universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Synthese und prosodischer Transfer zweier südbairischer Dialekte.
Sprachsynthese als Beispiel für ein interdisziplinäres Forschungsfeld der Linguistik."

verfasst von / submitted by
Carina Lozo

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Master of Arts (MA)

Wien, 2018 / Vienna, 2018

Studienkennzahl lt. Studienblatt/ degree programme code as it appears on the student record sheet:	A 066 899
Studienrichtung lt. Studienblatt/ degree programme as it appears on the student record sheet:	Angewandte Sprachwissenschaft
Betreut von / Supervisor:	Univ.-Prof. Dr. Daniel Büring

Danksagung

An dieser Stelle möchte ich mich herzlich bei Daniel Büring bedanken, der meine Arbeit in einer für mich schwierigen Situation ohne Zögern übernommen hat.

Mein besonderer Dank gilt Michael Pucher, der mir den Anstoß für diese Arbeit gegeben hat. Sein Ehrgeiz und Optimismus haben mich von Beginn an stets begleitet und mich auch abseits dieser Arbeit inspiriert. Die Zusammenarbeit mit ihm hat mir gezeigt, Problematiken aus verschiedenen Perspektiven zu beleuchten und wie essentiell ein disziplinübergreifender Ansatz für linguistische Themen sein kann.

Ebenso möchte ich Sylvia Moosmüller danken, die mich durch die Schwierigkeiten des phonetischen Transkribierens von Dialekten mit viel Geduld, Humor und umfangreichen Wissen manövrierte. Ebenso bestärkte sie mich, mich einem untypischen Thema in der Linguistik anzunehmen. Nicht zuletzt ist ihr und ihrer Forschung meine Begeisterung für die Phonetik und Sprachwissenschaft zu verdanken, die mich noch viele weitere Jahre begleiten wird.

Weiters möchte ich meinem lieben Kollegen Jan Luttenberger für sein unentbehrliches Korrekturlesen danken. Seine hilfreichen Anregungen und die unzähligen Diskussionen mit ihm über verschiedenste Themen haben mir letztendlich geholfen diese Arbeit fertigzustellen.

Schließlich gilt mein Dank auch meinem Probanden aus der Steiermark, der geduldig und konzentriert meinem Mikrophon seine Stimme spendete und der diese Studie erst ermöglichte.

Inhaltsverzeichnis

1 Einleitung **4**

2 Sprachsynthese **7**
2.1 Einführung in die Sprachsynthese 7
2.2 Forschungslandschaft 8
2.2.1 Historischer Überblick der wichtigsten Sprachsynthesesysteme . 12
2.2.2 Hidden-Markov-Modelle 18
2.2.3 Dialekt- bzw. Varietätensynthese 21

3 Der (süd-)bairische Dialektraum und seine Phonetik **24**
3.1 Allgemeine phonologische Merkmale des Bairischen 25
3.2 Die Eigenheiten von Osttirol und der Südweststeiermark 28

4 Empirischer Teil **30**
4.1 Experiment 31
4.1.1 Korpora 31
4.1.2 Synthese-Sätze 32
4.1.3 Phonsets 34
4.2 Phonmapping 37
4.3 Synthese 39
4.4 Evaluation und Ergebnisse 40

5 Conclusio **43**

6 Literatur **45**

Abbildungsverzeichnis **52**

Tabellenverzeichnis **53**

1 Einleitung

Dialekte kodieren längst nicht mehr primär geographische Herkunft, sondern vielmehr auch soziale Identität. Diese Entwicklung kann zum Anlass genommen werden, neue Erkenntnisse der Dialektologie über diese hinaus in anderen Disziplinen wirken zu lassen.

Die Dialektsynthese kann einen weiteren fruchtbaren Aspekt für die zeitgemäße Dialektologie darstellen, den sie als Chance wahrnehmen soll, sich als relevantes Forschungsfeld außerhalb der klassischen Sprachwissenschaft zu verorten und als wichtige Ergänzung in der Sprachtechnologie wahrgenommen zu werden.

Die Sprachsteuerung von Mobiltelefonen, vielseitige Sprachassistenten von Unterhaltungselektronik oder inzwischen auch der wachsende Bereich der Smart-Home-Anwendungen zeigen, dass die Sprachsynthese mit dem technologischen Fortschritt fortwährend Erfolge feiert und im Schnellschritt zu einem festen Bestandteil des Alltags heranwächst. Dass die Sprachsynthese sich nicht nur auf die Erzeugung von Standardsprache beschränken soll, zeigen bereits einige Studien: Synthesesysteme, die Akzentmodelle bzw. innersprachliche Varietäten erzeugen, werden einerseits von den SprecherInnen dieser Varietäten, andererseits auch in speziellen Anwendungsbereichen bevorzugt (vgl. Tamagawa et al. 2011; Wutiwiwatchai et al. 2011).

Während es bereits seit den 1990er Jahren erste Bemühungen gab, Spracherkennungssysteme für Dialekte zu adaptieren (Humphries et al. 1996; Huang et al. 2000), liegt in der Syntheseforschung erst mit der Arbeit von Pucher et al. 2010b eine Untersuchung vor, die die Möglichkeiten einer innersprachlichen Varietätensynthese untersucht.

Die aktuelle linguistische Forschung zu Varietäten innerhalb Österreichs bezieht sich derzeit vor allem auf phonetische, syntaktische oder morphologische Unterschiede (wie bspw. Lenz & Patocka (2016)); oder erörtert noch methodische Ansätze dazu (wie Breuer 2017; Lenz 2016; Lenz & Patocka 2016). Im deutschsprachigen Raum außerhalb Österreichs konzentrierte sich die Forschung an Varietäten eine längere Zeit fast ausschließlich auf segmentale Aspekte der Sprache; die Suprasegmentalia blieben weitgehend unerforscht, obwohl sich in den letzten Jahren mehr und mehr äußerte, dass

die Prosodie ein salienter Aspekt von dialektalen Varietäten zu sein scheint (Siebenhaar 2004: 353).

In dieser Arbeit wird mit der Prosodie ein Thema aufgegriffen, dessen Bedeutung für die Perzeption von innersprachlichen Varietäten intuitiv schon immer augenscheinlich war. Die prosodischen Unterschiede zwischen Varietäten können einerseits als regionaler Identifikator für die HörerInnen dienen, andererseits aber auch als bewusster Index für Authentizität und Natürlichkeit instrumentalisiert werden. Lange Zeit beschränkten sich Untersuchungen zur Prosodie des Deutschen bzw. deutscher Varietäten auf isolierte Sätze. Erst Arbeiten wie Selting (1995) und Selting (2004) beschäftigten sich mit der Prosodie von spontansprachlichen Daten. Dass die Forschung an natürlichen Daten vorteilhaft ist, zeigte u.a. ein DFG-Projekt zur regionalen Prosodievariation zwischen deutschen Varietäten (»Untersuchungen zur Struktur und Funktion regionalspezifischer Intonationsverläufe im Deutschen« (Peters et al. 2015). Hier konnte eine regionalspezifische Prosodie zumindest in Deutschland und der Schweiz belegt werden; ebenso, dass die varietätenspezifische Prosodie innerhalb eines Sprachraumes als Identifikator der Region gelten kann.

Wie ein Großteil der linguistischen Forschung, bezieht sich auch die Mehrheit der Prosodieforschung vor allem auf die Beschreibung von Standardsprachen bzw. Standardvarietäten (El Zarka et al. 2017; Schmid & Moosmüller 2017; Ulbrich 2005). Dabei gestehen bereits Arbeiten der frühen Dialektologie wie etwa Socin (1888) ein, dass die Intonation einer regionalen Varietät ein wichtiges Distinktionsmerkmal sei.

Hagmüller (2001) und Dizdarevic et al. (2004) untersuchten bereits Anfang des vorherigen Jahrzehnts, ob ein prosodiebasierter Ansatz für die automatische Spracherkennung (ASR) von österreichischen Varietäten geeignet ist. Darauf folgte erst mit Pucher et al. (2017a) eine Arbeit, die sich diesem Thema im Bereich der Sprachsynthese annahm.

Diese Arbeit möchte diese Erkenntnisse zum Anlass nehmen und die Anwendungsgebiete prosodiebasierter Ansätze für die Sprachsynthese anhand zwei südbairischer Varietäten (Südweststeirisch und Innervillgratisch) erforschen. Mit dem Transfer der Grundfrequenz und Dauer eines Originalsprechers soll gezeigt werden, dass ein methodischer Ansatz, der variationsspezifische Prosodie berücksichtigt, einen Vorteil für synthetisierten Sprachoutput bedeutet.

Diese Arbeit soll vor allem zwei Forschungsdesiderate erfüllen: Einerseits, wie phonetische Kenntnisse über Varietäten für effiziente neue Synthesizer zu nützen sind, andererseits, inwiefern Dialektprosodie bei der Erweiterung dieser Systeme hilfreich

sein kann. Diese Arbeit ist wie folgt gegliedert: Abschnitt zwei liefert einen Überblick über die allgemeinen Aspekte der Sprachsynthese, skizziert die Forschungslandschaft, sowie einen historischen Überblick über die relevantesten Sprachsynthesesysteme. Weiters wird in Abschnitt zwei noch auf Hidden-Markov-Modelle und Dialektsynthese eingegangen. Im dritten Abschnitt dieser Arbeit findet sich eine Beschreibung des südbairischen Dialektraums, wobei auch genauer auf die Merkmale der hier relevanten Dialekte Südweststeirisch und Innervillgratisch eingegangen wird. Bevor Abschnitt fünf die Arbeit mit abschließendem Kommentar zu Ende führt und reflektiert, werden in Abschnitt vier das Experiment, seine Methoden und Ergebnisse genauer beleuchtet.

Abschnitt vier dieser Arbeit wurde bereits in Pucher et al. (2017a) publiziert.

2 Sprachsynthese

2.1 Einführung in die Sprachsynthese

Als Sprachsynthese wird die künstliche Erzeugung von Sprache bezeichnet, wobei neben verschiedenen Synthesesystemen ebenso Methoden der Sprachsynthese zu unterscheiden sind. Grundsätzlich lassen sich fünf relevante Formen differenzieren, nämlich die mechanische, Formant-, konkatenative, artikulatorische und schließlich die parametrische Synthese. Mechanische Sprachsynthese zielt darauf ab, mithilfe von physiologischen bzw. anatomischen Erkenntnissen über die Lauterzeugung den menschlichen Sprachapparat nachzubilden, um dadurch zur Spracherzeugung zu gelangen. Im Gegensatz dazu basiert die Formantsynthese auf akustischen Erkenntnissen der Lauterzeugung. Hier werden über ein Quellsignal verschiedene Filter gelegt (»Quelle-Filter-Theorie«), die die akustischen Gegebenheiten des Vokaltrakts bei Lautproduktion simulieren sollen. In der konkatenativen Synthese wird durch die Aneinanderkettung bestimmter Lauteinheiten aus einer Datenbank (sogenannte »units«, daher auch »Unit-Selection«-Systeme) die gewünschte Lautfolge erzeugt. Die artikulatorische Synthese versucht, den gesamten Prozess der Sprachproduktion mathematisch zu modellieren. Schließlich werden bei der parametrischen Sprachsynthese einzelne Laute anhand ihrer akustischen Parameter beschrieben und in ihre Komponenten aufgeteilt, aus welchen man dann akustische oder statistische Modelle für die Synthese ableiten kann.

In der Regel handelt es sich bei den meisten Sprachsynthesizern um Text-to-Speech-Systeme (TTS), die Text in einen Sprachoutput verwandeln. TTS-Systeme können in zwei Komponenten aufgeteilt werden, nämlich Textanalyse und die tatsächliche Sprachsynthese. Die Textanalyse konvertiert einen gegebenen Text in die sogenannte »linguistische Spezifizierung«, wohingegen die Synthese diese Spezifizierung verwendet, um die Waveform zu generieren (King 2011: 837).

Im Prozess der Textanalyse wird dessen Form dekodiert und die Graphemsequenz in eine Phonsequenz umgewandelt, d.h. der Text wird in eine abstrakte Beschreibung des zu erzeugenden Signals transkribiert, die im Syntheseprozess dann in Sprachsignal

konvertiert wird. Die Transkription des Textes enthält Informationen hinsichtlich Wort- und Silbenakzent, Phrasengrenzen o.Ä. (Pfister & Kaufmann 2008: 197); sie muss nicht nur eine richtige phonetische Repräsentation der einzelnen Laute sein, sondern ebenso Silbenstrukturen und suprasegmentale Eigenschaften repräsentieren.

In der Synthese wird dann die phonologische Repräsentation, die im Analyseteil abgeleitet wurde, in ein konkretes Sprachsignal umgewandelt. In diesem Teil des Syntheseprozesses kommen zwei Komponenten des Systems zum Tragen, nämlich die Sprachsignalproduktion und die Prosodiesteuerung. Die Prosodiesteuerung leitet aus der gegebenen phonologischen Darstellung die prosodischen Parameter wie Dauer, Grundfrequenz und Intensität für jeden Laut ab. Zu den prosodischen Parametern zählt vor allem die richtige Akzentuierung auf lexikalischer sowie auf phrasaler Ebene. Während der Wortakzent in der Transkription bzw. im Lexikon, auf das der Synthesizer zugreift, inkludiert ist, muss das System beim Phrasenakzent semantische sowie pragmatische Parameter herbeiziehen, um die richtige Realisierung der gewünschten Prosodie zu erreichen. Da Synthesesysteme nicht in der Lage sind, spezielle Betonungsmuster aus einem gegebenen Text abzuleiten, wäre es diesem Fall erforderlich, diese im Inputtext speziell zu markieren. Aus der syntaktischen Struktur eines Satzes kann allerdings eine neutrale Prosodie abgeleitet werden; so befindet sich in der Regel in der rechten Satzperipherie bspw. das fokusakzenttragende Nomen. Neutrale Prosodie kann jedoch nur für Sätze angenommen werden, die ohne Zusammenhang zu anderen Sätzen stehen und ist daher für Anwendungen wie Vorleseautomaten bzw. Screenreader unbrauchbar (Pfister & Kaufmann 2008: 211).

Für die Produktion des Sprachsignals werden die einzelnen Laute laut linguistischer Spezifizierung in der durch den Text gegebenen Reihenfolge erzeugt und mittels prosodischen Parametern aus der Prosodiesteuerung modifiziert: Es resultiert das synthetische Sprachsignal (Pfister & Kaufmann 2008: 199 f.).

2.2 Forschungslandschaft

Das Feld der Sprachsynthese umfasst durch seine vielseitigen methodischen Ansätze viele unterschiedliche Forschungsbereiche und eine genaue Nachzeichnung der gegenwärtigen Forschung ist an dieser Stelle nur in Bezug auf die für diese Arbeit relevanten Einrichtungen möglich; dennoch sollen die in der Vergangenheit bedeutenden wie repräsentativsten Forschungseinrichtungen kurz beleuchtet werden.

Zahlreiche Workshops und Panels auf internationalen Konferenzen sowie Spezial-

ausgaben von Journals zum Thema Sprachsynthese spiegeln gegenwärtig eine aktive Forschungslandschaft wider. Durch den Aufschwung der Computertechnologie konnte man auch außerhalb der Telekommunikationsforschung seit Mitte der 1980er ein größeres Interesse an der digitalen Sprachverarbeitung wie automatische Spracherkennung oder Sprachsynthese bemerken. Die Entwicklung einzelner Komponenten weicht in den vergangenen Jahren jedoch immer mehr der ganzheitlichen Verbesserung des Computer-Mensch-Dialoges.

Die Nokia Bell Laboratories (ehemals AT&T Bell Laboratories und Bell Labs) in New Jersey, USA leiteten mit Homer Dudleys Forschung zur Sprachkodierung die moderne Sprachsynthese ein. Sie spielten jedoch nicht nur in Bezug auf die Sprachsynthese eine zentrale Rolle in der Vergangenheit, sondern ebenso mit dem ersten Spracherkenner *AUDREY* (Davis et al. 1952) für die automatische Spracherkennung, sowie aktuell für die Arbeit an informationstheoretischen Fragestellungen. Im Laufe des vergangenen Jahrhunderts beschäftigten die Bell Laboratories bedeutende Wissenschaftler wie James L. Flanagan (*1925 - †2015), die mit ihrer Arbeit die Basis für die heutige Spracherkennung sowie -synthese schafften.

