Erzeugung schnell gesprochener Sprache in der Unit-Selection-Sprachsynthese

Donata Moers[1,2], Petra Wagner[2] und Bernd Möbius[1,3]

[1]Abteilung Sprache und Kommunikation, Universität Bonn; [2]Fakultät für Linguistik und Literatur, Universität Bielefeld; [3]IMS, Universität Stuttgart
dmo@ifk.uni-bonn.de

Abstract: In unserem Beitrag wird ein neuer Ansatz zur Synthese schnell gesprochener Sprache in der Unit-Selection-Sprachsynthese vorgestellt. Schnell gesprochene Sprache unterscheidet sich hinsichtlich ihrer akustischen Eigenschaften von in normalem Tempo gesprochener Sprache. Um besonders unerwünschte Eigenschaften schneller Sprache bei der Erstellung eines schnell gesprochenen Bausteininventars für die Unit-Selection-Sprachsynthese weitestgehend zu vermeiden, wurde eine geeignete Sprecherin ausgewählt, die in der Lage war, bei maximalem Sprechtempo möglichst deutlich zu sprechen. Anschließend wurden zwei unabhängige, bezüglich des linguistischen Inhalts identische Synthesekorpora erstellt: eines in normalem (ca. 4 Silben pro Sekunde) und eines in schnellem und möglichst deutlichem Sprechtempo (ca. 8 Silben pro Sekunde). Für beide Korpora wurde die Verwendung so genannter *Phoxsy Units* [1] als Syntheseeinheiten untersucht. Die Ergebnisse einer perzeptiven Evaluation zeigen, dass *Phoxsy Units* insbesondere für die Synthese schnell gesprochener Sprache geeignet sind.

1 Einleitung

Für viele blinde und sehbehinderte Menschen ist die Nutzung von Sprachsynthesesystemen Teil ihres täglichen Lebens. Diese Nutzer bevorzugen häufig eine sehr schnelle Ausgabegeschwindigkeit [2, 3]. Architekturen wie Formant- oder Diphonsynthese sind in der Lage, Sprache in extrem hohen Geschwindigkeiten zu erzeugen; diese klingt aber sehr unnatürlich. Unit-Selection-Synthesesysteme dagegen können natürlicher klingende Sprache erzeugen, aber schnelle Sprechgeschwindigkeit wurde in diesen Systemen bisher nicht adäquat implementiert.

Schnell gesprochene Sprache unterscheidet sich hinsichtlich ihrer akustischen Eigenschaften von in normalem Tempo gesprochener Sprache. Je schneller jemand spricht, desto schlechter sind seine Äußerungen zu verstehen. Dies liegt daran, dass bei hoher Sprechgeschwindigkeit die für eine deutliche Artikulation notwendigen Zielstellungen von den Artikulatoren nicht mehr erreicht werden. Neben der insgesamt kürzeren Dauer der Laute ist daher häufig auch deren Assimilation an banachbarte Laute, ihre Reduktion oder gar vollständige Elision zu beobachten. Da derartige Phänomene die Verständlichkeit synthetischer Sprache noch mehr negativ beeinflussen, als sie dies bei natürlicher schneller Sprache ohnehin tun, sollten sie bei der Erstellung eines Bausteininventars für die Unit-Selection-Sprachsynthese so weit wie möglich vermieden werden. Aus diesem Grund wurde für das hier vorgestellte Vorhaben eine Sprecherin ausgewählt, die in der Lage war, bei maximalem Sprechtempo möglichst deutlich zu artikulieren. Diese Auswahl wurde durch eine perzeptive Evaluation der erstellten Korpusaufnahmen bestätigt [4].

In der Sprachsynthese gibt es verschiedene Möglichkeiten, schnelle Sprache zu generieren. Eine häufig verwendete Option ist dabei die Dauermanipulation mit Hilfe des TD-PSOLA-Algorithmus, der eine lineare Beschleunigung der Signale ermöglicht. Ab einem bestimmten Beschleunigungsfaktor treten jedoch häufig Artefakte auf, die die Qualität der erzeugten

