Prosodische Alternativeinheiten für ein silbenorientiertes chinesisches Sprachsynthesesystem

Jörg Helbig, Hongwei Ding
TU Dresden, Institut für Akustik und Sprachkommunikation, 01062 Dresden
Email: helbig@eakss2.et.tu-dresden.de, ding@eakss2.et.tu-dresden.de

1. Einleitung

Der Mangel an Natürlichkeit bei synthetischer Sprache ist gegenwärtig der wichtigste Kritikpunkt an der Ausgabequalität von Sprachsynthesesystemen. Insbesondere Systeme mit dem klassischen Diphon- oder Multiphonansatz sind davon betroffen. Die wichtigsten Ursachen dafür sind:

- Die ungenügende Qualität der Modellierung der prosodischen Parameter
- ein hoher Bedarf an prosodischer Manipulation der Basiseinheiten zur Erzeugung der geforderten Intonations- und Dauerwerte
- die große Anzahl von Verkettungsstellen im Synthesesignal aufgrund der geringen Bausteingrößen.

Schlußfolgernd daraus ergeben sich als Anforderungen zur Verbesserung der Sprachqualität für die Syntheseinventare folgende Anforderungen:

- Zurückdrängen des Einflusses einer (möglicherweise schlechten) Prosodiemodellierung durch Ausnutzen der prosodischen Varianz des natürlichen Sprachsignals,
- damit Minimierung bzw. Wegfall der Notwendigkeit prosodischer Manipulationen,
- größere und flexible Basiseinheiten.

Mit gestiegenen Systemressourcen sowie der Verfügbarkeit immer größerer Sprachdatenbanken begann deshalb die Entwicklung von Korpussynthesen, die den obigen Anforderungen besser gerecht werden. Das wesentlichste Merkmal solcher Systeme ist ein (meist sehr großes) phonetisch und prosodisch annotiertes Sprachdateninventar, das die Sprachbausteine in phonetischen, positionalen und prosodischen Varianten enthält, aus denen im Syntheseprozeß die geeigneten Kandidaten anhand von Optimierungskriterien ausgewählt und verkettet werden.

Als wichtigste Vertreter der korpusbasierten chinesischen Sprachsynthesen können gegenwärtig die Systeme in /1/ und /2/ angesehen werden. Auch die chinesische Sprachkomponente des Sprachsynthesesystems DreSS der TU Dresden verwendet ein Inventar mit phonetischen, koartikulatorisch motivierten Alternativeinheiten, das schrittweise um prosodische Varianten erweitert wird.

2. Die chinesische Sprachsynthese in DreSS

Ausgehend von der silbenorientierten Struktur der chinesischen Sprache wurde das chinesische Sprachdateninventar von DreSS als Silbeninventar konzipiert /3,4/. In seiner bisherigen Ausbaustufe enthielt es 2910 Silben, die aus phonetischer Sicht die Sprechsilben der chinesischen Sprache vollständig abdecken. Die Zusammensetzung der Datenbasis basierte auf folgenden Voraussetzungen:

- Die chinesische Sprache mit ihren mehr als 13000 Schriftzeichen verfügt nur über 408 phonologisch unterschiedliche Sprechsilben. Jede dieser Silben enthält einen zentralen Vokal, vor dem 21 verschiedene Konsonanten stehen können. Diese (Konsonant) – Vokal – Verbindungen können durch die nasale /n/ und /N/ ergänzt werden. Nicht alle phonotaktisch möglichen CVC-Varianten sind vorhanden.
- Das Chinesische ist eine Tonhöhensprache, bei der die Intonation im Silbenbereich bedeutungsunterscheidenden Charakter trägt. Die stimmhaften Bereiche der Silben können in bis zu 4 typischen Intonationsverläufen realisiert werden, die den Silben jeweils unterschiedliche Wortbedeutungen zuweisen. (Jede Silbe besitzt eine oder mehrere Wortbedeutungen, kann jedoch auch ein Teil von mehrsilbigen Wörtern sein.) Die typischen Intonationsverläufe werden als die Vokaltöne 1 – 4 (konstant, steigend, fallend/steigend und fallend) bezeichnet. Zusätzlich existiert noch der neutrale Ton 0, der in reduzierten Silben auftritt und sich in seinem Verlauf stark nach dem Vorgängerton richtet.