Der Ursprung der Sprachsynthese in Europa liegt an der Königlichen Technischen Hochschule (KTH) im KTH Speech, Music and Hearing in Stockholm (KTH THM). Gunnar Fant hat hier mit seiner auch für die Phonetik sehr zentralen akustischen Theorie (Fant 1960) grundlegende Arbeit für die Entwicklung von Sprachsynthesesystemen geleistet. Am KTH THM wurden ebenso die ersten Formantsynthesizer Orator Verbis Electris *OVE* und *OVE II* (Fant 1953; Fant & Mártony 1962) entwickelt. Der aktuelle Forschungsschwerpunkt am KTH THM liegt derzeit in der zwischenmenschlichen Kommunikation, die vor allem in multidisziplinären Forschungsaktivitäten angesiedelt ist.

Ungefähr zum gleichen Zeitpunkt wurde in den späten 1940er Jahren auch an der Universität Yale, in den Haskins Laboratories an verschiedenen Sprachsynthesesystemen gearbeitet. Franklin Cooper entwickelte gemeinsam mit Kollegen das System *Haskins Pattern Playback* (Cooper et al. 1951), das es erlaubte, Breitbandspektogramme in Laute zu konvertieren. Derzeitige Forschung in den Haskins Laboratories haben u.a. Sprachperzeption, Spracherwerb und Mehrsprachigkeit als Schwerpunkt, daher nimmt die Sprachsynthese hier keine große Rolle mehr ein.

Eine Forschungseinrichtung, die nach wie vor von großer Bedeutung in der Sprachsyntheseforschung ist, ist das Language Technology Institute (LTI) der Carnegie Mellon University in Pittsburgh, USA. Dieses Institut gehört zu jenen Einrichtungen, die Mitte

der 1980er als Antwort auf die vielseitigen und neuen Forschungsmöglichkeiten im Bereich der Sprachverarbeitung ins Leben gerufen wurden. Mit Alan Black beheimatet das Institut einen wichtigen Entwickler und Wissenschaftler im Sprachsynthesebereich; Black war am Centre for Speech Technology Research (CSTR) in Edinburgh u.a. verantwortlich für die Entwicklung des vielseitigen Synthesesystems *Festival* (Black & Taylor 1997).

Für erste Entwürfe von kommerziellen Synthesesystemen, aber auch für die Syntheseforschung selbst war Dennis H. Klatt (*1938 - †1988) am Massachusesetts Institute of Technology (MIT) von großer Bedeutung. Klatt war an zahlreichen einschlägigen Publikationen zur Sprachperzeption und Sprachproduktion, aber auch an der Entwicklung des Formantsynthesizers *MITalk* (Allen et al. 1979) und dessen bekannteren Nachfolger, das kommerzielle Synthesesystem *DECTalk* aus 1984 (ursprünglich aus dem Klattalk Text-to-Speech System von Klatt (1982) entstanden) beteiligt. Noch wichtiger für diese Arbeit tritt er aber als Autor der bedeutenden Publikation »Review of text-to-speech conversion for English« (Klatt 1987) in Erscheinung. Klatt vermittelt mit dieser Nachzeichnung der historischen Entwicklungen in der Sprachsyntheseforschung, sowie der ausführlichen Bibliographie einen beispiellosen Überblick über die Thematik der Sprachsynthese.

Bereits in den späten 1980er und frühen 1990er Jahren nahm die Forschung an Sprachsynthesesystemen auch in Japan zu. Das 1989 gegründete Advanced Telecommunications Research Institute in Kyoto (ATR) entwickelte bereits wenige Jahre nach seiner Gründung die Systeme *Nuu-Talk* (Sagisaka et al. 1992) und *CHATR* (Black & Taylor 1994); bei letzterem wirkten ebenso zwei zentrale Figuren der Sprachsynthese, Paul Taylor und Alan Black, mit. Aktuell arbeitet das ATR u.a. an künstlicher Intelligenz und Life Sciences.

Das Synthesesystem *HTS-2007/2008* (Yamagishi et al. 2008), das für Teile dieser Arbeit verwendet wurde, hat seinen Ursprung ebenso in Japan an der Technischen Universität Nagoya, am Institute of Technology im Speech Processing Laboratory (NITech). Die von Keiichi Tokuda entwickelte Sprachsynthesemethode durch Hidden-Markov-Modelle (Tokuda et al. 2013) erwies sich in der Vergangenheit als sehr erfolgreich und fand auch internationale Anerkennung, sowie Anwendung in anderen Synthesesystemen. Junichi Yamagishi, u.a. Entwickler des HTS-2007/2008 Synthesesystems, promovierte an diesem Institut und arbeitet derzeit u.a. im Bereich der Signalverarbeitung und Machine-Learning an statistischer Sprachsynthese an dem bereits genannten CSTR, ein weiteres bedeutendes Institut für die Forschung in diesem Bereich. Der Gegen-

wärtige Forschungsfokus des NITech liegt auf der Verbesserung von Voice-Interaction Systemanwendungen (Smartphones etc.).

Das CSTR wurde 1984 gegründet und kann somit auch in Europa als eine der Geburtsstätten der digitalen Sprachverarbeitung gelten. Aufgrund der zahlreichen internationalen bedeutenden Kollaborationen mit den relevantesten Instituten in der Sprachverarbeitung (wie LTI oder NITech) zählt dieses Institut im Bereich der Sprachsynthese zu den führenden Forschungseinrichtungen in Europa. Aktuell beschäftigt sich das CSTR mit methodischen Aspekten der Sprachsynthese, Spracherkennung, Sprachsignalverarbeitung, multimodalen Interfaces und somit ebenso mit dem gesamten Computer-Mensch-Dialog. Zu den wirkungsreichsten Publikationen dieser Einrichtung zählen Taylor (2009), Wu et al. (2016), die Synthesesoftwares Festival (Black & Taylor 1997) und HTS (Yamagishi et al. 2008). Die Zusammenarbeit des CSTR mit Pucher et al. am Projekt SALB (Pucher et al. 2017b) zeigt, wie konstruktiv eine internationale Zusammenarbeit auch für die österreichische Forschungslandschaft sein kann.

Das Institut für Phonetik der Universität des Saarlandes brachte bereits ab Mitte der 1980er Jahre vermehrt Forschungsarbeiten in Richtung Sprachverarbeitung hervor, beinahe ein Jahrzehnt vor der Gründung des Instituts für Computerlinguistik an der Universität des Saarlandes. Hier befinden sich mit Jürgen Trouvain und Bernd Möbius zwei Linguisten unter den treibenden Kräften, die einschlägige und vor allem interdisziplinäre Arbeiten rund um Sprachtechnologien im deutschsprachigen Raum veröffentlichten (wie Trouvain (2004) oder Möbius (2001)). Hinsichtlich der Sprachsynthese kann man am Institut für Computerlinguistik und Phonetik der Universität des Saarlandes auf eine lange Tradition interdisziplinärer Forschung im internationalen Rahmen blicken. Das von Schröder 2003 entwickelte Open-Source-Synthesesystem *Mary TTS* wird derzeit u.a. von Ingmar Steiner für Varietäten des Deutschen weiterentwickelt/angepasst (wie MaryLux (Steiner et al. 2015), gleichzeitig forscht Steiner in der von ihm geleiteten Multimodal Speech Processing Group u.a. mit dem CSTR an weiteren interdisziplinären Projekten.

Für die Sprachsyntheseforschung in Österreich nahm das außeruniversitäre Forschungszentrum Telekommunikation Wien (FTW) eine tragende Rolle ein. Das mit Ende 2015 aufgelassene Institut hinterlässt zahlreiche erfolgreich abgeschlossene Forschungsprojekte, in denen interdisziplinäre Vorhaben realisiert wurden, die den Fortschritt in der Wissenschaft, sowie die Verbesserung von Informations- und Kommunikationstechnik zum Ziel hatten. Zu den für diese Arbeit relevanten Beiträgen, die am FTW entstanden sind, zählt zweifellos Schabus (2009): Diese Diplomarbeit stellte sich als

unverzichtbare Grundlage für das Verständnis des Synthesesystems HTS-2007/2008 heraus, in Schabus (2009) konnte auch gezeigt werden, dass es mit der parametrischen Analyse-Resynthese-Technik einen flexibleren Ansatz zur (bis zu diesem Zeitpunkt weit verbreitete) Verkettungsmethode gibt, die weder Sprecher-Adaptierungen oder Sprecher- bzw. Varietäteninterpolation vorsieht.

Das FTW zeichnete sich auch durch immer wieder vorkommende Kollaborationen mit den bereits genannten führenden Einrichtungen CSTR oder NITech als wichtige Schnittstelle der österreichischen Forschung mit international bedeutenden Einrichtungen aus.

Weitere Forschungsbeiträge wie Pucher et al. (2010a), Pucher et al. (2012), Pucher et al. (2010b), Toman (2016), Toman & Pucher (2013) und Toman et al. (2015) und Schabus (2009), konnten nicht nur zeigen, dass Synthese des österreichischen Standarddeutsch (SAG), sowie für österreichische Varietäten ein breites Forschungsfeld öffnen konnte, sondern auch, wie wichtig lokale Forschung an Sprachverarbeitungstechnologien im sozialen Inklusionsprozess von Menschen mit Behinderung wie Sehbehinderte oder Blinde sein kann.

Neben einschlägigen Publikationen des FTW, konnten auch andere österreichische Forschungsarbeiten aus dem akademischen Bereich der Linguistik wie Schrank (2014) erst kürzlich zeigen, wie fruchtbar sprachtechnologische Verfahren und linguistische Kenntnisse zusammen wirken können.

Die o.g. österreichischen Forschungsarbeiten zum Thema Sprachsynthese sind stark von einer informatischen Perspektive geprägt. Obwohl in Pucher et al. (2012) und Pucher et al. (2010b) und Toman et al. (2015) mit Moosmüller, eine der wichtigsten Phonetikerinnen im deutschsprachigen Raum, ein linguistischer Hintergrund mit einbezogen wurde, kann man das Interesse der Linguistik in diesem Bereich nicht am internationalen Standard messen. Mit dieser Arbeit soll ein Blick aus der Linguistik auf die Sprachsynthese geworfen werden und auch die Notwendigkeit unterstreichen, gerade in klassischen Brückenfächern wie der Phonetik, interdisziplinär zu arbeiten, und auch die zukünftigen Schnittstellen dieser Disziplinen in beiden Bereichen hervorzuheben und bereits erfolgreiche Zusammenarbeiten in Erinnerung zu rufen.

2.2.1 Historischer Überblick der wichtigsten Sprachsynthesesysteme

Folgend wird ein historischer Abriss dargelegt, der die einflussreichsten Systeme kurz beleuchten und erlauben soll, die maßgeblichen Umstände bei der Entwicklung von

Sprachsynthesizern nachvollziehen zu können. Vor diesem Hintergrund soll auch hervorgehen, wie die Idee einer Maschine, die mit einer menschlich klingenden Stimme spricht, im vergangenen Jahrhundert ein ganzes Forschungsfeld entstehen ließ, welches bis heute durch eine wachsende, interdisziplinäre Forschungsgemeinschaft geprägt ist. Die Geschichte der Synthesesysteme spiegelt wider, wie neue Erkenntnisse aus unterschiedlichen wissenschaftlichen und technischen Disziplinen Paradigmenwechsel auslösen und Fragestellungen verschieben können.

Gegen Ende des 18. Jahrhunderts, im Jahr 1766, gab es die ersten relevanten Bemühungen, Sprache mithilfe eines mechanischen Apparats, der den Vokaltrakt nachbilden sollte, künstlich zu erzeugen. Die sogenannte Sprechmaschine von Wolfgang von Kempelen, der zuvor durch die Erfindung eines vorgeblichen Schachautomaten (»Schachtürke«) bekannt wurde, konnte nach über zehn Jahren Entwicklungszeit 1778 den ersten ernstzunehmenden Versuch einer Sprachsynthese vorstellen. Von Kempelen stützte die Konstruktion der Sprechmaschine auf in seinem über 500 Seiten umfassenden Buch »Mechanismus der menschlichen Sprache« (Kempelen et al. 2017) dargelegten Experimente, die u.a. beweisen sollten, dass sich die Lauterzeugung nicht wie bis dahin angenommen in der Larynx, sondern vielmehr im Vokaltrakt ereignet. Durch intensive Analysen der Artikulationsprozesse und die Durchführung sowie Dokumentation von Experimenten kann diesem Werk auch wissenschaftsgeschichtlich eine große Bedeutung zugeschrieben werden. Von Kempelen entschied sich dazu, seine Sprechmaschine aus simplifizierten Modellen von Sprechorganen zu konstruieren: Die Lunge wurde durch einen Blasebalg simuliert, die Glottis wurde aus einem Rohrblatt aus Elfenbein gefertigt und die Nase sowie der Mund wurde durch einen Gummitrichter imitiert. Durch die händische Manipulation des Trichters war es von Kempelen durchaus möglich, respektable Ergebnisse zu erzielen und auch verschiedene Vokalqualitäten zu erzeugen (Dudley & Tarnoczy 1950). Obwohl von Kempelen sich durch seine weit bekanntere Erfindung, den weit bekannteren (jedoch menschlich betriebenen und damit nur vortäuschenden) Schachtürken, durch Betrug selbst diskreditierte, waren seine Anstrengungen auf dem Gebiet der Synthese von menschlichen Sprachlauten nicht ohne Wirkung auf die Forschungsgemeinschaft (Schroeder 1993: 232). Mithilfe der Sprechmaschine wurden neue Forschungswege aufgezeigt, denen eine intensive Auseinandersetzung mit der Physiologie von Sprachproduktion und experimenteller Phonetik folgte. Hier wird ersichtlich, dass es zu Beginn der Forschung im Bereich der Sprachsynthese maßgeblich war, primär Erkenntnisse über die Lautproduktion

und akustischer Merkmale zu sammeln, um später sprachähnliche Laute generieren zu können.

160 Jahre nach von Kempelens Sprechmaschine stieg mit den technischen Möglichkeiten das wissenschaftliche Interesse an der Sprachsynthese wieder. In der ersten Hälfte des 20. Jahrhunderts konnte durch analog-elektrische Ansätze und später durch computerbasierte Methoden die Idee einer funktionierenden Sprachsynthesemaschine realisiert werden. Mit der Erfindung des Telefons war man auch immer stärker daran interessiert, die Daten von Sprachübertragung ohne Qualitätsverlust zu komprimieren. Diese Motivation führte zur Erfindung des ersten Synthesizers *Voder* (»Voice Operation Demonstrator«) durch Dudley et al. (1939), Mitarbeiter der Bell Laboratories. 1939 von Dudley auf der Weltausstellung in New York präsentiert, können mit ihm auch die Anfänge der parametrischen Synthese markiert werden. Das manuell bedienbare Synthesesystem beschrieb Sprachlaute anhand ihrer akustischen Parameter und erzeugte sie anschließend synthetisch. Die Grundlage des Voders bildete ein Kanalvocoder, der Sprache durch zehn Spannungsverteiler erzeugte. Ein Quellsignal, das durch zehn Resonatoren geführt wurde, konnte durch manuell bedienbare Spannungsverteiler kontrolliert werden und dadurch die Effekte des menschlichen Vokaltrakts imitieren und somit die Synthese von Sprachlauten ermöglichen. Obwohl die Qualität der Synthese hinsichtlich Verständlichkeit und Effizienz dürftig war, konnte man hinter dem Voder Potenzial erkennen (Klatt 1987: 741). Nur einige Jahre nach dem Voder wurde in den Haskins Laboratories der Yale Universität ein weiteres, für die Geschichte und Fortschritt der Sprachsynthese wichtiges System entwickelt: der Pattern Playback Synthesizer. Dieses System erlaubte es, mithilfe eines speziellen Scanners das Muster auf einem Breitband-Spektrogramm wieder in ein elektrisches Signal und somit Sprachlaute umzuwandeln (Cooper et al. 1951). Daraus folgten nicht nur Erkenntnisse über die Erzeugung von künstlicher Sprache, sondern genauso wichtige phonetische und akustische Einsichten, welche wiederum für die Entwicklung folgender Sprachsynthesesysteme maßgeblich waren.

Um nachfolgend die für die Forschungsgeschichte der Sprachsynthese relevanten Systeme besser verstehen bzw. verorten zu können, scheint an dieser Stelle die Unterscheidung zwischen Systemen, die auf Analyse/Resynthese basieren, und regelbasierten Systemen hilfreich zu sein. Analyse/Resynthese Systeme parametrisieren in der Regel eine gegebene Äußerung, um zu einer daten-reduzierten Version dieser Äußerung zu gelangen, während die Verständlichkeit nur gering gemindert wird. Im Gegensatz zu Analyse/Resynthese Systemen werden von regelbasierten Systemen keine bereits

bestehenden Laute benötigt. Sie sind so konstruiert, dass sie neue Laute von selbst generieren können, was sich allerdings negativ auf die Verständlichkeit auswirkt.

