Synthese von akzentspezifischen Sprachmerkmalen mittels phonetischer Transkriptionen

Stefan Taubert

Fakultät für Informatik
Professur Medieninformatik
Technische Universität Chemnitz
09107 Chemnitz, Deutschland
stefan.taubert@informatik.tu-chemnitz.de

Zusammenfassung: Ziel der Dissertation ist es, die Synthese von Sachtexten mittels phonetischer Transkriptionen zu ermöglichen. Die Ausgabe soll dabei gleichzeitig natürlich und verständlich sein, wobei sich mehrere verschiedene akzentspezifische Sprachmerkmale unter gleichbleibender Stimmidentität synthetisieren lassen. Der Fokus liegt auf denjenigen Sprachmerkmalen, welche sich durch die Silbenbetonung, die Lautdauer, die Satzbetonung und die Pausensetzung festlegen lassen. Hierbei wird eine bestehende neuronale Netzarchitektur um mehrere Komponenten erweitert und in Experimenten der Erfolg der jeweiligen Erweiterungen perzeptiv und objektiv evaluiert. Das Framework beinhaltet zudem eine Weboberfläche, in welcher sich in IPA transkribierte Eingabetexte mit verschiedenen Stimmen zu Sprache synthetisieren lassen. Linguisten und Psychologen sollen diese benutzen können, um beispielsweise Auswirkungen von verschiedenen Sprachmerkmalen auf die von Studienteilnehmern wahrgenommene Glaubwürdigkeit, Verlässlichkeit oder Kompetenz eines Sprechers (d.h. dessen synthetisierte Stimme) zu untersuchen.

Schlagwörter: Text-to-speech, Linguistik, IPA, Sprachverarbeitung

1 Einleitung

Text-zu-Sprache (engl. Text-to-Speech, TTS), auch Sprachsynthese (engl. speech synthesis, SS) genannt, bezeichnet die Umwandlung von Text in synthetische Sprache mittels Computer [O'M90, S. 17] [Dut97, S. 13] [Tay09, S. 1] [TQSL21, S. 1] und wird beispielsweise zum Lernen von Sprachen [Sha10], im e-Learning [ZGMC09], zur Unterstützung von sehbehinderten Menschen [dRP+10], als Sprachausgabe bei Computern und Handys (siehe Apple Siri, Amazon Alexa, etc.), für die Synthese von Hörbüchern oder in digitalen Wörterbüchern eingesetzt [Moo12].

1.1 Sonderforschungsbereich „Hybrid Societies"

Die Arbeit erfolgt im Rahmen des Sonderforschungsbereiches (SFB) „Hybrid Societies: Humans Interacting with Embodied Technologies" der Technischen Universität Chem-

nitz, dessen Ziel es ist, die Koordination von Menschen und verkörperten digitalen Technologien (engl. embodied digital technologies, EDTs) zu verbessern. Der SFB ist aufgeteilt in die vier Bereiche „Embodied Sensor and Motor Capabilities“, „Artificial Bodies“, „Shared Environments“ und „Intentionality in Hybrid Societies“. Die Arbeit ist hierbei dem letzten Bereich zugeordnet, welcher sich auf die Zuschreibung und Kommunikation von situationsspezifischen Intentionen zwischen Menschen und EDTs konzentriert. Der Fokus liegt in der Entwicklung eines pedagogischen Gesprächsagenten (engl. conversational pedagogical agent, CPA) für chinesisches Englisch. Agenten sind animierte Charaktere, welche mit Lernenden in einer computerbasierten interaktiven Lernumgebung interagieren und dadurch den Lernprozess stimulieren und fördern sollen [JRL00,Joh01]. Da die Erstellung eines solchen Agenten eine interdisziplinäre Aufgabe ist, sind ebenfalls ein Linguistiker und eine Psychologin im Projekt eingebunden. Ziel des Projektes ist es, Lernende mit chinesischer Muttersprache an deutschen Universitäten zu unterstützen, da sie den größten Anteil an ausländischen Studierenden in Deutschland ausmachen. In vergangenen Studien konnte gezeigt werden, dass sich der Lernerfolg bei Lernenden durch akzentuierte Sprache erhöhen lässt [RS13], da diese zu der Zuschreibung von mehr sozialer und fachlicher Kompetenz führt. In einer psychologischen Studie soll untersucht werden, welche chinesischen Sprachmerkmale den meisten Einfluss auf den Lernerfolg haben. Um diese Sprachmerkmale in der synthetisierten Ausgabestimme einzeln anpassen zu können, ist ein entsprechendes Synthesesystem notwendig, welches in der Dissertation erstellt werden soll.

2 Forschungslücke

Nachfolgend wird die Arbeit in das akademische Umfeld eingeordnet und auf die zentrale Forschungsfrage eingegangen.

2.1 Einordnung der Arbeit in das Akademisches Umfeld

Die Arbeit lässt sich in das akademische Feld der Sprachverarbeitung einordnen, welche wiederum ein Gebiet der Signalverarbeitung ist. Letztere ist ein Teilbereich der Elektrotechnik, der sich auf die Analyse, Bearbeitung und Synthese von Signalen fokussiert, wobei sich die Sprachverarbeitung mit der Verarbeitung von Sprachsignalen beschäftigt.

