Chapter 1

Introduction

1.1 Zusammenfassung

Die vorliegende Arbeit trägt zum aktuellen Forschungsstand der Sprachsynthese für europäisches Portugiesisch bei, indem ein Sprachsynthese- System entwickelt wurde, welches mit wenig Rechenleistung auskommt. Ebenso wurde ein spezielles Korpus hinzu entwickelt, mit dem das Sprachsythese-System trainiert werden kann. Weiterhin werden neue Ansätze zur Sprachsynthese eingeführt, die natürlich klingende und sehr verständliche Sprache aus Text generieren. Die Arbeit stellt innovative Lösungen vor, um Sprachsynthese auf Geräten mit wenig Rechenleistung, wie mobile Computer, oder Mobiltelefone auszuführen. Das entwickelte Sprachsynthese System ist das erste für europäisches Portugiesisch, das die Sprachparameter direkt aus Hidden-Markov Modellen erzeugt. Das Sprachsynthese System ist vollständig implementiert und läuft unter dem Linux Betriebsystem sowie unter dem Windows Betriebssystem. Die vorliegende Arbeit trägt auch zur Forschung bezüglich Sprachdatenkorpora für europäisches Portugiesisch bei. Es wurde ein kontext-basiertes Sprachdatenkorpus manuell ausgearbeitet und erstellt. Das Sprachdatenkorpus besteht aus je einem Satz, in dem jedes Diphon abhängig von der Koartikulation und zugehöriger Vokalreduktion in der gesprochen Sprache abgebildet mit zugehöriger phonetischer Transkription wird. Weiterhin wurde ein Text-Vorverarbeitungsmodul für europäisch portugiesische Sprachsynthese entwickelt, bestehend aus einem automatischen Graphem-Phonem Umsetzungsmodul, einer automatischen Silbengrenzen-Erkennung, sowie einer automatischen Silbenakzent Vorhersage. Die automatischen Vorhersagen werden mittels statistisch motivierten Modellen erzeugt. Für die Berechnung der Modelle wird der Maximum Entropie Algorithmus eingesetzt, der erfolgreich für natürlichsprachliche Textverarbeitung eingesetzt an anderer Stelle verwendet wird.

Es wird eine ausführliche Übersicht gegeben über Algorithmen und Methoden zum Einsatz des Quelle-Filter Modells, sowie Algorithmen, die die Vokaltrakt Funktion simulieren und manipulieren. Des Weiteren findet sich eine grundlegende Übersicht wieder, über konkatenative Sprachsynthese Ansätze, bei denen Sprachbausteine aus einem Sprachdatenkorpus ausgeschnitten werden und zu einem neuen Sprachsignal wieder zusammengesetzt werden. Es wird das Hidden-Markov Modell (HMM) basierte Sprachsynthese Verfahren eingeführt und die Verwendung von statistischem HMM Lernverfahren zur Generierung von Sprachsignalparametern erläutert, die dann automatisch als statistisches Modell trainiert werden können. Dieser Ansatz zur Generierung der Sprachsignalparameter mittels HMMs verwendet Spektralkoeffizienten, wie z.B. Linare Prädiktive Koeffizienten (LPC) oder Mel-Frequenz-Cepstrum-Koeffizienten (MFCC). Am häufigsten werden, wie auch in dieser Arbeit, MFCCs als Datenbasis für das Training des statistischen Modells herangezogen. Der Grund für die Auswahl der Parameter liegt im Ansatz, ob eine einfache Impulsfolge mit einem Gaussschen Rauschen als Anregung verwendet wird, wie in dieser Arbeit in Kapitel 5 beschrieben, oder aber LPCs, wenn zum Beispiel die Nutzung des verbleibenden Signals, das Residual-Signal, als Filter für die Anregung verwendet wird, wie es in Kapitel 6 als System-Erweiterung vorgeschlagen ist.