Ab Mitte des 20. Jahrhunderts gab man die analog-elektronischen Ansätze für Sprachsynthese zugunsten der sich neu darbietenden digitalen Methoden auf. Man arbeitete an den ersten computer-basierten Synthesemethoden, die man in zwei Gruppen unterteilen kann: System-Modelle, welche versuchen, das ganze Sprachproduktionssystem des Menschen zu modellieren, und Signal-Modelle, welche versuchen, nur das Sprachsignal selbst zu modellieren, wozu auch die bereits erwähnten regelbasierten Synthesesysteme zählen.

Gunnar Fant, schwedischer Elektroingenieur und Akustiker, machte mit der »akustischen Theorie zur Spracherzeugung« (Fant 1960) deutlich, in welcher spannungsreichen Wechselbeziehung Phonetik und Sprachverarbeitung bereits in der frühen Phase der Forschung zu Sprachsynthese standen. Fant legte damit auch eine theoretische Basis für die Formantsynthese, für die ein bereits existierendes Sprachsignal nicht mehr notwendig war. Neben seiner Forschung zur Perzeption und Produktion von Sprachlauten entwickelte Fant mit OVE I auch den ersten europäischen (Formant-)Synthesizer. Obwohl dessen Produktion auf Vokale beschränkt blieb, leitete er damit eine neue Ära in der Forschung zu Sprachsynthese ein. Formantsynthesizer zählen zu den regelbasierten Systemen und galten für lange Zeit als die erfolgreichste Methode um Sprache künstlich herzustellen. Für die Erzeugung des akustischen Signals verwenden Formantsynthesizer ein Quellsignal, das durch eine Anzahl von vokaltraktähnlichen variablen Resonanzfilter (Formanten) geführt wird (Donovan 1996: 6).

Die ersten Formantsynthese-Systeme verwendeten zwei Methoden zur Herstellung des Sprachsignals: In der Parallel-Formantsynthese wurde das Quellsignal auf alle Formantfilter angewendet und ihre Outputs wurden summiert, was den Vorteil der Spezifizierung jedes einzelnen Formanten mit sich bringt. Kaskaden-Systeme hingegen wendeten den Output eines ersten Formanten auf den darauf folgenden an (Donovan 1996: 7). *OVE* konkurrierte als Kaskadensystem mit dem Parallel-Formantsynthesizer *PAT* (»Parametric Artificial Talker«) von Lawrence (1953). Welche der zwei Methoden die bessere ist, war lange Gegenstand von Diskussionen. Beide Methoden weisen Stärken sowie Schwächen auf; so ist etwa die Kaskadenstruktur zwar besser für nichtnasale und stimmhafte Laute, die Parallelstruktur eignet sich jedoch besser für die Synthese von Nasalen, Frikativen und Plosiven. In den frühen 1980ern schließlich hatte Dennis H. Klatt mit der Verbindung beider Strukturen in einem System großen Erfolg. Sein DECTalk-System, das aus der Kommerzialisierung von MITalk von 1979 resultierte,

war bis Mitte der 2000er ein häufig verwendetes System, das durch seine Flexibilität hinsichtlich Lexika und Stimmen in unterschiedlichsten Kontexten wie etwa Wetter- oder Radiostationen, Anwendung fand.

In regelbasierten Systemen kann man weiters zwischen Waveform-Konkatenation und parametrischen Synthesesystemen unterscheiden. In den 1970er Jahren gewann die Synthese mit Waveform-Konkatenation bzw. später mit Unit-Selection immer mehr an Bedeutung. Konkatenative Systeme sind aufgrund der fixen Segmente, auf die sie zurückgreifen müssen, sehr unflexibel; ebenso leidet der Output häufig unter sogenannten Konkatenationsartefakten, da Transitionen zwischen den einzelnen Lautsegmenten schwer zu berechnen sind. Um diesem Problem entgegenzuwirken, greift man in der konkatenativen Synthese auf Diphone zurück, die sich von der Mitte eines Segments bis zur Mitte des folgenden Segments ziehen. Dennoch leidet die konkatenative Synthese zumeist unter einem unnatürlich klingenden Output, da die prosodischen Eigenschaften hier nicht auf Phrasenebene modelliert werden. Bis Ende der 1990er galt die Unit-Selection als der State-of-the-Art Ansatz in der Sprachsynthese, wichtige Systeme aus dieser Zeit waren das von Yoshinori Sagiasaka entwickelte Nuu-Talk-System (Sagisaka et al. 1992), CHATR (Hunt & Black 1996) oder auch das von Alan Black am CSTR entwickelte Festival-System (Black & Taylor 1997), das bis heute Anwendung findet.

2.1 zeigt einen Überblick über die oben besprochenen Synthesesysteme.

Im deutschsprachigen Raum wurde Anfang der 2000er Jahre das *MARY*-System entwickelt; ein Text-to-Speech-System, das als Unit-Selection sowie Hidden-Markov-Modell-System verfügbar ist. In der Mitte der 1990er begann man bereits statistisch basierte Ansätze für die Synthese zu diskutieren, wie etwa Robert Donovan in seiner Dissertation (Donovan 1996) oder auch Keiichi Tokuda, welcher 1995 auf der International Conference on Acoustics, Speech, and Signal Processing einen der ersten Beiträge zur statistisch basierten Synthese publizierte (Tokuda et al. 1995). Bis zur Veröffentlichung des ersten auf Hidden-Markov-Modellen (HMM) basierten Systems dauerte es jedoch noch weitere sieben Jahre; erst im Dezember 2002 wurde das erste HMM-Sprachsynthesesystem von NITech veröffentlicht (Tokuda et al. 2002).

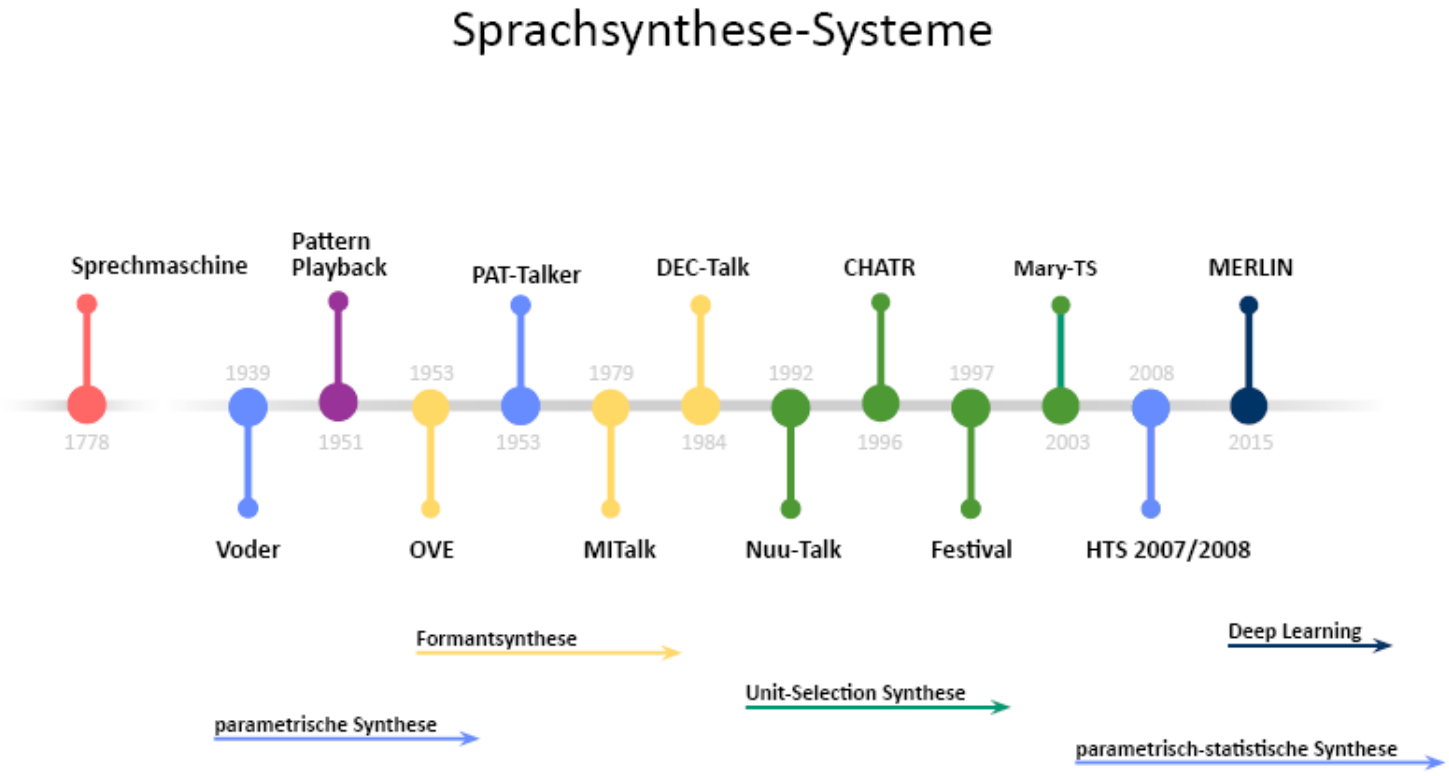

Abbildung 2.1: Überblick über die wichtigsten Sprachsynthesesysteme

2.2.2 Hidden-Markov-Modelle

Ein Hidden-Markov-Modell begreift sich als ein Sequenzmodell, das jede Einheit einer gegebenen Sequenz (Markov-Kette) labelt oder klassifiziert. Diese Sequenz wird durch das System mit unbeobachteten (»hidden«) Zuständen modelliert.

Im Fall der Sprachverarbeitung können diese Einheiten aus ganzen Sätzen, Wörtern, Morphemen oder Phonen bestehen. Weiters stellt ein HMM nicht nur ein reguläres Sequenzmodell dar, sondern ein probabilistisches, was bedeutet, dass es eine Wahrscheinlichkeitsverteilung über die möglichen Labelsequenzen berechnet und den wahrscheinlichsten Kandidaten auswählt.

Bevor ein HMM-System Sprache künstlich erzeugen kann, muss es mit einem umfangreichen akustischen Korpus trainiert werden, welches auf Lautebene segmentiert sowie transkribiert sein muss um einer akustischen Analyse unterzogen werden zu können. Durch das Training leitet das System die statistischen Modelle einzelnen Laute des gesamten Korpus ab, die es später für die Erzeugung des Sprachsignals benötigt. Schließlich kann das System durch einen gegebenen Text aus den abgeleiteten Modellen die zugehörigen Parameter generieren und somit die gewünschte bzw. die wahrscheinlichste Lautkette synthetisieren. Im Fall der HMM-Synthese ist die erwähnte Textanalyse, die eine Graphemsequenz verarbeitet und für die Synthese parametrisiert, um eine Komponente, die des Trainings erweitert. Im Training wird von einem akustischen Sprachkorpus bzw. Trainingsdaten durch Indexierung der parametrisierten akustischen Sprachdaten die linguistische Spezifizierung abgeleitet, mit der es dann möglich ist, das Sprachsignal zu erzeugen (King 2011: 838). Das heißt, ein HMM-System muss das akustische Korpus, mit dem es trainiert wurde, nicht speichern; es speichert allein das akustische Modell, das es durch die linguistische Spezifizierung abgeleitet hat. Die linguistische Spezifizierung, die ein HMM-System als Input erhält, kann aus einer simplen Phonsequenz bestehen, was allerdings in geringer Qualität hinsichtlich der Natürlichkeit des Outputs resultieren würde; daher inkludiert man ebenso suprasegmentale Informationen, wie die Prosodie des gewünschten Sprachsignals. Im Wesentlichen inkludiert die linguistische Spezifizierung jedoch Faktoren, die einen Einfluss auf die akustische Realisierung der zu erzeugenden Äußerung ausüben, wie etwa der phonetische Kontext eines Phons. 2.2 fasst ein HMM-Synthesesystem schematisch zusammen.

Das Synthesesystem, mit dem der Dialektoutput für diese Arbeit generiert wurde, ist eine Erweiterung des 2010 veröffentlichten Semi-Hidden-Markov-Modell (HSMM)-

Systems des EMIME-Projekts (Yamagishi & Watts 2010). Auch dieses System gliedert sich in die bereits erwähnten zwei Komponenten: Training und Synthese. Im Trainingsteil werden die akustischen Parameter Grundfrequenz, Spektrum und Dauer der Sprachdaten bzw. des Trainingskorpus extrahiert, sowie die linguistische Spezifizierung, in 2.2 als »Label« bezeichnet, abgeleitet. Mit den extrahierten akustischen Parametern und der linguistischen Spezifizierung werden anschließend HMMs trainiert. Im Syntheseteil wird der Textinput verarbeitet und jede Graphemsequenz wird in eine Phonsequenz, die mithilfe der linguistischen Spezifizierung aus dem Trainingsteil abgeleitet wurde, transformiert. Die wahrscheinlichste Phonsequenz wird errechnet und durch die Aneinanderkettung der akustischen Modelle der einzelnen Phone durch das HMM die akustischen Parameter generiert. Darauf folgt die Generierung des Anregungssignals, welches durch verschiedene Synthesefilter geführt wird und dann schließlich das akustische synthetische Sprachsignal ergeben.

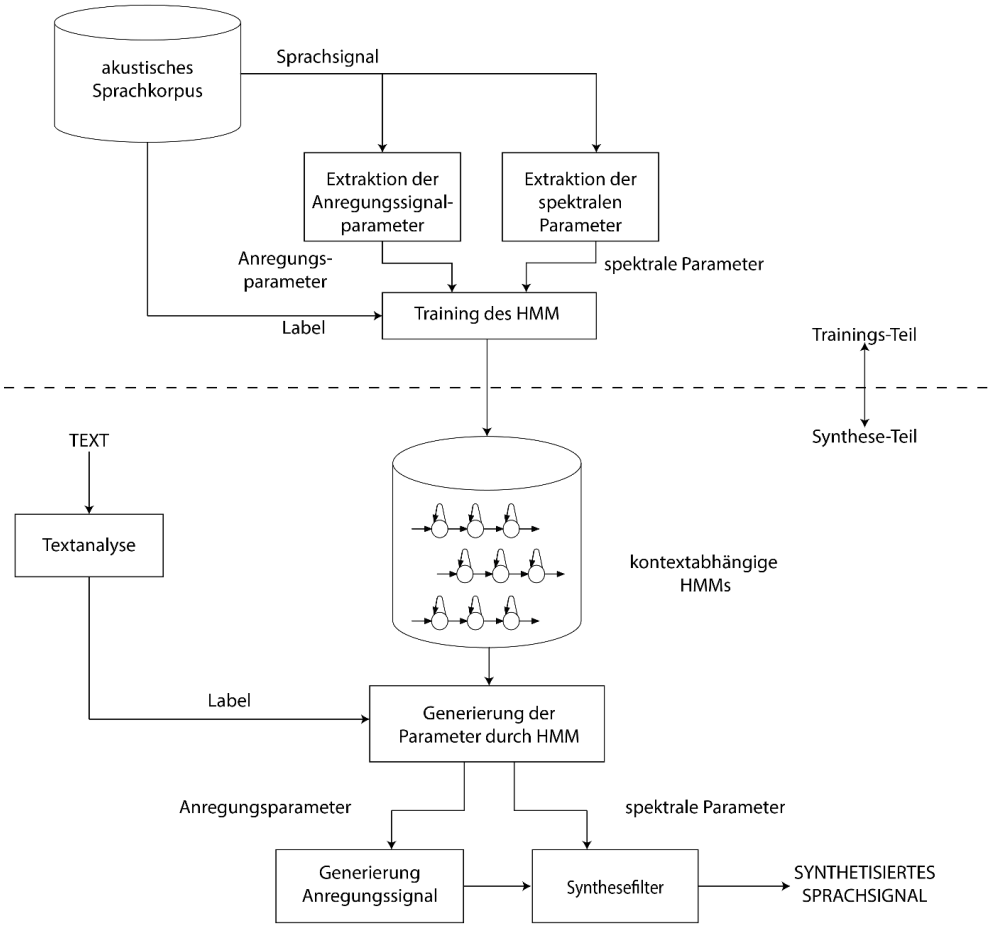

Abbildung 2.2: Struktur eines HMM-Synthesesystems

Zusammenfassung

Retrospektiv betrachtet werden die meisten Methoden, die zwischenzeitlich von Erfolg geprägt sind, in kürzester Zeit überholt. So auch bei der erfolgreichen HMM-Synthese, die seit 2008 als State-of-the-Art-Ansatz gelten konnte. Bereits 10 Jahre später verweisen Blicke in die Zukunft auf einen erneuten Paradigmenwechsel, bei dem die HMM-Synthese nach und nach von der vielversprechenden Deep-Neural-Networks-Synthese (DNN) abgelöst wird. Diese Systeme waren bereits in der automatischen Spracherkennung von Erfolg geprägt und können in der gesamten Sprachverarbeitung als neues Schlüsselmodell für die Zukunft gelten.