Insgesamt umfaßt das Chinesische 1218 Silben, die in der fließenden Sprache im Wortverbund bzw. an den Wortfugen typische Koartikulationsmuster aufweisen. Kategorisiert man diese Koartikulationsmuster nach den Artikulationsorten der sie bedingenden Folgekonsonanten, können die drei typischen Artikulationsverläufe nach labial, alveolar/palatal und velar unterschieden werden. Da diese koartikulatorischen Varianten nur bei auf Vokal endenden Silben perzeptive Relevanz besitzen, besteht das Basis-Syntheseinventar aus ein bis drei Vertretern pro Silbe.

Das Inventar wurde von einem männlichen Sprecher gesprochen. Zur Erzeugung der gewünschten Koartikulationsmuster wurden die Silben in Trägersätze eingebettet, bei denen die jeweiligen Folgesilben den erforderlichen Artikulationsort besaßen /3/.

Im Syntheseprozeß werden die passenden Vertreter entsprechend der geforderten Silbensequenz und den damit bekannten Folgesilben aus der Datenbasis extrahiert und verkettet. Zur Erzeugung der Satzprosodie wird der suprasegmantalen Intonation der Vokalton überlagert /5, 6/.

3. Prosodische Alternativeinheiten

Im Ergebnis von Evaluierungen der Synthesequalität wurde deutlich, daß die synthetisierte Sprache trotz hoher Verständlichkeit beim Hörurteil „fließende Sprache“ Mängel aufwies. Die Ausgabequalität wurde mit Sprechweise von Kindern, die das Lesen erlernen, verglichen. Durch einen Vergleich der Synthesedatenbasis mit Silben einer phonetisch und prosodisch annotierten Textdatenbank konnte als Ursache dafür ermittelt werden, daß die Synthesesilben – bedingt durch die Struktur der Trägersätze – mit Wortakzenten realisiert worden waren /7/. Damit gestaltete sich die Erzeugung von (häufig erforderlichen) unbetonten Silben als problematisch. Besonders Silben mit neutralem Ton waren davon betroffen. Die folgenden Abbildungen veranschaulichen die Problematik. Die Partikel „de0“ mit neutralem Ton hinter Ton1- bzw. Ton4-Silben aus der Textdatenbank des Sprechers zeigen stark unterschiedliche In-

tonationsverläufe, verbunden mit im Bild nicht sichtbaren spektralen Variationen, die mit den Mitteln der prosodischen Manipulation im Zeitbereich nur mit akustischen Störungen aus vorhandenen Syntheseeinheiten erzeugt werden können.

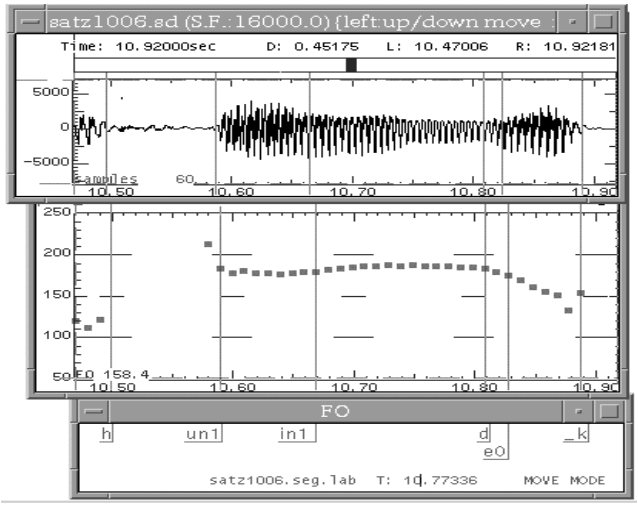

Bild 1: Neutraler Partikel „de“ hinter einer Ton1-Silbe in „hun1in1 de0“ (mittlere F0: ca. 160 Hz, F0-Bereich: 30 Hz)

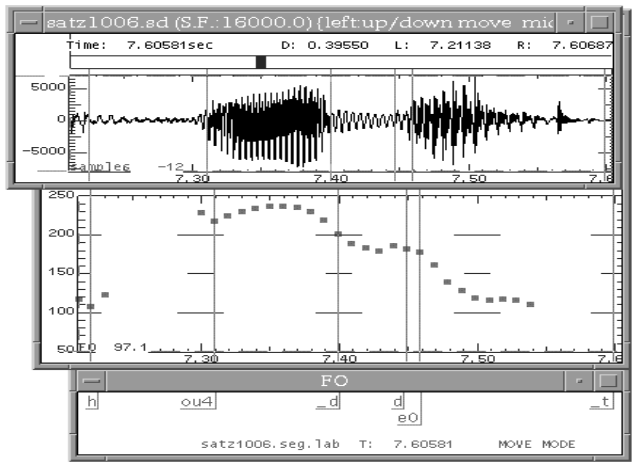

Bild 2: Neutraler Partikel „de“ hinter einer Ton4-Silbe in „yi3hou4 de0“ (mittlere F0: 140 Hz, F0-Bereich: 70 Hz)

