Die BITS Sprachsynthesekorpora – Diphon- und Unit Selection-Synthesekorpora für das Deutsche

Florian Schiel, Christoph Draxler, Tania Ellbogen, Klaus Jänsch, Sonja Schmidt
BAS Bayerisches Archiv für Sprachsignale
Institut für Phonetik und Sprachliche Kommunikation
Universität München
Schellingstr. 3
80799 München
draxler@phonetik.uni-muenchen.de

Abstract

Im Rahmen des BITS-Projekts hat das Bayerische Archiv für Sprachsynthese (BAS) zwei Korpora für die Sprachsynthese des Deutschen erstellt. Das erste Korpus erlaubt eine Sprachsynthese auf der Basis von Diphonen, das zweite eine Synthese mittels Unit Selection, d.h. auf der Basis einer Auswahl unterschiedlich langer Einheiten. Das Diphonkorpus umfasst 2795 Diphonkombinationen, gesprochen in neutralen Logatomen, das Unit Selection-Korpus 1683 gelesene Sätze.

Beide Korpora wurden von vier professionellen Sprechern (2 weiblich, 2 männlich) gesprochen und über zwei Mikrofone aufgenommen, zusätzlich wurde ein Laryngographsignal aufgezeichnet. Die Sprachsignale wurden automatisch vorsegmentiert, diese Segmentation wurde in zwei separaten Durchgängen manuell korrigiert. Damit sind die Korpora sowohl für die Forschung als auch für die Entwicklung qualitativ hochwertiger Sprachsynthesesysteme für das Deutsche geeignet.

Die Korpora werden vom BAS distribuiert.

1 Einleitung

Für das Deutsche gab es vor dem BITS-Projekt (BAS Infrastructure for Technical Speech Processing) keine allgemein verfügbaren Korpora für die Sprachsynthese. Ziel des BITS-Projekts war es daher, zwei solche Korpora zu erstellen, die dann sowohl für die Forschung als auch die praktische Entwicklung von konkatenativen Sprachsynthesesystemen verwendet werden können (Ellbogen et al., 2004).

Bei konkatenativer Sprachsynthese wird die zu produzierende Äußerung durch die Auswahl von Signalfragmenten aus einer Datenbank und die anschließende Aneinanderreihung dieser Fragmente erzeugt. Bei der Diphonsynthese bestehen diese Fragmente aus Lautpaaren, bei der Unit Selection aus unterschiedlich langen Lautketten.

Beim Zusammenfügen der Fragmente entstehen Brüche im Signal, die hörbar sind. Diese können durch geschickte Auswahl der Fragmente und anschließende Glättung des Signals minimiert werden. Die Unit Selection-Synthese reduziert potenziell die Anzahl dieser Brüche, vor allem dann, wenn die zu produzierende Äußerung bereits ganz oder teilweise in der Datenbank enthalten ist. Im schlechtesten Fall jedoch wird eine Äußerung aus Einzellauten zusammengesetzt, mit entsprechenden Folgen für die Sprachqualität.

Bei der Diphonsynthese ist die Natürlichkeit des synthetisierten Signals in der Regel geringer als bei der Unit Selection-Synthese, aber die Qualität ist gleichmäßiger.

Für die konkatenative Synthese gilt allgemein, dass Laute, die nicht in der Datenbank enthalten sind, auch nicht produziert werden können – die Korpora müssen also vollständig in Bezug auf ein Phoneminventar sein.

2 Spezifikation der Korpora

Zur Spezifikation der Synthesekorpora wurde am BAS im Frühjar 2002 ein Workshop organisiert, zu dem Sprachsyntheseforscher, -entwickler und potenzielle Nutzer eingeladen waren. Auf diesem Workshop wurde festgelegt, dass je ein Diphon- und ein Unit Selection-Korpus erstellt werden sollten. Die Auswahl der Sprecher sollte bei einem Casting durchgeführt werden.