Deep-Learning-Ansätze wie DNNs haben jedoch über die Sprachverarbeitung hinaus ein breites Anwendungsfeld. DNNs begreifen sich als selbstlernende Systeme, die durch einen umfangreichen Input, seien es visuelle, akustische oder andere digitale Daten, mithilfe von künstlichen neuronalen Netzwerken Merkmalsrepräsentationen erkennen und darauf basierend Output generieren können. DNN-Synthese-Systeme, wie das 2015 veröffentlichte Open-Source-Toolkit Merlin, verwenden künstliche neuronale Netzwerke, um akustische Merkmale eines akustischen Inputs vorhersagen zu können (Wu et al. 2016), welche dann an einen Vocoder weitergeleitet, der das Sprachsignal erzeugt.

Die HMM-Synthese sowie die übrigen bisher besprochenen Synthesemethoden dürfen hier jedoch nicht als redundant gelten; einerseits konnten sie den Weg für DNN-Methoden ebnen, andererseits kann daran gezeigt werden, wie sich Wissen einer fast 100-jährigen Forschungsgemeinschaft akkumuliert und aktuell in der DNN-Synthese gipfelt. Wie in diesem Abschnitt erkenntlich wurde, verschob sich die grundlegende Fragestellung zu Beginn von simpler Lautgenerierung (von Kempelen, Dudley), schrittweise darauf, wie man ein ganzheitliches System entwickeln kann, dass nicht nur natürlich klingt, sondern vor allem auch effizient und selbstlernend ist.

Die sich stetig erweiternden technischen Möglichkeiten werden von einem Verjüngungsprozess begleitet, dem sich die Forschungsgemeinschaft in Hoffnung auf das Annähern oder das Erfüllen des jahrhundertealten Wunsches nach einer menschlichen klingenden Stimme bedingungslos hingibt.

2.2.3 Dialekt- bzw. Varietätensynthese

Das Interesse an einem Dialektsynthesizer wächst nicht nur in der Sprachtechnologie-Gemeinschaft; mit den wachsenden Anwendungsbereichen von verschiedenen Arten

an Robotern zeigen sich auch privatwirtschaftliche oder öffentliche Sektoren interessiert an der Synthese von Sprachvarietäten. Besonders für Länder mit einer plurizentrischen Sprache (Clyne 1991), wie Österreich, stellt der Fortschritt im Bereich der Dialektsynthese ein besonderes Interesse dar, da hier die SprecherInnen in den Varietäten Dialekt und Standard kompetent sind und lebensweltlich beide regelmäßig verwendet werden (Moosmüller 1995).

Arbeiten von Ge et al. (2012) zeigen, dass die Synthese potentielle Anwendungsgebiete außerhalb der Telekommunikation hat, nämlich im wachsenden Feld der sozialen Robotik, vor allem im Gesundheitsbereich große Bedeutung vorhergesagt wird. Hier sollen Roboter – angefangen vom Spielzeug und Unterhaltungsgegenstand bis hin zum Pflegeassistenten – einfache Aufgaben erfüllen, aber gleichzeitig auch glaubhafte Interaktionspartner für den Menschen darstellen (Goodrich & Schultz 2007). Anwendungen, die die soziale Robotik leisten möchte, benötigt eine klare Verkörperung bzw. Vermenschlichung des sozialen Agenten, die sich auch durch die Stimme erschließen soll. Krenn et al. (2012) und Krenn et al. (2014) zeigten, dass HörerInnen vor allem die Systemstimme als Basis für die Beurteilung eines solchen sozialen Agenten heranziehen. Die künstliche Stimme ist also maßgeblich daran beteiligt, die Persona eines sozialen Agenten zu übermitteln. Das Konzept Persona kann als standardisiertes, mentales Abbild einer Persönlichkeit oder Charakters definiert werden, welches UserInnen von einer Systemstimme ableiten (Cohen et al. 2004). Verschiedene Anwendungen erfordern unterschiedliche Personas, wie bereits Pucher et al. (2018) zeigen konnte. Der Dialekt- bzw. Varietätensynthese wird hier in Zukunft eine besondere Rolle zugeschrieben, da diese bei Verwendung im Bereich der sozialen Robotik bei HörerInnen positivere Gefühle gegenüber dem Roboter auslösen, als bei Verwendung von Standardsynthese (Tamagawa et al. 2011).

Arbeiten wie Pucher et al. (2010a) und Pucher et al. (2010b) und Toman et al. (2015) ermöglichten die Synthese des Wienerischen und zweier weiterer bairischen Varietäten, der Dialekte von Innervillgraten und Bad Goisern. Pucher et al. (2018) konnte zeigen, dass eine Anwendung dieser Synthesesysteme auch in der Öffentlichkeit möglich ist. Langa et al. (2012) stellten ein Synthesesystem vor, das den nördlichen Sotho Dialekt mit einem limitierten Vokabular erzeugen konnte, auch Navas et al. (2014) beschäftigten sich mit einem Ansatz, der die Synthese des navarro-lapurdischen Dialekts des Baskischen vorsieht. Arbeiten wie diese zeigen, dass die Synthese neue Möglichkeiten für Minderheitensprachen bzw. nicht digitalisierte Sprachen eröffnen, sowie auch zur Sprachemanzipation beitragen kann. Die Dokumentation und das Sichtbarmachen,

welche mit der Synthese von (bedrohten) Sprachvarietäten einhergehen, stellt also auch ein wichtiges Desiderat der Sprachwissenschaft dar.

3 Der (süd-)bairische Dialektraum und seine Phonetik

Der südbairische Dialekt gliedert sich als einer der drei Großräume neben Nord- und Mittelbairisch in das bairische Dialektkontinuum ein. Der bairische Dialektraum gilt als der größte im deutschsprachigen Raum, dessen Grenzen sich von Ostösterreich in den Westen über ganz Österreich bis zum Arlberg, der Grenze zwischen Tirol und Vorarlberg und in den Norden bis Markneukirchen im südlichen Vogtland und in den Süden bis zur Salurnklamm, zwischen Bozen und Trient in Südtirol gelegen, erstrecken. Der südbairische Sprachraum ist auf seiner westlichen Seite durch den Arlberg begrenzt, erstreckt sich dann östlich über das Tiroler Oberland, südlich bis nach Südtirol und schließt weiters Osttirol ein.[1] Weiter nach Osten umfasst das Südbairische ganz Kärnten und reicht schließlich bis in die Weststeiermark hinein. Hier wird es nördlich vom oberen Murtal bei Knittelfeld begrenzt und führt in den Süden zur Staatsgrenze westlich an Graz vorbei. Die südbairische Varietät ist im Allgemeinen sehr von Heterogenität geprägt; es befinden sich einige Übergangszonen zu anderen Dialekten im Sprachraum wie bspw. die westlichste Region des Südbairischen bereits ein Übergangsgebiet zum Alemannischen darstellt.

3.1 zeigt eine Darstellung dieser Grenzen innerhalb Österreichs (nach Wiesinger (1990: 448)).

Eine genaue Beschreibung des Bairischen bzw. des Südbairischen soll an dieser Stelle nicht das Ziel des Abschnittes sein; mehr soll Nachstehendes einen Überblick über die prägnantesten phonologischen wie phonetischen Merkmale beschreiben, die den südbairische Raum unabhängig seiner Geographie kennzeichnen, um die Zusammenhänge der beiden Dialekte IVG und STY hervorzuheben bzw. zu belegen.

1 Das Tiroler Unterland ist kein Teil des Südbairischen, sondern gliedert sich in das Übergangsgebiet Südmittelbairisch, da es bereits wesentliche Charakteristika des Mittelbairischen aufweist.

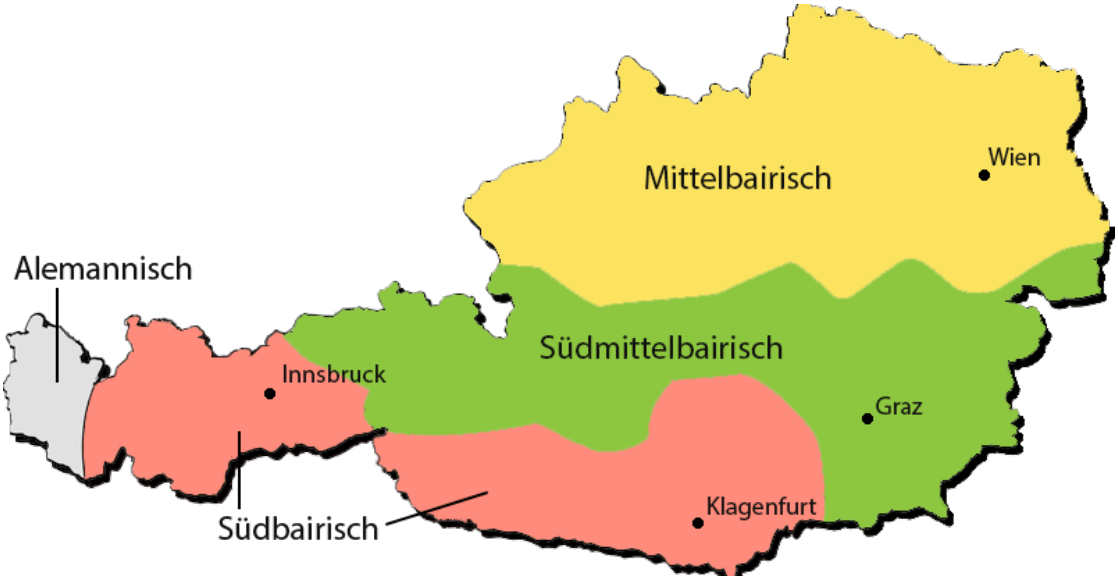

Abbildung 3.1: Die bairischen Dialekträume in Österreich

3.1 Allgemeine phonologische Merkmale des Bairischen

Folgende phonologische Merkmale sollen weniger als alleinige Charakteristika des Bairischen betrachtet werden, da diese teilweise auch in angrenzenden Dialekten belegt sind; mehr sollen diese Merkmale jedoch ein typisches Bild des Bairischen skizzieren (Wiesinger 1990: 452).

Die Entrundung der gerundeten Vorderzungenvokale [y] und [ø] ist im Bairischen bezeichnend, woraus Formen wie [ʃisl̩], *Schüssel* oder [meːg̊n̩], *mögen* entstehen. Der Diphthong <au>, sofern er auf dem MHD <ou> basiert, wird vor Labialen zum gelängten Vorderzungenvokal [aː] wie in [b̥aːm], *Baum*; [dʀaːm], *Traum* oder [kʰafɐ], *kaufen*.

Vor Velaren oder in wortfinaler Position bleibt der Diphthong jedoch erhalten wie in [aʊ̯g̊], *Auge* oder [fʀaʊ̯], *Frau*. Die Hebung und die Rückverlagerung des standardsprachlichen [a], MHD kurz und gelängtes <a> zu einem offenen Hinterzungenvokal [ɔ] tritt nur in nativen Wörtern, wie auch in alten Lehnwörtern und Eigennamen in Erscheinung: [b̥lɔtn̩], *Platte*; [ʃb̥ɐgɔːd̥], *Spagat*; [kɔɐ̯l], *Karl*. Zu [ɒ] wird es in Wörtern wie [mɒntl̩], *Mantel*; [ɒmpl̩], *Ampel*; [fʀɒn͡ts], *Franz*. Im Gegensatz dazu hat man bei jüngeren

Lehnwörtern, sowie Eigennamen ein palatales [a] bzw. [aː] wie in [kasɐ], *Kassa*; [klas], *Klasse*; [d̥aksɪ], *Taxi* oder [sandʀɐ], *Sandra*. Im Osten wie auch im Süden Österreichs wird für kurzes [a] der offene Hinterzungenvokal [ɔ] verwendet, was im angenäherten Standard als auch im Standard zum palatalen [a] führt. In Österreich ist der Unterschied zwischen [ɑ] und [a] weniger markiert als der Unterschied zwischen [ɒ] und [a] in Bayern; trotzdem kontrastieren sie sich in Wörtern wie [bɑŋk], *Sitzbank* und [baŋk], *Geldinstitut* oder [bɑl], *Spielball* und [bal], *Tanzveranstaltung* (Wiesinger 1990: 452-454).

Die Tilgung von <e> in unbetonten Silben führt zu Wörtern bzw. Wortformen ohne Endung im Bairischen. Maskuline und neutrale Nomen im Singular sind von diesem Prozess besonders betroffen, wie die Formen [ʔɔf], *Affe* oder [hɔːz], *Hase* zeigen. Vereinzelt gibt es allerdings auch Pluralformen mit diesem Merkmal, wie [fiɐ̯s], *Füße* oder [ʃʀɪt], *Schritte* zeigen. Die Tilgung des [ə] im <-en>-Suffix führt zu einer Assimilation des Nasals [n] an vorstehende labiale und velare Plosive [b], [p] und [g], [k] zu silbischen [m̩] und [ŋ̩]: [gɛː(b)m̩] , *geben*; [mɛː(g)ŋ̩], *mögen* und [ʀʊkŋ̩], *Rücken*. [n] nach den Dentalen [d] und [t] wird ebenso silbisch wie folgende Formen zeigen: [b̥ɪtn̩], *bitten* und [ʀeː(d)n̩], *reden*. Die Spirantisierung des standardsprachlichen <b> in intervokalischer Position und vor dem Lateral [l] ist ebenso ein kennzeichnendes Merkmal des Bairischen: [kɪːvə], Kübel oder [d̥ʀaɐ̯vɐ], Treiber. Weiters wird nach standardsprachlichem <r> der alveolare Frikativ [s] zu postalveolaren [ʒ] und [ʃ] wie in [fɛɐ̯ʒn̩], Ferse oder [t͡ʃɛɐ̯ʃt], zuerst; genauso wie mediales standardsprachliches <sp> als [ʃp] ausgesprochen wird: [ʀɔʃpɐ], Raspel oder [ʀɛʃpɛkt], Respekt. Zwischen dem Nasal [n] und dem Lateral [l] in derselben Silbe wird häufig der dentale Plosiv [d] eingeschoben, um mit dem damit erreichten Aufbruch der Silbenstruktur die Aussprache zu erleichtern. Daraus entstehen Formen wie [kɒndl̩], Kännchen oder [maɐ̯ndl̩], Meinl (Eigenname) (Wiesinger 1990: 454).

Das Diminutiv kann im Süd- sowie Mittelbairischen mit den Suffixen <-l> und <-le> gebildet werden. Der Plural dieser Formen unterscheidet sich innerhalb des Sprachraums: während in Tirol die Pluralformen des Diminutivs auf -[lən] enden, haben sie in Kärnten -[lɐn] im Auslaut. Die Weststeiermark und das obere Murtal stellen in diesem Fall ein Übergangsgebiet dar, das neben beiden Formen Wiesinger (1990: 475), auch Formen mit silbischen Lateral im Auslaut gebraucht: [dɪʀndlɐn : dɪʀndlən : dɪʀndl̩n], Dirndeln. Wortfinales <r> des Suffixes <-er> bleibt in einem Großteil des Südbairischen erhalten, teilweise sogar als silbentragendes Segment wie die Beispiele [fɔːtəʀ, fɔːtʀ̩], Vater oder [k͡xɪndəʀ, k͡xɪndʀ̩], Kinder zeigen. In den sprachlich konservativen Regionen, zu denen auch Osttirol gezählt wird, ist die Realisierung abhängig von ihrer historischen

Herkunft: Wenn die Endung von MHD <-er> vom AHD <-er, -ar, ir, êr> abgeleitet ist, dann folgt die Form -[ɐʀ] wie in [foːtɐʀ], Vater (AHD *vater*) oder [vɔsɐʀ], Wasser (AHD *wazzar*). Wenn die Endung <-er> allerdings auf das zweisilbige AHD <-iro/-ôro> der Komparativform und dem MHD <-œre> des Nomina Agentis zurückgeht, dann wird die Endung -[ɐʀ]: [gɾɛ̝ɐ̯sɐʀ], größer (AHD *grôzziro*) oder [ʒnaɐ̯dɐʀ], Schneider (MHD *snîdœre*) oder [liɐ̯vɐʀ], lieber (AHD *liobôro*). In einigen Teilen des südbairischen Sprachraums erscheinen diese in vokalisierter form (östliches Südtirol, Osttirol und nördliches Oberkärnten). Ostsüdtirol, Puster- und Lessachtal behalten den historischen Unterschied bei und alternieren zwischen -[ɔ] und -[a] wie in [foːtɔ], Vater und [liɐ̯va], lieber. In der restlichen Region von Osttirol und im Drau-, sowie Mölltal in Oberkärnten ist die Endung ein uniformes -[a] wie [fɔːta] (Wiesinger 1990: 475-476).

