Ein Toolkit zur Erstellung von Sprachkorpora

Hans Kruschke, Uwe Koloska, Guntram Strecha, Matthias Eichner, Diane Hirschfeld, Ulrich Kordon
Technische Universität Dresden, Institut für Akustik und Sprachkommunikation

Für alle Disziplinen der Sprachverarbeitung rücken große, nach verschiedenen Kriterien etikettierte Korpora in den Mittelpunkt der Forschungsarbeit. Die Erstellung und Aufbereitung von Sprachkorpora ist jedoch mit erheblichem Aufwand verbunden. Die ständig wachsenden Datenmengen und die stark abweichenden Zielrichtungen sind manuell nicht mehr beherrschbar. Das Dresdner Corpus-Kit, das vornehmlich für Arbeiten in der Sprachsynthese konzipiert ist, bietet deshalb komfortable Unterstützung bei der Korpusdefinition, Aufnahme, Aufbereitung, Etikettierung und Optimierung. Seine modulare Architektur ist offen für Erweiterungen und Verbesserungen und bietet für verschiedene Bearbeitungsschritte alternative Werkzeuge. Es werden zentrale Komponenten des Toolkits vorgestellt und ein Einblick in die verschiedenen Bearbeitungsstufen des Sprachmaterials gegeben.

1 Einleitung

Alle Gebiete der Sprachverarbeitung benötigen als Datenbasis nach verschiedenen Gesichtspunkten konzipierte und erstellte Sprachkorpora. So erfordern z.B. automatische Lernverfahren zur Prosodiegenerierung große Mengen natürlich gesprochener Äußerungen. Derzeitige Sprachsynthesesysteme basieren auf der Verkettung von Bausteinen natürlich gesprochener Sprachaufnahmen.
Um den ständig wachsenden Anforderungen an die Sprachkorpora gerecht zu werden, wird eine Automatisierung der zur Korpuserstellung erforderlichen Arbeitsschritte angestrebt. Konzept des Dresdner Corpus-Kit ist, zu diesem Zweck die im Rahmen verschiedener Aufgabenstellungen entstandenen Werkzeuge zu vereinen und für neue Anforderungen nutzbar zu machen. Im Mittelpunkt des Korpus stehen die Sprachdaten. Bei der Erstellung des Korpus ist deshalb zu planen, welche Daten wie aufgenommen werden sollen. Nach erfolgter Aufnahme sind die Sprachdaten aufzubereiten und zu etikettieren. Je nach Verwendungszweck erfolgt eine unterschiedliche weitere Verarbeitung der Daten. Der daraus resultierende Ablauf der Korpuserstellung ist in Abbildung 1 dargestellt. Für jeden Arbeitsschritt gibt es im Dresdner Corpus-Kit ein oder mehrere alternative Programmwerkzeuge. Die Schnittstellen sind einfach gestaltet. Erforderlichenfalls gibt es Konverter zur Umwandlung der Datenformate. Durch seinen einfachen modularen Aufbau ist das System sehr offen gestaltet und bietet deshalb den Vorteil, daß leistungsfähige spezialisierte Werkzeuge anderer Arbeiten leicht integrieren zu können.
Nachfolgend wird ein Überblick über die wichtigsten Arbeitsschritte der Korpuserstellung und -nutzung und insbesondere der dafür im Dresdner Corpus-Kit vorhandenen Werkzeuge gegeben.

2 Korpusdefinition

Die Festlegung der Texte und des Aufnahmeszenarios erfolgt in Abhängigkeit vom Verwendungszweck des Korpus. Für die Erstellung eines Syntheseinventars muß der Korpus beispielsweise alle benötigten Inventarbausteine, nach Möglichkeit noch in verschiedenen prosodischen bzw. kontextualen Varianten, enthalten. Dazu wird eine Bausteingröße, wie Diphone oder Silben, festgelegt. Mit den im Corpus-Kit vorhandenen Werkzeugen kann dann eine statistische Analyse großer Textmengen stattfinden. Dazu erfolgt die Umsetzung der Texte in Phoneme mittels Regelwerk des TTS-Systems oder Aussprachewörterbuch. Anschließend werden die entstandenen Phonemketten nach Vorkommen und Auftretenshäufigkeit der Bausteine durchsucht. Zur Definition der Aufnahmetexte des Korpus erfolgt in der zweiten Phase eine satzweise Zerlegung der Analysetexte. Jeder Satz wird dabei mit einem Abdeckungsindex entsprechend der enthaltenen Bausteine versehen. Schließlich erfolgt die Selektion von n Sätzen mit dem größten Abdeckungsindex und die Nachselektion aller Sätze, die die restlichen Bausteine enthalten [7].

3 Aufnahme

Die Aufnahme der Sprachdaten erfolgt im akustischen Speziallabor des Instituts. Die Randbedingungen der Aufnahme werden entsprechend den Anforderungen des Korpus gestaltet. So kann z.B bei der Aufnahmen von Trägerwörtern eines Diphonwortschatzes ein Referenzsignal mit der mittleren Grundfrequenz des Sprechers dargeboten werden, um ein Konstanthalten der Grundfrequenz zu erleichtern.