Sprache stark beeinträchtigen [5]. Eine weitere Möglichkeit ist die Nachahmung prosodischer Charakteristika natürlicher schnell gesprochener Sprache. Durch die Reduzierung der Anzahl und Dauer von Pausen, sowie durch die Verminderung der Anzahl und Stärke von Phrasengrenzen kann jedoch keine adäquate Beschleunigung der Sprache erreicht werden. Die Nachahmung weiterer Charakteristika schnell gesprochener Sprache führt zu abnehmender Verständlichkeit bei zunehmender Sprechgeschwindigkeit; eine zugrunde liegende deutliche Artikulation wird einer Synthese, die die Charakteristika schneller Sprache berücksichtigt, deutlich vorgezogen [6]. Unter Berücksichtigung dieser Gegebenheiten umfasst der hier vorgestellte Ansatz die Erstellung eines eigenen Bausteininventars für schnelle Sprache, um durch deren Verwendung die Natürlichkeit der synthetisierten Sprache zu erhöhen. Gleichzeitig werden zu starke Koartikulation und Reduktion durch die möglichst deutliche Aussprache in schneller Sprechgeschwindigkeit so weit wie möglich vermieden, um die Verständlichkeit nicht zu beeinträchtigen.

Die schnellen und fließenden Transitionen, die in natürlicher Sprache zu beobachten sind, sind auch für die Verständlichkeit synthetischer Sprache wichtig. Transitionen werden von herkömmlicher Diphonsynthese nicht adäquat behandelt, können aber mit Formantsynthese erzeugt werden. Dementsprechend bevorzugen blinde Hörer bezüglich der Verständlichkeit die weniger natürlich klingende Formantsynthese. Da die akustischen Transitionen aufeinander folgender Segment eine so wichtige Rolle für die Verständlichkeit von Sprache spielen, müssen die Diskontinuitäten, die bei konkatenativen Syntheseverfahren im erzeugten Sprachsignal auftreten, möglichst minimiert werden. Als Konsequenz wurde von Breuer und Abresch das Konzept der *Phoxsy Units* eingeführt [1]: Laute, die stark koartikuliert werden, werden zwar grundsätzlich als zwei einzelne Phone betrachtet, in der Synthese aber als untrennbare Syntheseeinheit behandelt. Die Verwendung von Phoxsy Units könnte somit ein Weg sein, schnell gesprochene Sprache in der Unit-Selection-Sprachsynthese zu erzeugen. Dadurch würde einerseits die Natürlichkeit der generierten Sprache erhöht werden, andererseits würde die Verständlichkeit durch die Verwendung von Einheiten, die typische Transitionen beinhalten, verbessert werden.

2 Phoxsy Units

Es ist bekannt, dass linguistisch motivierte Syntheseeinheiten wie beispielsweise Phone keine optimalen Eigenschaften für die Verwendung in der konkatenative Sprachsynthese besitzen. Ihr größter Nachteil ist die Missachtung akustischer und auditiver Kontinuität. *Phoxsy Units* (*pho*ne extensions for *sy*nthesis) wurden definiert, um Konkatenationsstellen, an denen die Kontinuität des Signals nicht optimal ist, systematisch zu verhindern, insbesondere an Stellen, an denen sie höchst unerwünscht sind. Diese Multiphon-Syntheseeinheiten bestehen aus einer Folge von Phonen, die stark koartikuliert werden, und umfassen damit auch die für die Perzeption so wichtigen Transitionen. Eine Konkatenationsstelle kann innerhalb von Phoxsy Units nicht auftreten. Tabelle 1 zeigt alle möglichen Phon-Kombinationen, wie sie in [1] definiert wurden. Die Spalte „IPA" zeigt dabei die Transkription der Einheitentranskription nach dem Internationalen Phonetischen Alphabet, die Spalte „BOSS-SAMPA" listet die an das Unit-Selection-Synthesesystem BOSS angepasste, bisherige Verwendung der Einheiten auf und die Spalte „Phoxsy" zeigt die Definition der Phoxsy-Einheiten.

2.1 Implementierung

Unter Berücksichtigung der Ergebnisse von Breuer und Abresch [1] wurden Phoxsy Units als eigenständige Einheiten-Auswahl-Ebene im Unit-Selection-Synthesesystem BOSS [7] implementiert, um einen robusten und zuverlässigen Zugriff zu gewährleisten. Dadurch konnte auch eine Verbesserung der Laufzeit erreicht werden.