Zur Sprachverarbeitung gehören unter anderem die Forschungsbereiche:

- Sprachsynthese (engl. speech synthesis, SS),
- Sprechererkennung (engl. speaker recognition),
- Sprecher-Diarisierung (engl. speaker diarization, SD),
- Stimmumwandlung (engl. voice conversion),

- Stimmklonen (engl. voice cloning),
- Spracherkennung (engl. automatic speech recognition, ASR),
- Sprachidentifikation (engl. language identification, LID),
- Prosodieerkennung (engl. prosody prediction) und
- Emotionserkennung (engl. emotion detection).

Den größten Anteil der Arbeit hat hierbei die Sprachsynthese, weitere Bereiche wie die Sprachumwandlung und das Sprachklonen sind jedoch ebenfalls relevant und sollen im weiteren Verlauf kurz eingeführt werden.

In der Sprachsynthese-Forschung besteht das Hauptziel darin, den Unterschied zwischen generierten und menschlichen Aufnahmen zu verringern. Die Evaluation von generierter Sprache erfolgt immer in Hinblick auf die Natürlichkeit (engl. naturalness) und Verständlichkeit (engl. intelligibility) [Tay09, SMK+17]. Die Natürlichkeit beschreibt den Grad an Menschlichkeit der Stimme und die Verständlichkeit bewertet wie gut der Inhalt des synthetisierten Textes verstanden wird.

Die Sprachsynthese-Forschung spezialisiert sich in weitere Gebiete, welche sich grob in Synthesesystemerweiterungen und -limitierungen gruppieren lassen. Zu ersteren gehören beispielsweise die Verbesserung der Robustheit der Systeme gegenüber einer breiten Masse an Eingabetexten, die Erweiterung der Möglichkeiten, die Ausgabe hinsichtlich Geschwindigkeit, Rhythmus, Tonlage, Stimmung, Sprachfluss, etc. zu beeinflussen, und die Fähigkeit mehrere Sprecher (engl. multi-speaker), Sprachen (engl. multi-language) oder Akzente innerhalb eines Systems synthetisieren zu können. Alle der Bereiche sind für das Synthesesystem der Arbeit relevant. Zu Gebieten, die sich mit der Limitation von Systemen in der Erstellung oder Synthese beschäftigen, gehören u.a. die Limitation der Datenmenge (engl. low-resource), Rechenleistung (z. B. mobiler Einsatz), Systemarchitektur oder die Begrenzung der Synthesedauer (z. B. um Echtzeitsynthesen zu ermöglichen). Um möglichst viele verschiedene Sprecher im Repertoire des finalen Systems zu haben, sollten die für die Synthese benötigten Trainingsdaten so gering wie möglich sein, um Aufnahme- und Transkriptionskosten zu senken. Die Synthesegeschwindigkeit und die Menge an benötigten Rechenressourcen spielen hingegen eine untergeordnete Rolle, da die Synthese nicht live und mobil erfolgen soll.

Die Stimmumwandlung befasst sich mit Methoden existierende Sprachaufnahmen mit einer anderen Stimmidentität zu versehen [ACP+18]. Zur Stimmidentität zählen Charakteristiken wie Tonhöhe, Geschlecht oder Artikulationsrate, jedoch keine linguistischen Informationen wie verwendete Phones. Stimmklonen hingegen beschäftigt sich mit dem Lernen einer Stimme von einem ungesehenen Sprecher anhand weniger Sprachbeispiele, um ungesehene Texte mit dessen Stimmidentität zu synthetisieren [ACP+18]. Ein Untergebiet der Stimmumwandlung ist die Akzentumwandlung (engl. accent conversion, AC; foreign accent conversion, FAC), welche darauf abzielt, eine neue Stimme zu schaffen,

die die stimmliche Identität eines bestimmten nichtmuttersprachlichen (L2) Sprechers hat, dabei aber den muttersprachlichen (L1) Akzent beibehält [DZG22, ZDG19, LG22]. Sprachenlernende können mit Hilfe solcher Systeme die muttersprachliche Betonung von Wörtern in ihrer eigenen Stimme anhören und die Aussprache besser verinnerlichen [FBG09]. Diese Akzentimitation wird auch im Stimmklonen untersucht. Da in der Arbeit verschiedene akzentspezifische Sprachmerkmale mit der gleichen Stimmidentität synthetisiert werden sollen, diese Merkmale jedoch von unterschiedlichen Sprechern kommen können, muss es dem Synthesesystem möglich sein, Sprecheridentität und linguistische Merkmale voneinander zu trennen. Hierfür werden Methoden des Stimmklonens und -umwandlung eingesetzt.