In dieser Arbeit liegt der Fokus auf statistischen Modellen, da diese Vorteile hinsichtlich der verwendeten Sprachressourcen haben. Es wird der Einsatz von HMMs in der Sprachsynthese, im Speziellen für Portugiesisch, untersucht und ein Sprachsynthesesystem entwickelt. HMMs zählen zu den prominentesten statistischen Sprachsynthese Ansätzen. Um die Vorteile der HMM basierten Sprachsynthese besser verstehen zu können, werden die unterschiedlichen Herangehensweisen von Unit-Selection basierten Synthesesystemen und der HMM basierten Sprachsynthese herausgearbeitet und dargestellt. Es wird die Entwicklung der HMM basierten Sprachsynthese aufgezeigt, bei denen zunächst ein Modell für alle Einheiten verwendet wurde, bis zu aktuellen Entwicklungen, bei denen pro Sprachsegment ein Modell generiert und verwendet werden. Ein wichtiger Schritt für den Erfolg der HMMs in der Sprachsynthese ist die Verwendung von dynamischen Eigenschaften von Sprache, welche durch Einschliessen der Vorgängerinformationen der jeweiligen Sprachsegmente verbessert wird. Es werden für den herausgearbeiteten HMM Ansatz die wichtigsten Techniken für den Einsatz von HMMs in der Sprachsynthese beschrieben und im entwickelten System eingesetzt. Die Mel-Cepstrum Analyse-Technik zur Extraktion von Sprachsignalparametern, wird verwendet um Entscheidungsbäume zu trainieren, die den sprachlichen Kontext abbilden und diesen in die einzelnen Zustände der HMMs mit einbeziehen. Weiterhin wird die mehrdimensionale Wahrscheinlichkeitsverteilung für die Unterscheidung von stimmhaften und stimmlosen Lauten zur Integration in die HMMs angewendet, sowie ein spezieller Algorithmus

für die Sprachsignalparameter Generierung, welche dann zu dem Sprachsignal syntheisiert werden.

Eine wichtige Methode der Sprachsynthese, der Unit-Selection basierte Ansatz, wurde untersucht und die Verfahren und Algorithmen dargestellt. Dieser Ansatz nutzt als Datenbasis groe Sprachdatenkorpora aus denen Sprachbausteine als Segmente durch Ausschneiden aus Trägersätzen verwendet werden, um dann die für das Ziel-Sprachsignal geeigneten besten Sprachbausteineinheiten zu verketten. Dieser Ansatz hat eine zweidimensionale Kostenfunktionen als Basis-Algorithmus zur Grundlage, die über manuelle Gewichtung oder durch eine statistische Übergangswahrschein- lichkeit die besten Übergänge der Sprachsegmente errechnet.

In der aktuellen Forschung und Anwendung beruhen Sprachsynthese Systeme meist auf sehr groen Sprachdaten-Korpora. Das hat den Vorteil, dass eine Vielzahl von kontextbasierten Sprachbausteinen repräsentiert werden. Soll nun eine neue Äuerung synthetisiert werden, so nutzt der Unit-Selection basierte Algorithmus eine bestimmte Zielvorgabe, die das Sprachsegment erfüllen muss, und sucht das geeignete Sprachsegment aus den Sprachdaten zur Verkettung mit den anderen Sprachsegmenten heraus. Diese Zielvorgaben werden über eine zweidimensionale Kosten-Funktion abgebildet, um eben die geeigneten Einheiten für die Verkettung zu identifizieren. Die Kostenfunktion selbst ist ein metrischer Algorithmus, der eine quasi Entfernung des phonetischen Zusammenhangs der aufeinanderfolgenden Sprachsegmente wie auch zusätzlich phonologische und prosodische Eigenschaften mit einbezieht. Weiterhin werden die Sprachsignaleigenschaften, die im Spektralbereich identifiziert wurden, in die metrische spektrale Abstandsberechnung der aufeinanderfolgenden Einheiten einbezogen. Der Vorteil eines solchen Systems ist, dass ein sehr natürlich klingendes Sprachsignal erzeugt werden kann. Der Nachteil liegt vor allem in den Expertenkosten bei der Gestaltung und Aufzeichnung der Sprachdatenkorpora, die in der Regel mehrere Stunden Sprache repräsentieren, und den Kosten für die Experten, die das Korpus nachbearbeiten. Darüber hinaus beeinflusst die Qualität der aufgenommen Sprachdaten die Qualität des erzeugten Sprachsignals. So kann es durchaus vorkommen, dass kein geeignetes Sprachsegment im Sprachdatenkorpus abgebildet ist und daher ein unpassendes Sprachsegment ausgewählt wurde. Dieser Umstand beeinträchtigt die Gesamtqualität des erzeugten Sprachsignals erheblich.