IPA	BOSS-SAMPA	Phoxsy
ʔ + Vokal	? + Vokal	? + Vokal
h/ɦ + Vokal	einzelne Phone	h + Vokal
j + Vokal	einzelne Phone	j + Vokal
ʋ/v + Vokal	einzelne Phone	v + Vokal
ʁ/ʁ/ɾ/r + Vokal	einzelne Phone	r + Vokal
l + Vokal	einzelne Phone	l + Vokal
ən/n	@n	@n
əm/m	einzelne Phone	@m
əl/l	einzelne Phone	@l
j/ʋ/v/ʁ/ʁ/ɾ/r/l + ən	einzelne Phone	j/v/r/l + @n
j/ʋ/v/ʁ/ʁ/ɾ/r/l + əm	einzelne Phone	j/v/r/l + @m
j/ʋ/v/ʁ/ʁ/ɾ/r/l + əl	einzelne Phone	j/v/r/l + @l

Tabelle 1 – Einheitendefinition in IPA, BOSS-SAMPA und als Phoxsy Unit

Die modulare Struktur von BOSS machte eine unproblematische Implementierung und Integration der neuen Einheiten-Auswahlebene in das bestehende System möglich. Das BOSS-Tool *blf2xml*, das Informationen aus den BOSS Label Format (BLF)-Dateien extrahiert [8] und eine XML-Datenbank erstellt, wurde erweitert, um Phoxsy-Einheiten mit Hilfe der BOSS_FSA-Klasse (Finite State Automaton) zu erkennen. Andere BOSS-Tools wurden ebenfalls an die neue Einheitenebene angepasst, um zusätzliche Informationen wie Kontextklassen, Phraseninformation und MFCCs für die Phoxsy Units zu berechnen. Diese Informationen wurden dann der XML-Datenbank hinzugefügt. Das Tool *blfxml2db* fügt die neue Einheiten-Auswahlebene in die MySQL-Datenbank ein und berechnet einen Einheiten-Index, mit dessen Hilfe jede Einheit des Korpus exakt identifiziert werden kann. Über Mapping-Tabellen werden die Einheiten benachbarter Auswahlebenen verknüpft. Die Auswahlebenen sind dabei hierarchisch angeordnet; oberste Ebene ist das Wort, gefolgt von der Silbe, den Phonen und zuletzt den Halbphonen. Die Ebene der Phoxsy-Einheiten wurde zwischen der Silben- und der Phon-Ebene implementiert. Die *syllable map* wurde angeglichen, um eine Verknüpfung zwischen Silben und Phoxsy-Einheiten zu erstellen. Eine entsprechende *phoxsy unit map* wurde ebenfalls erstellt, um die Verknüpfung zwischen Phoxsy-Einheiten und Phonen darzustellen. Darüber hinaus wurde eine neue Vorauswahldatei für die Phoxsy-Ebene erstellt. Die *BOSS_Unitselection*-Klasse wurde angepasst und eine neue Ebene *PHOXSY* wurde der *BOSS_Node*-Klasse hinzugefügt. Die *BOSS_Transcription*-Klasse wurde ebenfalls angepasst, um Phoxsy-Einheiten zu identifizieren und sie in die interne Kommunikationsstruktur des Systems einzufügen. Sie nutzt dabei denselben Mechanismus wie das *blf2xml*-Tool.

3 Synthese normaler Sprechgeschwindigkeit

Nachdem die bezüglich des linguistischen Inhalts identischen Korpora in den Sprechgeschwindigkeiten normal (ca. 4 Silben pro Sekunde) und schnell (ca. 8 Silben pro Sekunde) aufgenommen, aufbereitet und implemetiert worden waren [9], wurden zunächst 15 Sätze aus

den Domänen *Nachrichten* und *Erzählung* in normaler Sprechgeschwindigkeit synthetisiert. Dabei enthielt jeder Satz mindestens 3 Phoxsy-Einheiten. Zur Synthese dieser Sätze wurden verschiedene Strategien angewendet:

- Synthese nur aus Phonen
- Synthese nur aus Phoxsy Units
- Synthese unter Einbeziehung aller Einheiten-Auswahlebenen außer Phoxsy Units
- Synthese unter Einbeziehung aller Einheiten-Auswahlebenen inklusive Phoxsy Units

