hinsichtlich dieser Merkmale beschreiben. Diese werden als kontextuelle Labels bezeichnet, da je nach Kontext sich die quantitativen, prosodischen und phonologischen Merkmale ändern. Die trainierten Modelle werden dann während der Laufzeit verwendet, um die Sprachparameter zu generieren, die daraufhin verkettet werden und letztlich durch das MLSA-Filter (Mel Log Spectrum Approximation) zu einem akustischen Signal geformt werden. Das MLSA-Filter ist beschrieben in Imai (1983) und stellt eine Technik zur Verfügung, mit Hilfe derer Sprache direkt durch das Mel-verzerrte Cepstrum erzeugt werden kann. Siehe hierzu auch: <http://lima.lti.cs.cmu.edu/mediawiki/index.php/MLSA>. Dies geschieht zur Laufzeit und ist unabhängig von dem Trainingsabschnitt. Abbildung 20 zeigt einen Überblick, wie das Sprachsignal zur Laufzeit erzeugt wird. Nachfolgend wird das kontextuelle, sprachabhängige Training der HMMs für das Deutsche beschrieben, wie es auch für die Sprachsignalgenerierung verwendet wird.

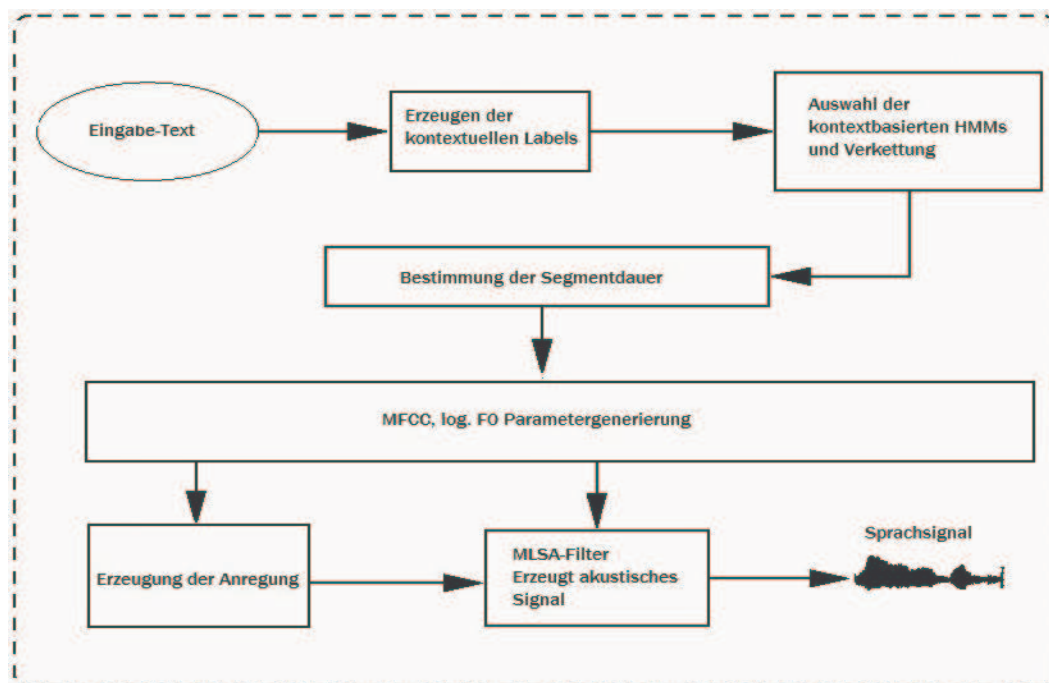


Abbildung 20: Blockdiagramm: Erzeugung eines Sprachsignals zur Laufzeit mit dem HTS-Sprachsynthese-System

5.3.2 Sprachabhängiges Training der kontext-basierten HMMs

Das HTS-System wurde für das Deutsche angepasst und trainiert unter Verwendung von 500 Sätzen gesprochener Sprache. Die verwendeten Sprachdaten sind aus dem zugrunde liegenden Sprachdaten-Korpus extrahiert. Die 500 Sätze ergeben Sprachdatenmaterial von ca.15 Minuten Sprache. Die kontextuelle Erzeugung der Labels wurde aus dem annotierten Sprachdaten-Korpus gewonnen. Beispiele: Die Labels für /d a n/ haben die Form:

$$0^0-d+a=n/M2:1_3/S1:1-3+1_1/W1:44_#1/W5:1/U:1_&1$$

$$0^d-a+n=0/M2:2_2/S1:1-3+1_1/W1:44_#1/W5:1/U:1_&1$$

$$d^a-n+0=0/M2:3_1/S1:1-3+1_1/W1:44_#1/W5:1/U:1_&1$$

wobei ein Kontext von zwei Phonemen rechts und zwei links berücksichtigt wird. Die Sonderzeichen (^, -, +, =, /, :, _, #, &) geben die Trennung zwischen den Phonemen und den Merkmalen an. 0 bezeichnet ein Dummy-Phonem. M gibt Metamerkmale, wie Part-of-Speech oder Satztyp wieder. S bezeichnet die Merkmale bezogen auf die Silbe, W bezogen auf das Wort und U bezogen auf die Gesamtäußerung. Die einzelnen Merkmale, welche bei der Erstellung der Labels Verwendung finden, wurden in vorherigen Abschnitt 5.1.1 Tabelle 11 aufgeführt. In den oben angeführten Beispielen wird jeweils zuerst das /d/ betrachtet, danach das /a/ und zuletzt das /n/.

Das sprachabhängige Training des HMM-basierten Sprachsynthese-Systems wird anhand der erstellten Labels mit den dazugehörigen prosodischen Merkmalen Dauer und F0 durchgeführt. Diese repräsentieren die Zustände s_t des HMM. Die Beobachtungen O zu den Zuständen werden durch die MFC-Koeffizienten mit ihren Delta- und Delta-Delta-Koeffizienten x_t wiedergegeben

$$o_t = [x_t, \Delta x_{t-1}, \Delta x_{t-2}].$$

Die Beobachtungen folgen einer einfachen Gauss-Verteilung der Merkmalsvektoren X . Die Aufgabe eines HMM λ besteht nun darin, zu den gegebenen Zuständen

$s = (s_1, s_2, \dots, s_n)$ die geeignete Beobachtung O zu generieren. Formal wird die Wahrscheinlichkeit $P(O | a, \lambda)$ maximiert unter der Berücksichtigung von X .

Die Dauer-Modellierung wird anhand von Entscheidungsbäumen gelöst. Jedem Zustand des HMMs wird eine einfache Gauss-Verteilung der Dauer zugewiesen. Für die Entscheidungsbaum-basierte Dauer-Modellierung sei auf den nachfolgenden Abschnitt 5.4.2 verwiesen, da bei der in dieser Arbeit entwickelten Unit-Selection-basierten Synthese mittels akustischen Einheiten, die Dauer als dynamisches Merkmal anhand eines Entscheidungsbaumes vorhergesagt wird.

5.4 Korpusbasierte Sprachsynthese

Im Gegensatz zu der HMM-basierten Sprachsynthese wird bei der korpusbasierten Sprachsynthese das Sprachsignal nicht mittels eines Parametergenerierungsalgorithmus und eines Filters erzeugt, sondern durch die Auswahl und Verkettung von akustischen Sprachbausteinen. Die jeweiligen Sprachbausteine für die Verkettung haben unterschiedliche prosodische Realisierungen und unterliegen somit keiner nachträglichen Signalmanipulation, um eine gewünschte prosodische Realisierung erreichen zu können. Hieraus ergibt sich auch der wesentliche Vorteil dieser Herangehensweise. Dadurch, dass verschiedene prosodische Realisierungen ein- und desselben Sprachsegments vorliegen, kann man das am besten geeignete Sprachsegment auswählen und vermeidet somit die Signalstörungen, wie sie zum Beispiel bei der Diphon-Synthese durch Manipulation der Segmente auftreten. Korpusbasierte Sprachsynthese liefert natürlich klingende Sprache, die oftmals von der Original-Aufnahme nicht zu unterscheiden ist. Nachfolgend werden die Einheiten definiert, wie sie zur Auswahl zur Verfügung stehen. Der Ansatz folgt dem Prinzip, dass die Sprachsegmente in unterschiedlicher Bausteingröße verwendet werden können und das längste zusammenhängende Segment ausgewählt und konkateniert wird zum eigentlichen Sprachsignal.

5.4.1 Korpusbasierte Sprachsegmente

Die Sprachbausteine gliedern sich in hierarchischer Struktur. Dies ermöglicht eine mög-

lichst optimale Auswahl der Sprachsegmente, die in ihrer natürlichen Umgebung stehen. Mit natürlicher Umgebung sei hier gemeint, dass die Sprachbausteine in dieser Abfolge im zugrunde liegenden Sprachdaten-Korpus auftreten und somit keinerlei Verkettungsstörungen unterliegen. Die Einheiten wurden so definiert, dass eine Top-Down-Auswahl ermöglicht wird. Bei der Synthese des Eingabetextes werden entsprechend der phonetischen Transkription zuerst ganze Satzphrasen selektiert. Dies hat den Vorteil, dass eine natürliche prosodische Realisierung erhalten bleibt und innerhalb dieser Satzphrase keinerlei Störungen auftreten, welche an den Verkettungsstellen üblicherweise auftreten können. Die Sprachsegmente gliedern sich in die Einheiten Wort-Bigramm, Wort, Silbe, Triphon, Diphon und Phon. Jeder Sprachbaustein wird durch einen Merkmalsvektor beschrieben. Zur Beschreibung werden dieselben Merkmale verwendet wie zuvor bei der HMM-basierten Sprachsynthese. Diese sind in Abschnitt 5.1.1 in Tabelle 11 aufgeführt.

5.4.2 Datenbasierte Dauer-Modellierung

Die Dauer hat eine wesentliche Bedeutung für die Realisierung der prosodischen Merkmale und ist beeinflusst von quantitativen wie auch von linguistischen Merkmalen. So kann ein und dasselbe Lautsegment je nach Position unterschiedlich realisiert sein. Konventionelle Sprache produziert 150 bis 200 Wörter pro Minute inklusive Pausen, die mit durchschnittlich 650 Millisekunden angegeben werden (O'Shaughnessy 2000). Die Dauer der Sprachsegmente ist abhängig vom Sprechstil. Unterschieden werden kann z.B. ein Vorlesestil oder natürliche Konversation. Weiterhin ist die Dauer abhängig von Akzenten, Ort der Pausengrenzen, Ort und Art der Artikulation sowie dem Sprechrhythmus. Die typische Dauer einer Silbe beträgt durchschnittlich 200 Millisekunden. In Abbildung 21 wird exemplarisch die Dauerverteilung von /d a s/, wie das Segment im Korpus auftritt, dargestellt. Silben, die in finaler Position stehen, haben eine signifikant längere Dauer als Silben, die anderswo im Satz realisiert wurden. Bei den Phondauern variiert die Dauer zwischen Vokalen und Diphthongen und Sonoranten, wobei Diphthonge durchschnittlich 75 Millisekunden länger sind als Vokale, die wiederum 30 Millisekunden länger als Sonoranten sind. Die Segmentdauer hat einen entscheidenden Einfluss auf die Qualität der Sprachausgabe und führt bei Nichtbeachtung zu einem unnatürlichen Sprachsignal. Die Segmentdauer ist ein wichtiges Merkmal zur Identifikation des geeigneten Sprachbau-

steins im zugrunde liegenden Sprachdaten-Korpus. Für die Auswahl der Sprachbausteine wird die Dauer für die Einheiten Wort, Silbe und Phon angegeben und vorhergesagt. Da die Dauer sich über eine große Werteskala erstreckt, wurden zur Verbesserung des Trainings die Dauerwerte mittels des natürlichen Logarithmus logarithmiert.

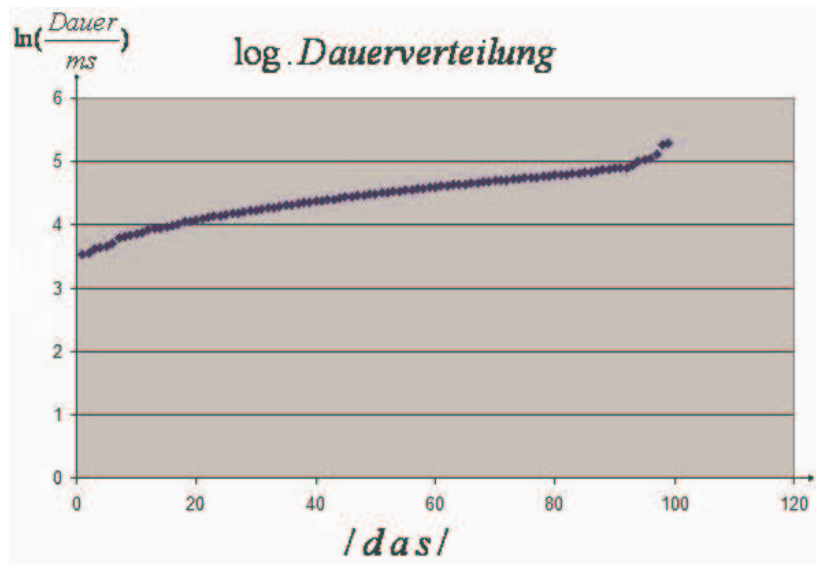


Abbildung 21: Dauerverteilung in Millisekunden für /d a s/ im zugrunde liegenden Sprachdaten-Korpus.

Die x-Achse repräsentiert die Anzahl der Sprachsegmente.

Die Modellierung des Segmentdauermodells beruht auf einer statistischen Vorhersage abhängig vom Kontext, in dem das Sprachsegment auftritt. Dadurch werden die oben angeführten unterschiedlichen Dauerrealisierungen berücksichtigt. Zur Prädiktion der Segmentdauer für die Einheiten wird ein Entscheidungsbaum-basiertes Lernverfahren verwendet (Breiman et al. 1984). Die Merkmalsvektoren für die einzelnen Sprachsegmente werden aus den Merkmalen, wie sie im vorhergehenden Abschnitt aufgezeigt wurden, gebildet. Abbildung 22 zeigt den schematischen Aufbau eines Entscheidungsbaumes. Die Entscheidungen sind binäre Entscheidungen, die jeweils abfragen, ob ein bestimmtes Merkmal im Merkmalsvektor zutrifft oder nicht. Nach diesen Entscheidungen wird der Baum von der Wurzel nach unten durchschritten bis ein Blatt mit einem diskreten Wert erreicht wurde. Der Dauerwert der jeweiligen Einheit ergänzt den Merkmalsvektor, welcher den Sprachbaustein beschreibt.

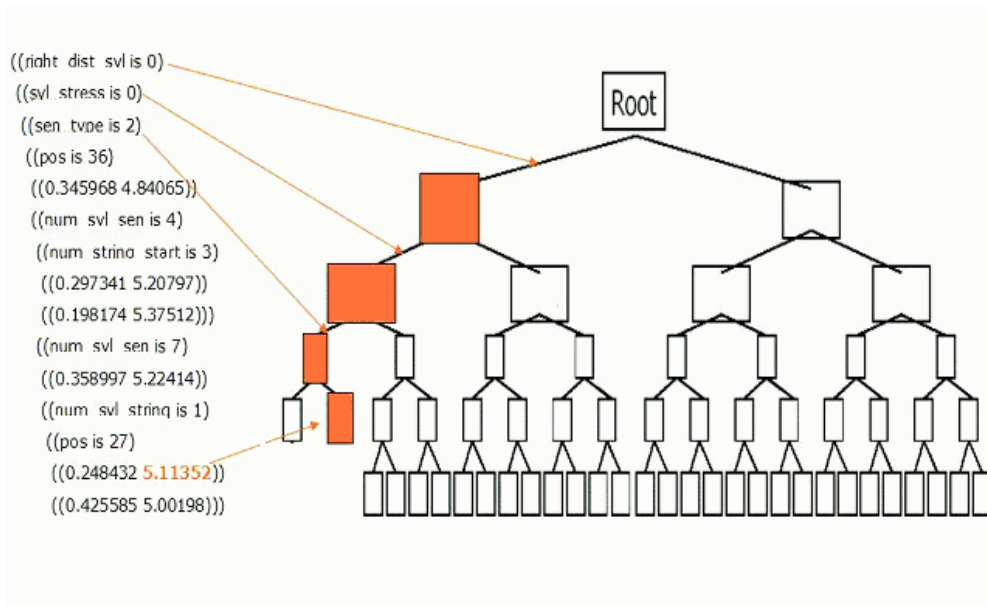


Abbildung 22: Schematische Darstellung eines Entscheidungsbaumes zur Vorhersage von Silbendauern. Die linksseitig in Klammern stehenden Angaben repräsentieren die Fragen, welche während eines Baumdurchlaufes abgefragt werden.

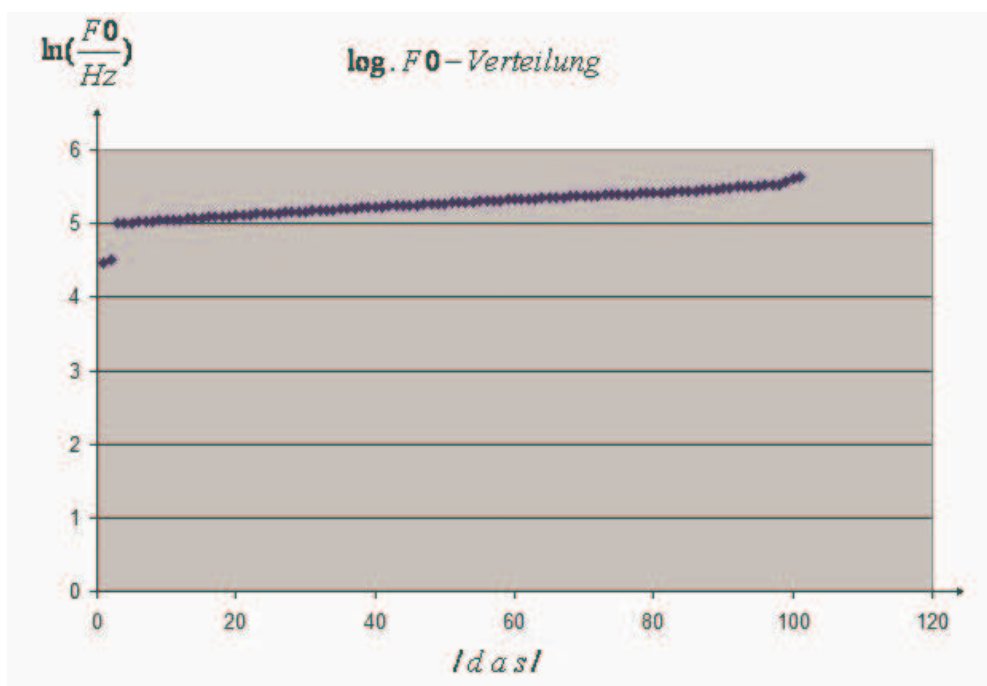


Abbildung 23: Durchschnittlicher $\log_e F_0$ Wert für /d a s/ im zugrunde liegenden Sprachdaten-Korpus. Die x-Achse repräsentiert die Anzahl der Sprachsegmente.

5.4.3 Datenbasierte Grundfrequenz-Generierung

Die physikalische Größe Grundfrequenz (F0) ist neben der Dauer eines Lautes einer der wichtigsten Parameter hinsichtlich der Prosodie. Dies wirkt sich auch auf das Beschreibungsmerkmal eines Sprachsegments aus, da Dauer und F0 entscheidend für die Auswahl der geeigneten Einheit sind. F0 wird wie die Dauer anhand eines Entscheidungsbaumes vorhergesagt. Abbildung 23 zeigt exemplarisch die F0-Verteilung von / d a s/.

5.4.4 Karhunen-Loeve Transformation und Eigenspektrum

Die nach Karhunen und Loeve benannte Hauptkomponentenanalyse (engl. PCA: principle component analysis) ist ein statistisches Mittel, um eine Merkmalsreduktion durchzuführen. Mit Hilfe der Hauptkomponentenanalyse lassen sich aus Daten mit vielen Eigenschaften die Hauptmerkmals-Faktoren extrahieren, die für die Eigenschaften bestimmend sind. Es wird eine Hauptachsentransformation durchgeführt, indem man die Korrelation mehrdimensionaler Merkmale minimiert, durch Überführung in einen Vektorraum mit neuer Basis. Die MFCC wurden in dieser Arbeit in einem 25 Millisekunden Zeitfenster abgetastet mit 5 Millisekunden Fensterverschiebung von links nach rechts. Durch diese Methode erhält man für ein Phon eine Große Anzahl an Merkmalen. Durch die Anwendung der Hauptachsentransformation wird der mehrdimensionale Variablenraum auf einen zweidimensionalen reduziert. Die x-Achse gibt dann die Hauptkomponenten der MFCC wieder und die y-Achse repräsentiert die Anzahl der Kurzeitanalysen, welche sich auf 1 reduziert. Abbildung 24 zeigt diese Hauptachsentransformation anhand der MFC-Koeffizienten des Lautes /a:/. Im ersten Bild sieht man 24 MFC-Koeffizienten (12 MFCC zuzüglich der Delta MFCC), wie sie auch im Korpus für ein bestimmtes /a:/ vorkommen. Auf der x-Achse ist die zeitliche Abfolge der MFC-Koeffizienten abgetragen und auf der y-Achse der zugehörige MFC-Koeffizienten-Wert. Extrahiert man jetzt für jedes Zeitfenster die MFCC, bekommt man eine große Anzahl von Merkmalen. Die Hauptachsentransformation führt eine Merkmalsreduktion, der MFC-Koeffizienten für dieses /a:/ durch.