Ein Ansatz, der die Nachteile der Unit-Selection basierten Synthesen überwindet, besteht in der Verwendung eines Quelle-Filter-Modells. Die Quelle des Filters kann durch eine einfache Pulsfolge mit akustischen Parametern, wie der Grundfrequenz (F0), angeregt und das Filter mittels der spektralen Parameter realisiert werden, das dann die gesamte Eingabe der Signalparameter zu einer sprachlichen Äuerung resyn-

thetisiert. Um nun die richtigen Parameter für die Quelle und das Filter auszuwählen, wird ein HMM trainiert. Dieser Ansatz ist für die Generierung der Sprachsignale des entwickelten Systems umgesetzt. Die extrahierten Parameter, die als Trainingsdaten für die HMMs verwendet werden, sind die Mel-Cepstrum Parameter sowie F0 und Dauer in Millisekunden. Mit Hilfe von statistisch motivierten Entscheidungsbäumen werden Kontextabhängigkeiten, wie sie in gesprochener Sprache vorkommen, miteinbezogen und die Sprachsegmente in Cluster zusammengefasst. Der Vorteil eines HMM basierten Sprachsynthese-Systems ist, dass es verständliche Sprache mit einer kleinen Menge von Sprachdaten erzeugen kann. Ebenso ist die HMM basierte Sprachsignalgenerierung weniger anfällig für Inkonsistenzen in der Sprachdatenbank. Aufgrund der geringen Sprachdaten sind auch die Expertenkosten zur Erstellung der Sprachdaten- Trainingskorpora geringer. Ein weiterer Vorteil ist die Adaption an neue Sprachen bzw. Stimmen. Der Nachteil eines solchen Systems ist immer noch der Vocoder-Klang der synthetisierten Äusserung, der sich aus dem Quelle-Filter-Modell bzw. aus der Transferfunktion ergibt.

Das entwickelte Sprachsynthese System wurde für in Europa gesprochenes Portugiesisch entwickelt und implementiert. Der Bedarf für ein Sprachsynthese System für Portugiesisch resultiert aus der Anforderung, dass Portugiesisch die siebthäufigste Sprache der Welt in Bezug auf die Zahl der Muttersprachler ist, und mit rund 178 Millionen Muttersprachlern, und es ist die zweithäufigste gesprochene Sprache in Lateinamerika. Portugiesisch wird unter anderem in Angola, Brasilien, auf den Kap Verdischen Inseln, China (Macau) und in Guinea-Bissau gesprochen. Weiterhin in der Indische Union (Daman, Diu und Goa), Indonesien (Flores Island), Ara, Malaysia (Mallaca), Mosambik, Portugal, Sao Tome & Principe, Timor Lorosa'e und Uruguay. Wobei es Amtssprache in acht Ländern ist: Angola, Brasilien, Kap Verde, Guinea-Bissau, Mosambik, Portugal, Sao Tome & Principe, Timor Lorosa'e.