2.2 Forschungsfrage

Die zentrale Forschungsfrage der Dissertation ist die folgende: *Wie lassen sich Sachtexte gleichzeitig natürlich, verständlich und gezielt mit mehreren verschiedenen akzentspezifischen Sprachmerkmalen unter gleichbleibender Stimmidentität synthetisieren?*

Im Nachfolgenden sollen alle relevanten Teile der Frage genauer untersucht werden:

- *„Sachtexte“*: Zu Sachtexten gehören expositorische, sachinformierende und normative Texte, welche u. a. keine direkte Rede beinhalten. Die Synthese von poetischen Texten, Liedern, auffordernden Texten oder spontaner Rede soll nicht Ziel der Arbeit sein. Da bei den Zieltexten auch Schachtelsätze vorkommen können, ist es notwendig, dass die Satzbetonung für verschiedene Satzlängen korrekt gesetzt wird, z. B. ein sinkender Tonhöhenverlauf vor einem Komma oder Punkt.
- *„synthetisieren“*: Der Fokus der Arbeit soll ausschließlich auf der Synthese von Text bestehen. Das automatische Erkennen von linguistischen Sprachmerkmalen anhand von transkribierten Aufnahmen und das automatische Anwenden von solchen Merkmalen auf Texte ist nicht vorgesehen. Die Eingabe in das Synthesesystem ist bereits mit allen Merkmalen versehen, das System muss diese unterstützen. Das heißt auch, dass die Eingabe bereits normalisiert ist, also z. B. Zahlen ausgeschrieben sind, und in IPA [The99] transkribiert wurde.
- *„natürlich“*: Die Ausgabestimme des Systems muss sich möglichst menschlich anhören.
- *„verständlich“*: Der Originaltext der Systemausgabe muss erkennbar sein.
- *„gezielt“*: Gezielt bedeutet in dem Kontext, dass der Akzent nicht 1:1 entsprechend einer Referenzstimme nachgebildet wird, sondern, sich gezielte Akzentmerkmale in der Ausgabe setzen lassen.
- *„akzentspezifische Sprachmerkmale“*: Es sollen vorwiegend akzentspezifische Sprachmerkmale anpassbar sein. Das betrifft die Kontrolle über die Grammatik, das Voka-

bular, die Vokalbetonung, den lexikalischen Ton, die Lautdauer und die Pausensetzung. Nichtmuttersprachler produzieren beispielsweise aufgrund von Unsicherheiten und Überlegungen zur Aussprache einzelner Wörter viel häufiger Pausen als Muttersprachler [AH98, S. 29]. Das Wiedergeben von Emotionen oder das Nachbilden von Gesang ist nicht vorgesehen. Auch bestimmte paralinguistische Merkmale wie das Alter des Sprechers oder stimmqualitative Eigenschaften wie die Rauhheit der Stimme sollen nicht explizit anpassbar sein.
- *„mehrere verschiedene“*: Um neue authentische Akzente generieren zu können, muss eine Kombination von mehreren Merkmalen in der Ausgabestimme möglich sein.
- *„gleichzeitig“*: Die drei Anforderungen: Natürlichkeit, Verständlichkeit und das Vorhandensein der gesetzten Sprachmerkmale müssen für ein erfolgreiches Synthesesystem gleichzeitig in der Ausgabe erfüllt sein.
- *„gleichbleibende Stimmidentität“*: Unabhängig davon, welche Sprachmerkmale gesetzt wurden, soll die Stimme immer die gleiche Identität haben. Dies ist besonders wichtig, um die Wirkung von verschiedenen Akzenten ohne den Einfluss der Stimmpräferenz bei Studienteilnehmern messen zu können.

Aus dieser Forschungsfrage lassen sich für die einzelnen Sprachmerkmale folgende Hypothesen formulieren:

1. **Silbenbetonung**: Einzelne Silben lassen sich mit gleicher Stimmidentität natürlich, verständlich und gezielt unbetont, hauptbetont und nebenbetont oder mit den Tonhöhenverlaufsstufen steigend, fallend, hoch steigend, tief steigend und steigend-fallend synthetisieren.
2. **Lautdauer**: Einzelne Laute lassen sich mit gleicher Stimmidentität natürlich, verständlich und in den Stufen lang, halblang, normal und sehr kurz synthetisieren.
3. **Pausensetzung**: Texte lassen sich mit gleicher Stimmidentität natürlich, verständlich und mit keinen, kleinen mittleren und großen Pausen zwischen Wörtern synthetisieren. Außerdem lassen sich Wörter gebunden synthetisieren (Liaison).

Aus der Anforderung Sachtexte synthetisieren zu können, lässt sich folgende Hypothese ableiten:

4. **Satzbetonung**: Sätze lassen sich mit gleicher Stimmidentität natürlich, verständlich und mit einem Tonhöhenverlauf in den Stufen normal, steigend und fallend synthetisieren. Dabei lassen sich für einen Satz gleichzeitig mehrere Verläufe an verschiedenen Positionen festlegen.

Da eine Synthese von mehreren Merkmalen gleichzeitig notwendig ist, ergibt sich letzte die Hypothese:

5. **Kombination**: Texte lassen sich mit gleicher Stimmidentität natürlich, verständlich und unterschiedlich anpassbarer Silbenbetonung, Lautdauer, Pausensetzung und Satzbetonung synthetisierten.

Per IPA lässt sich direkt anpassen welche Phoneme wann gesprochen werden sollen. Darin enthalten ist die Anpassung der Grammatik des Satzes und das verwendete Vokabular. Deshalb wurden für die Phoneme-, Grammatik- und Vokabularanpassung keine separaten Hypothesen formuliert.