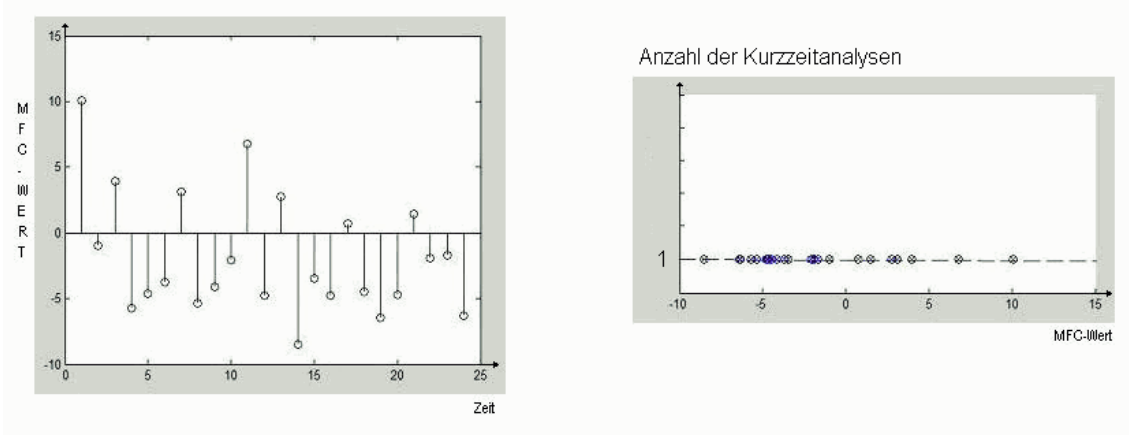


Abbildung 24: Hauptkomponentenanalyse der Mel-Cepstrum-Koeffizienten.

Bild 1: MFC-Extraktion innerhalb eines Zeitfensters (x-Achse = Zeit; y-Achse = MFC-Wert);

Bild 2: nach KL-Transformation aller MFC (x-Achse = MFC-Wert)

Abbildung 24 zeigt dies anhand der Punkte entlang der x-Achse. Die Hauptachsentransformation lässt sich durch eine Matrix angeben, die aus den Eigenvektoren der Kovarianzmatrix gebildet wird. Formal berechnen sich die Hauptkomponenten nach:

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad \mathbf{X} - \bar{\mathbf{X}} :$$

Hierbei wird zuerst der Mittelwert berechnet und durch Subtraktion des Mittelwertes $\bar{\mathbf{X}}$ von den originalen Werten \mathbf{X} wird ein Normalisieren (mittelwertsfrei) der Daten erreicht. \mathbf{X} repräsentiert die Menge der ursprünglichen Werte, hier also die MFC-Koeffizienten. Danach wird die Berechnung der Varianz und Kovarianz durchgeführt. Die Varianz berechnet sich nach:

$$\text{var}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})$$

Die Kovarianz berechnet sich nach:

$$\text{cov}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{Y}_i - \bar{\mathbf{Y}})$$

Die letzte Hauptachsentransformation sowie die entsprechenden Hauptkomponenten erreicht man durch die Berechnung des Eigenvektors multipliziert mit den normalisierten Daten. Hierbei wird der Eigenvektor wie auch die normalisierten Daten transponiert. Formal ergibt sich:

$$\text{Hauptkomponenten} = \text{Eigenvektor}^T \times \text{normalisierte Daten}^T$$

Die im Korpus aufgenommenen Sprachsegmente besitzen wie im Abschnitt zuvor erläutert unterschiedliche spektrale Eigenschaften, welche in dieser Arbeit durch die Mel-Cepstrum-Koeffizienten repräsentiert sind. Bei der Auswahl der Sprachsegmente ist der spektrale Abstand entscheidend für eine störungsverminderte Konkatenation. Zur Berechnung der spektralen Übergangskosten sollen die Sprachsegmente ausgewählt werden, die in ihrem Abstand zwischen den Mel-Cepstrum-Koeffizienten minimale Werte haben. In dieser Arbeit wird zur Reduktion der Mel-Cepstrum-Koeffizienten eine Hauptkomponentenanalyse durchgeführt. Zur Dimensionsreduktion werden dann die ersten 8 Mel-Cepstrum-Koeffizienten verwendet.

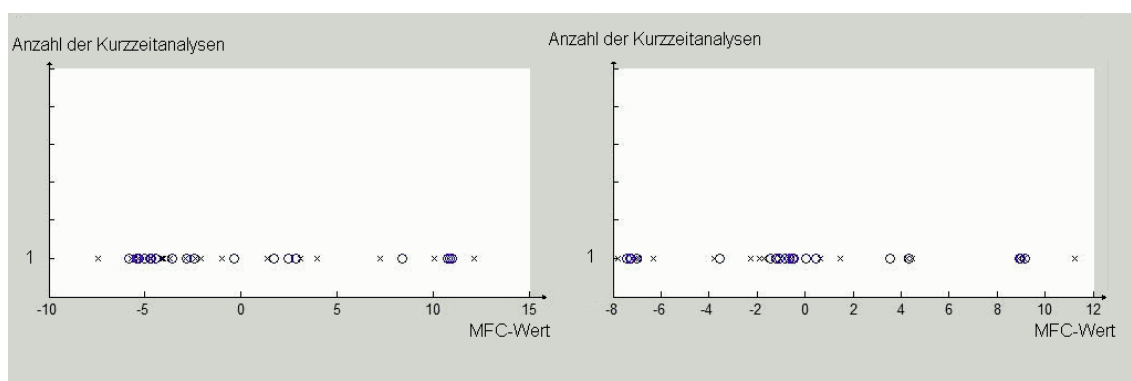


Abbildung 25: Hauptkomponentenanalyse von zwei unterschiedlich realisierten /a:/

Abbildung 25 zeigt die Transformation der MFCC von zwei verschiedenen /a:/, die im Sprachdaten-Korpus an unterschiedlicher Stelle auftreten. Hierzu wurde eine Hauptkomponenten Analyse der MFCC zweier /a:/ durchgeführt. Da zu jedem Zeitfenster (25 Millisekunden) im Sprachsignal /a:/ 12 MFCCs extrahiert werden, bekommt man bei einer durchschnittlichen Phonlänge von 50 Millisekunden, bei einer Verschiebung des Fensters um 5 Millisekunden von links nach rechts, 108 Koeffizienten. Diese MFCC werden aus einer mehrdimensionalen Ebene (9 x 12) auf eine zwiedimensionale Ebene transformiert

(1 x MFCC). Diese merkmalsreduktion eignet sich zur Berechnung des Abstandes zwischen den MFCC bei der Konkatenation der Sprachsegmente. Der Unterschied der zwei /a:/ ist hierdurch ebenso graphisch gut darstellbar, was eine gute Analysemöglichkeit bietet. Entlang der x-Achse werden die MFCC dargestellt und die y-Achse repräsentiert die Anzahl der Kurzzeitanalysen, die von ursprünglich neun auf eins reduziert wurden. Für die Weiterverarbeitung werden dann die ersten 8 MFC-Koeffizienten verwendet.

5.5 Bedingte Wahrscheinlichkeitsmodelle zur Segmentauswahl

Der Ausgangspunkt der Überlegungen für eine Auswahl der Sprachsegmente mittels nicht generativer Wahrscheinlichkeitsmodelle stützt sich auf der Annahme, dass die zeitliche Abfolge bestimmter Segmente durch eine bedingte Wahrscheinlichkeit erfasst werden kann. Bei einem generativen Modell wie dem HMM wird die Verbundwahrscheinlichkeit durch $p(x, y)$ berechnet, wobei x die Variablen der Zustände und y die Beobachtung entsprechend zu x widerspiegelt. Es werden nachfolgend Modelle für die Auswahl der geeigneten Sprachsegmente beschrieben, die dieses Problem übergehen und ebenso die Voraussetzung erfüllen, keine Unabhängigkeitsannahmen zwischen den Variablen besitzen zu müssen. Ein Modell, welches beide Voraussetzungen realisieren kann, ist mit der bedingten Wahrscheinlichkeit gegeben. Die bedingte Wahrscheinlichkeit modelliert die zeitliche Abfolge der Sprachsegmente und der sich ergebenden Beobachtung durch $p(x|y)$. Dies hat zum Vorteil, dass die Ausgaben nicht erst modelliert werden müssen. Modelle, die auf der bedingten Wahrscheinlichkeit beruhen, weisen einer neuen Abfolge von Labels x eine Ausgabe Sprachsegment-ID y zu, indem die bedingte Wahrscheinlichkeit $p(x|y)$ maximiert wird. Wie auch bei der bisher verwendeten Strategie der Kostenterme für die Auswahl der Sprachsegmente werden diese mit Hilfe eines Graphen repräsentiert. Jeder Knoten spiegelt ein Sprachsegment wider.

Die nachfolgenden Modelle basieren auf gerichteten Graphen bzw. ungerichteten Graphen. Ein gerichteter Graph umfasst eine endliche Menge von Knoten V und eine endliche Menge von gerichteten Kanten E . Dabei verbindet jede Kante $e \in E$ genau einen Startknoten $v \in V$ mit genau einem Endknoten $w \in W$. Man sagt auch, Kante e führt von Knoten v nach Knoten w . Falls $v = w$ gilt, heißt e auch Schlinge. Dies trifft in unserem Falle der

Auswahl der Sprachbausteine nicht zu. Im Gegensatz dazu wird in Abschnitt 5.5.3 das Modell der bedingten Zufallsfelder, „Conditional Random Fields“ (CRF), eingeführt, welches auf ungerichteten Graphen basiert. Ein ungerichteter Graph umfasst eine endliche Menge von Knoten V und eine endliche Menge von ungerichteten Kanten E . Jede ungerichtete Kante $e \in E$ verbindet entweder zwei verschiedene Knoten $u, v \in V$ miteinander oder im Falle einer ungerichteten Schlinge einen Knoten $v \in V$ mit sich selbst. Die Beziehungen zwischen einer ungerichteten Kante und den durch sie verbundenen Knoten werden durch die Funktionen $a, b: E \rightarrow V$ ausgedrückt. $a(e)$ ist einer der beiden durch Kante e verbundenen Knoten, $b(e)$ der andere Knoten. Falls e eine Schlinge ist, gilt $a(e) = b(e)$.

Nachfolgend werden ein probabilistischer endlicher Automat sowie ein bedingtes Entropie-Modell beschrieben, die für die Auswahl der Sprachbausteine entwickelt wurden. Diese Modelle beziehen sich auf gerichtete Graphen. Das Modell der bedingten Zufallsfelder (CRF) bezieht sich auf einen ungerichteten Graphen. Bei beiden Modellen sind die Labels der Sprachbausteine durch die Knoten repräsentiert und die Kanten durch die bedingte Wahrscheinlichkeit.

5.5.1 Probabilistischer endlicher Automat

Für die Konkatenation der Sprachbausteine wird eine zeitliche Abfolge der Sprachlaute betrachtet. Dies kann mittels eines gerichteten Graphen gut rekonstruiert werden. Das graphische Modell kann mathematisch durch einen endlichen Automaten dargestellt werden. Der endliche Automat hat eine endliche Anzahl von Zuständen, wobei jeder Zustand einem Sprachsegment entspricht. Für die Unit-Selection-basierte Sprachsynthese wird das Modell des endlichen Automaten durch eine endliche Menge an Zuständen erfüllt, da das Korpus eine endliche Menge von Phonemen, Silben und Wörtern besitzt, mit denen eine sinnvolle Äußerung erreicht werden kann. Die Zustände sind die zeitlich aufeinander folgenden Labels, also das, was letztlich synthetisiert werden soll. Die zweite Bedingung für einen endlichen Automaten ist die zeitliche Ordnung der Zustände. Weiterhin wird der endliche Automat durch eine endliche Menge von Eingabesymbolen (Alphabet), determiniert, die nur phonetisch transkribierten Text als Eingabe zulässt. Der endliche Automat liest bei jedem Schritt ein Eingabesymbol und springt entsprechend seiner Übergangs-

relation von einem Zustand zu dem nächsten Zustand. Bei der Unit-Selection ist dies der Übergang von einem Sprachsegmentlabel zum nächsten, welches zur Auswahl steht. Abbildung 26 zeigt einen endlichen Automaten. Wenn der endliche Automat, wie er hier verwendet wird, eine Ausgabe erzeugt, spricht man von einem Transduktor, der ein Eingabealphabet in ein Ausgabealphabet überführt. Das Eingabealphabet sind wie zuvor erwähnt die Labels des Eingabetextes, der synthetisiert werden soll. Das Ausgabealphabet sind die jeweilig zugehörigen IDs, die die Position der Sprachsegmente im zugrunde liegenden Sprachdaten-Korpus widerspiegeln.

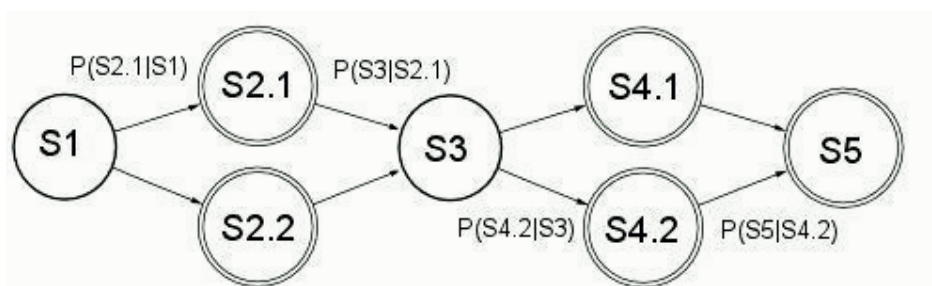


Abbildung 26: Probabilistischer endlicher Automat: S1 gibt das erste Segment an.

Formal wird der Transduktor durch ein 6-Tupel der Form $\langle \Sigma, \Delta, S, s_0, F, E \rangle$ bestimmt, wobei folgende Bedingungen erfüllt sein müssen.

Σ und Δ sind zwei endliche Mengen. Σ repräsentiert den phonetisch transkribierten Eingabetext und Δ die Merkmalsvektoren der Sprachsegmente mit der entsprechenden Identifikation zur Auffindung im Korpus.

S ist eine endliche Menge an Zuständen

$s_0 \in S$ gibt den Startzustand an

$F \subseteq S$ ist die Menge der Endzustände

$E \subseteq S \times \Sigma^* \times \Delta^* \times S$ ist eine endliche Menge an Kanten.

Der endliche Automat ist ein gerichteter Graph mit Kanten, die den Übergang von einem Zustand zum nächsten bilden. Die Kanten sind gewichtet mit Übergangs-

Wahrscheinlichkeiten, welche Aussagen treffen, wie wahrscheinlich der Übergang von einem Zustand, also Label, zum aktuellen Zustand ist. Die Wahrscheinlichkeiten ergeben sich aus der Verteilung der Sprachsegmente und wie sie im Korpus auftreten. Als Grundlage dient hier die bedingte Wahrscheinlichkeit, wie sie in Kapitel 3 dargestellt wurde. Die Kanten spiegeln also die Wahrscheinlichkeit wider, mit der die Sprachsegmente aufeinander folgen, abhängig vom nächsten Label. Die Verbundwahrscheinlichkeit kann faktorisiert werden zu einem Produkt bedingter Wahrscheinlichkeiten.

$$p(v_1^s, v_2^s, \dots, v_n^s) = \prod_{i=1}^n p(v_i^s | v_{\pi_i}^s)$$

wobei v_i^s jeweils einem Sprachbaustein-Label entspricht und $v_{\pi_i}^s$ die Labels kennzeichnet die, die v_i^s vorausgegangen sind.

5.5.2 Bedingte-Entropie-basierte Segmentauswahl

Mittels der bedingten Entropie wird ein graphisches Modell entwickelt, welches die Wahrscheinlichkeiten der Sprachsegmentabfolge abhängig von dem jeweils vorhergehenden Sprachsegment berechnet und ähnlich dem probabilistischen endlichen Automaten ist. Wie in Abschnitt 3.4 Entropie respektive 3.4.1 Bedingte-Entropie erläutert, kann mit Hilfe der Entropie der Informationsgehalt wahrscheinlichkeitstheoretisch angegeben werden. Der Informationsgehalt aufeinander folgender Sprachsegmente wird gemessen, indem nicht die Entropie einer bestimmten Variablen, welche ein bestimmtes Sprachsegment repräsentiert, berechnet wird, sondern die Entropie des Sprachsegments abhängig von den vorausgehenden Sprachsegmenten. Im Gegensatz zu dem generativen HMM, bei dem $P(X, Y)$ berechnet wird, beschreibt die bedingte Entropie das Modell durch $P(Y | X)$. Die bedingte Entropie berechnet sich nach:

$$H(Y | X) = - \sum_{x \in X} \sum_{y \in Y} P(x, y) \log P(y | x)$$

wobei „ \log “ den „logarithmus dualis“ wiedergibt. Die Sprachsegmentabfolgen werden in ihrem zeitlichen Auftreten durch Merkmalsvektoren beschrieben, wie sie in den vorange-

gangenen Abschnitten angeführt wurden. Der Unterschied zu dem probabilistischen endlichen Automaten ist, dass das Bedingte-Entropie-Modell einem probabilistischen endlichen Akzeptor entspricht, bei dem die maximale Entropie eine Sequenz an Sprachbaustein-IDs ausgibt. Bei dem probabilistischen endlichen Automaten kann es häufig der Fall sein, abhängig von der Größe des verwendeten Sprachdatenmaterials, dass der Übergang von einem zum nächsten Sprachsegment für unterschiedliche Sprachbausteine die gleiche Wahrscheinlichkeit besitzt. Dies erschwert die bestmögliche Auswahl eines geeigneten Sprachsegments. Aus diesem Grund wurde die bedingte Entropie als Gewichtung für die Kanten verwendet, die dieses Phänomen verringert. Die Auswahl der Sprachbausteine ergibt sich dann durch die Maximierung der Entropie über den Gesamtgraphen.

Es ergibt sich also:

$$S = \arg \max H(Y | X)$$

S repräsentiert die Sprachbausteine, die letztlich die zu synthetisierende Äußerung erzeugen.

Abbildung 27 zeigt die Einhüllende des Sprachsignals im zeitlichen Verlauf, das Spektrum, sowie die F0 Kurve des synthetisierten Signals. Die Darstellung zeigt das synthetisierte Sprachsignal, wobei die Sprachbausteine mittels bedingter Entropie ausgewählt und konkateniert wurden. Die untere Darstellung zeigt das generierte Sprachsignal, wobei die Sprachsegmente mit der zwei-dimensionalen Kostenfunktion (Hunt, Black 1996; Stöber 2002) ausgewählt wurden. Es wurde der Satz: „Nehmen Sie dann um zehn Uhr das Flugzeug nach Frankfurt?“ synthetisiert. Es ist in der Abbildung gut zu erkennen, dass am Ende der Grundfrequenzverlauf in der oberen Darstellung dem Fragesatz entspricht. Im Spektrogramm erkennt man, dass nach „um“ der Übergang an der Konkatenationsstelle in der oberen Darstellung weicher ist und somit der Übergang von einem Sprachsegment zum anderen keiner größeren Störung unterliegt. Daraus lässt sich schließen, dass die ausgewählten Segmente besser zueinander passen.