Da ein paarweiser Vergleich zwischen allen 4 Versionen eines Satzes den vertretbaren Umfang einer perzeptiven Evaluation überschritten hätte, wurde entschieden, das Experiment in zwei Teilexperimente aufzuteilen. Im ersten dieser Teilexperimente wurden diejenigen Stimuli paarweise miteinander verglichen, die beide aus einer einzelnen Einheiten-Auswahl-Ebene generiert wurden: nur Phone gegenüber nur Phoxsy-Einheiten. Das zweite Teilexperiment bestand aus dem paarweisen Vergleich derjenigen Stimuli, die entweder unter Einbeziehung aller Einheiten-Auswahlebenen außer Phoxsy Units oder unter Einbeziehung aller Einheiten-Auswahlebenen inklusive Phoxsy Units generiert wurden.

Am ersten Teilexperiment nahmen 23 Versuchspersonen teil. Alle waren naive Hörer und nicht mit der Nutzung von Sprachsynthesesystemen vertraut. Die Versuchspersonen wurden gebeten, von einem Stimuluspaar diejenige Version anzugeben, die verständlicher war, sowie diejenige Version, die sich natürlicher anhörte. Jeder der 15 Sätze wurde den Versuchspersonen zweimal präsentiert, um die Verlässlichkeit der Beurteilungen zu erfassen. Abbildung 1 zeigt auf der linken Seite die „verständlicher" und „natürlicher" Beurteilungen für Stimuli, die nur aus Phonen generiert wurden (dunkelgrau) im Vergleich zu Stimuli, die nur aus Phoxsy-Einheiten erzeugt wurden (hellgrau). Die Ergebnisse zeigen, dass die Verständlichkeit für Stimuli, die nur aus Phoxsy-Einheiten bestanden, als signifikant besser beurteilt wurde als die Verständlichkeit von Stimuli, die nur aus Phonen bestanden (χ^2, $p < 0.05$). Bezüglich der Natürlichkeit wurde hingegen kein signifikanter Unterschied in den Beurteilungen festgestellt.

An der zweiten Evaluation nahmen 14 Versuchspersonen teil, die ebenfalls alle naive Hörer waren und nicht geübt in der Nutzung von Sprachsynthesesystemen. Die Versuchspersonen wurden auch in diesem Experiment gebeten, diejenige Version eines Stimuluspaares anzugeben, die verständlicher war sowie diejenige Version, die sich natürlicher anhörte. Jedes

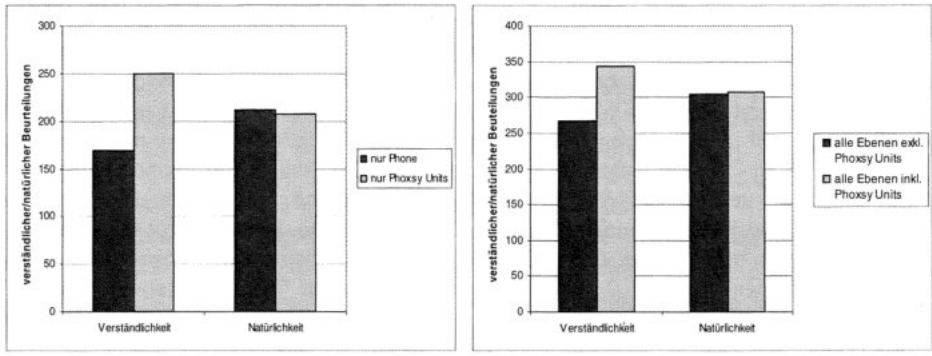

Abbildung 1 – Beurteilungen der Verständlichkeit und Natürlichkeit von Stimuli in normaler Sprechgeschwindigkeit, die aus nur einer Einheiten-Auswahlebene generiert wurden (links) und von Stimuli, die aus allen Ebenen mit oder ohne Einbeziehung von Phoxsy Units generiert wurden (rechts).

der 15 Satzpaare wurde den Hörern zweimal in randomisierter Reihenfolge vorgespielt. Auf der rechten Seite von Abbildung 1 sind die „verständlicher" und „natürlicher" Beurteilungen für Stimuli abgebildet, die entweder unter Einbeziehung aller Einheiten-Auswahlebenen außer Phoxsy Units (dunkelgrau) oder unter Einbeziehung aller Einheiten-Auswahlebenen inklusive Phoxsy Units (hellgrau) generiert wurden. Auch hier zeigte sich ein signifikanter Unterschied in der Beurteilung der Verständlichkeit zugunsten der Verwendung von Phoxsy-Einheiten (χ^2, $p < 0.005$), jedoch kein signifikanter Unterschied zwischen beiden Versionen bezüglich der Natürlichkeit.