Die folgenden Aspekte von Akzentmerkmalen wurden ausgelassen, da sie entweder zu spezifisch sind oder den Rahmen der Arbeit sprengen würden:

- Artikulationsrate: Die Artikulationsrate lässt sich zwar leicht durch Time-Stretching anpassen, diese Anpassung entspricht jedoch nicht dem natürlichen Sprachverhalten, in welchem beispielsweise Vokale bei einer Verringerung der Artikulationsrate stärker verlängert werden als Konsonanten [Kuw96, MTAL97] [BBJ+20, S. 4402].
- Diakritika: stimmlose, stimmhafte, behauchte, etc. Betonung
- Tonhebenen: extrahoch, hoch, mittel, tief, extratief
- Tonkontur: Downstep, Upstep

3 Methode

Moderne Synthesesysteme bestehen aus den drei Komponenten Textanalysemodul, akustisches Modell und Vokoder (s. Abb. 1) [TQSL21]. Ersteres wandelt eine Textsequenz in linguistische Features um, das akustische Modell generiert aus diesen akustische Features und der Vokoder synthetisiert ein Audiosignal aus den akustisches Features.

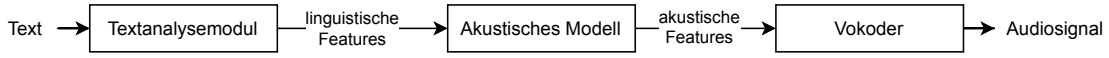

Abb. 1: Komponenten neuronale Sprachsynthese

Da das Synthesesystem der Dissertation bereits linguistische Features in Form von IPA Transkriptionen als Eingabe erhält, wird diese Komponente nicht benötigt. Der Vokoder wird zur Synthese benötigt (s. 3.2), Anpassungen an diesem sind jedoch nicht notwendig, da ausschließlich das akustische Modell für die Transformation aller akzentspezifischen Merkmale in akustische Features verantwortlich ist. Daher liegt der Fokus der Arbeit klar in der Entwicklung diesen akustischen Modells (s. 3.1). Dazu wird ein bestehendes Modell um die benötigten Funktionen erweitert und es wird anschließend in Experimenten evaluiert, wie erfolgreich die vorgenommenen Erweiterungen waren (s. 3.3). Am Ende

soll eine Weboberfläche geschaffen werden, in welcher sich mit Hilfe des entstandenen Synthesesystems in IPA transkribierte Texte synthetisieren lassen (s. 3.4).

3.1 Akustisches Modell

Das akustische Modell wird basierend auf dem State-of-the-Art neuralen Modell Tacotron [WSS+17, SPW+18] entwickelt. Dieses wird um die Funktionen erweitert, die Silbenbetonung, die Lautdauer, die Pausensetzung und die Satzbetonung zu lernen und gezielt anpassen zu können. Die Anpassung soll hierbei direkt durch den Eingabetext (IPA) möglich sein. Geplant ist der Modellarchitektur verschiedene Embeddings für Sprache, Akzent, Silbenbetonung und Lautdauer hinzuzufügen. Es wird bewusst eine neurale Synthesearchitektur verwendet, da diese im Vergleich zur zuvor vorwiegend eingesetzten parametrischen Synthese mehr Natürlichkeit erzielen kann [WEHFV19, S. 1] [TQSL21, S. 8].

Für das Training des Modells sollen mehrere Sprachdatensätze bestehend aus Aufnahmen in unterschiedlichen Sprachen zum Einsatz kommen. Das Ziel ist es, eine möglichst große Menge an Vokalen und Konsonanten synthetisieren zu können, weshalb Aufnahmen von mehreren Sprachen kombiniert werden müssen.

3.2 Vokoder

Für die Transformation der akustischen Features in das Audiosignal wird der State-of-the-Art Vokoder WaveGlow [PVC19] eingesetzt. Architekturelle Anpassungen müssen an diesem nicht vollzogen werden. Die öffentliche Implementation [Nvi18], inkl. einem von Nvidia vortrainierten Modell [Nvi19] werden für die Synthese verwendet.

3.3 Evaluation

Evaluiert werden sollen Natürlichkeit, Verständlichkeit und das Vorhandensein der Akzentmerkmale in der synthetisierten Sprache.

Für die Evaluation der Natürlichkeit wird üblicherweise der Mean Opinion Score (MOS) berechnet [WSS+17, SPW+18, SWC+20, XTR+20]. Dieser soll auch in der Dissertation als Metrik eingesetzt werden. Er wird berechnet, indem Muttersprachler synthetisierte und originale Sprachäußerungen auf einer 5-stufigen Likert-Skala in 0,5er Schritten bewerten. Die Auswertung erfolgt mit einem Konfidenzinterval von 95%. Dabei sollte eine Äußerung von so vielen verschiedenen Sprechern wie möglich bewertet werden. Die Verständlichkeit lässt sich mit Hilfe der Verständlichkeitsrate messen, welche das Verhältnis an verständlichen Worten zu der Gesamtmenge an Worten angibt [XTR+20]. Dazu

werden synthetisierte Sprachäußerungen von Muttersprachlern bewertet, indem diese die Worte angeben, die verständlich waren. Diese Metrik soll bei der Evaluation des Synthesesystems ebenfalls zum Einsatz kommen. Zusätzlich soll die Mel-Cepstral Distanz (MCD) berechnet werden [Kub93], welche die originalen und synthetisierten Spektrogramme miteinander vergleicht.

