Evaluation eines Sprechers für schnell gesprochene Sprache in der Unit-Selection basierten Sprachsynthese

Donata Moers, Petra Wagner

Institut für Kommunikationswissenschaften, Abteilung Sprache und Kommunikation, Rheinische Friedrich-Wilhelms-Universität Bonn, Poppelsdorfer Allee 47, 53115 Bonn
E-Mail: {dmo,pwa}@ifk.uni-bonn.de
Web: http://www.ifk.uni-bonn.de/

Zusammenfassung

Unser Beitrag befasst sich mit der akustischen und perzeptiven Evaluation eines Sprechers, der aufgrund einer globalen Vorauswahl für das Aufsprechen eines schnell gesprochenen Bausteininventars für die Unit-Selection basierte Sprachsynthese als geeignet erscheint. Hierzu wird zunächst ein Überblick über die phonetischen Eigenschaften schnell gesprochener Sprache gegeben. Danach wird die H&H-Theorie dargelegt, welche verschiedene Strategien schnellen Sprechens erläutert, aus denen Anforderungen an den Sprecher und damit wichtige Voraussetzungen für seine Eignung abgeleitet werden. Anschließend wird ein Perzeptionsexperiment vorgestellt, dessen Ergebnisse ebenso wie die Ergebnisse einer akustischen Analyse der Aufnahmen die aus der Vorauswahl gewonnenen Eindrücke sowie die aus der H&H-Theorie abgeleiteten Anforderungen untermauern.

1 Einleitung

Insbesondere Blinde und Sehbehinderte bevorzugen bei häufiger Anwendung von Sprachsynthesesystemen eine hohe Sprechrate [1, 2, 3]. Diese wird in Unit-Selection basierten Synthesesystemen bisher aber nur unzureichend modelliert. Andere Synthesearchitekturen wie Formant- oder Diphon-Synthese können zwar hohe Sprechraten erzeugen; die dadurch generierte Sprache enthält aber nicht die in natürlicher schnell gesprochener Sprache zu beobachtenden Phänomene.

Je schneller ein Sprecher spricht, desto unverständlicher werden seine Äußerungen meist für den Zuhörer. Dies liegt unter anderem daran, dass die einzelnen Laute bei hoher Sprechgeschwindigkeit zeitlich stärker überlappen und benachbarte Laute einander mehr angeglichen werden (Assimilation) [4]. Auch werden die für eine deutliche Artikulation notwendigen Zielstellungen von den jeweiligen Artikulatoren nicht mehr vollständig realisiert [4, 5]. Generell lassen sich diese Phänomene mit den Begriffen Koartikulation und Reduktion umschreiben. Bei Vokalen äußern sie sich hauptsächlich in einer Verkürzung der Dauer und einer Änderung der Formantfrequenzen [6, 7]. Konsonanten hingegen werden häufiger an ihren Kontext angeglichen und gehen dabei ggf. in eine andere Konsonantenklasse über; auch werden sie mit weniger Intensität und/oder nicht mehr komplett realisiert bis hin zu ihrem vollständigen Wegfall (Elision) [6, 8]. Größere Einheiten wie Silben oder Intonationsphrasen sind in schnell gesprochener Sprache ebenfalls von Veränderungen betroffen; so werden Silben verkürzt und die Anzahl betonter Silben nimmt insgesamt ab [9]. Die Anzahl und Stärke der Phrasengrenzen ist ebenso rückläufig [10], die Grundfrequenz weist insgesamt einen flacheren Verlauf auf [11].

Bei der Modellierung schneller Sprache in der Unit-Selection basierten Sprachsynthese sind diese Phänomene größtenteils unerwünscht, da sie nicht nur die Verständlichkeit der erzeugten Sprache negativ beeinflussen [11, 12], sondern auch die Definition der Auswahleinheiten wesentlich erschweren [13].

2 Die H&H-Theorie

Da sowohl Koartikulation als auch Reduktion die Verständlichkeit von Sprache negativ beeinflussen, stellt sich nun die Frage, ob es grundsätzlich eine Möglichkeit gibt, diese Phänomene beim Sprechen weitestgehend zu vermeiden. Eine Antwort auf diese Frage liefert die von Lindblom aufgestellte Hyper- und Hypoartikulationstheorie (kurz: H&H-Theorie) [14]. Sie besagt, dass trotz des kontinuierlichen Sprachverlaufs und dadurch bedingter Koartikulation eine ausreichende Kontrastierung akustischer Signale für den Sprecher möglich ist. Gleichzeitig ist dies auch notwendig, um von einem Hörer verstanden zu werden und somit erfolgreich zu kommunizieren. Die vom Sprecher ausgehenden Signale müssen hierfür einen ausreichenden Kontrast für den lexikalischen Zugriff im mentalen Lexikon des Hörers aufweisen [ibid.].