Die standardsprachlichen Vokale [eː], [ø], [o] für MHD <ê, œ, o> resultieren in [ɛɐ̯] und [ɔɐ̯] wie in [ʃnɛɐ̯], Schnee; [lɛɐ̯tn̩], löten oder [ʀɔɐ̯t], rot gezeigt werden kann. Diese zentrierende Diphthongierung ist auf den südbairischen Raum begrenzt (Wiesinger 1990: 457).

Das nördlichste Merkmal des südbairischen ist die Erhaltung der Affrikate [k͡x] bzw. stark aspiriertes [kʰ] für standardsprachlich und MHD <k, ck>. Diese Realisierungen stehen im Gegensatz zu dem im Mittel- und Nordbairischen an dieser Stelle auftretenden, geschwächten [g] und [k], was in folgenden Formen deutlich wird: [k͡xnɛxt : gnɛxt], *Knecht* oder [tʀɪŋk͡xŋ : dʀɪŋgɐ], *trinken*. Während im Mittelbairischen vor einem Konsonanten die Stimmhaftigkeit zwischen [k] und [g] neutralisiert wird und in einem [g] resultiert, bleibt dieser Unterschied im Südbairischen erhalten Wiesinger (1990: 458).

Das Südbairische bewahrt auch den Unterschied zwischen MHD und standardsprachlichen <t, d> in wortinitialer und medialer Position als [t] und [d], wohingegen dieser Unterschied im Mittelbairischen ebenso neutralisiert wird wie bei [k] und [g]: [vaɐ̯tɐʀ : vaɐ̯dɐ], *weiter*; [ʃnaɐ̯dɐʀ] : ʒnaɐ̯dɐ], Schneider. Das Südbairische unterscheidet neben der Affrikate [k͡x] und dem Lenisplosiv [g] in der velaren Reihe auch noch den unaspirierten Fortisplosiv [k], welcher standardsprachlich in wortmedialer Position als <ck> und in manchen Fällen initial als <k> oder <g> auftritt. Daher kann im Südbairisch der Erhalt des Unterschieds [k͡x : k : g], welche im Mittel- und Nordbairischen fehlt, als weiterer kennzeichnender Unterscheid zum Mittelbairischen gezählt werden: [bʊk͡xn̩], *bücken*; [ʀʊkn̩], *Rücken* und [sɔːgn̩], *sagen*. Im Gegensatz zum Südbairischen treten im Mittelbairischen nach kurzen Vokalen diese Formen mit dem Fortisplosiv [k] auf (Wiesinger 1990: 458).

Weiters unterscheidet sich das Südbairische in Südtirol, Zentral-Nordtirol, Osttirol

und Kärnten ohne Lavanttal, noch in der Erhaltung von wortfinalem <n> vom Mittelbairischen, welches die Tilgung mit der Nasalierung des vorhergehenden Vokals kompensiert: [vaɐ̯n : vaɛ̃]. Prominentester Unterschied zwischen der südbairischen und der mittelbairischen Varietät dürfte die Realisierung der l-Vokalisierung sein. Vor Konsonanten wird der Lateral [l] vokalisiert wie folgende kontrastierende Beispiele zeigen: [vɔld : vɔɪ̯d], *Wald*; [hɔls : hoɪ̯z], *Hals* oder [tɔːl : dɔɪ̯], *Tal* (Wiesinger 1990: 459).

Hinsichtlich der l-Vokalisierung, die ein prestigereiches Merkmal innerhalb der Bairischen Varietät darstellt, muss erwähnt werden, dass sich diese mehr und mehr in den südbairischen Sprachraum auszubreiten scheint, wie aktuellere Studien (Vollmann et al. 2017) zeigen.

Im gesamten tirolerischen Raum, ausgenommen das Unterinntal- und die Kitzbühelregion, kommt es zur palatalisierten Aussprache des standardsprachlichen <st> [ʃt]. Die Hebung der zentrierenden Diphthonge, die vom MHD <ê, ô> herrühren ist vor Nasalen sehr verbreitet: [gɪə̯n], *gehen*; [lʊə̯n], *Lohn*. Im Gegensatz dazu steht die velare Aussprache von der bereits erwähnten Affrikate [k͡x] und [x] unbeirrt von ihrer Umgebung[2] (Wiesinger 1990: 479).

Das Präfix <ge-> für die Bildung des Partizip Perfekts wird in Osttirol vor Plosiven als [gɪ] realisiert: [gɪtʁiːbm̩] , getrieben oder [gɪb̥ɔːdət], gebadet. Ebenso vor den Nasalen [m], [n] und den Approximanten [l] und [ʊ] bleibt dieses Präfix erhalten, was allerdings nicht auf das restliche Südbairische in Tirol zutrifft: [gɪmɔχt], gemacht; [gɪlɔχt], gelacht. Ein Merkmal, das den Westen vom Osten Tirols abgrenzt ist, die Hebung des [ɔː] zu [uː] vor Nasalen, wobei [ɔː] standardsprachliches <a> repräsentiert, welchem MHD <â> und MHD gelängten <a> bzw. vom MHD <o> vorausgeht, wie bspw. [gɪtuːn], getan oder [muːn], Mann. Im Gegensatz dazu wird im Westen die Originalstufe der Entwicklung beibehalten bzw. zu [oʊ̯] oder [õʊ̯̃] modifiziert: [mɔːn : moʊ̯n : mõʊ̯̃] (Wiesinger 1990: 479).

3.2 Die Eigenheiten von Osttirol und der Südweststeiermark

Zu den wichtigeren Charakteristika des Steirischen zählen eine steigend-fallende Intonation, die Hebung des standardsprachlichen bzw. MHD <o> vor [l] zu [u] und die Hebung des Diphthongs [ɔɪ̯] (MHD <iu> standardsprachlich <ie, eu> zu [ʊɪ̯] wie

2 Hier muss natürlich die komplementäre Verteilung des <ch>-Lautes berücksichtigt werden, die abhängig von seiner Umgebung ist: nach Hinterzungenvokalen wird er zu [x] oder [χ], nach Vorderzungenvokalen zu [ç]

die Beispiele [huɭts], *Holz*; [kx͡xʊɭn], *Kohle* oder [flʊ̯ɪ̯gn̩], *Fliege* belegen. Besonders kennzeichnend für die südbairische Region der Steiermark ist die retroflexe Realisierung des Laterals in diesen präkonsonantischen Positionen. Die zentrierenden Diphthonge [ɛɐ̯] und [ɔɐ̯] (standardsprachlich <e, ö, o> und MHD <ê, œ, ô>) sind vor allem für die westliche Obersteiermark bzw. Weststeiermark üblich. Die sogenannte »Kärntner Längung« ist bereits bis zum oberen Murtal vorgedrungen und im Fall einer Fortisschwächung bewahrt die südliche Region der Weststeiermark noch die ursprünglichen Langvokale: [ʒdʀɔːzn̩], *Straße* oder [ʒlɔːvm̩], *schlafen*. Ein Merkmal, das für die West- und Mittelsteiermark zutrifft ist die Hebung von standardsprachlichen <e, ö, o> bzw. MHD <ê, œ, ô> vor Nasalen zu [ɪɐ̯̃] und [ʊɐ̯̃], woraus Formen wie [gɪɐ̯̃], *gehen* oder [ʒɪɐ̯̃], *schön* entstehen. Ein weiteres auffälliges Merkmal für die (Süd-)Weststeiermark ist der retroflexe Tap [ɽ] vor Labialen und Velaren, wo er teilweise als Silbenträger fungiert: [ɛɐ̯ʃtɪɽp], *er stirbt* oder [pɪɽkx͡xɐn], *Birke* (Wiesinger 1990: 471-472).

Im Gegensatz dazu verhält sich MHD <ô> in Innervillgraten als [ɛ]: [ʀɛt], *rot* oder [pʀɛt], *Brot* (Hornung 1964: 126).

MHD <ê> wird zu [ɛːɐ̯], wie [pɛːɐ̯də], *beide* (MHD *bêde*) oder auch vor dem dentalen Nasal wie [gɛɐ̯n], *gehen* zeigt. Das MHD <ou> hat sich zu [ʊɪ̯] entwickelt: [gʊɪ̯t], *gut* oder [plʊɪ̯t], *Blut*; während die Palatalisierung des MHD <o> allerdings erhalten blieb: [kx͡xœp͡f], *Kopf*; [gɪhœlfm̩], *geholfen*. In Innervillgraten ist dieser Prozess so stark ausgeprägt, dass auch [ø] angenommen werden kann (Hornung 1964: 126).

Die Realisierung des Trills [r] mit seinen Allophonen ist in Osttirol sehr vielseitig. Vor dentalen Plosiven wird spirantisiert und zu [x]: [hɛːɐ̯xt], *Herd*. In der MHD Endung <-er> wird der Trill zu einem geschwächten [ɔ] mit darauffolgenden velaren Frikativ [x]: [zʊmɔˣ], *Sommer*. Die Frikative [f] und [s] wurden in Innervillgraten stimmhaft als [v] und [z] realisiert: [vɔːʀe], *wilde Fahre*, im aktuellen Phonset des Innervillgratischen ist der stimmhafte Frikativ [z] nicht mehr enthalten (Hornung 1964: 127).

4 Empirischer Teil

Für die erfolgreiche Synthese von Sprache benötigt man qualitativ hochwertige Sprachkorpora, welche auf segmentaler Ebene bereits vorverarbeitet sind. So ist es erforderlich Sprachdaten auf phonetischer Ebene zu transkribieren und zu segmentieren. Die Zusammenstellung eines solchen Korpus ist vor allem für die Dialektsynthese, bedingt durch den vielfältigen Charakter von Dialekten, ein sehr zeitintensives Unterfangen. In diesem Abschnitt der Arbeit wird eine einfache Phonmapping-Methode getestet, um die Synthese eines Dialekts zu realisieren, für den kein qualitativ hochwertiges Korpus verfügbar ist. Stattdessen wird ein bereits trainiertes akustisches Hidden-Markov-Modell (HMM) eines ähnlichen Dialekts verwendet um die Synthese durchzuführen. Als Zieldialekt wird hier das Südweststeirische aus der Ortschaft Zelko (Sty) gewählt; der Quelldialekt Innervillgratisch aus Osttirol (Ivg) wird zwar zur selben Dialektgruppe (Südbairisch) gezählt, befindet sich jedoch 225km (siehe 4.1) entfernt. Beide Dialekte überschneiden sich hinsichtlich charakteristischen Phonen wie etwa im retroflexen Lateral [l]. Es besteht die Absicht die Ähnlichkeit auf Lautebene dieser Dialekte für die Realisierung eines authentischen synthetischen Outputs für Sty zu instrumentalisieren.

Es soll darauf aufmerksam gemacht werden, dass die Aussprache des Sprechers, bedingt durch eine Zahnlücke, eine vermehrte Spirantisierung aufweist. Das hat zur Folge, dass die Transkription der Originalsätze unter Berücksichtigung dieses Umstandes, nur als ein approximiertes Transkript des Sprechers gelten kann, da es vorrangig war, die Charakteristik des Sty-Dialekts und nicht des Sprechers festzuhalten.

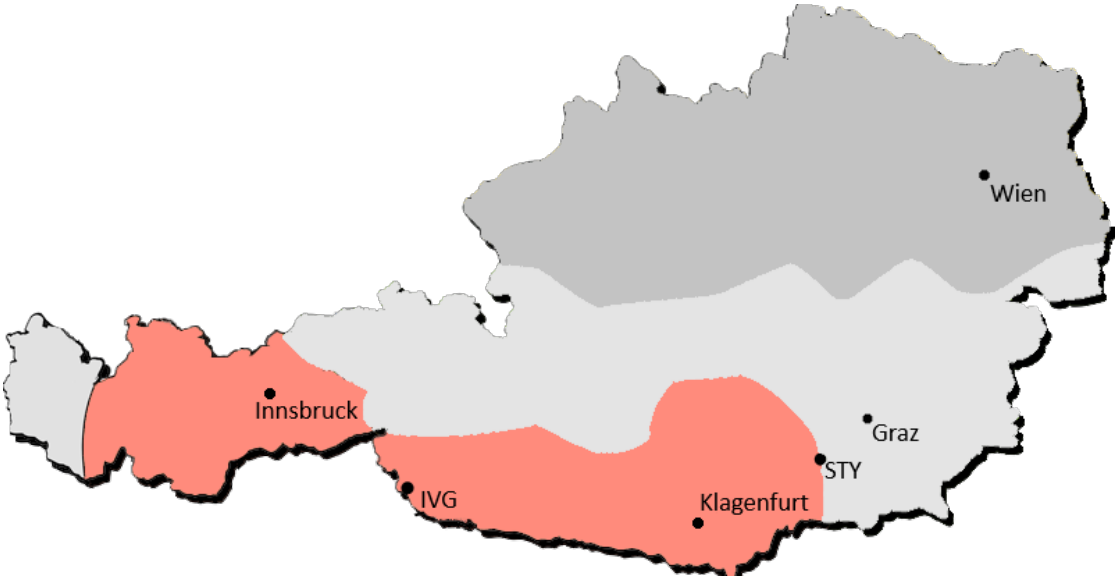

Abbildung 4.1: Der südbairische Raum in Österreich mit den beiden Orten Innervillgraten (IVG) und Zelko (STY)

4.1 Experiment

4.1.1 Korpora

Für dieses Experiment fanden zwei verschiedene Korpora Anwendung. Das Synthesesystem für IVG greift auf ein bereits für Toman et al. (2015) erstelltes Korpus zurück. Für die Determination des STY-Phonsets musste jedoch ein Korpus erstellt werden. Die Erstellung des IVG-Korpus war zweiphasig: Die erste Phase bestand darin, Sprachaufnahmen mit zehn DialektsprecherInnen aus Innervillgraten (Osttirol) im Feld, also ohne Studioaufnahmequalität, durchzuführen. Daraus resultierte für IVG ein 20-stündiges Korpus, aus dem 660 phonetisch ausbalancierte Sätze ausgewählt und manuell transkribiert wurden. In der zweiten Phase wurden vier dieser zehn Personen (zwei männl., zwei weibl.), aufgrund ihrer linguistischen Eignung (aufgewachsen in Innervillgraten, durchgehende Verwendung von charakteristischen phonologischen Prozessen, lexikalisches Wissen und morpho-syntaktische Kompetenz) von einer Linguistin ausgewählt. Die Sprachaufnahmen dieser selektierten ProbandInnen wurden dann in einem Aufnahmestudio durchgeführt. Die Aufnahme beinhaltete ein durch

Stichworte elizitiertes Spontangespräch, Leseaufgaben, eine Bildbenennungsaufgabe und Übersetzungsaufgaben von Standardsprache in den Dialekt (Toman et al. 2015).

Für das wesentlich kleinere STY Korpus wurde ein Sprecher aus dem südweststeirischen Ort Zelko, welcher politisch dem Bezirk Deutschlandsberg zugeordnet wird, aufgenommen. Beim Sprecher handelt es sich um einen 75-jährigen Mann, welcher in Zelko aufgewachsen ist und zeit seines Lebens auch dort gelebt hat. Für die Sprachaufnahme wurden von der Aufnahmeleiterin ein durch Stichworte elizitiertes Spontangespräch, Bildbenennungsaufgaben und Übersetzungsaufgaben von Standard in den Dialekt durchgeführt (n=48). Die Übersetzungsaufgaben gleichen sich mit denen IVG-Lesekorpus gleich (Tomann et al. 2015), welche schließlich auch den Kern des STY-Korpus bilden. Insgesamt hat das STY-Korpus eine Länge von etwa einer Stunde, die Aufnahme erfolgte bei 44100 Hz, 16 bits/sample.

Die Aufnahme fand im Haus des Sprechers statt; das alte Bauernhaus mit großen halligen Räumen barg einige akustische Herausforderungen für das Aufnahmesetting, denen jedoch mit Akustikschaumstoff begegnet werden konnte (Pucher et al. 2017a: 181).