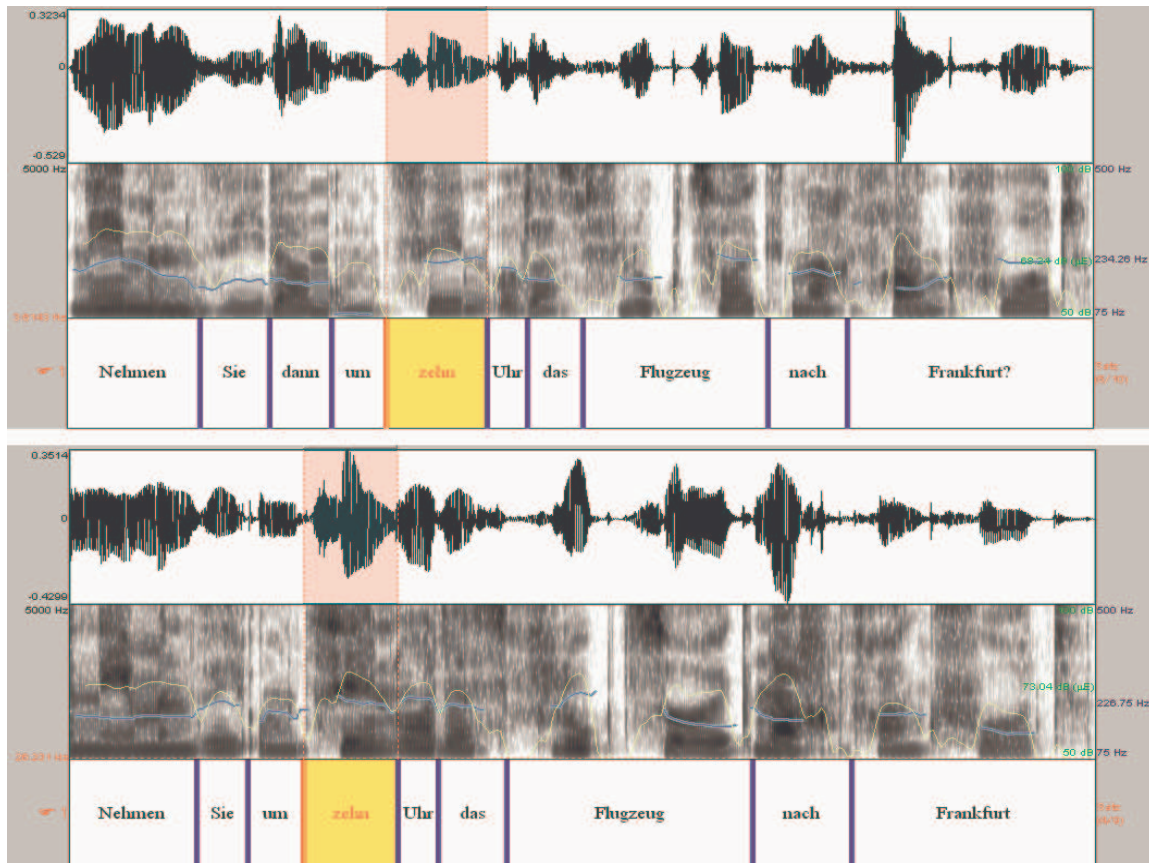


Abbildung 27: Darstellung zweier unterschiedlich generierter Sprachsignale.
 Obere Darstellung: Segmentauswahl und Konkatination mittels bedingter Entropie;
 Untere Darstellung: Segmentauswahl und Konkatination mit zwei-dimensionaler Kostenfunktion

5.6 Bestimmungen optimaler Sprachbausteinfolgen

Die Bestimmung der optimalen Sprachbausteinfolgen kann als Suchproblem verstanden werden. Bei allen entwickelten und angewendeten Verfahren wird die optimale Bausteinfolge anhand einer Maximierung von Wahrscheinlichkeiten durchgeführt. Bei der graphen-basierten Repräsentation der Sprachsegmente werden die Knoten durch Kanten verbunden, wobei jede Kante eine Wegstrecke zwischen den Knoten darstellt. Um den optimalen, bzw. kürzesten Weg in einem graphen-basierten Netzwerk zu finden bedient man sich den Kürzeste-Pfad-Algorithmen. Ein bekannter und häufig eingesetzter Algorithmus zur Berechnung des besten Pfades (hier: bester Pfad = kürzester Pfad) ist der A*-Algorithmus. Eine genaue Betrachtung der Funktionsweise des A*-Algorithmus wird in Abschnitt 5.6.2 wiedergegeben. Bei der Problemstellung, die besten Sprachsegmente in einem Sprachsegmentnetzwerk zu finden, wird die Wegstrecke zwischen den Knoten

durch Wahrscheinlichkeiten beschrieben. Die Wahrscheinlichkeiten zwischen den Knoten kann beim A*-Algorithmus also als Wegstrecke betrachtet werden, wie weit ein Knoten bzw. Sprachsegment vom anderen entfernt liegt. Der Viterbi-Algorithmus wie auch der A*-Algorithmus wurden erfolgreich u.a. in der Spracherkennung eingesetzt. Der Viterbi-Algorithmus stammt aus der Gruppe der dynamischen Programmierung. Beide Algorithmen wurden verwendet, wobei der Viterbi-Algorithmus bei dem CRF-basierten Unit-Selection-Ansatz verwendet wurde und der A*-Algorithmus bei dem Ansatz des probabilistischen endlichen Automaten und dem Ansatz der Auswahl von geeigneten Sprachbausteinen mittels der bedingten Entropie.

5.6.1 Viterbi-Algorithmus

Der Viterbi-Algorithmus dient dazu, auf effiziente Weise die beste Pfadsequenz durch ein Graphen-Modell zu finden. Die formale Spezifikation des Algorithmus nach Manning et al. (2000, S. 350) lautet:

```

Initialization
 $\delta_1(\Omega) = 1.0$ 
 $\delta_1(t) = 0.0$  for  $t \neq \Omega$ 
Induction
for  $i := 1$  to  $n$  step 1 do
  for all tags  $t^j$  do
     $\delta_{i+1}(t^j) := \max_{1 \leq k \leq T} [\delta_i(t^k) P(w_{i+1}|t^j) P(t^j|t^k)]$ 
     $\psi_{i+1}(t^j) := \arg \max_{1 \leq k \leq T} [\delta_i(t^k) P(w_{i+1}|t^j) P(t^j|t^k)]$ 
  end
end
Termination
 $X_{n+1} = \arg \max_{1 \leq j \leq T} \delta_{n+1}(j)$ 

for  $j := n$  to 1 step -1 do
   $X_j = \psi_{j+1}(X_{j+1})$ 
End
 $P(X_1, \dots, X_n) = \max_{1 \leq j \leq T} \delta_{n+1}(t^j)$ 

```

Der Viterbi-Algorithmus ist ein weitverbreiteter Algorithmus und wird in unterschiedlichsten Anwendungen eingesetzt. Die Informationstheorie, Spracherkennung, Bioinfor-

matik und Computerlinguistik verwenden den Viterbi-Algorithmus. So ist er in der Spracherkennung nicht mehr wegzudenken. Der Viterbi-Algorithmus nutzt die Eigenschaft, welche aus der dynamischen Programmierung stammt, und berechnet die einzelnen Teilpfade, wobei nur der beste Teilpfad verwendet und gespeichert wird.

5.6.2 A*-Algorithmus

Der A*-Algorithmus dient der Berechnung des kürzesten Pfades zwischen zwei Knoten in einem Graphen. Hier entspricht jeder Knoten einem Sprachsegment und jede Kante, die zwei Knoten verbindet, entspricht der Übergangswahrscheinlichkeit. Der A*-Algorithmus implementiert eine so genannte informierte Suche. Der A*-Algorithmus greift auf eine Heuristik zurück, um zielgerichtet zu suchen, und untersucht dabei zuerst jene Knoten, welche am wahrscheinlichsten zum Endsegment führen. Abbildung 28 zeigt einen Sprachsegment-Graphen, auf dem der A*-Algorithmus operiert.

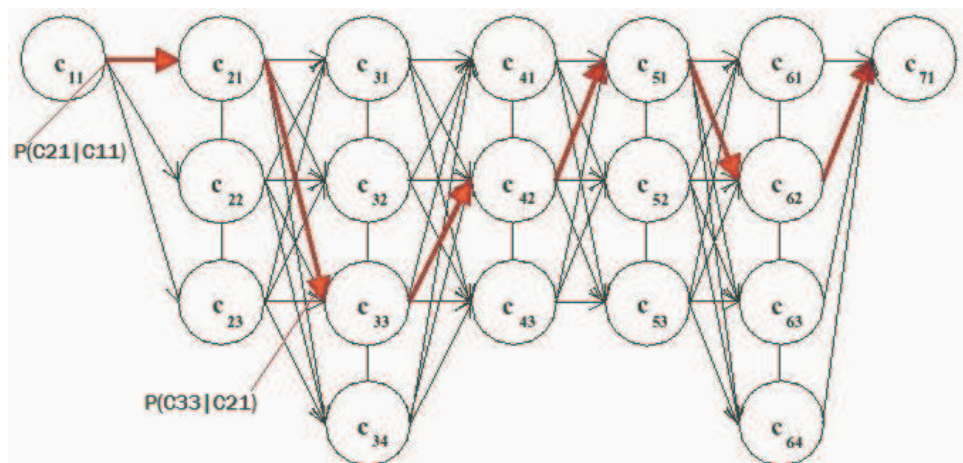


Abbildung 28: Sprachsegment-Graph auf welchem der A*-Algorithmus die Suche vom Startsegment (C11) zum Endsegment (C71) durchführt.

Wendet man den A*-Algorithmus auf einem Knoten S an, so werden zuerst alle von diesem Knoten aus erreichbaren Nachbarn berechnet. Danach wird durch die Heuristik für jeden dieser Nachbarn eine Schätzung abgegeben, wie „teuer“ es ist, von ihm aus zum Ziel zu kommen. Für jeden dieser Nachbarknoten S_n addiert der A*-Algorithmus nun die von der Heuristik geschätzten Kosten bis zum Ziel, um vom Knoten S zu jenem Knoten S_n zu kommen. Die Vorgehensweise des Algorithmus lässt sich dabei durch folgende Gleichung beschreiben:

$$F(S_n) = G(S_n) + H(S_n)$$

Hierbei steht $H(S_n)$ für die von der Heuristik für Knoten S_n geschätzten Kosten bis zum Ziel, $G(S_n)$ für die bisherigen gesamten Wegkosten, um vom Knoten, bei dem man den A*-Algorithmus gestartet hat, zum Knoten S_n zu kommen, und $F(S_n)$ steht für die wahrscheinlich zu erwartenden Kosten, wenn man vom Startknoten aus über seine aktuelle Position (S) zu dem entsprechenden Nachbarn (S_n) weitergeht, um von dort aus irgendwie weiter zum Ziel zu gelangen. Im nächsten Schritt wird nun der Knoten weiter untersucht, welcher den geringsten F-Wert besitzt. Nachfolgend die formale Spezifikation des A*-Algorithmus (vgl. Schukat-Talamazzini 1995, S. 246):

```

Sentence of length n
Initialization
 $O = K_\alpha$  (Menge aller Startknoten);  $K_\alpha \in K$ 
Remove O from best vertices k
IF  $k \in K$ 
    Return k
End
Else
     $\hat{f}(k')$  for all  $k' \in K, k < k'$ 

IF
 $k' \notin O$ 
Sort  $k'$  in  $O$ 
IF
 $k' \in O$ 
Correct costs  $\hat{g}(k'')$  of all vertices  $k'' \in O$ 
Repeat

```

$K = \{k_1, k_2, k_3, \dots\}$: beliebige Knotenmenge

$T \subseteq K \times K$: Menge von Kanten

$d : T \rightarrow \mathbb{R}^+$: nicht negative Kostenfunktion

$\hat{f}(k')$ for all $k' \in K, k < k'$ wobei $f(k)$ eine nicht näher spezifizierte heuristische Funktion bezeichnet, die die Erfolgchance abschätzt, um die geordnete Suche weiterzuführen oder nicht.

$k < k'$ bezeichnet die Präzedenzrelation: $k_1 < k_2$ falls Kante von k_1 zu k_2 führt.

Der A*-Algorithmus braucht, um einen kürzesten Weg zu finden und zurückzugeben, fünf Eingaben: den Graph, welcher die Sprachsegmente repräsentiert und auf welchem er operieren soll, des Weiteren das Startsegment, von dem aus die Suche gestartet werden soll, das Endsegment, welches das Ende der zu synthetisierenden Äußerung repräsentiert, zu dem ein kürzester Pfad gefunden werden soll. Die Werte der Segmente, die der Algorithmus bereits kennt (und deren F-Werte daher bereits bekannt sind), aber noch nicht besucht hat, werden in einer Prioritätswarteschlange (W) gespeichert. Die Heuristik H schätzt für alle Segmente die Entfernung bis zum Endsegment ab.

5.6.3 Spektrale Abstandsmaßberechnung

Die spektrale Abstandsmaßberechnung dient einer optionalen Komponente, da Tests gezeigt haben, dass die in Abschnitt 5.5.1 und Abschnitt 5.5.2 entwickelten Verfahren unter bestimmten Umständen mehrere Sprachsegmente mit der gleichen Wahrscheinlichkeit ausgeben. Dies kommt vor allem bei Sprachsegmenten vor, die eine hohe Häufigkeit im Sprachdatenkorpus besitzen. Um die Auswahl des optimalen Sprachbausteins zu gewährleisten, wurde zusätzlich der spektrale Abstand der zu verkettenden Sprachsegmente errechnet. Hierbei wurde zum einen der Euklidische Abstand verwendet und als Alternative der Mahalanobis-Abstand.

Der Euklidische Abstand errechnet sich nach:

$$d(x, y) \equiv \sqrt{\sum_{r=1}^n (a_r(x) - a_r(y))^2}$$

und ist sehr leicht zu berechnen. Mittels des Euklidischen Abstandes kann ein absolutes Maß errechnet werden, das allerdings keine Varianz und Kovarianz der einzelnen Merkmale berücksichtigt. Hierzu eignet sich der statistisch motivierte Mahalanobis-Abstand.

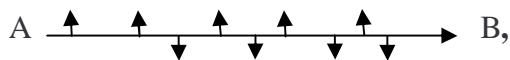
Der Mahalanobis-Abstand errechnet sich nach:

$$d(x, y) = (\underline{x} - \underline{y})^T S^{-1} (\underline{x} - \underline{y})$$

wobei x der (MFCC)-Datenvektor ist und y den Mittelwert des (MFCC)-Datenvektors des vorangegangenen Segments repräsentiert. Beide Vektoren werden transponiert und mit S^{-1} , der inversen Kovarianz von x , faktorisiert. Die m Koeffizienten werden als m -dimensionaler Spaltenvektor dargestellt. Diese werden als Realisation eines Zufallsvektors X mit der Kovarianzmatrix Σ verwendet. Die Kovarianzmatrix ergibt den Abstand zweier so verteilter x und y .

5.7 Conditional-Random-Field-basierte Segmentauswahl

Zufallsfelder (Random Fields) kommen aus der Theoretischen Physik und erlangten große Bekanntheit durch das sogenannte Ising-Modell. Das Ising-Modell ist benannt nach dem Physiker Ernst Ising⁷. In diesem Modell wird vereinfacht dargestellt, eine Zeitreihe von A nach B betrachtet, die bei jeder Zeiteinheit eine binäre Ausrichtung -1, +1, besitzt.



der Abschnitt von A nach B besitzt eine bestimmte Konfiguration an +1 und -1 Ausrichtungen. Jede Konfiguration besitzt eine Wahrscheinlichkeit bezüglich aller potentiell eintretenden Konfigurationen. Eine solche Konfiguration wird dann als Zufallsfeld bezeichnet.

Ein bedingtes Zufallsfeld (engl. Conditional Random Field: CRF) (Sutton, McCallum, 2006) leitet sich aus den bedingten Wahrscheinlichkeitsmodellen ab. CRFs sind Modelle, die im Gegensatz zu den zuvor genannten Modellen auf ungerichteten Graphen basieren. Formal ist ein Zufallsfeld, engl.: Random Field, ein ungerichteter Graph $G = (V, E)$ mit den Zufallsvariablen $Y = (Y_v)_{v \in V}$. Abbildung 29a zeigt schematisch einen ungerichteten Graphen. In Abbildung 29b ist der Zusammenhang zwischen dem Beobachtungszustand und der Ausgabe dargestellt.

⁷ Ernst Ising (05/1900 – 05/1998): Deutscher Physiker, entwickelte das Ising-Modell (1925) zur Beschreibung von Ferromagnetismus. Eine Abhandlung findet sich in: Brush, S.: The Lenz-Ising Model, 1967

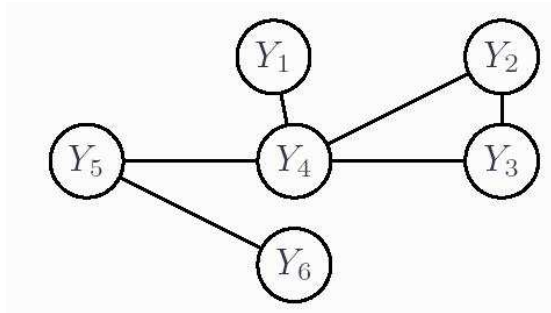


Abbildung 29a: Ungerichteter Graph

Während bei generativen Modellen, wie dem HMM, die Ausgabe unmittelbar von der Beobachtung abhängt, ist bei den ungerichteten Graphen-Modellen wie dem CRF der Zustand abhängig von der Beobachtung (X) und zusätzlich von den vorhergehenden Ausgaben (Y). Ein Zustand hängt also von allen benachbarten Zuständen ab. Dies ermöglicht, vorhergehende Sprachsegment-Labels über die gesamte zeitliche Abfolge der Sprachsegment-Labels hinweg zu betrachten.

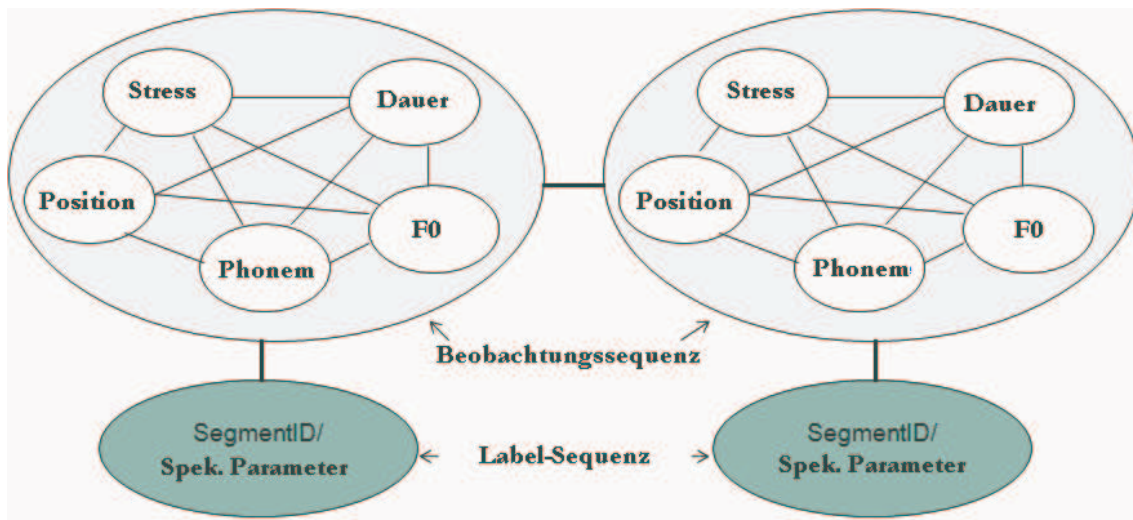


Abbildung 29b: Beobachtungs- und Zustandssequenz mit zugehöriger Konfiguration

Einzig das Zustands- und Beobachtungsfenster muss vorher bestimmt werden, um den Kontext festzulegen. Dies hat zum Vorteil, dass die vorangegangenen Labels mitberücksichtigt werden und somit eine Abschätzung der zu generierenden Ausgabe möglich wird.

CRFs berechnen sich exponentiell und ordnen unterschiedliche Gewichte den einzelnen Merkmalen zu. Lafferty (2001) geben die Berechnung der Wahrscheinlichkeit einer Segmentabfolge wieder durch:

$$p(y | x, \lambda) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j F_j(y, x)\right)$$

wobei $Z(x)$ ein Normalisierungsfaktor ist und λ das zugehörige Modell der Merkmalsfunktion, welche die Parameter für X und Y schätzt. Die Beobachtungssequenz, Sprachsegment (siehe Abbildung 29b, SegmentID: dieses Sprachsegment kann je nach Ebene Wort, Silbe oder Phonem sein) mit der zusätzlichen Information für Startzeit und Endzeit des Sprachsegments, wird mit x angegeben. y ist die Sprachsegment-Label-Sequenz bei gegebenem x (vgl.: Abbildung 29b). Die Funktion, welche die einzelnen Merkmale des Sprachbausteins beinhaltet, wird durch

$$F_j(y, x) = \sum_{i=1}^n f_j(y_{i-1}, y_i, x, i)$$

angegeben und trifft binäre Entscheidungen

$$f_j(y_{i-1}, y_i, x, i) = \begin{cases} b(x, i) \\ 0 \end{cases}$$

Die Merkmalsfunktion $b(x, i)$ beschreibt das Beobachtungsmerkmal x für eine Variable des Segments i wie dieses im Trainingsdaten-Korpus auftritt. Es wird die maximale Wahrscheinlichkeit berechnet abhängig von Bedingungen. Nachfolgend für das Merkmal Phonem dann für das Merkmal Dauer, etc..., in der Form:

$$f_{phoneme}^{PR}(s^{t-2}, s^{t-1}, s, s^{t+1}, s^{t+2}, o, t) = \begin{cases} 1 & \text{if } t = 0 \wedge Pho(s^t) = dh \wedge Pho(s^{t-2}) = \dots \\ & \wedge Pho(s^{t-1}) = \dots \wedge Pho(s^{t+1}) = ax \wedge Pho(s^{t+2}) = \dots \\ 0 & \text{otherwise} \end{cases}$$

dann für das Merkmal Dauer, in der Form:

$$f_{duration}^{PR}(s^{t-2}, s^{t-1}, s, s^{t+1}, s^{t+2}, o, t) = \begin{cases} 1 & \text{if } t = 0 \wedge Dur(s^t) = 4.6 \wedge Dur(s^{t-2}) = \dots \\ & \wedge Dur(s^{t-1}) = \dots \wedge Dur(s^{t+1}) = 4.2 \wedge Dur(s^{t+2}) = \dots \\ 0 & \text{otherwise} \end{cases}$$

Es wird also abhängig von den Merkmalen eines Segments innerhalb des gewählten Beobachtungsfensters die maximale Wahrscheinlichkeit eines Merkmals berechnet, abhängig von vorhergehenden Ereignissen. Die verwendeten Merkmale für jeden Sprachbaustein entsprechen den Merkmalen, wie sie auch für das HMM-basierte Training verwendet wurden. Das System wurde mit Hilfe der Mallet- Bibliothek (McCullum et al. 2002) trainiert. Es werden jeweils kontext-basierte Modelle auf Wortebene, auf Silbenebene und auf Phonebene erstellt. Der Kontext umfasst den vorangehenden Merkmalsvektor und den nachfolgenden. Insgesamt wird für das kontext-abhängige Training der CRFs ein 15 dimensionaler Merkmalsvektor verwendet. In Abbildung 30 ist der Ablauf zum Erstellen der kontextbasierten CRF-Modelle schematisch dargestellt. Aus dem Sprachdatenkorpus werden die kontextbasierten Etiketten erstellt. Die dynamischen Parameter Dauer und F0 werden vorhergesagt und in den Merkmalsvektor integriert. Das Training der CRF-Modelle wird anhand des GIS-Algorithmus (Iterative Scaling, Berger, DellaPietra 1996) durchgeführt. Hierbei werden die CRF-Parameter in einem iterativen Verfahren jeweils neu berechnet.