Beide Teilexperimente zeigten einen signifikanten Unterschied in der Beurteilung der Verständlichkeit von Stimuli, bei denen Phoxsy-Einheiten in die Synthese einbezogen wurden bzw. ausschließlich für die Synthese benutzt wurden. Die Ergebnisse von Breuer und Abresch [1] konnten somit bestätigt werden.

4 Synthese schneller Sprechgeschwindigkeit

Starke Koartikulation und Reduktion sind in schnell gesprochener Sprache nicht vollständig vermeidbar, auch wenn diese mit großer Genauigkeit und erhöhtem artikulatorischem Aufwand realisiert wird. Da Phoxsy-Einheiten als eine Folge von Phonen, die stark koartikuliert werden, definiert sind, könnte ihre Verwendung für die Synthese schnell gesprochener Sprache großen Einfluss auf deren Verständlichkeit und Natürlichkeit haben. Dieser Einfluss könnte zum einen stark negativ sein, da Phoxsy Units natürliche Koartikulations- und Reduktionsphänomene beinhalten und somit die Verständlichkeit der synthetisierten Sprache drastisch verschlechtern könnten. Andererseits könnte die Verwendung dieser Multiphon-Syntheseeinheiten sowohl die Verständlichkeit als auch die Natürlichkeit synthetisierter schnell gesprochener Sprache signifikant erhöhen, da sie neben den natürlichen Koartikulations- und Reduktionsphänomenen auch die für die Perzeption so wichtigen natürlichen Transitionen enthalten und darüber hinaus mehr Kontextinformation anbieten als einzelne Phone.

Für die Evaluation der Synthese schnell gesprochener Sprache wurden erneut 15 Sätze aus den Domänen *Nachrichten* und *Erzählung* synthetisiert, die mindestens 3 Phoxsy Units beinhalteten. Auch hier wurden unterschiedliche Strategien zur Generierung der schnellen synthetischen Sprache angewendet:

- Synthese nur aus Phonen
- Synthese nur aus Phoxsy Units
- Synthese unter Einbeziehung aller Einheiten-Auswahlebenen außer Phoxsy Units
- Synthese unter Einbeziehung aller Einheiten-Auswahlebenen inklusive Phoxsy Units

Das erste Perzeptionsexperiment war ein paarweiser Verglich von Stimuli, die entweder nur aus Phonen oder nur aus Phoxsy Units generiert wurden. 22 Versuchspersonen nahmen an dem Experiment teil. Wie bereits zuvor waren alle Versuchspersonen naive Hörer und nicht an die Verwendung von Sprachsynthese gewöhnt. Alle Versuchspersonen wurden gebeten, von einem Stimuluspaar diejenige Version anzugeben, die verständlicher war und diejenige Version, die sich natürlicher anhörte. Jeder der 15 Sätze wurde den Hörern zweimal präsentiert, um die Zuverlässigkeit der Beurteilungen zu erfassen. Auf der linken Seite von Abbildung 2 sind die „verständlicher" und „natürlicher" Beurteilungen für Stimuli dargestellt, die nur aus Phonen generiert wurden (dunkelgrau) im Vergleich zu Stimuli, die nur aus Phoxsy-Einheiten erzeugt wurden (hellgrau). Die Ergebnisse zeigen, dass die Verständlichkeit von Stimuli, die nur aus Phoxsy-Einheiten bestanden, als signifikant besser beurteilt wurde als die Verständlichkeit von Stimuli, die nur aus Phonen bestanden (χ^2, $p < 0.001$). Auch bezüglich