4.1.2 Synthese-Sätze

Aus den bereits genannten Übersetzungsaufgaben wurden 14 Sätze ausgewählt, welche für die Synthese weiterverarbeitet wurden, indem sie in IPA und SAMPA transkribiert und im Weiteren auf Phonebene in der phonetischen Analysesoftware Praat (Boersma & Weenink 2018) segmentiert wurden. Die Transkription für den Synthesizer, der für diese Arbeit verwendet wurde (Toman et al. 2015), basiert auf der SAMPA-Lautschrift (Wells 1997), die es erlaubt IPA-Zeichen in ASCII darzustellen und somit für Computerprogramme lesbar zu machen. Wortgrenzen werden mit ».«, Silben mit »'« und akzenttragende Silben mit »1« dargestellt bzw. markiert, wie Beispiel (1) aus dem STY Korpus illustriert:

(1)

<orth.>	»Wir fahren morgen nach Silian.«
<IPA>	[mɪ.vɔ.mɐ.ˌmoɪ.ŋ.nɔ.ˈʃil.jɑn]
<SAMPA>	`'mI.'1vO'mP6.'1moh'Nl.'nOh.'1Sil'jan`

In Tabelle 4.1 finden sich nur die IPA-Transkriptionen der synthetisierten Sätze. An dieser Stelle muss angemerkt werden, dass es sich bei der Transkription nicht um eine enge phonetische handelt. Diese wurde bewusst vermieden, da sprecherabhängige phonetische Prozesse wie die Labialisierung einzelner Vokale für die Synthese nicht berücksichtigt werden konnten.

Sty_01	Wir sagen zu der Mame und zum Vater Des.
	[ɪzɔŋ͡tsɪmɑ̃ːmɛɔntʊnvɔːdɐˈdɛiʃ]
Sty_02	Wir fahren morgen nach Sillian.
	[mɪvɔmɐˌmõːŋnɔˈʃiljɑn]
Sty_03	Das ist fast wie die Königin.
	[dɛsˌfɔʃdɣɪdɪˈçœːnɪɣɪn]
Sty_04	Tun wir morgen Erdäpfel stecken?
	[d̥ʊmɔ̃ˌmɔ̃ɐ̯̃ŋɛɑ̃dɛb̥͡fl̥ʃdœɪ̯kɤ͡xən]
Sty_05	Ihr glaubt gar nicht, wie schön Urlaub machen ist.
	[d̥ɛzvɔːʃgɔnɛdɣʏfiɛ̃ːzʉːlɑmɔfiɐnɛʃ]
Sty_06	Gehen wir morgen in die Kirche?
	[gɛmɐmɔ̃ɑ̃ŋˈkɤ͡xøːl̥ɑ̃n]
Sty_07	Du bäckst mir morgen einen Kuchen.
	[d̥ʊbɔʊ̯xsmɛ̃ˌmɔ̃ːɐ̯ŋkʰɔʊ̯fiɛ̃n]
Sty_08	Du gehst morgen für mich einkaufen.
	[d̥ʊˌgɛstʰfœmɪɛ̯mɔ̃ːŋˈaɛ̯kɤ͡xɑːβœn]
Sty_09	Gehst du morgen irgendwo hin?
	[(g̊)ɛd̥ˌmɔːŋːɪːɐ̯nˈbɔːfiɑnː]
Sty_10	Hast du das Buch schon gelesen?
	[hɔʃˌb̥iɐ̯lʃʊˈglɛːzn̩]
Sty_11	Das Bett ist zu weich.
	[ˌsb̥ɛi̯tɪsˈt͡svoːxː]
Sty_12	Der Vater hat seinen Hut vergessen.
	[nˌvɑːdœn̩hʊɑ̯tʰfɐˌɣɛːzn̩]
Sty_13	Ich will sie nicht sehen.
	[ɪˌbyznɛˈt͡sɛːfiɑn]
Sty_14	Meine Eltern hätten das so gesagt.
	[mɑ̃ɛ̯̃nɛˈœl̥dɐnfiɛndəsɔʊ̯ˌg̊sɔːkʰ]

Tabelle 4.1: Synthesesätze mit dazugehöriger phonetischer Lautschrift (IPA)

4.1.3 Phonsets

Das STY-Phonset besteht aus insgesamt 61 Lauten, die aus dem kleineren STY-Korpus extrahiert werden konnten. Im Gegensatz dazu, enthält das IVG-Phonset 85 Laute, von welchen 50 auch im STY zu finden sind, d.h. die beiden Phonsets überschneiden sich zu 82% (siehe 4.2).

4.3 zeigt die Phonsets der österreichischen Standardsprache ((R)SAG); Innervillgraten (IVG) und Südweststeirisch (STY). Nachstehend werden die Phonsets des IVG und STY kurz beschrieben.

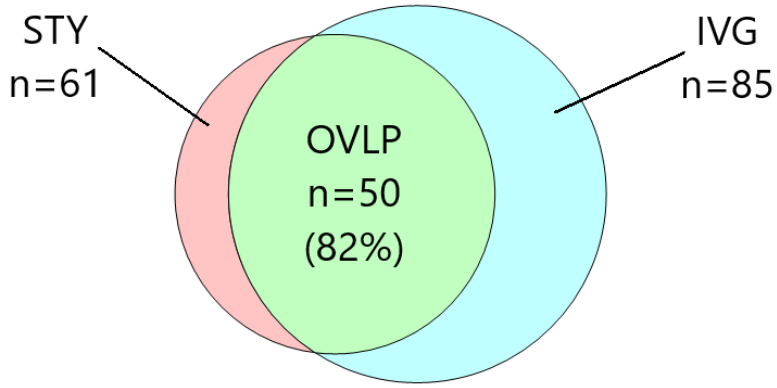

Abbildung 4.2: Überlappung der Phonsets

Kategorie	(R)SAG	IVG	STY
Vokale	ɑ ɑː ɒ ɒː ɐ e ɛ eː i ɪ iː ɔ o oː øː æː ɐ œ œː ə u ʊ uː ʏ y yː	ɑ ɑː ɒ ɒː e ɛ eː ɛː i ɪ iː ɔ o ɔː oː ɐ œ ə u ʊ uː ʏ y yː	ɑ ɑː ɒ ɛ e ː i ɪ ɔ œ ø œː ɐ ə ʉ ʊ u ʏ y
Nasalierte Vokale	ɒ̃ː ɔ̃ː æ̃ː œ̃ː	-	-
Diphthonge	aɪ̯ ɒːɐ̯ ɑːɐ̯ ɒɪ̯ ɑʊ̯ ɛɐ̯ ɛːɐ̯ ɪɐ̯ iɐ̯ iːɐ̯ ɔɐ̯ ɔːɐ̯ oːɐ̯ ɔʏ̯ øːɐ̯ œɐ̯ ʊɐ̯ uːɐ̯ ʊːɐ̯ ʏɐ̯ yːɐ̯	aẹ̯ aɛ̯ ɑɔ̯ ɑɔɐ̯ ɑoɐ̯ eɑ̯ ɛɑ̯ eɛ̯ ɛẹ̯ ɛi̯ ɛɐ̯ iɛ̯ ɪi̯ iɐ̯ ɔɑ̯ oɑ̯ ɔɛ̯ oẹ̯ ɔo̯ ɔɐ̯ ɔu̯ ɛɔ̯ ɐu̯ ui̯ ʊi̯ ʊu̯	ɑɔ̯ aɛ̯ ɑɛ̯ ɛi̯ ɛɑ̯ ɐɪ̯ ɪɐ̯ oɐ̯ ɔu̯ ʊɑ̯
Plosive	b̥ d̥ g̥ k ʔ p t	b̥ b d̥ d ɖ̥ː g̥ g k kʰ ʔ t tʰ	b̥ b d̥ d g̥ g k kʰ t tʰ
Nasale	m n ŋ	m n ŋ	m n ŋ
Frikative	ç x f h s ʃ v z ʒ	β ð ç x f ɣ h s ʃ v	β ç x f ɣ h ɦ s ʃ v z
Affrikaten	-	b̥͡f kˣ k͡s t͡ʃ t͡s	b̥͡f kˣ k͡s t͡s
Approximant	j	j	j
Trills	r	ʀ ʀˣ	-
Lateralapproximant	l	l ɭ	l ɭ

Abbildung 4.3: Phonsets des österreichischen Standarddeutsch (SAG), Innervillgratisch (IVG) und südweststeirische Varietät aus Zelko (STY) (Pucher et al. 2017a: 181)

Vokale

Die 18 Vokale des STY heben sich im Wesentlichen nicht sehr vom IVG-Vokalsystem ab. Der perzeptive Abstand zwischen den Vokalen ist nicht von besonderer Bedeutung. Allgemein kann beobachtet werden, dass der STY- Sprecher zu einer eher offenen statt geschlossenen Artikulation der Vokale neigt.
Der für die österreichische Aussprache typische ungerundete offene Hinterzungenvokal [ɑ] (Moosmüller 2007) wird auch im STY anstelle des ungerundeten Vorderzungenvokals [a] bevorzugt realisiert. Die weiteren ungerundeten Vorderzungenvokale [e] und [ɛ] treten beide auf, wobei angemerkt werden muss, dass es eine klare Präferenz – selbst in betonter Position – des STY-Sprechers zum halboffenen Vorderzungenvokal [ɛ] gibt.

In Positionen, die den Vokal [ɛ] erwarten lassen, tritt häufig seine gerundete Variante, der gerundete halboffene Vorderzungenvokal [œ] auf. Untypisch ist der gerundete, geschlossene Zentralvokal [ʉ], der jedoch nur mit einer Notation vorkommt und bedingt durch einen Zentrierungsprozess nicht als festes Inventarelement des STY gelten kann. Selbiges kann über den halbgeschlossenen gerundeten Vorderzungenvokal [ø] gesagt werden, und ist wahrscheinlich auf die Rundungsfreudigkeit des Sprechers zurückzuführen.

In unbetonten Nebensilben von Wörtern, die den Fokusakzent der Phrase tragen, werden Vollvokale bevorzugt, daneben gibt es in unbetonten Silben klarerweise auch die beiden Schwa-Laute [ə] und [ɐ].

Zusammenfassend unterscheidet sich das STY-Vokalsystem in nur zwei Instanzen - [ø] und [ʉ] - von dem des IVG. Der Quantitätsunterschied in [œː] ist an dieser Stelle nicht relevant, da die Segmentdauer des STY-Sprechers auf den synthetisierten Output übertragen wurde.

Diphthonge

Es konnten aus dem STY-Korpus insgesamt 12 Diphthonge extrahiert werden. Während es im IVG-Phonset nicht nur steigende oder fallende Diphthonge gibt, sondern auch horizontale Di- und auch Triphthonge, können im STY nur steigende oder fallende Diphthonge belegt werden. Die Laute [ɛɑ̯] und [oɐ̯] zählen zu den kennzeichnenden Diphthongen der Südweststeiermark, vor allem weil sie von einer steigend-fallenden Intonation begleitet werden.

Die Diphthonge stellen im Experiment eine kritische Kategorie dar, da sechs der 12

Diphthonge nicht im IVG-Phonset enthalten sind. Das Mapping dieser Laute stellte einige Herausforderungen, auf die in Abschnitt 4.2. näher eingegangen wird.

Konsonanten

Die Plosive und Nasale fallen wie erwartet aus, auffällig bei den Konsonanten ist das Fehlen eines uvularen Trills [ʀ] oder Frikativ [ʁ], was jedoch darin begründet scheint, dass einerseits der Korpus sehr klein war und andererseits der Sprecher durch seine physiologische Merkmale den Trill durch den stimmhaften velaren Frikativ [ɣ] substituiert hat. Ebenso dürfte das Auftreten der stimmhaften Frikative [ɦ] und [z], sowie das Fehlen des glottalen Plosivs [ʔ] am Alter des Sprechers liegen, das zu einer muskulären Schwäche der Artikulatoren führt. Wider Erwarten, konnte ein charakteristischer Laut für die (Süd-)Weststeiermark, der retroflexe Tap [ɽ] vor Labialen und Velaren, leider nicht im Korpus gefunden werden.

Im STY-Korpus konnte wie erwartet der alveolare und der retroflexe Lateralapproximanten [l] und [ɭ] belegt werden.

Wie bereits Wiesinger (1990: 458) feststellte, gibt es im südbairischen Dialektraum noch die Unterscheidung zwischen den Velaren [g]; [k] und [k^h], was auch auf das STY zutrifft. Besonders in betonten bzw. wortinitialen Positionen sind die stimmlosen oder aspirierten Velarplosive belegt.

Die Frikative im STY-Phonset sind vielseitig, gleichen sich jedoch in großer Mehrheit mit dem IVG-Phonset. [ɦ] und [z] sind im IVG-Phonset nicht enthalten, und werden auf die vermehrte Stimmhaftigkeit des STY-Sprechers zurückgeführt.

4.2 Phonmapping

Die Unterschiede zwischen den Phonsets beider Varietäten bestehen vorwiegend in der Stimmhaftigkeit der einzelnen Laute.
Durch das Mapping von stimmhaften Lauten wie [ɦ], [z] auf ihre stimmlosen Pendants, wird kein großer Einschnitt in die Dialektcharakteristik erwartet. Im Gegensatz dazu, werden beim Mapping der Diphthonge einige Einschränkungen in der Realisierbarkeit der Synthese bzw. auch der Authentizität des Dialekts erwartet. Es wird angenommen, dass Diphthonge in Verbindung mit Intonation im STY charakteristische Merkmale transportieren.

Fünf der sechs Diphthonge, die im IVG-Phonset nicht enthalten sind, konnten unter

Einhaltung der Mapping-Prämisse auf einen entsprechenden IvG-Diphthong zugewiesen werden. Die Prämissen für das Mapping der Vokale beziehen sich auf den Verengungsort, welcher ähnlich sein musste; und auch auf die privativen Merkmale, die ein Laut mit dem zugewiesenen teilen musste (Pucher et al. 2017a: 182). Weitere Unterschiede im Phonset fanden sich in den Diphthongen [ɪʊ̯] und [oʊ̯], die, vor Nasalen, als charakteristisch für die Region um STY gelten soll (Wiesinger 1990: 471).

[ɪʊ̯] wurde dem IvG-Diphthong [iʊ̯] zugewiesen, da der erste Teil des Diphthongs [i] keinen großen phonetischen Abstand zum ursprünglichen [ɪ] aufweist. Für [oʊ̯] wurde ähnlich vorgegangen; er wurde dem Laut [ɔʊ̯] zugewiesen. Etwas kritischer war die Zuweisung von [œɪ̯] auf [ɛi̯], da hier beide Teile des Diphthongs ersetzt werden mussten; der erste wurde entrundet, weist allerdings dieselbe Zungenhöhe auf, der zweite wurde zu [i] gehoben. Ein weiterer kritischer Punkt ist die Doppelbesetzung des zugewiesenen Lautes [ɛi̯]. Dieser wurde nämlich auch als Ersatz für [ɛɪ̯] verwendet, wobei hier wiederum der phonetische Abstand als nicht groß gilt und daher keine großen Einschnitte in der Verständlichkeit bzw. Charakteristik des Synthetisierten erwartet wurden.

Für den Diphthong [ʊɑ̯] konnte allerdings kein zufriedenstellendes Mapping durchgeführt werden. In diesem Fall wurde versucht den Diphthong aus zwei getrennten Lauten herzustellen, die Modelle für [ʊ] und [ɑ] wurden mit einer verkürzten Segmentdauer aneinandergereiht und wie die restlichen Laute separat synthetisiert.

4.2 zeigt das Mapping, wie es für die Synthese angewendet wurde.

STY	IvG
ø	œ
ʉ	ʊ
ɑɛ̯	aɛ̯
ɛɪ̯	ei̯
œɪ̯	ɛi̯
ɪʊ̯	iʊ̯
oʊ̯	ɔʊ̯
ɦ	h
z	s

Tabelle 4.2: Mapping. Links die STY-Laute und rechts die ihnen zugewiesenen IvG-Laute (Pucher et al. 2017a: 182)

4.3 Synthese

Für die Synthese der drei verschiedenen Versionen der Prompts wurden die phonetischen Transkriptionen des STY in IVG-Transkriptionen mithilfe des oben beschriebenen Mapping-Prozesses transformiert. Für das HMM-basierte Synthesesystem wird die phonetische Transkription, die neben dem phonetischen auch den prosodischen Kontext beinhaltet, in für den Synthesizer verarbeitbare *full-context-labels* transformiert. Mithilfe der full-context-labels und der bereits entwickelten sprecherabhängigen Synthesestimme des IVG-Sprechers CSC (Toman et al. 2015) wurde das erste Set an Prompts, vorerst ohne Übertragung der Prosodie des STY synthetisiert (Pucher et al. 2017a: 182). Diese Prompts erhalten die Abkürzung *syn*.