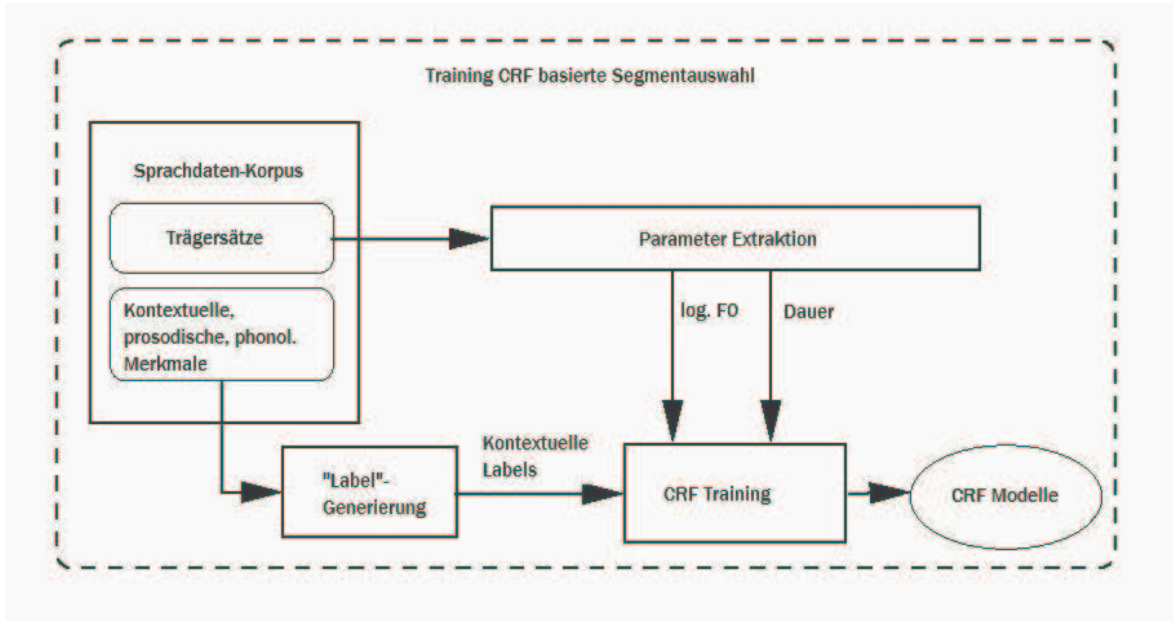


Abbildung 30: Schematische Darstellung des Trainings der kontextbasierten CRF Modelle

Die Eingabe für den iterativen Algorithmus zur Parameterschätzung sind die Merkmalsfunktionen $F_j(y, x) = \sum_{i=1}^n f_j(y_{i-1}, y_i, x, i)$. Die Ausgabe liefert die besten Parameterwerte $\lambda_1, \dots, \lambda_m$ für die jeweilige Funktion.

Formal wird der Algorithmus beschrieben durch:

1. Start with $\lambda_i = 0$ for each $i \in \{1, 2, 3, \dots, n\}$
2. Do foreach $i \in \{1, 2, 3, \dots, n\}$:

$$\Delta\lambda_i = \frac{1}{C} \log \frac{\tilde{E}(f_i)}{E(f_i)}$$

$$\text{Update } \lambda_i \leftarrow \lambda_i + \Delta\lambda_i$$

3. Goto Step2 if not all λ_i have converged.

Zu beachten ist, dass C eine Normalisierungskonstante ist, die für die empirische Verteilung $p(x, y)$ berechnet werden muss.

Zur Laufzeit werden aus dem Eingabetext die Merkmale extrahiert, die zum Aufbau der kontextbasierten Merkmalsvektoren verwendet werden. Die Merkmalsvektoren werden in einem hierarchisch organisierten Suchprozess den jeweiligen Modellen übergeben, die das entsprechende Sprachsegment anhand ID - Satz sowie Start- und Endzeit – ausgeben. Der Suchprozess beginnt auf der Wortebene.

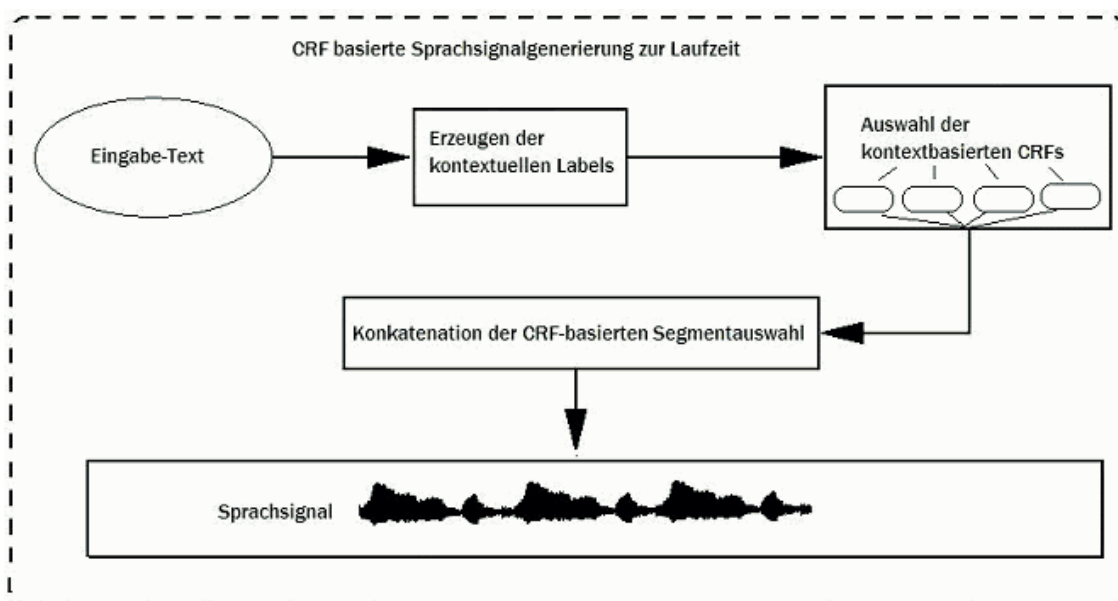


Abbildung 31: Schematische Darstellung der Sprachsignalgenerierung mittels CRFs

Wird kein Modell zu dem entsprechenden phonetisch transkribierten Eingabesegment gefunden, wird es auf Silbenebene weiterverarbeitet. Wird auch hier kein entsprechendes Etikett gefunden, wird auf Triphon, Diphon bis zur Phonebene das entsprechende Modell für das Segment ausgewählt. Die ausgewählten Segmente werden konkateniert und das gewünschte Sprachsignal wird ausgegeben. Abbildung 32 zeigt zwei generierte Sprachsignale mit ihrem Spektrum, F0-Verlauf (blaue Kurve) und Intensität (gelbe Kurve). Es wurde der Satz „Wollen Sie um zehn nach Frankfurt“ mit einem Synthesystem generiert, welches auf die zweidimensionale Kostenfunktion von Black et al (Black 1995, 1996) zurückgreift und mit dem CRF-basierten Algorithmus.

In der oberen Darstellung ist das mit CRF generierte Sprachsignal zu sehen und die untere Darstellung zeigt das Sprachsignal, welches mit der herkömmlichen Methode synthetisiert

wurde. Es ist in der Abbildung gut zu erkennen, dass die Dauer, und Intensität des Wortes „nach“ keine natürlichen Werte hat.

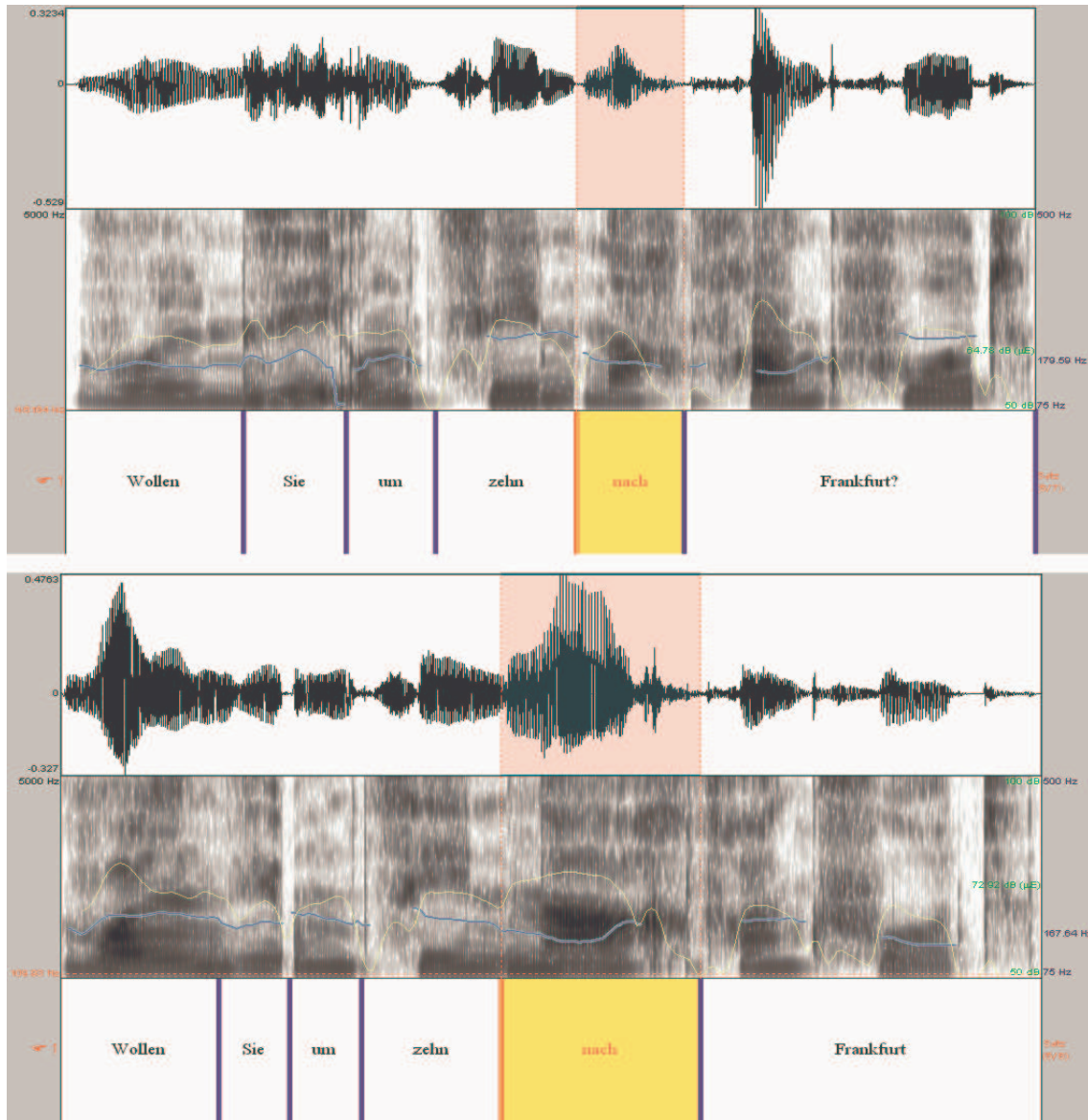


Abbildung 32: Darstellung zweier unterschiedlich generierter Sprachsignale.

Obere Darstellung: Segmentauswahl und Konkatination mittels Conditional Random Fields;

Untere Darstellung: Segmentauswahl und Konkatination mit zwei-dimensionalen Kostenfunktion

Auch die prosodische Realisierung am Satzende entspricht nicht der Realisierung, wie es das Fragezeichen in diesem Satz erfordert. Hier ist zu sehen, dass die mit CRF ausgewählten Einheiten ein eindeutig besseres Synthesergebnis liefern.

5.8 Fazit

Mittels der in Abschnitt 5.4 bis 5.7 entwickelten Verfahren werden Sprachsegmente ausgewählt, die hinsichtlich der Zielvorgaben geeignet sind, die gewünschte Äußerung durch Konkatenation der Sprachsegmente wiederzugeben. Das Erstellen eines probabilistischen endlichen Automaten für die Auswahl der zu verkettenden Sprachsegmente gestaltet sich recht einfach und effizient. Der Nachteil bei diesem Verfahren liegt darin, dass die Übergangswahrscheinlichkeiten mehrerer zur Auswahl stehender Sprachsegmente gleich hoch sein können. Dies führt zu einer nicht kontrollierbaren Auswahl und oftmals zur Auswahl von Sprachsegmenten, die nicht so gut geeignet sind als andere es wären. Aus diesem Grund wurde das Modell der bedingten Entropie entwickelt. Mittels des Entropie-Maßes kann eine feinere Abstufung gewährleistet werden und es können somit geeignetere Sprachsegmente ausgewählt werden. In Experimenten hat sich auch hier gezeigt, dass die Auswahl der Sprachsegmente hinsichtlich ihrer Merkmale nicht steuerbar ist. So müsste zusätzlich eine optimale Gewichtung der einzelnen Merkmale durchgeführt werden. Die Frage nach der optimalen Gewichtung der einzelnen Merkmale ist in der Sprachsyntheseforschung noch nicht eindeutig geklärt. So ist bisher nicht eindeutig geklärt, ob dem Merkmal Grundfrequenz ein höherer Stellenwert einzuräumen ist als dem Merkmal Dauer. Zwar wurde als Alternative zu der Auswahl mittels bedingter Entropie noch zusätzlich der spektrale Abstand zwischen aufeinanderfolgenden Segmenten berechnet, doch dies führt wiederum zu einer Erhöhung der Rechenzeit, welche zuvor mit den erstellten Modellen eingespart wurde. Der Nachteil, dass die einzelnen Merkmale nicht in die bestehenden Modelle so integriert werden können, dass hinsichtlich dieser eine verbesserte Auswahl möglich ist, hat zu der Entscheidung geführt, Conditional Random Fields einzusetzen. CRF's haben den Vorteil, dass sie nicht nur den vorhergehenden Zustand betrachten, sondern die gesamte Sequenz und die einzelnen Merkmale der Merkmalsvektoren berücksichtigen. Im Gegensatz zu konventionellen HMMs erster Ordnung, bei welchen nur der unmittelbar vorhergehende Zustand mit einbezogen wird.

Conditional Random Fields brauchen wegen ihrer exponentiellen Komplexität eine lange Trainingsphase. Das Training der CRFs auf Wortebene beanspruchte mit einer P-IV 512 MB Maschine ca. vier Tage.

KORPUSBASIERTE VISUELLE SYNTHESE

In diesem Kapitel wird die visuelle Sprachsynthese auf Basis von 2D Bilddaten beschrieben, wie sie für die audio-visuelle Synthese AVISS entwickelt wurde. Ausgehend von einer Abgrenzung verschiedener Algorithmen zur Erstellung von Talking-Heads wird das entwickelte Verfahren vorgestellt. Nachfolgende Abschnitte beschreiben den KNN-Algorithmus zur Bild-Segmentauswahl, die Parametergewinnung aus den Bilddaten, welche als Parameter für die metrische Distanz dienen und die abschließende Synchronisation des Audiosignals mit dem generierten Videosignal.

6.1 Visuelle Sprachsynthese

Um einen Talking-Head darzustellen muss zum einen eine Äußerung aus dem zugrundeliegenden Text erzeugt werden und zum anderen die korrespondierende Animation des Gesichtes bzw. das Videosignal mit den entsprechenden Videoframes, die die passende Mundbewegung generieren. Das Videosignal und das Sprachsignal werden dann Lippen-synchron ausgegeben. Realistische 2D-Darstellung so zu animieren, dass neue Äußerungen erzeugt werden können, ist eine nicht triviale Aufgabe. So müssen viele Einzelheiten wie Kopfstellung, Lippenstellung und Synchronisation berücksichtigt werden. Beim Menschen werden während eines Sprechvorgangs unterschiedliche Gesichtsaktivitäten ausgelöst. Dies ist darauf zurückzuführen, dass das menschliche Gesicht sehr komplex ist und eine sehr große Anzahl von Gesichtsmuskeln beim Kommunizieren beansprucht wird. So variiert die Lippenstellung jeweils im Kontext der zu sprechenden Äußerung, die Augenstellung sowie die Augenbrauen verändern ihre Position und die Wangenknochen und der Unterkiefer werden bewegt. Diese natürlichen Gegebenheiten erschweren die Modellierung eines Talking-Heads mit 2D fotorealistischen Sequenzen. Bei der Rekonstruktion von Videosequenzen durch neu zusammengesetzte Bildsequenzen werden Störungen an den Konkatenationsstellen sehr leicht vom Menschen wahrgenommen. Ist die Lippenstellung bzw. Kopfstellung nicht in einem natürlichen Fluss der Bildfolge wieder-

gegeben, so entsteht an dieser Verkettungsstelle ein Sprung. Dies lässt die Sequenz als unnatürlich erscheinen und stört die Gesamtwahrnehmung beim Nutzer. Für diese Art von Schwierigkeiten muss ein Verfahren genau das Folgebild in einer Sequenz identifizieren, welches die Sprünge an den Konkatinationsstellen minimiert. Bei dem nachfolgend beschriebenen Verfahren wurde eine solche Methode entwickelt, welche Bildsegmente aus dem zugrunde liegenden Videodatenkorpus extrahiert und zu einer neuen Äußerung zusammensetzt, indem ein KNN-Algorithmus die Segmente auswählt, welche eine geringe Distanz zueinander aufweisen. Dieses Verfahren hat den Vorteil, dass die Natürlichkeit des Sprechers erhalten bleibt und die Störungen an den Konkatinationsstellen im Videosignal, die Gesamtwahrnehmung des Nutzers nicht zu stark beeinträchtigt.

6.2 KNN-basierte Auswahl der Video-Frame-Segmente

Der K-Nächster-Nachbar Algorithmus (engl.: K-Nearest-Neighbor), weiterhin als KNN bezeichnet, ist ein Algorithmus, welcher im Gegensatz zu HMM ohne ein zugrunde liegendes Modell auskommt und deshalb oft auch als „Memory based learning“ oder Prototypen-basiertes Lernen bezeichnet wird. Die Funktionsweise von KNN basiert darauf, zu einem gegebenen Punkt x_0 die k Trainingspunkte $x_{(r)}, r=1, \dots, k$ zu finden, die den geringsten Abstand zu x_0 haben.

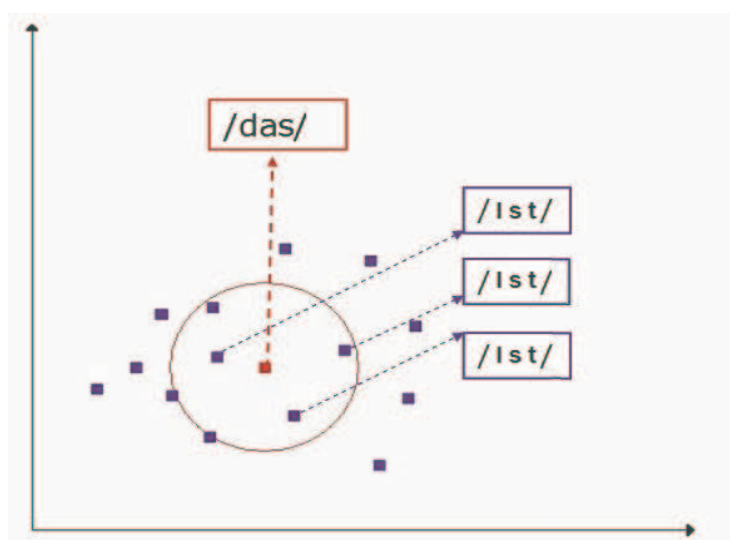


Abbildung 33: Schematische Darstellung der „Nächsten Nachbarn“, der in Betracht kommenden Segmente

Der KNN-Algorithmus dient zur Auswahl der Bildsequenzen für die Rekonstruktion des Talking-Heads aus fotorealistischen 2D Bilddaten, welche aus dem Videokorpus zur Laufzeit extrahiert werden. Das KNN-Verfahren ist eine Methode, die Beobachtungen im Trainingsdatenkörper τ als Ausgangspunkt für die Eingabevektoren verwendet, um das Videosegment durch \hat{Y} zu lokalisieren, wobei \hat{Y} die Kennzahl wiedergibt. Formal kann dies durch:

$$\hat{Y}(x) = \frac{1}{k} \sum_{\{i|x_i \in N_k(x)\}} y_i$$

wobei $N_k(x)$ die Nachbarn zu x definiert abhängig von den k nächsten Punkten. Zur Berechnung der nächsten Nachbarn wird der Euklid-Abstand als metrisches Abstandsmaß verwendet. In Abbildung 33 ist das KNN-Auswahlverfahren schematisch dargestellt. Das Segment, welches als /d a s/ die vorangehende Bildsequenz repräsentiert sowie die Segmente /I s t/, welche die zur Auswahl stehenden nachfolgenden Bildsequenzen wiedergeben. Mittels des KNN-Algorithmus muss nun die potentielle Bildsequenz des visuellen Segmentes /I s t/ ermittelt werden. Die Videosequenz /d a s/ endet mit einem 2D-Bild, welches als Referenzbild für die Distanzmetrik zwischen eben diesem 2D-Bild und dem Anfangsbild der Videosequenz von /I s t/ dient. Mit Hilfe des KNN wird das entsprechende visuelle Segment identifiziert, welches dann aus der Videospur extrahiert wird. Die nachfolgenden 2D-Bilddaten spiegeln den visuellen Verlauf der Äußerung „in Hannover“ wider. Hannover ist zusammengesetzt aus den Trivisem-Segmenten „Han“ und „nov“ und „er%“, wobei % ein Dummy darstellt, welches aus Gründen der Implementierung verwendet wurde, damit eine Einheitliche Trivisem-Abfolge simuliert werden kann. Ein Trivisem ist eine Abfolge von drei Visemen wie sie in Abschnitt 4.4.3 Tabelle 10 aufgeführt sind, und jeweils von einer Videoframesequenz dargestellt werden.