Das portugiesische Alphabet besteht aus dem ursprünglichen lateinischen Alphabet mit 23 Buchstaben. Das Europäische Portugiesisch besitzt ein phonetisches Inventar mit achtunddreiig Phonemen. Besonderheiten des Portugiesischen in der gesprochenen Sprache sind vor allem die Koartikulation zwischen Wrtern und Vokalreduktionen, die sehr häufig auftreten. Beide Effekte werden in dieser Arbeit behandelt und bezüglich der Korporaerstellung für das Portugiesische Sprachsynthesesystem mit einbezogen. Die Besonderheit der Koartikulation zwischen Wrtern wirkt sich zum Beispiel so aus, dass die phonetischen Transkriptionen von aufeinanderfolgenden Worten beeinflusst werden gegenüber den phonetischen Standard Transkriptionen. Die Auswirkungen der Vokalreduktion sind bezüglich gesprochener Sprache und synthetisierter Sprache in der Weise unterschiedlich, dass die synthetisierte Sprache flschlich Vokale hörbar macht, die in der gesprochenen Sprache nicht hörbar wären.

In den Abschnitten 3 bis 5 wird die Implementierung eines vollständigen Sprachsynthesesystems beschrieben, beginnend von der Eingabe des Textes bis zur Ausgabe einer sprachlichen Äuerung. Zu Beginn kommt ein Text Vorverarbeitungs- modul zum Einsatz, welches den eingegebenen Text mit dem Text Vorverarbeitungs- modul Textnormalisierung bearbeitet. Textnormalisierung heisst, dass Abkürzungen, Datum, Telefonnummern, Zahlen, Akronyme und andere Symbole, in lesbaren Text graphemisch umgesetzt werden muss. Wurde der Text aufbereitet und alle nicht graphemischen Textstellen transformiert, wird eine linguistische Analyse der Texteingabe erstellt. Das linguistische Analyse Modul führt zum Beispiel eine morphosyntaktische Analyse durch, die sehr hilfreich ist, um Homographen auszulösen. Auch für die prosodische Vorverarbeitung und die Extraktion prosodischer Merkmale ist das linguistische Analysemodul notwendig. Während die Textnormalisierung und das linguistische Analysemodul auf Graphembasis arbeiten, kommt für die eigentliche Eingabe in das Synthesesystem ein phonetisch transkribierter Text zum Einsatz. Die Umwandlung der Eingabe von Text in ihre korrespondierende phonetische Transkription übernimmt ein Verarbeitungsmodul für natürliche Sprache, kurz NLP-Modul, welches einen Graphem-Phonem Umsetzungsalgorithmus enthält. Das NLP-Modul erfüllt noch weitere wichtige Aufgaben, um Merkmale aus der Texteingabe zu erhalten. So werden weitere nützliche Informationen wie Wörter- und Silbengrenzen erkannt, sowie Wort- und Silbenprominenz bzw. Akzent markiert. Damit die Prosodie auch in der gesprochenen Sprache in der Sprachsynthese wiedergegeben werden kann, müssen auch die prosodischen Muster bestimmt werden. Diese werden vor allem durch die Grundfrequenz und die Dauer bestimmt. Hierzu kommt ein Prosodie-Modul zum Einsatz, dass in das Sprachsynthese System integriert wird. Information, wie F0 und segmentale Dauer werden hier über statistische Verfahren geschätzt. Das letztes Modul, welches dann die extrahierten Merkmale und Sprachsignalparameter verarbeitet und den Text in eine sprachliche Äuerung überträgt, ist das Signalgenerierungsmodul.

Von groer Wichtigkeit für die Qualität von Sprachsynthese Systemen ist die Gestaltung der Sprachdatenkorpora. Kapitel 5 befasst sich ausführlich mit der Gestaltung und Entwicklung eines geeigneten Sprachdatenkorpus für die HMM basierten Synthese-Systeme, wie in dieser Arbeit umgesetzt. Statistische Systems wie das HMM System berechnen die Übergangswahrscheinlichkeiten zwischen den einzelnen Zuständen. Von daher ist das Einbeziehen von sprachlichem Kontext in der Korpuserstellung notwendig, um alle Ereignisse, die auftreten können, abzudecken. Gewöhnlich ist es nicht möglich alle auftretenden Ereignisse in der gesprochenen Sprache mit einem kleinen Sprachdatenkorpus zu berücksichtigen und diese aufzunehmen. Bei der Erstellung eines geeigneten Sprachdatenkorpus wurde speziell auf diese Einschränkungen eingegangen und ein Korpus entwickelt, dass mit wenigen