Die akzentspezifischen Merkmale müssen separat evaluiert werden. Geplant ist dafür eine linguistische Transkription der Syntheseergebnisse und ein anschließender Vergleich mit der Ground Truth mittels Berechnung einer Phonfehlerrate in Anlehnung an die Wortfehlerrate (engl. word error rate, WER), bezogen auf einzelne Phone anstelle von Wörtern. In verbundener Sprache berechnet sich die Wortfehlerrate durch die Summe der Anzahl an Wortersetzungen, -löschungen und -einfügungen dividiert durch die Anzahl an Eingabewörtern [MMG04, S. 2] [BJ75, Moo77]. Zur Evaluation der Lautdauer und Pausensetzung soll die durchschnittliche und mittlere Dauerdifferenz anhand der Transkription berechnet werden.

3.4 GUI

Im Rahmen der Arbeit soll eine auf Angular basierende Weboberfläche (GUI) entwickelt werden, in welcher sich in IPA transkribierter Text zu Sprache synthetisieren lässt. In Abbildung 2 ist der geplante Aufbau der GUI dargestellt (Top-down nach EVA-Prinzip konstruiert).

In der Oberfläche kann ein Text jeweils in normaler Schreibweise und transkribiert in IPA angegeben werden. Der normale Text dient zur besseren Orientierung im IPA Text, welcher wiederum ausschlaggebend für die Synthese ist. Als Eingabehilfe ist eine Tabelle an unterstützten Symbolen vorgesehen, mit welcher Symbole an der aktuellen Cursorposition eingefügt werden können. Vor der Synthese können verschiedene Parameter eingestellt werden, z.B. Stimme, Sprache, Akzent und Pausenlängen. Zudem lässt sich kontrollieren, ob der gesamte Text synthetisiert werden soll oder nur ein bestimmter Satz. Dadurch ist es möglich die Transkription einzelner Sätze schneller zu optimieren. Die resultierende Audiodatei kann im Browser angehört werden und lässt sich bei Bedarf herunterladen. Alle Eingabefelder, inkl. Titel des Projektes lassen sich in einer Datei speichern und aus dieser laden, um den Arbeitsstand zu sichern.

Um den Open-Science Maßstäben gerecht zu werden und Reproduzierbarkeit zu ermöglichen, erfolgt eine Veröffentlichung der Anwendungen auf GitHub, Zenodo[1] und dem Python Package Index[2] (PyPI), dem offiziellen Repository von Drittanbietersoftware für Python.

[1] `https://zenodo.org`
[2] `https://pypi.org`

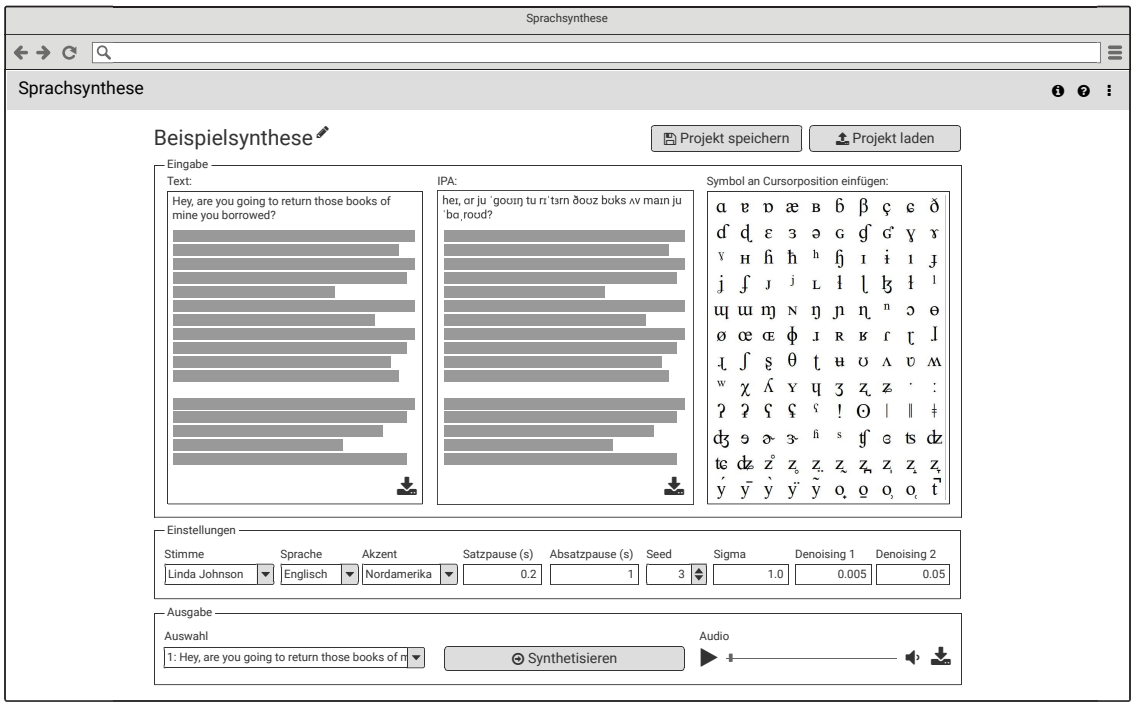

Abb. 2: Geplanter Aufbau der GUI

4 Erste Ergebnisse

In dieser Sektion werden durchgeführte Experimente und Studien kurz vorgestellt und eine Übersicht über bisher entwickelte Software gegeben.