Abbildung 34 zeigt die Segmentabfolgenauswahl, wie sie von dem KNN-Algorithmus vorgenommen wird. Die schwarze Linie zeigt die letztlich ausgewählte Segmentabfolge, die graue Linie die potentiell in Frage kommenden Segmentabfolgen. Das letzte Bild in der Bildsequenz des ersten identifizierten Segments, welches anhand seiner visemischen

Transkription identifiziert wurde, dient als Referenz für die potentiellen nachfolgenden Segmente.

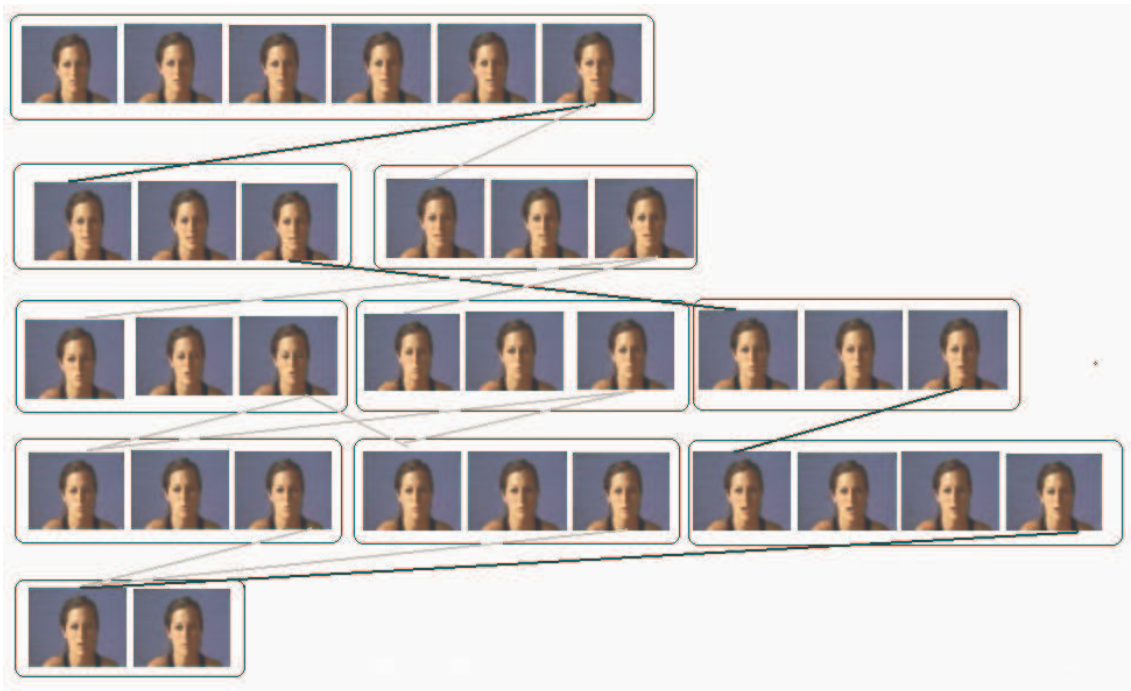


Abbildung 34: Schematische Darstellung der KNN Segmentabfolgenauswahl: „in Hannover“

Anhand dieser werden die k nächsten Nachbarn ausgewählt und eine Euklidische Distanzmetrik auf die Merkmalsvektoren angewendet, um das Folgesegment zu wählen, welches eine minimale Störung des Bildflusses ergibt. Im Idealfall wird eine Segmentabfolge im Videodatenkorpus gefunden, die genau der erwünschten Kombinationsfolge entspricht. Die Distanzmetrik des KNN-Algorithmus wird durch die extrahierten Pixel-Koordinaten-basierten Parameter vorgenommen. Die Parametergewinnung ist nachfolgend beschrieben.

6.3 Parametergewinnung für die Video-Segmentauswahl

Die Auswahl der Videosegmente, welche für die Rekonstruktion des Videosignals benötigt werden, müssen in einem Vorverarbeitungsschritt analysiert werden. Der Trägersatz, welcher als Videosignal vorliegt, wird hierzu annotiert. In einem ersten Schritt wird der Trägersatz visemisch transkribiert. Mit den in Kapitel 4 vorgestellten Visemklassen und

Transkriptionsverfahren werden die Grapheme in ihre entsprechenden Viseme umgesetzt und anhand der visemischen Transkription werden die potentiellen Segmente identifiziert. Die Annotation des zugrunde liegenden Videodatenkorpus umfasst ebenso die Festlegung der Zeitmarken, also Start- und Endzeitpunkt der entsprechenden Bildsequenz. Hier wurden die Zeitmarken gesetzt für die Wortebene und Trivisemebene. Tests haben gezeigt, dass die Auswahl von Einzelbildern und deren Konkatenation nicht die gewünschten Ergebnisse liefert und aufgrund der visuellen Koartikulation als unterste Auswahlenebene Triviseme dienen. Neben den Zeitmarken sind geometrische Merkmale entscheidend für eine geeignete Identifizierung von potentiellen visuellen Segmenten. Da geringe Störungen im zeitlichen Ablauf im Videosignal vom menschlichen Auge sofort registriert werden, ist es das Ziel, möglichst die Segmente auszuwählen, bei denen die Kopfposition nicht zu sehr von dem Ausgangsbild zum Folgebild abweicht. Hierzu wurde eine geometrische und pixelbasierte Merkmalsextraktion vollzogen. Aus den Videosignalen wurden die Einzelbilder extrahiert und ein mehrstufiger Filter darauf angewendet. In Abbildung 35 wird dieser Prozess schematisch dargestellt.

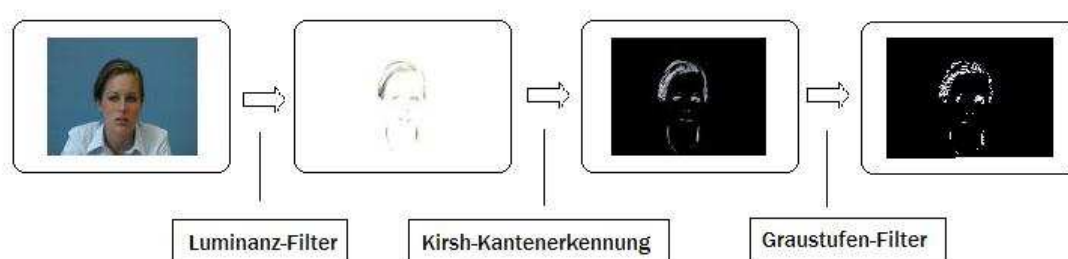


Abbildung 35: Schematische Darstellung Filter zur geometrischen und pixelbasierten Merkmalsextraktion

Auf das Originalbild wird zuerst ein Luminanz-Filter angewendet. Danach kommt ein Kantenerkennungsfiler zum Einsatz und zuletzt wird eine Umwandlung in ein Graustufenbild vollzogen. Für eine detaillierte Beschreibung der Arbeitsweise der Filter sei auf Jähne (1993) verwiesen. Aus dem verbliebenen Graustufenbild werden die Pixel-Koordinaten der weißen Farbpixel extrahiert. Zur Identifikation dieser dient der RGB-Wert 255, 255, 255 für die Pixel. Die X-Y Koordinaten der Pixel werden einer Hauptkomponentenanalyse unterzogen.

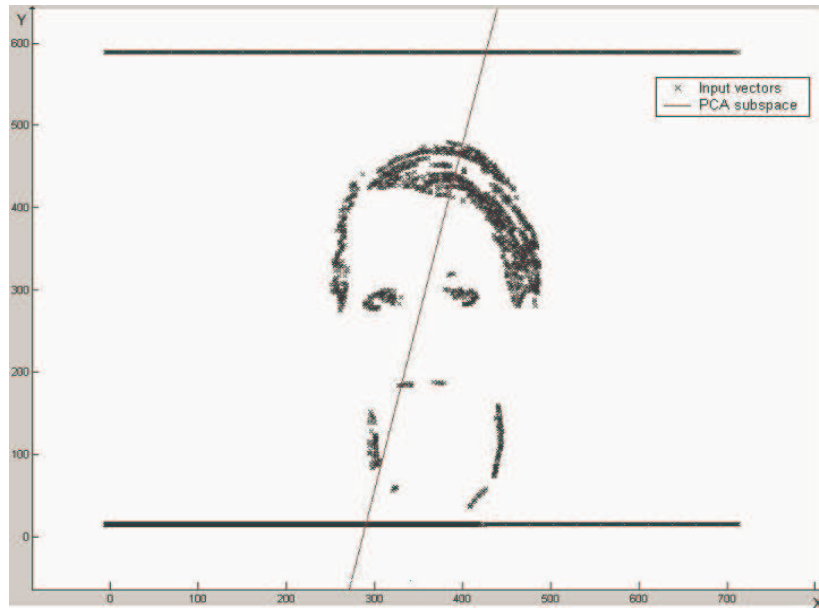


Abbildung 36: Eingabebild für die Hauptkomponentenanalyse und die durchgeführte Hauptachsentransformation.

Mit dieser Technik wird eine Hauptachsentransformation durchgeführt. Die verbleibenden Hauptkomponenten dienen als Merkmalsvektor für die Distanzmetrik des KNN-Algorithmus. Durch Minimierung der Distanz zwischen den auf End- und Start-Bild basierenden Parametern, der aufeinanderfolgenden Videosequenzen, wird das optimale visuelle Segment für die anschließende Konkatenation identifiziert. Abbildung 36 zeigt den Eingabe-Vektor und die resultierende Hauptachsentransformation mit dem Hauptkomponenten-Unterraum.

6.4 Audio-Video Konkatenation und Synchronisation

Um einen audio-visuellen Ausgabestrom zu erhalten, werden die Bilddaten mit dem synthetisierten Sprachsignal zusammengefügt. Die durch 2D-Bildabfolgen-Konkatenation generierte Videosequenz wird linear mit dem akustischen Signal synchronisiert. Die Synchronisation des Audiosignals und des Videosignals ist für die Natürlichkeit und der damit verbundenen Akzeptanz synthetischer audiovisueller Synthese entscheidend. Damit bei der visuellen Ausgabe die Lippenbewegungen mit den gesprochenen Einheiten übereinstimmen, wurde ein Verfahren verwendet, welches zur Laufzeit einen linearen Transitionsfaktor für die 2D-Bild-Übergänge im Videosignal berechnet. Dieser Faktor

liegt, bei einer Standardvideosequenz von einer Sekunde, zwischen 25 und 30. Der Transitionsfaktor ist definiert als der Wert des Quotienten Zeit/Framerate.

Die Synchronisation ist abhängig von der Anzahl der Bilddaten und der Zeit des Audiosignals. Zur Identifikation und der Extraktion der gewünschten Bilddaten wird die Zeit des Audiosignals berechnet:

$$T_g = \sum_{j=a}^e \Delta t(S_{ij})$$

T_g gibt die Zeit an und S repräsentiert das entsprechende Sprachsegment j in Trägersatz i . Die Startzeit des Sprachsegments wird durch a wiedergegeben und e gibt die Endzeit des Sprachsegments an.

Die Zeit wird mit der Anzahl der Bilder pro Sekunde multipliziert, und man erhält die Anzahl der Bilddaten welche extrahiert werden sollen.

$$F_g = T_g \cdot f_{ps}$$

F_g spiegelt in obiger Formel die Gesamtanzahl der zu extrahierenden Bilddaten wider. T_g ist hierbei die Zeit, die mit dem Standardfaktor f multipliziert wird. Bei normalen Videoaufnahmen variiert dieser Faktor zwischen 24 und 30 Bildern pro Sekunde. Hier wurde der Faktor f_{ps} mit dem Wert 25 verwendet.

Da sich die Sprechgeschwindigkeit der Sprecher von Audio- und Videokorpus unterscheiden, muss dieser Transitionsfaktor TF dynamisch errechnet werden.

$$TF = T_{Sij} / C_{FpVS}$$

Der Transitionsfaktor ist der Quotient aus der Zeit T des entsprechenden Audiosignals in Millisekunden und der Anzahl C der aus der Datenbasis des Videokorpus extrahierten 2D-Bilddaten. Die Synchronisation erfolgt linear über die Zeit und das Videosignal wird auf das Audiosignal synchronisiert.

DAS AUDIO-VISUELLE SYNTHESE-SYSTEM AVISS

Die in den vorangehenden Kapiteln beschriebenen Verfahren zur Umsetzung eines audio-visuellen Synthese Systems wurden in der Software AVISS (Audio-Visuelles Synthese System) implementiert. Mit Hilfe der entwickelten Software ist es möglich, eine audio-visuelle Synthese unabhängig von den ursprünglich gesprochen Sätzen zu erzeugen. Hierfür sind lediglich die annotierten Sprach- und Videodaten der Originalsprache notwendig, die zuvor unter bestimmten Einstellungen aufgenommen werden müssen. Die Verfahren erlauben eine schnelle Anpassung an andere Sprachen als das Deutsche und kann mit der AVISS-Software durchgeführt werden. Nachfolgend werden die Module der AVISS-Software vorgestellt und erläutert, sowie die Generierung eines synthetischen audio-visuellen Ausgabevideos beschrieben.

7.1 AVISS-Software-Implementierung

Die Software ist vollständig in C# .Net geschrieben und umfasst mehr als 10000 Zeilen Code. Die Software kann unterteilt werden in 2 Instanzen: Eine Vorverarbeitungsstufe zur Vorbereitung der Sprach- und Videodaten sowie zum Errechnen der statistischen Modelle und in eine Prozessstufe, welche die eigentliche audio-visuelle Synthese durchführt und zur Laufzeit das audio-visuelle Signal erzeugt. Für die einzelnen Arbeitsschritte wurden die Verfahren und Algorithmen jeweils als entsprechende interagierende Module implementiert. In den nachfolgenden Tabellen 12 und 13 sind die einzelnen Modulklassen aufgeführt, die für die erfolgreiche Umsetzung der audio-visuellen Synthese verwendet werden. Zur Unterscheidung werden in Tabelle 12 die Modulklassen angeführt, die zur Vorverarbeitung und Training der Modelle verwendet werden und in Tabelle 13 die Modulklassen, die während der Laufzeit zur Generierung eingesetzt werden. Die Vorbereitung der annotierten Sprachdaten umfasst die Extraktion von kontextabhängigen Merkmalen der Sprachsegmente, die zum Training der statistischen Modelle erforderlich sind.

Modellerstellung	<ul style="list-style-type: none"> ▪ Extraktion der Parameter und Erstellung der kontextbasierten Merkmalsvektoren ▪ Wahrscheinlichkeitsmodelle erstellen ▪ CRF-Modelle trainieren
Visuelle Vorverarbeitung	<ul style="list-style-type: none"> ▪ Video-Audiospur extrahieren ▪ Graphem-Visem Umsetzung ▪ Erstellen der XML basierten Annotation

Tabelle 12: Übersicht der Module zur Vorbereitung der Sprachdaten, Videodaten und statische Modelle

Diese Merkmalsextraktion gestaltet sich nach den in Kapitel 5 angeführten Merkmalen, die zum Aufbau eines kontextabhängigen Merkmalsvektors zur Beschreibung des Sprachsegments dienen. Anhand der Etikettierung der Sprachsegmente mit den kontextbasierten Merkmalsvektoren werden bedingte Wahrscheinlichkeitsmodelle zu den jeweiligen Sprachsegmenten erstellt. Die Modelle folgen den in Kapitel 5 beschriebenen statistischen Modellen. Die Erstellung der kontextabhängigen CRF-Modelle basiert ebenfalls auf diesen Merkmalsvektoren der Sprachsegmente, wobei die Klasse zu den zugehörigen Merkmalsvektoren durch die Segment-ID, einschließlich Start- und Endzeit des zugehörigen Sprachsegments, gebildet wird. Das Training der CRFs wird mittels der Mallet-Bibliothek⁸ durchgeführt, die in die bestehende Software integriert wurde.

Für die Vorverarbeitung der Videosignale wird aus den Videodaten zum einen die Audiospur und zum anderen die Videospur extrahiert und getrennt abgespeichert. Die Videospur enthält die Einzelbilddaten, welche später zur Laufzeit extrahiert und zu einer neuen Äußerung konkateniert werden. Die Audiospur dient zur Bestimmung der Zeitmarken, welche für die Segmentierung der Videodaten benötigt werden. Ein Modul zur automatischen Erstellung einer XML-basierten Annotationsdatei wurde integriert. Diese XML-Annotationsdatei bildet die Basis für die spätere Detektion der Träger-Videodatei, aus der dann die erwünschten Bilddaten extrahiert werden. Hierbei wird die Identifikation der Segmente durch die visemische Transkription des Eingabetextes vollzogen. Insgesamt enthält die XML-Datei die visemische Transkription, Angaben zur Start-, Endzeit in Mil-

⁸ http://mallet.cs.umass.edu/index.php/Main_Page

lisekunden, den Start-, Endframe, die ID zur Identifikation des Videosignals im Korpus, sowie die pixelbasierten Merkmale, anhand derer der KNN-Algorithmus die metrische Distanz zwischen den Bilddaten errechnet. Nachfolgend wird ein Auszug einer solchen XML-Datei beispielhaft dargestellt.

```
<SENTENCE Path=" BBTestsätze" Id="s0005" NFrame="48" Length="1920"
Orth="guten Tag" STKey="g u: t @n |t a: k " SVKey="g u: t @n |t a: k " Type=".">
```

```
<WORD WTKey="g u: t @n " WVKey="TUN" Pos="ADJA" AStart="1068"
AEnd="1358" VStart="27" VEnd="34" Fall="7" HZ="" VK/>
```

```
<WORD WTKey="t a: k " WVKey="TAT" Pos="NN" AStart="1358"
AEnd="1820" VStart="34" VEnd="46" Fall="12" HZ="" VK=""/>
```

```
</SENTENCE>
```

In Tabelle 13 sind die Module der Prozessstufe aufgeführt, die zur Laufzeit verwendet werden, um die eigentliche audio-visuelle Synthese durchzuführen. Hierbei sind nochmals die Sprachsynthese und die visuelle Synthese unabhängig voneinander durchführbar.

Textvorverarbeitung	<ul style="list-style-type: none"> ▪ Textnormalisierung ▪ Graphem-Phonem-Umsetzung ▪ Graphem-Visem-Umsetzung
Audio-Visuelle Synthese	<ul style="list-style-type: none"> ▪ Wortklassen-Vorhersage ▪ Dauer- und F0-Prädiktion ▪ KNN-basierte Segmentidentifikation und Auswahl ▪ Extraktion der Bildabfolgen und Konkatination ▪ Audio-Video-Synchronisation ▪ Generierung von Ausgabevideo und Kompression
Client-Server Kommunikation	<ul style="list-style-type: none"> ▪ Socket-basierte Client-Server-Kommunikation

Tabelle 13: Übersicht der Module, die zur Laufzeit zur Synthese verwendet werden.

Die Sprachsynthese folgt der klassischen Dreiteilung: Textvorverarbeitung, Prosodie und akustische Synthese. Die Textvorverarbeitung löst durch Nachschauen in einem Lexikon vorhandene Abkürzungen im Eingabetext auf. Danach werden numerische Datums- und Zeitangaben in Text umgewandelt. Weiterhin werden Satzzeichen erkannt und entfernt. Im Falle des Satzendezeichens wird dieses als Merkmal gespeichert. Nachdem der Eingabetext normalisiert wurde, kommt das Graphem-Phonem-Transkriptionsmodul zum Einsatz und wandelt den Eingabetext in seine phonemische Entsprechung um. Auf Basis der phonetischen Transkription wird die visemische Transkription des Eingabetextes durchgeführt. Nach Beendigung dieses Vorgangs ist die symbolische Vorverarbeitung abgeschlossen. Die symbolische Vorverarbeitung wird getrennt von den Prosodiemodulen und den Modulen zur akustischen, bzw. visuellen Synthese als eigenständig adaptierbares Modul implementiert, da die symbolische Vorverarbeitung stark sprachenabhängig ist und Anpassungen für eine andere Sprache als Deutsch notwendig sind. Hingegen sind das Prosodiemodul und die akustische und visuelle Synthese-Module in diesem System unabhängig von der Quellsprache (ausgenommen Tonsprachen u. Ä.) und somit auf jede Sprache trainierbar.

7.2 Generierung der audio-visuellen Synthese

In einer zweiten Stufe werden die dynamischen Merkmale, wie Wortklassenbestimmung, Dauer und F0 vorhergesagt. Hier beginnt auch die eigentliche Generierung des audio-visuellen Synthesesignals, die nachfolgend beschrieben ist.