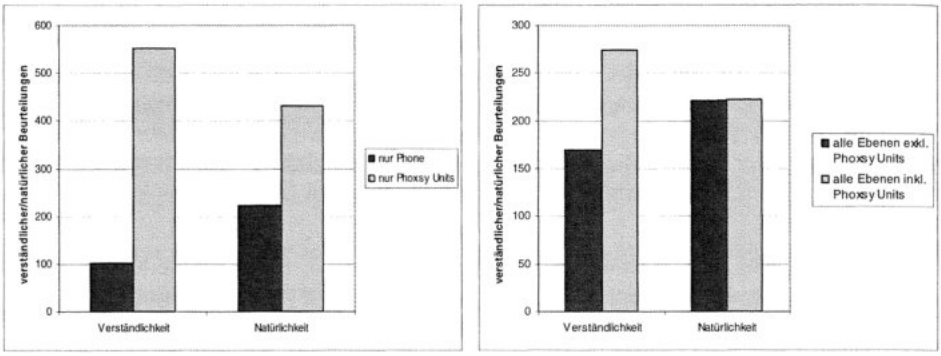

Abbildung 2 – Beurteilungen der Verständlichkeit und Natürlichkeit von Stimuli in schneller Sprechgeschwindigkeit, die aus nur einer Einheiten-Auswahlebene generiert wurden (links) und von Stimuli, die aus allen Ebenen mit oder ohne Einbeziehung von Phoxsy Units generiert wurden (rechts).

der Natürlichkeit wurde ein signifikanter Unterschied (χ^2, $p < 0.001$) in den Beurteilungen zugunsten der Phoxsy Units festgestellt.

Das zweite Teilexperiment bestand auch hier aus einem paarweisen Vergleich zwischen Stimuli, die entweder unter Einbeziehung aller Einheiten-Auswahlebenen außer Phoxsy Units oder unter Einbeziehung aller Einheiten-Auswahlebenen inklusive Phoxsy Units generiert wurden. An dem Experiment nahmen 15 Versuchspersonen teil. Diese waren wiederum alle naive Hörer und nicht geübt in der Nutzung von Sprachsynthese. Die Versuchspersonen wurden gebeten, von einem Stimuluspaar diejenige Version anzugeben, die verständlicher war sowie diejenige Version, die sich natürlicher anhörte. Abbildung 2 zeigt rechts die „verständlicher" und „natürlicher" Beurteilungen für Stimuli, die unter Einbeziehung aller Einheiten-Auswahlebenen exklusive Phoxsy Units (dunkelgrau) und unter Einbeziehung aller Einheiten-Auswahlebenen inklusive Phoxsy Units (hellgrau) erzeugt wurden. Die Ergebnisse zeigen, dass die Verständlichkeit für Stimuli, die unter Einbeziehung aller Einheiten-Auswahlebenen inklusive Phoxsy Units generiert wurden, als signifikant besser beurteilt wurde als die Verständlichkeit von Stimuli, die unter Einbeziehung aller Einheiten-Auswahlebenen exklusive Phoxsy Units (χ^2, $p < 0.001$) erzeugt wurden. Bezüglich der Natürlichkeit wurde kein signifikanter Unterschied in den Beurteilungen festgestellt.

Die perzeptive Evaluation der vom schnell gesprochenen Korpus synthetisierten Sprache hat gezeigt, dass Stimuli, die nur aus Phoxsy-Einheiten synthetisiert wurden, sowohl bezüglich ihrer Verständlichkeit als bezüglich ihrer Natürlichkeit als signifikant besser beurteilt wurden als Stimuli, die nur aus Phonen synthetisiert wurden. Für Stimuli, die unter Berücksichtigung aller Einheiten-Auswahlebenen mit oder ohne Berücksichtigung von Phoxsy Units synthetisiert wurden, wurde ein signifikanter Unterschied bezüglich der Verständlichkeit festgestellt. Multiphon-Einheiten sind somit nicht nur für die Synthese von Sprache in normalem Sprechtempo geeignet, sondern insbesondere auch für die Synthese schnell gesprochener Sprache von einem schnell gesprochenen Bausteininventar.

5 Diskussion

Für die Untersuchung von Synthese schnell gesprochener Sprache wurden Multiphon-Einheiten, so genannte *Phoxsy Units,* als eigenständige Einheiten-Auswahlebene in das Unit-Selection-Synthesesystem BOSS implementiert, um einen robusten und zuverlässigen Zugriff zu gewährleisten. Die Evaluation synthetisierter Sprache, die aus einem in normalem Tempo gesprochenen Bausteininventar generiert wurde, zeigte eine signifikant bessere Verständlich-