4.1 Experimente

In bisherigen Experimenten konnte gezeigt werden, dass der Einsatz von Textselektionsalgorithmen die zur Erstellung eines Synthesesystems benötigte Datenmenge reduzieren kann. Verglichen wurde eine zufällige Auswahl und zwei gierige Algorithmen. Bei einem der Algorithmen wurden unter Verwendung der Kullback-Leibler Distanz [CT91] iterativ Sätze ausgewählt, sodass am Ende eine Gleichverteilung aller Phoneme des Skripts angenähert wurde.

Weitere Experimente konnten zeigen, dass es mit Hilfe von Transferlernen möglich ist, ein Synthesesystem anhand von nur 30 Minuten an Aufnahmen zu erstellen. Dabei wird ein Modell auf einen großen Datensatz mit rund 24 Stunden vortrainiert und im Anschluss auf einen 30-minütigen Datensatz adaptiert.

In einer in Italien durchgeführte Studie wurde eine roboterartige und eine natürlichklingende Stimme von den Teilnehmern anhand von 15 Sätzen bewertet. Die roboterartige Stimme war ähnlich verständlich wie die natürliche Stimme, allerdings wurde sie erwartungsgemäß deutlich weniger menschlich, ansprechend, sympathisch, ausdrucksstark, kompetent, vertrauenswürdig, zuverlässig und glaubwürdig wahrgenommen.

In einem anderen Teil dieser Studie wurde ebenfalls evaluiert, ob ein synthetisierter Lerntext mit nordamerikanischen oder italienischen Akzent besser verstanden wurde. Dazu wurden zwei Lerntexte jeweils in den beiden Akzenten synthetisiert und den Studienteilnehmern vorgespielt, welche dann Fragen dazu beantworten sollten. Die eine Hälfte bekam Text 1 in nordamerikanischen Englisch und Text 2 in italienischen Englisch, die andere Hälfte bekam die Texte im jeweils anderen Akzent. Von den 45 Teilnehmern, schlossen nur 13 den Fragebogen ab, daher ist die Anzahl zu gering, um statistisch signifikante Aussagen zu treffen. Die Ergebnisse deuten jedoch darauf hin, dass die Wahrnehmung von Merkmalen wie Wärme, Aufrichtigkeit und Glaubwürdigkeit des italienischen Akzents eher neutral eingeschätzt wurde. Den nordamerikanischen Akzent bewerteten die Teilnehmer als sehr zustimmungsfähig. Beide Akzente wurden als glaubwürdig, authentisch und genau eingeschätzt. Die interessantesten Ergebnisse waren, dass die Teilnehmer bei dem Text mit italienischen Akzent ausführlicher und auch bei mehr Fragen antworteten, obwohl sie den Akzent eher neutral einstuften. Dies könnte daran liegen, dass die Teilnehmer an den nichtmuttersprachliche Akzent besser gewöhnt sind und sich deshalb leichter an die Informationen erinnern können.

In einem anderen Experiment wurde untersucht, ob die Betonung der Vokale im Satz gezielt angepasst werden kann. Dafür wurde ein aktuelles Synthesesystem um die Funktion erweitert, die Betonung separat von den Phonemen zu lernen. Das Training war erfolgreich, da sich Vokale gezielt unbetont, primär betont und sekundär betont synthetisieren ließen.

Um die Pausen besser zu kontrollieren, wurden neben den Phonemen auch die Pausen transkribiert. Festgelegt wurden eine kurze Pause (≤ 0.1 Sekunden) und eine lange Pause (> 0.1 Sekunden). Das Lernen der Pausen war erfolgreich und führte zu einer verbesserten Kontrolle der Aussprache. Es wurde zudem überprüft, ob ein Worttrenner zwischen aufeinanderfolgenden Wörtern, bei denen keine Pause gesprochen wurde (Leerzeichen), für die korrekte Aussprache notwendig war. Es zeigte sich, dass die Aussprache ohne Worttrenner deutlich schlechter wurde, da Wörter zusammengesprochen wurden, welche nicht zusammengesprochen werden sollten.

Es sind weitere Experimente geplant, in welchen die Lernfähigkeit der Lautdauer und Satzbetonung untersucht werden soll.