Die phonetische und visemische Umsetzung des Eingabetextes wird benötigt, um die Segmente im Korpus eindeutig zu identifizieren und einen entsprechenden kontextabhängigen Merkmalsvektor für die Segmente zu generieren. Hierfür werden die quantitativen Merkmale, wie Position im Satz oder im Wort extrahiert (siehe Kapitel 5). Für die dynamischen Merkmale werden, wie zuvor erwähnt, die Merkmale anhand statistischer Lernverfahren vorhergesagt und dem Merkmalsvektor hinzugefügt. Wurde nun z.B. für ein Wort ein solcher Merkmalsvektor aufgebaut, beginnt der Auswahlschritt. Es wird je nach angewendetem Verfahren, bedingte Entropie bzw. CRF, das entsprechende Modell für das Zielsegment aufgerufen. Im Idealfall wird ein Bigramm-Modell gefunden. Ist dies

nicht der Fall, wird eine hierarchische „top-down“ Suche angewendet. Zuerst auf Wortebene, dann auf Silbenebene bis hinunter auf die Phonebene. Für jedes Sprachsegment prädiziert das statistische Modell eine ID, welche angibt, in welchem Trägersatz sich das Sprach-, bzw. visuelles Segment befindet. Die ID beinhaltet ebenso die Start- und Endzeit der Segmente, die benötigt wird, um das Segment zu extrahieren. Nachdem die entsprechenden Segmente ausgeschnitten wurden, werden sie neu konkateniert und ergeben so das erwünschte Sprachsignal. Für die visuelle Synthese wird der visemisch transkribierte Eingabetext verwendet. Anhand des transkribierten Textes können die visuellen Segmente im Korpus identifiziert werden. Hierbei wird zuerst auf Wortebene, dann auf Trivisebene nach geeigneten visuellen Segmenten gesucht. Sind die potentiellen Segmente identifiziert, kommt der KNN-Algorithmus zum Einsatz und wählt dasjenige Segment aus, welches den geringsten Abstand zum letzten Bild im visuellen Ausgangssegment hat. Wurden die geeigneten Bildsequenzen ausgewählt und konkateniert, wird noch die Synchronisation mit dem Sprachsignal vollzogen und die beiden Quellen werden zu einem audio-visuellen Synthesesignal vereint.

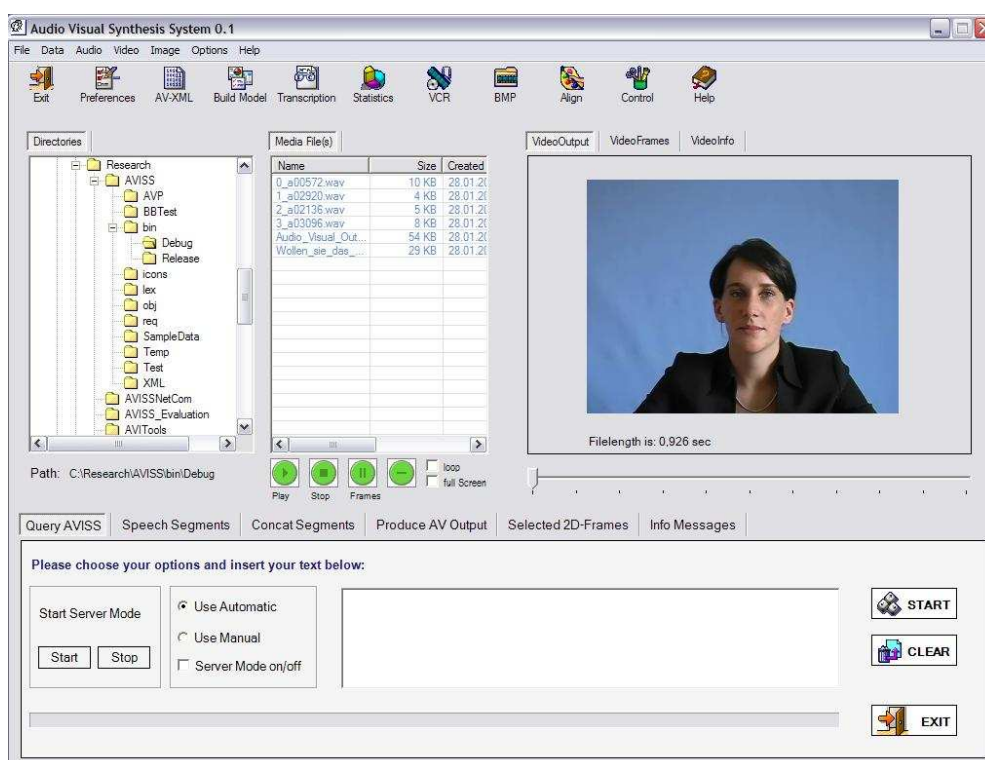


Abbildung 37: Benutzeroberfläche der AVISS Software

Das audio-visuelle Signal wird komprimiert und als Video ausgegeben. Hierbei kann dies über eine Client-Server Architektur generiert werden oder unmittelbar auf dem Client-PC. Abbildung 37 zeigt die implementierte Benutzerschnittstelle, mittels derer die audio-visuelle Synthese durchgeführt werden kann.

EVALUATION

In Kapitel 5 wurden statistisch motivierte Verfahren entwickelt und eingesetzt, die Eingabetexte in ein synthetisiertes Sprachsignal und ein synthetisiertes audio-visuelles Videosignal umwandeln. Die Verfahren wurden als akustisches und visuelles Synthese-Modul in die in Kapitel 7 vorgestellte Software AVISS integriert.

In diesem Kapitel werden die Ergebnisse der Evaluation beschrieben. Die Evaluation wurde durch eine Onlineabstimmung durchgeführt, bei der die Probanden eine Auswahl an Sprachsignalen bzw. audio-visuellen Signalen bewertet haben. Diese mit der Software synthetisierten Sprachsignale und audio-visuell synthetisierten Videodateien, wurden mit einem MOS (Mean Opinion Score) bewertet. Nachfolgend werden in Abschnitt 8.1 die Ergebnisse der Evaluation des Sprachsynthesemoduls sowie in Abschnitt 8.2 die Ergebnisse der audio-visuellen Synthese angeführt. Die zur Evaluation verwendeten Sätze sind im Anhang A angegeben.

8.1 Evaluationen der Sprachsynthese-Ausgabe

Das statistische motivierte Synthesemodul folgt einem kaskadierten Modell, in dem unterschiedliche Ebenen von Bausteingrößen verwendet werden, um synthetische Sprache zu erzeugen. Um eine Bewertung der Verfahren zu erzielen, wurde ein kombinierter Perzeptionstest verwendet. Zum einen wird die Präferenz erfragt, und zum anderen wurden auf Basis des MOS-Testes (Mean Opinion Score) verschiedene Kategorien zur Bewertung der synthetisierten Sprachsignale verwendet. Dieser Test hat eine Skala von 1 bis 5, wobei 1 als schlechteste und 5 als beste Bewertung gilt. Es wurden synthetisierte Sätze mittels Markov-Entropie (ME) und CRF generiert. Diese wurden jeweils zusammen mit Sprachsignalen präsentiert, die mit dem auf 2-dimensionalen Kostenfunktionen basierten Ansatz synthetisiert wurden. Den Teilnehmern der Evaluation wurden die Sätze über eine

WWW-Seite zugänglich gemacht. Die Bewertung der Stimuli mussten die Teilnehmer ebenfalls auf der WWW-Seite eintragen.

Die Kategorien, welche bewertet werden sollten, waren:

- Gesamteindruck
- Natürlichkeit
- Intonation
- Pausenverteilung und Phrasengrenzen
- Qualität

Den Probanden wurde jede Kategorie erklärt. Unter Gesamteindruck sollte der „erste Eindruck“ des synthetisierten Satzes bewertet werden. Unter Natürlichkeit, wie natürlich, angelehnt an die menschliche Sprache, der synthetisierte Satz klingt. Die Intonation beinhaltet den F0-Verlauf der Äußerung wie auch die Akzentverteilung. Die Kategorie Pausen, Pausenverteilung und Phrasengrenzen soll abfragen, wie gut die einzelnen Intonationsphrasen realisiert wurden. Unter Qualität sollte bewertet werden, wie gut das Sprachsignal hinsichtlich Störungen ist. Störungen können Konkatenationsfehler sein oder Störungen im Sprachsignal selbst. Es wurden insgesamt 75 Hörer berücksichtigt, wobei die Einträge mit durchgängig minimaler oder maximaler Bewertung aus der Gesamthörerzahl herausgenommen wurden. Durchgängig minimale Bewertung wurde festgestellt, wenn nur Werte von 1 eingetragen wurden und durchgängig maximale Bewertung wurde festgestellt, wenn nur Werte von 5 eingetragen wurden.

Die Hörer bilden eine heterogene Gruppe, wobei einige aus dem Bereich Sprachverarbeitung kommen, andere sich mit diesem Thema aktuell im Studium damit befassen. Eine weitere Gruppe Hörer hatte bisher keinen Bezug zu Sprachverarbeitung. Abbildung 38 zeigt die Ergebnisse der Bewertung der Sprachsignale, die mit dem Markov-Entropie basierten Modell und dem konventionellen 2-dimensionalen Kostenfunktionsansatz synthetisiert wurden.

Es wurde jeweils der Mittelwert gebildet und auf der Y-Achse im Diagramm für jede Kategorie eingetragen. Wie in dem Ergebnisdiagramm zu erkennen ist, wurde der Gesamteindruck bei dem ME-basierten Ansatz geringfügig besser als der Gesamteindruck des konventionellen Ansatzes bewertet.

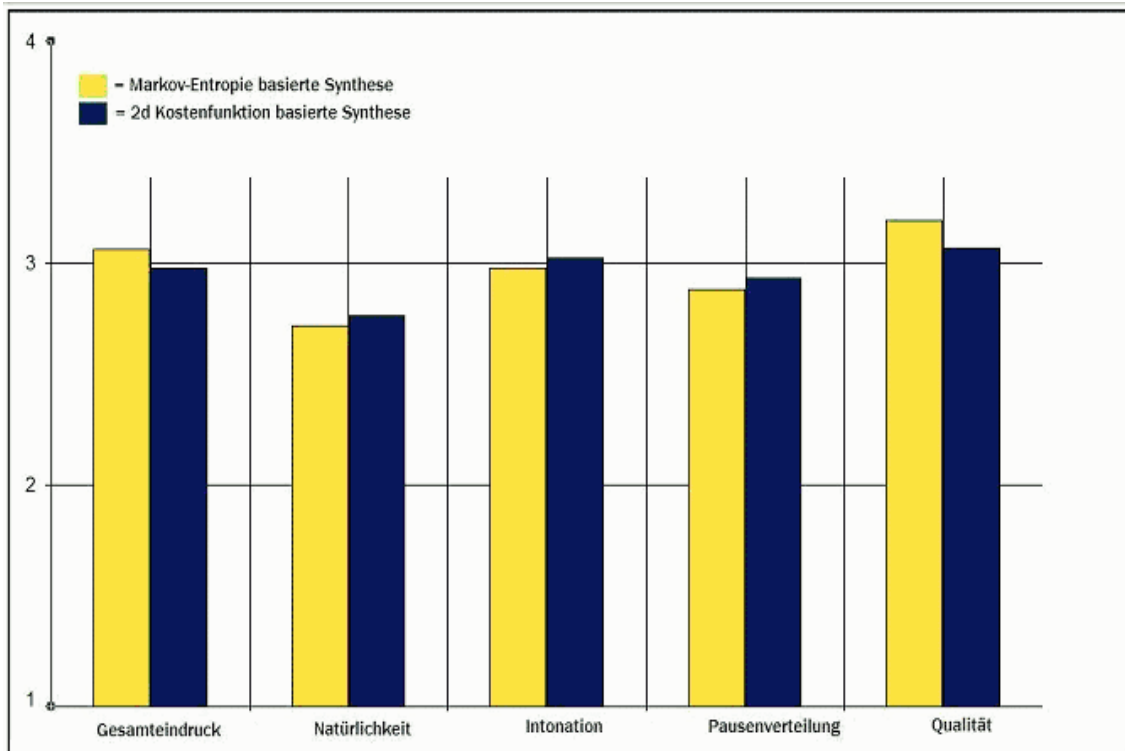


Abbildung 38: Diagramm der Evaluationsergebnisse: Markov-Entropie basierte Synthese

In den Kategorien Natürlichkeit, Intonation und Pausenverteilung wurden die Hörbeispiele der kostenfunktionsbasierten Synthese im Schnitt 0.15 besser bewertet. Bei der Qualität wurden die synthetisierten Sätze, welche mit dem ME-Ansatz synthetisiert wurden, als besser bewertet.

Da Gesamteindruck und Qualität besser bewertet wurden als Natürlichkeit, Intonation und Pausenverteilung, kann nicht davon ausgegangen werden, dass das ME-basierte Verfahren eine bessere Ausgabe von synthetisierter Sprache liefert als das konventionelle kostenfunktionsbasierte Verfahren.

Als nächstes wurde eine Evaluation durchgeführt, die den CRF-basierten Ansatz mit dem konventionellen Ansatz vergleicht und Bewertungen für die einzelnen Kategorien abgegeben werden mussten. Es wurden jeweils mit dem gleichen Ausgangstext Sätze synthetisiert und den Probanden diese, über eine WWW-Seite zugänglich gemacht. Es wurden die gleichen Kategorien verwendet, wie bei der Markov-Entropie-basierten Evaluation.

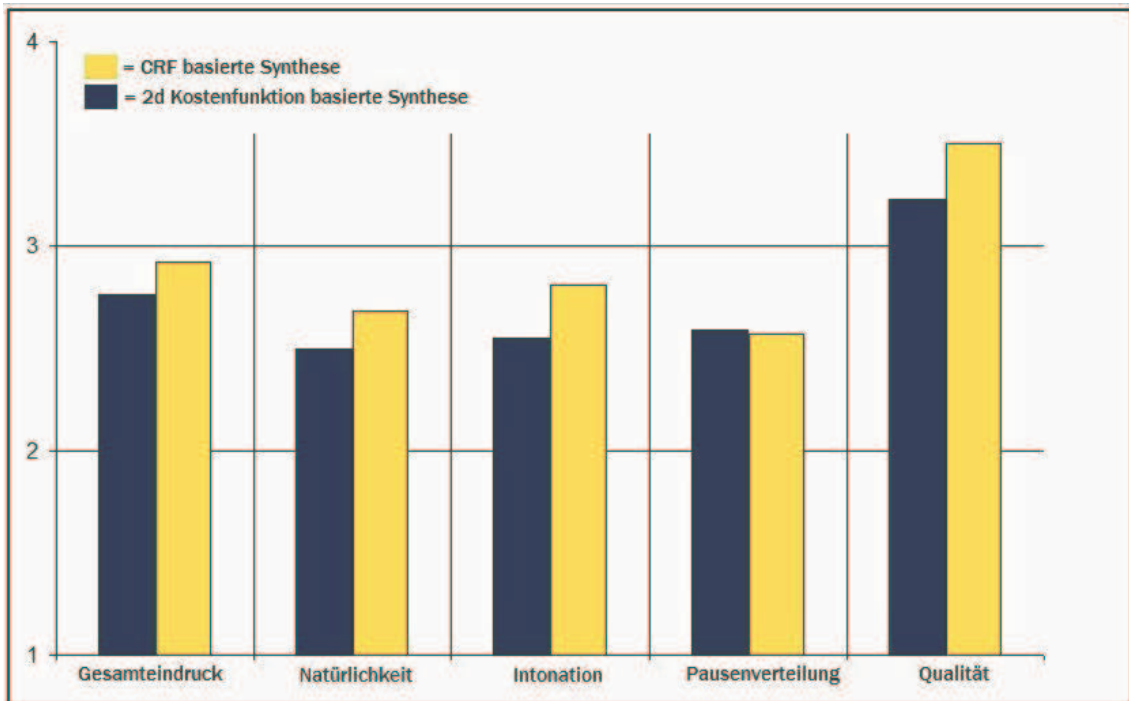


Abbildung 39: Diagramm der Evaluationsergebnisse: CRF basierte Synthese

Abbildung 39 stellt die Ergebnisse in einer Diagrammübersicht dar. Der CRF-basierte Syntheseansatz erzielt in 4 von 5 Kategorien bessere Ergebnisse als der konventionelle kostenbasierte Ansatz. In der Kategorie Pausenverteilung wurde der konventionelle Ansatz besser bewertet.

Die Ergebnisse zeigen der Bewertung zeigen, dass Segmentauswahl mittels der kontextbasierten CRF-Modelle besser Bewertet wurde, als der konventionelle Ansatz. Dies könnte auf die statistisch motivierte Gewichtung der Merkmale zurückgeführt werden, die bei Segmenten mit hoher Häufigkeit zu einer besseren Auswahl des potentiell geeignetsten Segments führt. In der Forschung besteht jedoch bisher keine Einigkeit darüber wie die einzelnen Merkmale zu gewichten sind. Hier bedarf es weitergehende Untersuchungen, die im Rahmen dieser Arbeit jedoch nicht durchgeführt werden konnten.

8.2 Evaluationen der audio-visuellen Synthese-Ausgabe

Im Gegensatz zur Sprachsynthese gibt es bei der audio-visuellen Synthese keine allgemein anerkannten Verfahren, wie die audio-visuellen Stimuli evaluiert werden und vor allem was evaluiert werden soll. Daher wurde zur Überprüfung der Ergebnisse der audio-

visuellen Syntheseausgabe die Evaluation an das Evaluationsverfahren der Sprachsynthese angelehnt. Ein 5-wertiger MOS-Test wurde hierfür verwendet, mit dem die audio-visuellen Signale bewertet werden. Die Bewertungsskala reicht von 1, für schlecht, bis 5, für sehr gut. Die Evaluation wurde jeweils für unterschiedliche Kategorien vorgenommen. Folgende Bewertungskategorien wurden für die Evaluation verwendet:

- Gesamtqualität: evaluiert die Signalqualität hinsichtlich Störungen
- Natürlichkeit: evaluiert die Natürlichkeit, in wie weit das Videosignal von einem original Video unterscheidbar ist
- Synchronität: evaluiert die Synchronität von Audiosignal und Lippenbewegung
- Verständlichkeit: evaluiert die Verständlichkeit des audio-visuellen Signals
- Akzeptanz: evaluiert die Akzeptanz beim Nutzer

Das Verfahren der Evaluation wurde folgendermaßen durchgeführt. Es wurden audio-visuelle Stimuli anhand von Textsequenzen, wie sie im zugrundeliegenden Korpus vorkommen, jedoch nicht in dieser Abfolge, synthetisiert. Diese wurden mittels einer WWW-Benutzerschnittstelle den Teilnehmern als Videodatei zugänglich gemacht.

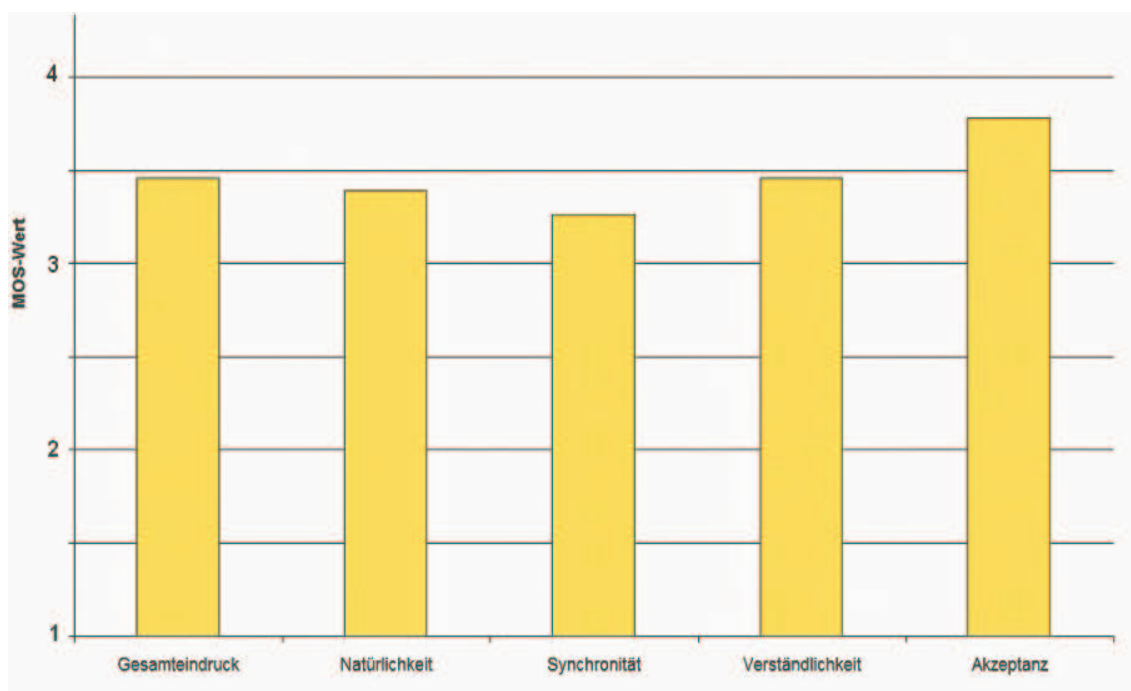


Abbildung 40: Diagramm der Evaluationsergebnisse audio-visuelles Synthesesignale

Die Teilnehmer mussten sich die Videodatei anschauen und die Bewertung in ein WWW-Formular eintragen. Für jede Videodatei stand ein eigenes Bewertungsformular bereit. Bei

der Evaluation wurden die Bewertungen von 76 Teilnehmern ausgewertet. Es wurden die Einträge von Teilnehmern nicht berücksichtigt, die durchgängig eine minimale bzw. maximale Bewertung abgaben. Abbildung 40 zeigt die Ergebnisse der Evaluation. Es ist zu sehen, dass die Teilnehmer das audio-visuell synthetisierte Videosignal als akzeptabel einstufen. Diese Kategorie hat die höchste Bewertung erhalten. Dies ist vielversprechend, da Nutzerakzeptanz eine entscheidende Rolle spielt, wenn audio-visuelle Synthese als Mensch-Maschine-Schnittstelle eingesetzt werden soll. Gesamteindruck und Verständlichkeit erhielten die gleiche Bewertung. Auch diese Kategorien sind entscheidend für den Einsatz dieser Technologie. Natürlichkeit erhielt die zweitschlechteste Bewertung. Diese Bewertung gibt Aufschluss darüber, dass das generierte Signal noch Störungen z.B. an den Konkatenationsstellen aufweist und daher als unnatürlich betrachtet wird. Diese Störungen können mit einer nachträglichen Signalmanipulation, z.B. mit einer Glättung der Segmentübergänge mittels Morphing, verbessert werden. Die Bewertung der Natürlichkeit hängt auch mit der schlechten Bewertung der Synchronität zusammen. Dies ist darauf zurückzuführen, dass die Synchronität linear über die Zeit des Audiosignals berechnet wurde. Dennoch haben die audio-visuell synthetisierten Stimuli eine hohe Akzeptanz und eine gute Gesamtqualität, mit einer guten Verständlichkeit, was zu einer positiven Einschätzung der Evaluationsergebnisse führt. Abschließend kann man festhalten, dass vor allem an der Natürlichkeit, der Verständlichkeit und der Synchronität Verbesserungen erarbeitet werden müssen, um hier bessere Ergebnisse zu erzielen und den Gesamteindruck zu steigern, welches dann zu einer noch größeren Akzeptanz beim Nutzer führen kann.