Daten möglichst alle sprachlichen Ereignisse abdeckt. Dieses Korpus wurde manuell zusammengestellt. Das neue Sprachdatenkorpus wurde auch unter Berücksichtigung der speziellen Anforderungen und den bisherigen wissenschaftlichen Erkenntnissen zur Erstellung von Sprachsynthesekorpora entwickelt. Hier wurde vor allem auf das Korpus der Ingenieursfakultät der Universität Porto, Portugal, FEUP-IPB-Datenbank zurückgegriffen. Es wurde ein professioneller männlicher Sprecher aufgenommen, der im Vorlesestil die aufzunehmenden Sätze gesprochen hat. Das neue Sprachdatenkorpus ist folgendermaen repräsentiert: Graphem-Sätze, phonetische Transkription, typische Phänomen mit Markierungen der in Europäischem Portugiesisch typischen Effekte der Koartikulation und Vokalreduktion. Das Korpus wird für den späteren Einsatz durch die Verteilung der Phoneme im Gesamten, sowie die Verteilung von Phonemen innerhalb der Sätze und Wörter repräsentiert. Es wurde eine Analyse zwischen der Anzahl der Einheiten in der phonetischen Transkription mit und ohne Berücksichtigung der Vokalreduzierung vorgelegt, um die Bedeutung dieses Effektes im Europäischen Portugiesisch nachvollziehen zu können, sowie eine Analyse der verwendeten Diphone. Zur Darstellung wurde eine Verwechslungsmatrix generiert. Für die Graphem-Phonem Umsetzung wurde ein statistisches Werkzeug auf Basis des Maximum Entropie Frameworks entwickelt und eingesetzt. Maximum Entropie ist ein statistisch motiviertes Modell, das u.a. erfolgreich für das Tagging von sequentiellen Daten wie Part-of-Speech (POS) Tagging oder für das syntaktische Parsing angewandt wurde. Die Entscheidung für statistisch motivierte Werkzeuge basiert auf einer Kostenabschätzung für die einzelnen Teilschritte. Regel-basierte Systeme erfordern Expertenwissen, wohingegen statistisch motivierte Systeme auch ohne linguistisches Fachwissen umgesetzt und betrieben werden können. Vor allem bei der Umsetzung von Graphemen in entsprechende Phoneme ist Expertenwissen und Erfahrung des Experten entscheidend. Statistische Systeme zeigen sich hier flexibler, sofern die statistischen Modelle mit ausreichenden und sinnvollen Daten trainiert werden, wie zuvor angedeutet mit Berücksichtigung von Effekten der gesprochenen Sprache, Koartikulation und Vokalreduktion. In europäischem Portugiesisch existieren achtunddreiig Phoneme, da einige Grapheme in den Transkriptionen durch eine Kombination von mehr als einem Phonem repräsentiert werden, entstehen vierundvierzig Phon-Klassen statt achtunddreiig.

Bei der Vorhersage des Wort- und Silbenakzents kommt eine binäre Entscheidung zum Tragen, indem eine Klasse die betonten Silben widerspiegelt und eine Klasse für keinen gesetzten Akzent steht. Die gleiche binäre Klassifikation kommt bei der Entscheidung Silbengrenze oder nicht zum Einsatz. Zur Einschätzung der Qualität der Vorhersagen der statistischen Werkzeuge wird das Qualitätsma Word Error Rate (WER) und Phoneme Error Rate (PER) eingeführt, wobei sich ersteres auf die Fehlerrate der Worte bezieht und das zweite Ma auf die Substitutionsfehler der