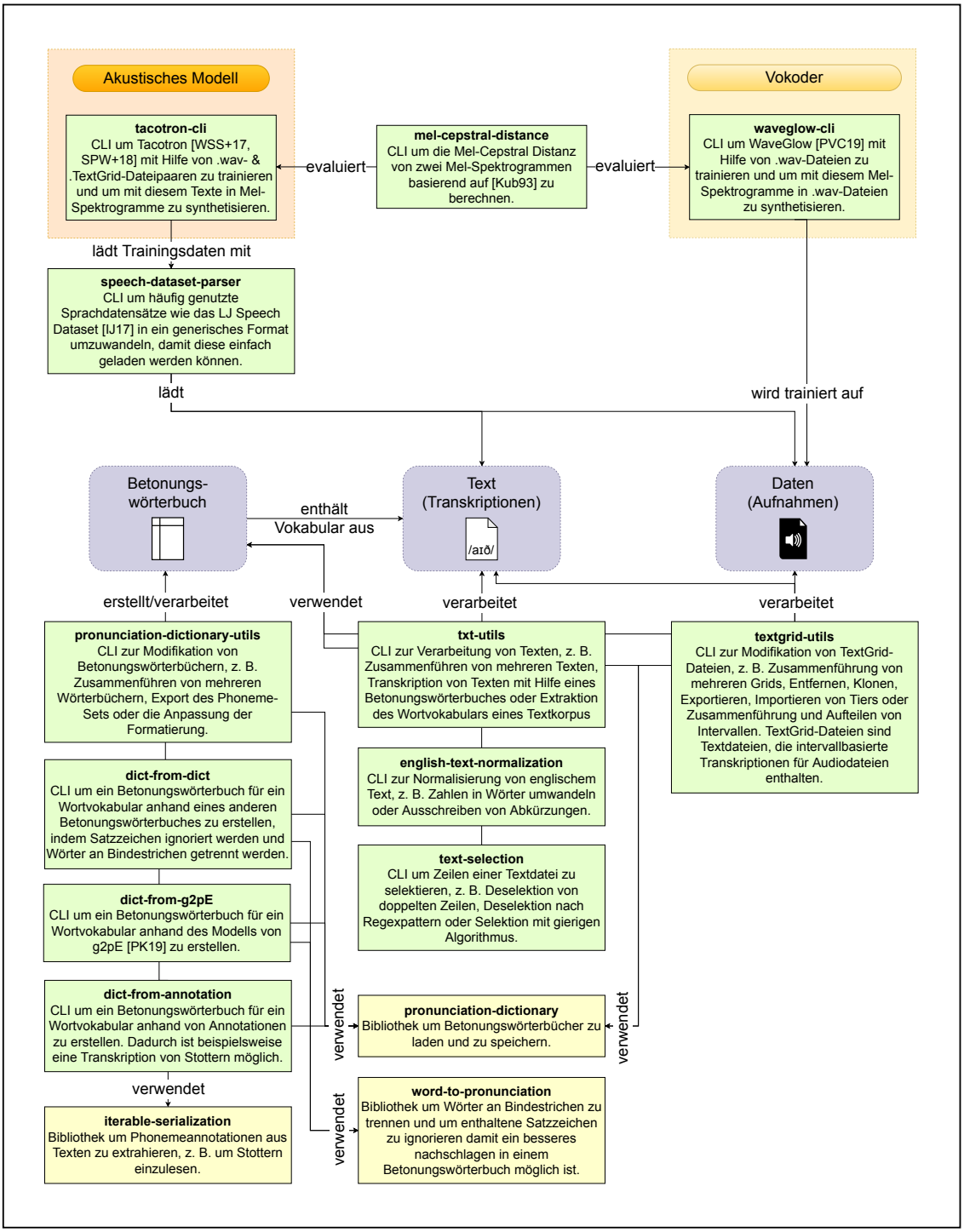

Abb. 3: Überblick über die bisher entwickelten Pythonprogramme (grün) und -bibliotheken (gelb) inkl. ihrem Verwendungszweck (lila).

4.2 Entwickelte Software

Es wurden mehrere Kommandozeileninterfaces (engl. command-line interface, CLI) basierend auf Python entwickelt und auf GitHub[3] und PyPI veröffentlicht[4]. In Abb. 3 sind die entsprechenden Programme (grün) und Bibliotheken (gelb) inklusive ihrem Verwendungszweck (lila) dargestellt.

Danksagung

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 416228727 – SFB 1410

Literatur

[ACP+18] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural Voice Cloning with a Few Samples. page 11, 2018.

[AH98] Levent M. Arslan and John H. L. Hansen. A study of temporal features and frequency characteristics in American English foreign accent. *The Journal of the Acoustical Society of America*, 102(1):28, August 1998.

[BBJ+20] Jae-Sung Bae, Hanbin Bae, Young-Sun Joo, Junmo Lee, Gyeong-Hoon Lee, and Hoon-Young Cho. Speaking Speed Control of End-to-End Speech Synthesis Using Sentence-Level Conditioning. In *Interspeech 2020*, pages 4402–4406. ISCA, October 2020.

[BJ75] L. Bahl and F. Jelinek. Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory*, 21(4):404–411, July 1975.

[CT91] T. M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley series in telecommunications. Wiley, New York, 1991.