SCHLUSSBETRACHTUNG UND AUSBLICK

In dieser Arbeit wurden Algorithmen und Verfahren entwickelt und angewendet, die es ermöglichen eine video-realistische audio-visuelle Synthese durchzuführen. Das generierte audio-visuelle Signal zeigt einen Talking-Head, der aus zuvor aufgenommenen Videodaten und einem zugrunde liegenden TTS-System konstruiert wurde. Es wurden statistisch motivierte Trainingsverfahren für Unit-Selection-basierte TTS-Systeme entwickelt, die eine schnelle Adaption an unterschiedliche Sprachen zulassen. Zusätzlich wurde ein Nächster-Nachbar-Algorithmus für die visuelle Synthese eingesetzt, der ebenfalls unabhängig von der Quellsprache ist. Das Training des Systems benötigt ein annotiertes Sprachdatenkorpus, sowie ein annotiertes Videodatenkorpus. Alle Schritte bis zur letzten Signalausgabe erfolgen automatisch und bedürfen keinerlei manuellen Eingriffs.

Die Sprachsynthese folgt dem klassischen 3-stufigen Aufbau, Textvorverarbeitung, Prosodie und akustische Synthese. Es wurden die Textvorverarbeitung, die Prosodiegenerierung und die akustische Synthese jeweils unter dem Gesichtspunkt der statistischen Modellierung betrachtet und statistisch motivierte Lernverfahren für die einzelnen Aufgabenstellungen verwendet. Es zeigte sich, dass anhand der eingesetzten Verfahren ein vollständiges Unit-Selection-basiertes Sprachsynthese-System erstellt werden kann. Voraussetzung hierfür ist ein entsprechendes Sprachdatenkorpus mit zugehöriger Annotation. Als Erweiterung der Sprachsynthese wurde eine visuelle Synthese entwickelt. Die visuelle Synthese wurde mit der Sprachsynthese zusammengeschaltet und ein audio-visuelles Synthese-System erstellt.

Bei der Textvorverarbeitung wurde für die Graphem-Phonem-Umsetzung ein Maximum-Entropie basiertes Modell trainiert. Es hat sich gezeigt, dass das eingesetzte Verfahren, trotz geringer Trainingsdaten, vernünftige Ergebnisse liefert. Zur Verbesserung der automatischen Graphem-Phonem-Umsetzung wird ein Trainingsdatenkorpus benötigt, das alle

möglichen Kontexte in mehrfacher Auflistung enthält, so dass jede Klasse ausreichend Trainingsbeispiele zur Verfügung hat. Es kann davon ausgegangen werden, dass ein Trainingsdatenkorpus mit mehr als 100000 Ereignissen ein zufriedenstellendes Trainingskorpus für die statistische Graphem-Phonem Zuordnung ist. Ein Ereignis spiegelt hier einen Merkmalsvektor wider, der einer Phonemklasse zugeordnet wird.

Zur Erinnerung sei nochmals auf Abschnitt 4.4.1 verwiesen, in dem das Trainingsdatenkorpus dieser Arbeit beschrieben ist. Ebenso wie bei der Graphem-Phonem-Umsetzung verhält es sich mit der Vorhersage von Akzent und Silbengrenzen. Hier ist ebenso die Trainingsdatenmenge entscheidend. Aufgrund der vorhandenen Ressourcen konnten keine besseren Ergebnisse erzielt werden. Anders bei der Vorhersage der Wortklassen. Hier konnte auf eine ausreichende Trainingsdatenmenge zurückgegriffen werden. Dies spiegelt sich in den Ergebnissen der Klassifikationsleistung wider. Bei der Vorhersage der prosodischen Parameter, Dauer und F0, wurde ein Entscheidungsbaum-basiertes Lernen verwendet. Die Dauer-Werte und F0-Werte wurden hierfür logarithmiert, was sich als sinnvoll erwiesen hat, da durch das Logarithmieren eine Minimierung der Klassen erreicht wird. Dies hat zum Vorteil, dass mehr Trainingsbeispiele für eine Klasse zu Verfügung stehen. Hier wurden vernünftige Ergebnisse erzielt.

Für die Generierung des akustischen Signals wurden 2 unterschiedliche Ansätze verfolgt. Zum einen der HMM-basierte Ansatz, der das Sprachsignal aus Sprachparametern approximiert, indem das MLSA-Filter verwendet wird, und zum anderen der statistisch motivierte Ansatz, bei dem akustische Einheiten unterschiedlicher Größe zur Sprachsignalgenerierung konkateniert werden. Bei der HMM-basierten Sprachsynthese hat sich gezeigt, dass mit einem Sprachdatenkorpus von 20 Minuten Sprache eine akzeptable Sprachsynthese erstellt werden kann. Das Sprachsignal selbst klingt durch das Filter etwas metallisch. Neuere Forschungsarbeiten befassen sich mit Filtern, die eine gemischte Anregung als Eingabeparameter verwenden. Dies kann unter Umständen zu einer Verbesserung der Sprachsignalqualität führen. Der HMM-basierte Sprachsynthese-Ansatz ist leicht an neue Sprachen anzupassen und benötigt nur ein zugrunde liegendes Sprachdatenkorpus mit entsprechender Annotation. Das Training der kontextabhängigen HMMs hängt von der Größe des Sprachdatenkorpus (20 Minuten bis 3 Stunden Sprache) ab, dauert aber im Allgemeinen auf handelsüblichen PCs nicht länger als 24 Stunden.

Die Sprachsignalgenerierung, basierend auf den entwickelten graphischen Modellen, bedingte Entropie und CRF, konkatenieren die Sprachsegmente, wie sie von den statistischen Lernverfahren prädiziert werden und generieren das Sprachsignal ohne zusätzliche Signalmanipulation. Wie die Evaluationen in Kapitel 8 gezeigt haben, wurde mit beiden Ansätzen natürlich klingende und verständliche Sprache erzeugt. Bei der Sprachsignalgenerierung, die durch Segmentauswahl anhand der bedingten Entropie und anschließender Konkatenation durchgeführt wurde, wurde analog zu Sprachsignalen, die mit dem konventionellen auf 2-dimensionalen Kosten basierten Ansatz generiert wurden, vergleichbare Ergebnisse erzielt. Sprachsignale die mit dem CRF-basierten Ansatz erzeugt wurden, haben, im Vergleich zu den konventionell generierten Sprachsignalen, in der Evaluation besser abgeschnitten. Dies kann darauf zurückgeführt werden, dass bei dem konventionellen Ansatz die Gewichtung der Kosten nicht den entsprechenden Faktoren Rechnung tragen, die zu der besten Segmentauswahl führen. Bei dem CRF-basierten Ansatz werden die Gewichte entsprechend der Trainingsdaten errechnet und ermöglichen so eine bessere Auswahl je nach Kontext. Beide Ansätze lassen sich leicht an neue Sprachen anpassen. Hierzu bedarf es einer sprachabhängigen Textvorverarbeitung und eines sprachabhängigen Prosodiemoduls. Dies ermöglicht eine schnelle und flexible Erstellung eines Sprachsynthese-Systems für neue Sprachen. Das Erstellen der Modelle, für die Sprachsegmentidentifikation mittels bedingter Entropie und anschließender konkatenationsbasierte Sprachsignalgenerierung, hat eine Laufzeit von ca. 6 Stunden. Im Gegensatz hierzu dauerte das Training der kontextabhängigen CRF-Modelle, auf einem handelsüblichen PC, 96 Stunden. Bei den entwickelten Verfahren, Bedingte-Entropie-basierte und CRF-basierte Sprachsignalgenerierung, liegt die Antwortzeit, bis das synthetisierte Signal ausgegeben wird, über der Antwortzeit des konventionellen auf 2-dimensionalen Kosten basierten Systems. Die Antwortzeit beträgt zwischen einer Sekunde mehrerer Sekunden bei langen Sätzen. Hier muss eine effektive Implementierung der Suchstrategie erfolgen. Weiterhin kann eine nachträgliche Signalmanipulation die Sprachsignalqualität erhöhen. Beides liegt außerhalb der Themenstellung dieser Arbeit und wurde daher nicht berücksichtigt.

Die Erstellung der visuellen Synthese folgt dem Ansatz der Unit-Selection-basierten Sprachsynthese und extrahiert aus einem bestehenden Korpus, entsprechend den Zielvor-

gaben, Videosegmente, die anschließend zu einem neuen Videosignal konkateniert werden. Bei diesem Ansatz ist eine sorgfältige Korpuserstellung notwendig. Der Sprecher muss angewiesen werden, ständige Kopfbewegungen zu vermeiden. Ausgangsüberlegung ist ein Nachrichtensprecher, der frontal in eine Kamera schaut. Das Verfahren liefert gute Ergebnisse abhängig vom zugrunde liegenden Videokorpus. Die Vorverarbeitung des Videokorpus wird automatisch vollzogen. Manuelle Korrektur muss bei der Segmentierung erfolgen. Der KNN-Algorithmus hat sich als robust erwiesen und wählt das geeignetste Segment zur Konkatenation anhand einer Distanzmetrik aus. Wie die Evaluation der audio-visuell synthetisierten Videostimuli gezeigt hat, besteht eine hohe Akzeptanz, wie auch eine gute Bewertung der Gesamtqualität und Verständlichkeit. Die schlechteren Bewertungen von Natürlichkeit und Synchronität lassen den Schluss zu, dass die Natürlichkeit mit der Synchronität korreliert. Die visuelle Synthese wird linear zum Sprachsignal synchronisiert. Eine Verbesserung wäre hier der Einsatz einer dynamischen, der Segmentgröße entsprechenden Synchronisation von Audio- und Videospur. Ebenso wie bei der Sprachsynthese kann eine nachträgliche Signalmanipulation die Segmentübergänge an den Konkatenationsstellen glätten und somit eine bessere Qualität des audio-visuellen Synthesesignals erbringen.

In dieser Arbeit wurden Verfahren entwickelt und eingesetzt, die eine Grundlage für audio-visuelle Sprachsynthese bilden. Hauptaugenmerk lag auf den statistisch motivierten Lernverfahren für eine sprachenunabhängige Modellierung audio-visueller Synthese. Es hat sich gezeigt, dass die neu entwickelten und eingesetzten Verfahren, vor allem der CRF-basierte Syntheseansatz, eine Verbesserung der Sprachsynthese erbracht hat. Bei der visuellen Synthese wurde ein neues Verfahren eingeführt, das video-realistische audio-visuelle Synthesesignale erzeugen kann. Weitere Arbeiten sind nötig, um die Qualität zu verbessern und um eine 3-dimensionale Ansicht der audio-visuellen Sprachsynthese-Ausgabe zu erhalten. Arbeiten hierzu, wie auch Forschung zur visuellen Prosodie kann dazu beitragen, die Akzeptanz, Qualität und das Verständnis der Relation Sprache und Mimik zu erhöhen.

LITERATURVERZEICHNIS

- Allen, J., Hunnicutt, S., Klatt, D.: From text to speech: the MITalk system. MIT Press, Cambridge, MA, 1987.
- Andre E, Rist E, Muller J.: Guiding the user through dynamically generated hypermedia presentations with a life-like character. Intelligent User Interfaces. San Francisco, CA, 1998.
- Arb, A., Gustafson, S., Anderson, T., Slyh, R.: Hidden markov models for visual speech synthesis with limited data. In Proceedings of the Auditory-Visual Speech Processing Workshop, Aalborg, Denmark, September 2001.
- Bailly, G., Béjar, M., Elisei, F., Odisio, M.: Audiovisual speech synthesis. International Journal of Speech Technology, 6, October 2003.
- Berger, A., Della Pietra, S., Della Pietra, V.: A maximum entropy approach to natural language processing. Computational Linguistics, 22(1), 1996.
- Beskow, J.: Animation of talking agents. In: Proceedings of the Auditory-Visual Speech Processing Workshop, Rhodes, Greece, September 1997.
- Beskow, J.: Talking Heads. Models and applications for multimodal speech synthesis. PhD thesis, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, 2003.
- Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A.: The AT&T Next-Gen TTS System, Joint meeting of ASA, EAA, and DAGA, Berlin, March 1999.
- Beutnagel, M., Mohri, M., Riley, M.: Rapid unit selection from a large speech corpus for concatenative speech synthesis. In: Proceedings of the European Conference on Speech Communication and Technology, EuroSpeech, Budapest, Ungarn, 1999.
- Beymer, D., Poggio, T.: Image Representation for Visual Learning, Science, vol. 272, pp.1905-1909, 28 June 1996.
- Black, A., Taylor, P.: CHATR: a generic speech synthesis system. In: Proceedings of the International Conference on Computational Linguistics, Band 2, Kyoto, Japan, 1994.

- Black, A., Campbell, N.: Optimizing selection of units from speech databases for concatenative synthesis. In: Eurospeech, Vol 1, Madrid, Spain 1995.
- Black, A., Taylor, P.: Automatically Clustering Similar Units for Unit Selection in Speech Synthesis, Proc. Eurospeech, Rhodes 1997.
- Black, A., Lenzo, K.: Limited Domain Synthesis. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, China, 2000.
- Blanz, V., Basso, C., Poggio, T., Vetter, T.: Reanimating faces in images and video. In: Proceedings of the Annual Conference of the European Association for Computer Graphics, Granada, Spain, 2003
- Breen, A. P., Jackson, P.: Non-uniform unit selection and the similarity metric within BT's Laureate TTS system. In: Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australien, 1998.
- Bregler, C., Covell, M., Slaney, M.: Video Rewrite: Driving Visual Speech with Audio. In: Proc. SIGGRAPH, ACM SIGGRAPH, July 1997.
- Breiman, L., Friedman, J., Olshen, R., Stone C.: Classification and regression trees, Chapman Hall, 1984.
- Brooke, N. , Scott, S.: Two and three-dimensional audiovisual speech synthesis. In Proceedings AVSP, Sydney, Australia, 1998.
- Cohen, M.M., Massaro, D.W.: Modeling Coarticulation in Synthetic Visual Speech, Models and Techniques in Computer Animation, Springer Verlag, 1993.
- Cohen, M.M., Walker, R.L., Massaro, D.W.: Perception of synthetic visual speech. In: Speech reading by humans and Machines, D.G. Stroke and M.E. Hennecke (Eds.), New York: Springer 1996.
- Cohen, M.M., Massaro, D. W., Clark, R.: Training a talking head. In: Proceedings of the IEEE International Conference on Multimodal Interfaces, Pittsburgh, USA, October 2002.
- Conkie, A., Beutnagel, M., Sydral, A., Brown, P.: Preselection of candidate units in a unit selectionbased text-to-speech synthesis system. In: Proceedings of the Inter-

- national Conference on Spoken Language Processing (ICSLP), Beijing, China, 2000.
- Cosatto, E., Graf, H.-P.: Photo-Realistic Talking-Heads from Image Samples. In: IEEE Transactions on Multimedia, Vol. 2, No. 3, 2000.
- Cosatto, E., Graf, H.P.: Sample-Based Synthesis of Photo-Realistic Talking-Heads, IEEE Computer Animation, 1998.
- Cosker, D., Marshall, D., Rosin, P., Hicks, Y. A.: Video realistic talking heads using hierarchical non-linear speech-appearance models. In: Proceedings of MIRAGE, INRIA Rocquencourt, France, 2003.
- Cosker, D., Marshall, D., Rosin, P., Paddock, S., Rushton, S.: Towards perceptually realistic talking heads: models, methods and McGurk. In: Symposium on Applied Perception in Graphics and Visualization, Los Angeles, USA, August 2004.
- Demberg, V.: Letter-to-Phoneme Conversion for a German Text-to-speech System. Diplomarbeit. IMS, Stuttgart, 2006.
- Donovan, R.E.: A new distance measure for costing spectral discontinuities in concatenative speech synthesizers, The 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.
- Donovan, R.E.: Segment pre-selection in decision-tree based speech synthesis systems. Proc. ICASSP, Istanbul, Turkey, 2000.
- Donovan, R.E., Woodland, P.: A hidden Markov-Model-based trainable speech synthesizer. Computer, Speech and Language, 13, 1999.
- Donovan, R.E., Eide, E.M.: The IBM Trainable Speech Synthesis System, Proc. ICSLP, Sydney 1998.
- Donovan, R.E.: Trainable Speech Synthesis, PhD. Thesis, Cambridge University Engineering Department, 1996.
- Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, New York, 2nd edition, 2001.
- Dunteman, G., H.: Principal Component Analysis, Sage Publications, 1989.
- Dutoit T.: An Introduction to Text-to-Speech Synthesis. Kluwer, Dordrecht 1997.

- Ekman, P., Friesen, W. V. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues* New Jersey: Prentice Hall, 1975.
- Ekman, P., Friesen, W.V., Hager, J.C.: *The Facial Action Coding System*. Second edition, Salt Lake City, 2002.
- Ekman, P.: *Gefühle lesen - Wie Sie Emotionen erkennen und richtig interpretieren*, Spektrum Akademischer Verlag, München 2004.
- Eichner, M., Wolff, M., Ohnewald, S., Hoffmann, R.: *Speech synthesis using stochastic Markov graphs*. Proc. ICASSP, Salt Lake City 2001.
- Eichner, M., Wolff, M., Hoffmann, R.: *A unified approach for speech synthesis and speech recognition using Stochastic Markov Graphs*", Proc. ICSLP, Beijing, vol. 1, 2000.
- Eisert P., Chaudhuri S., Girod B.: *Speech driven synthesis of talking head sequences, 3D Image Analysis and Synthesis*. Erlangen, 1997.
- Elisei, F., Odisio, M., Bailly, G., Badin, P.: *Creating and controlling video-realistic talking heads*. In: *Proceedings of the Auditory-Visual Speech Processing Workshop Aalborg, Denmark, September 2001*.
- Engl-Müller, G., Schäfer, W., Trippler, G.: *Kompaktkurs Ingenieurmathematik*, Hanser, Leipzig 2001
- Ezzat, T., Geiger, G., Poggio, T.: *Trainable Videorealistic Speech Animation*. In: *Proceedings of ACM SIGGRAPH, San Antonio, Texas, 2002*
- Ezzat, T., Poggio, T.: *MikeTalk: Facial Display on Morphing Visemes*. In: *Proceedings of the Computer Animation Conference Philadelphia, PA, June 1998*.
- Ezzat, T., Poggio, T.: *Videorealistic Talking Faces: A Morphing Approach*. In: *Proceedings of the Audiovisual Speech Processing Workshop, Rhodes, Greece, September 1997*.
- Fant, G.: *Acoustic theory of speech production*. The Hague, Netherlands: Mouton, 2nd edition. 1960,
- Günther, C.: *Prosodie und Sprachproduktion*. Niemeyer, Tübingen 1999.

- Hamon, C., Moulines, E., Charpentier, F. : A diphone synthesis system based on time-domain modifications of speech. In Proceedings of ICASSP, New York, USA, 1989.
- Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, inference and Prediction. Springer series in Statistics, New York, Berlin, Heidelberg, 2001.
- Hess, W.: Grundlagen der Phonetik. Kapitel 1-4, http://www.ikp.uni-bonn.de/dt/lehre/materialien/grundl_phon/index.html, 2005.
- Hirai, T., Iwahashi, N., Higuchi, N., Sagisaka, Y.: Automatic extraction of F0 control rules using statistical analysis. Progress in Speech Synthesis, J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, N.Y., Springer, 1997.
- Hirai, T., Tenpaku, S., Shikano, K.: Speech unit selection based on target values driven by speech data in concatenative speech synthesis. Proc. IEEE 2002 Workshop on Speech Synthesis, Santa Monica, U.S.A., Sep. 2002.
- Huang, X., Acero, A., Adcock, J., Hon, H-W., Goldsmith, J., Liu, J., Plumpe, M.: Whistler: A Trainable Text-to-Speech System, Proc. ICSLP, Philadelphia 1996.
- Huang, X., Acero, A., Hon, H-W.: Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall 2001.
- Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of ICASSP, Vol. 1, Atlanta, Georgia, 1996.
- Imai, S.: Cepstral analysis synthesis on the mel frequency scale. In: Acoustics, Speech, and Signal Processing, IEEE International Conference, ICASSP, 1983.
- Jähne, B., Digital image processing. Berlin: Springer-Verlag, 1993.
- Jelinek, F.: Statistical Methods for Speech Recognition. Language, Speech, and Communication. MIT Press, London, 1997.
- Kawai, H., Yamamoto, S., Higuchi, N., Shimizu, T.: A design method of speech corpus for text-to-speech synthesis taking account of prosody. Proc. ICSLP, Vol. 3, Beijing, China, 2000.