Im zweiten Prompt-Set wurde die ungefähre Phondauer des STY-Sprechers, jedoch die exakt selbe Samplelänge des STY übernommen (*syn_dur*).

Für das dritte Prompt-Set wurde neben der Segmentdauer ebenso der Grundfrequenzverlauf der Äußerungen des STY-Sprechers auf den Syntheseoutput übertragen (*syn_dur_f0*). Um den IVG-Grundfrequenzverlauf dem des STY-Sprechers anzunähern, musste die F0-Kurve des STY-Sprechers zuerst auf Phrasenebene modifiziert werden (Pucher et al. 2017a: 183). Ein Algorithmus[1] wurde angewendet um die F0-Dynamik des STY-Sprechers zu übernehmen, den ursprünglichen F0-Umfang der IVG-Stimme jedoch zu behalten. Dies war notwendig, da es zwischen der STY- und IVG-Stimme es einen signifikanten Unterschied im Grundfrequenzumfang gab (STY weist eine höhere F0 als IVG auf) (Pucher et al. 2017a: 183). 4.4 und 4.5 zeigen einen Vergleich des Satzes »Du bäckst mir morgen einen Kuchen« (STY_07). Hier wird einerseits der Unterschied zwischen den Grundfrequenzverläufen beider Samples deutlich, andererseits ist hier auch erkenntlich, dass die Qualität des STY-Samples um einiges geräuschhafter ist, als die des Synthesizers.

1 Für den genauen Algorithmus siehe Pucher et al. (2017a: 183)

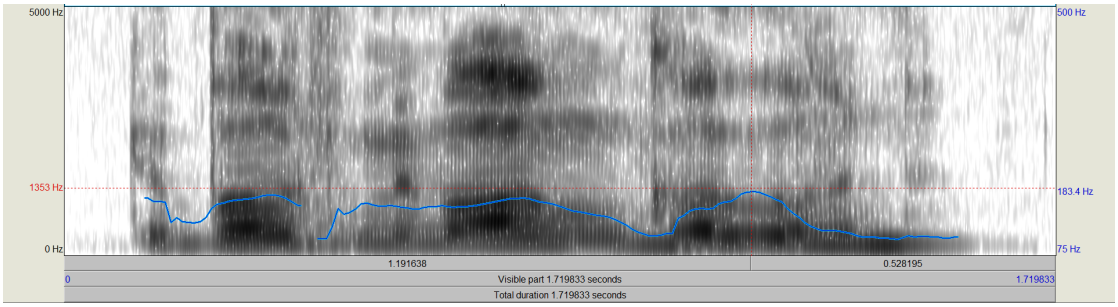

Abbildung 4.4: Spektrogramm mit Grundfrequenzverlauf von STY_07 »Du bäckst mir morgen einen Kuchen« des STY-Sprechers.

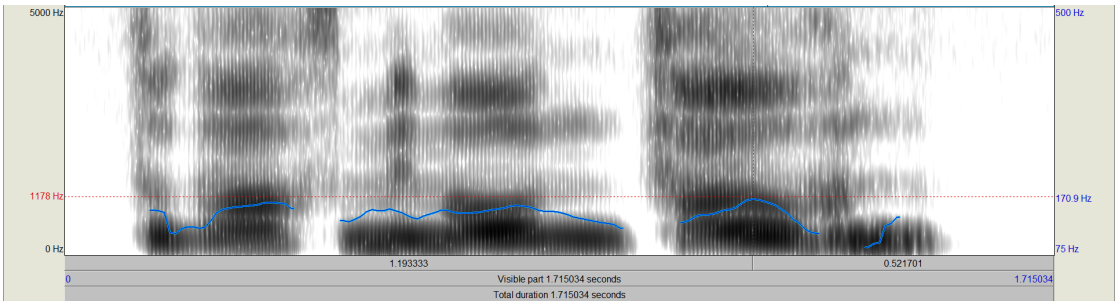

Abbildung 4.5: Spektrogramm mit Grundfrequenzverlauf von STY_07 »Du bäckst mir morgen einen Kuchen« der synthetischen IVG-Stimme.

4.4 Evaluation und Ergebnisse

Für die Evaluation der synthetisierten Prompts wurden zehn HörerInnen aus vier verschiedenen Regionen Österreichs befragt. Ein Hauptkriterium für die Auswahl der HörerInnen war die Vertrautheit mit dem STY bzw. mit südbairischen Varietäten. Die Hörer waren im Alter von 22 und 66 Jahren und kamen aus der Steiermark (Bezirke Graz, Eisenerz, Bruck an der Mur und Deutschlandsberg), Kärnten (Spittal an der Drau und Villach Land), Tirol (Kufstein) und Wien (Pucher et al. 2017a: 183).

Die drei-teilige Evaluation bestand aus einem Verständlichkeitstest, einem Mean-Opinion Score-Test (MOS) und einem paarweisen Vergleich der synthetisierten Stimmen.

Im Verständlichkeitstest wurden den HörerInnen die Originalaufnahmen und eine

Version der synthetisierten Prompts (*syn, syn_dur oder syn_dur_fo*) einmalig vorgespielt und diese danach gebeten, das Verstandene in einem Textfeld niederzuschreiben. Aus der Differenz zwischen der Niederschrift der HörerInnen und dem zugrunde liegenden Textinput ergibt sich die Word-Error-Rate (WER), eine Quote, die in der Evaluation von Synthese-, aber auch Spracherkennungssystemen den Fehler des/der Hörers/-in bzw. des Systems bei ASR-Systemen in Prozent angibt. Hinsichtlich Synthesesystemen sagt die WER aus: Je höher die WER, desto unverständlicher ist der synthetisierte Output. 4.3 liefert einen Überblick über die Ergebnisse des Verständlichkeitstests; hier wird deutlich, dass die Originalaufnahme des Sty-Sprechers mit einer WER von 32.6% deutlich niedriger und somit verständlicher als die synthetisierten Outputs ist. Die Unterschiede zwischen den synthetisierten Outputs sind gering. .

Methode	*orig*	*syn*	*syn_dur*	*syn_dur_fo*
Wer	32.6	46.8	51.2	51.5

Tabelle 4.3: Ergebnisse des Verständlichkeitstests der verschiedenen Methoden, angegeben mit der Word-Error-Rate in % (Pucher et al. 2017a: 183).

Der zweite Teil der Evaluation sah vor, mittels eines MOS-Tests die Authentizität des Dialekts und des synthetisierten Outputs zu ermitteln. Den HörerInnen wurden alle Audiosamples, d.h. synthetisierte und Originalsamples, vorgespielt. Anschließend waren die Samples auf einer Ordinalskala (1 – »schlecht«, 2 – »dürftig«, 3 – »mittelmäßig«, 4 – »gut«, 5 - »herrvorragend«) zu bewerten. 4.6 zeigt die durchschnittliche Bewertung der einzelnen Methoden, woraus auch hervorgeht, dass die prosodiebasierten Methoden, etwas besser abschnitten als die Synthese ohne Prosodietransfer (Pucher et al. 2017a: 184).

Im dritten Teil der Evaluation wurde ein paarweiser Vergleich aller Audiosamples (original und die drei synthetisierten Methoden) durchgeführt. Die Samples des Sty-Sprechers wurden immer als besser bewertet als die synthetisierten Samples, d.h. *orig* gewann in jedem Vergleich. Die Basissynthese (*syn*) hat in 30% der Vergleiche gewonnen, die Dauersamples (*syndur*) im Durschnitt 50% und die Dauer-Fo-Methode (*syn_dur_fo*) in 70% der Vergleiche (Pucher et al. 2017a: 184).

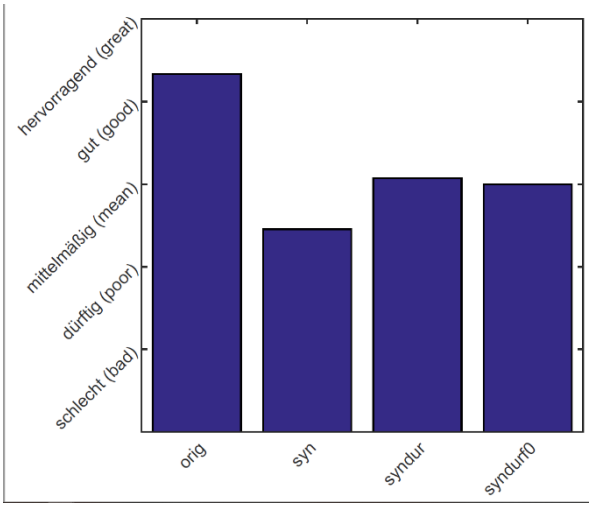

Abbildung 4.6: Ergebnisse des Mean-Opinion-Score-Tests (Pucher et al. 2017a: 184)

Ein zwei-seitiger t-Test konnte zeigen, dass sich die *orig*-Methode statistisch von den anderen Methoden unterscheidet, genauso wie *syn* von *syn_dur* und *syn_dur_fo*, während zwischen den prosodietransfer-Methoden kein statistischer Unterschied festgestellt werden konnte.

5 Conclusio

Die Idee einer Maschine, die mit einer menschlich klingenden Stimme spricht, ließ im vergangenen Jahrhundert ein ganzes Forschungsfeld entstehen, welches bis heute durch eine wachsende interdisziplinäre Forschungsgemeinschaft geprägt ist. Ein Blick auf die Geschichte der Synthesesysteme konnte illustrieren, wie der Wunsch nach einer natürlich klingenden, künstlichen Stimme als Fixpunkt bestehen bleibt. Er zieht sich wie ein roter Faden durch ein Jahrhundert an Forschung, während sich um ihn herum Fragestellungen verschieben, grundsätzlich neue entstehen und durch anhaltenden technischen Fortschritt ebenso Paradigmenwechsel ereignen.

Man darf die Entwicklungen im Gebiet der Sprachsynthese nicht isoliert vom Geschehen in anderen Wissenschaften betrachten. So war die automatische Spracherkennung (ASR) bspw. maßgeblich am Erfolg von TTS-Systemen beteiligt. Natürlich kann eine vielversprechende Methode in der ASR nicht eine eigenständige Exploration dieser Themen oder Methodik in der Sprachsynthese ersetzen. Für die ASR erfolgreichen Hidden-Markov-Modelle oder aktuell auch Deep-Neural-Networks können nicht durch simple Analogie einen funktionierenden Lösungsansatz für ein gänzlich unterschiedliches Problemfeld wie die Synthese darstellen (Pucher 2017). Ebenso war es zu Beginn der Syntheseforschung maßgeblich, Erkenntnisse über die Lautproduktion und akustischer Merkmale (von Kempelen) zu sammeln, um primär einmal nur sprachähnliche Laute zu generieren. Dies hatte eine intensive Auseinandersetzung mit der menschlichen Lauterzeugung und den akustischen Merkmalen von Lauten zur Folge (Fant). Die neuen Erkenntnisse in Lautproduktion, Lautperzeption sowie akustischer Phonetik offenbarten sich zunehmend als Dringlichkeit, dieses Wissen in die Synthesesysteme einzuarbeiten. Aus anderen wissenschaftlichen Feldern, wie der Informatik, kamen zunehmend mehr Möglichkeiten in die Syntheseforschung, dieses Wissen umzusetzen. Mehr Rechenkapazität, die auch seit den frühen 1980er Jahren immer leistbarer wurde, und nicht mehr ausschließlich den privaten Telekommunikationsforschungseinrichtungen verfügbar waren, ließ mit Unit-Selection-Ansätzen eine bis heute qualitativ

hohe Synthesemethode entstehen. Ebenso daraus resultierend ist der aktuelle und sehr erfolgreiche DNN-Ansatz.

Die künstliche Erzeugung von Sprache und die Erkenntnisse anderer Forschungsfelder stehen somit in einer dynamischen Wechselbeziehung zueinander, welche ein sich ständig erweiterndes Forschungsfeld zum interdisziplinären Austausch auf mehreren Ebenen anregt.

Mit dem Phonmapping-Ansatz konnte gezeigt werden, wie wichtig phonetische sowie dialektologische Kenntnisse für die Sprachsynthese sein können. Es konnte auf eine effiziente Weise ein neuer Dialekt synthetisiert werden, ohne dabei einen sonst notwendigen großen Korpus zu sammeln oder zeitintensiv aufzubereiten. Diese Arbeit konnte zeigen, dass für die Erzeugung eines noch nicht digitalisierten Dialekts der vorgestellte Phonmapping-Ansatz eine gute Methode darstellt.

Weiters konnte gezeigt werden, dass der Transfer der Prosodie der Zielvarietät auf eine bereits existierende Synthesestimme die Qualität der Synthese positiv beeinflusst.

Wie Pucher et al. (2018) zeigen konnte, ist es aber für die zukünftige Forschung bedeutend, Systeme, die sich mit der Synthese von Varietäten beschäftigen, vor dem Hintergrund ihres potentiellen Anwendungsgebietes zu evaluieren. Evaluierungsmethoden werden in Zukunft immer mehr an Bedeutung gewinnen, da die Anwendung von Synthesesystemen schließlich im Alltag eines jeden angekommen ist.

6 Literatur

Allen, Jonathan, Sharon Hunnicutt, Rolf Carlson & Bjorn Granstrom. 1979. MITalk-79: The 1979 MIT text-to-speech system. *The Journal of the Acoustical Society of America* 65(S1). S130–S130.

Black, Alan & Paul Taylor. 1994. CHATR: a generic speech synthesis system. In *Proceedings of COLING*, 983–986. Kyoto, Japan.

Black, Alan & Paul Taylor. 1997. Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings of Eurospeech*, 601–604. Rhodes, Greece.

Boersma, Paul & David Weenink. 2018. *Praat: doing phonetics by computer*. O. O.

Breuer, Ludwig M. 2017. Computers & Coffee: Computergestützte Sprachproduktionstests zur syntaktischen Variation des »unbestimmten Artikels vor Massennomen« im »Wienerischen«. In Helmut Kowar (Hg.), *International Forum on Audio-Visual Research: Jahrbuch des Phonogrammarchivs*, Bd. 7, 86–111. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Clyne, Michael (Hg.). 1991. *Pluricentric languages: Differing norms in different nations*. English (Contributions to the Sociology of Language 62). Berlin, Boston.

Cohen, Michael H., James P. Giangola & Jennifer Balogh (Hgg.). 2004. *Voice user interface design*. en. Boston: Addison-Wesley.

Cooper, Franklin S., Alvin M. Libermann & John M. Borst. 1951. The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Science* (37). 318–325.

Davis, K. H., R. Biddulph & S. Balashek. 1952. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America* 24(6) (November). 637–642.

Dizdarevic, V., M. Hagmuller, G. Kubin, E. Pernkopf & M. Baum. 2004. Prosody-based recognition of spoken German varieties. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Bd. 1, I–929–32. Montreal, QC, Canada: IEEE.

Donovan, Robert Edward. 1996. *Trainable speech synthesis*. Cambridge: University of Cambridge Dissertation.

Dudley, Homer, R.R. Riesz & S.S.A. Watkins. 1939. A synthetic speaker. *Journal of the Franklin Institute* 227(6). 739–764.

Dudley, Homer & T. H. Tarnoczy. 1950. The Speaking Machine of Wolfgang von Kempelen. *The Journal of the Acoustical Society of America* 22(2). 151–166.

El Zarka, Dina, Barbara Schuppler, Carina Lozo, Wolfgang Eibler & Patrick Wurzwallner. 2017. Acoustic correlates of stress and accent in Standard Austrian German. In Sylvia Moosmüller, Manfred Sellner & Carolin Schmid (Hgg.), *Phonetik in und über Österreich*. Wien: Österreichische Akademie der Wissenschaften.

Fant, Gunnar. 1953. Speech Communication Research. *IVA, Royal Swedish Academy of Engineering Sciences* 24. 331–337.

Fant, Gunnar. 1960. *Acoustic theory of speech production*. Den Hag: Mouton.

Fant, Gunnar & J. Mártony. 1962. Speech synthesis. Instrumentation for parametric synthesis (OVE II). *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)* 3(2). 18–24.

Ge, Shuzhi Sam, Oussama Khatib, John Cabibihan, Reid Simmons & Mary-Anne Williams (Hgg.). 2012. *Social robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012: proceedings* (Lecture notes in artificial intelligence 7621). Berlin; New York: Springer.

Goodrich, Michael A. & Alan C. Schultz. 2007. Human-Robot Interaction: A Survey. *Foundations and Trends® in Human-Computer Interaction* 1(3). 203–275.