2.1 Korpusinventar

Für die Diphonsynthese wurden alle Phoneme gemäß dem SAM-PA Inventar für das Deut-

sche (Wells, 1997) sowie die wichtigsten, im deutschen Inventar nicht vorhandenen englischen und französischen SAM-PA Phoneme verwendet (siehe Tabelle 1). Die Phonemkombinationen wurden so in Logatome eingebettet, dass möglichst keine koartikulatorischen Effekte auftreten. So wurden Vokale generell in einen */d/* oder */t/* Kontext, Konsonanten in einen */a/* oder */@/* Kontext gestellt, z.B. "DATA*FAH*TADEU" für das Diphon */fa:/* oder "AD*EUSCH*ADEI" für */OYS/*.

Die Logatome wurden mit neutraler Intonation gesprochen.

Tabelle 1: Phoneminventar des Diphon-Sprachsynthesekorpus in SAM-PA-Notation

Sprache	Phoneme
deutsch	/I/, /E/, /a/, /O/, /U/, /Y/, / 9/, /i:/, /e:/, /E:/, /a:/, /o:/, /u:/, /y:/, /2:/, /aI/, /aU/, /OY/, / @/, /6/, /?/, /p/, /b/, /t/, /d/, /k/, /g/, /pf/, /ts/, /tS/, /f/, /v/, /s/, /z/, /S/, /Z/, /C/, /x/, /j/, /h/, /m/, /n/, /N/, /l/, /R/
englisch	/EI/, /@U/, /T/, /D/, /r/, /L/, /w/
französisch	/E∼/, /a∼/, /o∼/

Es sollten alle Phonemkombinationen gelesen werden, auch solche, die aus phonotaktischen Gründen im Deutschen innerhalb von Wörtern nicht vorkommen. Bei Wortfolgen sowie in zweisprachigen Texten können diese Kombinationen jedoch nicht ausgeschlossen werden. Einige Phonemkombinationen wurden wegen Unaussprechlichkeit nicht in das Korpus aufgenommen, so dass insgesamt 2795 Diphone aufzunehmen waren.

Als Grundlage für das Unit Selection-Korpus wurde ein Satzkorpus des IMS Stuttgart herangezogen, das bereits prosodisch vorannotiert war. Es umfasst 1683 Sätze, die geringfügig überarbeitet wurden, z.B. Aktualisierung von Datumsangaben oder Währungsbeträgen.

2.2 Sprecher

Für das Casting wurden 45 Sprecher eingeladen, die je 90 Logatome lesen mussten. Aus diesen Logatomen konnten drei Testsätze synthetisiert werden, die von 18 Testhörern auf ihre Natürlichkeit und Sympathie hin angehört wurden. 10 Sprecher kamen in die zweite Runde, in der sie von Sprachsynthese-Experten beurteilt wurden. Aus diesen wurden dann die in Tabelle 2 aufgeführten Sprecher ausgewählt.

Tabelle 2: BITS Synthesekorpus-Sprecher

	spr1	spr2	spr3	spr4
Geschlecht	w	w	m	m
Alter	47	55	40	38
Raucher	ja	ja	ja	nein
L2	E, F	E, I, GR	E, F	E, I
Beruf	Rundfunksprecher			Schauspieler

3 Aufnahmetechnik

Die Aufnahmen erfolgten in einer schallgedämmten Kabine vom Typ StudioBox Professional.

Aufgenommen wurden drei Kanäle (Abb. 1):

1. Nahbesprechungsmikrofon Beyerdynamic NEM192, Position 7cm rechts von der Mundmitte und auf Höhe der Oberlippe,
2. Großmembran Kondensatormikrofon Neumann Typ TLM 103, Position 60 cm vom Mund entfernt, und
3. Laryngograph-Signal vom Typ LaryngoGraph PCLX.

Die Samplerate beträgt 48 kHz bei 16 Bit Quantisierung. Für die Aufnahme wurde die Software SpeechRecorder (Draxler and Jänsch, 2004) verwendet.

Die Dauer der Sprachaufnahmen im Studio betrug einschließlich der Sitzungen zur Fehlerkorrektur insgesamt 134 Stunden. Bei jeder Sitzung waren drei Mitglieder des Aufnahmeteams anwesend, die die korrekte Aussprache, Intonation, Nebengeräusche beim Sprechen und die technische Qualität überwachten.