[dRP+10] J. M. H. du Buf, J. M. F. Rodrigues, Hugo Paredes, João Barroso, Miguel Farrajota, João José, Victor Teixeira, and Mário Saleiro. The SmartVision navigation prototype for the blind. 2010.

[Dut97] Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*, volume 3 of *Text, Speech and Language Technology*. Springer Netherlands, Dordrecht, 1997.

[3] `https://github.com/stefantaubert`
[4] `https://pypi.org/user/stefantaubert`

[DZG22] Shaojin Ding, Guanlong Zhao, and Ricardo Gutierrez-Osuna. Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech & Language*, 72:101302, March 2022.

[FBG09] Daniel Felps, Heather Bortfeld, and Ricardo Gutierrez-Osuna. Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*, 51(10):920–932, October 2009.

[IJ17] Keith Ito and Linda Johnson. The LJ Speech Dataset, 2017.

[Joh01] W. Lewis Johnson. Pedagogical Agent Research at CARTE. *AI Magazine*, 22(4):85–85, December 2001.

[JRL00] W. Lewis Johnson, Jeff W. Rickel, and James C. Lester. Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of Artificial InTELligence in Education*, 11:47–78, 2000.

[Kub93] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128, Victoria, BC, Canada, 1993. IEEE.

[Kuw96] H. Kuwabara. Acoustic properties of phonemes in continuous speech for different speaking rate. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 4, pages 2435–2438, October 1996.

[LG22] Christopher Liberatore and Ricardo Gutierrez-Osuna. Minimizing Residuals for Native-Nonnative Voice Conversion in a Sparse, Anchor-Based Representation of Speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7002–7006, May 2022.

[MMG04] Andrew Cameron Morris, Viktoria Maier, and Phil Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768. ISCA, October 2004.

[Moo77] R. Moore. Evaluating speech recognizers. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(2):178–183, April 1977.

[Moo12] Dosik Moon. Web-Based Text-to-Speech Technologies in Foreign Language Learning: Opportunities and Challenges. In Tai-hoon Kim, Jianhua Ma, Wai-chi Fang, Yanchun Zhang, and Alfredo Cuzzocrea, editors,

Computer Applications for Database, Education, and Ubiquitous Computing, Communications in Computer and Information Science, pages 120–125, Berlin, Heidelberg, 2012. Springer.

[MTAL97] Fernando Martinez, Daniel Tapias, Jorge Alvarez, and Paloma Leon. Characteristics of Slow, Average and Fast Speech and Their Effects in Large Vocabulary Continuous Speech Recognition. In *EUROSPEECH-1997*, pages 469–472, 1997.

[Nvi18] Nvidia. WaveGlow: a Flow-based Generative Network for Speech Synthesis. Nvidia Corporation, 2018.

[Nvi19] Nvidia. WaveGlow LJS 256 channels. `https://catalog.ngc.nvidia.com/orgs/nvidia/models/waveglow_ljs_256channels`, 2019.

[O'M90] M.H. O'Malley. Text-to-speech conversion technology. *Computer*, 23(8):17–23, August 1990.

[PK19] Kyubyong Park and Jongseok Kim. g2pE. `https://github.com/Kyubyong/g2p`, 2019.

[PVC19] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: A Flow-based Generative Network for Speech Synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621, Brighton, United Kingdom, May 2019. IEEE.

[RS13] Günter Daniel Rey and Nadine Steib. The personalization effect in multimedia learning: The influence of dialect. *Computers in Human Behavior*, 29(5):2022–2028, 2013.

[Sha10] Guoquan Sha. Using TTS voices to develop audio materials for listening comprehension: A digital approach. *British Journal of Educational Technology*, 41(4):632–641, 2010.

[SMK+17] Jose M. R. Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron C. Courville, and Yoshua Bengio. Char2Wav: End-to-End Speech Synthesis. In *ICLR*, 2017.

[SPW+18] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, April 2018.

[SWC+20] Aolan Sun, Jianzong Wang, Ning Cheng, Huayi Peng, Zhen Zeng, and Jing Xiao. GraphTTS: Graph-to-Sequence Modelling in Neural Text-to-Speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6719–6723, May 2020.

[Tay09] Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[The99] The International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, U.K., 1999.

[TQSL21] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A Survey on Neural Speech Synthesis, July 2021.

[WEHFV19] Oliver Watts, Gustav Eje Henter, Jason Fong, and Cassia Valentini-Botinhao. Where do the improvements come from in sequence-to-sequence neural TTS? In *10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 217–222. ISCA, September 2019.

[WSS+17] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. In *Interspeech 2017*, pages 4006–4010. ISCA, August 2017.

[XTR+20] Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812, Virtual Event CA USA, August 2020. ACM.

[ZDG19] Guanlong Zhao, Shaojin Ding, and Ricardo Gutierrez-Osuna. Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *Interspeech 2019*, pages 2843–2847. ISCA, September 2019.

[ZGMC09] Wu Zhiyong, Cao Guangqi, M. Helen Meng, and Lianhong Cai. A unified framework for multilingual text-to-speech synthesis with SSML specification as interface. *Tsinghua Science and Technology*, 14(5):623–630, October 2009.