- Keller, E., Bailly, G., Monaghan, A., Terken, J., Huckvale, M.: Improvements in Speech Synthesis Wiley & Sons, Chichester, UK, 2001
- Keller, E.: Fundamentals of Speech Synthesis and Speech Recognition. John Wiley & Sons, 1994.
- Klabbers, E., Veldhuis, R.: Reducing audible spectral discontinuities, IEEE Transactions on Speech and Audio Processing, vol. 9, January 2001.
- Klabbers, E., Veldhuis, R.: On the Reduction of Concatenation Artefacts in Diphone Synthesis, Proc. ICSLP, Sydney 1998.
- Kraft, V.: Verkettung natürlichsprachlicher Bausteine zur Sprachsynthese: Anforderungen, Techniken und Evaluierung. VDI Fortschrittsberichte Informatik / Kommunikationstechnik 10, Nr. 468, Düsseldorf 1997.
- Lafferty, J., McCallum, A. and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. ICML, 2001.
- Lee, Y., Terzopoulos, D., Waters, K.: Realistic modeling for facial animation. In: Proc. SIGGRAPH, ACM Press/ACM SIGGRAPH, Los Angeles, 1995.
- Le Goff, B., Benoit, C.: A text-to-audiovisual-speech synthesizer for French. In: Proceeding of the International Conference on Spoken Language Processing, ICSLP, 1996.
- Magnenat-Thalmann, N., Primeau, E. and Thalmann, D.: Abstract muscle action procedures for human face animation. The Visual Computer 3(5), 1988.
- Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press 1999.
- Massaro, D.,W.: Perciving talking faces: From speech perception to a behavioral principle. Cambridge, MA: The MIT Press 1998.
- Masuko, T., Kobayashi, T., Tamura, M., Masubuchi, J., Tokuda, K.: Text-to-visual speech synthesis based on parameter generation from HMM, ICASSP, 1998.
- McCallum, A., K.: MALLET: A Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu>, 2002.
- McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature 264, 1976.

- Möbius, B.: German and Multilingual Speech Synthesis. Habilitationsschrift, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 7 2001.
- Möbius, B.: Corpus-Based Speech Synthesis: Methods and Challenges. Band 6(4) der Reihe Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Universität Stuttgart, 2000.
- Möbius, B.: Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis. *International Journal of Speech Technology*, 2003.
- Moulines, E., Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 1990.
- Narayanan, S., Alwan, A.: *Text To Speech Synthesis. New Paradigms and Advances.* Prentice Hall, New Jersey 2004
- Nikleczy, P. / Olaszy, G.: "A reconstruction of farkas kempelen's speaking machine", In *Proceedings of Eurospeech*, Geneva 2003.
- Olive, J. P.: *The Talking Computer: Text to Speech Synthesis.* In: D. G. Stork (Ed.), *HAL's Legacy: 2001's Computer as Dream and Reality.* MIT Press, Cambridge, MA, 1997.
- O'Shaughnessy, D.: *Speech Communication: Human and Machine.* Addison-Wesley, Reading, PA, 1987.
- Ostermann, J., Beutnagel, M., Fischer, A., Wang, Y.: *Intergration of Talking Heads and Text-to-Speech Synthesizers for Visual TTS.* In: *Proceedings of the ICSLP*, Sydney 1998.
- Pandzic, I., Ostermann, J., Millen, D.: *User evaluation: Synthetic talking faces for interactive services,* *The Visual Computer*, Volume 15, Issue 7/8, 11/04, 1999.
- Parke, F. I., Waters, K.: *Computer facial animation,* Wellesley MA: A K Peters, 1996.
- Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.: *Synthesizing realistic facial expressions from photographs.* In *Computer Graphics Proceedings SIGGRAPH'98*, 1998.

- Pighin, F., Szeliski, R., Salesin, D.: Modelling and animating realistic faces from images. *International Journal of Computer Vision*, 50, 2002.
- Platt, S.M., Badler, N.I. : Animating facial expressions. *Computer Graphics* 15(3), 1981.
- Pompino-Marschall, B.: Einführung in die Phonetik. DeGruyter Studienbuch, Berlin, New York 1996.
- Potamianos, G., Neti, C., Luettin, J., matthews, I.: Automatic Audio-Visual Speech Recognition: An overview, in: *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds., MIT Press, 2004.
- Rabiner, L., Juang, B-H.: *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey 1993.
- Ratnaparkhi Adwait: *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. Dissertation. University of Pennsylvania, 1998.
- Sagisaka, Y.: Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, 1988.
- Sagisaka, Y., N. Kaiki, et al.: ATR-talk speech synthesis system. In *Proceedings of the Intl. Conf. on Spoken Language Processing*, vol. 1, 1992.
- Schukat-Talamazzini, Günther, E.: *Automatische Spracherkennung*. Vieweg, 1995.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.: ToBI: a standard for labeling English prosody. *Proc. ICSLP*, Canada, 1992.
- Spencer, A.: *Phonology*, Blackwell, Oxford, 1996.
- Sproat, R.: *Multilingual Text-to-Speech Synthesis*. Kluwer, Dordrecht 1998.
- Stork, D., G.: *HAL's Legacy 2001. Computer as a Dream and Reality*. MIT Press, Cambridge, 1998.
- Stöber, K., Hess, W.: Additional Use of Phoneme Duration Hypothesis in Automatic Speech Segmentation. In: *Proceedings of the ICSLP*, Sydney 1998.

- Stöber, K., Wagner, P., Helbig, J., Köster, S., Stall, D., Thomae, M., Blauert, J., Hess, W., Hoffmann, R., Mangold, H.: Speech Synthesis by Multilevel Selection and Concatenation of Units from Large Speech Corpora. In: Wolfgang Wahlster (ed.), *VerbMobil: Foundations of Speech-to-Speech Translation, Symbolic Computation*, Springer, Berlin.
- Stöber, K.: Bestimmung und Auswahl von Zeitbereichseinheiten für die konkatenative Sprachsynthese. Dissertation. IKP, Universität Bonn 2002.
- Stylianou, Y., Syrdal, A. K.: Perceptual and objective detection of discontinuities in concatenative speech synthesis. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, USA, 2001.
- Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning. In: *Introduction to Statistical Relational Learning*, Lise Getoor, Ben Taskar (eds), MIT Press, 2006.
- Syrdal, A. K., Wightman, C.W., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Strom, V., Lee, K-S., Makashay, M.J.: Corpus-based techniques in the AT&T NextGen synthesis system. *Proc. ICSLP*, Vol. 3, Beijing, China, 2000.
- Terzopoulos, D., K. Waters: Physically-based facial modeling, analysis and animation. *The Journal of Visual and Computer Animation* 1, 1990.
- Tokuda, K. Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. ICASSP*, Istanbul 2000.
- Tokuda, K., Zen, H., and Black, A.: An HMM-Based Speech Synthesis System applied to English *IEEE TTS Workshop*, Santa Monica, CA 2002
- Tsuzaki, M., Kawai, H.: Feature extraction for unit selection in concatenative speech synthesis: comparison between AIM, LPC, and MFCC. *Proc. ICSLP*, Denver, U.S.A., Sep. 2002.
- Vary, P.; Heute, U.; Hess, W.: *Digitale Sprachsignalverarbeitung*. Teubner, Stuttgart, 1998.

- Vatikiotis-Bateson, E., Kroos, C., Kuratate, T., Munhall, K. G., Rubin, P., Yehia, H.: Building talking heads: Production based synthesis of audiovisual speech. In Proceedings of the First IEEE-RAS International Conference on Humanoid Robots, Cambridge, USA, September 2000.
- Vatikiotis-Bateson, E., Kuratate, T., Munhall, K. G., Yehia, H.: The production and perception of a realistic talking face. In: Proceedings of LP'98, Item order in language and speech, volume 2, pages 439-460, Columbus, USA, September 2000.
- Von Kempelen, W.: Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine, Wien: J.V. Degen, daselbst auch in Französisch erschienen, *Le Mécanisme de la parole, suivi de la description d'une machine parlante*. Ein Faksimile-Neudruck der deutschsprachigen Version.
- Herbert E. Brekle und Wolfgang Wildgren, ist 1970 bei Frommann-Holzboog in Stuttgart erschienen. Pressburg, der damaligen Hauptstadt von Ungarn, geboren und starb 1804 in Wien.
- Wahlster, W., Reithinger, N., Blocher, A.: SmartKom: Multimodal communication with a life-like character. In: Proceedings of the European Conference on Speech Communication and Technology (EuroSpeech), Band 3, Aalborg, Dänemark, 2001.
- Waters, K. (1987): A muscle model for animating three-dimensional facial expression. *Computer Graphics* 21(4), pp. 17-24.)
- Wouters, J., Macon, M.: A perceptual evaluation of distance measures for concatenative speech synthesis, in Proc. 5th Int. Conf. Spoken Language Processing ICSLP, vol. 6, Sydney, 1998.
- Wouters, J., Macon, M.: Control of spectral dynamics in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, 2001.
- Yang, J., Xiao, J., Ritter, M.: Automatic Selection of Visemes for Image-based Visual Speech Synthesis. In: Proceedings of First IEEE International Conference on Multimedia IEEE ME 2000.
- Zwicker, E.: Psychoakustik. Hochschultext, Springer, Berlin, 1982.

Ananova. <http://www.ananova.com/video/>

AVISS. <http://www.ikp.uni-bonn.de/~cwe>

HTK, Hidden Markov Model Toolkit: <http://htk.eng.cam.ac.uk>

http://www.ikp.uni-bonn.de/dt/lehre/materialien/grundl_phon/index.html

<http://www.speech.kth.se/multimodal>

<http://www.phon.ucl.ac.uk/home/sampa/home.htm>

<http://www.arts.gla.ac.uk/IPA/ipa.html>

ABKÜRZUNGSVERZEICHNIS

2D Zweidimensional

3D Dreidimensional

AVS Audio-visuelle Synthese

CART Classification and regression tree.

C4.5 Software für Entscheidungsbaumlernen.

CTS Content-to-Speech

DFT Diskrete Fourier Transformation

F0 Grundfrequenz

FFT Fast Fourier Transformation

HMM. Hidden-Markov-Modell

IKP Institut für Kommunikationsforschung und Phonetik

IPA.Internation Phonetic Association .

KNN K nächster Nachbar (K-Nearest-Neighbor)

LPC. Linear Predictive Coding

MFCC Mel Frequency Cepstral Coefficients

MLSA Mel log spectrum approximation

POS. Part-of-Speech

PSOLA. Pitch Synchronous Overlap Add.

SAMPA.

ToBI. Tone and Break Indices

TTS. Text-to-Speech

XML Extensible Markup Language

ANHANG A

Evaluation Markov-Entropie basierter Sprachsegmentauswahl:

1. Sie haben um sieben Uhr einen Termin in Hannover.
2. Sie haben um sieben Uhr einen Termin in Hamburg und um acht Uhr in Hannover.
3. Wollen Sie mit dem Flugzeug nach Hannover oder fahren Sie mit dem Auto?
4. Nehmen Sie dann um zehn Uhr das Flugzeug nach Frankfurt?
5. Sie haben kein Auto bei uns reserviert.
6. Guten Tag Frau Müller, wollen Sie auch am Freitag arbeiten?

Evaluation Conditional-Random-Field-basierte Sprachsegmentauswahl:

1. Guten Tag Herr Metze, fliegen Sie heute nach Frankfurt?
2. Ich habe hier drei Hotels zur Auswahl.
3. Welches Hotel in Frankfurt haben Sie gebucht?
4. Wollen Sie um zehn nach Frankfurt?
5. Guten Tag, Sie haben um sieben einen Termin.
6. Wollen sie nach Frankfurt oder nach Hannover fliegen?

Evaluation der audio-visuellen Synthese:

1. Guten Tag, haben Sie einen Termin?
2. Das ist das Flugzeug nach Hannover.
3. Gut, dann nehmen Sie das Flugzeug nach Hannover.
4. Guten Tag, wir haben Ihr Auto in Frankfurt.
5. Wollen Sie nach Frankfurt oder nach Hannover fliegen?
6. Sie haben kein Auto bei uns reserviert.

SAMPA

P	Pein / paIn
b	Bein / baIn
t	Teich / taIC
d	Deich / daIC
k	Kunst / kUnst
g	Gunst / gUnst
ʔ	Verein / fɛ6"ʔaIn
pf	Pfahl / pfa:l
ts	Zahl / tsa:l
tS	deutsch / dOYtS
dZ	Dschungel / "dZUN=l
f	fast / fast
v	was / vas
s	Tasse / "tas@
z	Hase / "ha:z@
S	waschen / "vaS=n
Z	Genie / Ze"ni:
C	sicher / "zIC6
j	Jahr / ja:6
x	Buch / bu:x
h	Hand / hant
m	mein / maIn
n	nein / naIn
N	Ding / dIN
l	Leim / laIm
R	Reim / RaIm
l	Sitz / zIts
E	Gesetz / g@"zEts
a	Satz / zats
O	Trotz / trOts
U	Schutz / SUts
Y	hübsch / hYpS
9	plötzlich / "pl9tsIIC
i:	Lied / li:t
e:	Beet / be:t
E:	spat / SpE:t
a:	Tat / ta:t

o:	rot / ro:t
u:	Blut / blu:t
y:	süß / zy:s
2:	blöd / bl2:t
aI	Eis / aIs
aU	Haus / haUs
OY	Kreuz / krOYts
@	bitte / "bIt@
6	besser / "bEs6
i:6	Tier / ti:6
I6	Wirt / vI6t
y:6	Tür / ty:6
Y6	Türke / "tY6k@
e:6	schwer / Sve:6
E6	Berg / bE6k
E:6	Bär / bE:6
2:6	Föhr / f2:6
96	Wörter / "v96t6
a:6	Haar / ha:6
a6	hart / ha6t
u:6	Kur / ku:6
U6	kurz / kU6ts
o:6	Ohr / o:6
O6	dort / dO6t

STTS (Stuttgart-Tübinger Tagset, <http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html>)

ADJA: attributives Adjektiv	(das) große (Haus)
ADJD: adverbiales / prädikatives Adjektiv	(er fährt) schnell / (er ist) schnell
ADV: Adverb	schon, bald, doch
APPR: Präposition; Zirkumposition links	in (der Stadt), ohne (mich)
APPRART: Präposition mit Artikel	im (Haus), zur (Sache)
APPO: Postposition	(ihm) zufolge, (der Sache) wegen
APZR: Zirkumposition rechts	(von jetzt) an
ART: bestimmter oder unbestimmter Artikel	der, die, das / ein, eine

CARD: Kardinalzahl.....	zwei (Männer), (im Jahre) 1994
FM: Fremdsprachliches Material.....	(Er hat das mit ``) A big fish (" übersetzt)
ITJ: Interjektion	mhm, ach, tja
ORD: Ordinalzahl.....	(der) neunte (August)
KOUI: unterordnende Konjunktion	um (zu leben), mit ``zu" und Infinitiv anstatt
KOUS: unterordnende Konjunktion	weil, dass, damit, mit Satz wenn, ob
KON: nebenordnende Konjunktion	und, oder, aber
KOKOM: Vergleichskonjunktion	als, wie
NN: normales Nomen.....	Tisch, Herr, (das) Reisen
NE: Eigennamen.....	Hans, Hamburg, HSV
PDS: substituierendes Demonstrativpronomen.....	dieser, jener
PDAT: attribulierendes Demonstrativpronomen.....	jener (Mensch)
PIS: substituierendes Indefinitpronomen.....	keiner, viele, man, niemand
PIAT: attribulierendes Indefinitpronomen ohne Determiner.....	kein (Mensch), irgendein (Glas)
PIDAT: attribulierendes Indefinitpronomen mit Determiner	(ein) wenig (Wasser),
PPER: irreflexives Personalpronomen	ich, er, ihm, mich, dir
PPOSS: substituierendes Possessivpronomen	meins, deiner
PPOSAT: attribulierendes Possessivpronomen	mein (Buch), deine (Mutter)
PRELS: substituierendes Relativpronomen.....	(der Hund ,) der
PRELAT: attribulierendes Relativpronomen.....	(der Mann ,) dessen (Hund)
PRF: reflexives Personalpronomen	sich, einander, dich, mir
PWS: substituierendes Interrogativpronomen	wer, was
PWAT: attribulierendes Interrogativpronomen	welche (Farbe), wessen (Hut)
PWAV: adverbiales Interrogativ oder Relativpronomen	warum, wo, wann,
PAV: Pronominaladverb.....	dafür, dabei, deswegen, trotzdem
PTKZU: ``zu" vor Infinitiv	zu (gehen)
PTKNEG: Negationspartikel	nicht
PTKVZ: abgetrennter Verbzusatz	(er kommt) an, (er fährt) rad
PTKANT: Antwortpartikel	ja, nein, danke, bitte
PTKA: Partikel bei Adjektiv oder Adverb	am (schönsten), zu (schnell)
SGML:	SGML Markup
SPELL: Buchstabierfolge	S-C-H-W-E-I-K-L
TRUNC: Kompositions-Erstglied	An- (und Abreise)
VVFIN: finites Verb, voll.....	(du) gehst, (wir) kommen (an)
VVIMP: Imperativ, voll	komm (!)
VVINF: Infinitiv, voll	gehen, ankommen
VVIZU: Infinitiv mit ``zu", voll	anzukommen, loszulassen
VVPP: Partizip Perfekt, voll	gegangen, angekommen

VAFIN: finites Verb, aux (du) bist, (wir) werden
 VAIMP: Imperativ, aux sei (ruhig !)
 VAINF: Infinitiv, aux werden, sein
 VAPP: Partizip Perfekt, aux gewesen
 VMFIN: finites Verb, modal dürfen
 VMINF: Infinitiv, modal wollen
 VMPP: Partizip Perfekt, modal gekonnt, (er hat gehen) können
 XY: Nichtwort, Sonderzeichen enthaltend 3:7, H2O, D2XW3

GToBI (German Tone and Break Indices,
<http://www.uni-koeln.de/phil-fak/phonetik/gtobi/index.html>)

Übersicht über die sechs Akzenttypen und Grenztöne

H *	Gipfelakzent-Ton
H * + L	Anstieg von einem tiefen L- Ton zu einem Gipfelakzent-Ton
L *	Tiefer Akzent-Ton
L * + H	Tiefer Akzent-Ton mit anschließendem Anstieg des F0-Wertes.
H + L *	Abfall von hohem F0-Wert zu einem tiefen Akzent-Ton
H + ! H	Downstepping von einem hohen F0-Wert zu einem mittleren Akzent-Ton
L -	L – kennzeichnet einen tiefen F0-Wert auf der Baseline
H -	H – wird auf der Topline mit einem hohen F0-Wert wiedergegeben
! H -	H-Ton mit einem Downstepfaktor.
H - %	Hoher F0-Wert zum Ende einer Phrase
H - ^ H %	Hoher F0-Wert mit einem starken Anstieg zum Ende der Phrase
L - H %	Tiefer F0-Wert mit einem mittlern Anstieg zum Phrasenende
L - %	Tiefer F0-Wert, wenn nötig mit einem absinkenden F0-Wert
% H	Initiale Markierung mittels eines hohen F0-Wertes

PUBLIKATIONEN

Weiss, C., Hess, W.: Conditional Random Fields for Hierarchical Segment Selection in TTS. In: Proceedings of Interspeech 2006, Pittsburgh, USA, 2006.

Weiss, C., Da Silva, M.R., Tokuda, K., Hess, W.: Low Resource HMM-based Speech Synthesis applied to German. In Hoffmann, R., Proceedings 16. Conference „Elektronische Sprachsignalverarbeitung“ (ESSP 2005), Prag, Czech Republic 2005.

Weiss, C.: FSM and K-Nearest-Neighbor for Corpus based Video-Realistic Audio-Visual Synthesis. In: Proceedings of Interspeech, Lisbon, Portugal 2005.

Weiss, C., Aschenberger, B.: A German Viseme-Set for Automatic Transcription of Input Text Used for Audio-Visual-Speech-Synthesis. In: Proceedings of Interspeech, Lisbon, Portugal 2005.

Weiss, C.: Audio-visuelle Synthese mittels HMM basierter Segmentauswahl, In: Fortschritte der Akustik - DAGA, München, 2005.

Weiss, C.: Markov-Entropie basierte Auswahl geeigneter Sprachsegmente für Korpusbasierte Sprachsynthese Systeme, in: Fellbaum, K., Tagungsband 15. Konferenz „Elektronische Sprachsignalverarbeitung“, ESSV, Cottbus 2004.

Weiss, C.: Framework for data-driven video-realistic audio-visual speech synthesis, in Proceedings of Fourth Int. Conf. on Language Resources and Evaluation LREC, Lisbon, Portugal 2004.

Weiss, C.: Using video realistic audio-visual synthesis output in dialogue systems. 5th SIGdial Workshop on Discourse and Dialogue Cambridge, MA, 2004.

Weiss, C.: Videorealistische audiovisuelle Synthese basierend auf Unit-Selection, in Krotschel, C., Tagungsband 14. Konferenz "Elektronische Sprachsignalverarbeitung", ESSV, Karlsruhe 2003.