4 Annotation

Das Logatom-Korpus wurde nur phonetisch segmentiert, das Unit Selection-Korpus zusätzlich noch prosodisch annotiert.

Abbildung 1: Position der Mikrofone und der Laryngographen-Elektroden bei den Syntheseaufnahmen in der StudioBox

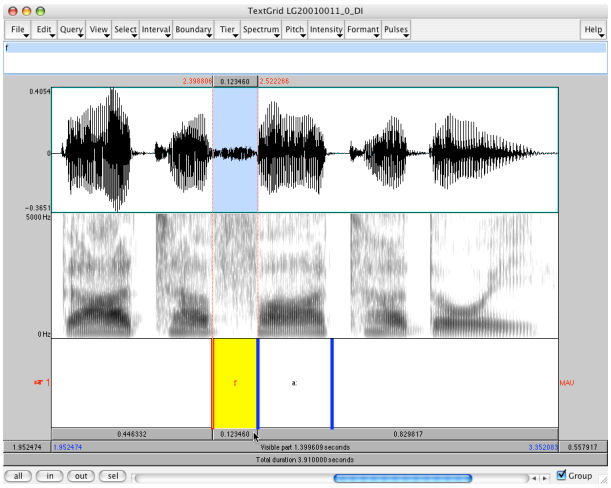

Abbildung 2: Segmentiertes Logatom DATA*FAH*TADEU von Sprecherin 1. Das Diphon */fa:/* ist in zwei Segmente unterteilt, */f/* ist ausgewählt.

Beide Annotationen erfolgten in mehreren aufeinanderfolgenden Schritten mit einer Feedback-Schleife zur Meldung von in späteren Schritten gefundenen Fehlern. Die Annotierer waren den einzelnen Schritten Erst-, Zweit- und Abschlussannotation fest zugeordnet, so dass in jedem Schritt ein anderer Annotierer eine Annotation bearbeitete. Zur Verbesserung der Konsistenz der Annotationen gab es regelmässige Treffen der Annotierer zum Erfahrungsaustausch und zur Besprechung problematischer Fälle.

4.1 Phonetische Transkription und Segmentierung

Beide Synthesekorpora wurden mit MAUS automatisch vorsegmentiert (Schiel, 1999) und anschliessend mit Praat (Boersma, 2001) bearbeitet.

Das Diphonmaterial enthält normalerweise drei Segmentgrenzen: Beginn des Diphons, Grenze zwischen den Einzellauten, und Ende des Diphons (Abb. 2). Bei Diphonen mit Plosiven werden innerhalb eines Lautes zusätzliche Grenzen gesetzt, um Stille- und Burstphase abzugrenzen.

Im Unit Selection-Material sind alle Sätze komplett phonetisch segmentiert.

Diese phonetische Transkription wurde von geschulten Transkribierern manuell überprüft. Beim Diphonmaterial wurden nur die Grenzen verschoben – wenn ein Laut nicht den Vorgaben entsprach, wurde die Äußerung als defekt markiert und erneut aufgenommen. In den Unit Selection-Sätzen konnten neben den Grenzen auch die Lautvorgaben entsprechend den real produzierten Äußerungen verändert werden.

Alle Transkriptionen wurden in einem zweiten Transkriptionsschritt überprüft und gegebenenfalls überarbeitet, und in einem letzten Korrekturschritt abgeschlossen.

4.2 Prosodische Annotation

Ausgangspunkt der prosodischen Annotation des Unit Selection-Korpus war eine automatisch erstellte Vorgabe nach GToBI light. Diese wurde in Praat als zusätzliche Spur angezeigt, ebenso der $F0$-Verlauf der Äusserung. Diese Vorgabe wurde in einem ersten Schritt manuell an die tatsächliche Realisierung der Äusserung angepasst und in einem abschliessenden Schritt von anderen Annotierern nochmals überprüft.

4.3 Annotationsformate

Sämtliche Annotationsdaten sind im BAS Partiturformat (Schiel et al., 1997), im Annotation Graph Format (Bird and Liberman, 1999) und als Praat Textgrids (Boersma, 2001) gespeichert.