Schlußfolgernd daraus wurden die wichtigsten Neutralpartikel durch Nachaufnahmen in das Syntheseinventar integriert. Dabei wurde wieder die Strategie der koartikulatorischen Varianten verfolgt. Außerdem wurden je 4 Tonkoartikulationsmuster pro Silbe in Abhängigkeit vom Vorgängerton generiert. Abbildung 3 zeigt am Beispiel von „de0“ mit labialem Folgelaut die Zusammensetzung der neuen Trägersätze. Insgesamt wurden 139 neutrale Silben zusätzlich aufgenommen.

4. Diskussion

Mit der Integration von unbetonten Silben in das chinesische Syntheseinventar konnte die Notwendigkeit umfangreicher prosodischer Manipulationen verringert und die

Zhe4 shi4 ta1 de5 bai2 zhi3.
这 是 他 **(ta1)** 的 白 纸.
Das ist sein weisses Papier. (Einheit: "de513")

Zhe4 shi4 nan2 de5 bai2 zhi3.
这 是 男 **(nan2)** 的 白 纸.
Das ist maennliches weisses Papier. (Einheit: "de523")

Zhe4 shi4 ni3 de5 bai2 zhi3.
这 是 你 **(ni3)** 的 白 纸.
Das ist dein weisses Papier. (Einheit: "de533")

Zhe4 shi4 huai4 de5 bai2 zhi3.
这 是 坏**(huai4)** 的 白 纸.
Das ist defektes weisses Papier. (Einheit: "de543")

Bild 3: Trägersätze zur Generierung der tonkoartikulatorischen Varianten von „de0“ in labialem Kontext; die Ziffern nach dem Einheitennamen (Bsp. „de513“) bedeuten nach den Konventionen der Synthesedatenbasis von DreSS:
5- neutraler Ton, 1-Vorgängerton 1, 3- labialer Folgelaut

Qualität der synthetisierten Sprache verbessert werden. Die Verwendung von unbetonten Silben (Ton 1-4) aus der fließend gelesenen Textdatenbasis führte dagegen nicht zu einer Verbesserung der Ausgabequalität. Obwohl sowohl die Trägersätze als auch die Textdatenbasis vom gleichen Sprecher stammen und unter identischen akustischen Voraussetzungen aufgenommen sind, bestehen erhebliche Kontraste zwischen den fließend gelesenen, teils sehr flüchtig artikulierten Texten und den Trägersätzen hinsichtlich spektraler Eigenschaften, Lautdauer sowie Tonkontur. Die gemeinsame Verwendung derartig unterschiedlicher Einheiten führt zu starken Schwankungen in der Ausgabequalität, die den Hörer irritieren. Zukünftig ist eine Ergänzung der Synthesedatenbasis mit unbetonten Silben aus deutlich artikulierten Fließtexten vorgesehen.

5. Literatur

/1/ Chou F-C., Tseng C-Y, Lee L-S.: Selection of Waveform units for Corpus-Based Mandarin Speech Synthesis based on Decision Trees an d prosodic Modification costs. Proc. EUROSPEECH 99, Budapest.
/2/ Wang, R. et.al.: “A New Chinese Text-to-Speech System with High Naturalness”. Proc. ICSLP 96. P.1441-1444
/3/ Helbig, J., Ding, H.: A Syllable-based Mandarin Chinese Speech Synthesis regarding Cross-syllable Coarticulation Effects. ICSP´97 Aug. 1997, Seoul, Korea. 173-176
/4/ Ding, H.: A syllable based Chinese Synthesis System with a coaticulation-oriented Inventory, Diss. TU Dresden, 1999
/5/ Ding, H., Helbig, J.: Natural Tone Contours in a Mandarin Chinese Speech Synthesizer. Proc. ESCA Workshop on Intonation, Athens, Greece, Sept. 1997, 95-98
/6/ Ding H., Helbig J.: Modeling Duration and Tonal Coarticulation in a Mandarin Chinese Speech Synthesis. Proc. 1. ISCSLP, Singapore, Dez. 1998, 243-248
/7/ Ding, H., Helbig, J.: Untersuchungen zur Dauersteuerung für ein chinesisches Sprachsynthesesystem. Proc. DAGA 98, Zürich 1998, 378,379