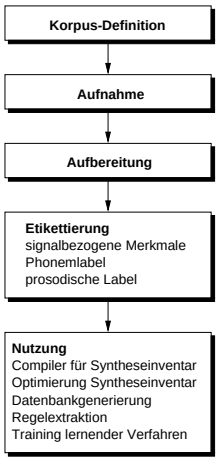

Abb. 1: Arbeitsschritte zur Erstellung eines Sprachkorpus.

4 Aufbereitung und Analyse

Zur weiteren Verwertung der Sprachaufnahmen ist meist eine Nachbearbeitung der Daten notwendig. Dazu zählt z.B. das Ändern der Abtastfrequenz, für das verschiedene Programme bereitstehen. Weiterhin erfolgt oft ein Ausschneiden der interessierenden Passagen aus den digitalisierten Aufnahmen. Dazu werden die Spracheinheiten mittels Pausendetektor markiert [5] und nach manueller Korrektur der Label im Corpus-Kit automatisch geschnitten.
In weiteren Verarbeitungsschritten werden signalbezogene Merkmale extrahiert. Dazu wird vor allem das Programmpaket *esps* genutzt. Mit ihm erfolgt z.B. die Formantanalyse. Das verwendete Formantsuchverfahren basiert auf der Suche von Maxima im LPC-Spektrum.
Soll der Sprachkorpus als Syntheseinventar verwendet werden, so ist für die Signalmanipulation mittels PSOLA das Setzen von Periodenmarken erforderlich. Dafür stehen im Dresdner Corpus-Kit verschiedene Programme zur Verfügung. Die besten Erfahrungen wurden bisher mit dem in [8] vorgestellten Verfahren gemacht.

5 Etikettierung

Ein wesentlicher Punkt für die weitere Verwendung des Korpus ist die Etikettierung der Aufnahmen nach verschiedenen Kriterien. Im Mittelpunkt stehen dabei segmentale und suprasegmentale Merkmale. Im Wesentlichen gibt es dazu derzeit zwei Vorgehensweisen: die Etikettierung erfolgt durch Messen bzw. Erkennen der entsprechenden Merkmale aus dem Signal oder durch Vorgabe der Merkmale,

z.B. einer Phonemfolge, und Alignen der Vorgabe mit den aktuellen Daten.

5.1 Lautetiketten

Beim Segmentieren der Sprachdaten in Lauteinheiten und der Zuordnung von Phonemnamen zu den Segmenten wird im Dresdner Corpus-Kit derzeit vor allem die Zielrichtung verfolgt, die gesprochene Symbolfolge vorzugeben und mit den aktuellen Sprachdaten zu alignen. Dazu werden zwei alternative Verfahren verwendet: Ein DTW-Aligner und ein HMM-Labeler.
Beim DTW-Aligner [9] wird ein zuvor synthetisiertes (oder, falls vorhanden, ein natürliches) Sprachsignal als Referenzsignal verwendet. Die Etikette des Referenzsignals werden auf das zu etikettierende Signal abgebildet, indem beide Sprachsignale mittels Dynamischer Programmierung synchronisiert werden. Das Verfahren ist sprecher- und sprachunabhängig.
Der HMM-Labeler arbeitet nach dem allgemein bekannten Prinzip. Die verwendeten HMM-Modelle wurden mit Spontansprache trainiert.

5.2 Prosodische Etikettierung

Neben der Segmentierung der Sprachdaten auf der Phonemebene besitzt die Etikettierung prosodischer Merkmale eine zunehmende Bedeutung. Die physikalischen Parameter der Prosodie, Grundfrequenz, Lautdauer und Amplitude, können direkt aus dem Signal bzw. aus den Ergebnissen der Segmentierung gewonnen werden. Entscheidend für die weitere Verwertung der Ergebnisse ist die Frage, nach welchem Modell die Analyse der Prosodie vorgenommen wird. Dementsprechend müssen, wie im Falle der Grundfrequenz, aus den physikalischen Merkmalen die Parameter des Analysemodells extrahiert werden.
Zur Analyse der Grundfrequenz aus den Sprachdaten sind eine Reihe von Verfahren entwickelt worden. Da keiner dieser Algorithmen fehlerfrei arbeitet, ist ein anschließendes Aufbereiten der Meßergebnisse erforderlich. Dazu wurde ein automatisches Verfahren zur Fehlerkorrektur entwickelt, das auf der Verknüpfung von statistischen und wissensbasierten Annahmen beruht [1].
Für die Generierung natürlicher Grundfrequenzkonturen hat sich im Dresdner multilingualen Sprachsynthesesystem DreSS das quantitative Modell von Fujisaki etabliert. Von entscheidender Bedeutung war dabei die Entwicklung von automatischen Verfahren zur Extraktion der Modellparameter [6] [3].
Das Messen der Lautdauern kann anhand der Phonemlabel direkt vorgenommen werden. Als übergeordnetes Analysemodell wird der zscore verwendet. Zur Analyse der zscore-Parameter und Generierung der zscore-Verläufe für ein konkretes Sprachsignal stellt der Corpus-Kit entsprechende Werkzeuge zur Verfügung [2].
Das Messen der Amplitude erfolgt durch Berechnung des RMS aus den Werten der Zeitfunktion. Da eine Steuerung der Amplitude während der Synthese derzeit nicht erfolgt, wird auch die Extraktion der Parameter eines entsprechend übergeordneten Modells momentan nicht vorgenommen.
Perzeptive Merkmale der Prosodie, wie Akzentuierung, werden aufgrund der Prosodiemodelle, wie dem Fujisaki-Modell, gelabelt. Alternativ können diese Parameter auch mit dem TTS-System generiert und mit dem Sprachsignal aligned werden.