5 Validierung

Nach Abschluss der Arbeiten wurden beide Korpora validiert. Dazu wurden Vorabversionen sowohl institutsintern als auch an Teilnehmer des initialen Workshops verteilt mit der Bitte um Fehlermeldungen und Kommentare.

Zusätzlich wurden beide Korpora automatisch auf formale Fehler überprüft. Dazu wurden Parser für die phonetische Transkription bzw. die Partiturfiles eingesetzt.

Bei der Validierung hat sich gezeigt, dass einige deutsche Vokale, namentlich die gespannten kurzen Vokale, fehlen (sie sind in SAM-PA nicht enthalten). Außerdem enthält das Unit Selection-Korpus nur wenige Fragesätze, so dass für eine Synthese solcher Sätze vermutlich nicht ausreichend viele prosodische Muster vorliegen.

Die bei der Validierung gefundenen formalen und technischen Fehler wurden für die endgültige Distribution korrigiert. Die Dokumentation wurde überarbeitet und es wurden Abweichungen von der ursprünglichen Spezifikation in einem Abschlussbericht festgehalten.

6 Verfügbarkeit

Der Umfang des Diphonkorpus beträgt ca. 13 GB, der des Unit Selection-Korpus ca. 14 GB.

Die Korpora können in verschiedenen Varianten zu Forschungs-, Entwicklungs- oder Anwendungszwecken lizensiert werden.

Auf der Webseite des BAS `www.bas.uni-muenchen.de` können kurze Äußerungen synthetisiert werden, wobei es sich dabei nur um eine Rohsynthese ohne jede Glättung des Signals oder weitergehende Signalverarbeitungsschritte handelt.

Außerdem befindet sich auf der Webseite ein Video, das einen Einblick in die Sprachaufnahmen im Studio gibt.

7 Ausblick

Mit den beiden Synthesekorpora stehen der Syntheseforschung und -entwicklung für das Deutsche nun endlich interessante und ausreichend große Korpora zur allgemeinen Verwendung bereit.

Die Sprecher der Synthesekorpora stehen für weitere Aufnahmen zur Verfügung. Dies ist vor allem dann interessant, wenn für ein bestimmtes Anwendungsgebiet spezielles Sprachmaterial aufgenommen werden soll, z.B. für Auto-Navigationssysteme oder die Synthese von Ansagediensten wie Wetterbericht oder Bahnauskunft. Mit solchen zusätzlichen Aufnahmen bietet sich dann die einmalige Chance, mit relativ geringem Aufwand qualitativ hochwertige Sprachsynthese auf der Basis zweier allgemein verwendbarer Korpora und eines kleinen, anwendungsspezifischen Korpus zu realisieren.

Danksagung

Besonderer Dank gebührt den studentischen Hilfskräften am Institut für Phonetik der LMU München, die die enge phonetische Transkription und prosodische Annotation durchgeführt haben, außerdem dem Aufnahmeteam im Studio, das sich geduldig über 11.000 Logatome angehört hat.

Die Arbeit an den Synthesekorpora wurde mit Mitteln des BITS Projektes des BMBF (Kennzeichen #01IVB01) gefördert.

References

St. Bird and M. Liberman. 1999. Annotation graphs as a framework for multidimensional linguistic data analysis. Technical report, Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Chr. Draxler and K. Jänsch. 2004. Speechrecorder – a universal platform independent multichannel audiorecording software. In *Proceedings. of 4th Intl. Conference on Language Resources and Evaluation*, pages 559–562, Lisbon.

T. Ellbogen, F. Schiel, and A. Steffen. 2004. The BITS speech synthesis corpus for german. In *Proceedings of LREC 2004*, Lisbon.

F. Schiel, S. Burger, A. Geumann, and K. Weilhammer. 1997. The Partitur Format at BAS. FIPKM report #35, Institut für Phonetik und Sprachliche Kommunikation, Universität München.

F. Schiel. 1999. Automatic phonetic transcription of non-prompted speech. In *Proceedings of ICPhS*, pages 607–610, San Francisco.

J. Wells. 1997. Standards, Assessment, and Methods: Phonetic Alphabets. Technical report, University College, London.