Hagmüller, Martin. 2001. *Recognition of regional variants of German using prosodic features*. Graz: Technische Universität Graz Diplomarbeit.

Hornung, Maria. 1964. *Mundartkunde Osttirols: eine dialektgeographische Darstellung mit volkskundlichen Einblicken in die altbäuerliche Lebenswelt; mit laut- und wortkundlichen Karten*. Graz, Wien: Böhlau.

Huang, Chao, Eric Chang, Jianlai Zhou & Kai-Fu Lee. 2000. Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. In *Sixth International Conference on Spoken Language Processing (ICSLP 2000)*. Beijing, China.

Humphries, J.J., P.C. Woodland & D. Pearce. 1996. Using accent-specific pronunciation modelling for robust speech recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Bd. 4, 2324–2327. Philadelphia, PA, USA: IEEE.

Hunt, Andrew & Alan Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acou-*

stics, Speech, and Signal Processing Conference Proceedings, Bd. 1, 373–376. Atlanta, GA, USA: IEEE.

Kempelen, Wolfgang von, Fabian Brackhane, Richard William Sproat, Jürgen Trouvain & Wolfgang von Kempelen. 2017. *Der Mechanismus der menschlichen Sprache* (Studientexte zur Sprachkommunikation Band 87-88). Dresden: TUD Press.

King, Simon. 2011. An introduction to statistical parametric speech synthesis. *Sadhana* 36(5) (Oktober). 837–852.

Klatt, Dennis H. 1982. The Klattalk text-to-speech conversion system. In *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Bd. 7, 1589–1592. Paris, France: Institute of Electrical & Electronics Engineers.

Klatt, Dennis H. 1987. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*. 737–793.

Krenn, Brigitte, Birgit Endrass, Felix Kistler & Elisabeth André. 2014. Effects of language variety on personality perception in embodied conversational agents. In *Human-Computer Interaction. Advances Interaction Modalities and Techniques - 16th International Confrence*, 429–439. Heraklion, Griechenland.

Krenn, Brigitte, Stephanie Schreitter, Friedrich Neubarth & Gregor Sieber. 2012. Social Evaluation of Artificial Agents by Language Varieties. In Yukiko Nakano, Michael Neff, Ana Paiva & Marilyn Walker (Hgg.), *Intelligent Virtual Agents*, Bd. 7502, 377–389. Berlin, Heidelberg: Springer Berlin Heidelberg.

Langa, Ranta, Jonas Manamela & Nalson Gasela. 2012. Synthesis of dialect speech for an underresourced language. In *SATNAC*, 2. Western Cape, South Africa.

Lawrence, Walter. 1953. The synthesis of speech from signals which have a low information rate. *Communication Theory*. 460–469.

Lenz, Alexandra N. 2016. On eliciting dialect-syntactic data. Comparing direct and indirect methods. In Augustin Speyer & Philipp Rauth (Hgg.), *Syntax aus Saarbrücker Sicht. Beiträge der SaRDiS-Tagung zur Dialektsyntax (Zeitschrift für Dialektologie und Linguistik Beihefte. 165).* 187–219. Stuttgart: Steiner.

Lenz, Alexandra N. & Franz Patocka (Hgg.). 2016. *Syntaktische Variation: areallinguistische Perspektiven*. ger (Wiener Arbeiten zur Linguistik Band 2). Göttingen: V&R unipress.

Möbius, Bernd. 2001. German and Multilingual Speech Synthesis. *AIMS*. Phonetik AIMS 7(4).

Moosmüller, Sylvia. 1995. Evaluation of language use in public discourse. In Patrick Stevenson (Hg.), *The German Language and the Real World. Sociolinguistic, Cultural, and Pragmatic Perspectives on Contemporary German*, 257–278. Oxford: Clarendon Press.

Moosmüller, Sylvia. 2007. *Vowels in Standard Austrian German. An acoustic-phonetic and phonological analysis.* Wien: Habilitationsschrift.

Navas, Eva, Inma Hernaez, Daniel Erro, Jasone Salaberria, Beñat Oyharçabal & Manuel Padilla. 2014. Developing a Basque TTS for the Navarro-Lapurdian Dialect. In Juan Luis Navarro Mesa, Alfonso Ortega, António Teixeira, Eduardo Hernández Pérez, Pedro Quintana Morales, Antonio Ravelo García, Iván Guerra Moreno & Doroteo T. Toledano (Hgg.), *Advances in Speech and Language Technologies for Iberian Languages*, Bd. 8854, 11–20. Cham: Springer International Publishing.

Peters, Jörg, Peter Auer, Peter Gilles & Margret Selting. 2015. Untersuchungen zur Struktur und Funktion regionalspezifischer Intonationsverläufe im Deutschen. Rückblick auf ein Forschungsprojekt. *Regionale Variation des Deutschen. Projekte und Perspektiven*. 53–80.

Pfister, Beat & Tobias Kaufmann. 2008. *Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Pucher, Michael. 2017. *Speech processing for multimodal and adaptive systems*. Graz: Technische Universität Graz Habilitationsschrift.

Pucher, Michael, Carina Lozo & Sylvia Moosmüller. 2017a. Phone mapping and prosodic transfer in speech synthesis of similar dialect pairs. In *28th Conference on Electronic Speech Signal Processing*, 180–185. Saarbrucken, Germany.

Pucher, Michael, Carina Lozo & Sylvia Moosmüller. 2018. Evaluation methods for dialect speech synthesis of similar dialect pairs. In *DAGA 2018 - 44. Jahrestagung für Akustik*, 515–517. Munich, Germany.

Pucher, Michael, Friedrich Neubarth, Volker Strom, Sylvia Moosmüller, Gregor Hofer, Christian Kranzler, Gudrun Schuchmann & Dietmar Schabus. 2010a. Resources for speech synthesis of Viennese varieties. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, 105–108. Valetta, Malta.

Pucher, Michael, Dietmar Schabus, Gregor Hofer, Nadja Kerschhofer-Puhalo & Sylvia Moosmüller. 2012. Regionalizing Virtual Avatars - Towards Adaptive Audio-Visual Dialect Speech Synthesis. In *CogSys 2012, 5th International Conference on Cognitive Systems*, 95. Vienna, Austria.

Pucher, Michael, Dietmar Schabus, Junichi Yamagishi, Friedrich Neubarth & Volker Strom. 2010b. Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. *Speech Communication* 52(2). 164–179.

Pucher, Michael, Bettina Zillinger, Markus Toman, Dietmar Schabus, Cassia Valentini-Botinhao, Junichi Yamagishi, Erich Schmid & Thomas Woltron. 2017b. Influence

of speaker familiarity on blind and visually impaired children's and young adults' perception of synthetic voices. 46. 179–195.

Sagisaka, Yoshinori, Nobuyoshi Kaiki, Naoto Iwahashi & Katsuhiko Mimura. 1992. ATR -Talk Speech Synthesis System. In *Second International conference on Spoken Language Processing (ICSLP '92)*, 483–486. Banff, AB, Canada.

Schabus, Dietmar. 2009. *Interpolation of Austrian German and Viennese Dialect/Sociolect in HMM-based speech synthesis.* Wien: Technische Universität Wien Diplomarbeit.

Schmid, Carolin & Sylvia Moosmüller. 2017. An acoustic comparison between stressed and unstressed vowels in Standard Austrian German and Standard German German. In Sylvia Moosmüller, Manfred Sellner & Carolin Schmid (Hgg.), *Phonetik in und über Österreich*, 45–59. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Schrank, Tobias. 2014. *Relevante Merkmal zur Erkennung von Unsicherheit im Dialog*. Graz: Karl-Franzens Universität Graz Diplomarbeit.

Schröder, Marc und Trouvain, Jürgen. 2003. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology* 6(4). 365–377.

Schroeder, Manfred R. 1993. A brief history of synthetic speech. *Speech Communication* 13(1-2). 231–237.

Selting, Margret. 1995. *Prosodie im Gespräch: Aspekte einer interaktionalen Phonologie der Konversation* (Linguistische Arbeiten 329). Tübingen: Max Niemeyer.

Selting, Margret. 2004. Regionalized intonation in its conversational context. In Peter Gilles & Jörg Peters (Hgg.), *Regional variation in intonation* (Linguistische Arbeiten 492), 49–73. Tübingen: Niemeyer.

Siebenhaar, Beat. 2004. Comparing timing models of two Swiss German dialects. *Language Variation in Europe. Papers from ICLaVE* 2. 353–365.

Socin, Adolf. 1888. *Schriftsprache und Dialekte im Deutschen nach Zeugnissen alter und neuer Zeit: Beiträge zur Geschichte der deutschen Sprache*. Heilbronn: Gebrüder Henninger.

Steiner, Ingmar, Jürgen Trouvain, Judith Manzoni & Peter Gilles. 2015. MaryLux – Luxemburgische Sprachsynthese. In *Proceedings of 5. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen e. V. (IGDD)*. Luxembourg.

Tamagawa, Rie, Catherine I. Watson, I. Han Kuo, Bruce A. MacDonald & Elizabeth Broadbent. 2011. The Effects of Synthesized Voice Accents on User Perceptions of Robots. *International Journal of Social Robotics* 3(3). 253–262.

Taylor, Paul. 2009. *Text-to-speech synthesis*. Cambridge, UK: Cambridge University Press.

Tokuda, Keiichi, T. Kobayashi & S. Imai. 1995. Speech parameter generation from HMM using dynamic features. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, Bd. 1, 660–663. Detroit, MI, USA: IEEE.

Tokuda, Keiichi, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi & Keiichiro Oura. 2013. Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE* 101(5). 1234–1252.

Tokuda, Keiichi, Heiga Zen & Alan Black. 2002. An HMM-based speech synthesis system applied to English. In *IEEE Speech Synthesis Workshop*, 227–230. Santa Monica, CA, USA.

Toman, Markus. 2016. *Transformation and interpolation of language varieties for speech synthesis*. Wien: Technische Universität Wien. Dissertation.

Toman, Markus & Michael Pucher. 2013. Structural KLD for Cross-Variety Speaker Adaptation in HMM-based Speech Synthesis. In *Computer Graphics and Imaging / 798: Signal Processing, Pattern Recognition and Applications*. Innsbruck, Austria: ACTAPRESS.

Toman, Markus, Michael Pucher, Sylvia Moosmüller & Dietmar Schabus. 2015. Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis. *Speech Communication* 72. 176–193.

Trouvain, Jürgen. 2004. *Tempo Variation in Speech Production. Implications for Speech Synthesis.* Saarbrücken: Universität des Saarlandes. Dissertation.

Ulbrich, Christiane. 2005. *Phonetische Untersuchungen zur Prosodie der Standardvarietäten des Deutschen in der Bundesrepublik Deutschland, in der Schweiz und in Österreich* (Hallesche Schriften zur Sprechwissenschaft und Phonetik Bd. 16). Frankfurt: P. Lang.

Vollmann, Ralf, Thorsten Seifter, Bettina Hobel & Florian Pokorny. 2017. l-vocalization in Styria. In Sylvia Moosmüller, Manfred Sellner & Carolin Schmid (Hgg.), *Phonetik in und über Österreich* (Veröffentlichung zur Linguistik und Kommunikationsforschung 31), 123–136. Wien: Österreichische Akademie der Wissenschaften.

Wells, J.C. 1997. SAMPA computer readable phonetic alphabet. In *Handbook of Standards and Resources for Spoken Language Systems.* Part IV, section B. Berlin; New York: Mouton de Gruyter.

Wiesinger, Peter. 1990. The Central and Southern Bavarian Dialects in Bavaria and Austria. In Charles V. J. Russ (Hg.), *The Dialects of modern German: a linguistic survey*, 438–519. O. O.: Routledge.

Wu, Zhizheng, Oliver Watts & Simon King. 2016. Merlin: An Open Source Neural Network Speech Synthesis System. In *9th ISCA Speech Synthesis Workshop (2016)*, 218–223. Sunnyvale, CA, USA.

Wutiwiwatchai, Chai, Ausdang Thangthai, Ananlada Chotimongkol, Chatchawarn Hansakunbuntheung & Nattanun Thatphithakkul. 2011. Accent level adjustment in bilingual Thai-English text-to-speech synthesis. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 295–299. Waikoloa, HI, USA: IEEE.

Yamagishi, Junichi & Oliver Watts. 2010. The CSTR/EMIME HTS system for Blizzard challenge 2010. In *Proceedings Blizzard Workshop*, 1–6. Kansai Science City, Japan.

Yamagishi, Junichi, Heiga Zen, Yi-Jian Wu, Tomoki Toda & Keiichi Tokuda. 2008. The HTS-2008 System: Yet Another Evaluation of the Speaker-Adaptive HMM-based Speech Synthesis System in The 2008 Blizzard Challenge. In *Proceedings Blizzard Workshop*, 6. Brisbane, Australia.

Abbildungsverzeichnis

2.1 Überblick über die wichtigsten Sprachsynthesesysteme 17
2.2 Struktur eines HMM-Synthesesystems 20

3.1 Die bairischen Dialekträume in Österreich 25

4.1 Der südbairische Raum in Österreich mit den beiden Orten Innervillgraten (IVG) und Zelko (STY) 31
4.2 Überlappung der Phonsets 34
4.3 Phonsets des österreichischen Standarddeutsch (SAG), Innervillgratisch (IVG) und südweststeirische Varietät aus Zelko (STY) (Pucher et al. 2017a: 181) 35
4.4 Spektrogramm mit Grundfrequenzverlauf von STY_07 »Du bäckst mir morgen einen Kuchen« des STY-Sprechers. 40
4.5 Spektrogramm mit Grundfrequenzverlauf von STY_07 »Du bäckst mir morgen einen Kuchen« der synthetischen IVG-Stimme. 40
4.6 Ergebnisse des Mean-Opinion-Score-Tests (Pucher et al. 2017a: 184) . . 42

Tabellenverzeichnis

4.1 Synthesesätze mit dazugehöriger phonetischer Lautschrift (IPA) 33
4.2 Mapping. Links die STY-Laute und rechts die ihnen zugewiesenen IVG-Laute (Pucher et al. 2017a: 182) 38
4.3 Ergebnisse des Verständlichkeitstests der verschiedenen Methoden, angegeben mit der Word-Error-Rate in % (Pucher et al. 2017a: 183). 41

Abstract

Dialekte kodieren längst nicht mehr primär geographische Herkunft, sondern vielmehr auch soziale Identität. Diese Entwicklung kann zum Anlass genommen werden, neue Erkenntnisse der Dialektologie über diese hinaus in anderen Disziplinen wirken zu lassen. Die Dialektsynthese kann einen weiteren fruchtbaren Aspekt für die zeitgemäße Dialektologie darstellen, den sie als Chance wahrnehmen soll, sich als relevantes Forschungsfeld außerhalb der klassischen Sprachwissenschaft zu verorten und als wichtige Ergänzung in der Sprachtechnologie wahrgenommen zu werden.

Für die erfolgreiche Synthese von Sprache benötigt man qualitativ hochwertige Sprachkorpora, welche auf segmentaler Ebene bereits vorverarbeitet sind. So ist es erforderlich Sprachdaten auf phonetischer Ebene zu transkribieren und zu segmentieren. Die Zusammenstellung eines solchen Korpus ist ein sehr zeitintensives Unterfangen.

In dieser Arbeit wird eine einfache Phonmapping-Methode getestet, um die Synthese eines Dialekts zu realisieren, für den kein qualitativ hochwertiges Korpus verfügbar ist. Stattdessen wird ein bereits trainiertes akustisches Hidden-Markov-Modell (HMM) eines ähnlichen Dialekts verwendet um die Synthese durchzuführen. Als Zieldialekt wird hier das Südweststeirische aus der Ortschaft Zelko (STY) gewählt; der Quelldialekt Innervillgratisch aus Osttirol (IVG) wird zwar zur selben Dialektgruppe (Südbairisch) gezählt, befindet sich jedoch 225km entfernt. Beide Dialekte überschneiden sich hinsichtlich charakteristischen Phonen, wie etwa dem retroflexen Lateral []. Es besteht die Absicht die Ähnlichkeit auf Lautebene dieser Dialekte für die Realisierung eines authentischen synthetischen Outputs für STY zu instrumentalisieren.

Diese Arbeit erforscht die Anwendungsgebiete prosodiebasierter Ansätze für die Sprachsynthese anhand der zwei südbairischen Varietäten (Südweststeirisch und Innervillgratisch). Mit dem Transfer der Grundfrequenz und Dauer eines Originalsprechers wird gezeigt, dass ein methodischer Ansatz, der variationsspezifische Prosodie berücksichtigt, einen Vorteil für synthetisierten Sprachoutput bedeutet.