6 Nutzung

Die Nutzung des Korpus ist natürlich stark vom Verwendungszweck abhängig. Soll der Korpus als Einheiteninventar des Sprachsynthesesystems dienen, so existieren weitere Werkzeuge zum Compilieren der Einheitendatenbasis. Die dafür benötigten Eingangsinformationen sind mit dem Corpus-Kit im Wesentlichen bereits erzeugt.

Sollen die Daten zur Bildung von Regelwerken dienen, wie zum Beispiel dem Aufstellen von Modellen zur Prosodiesteuerung, so stehen weitere Tools zum Zusammenstellen der Daten als relationale Datenbank zur Verfügung. Andere Programme dienen dem Training datenbasierter Verfahren mit den Informationen des Korpus.

7 Optimierung

Wurde aus dem Korpus ein Syntheseinventar erstellt, besteht ggf. der Wunsch nach dessen nachträglicher Optimierung. Dazu enthält der Corpus-Kit Programme, die die Inventarbausteine mit den während der Korpuserstellung generierten Etiketten versehen. Anhand des Vergleiches der einzelnen über die Etiketten beschriebenen Merkmale erfolgt die Erstellung von “Idealbausteinen”. Für jeden Alternativbaustein wird dann der gewichtete Abstand zum jeweiligen Idealbaustein berechnet und die Alternativen entsprechend ihres Abstands geordnet. Die endgültige Optimierung erfordert allerdings noch eine manuelle Nachbearbeitung.

8 Schnittstellen

Um einen hohen Grad an Kompatibilität zu gewährleisten, wird die Definition der Schnittstellen und Dateiformate möglichst einfach und universell gehalten. Die Etiketten werden z.B. als *esps*-Labeldateien gespeichert. Dabei handelt es sich um ASCII-Dateien mit einfachem Dateiformat. Bei Erfordernis können auch Daten anderer Formate leicht in dieses Format umgewandelt werden. Voraussetzung ist lediglich ein analoges Labelschema und gleiche Labelbezeichner.
Für das Exportieren der zusammengestellten Daten werden einfache Tabellenformate verwendet, die leicht in verschiedene andere Verarbeitungsprogramme importiert werden können.

9 Grafische Oberfläche

Für spezialisierte Aufgabenbereiche des Corpus-Kits wurde die graphische Benutzeroberfläche *wige* erstellt [4]. Da sie die script-Sprache Tcl / Tk nutzt, ist sie auf verschiedenen Rechnerplattformen lauffähig.

Literatur

[1] Diane Hirschfeld, Joachim Mersdorf, Hans Kruschke, and Oliver Jokisch. Ökonomische repräsentation natürlicher intonationskonturen. In *DAGA 2000*, Oldenburg, 2000.

[2] Oliver Jokisch et al. Multi-level rhythm control for speech synthesis using hybrid data driven and rule-based approaches. In *Proccedings of the International Conference of Spoken Language Processing*, 1998.

[3] Oliver Jokisch and Hans Kruschke. Predicting prosodic parameters: Evolutionary parameter extraction and hybrid neural network, rule based modelling. In *Proccedings of ISCA Workshop Prosody 2000*, Krakow, 2000.

[4] Uwe Koloska. Ein interaktives automatisches System zur Inventarerstellung und -optimierung. In *DAGA 2000*, Oldenburg, 2000.

[5] Thomas Kremer. Entwicklung und Test eines echtzeitfähigen Pausendetektors für die Wort- und Satzerkennung. Diplomarbeit, TU Dresden, 1996.

[6] Hansjörg Mixdorff. A novel approach to the fully automatic extraction of fujisaki model parameters. In *Proceeding of the International Conference on Acoustics, Speech and Signal Processing*, 1999.

[7] Thomas Richter. Statistische Untersuchungen zu Bausteingröße und -variantenzahl für die Synthese deutscher Sprache. Studienarbeit, TU Dresden, 2000.

[8] Ansgar Rinscheid. Automatische bestimmung von periodenmarken mit dem emark-algorithmus. In *Fortschritte der Akustik - DAGA ’93*, pages 1048 – 1051. DPG Verlag Bad Honnef, 1993.

[9] Guntram Strecha. Multilinguale Etikettierung natürlicher Sprachsignale auf Basis synthetischer Referenzsignale. Diplomarbeit, TU Dresden, 2000.