keit von Stimuli, die nur aus Phoxsy-Einheiten erzeugt wurden, gegenüber Stimuli, die nur aus Phonen synthetisiert wurden. Stimuli, die aus allen Einheiten-Auswahlebenen inklusive Phoxsy Units generiert wurden, wiesen ebenfalls eine signifikant bessere Verständlichkeit auf als Stimuli, bei denen alle Ebenen außer Phoxsy Units zur Synthese genutzt wurden. Diese Signifikanz war allerdings nicht so hoch wie in der Bedingung, bei der einzelne Einheiten-Auswahlebenen miteinander verglichen wurden. Bezüglich der Natürlichkeit wurde weder für diese Bedingung noch für die Bedingung, bei der die Stimuli aus mehreren Einheiten-Auswahlebenen erzeugt wurden, ein signifikanter Unterschied zwischen den jeweiligen Versionen festgestellt. Die Ergebnisse von Breuer und Abresch [1] konnten somit für normale Sprechgeschwindigkeit in ähnlicher Ausprägung bestätigt werden.

Für die Synthese schneller Sprache wurde ein schnell gesprochenes Bausteininventar verwendet; die schnelle Sprache wurde hierbei besonders sorgfältig artikuliert. Die Verwendung von Phoxsy Units hatte in dieser Bedingung einen großen Einfluss auf die Qualität der synthetisierten Sprache, sowohl bezüglich der Verständlichkeit als auch bezüglich der Natürlichkeit. Dies liegt möglicherweise daran, dass Phoxsy Units als Multiphon-Syntheseeinheiten die für die Perzeption wichtigen Transitionen beinhalten und so der negative Einfluss zu starker Koartikulation und Reduktion überdeckt wird. Wie erwartet zeigte eine perzeptive Evaluation auch hier einen signifikanten Vorteil der Stimuli, die unter Verwendung von Phoxsy-Einheiten synthetisiert wurden gegenüber Stimuli, bei deren Erzeugung Phoxsy Units nicht genutzt wurden. Die Verwendung von Phoxsy Units erhöhte insbesondere bei der schnellen Sprache sowohl die Verständlichkeit, als auch – zumindest zum Teil – die Natürlichkeit der generierten Sprache. Phoxsy Units sind somit für die Synthese schnell gesprochener Sprache von einem schnell und präzise artikulierten Korpus besonders geeignet.

Literatur

[1] Breuer, S., Abresch, J.: Phoxsy: Multi-phone Segments for Unit Selection Speech Synthesis. In: Proceedings Interspeech 2004 – ICSLP, Jeju Island, Korea, 2004.

[2] Moers, D., Wagner, P., Breuer, S.: Assessing the Adequate Treatment of Fast Speech in Unit Selection Speech Synthesis Systems for the Visually Impaired. In: Proceedings 6th ISCA Workshop on Speech Synthesis (SSW-6), Bonn, 2007.

[3] Moos, A., Trouvain, J.: Comprehension of Ultra-Fast Speech – Blind vs. 'Normally Hearing' Persons. In: Proceedings ICPhS XVI: pp. 677-684, Saarbrücken, 2007.

[4] Moers, D., Wagner, P.: Assessing a Speaker for Fast Speech in Unit Selection Speech Synthesis. In: Proceedings Interspeech 2009, Brighton, 2009.

[5] Chen, S.-H., Chen, S.-J., Kuo, C.-C.: Perceptual Distortion Analysis and Quality Estimation of Prosody-Modified Speech for TD-PSOLA. In: Proceedings of the ICASSP'06, Toulouse, 2006.

[6] Janse, E.: Production and Perception of Fast Speech. Dissertation Universiteit Utrecht, 2003.

[7] Klabbers, E. et al.: Speech synthesis development made easy: The Bonn Open Synthesis System. In: Proceedings Eurospeech, Aalborg, 2001.

[8] Breuer, S., Abresch, J., Wagner, P., Stöber, K.: BLF - ein Labelformat für die maschinelle Sprachsynthese mit BOSS II. In: Hess, W., Stöber, K. (Ed.): Tagungsband Elektronische Sprachsignalverarbeitung ESSV, Studientexte zur Sprachkommunikation. Bonn, 2001.

[9] Moers, D., Wagner, P., Möbius, B., Müllers, F., Jauk, I.: Integrating a Fast Speech Corpus in Unit Selection Speech Synthesis: Experiments on perception, segmentation and duration prediction. In: Proceedings Speech Prosody 2010 Chicago, IL, 2010.