Der Sprecher befindet sich folglich in einem Dilemma: Auf der einen Seite ist er bemüht, seine Information mit möglichst geringem Energieaufwand, z. B. durch Verringerung des Artikulationsaufwands, zu übermitteln. Dieses ökonomische Verhalten manifestiert sich in eher undeutlicher Sprache (Hypospeech). Auf der anderen Seite bedingen Zweckorientiertheit und Plastizität überdeutliche Sprache (Hyperspeech). Ein Sprecher muss also ständig zwischen der Ökonomie der Artikulationsaufwendungen und dem Erreichen des Kommunikationsziels abwägen und bewegt sich dabei permanent entlang eines Kontinuums zwischen Hypo- und Hyperspeech. Lindblom umschreibt dies folgendermaßen: „Hence speakers are expected to vary their output along a continuum of hyper- and hypospeech“ [ibid.: 403].

Macht das Kommunikationsziel (oder die Aufgabenstellung) dies erforderlich, könnte es einem geeigneten Sprecher somit durchaus möglich sein, sehr schnell und trotzdem deutlich zu sprechen.

3 Anforderungen an den Sprecher

In der Sprachsyntheseforschung hat sich immer wieder gezeigt, dass die Qualität synthetischer Sprache zu einem Großteil vom Sprecher des Bausteininventars determiniert wird. Ausgebildete Sprecher, die über einen längeren Zeitraum mit gleich bleibender Stimmqualität und hoher artikulatorischer Präzision zu sprechen gelernt haben, produzieren in der Regel hochwertigere Synthesebausteine [15].

Bauen die Inventare auf schnell gesprochener Sprache auf, verschärfen sich die möglicherweise auftretenden Probleme der Genauigkeit der Artikulation sowie gleich bleibender Stimmqualität und Sprechgeschwindigkeit, da zunächst davon auszugehen ist, dass ein Sprecher bei diesem Sprechstil seine Artikulationsgenauigkeit der Sprachökonomie zumindest teilweise opfern wird. Ungeübte Sprecher zeigen diese Tendenz vermutlich noch stärker. Um die zentrale Frage nach der Realisierbarkeit schnell gesprochener synthetischer Sprache nicht aufgrund eines ungeeigneten Inventarsprechers negativ beantworten zu müssen, sollte ein Sprecher also folgenden Kriterien genügen:

- Er sollte in der Lage sein, sehr schnell bei maximaler Verständlichkeit zu sprechen. Bisherige Untersuchungen für Deutsch [16] und Niederländisch [17] zeigen dabei eine maximale Artikulationsrate bei ca. 8 Silben/Sekunde auf, sofern die Sprache noch verständlich sein soll.
- Die gesammelten Sprecherfahrungen des Sprechers sollten möglichst domänenübergreifend sein, damit er beim Schnellsprechen nicht in einen speziellen Sprechstil verfällt, wie bspw. den in Auktionshäusern gebräuchlichen Stil, der nicht auf andere Domänen übertragbar ist.

Ausgehend von diesen Voraussetzungen wurde für die Suche nach potentiellen Sprechern zunächst auf Sprecher zurückgegriffen, die aufgrund vorheriger Korpusarbeiten und Syntheseevaluationen als geeignet erschienen. Dazu kamen einige freiwillige Sprecher, die Sprecherfahrung aus anderen Bereichen aufweisen konnten. Es wurden Voraufnahmen auf der Grundlage verschiedener Aufgabenstellungen erstellt. Dazu gehörten das Vorlesen eines 5 Sätze umfassenden Textes in normaler und schneller Sprechgeschwindigkeit sowie die Realisierung zweier weiterer Sätze, welche mehrere englische Wörter enthielten, ebenfalls in beiden Tempovariationen. Aus dem Pool von insgesamt 6 weiblichen und 3 männlichen potentiellen Sprechern wurde dann von geschulten Hörern in einer Vorauswahl mittels eines globalen Präferenztests derjenige für das Sprechen des Syntheseinventars ausgewählt, dessen schnelle Sprache am verständlichsten und dabei noch am natürlichsten war. Dies ist zudem ein Garant für hohe Natürlichkeit bei synthetischer Sprache [15].

Insbesondere die Kontinuität von Stimmqualität und Lautstärke, die Deutlichkeit der Artikulation sowie die Natürlichkeit von Aussprache und Intonation – vor allem in schnell gesprochener Sprache – waren die zentralen Beurteilungskriterien. Weiterhin wurde auch der subjektive Klangeindruck jeder Stimme von den Hörern festgehalten. Anhand dieser Kriterien wurden 2 weibliche und 1 männlicher Sprecher in die engere Wahl gezogen. Es zeigte sich, dass eine der Sprecherinnen sehr viel besser als die beiden anderen Kandidaten dazu in der Lage war, extrem schnell bei maximaler Deutlichkeit zu sprechen, so dass die endgültige Wahl auf sie fiel. Außerdem stand sie für die Aufnahmen eines Korpus, welche in der Regel recht zeitaufwändig sind und oft auch Nachaufnahmen erfordern, zeitlich unbegrenzt zur Verfügung, so dass dies ein letztes Kriterium für ihre Auswahl darstellte.

4 Sprecherevaluation

Von der ausgewählten Sprecherin wurden zunächst jeweils 3 Aufnahmen in normaler sowie in intendiert zunehmender Sprechgeschwindigkeit gemacht. Grundlage für diese Aufnahmen war ein Text, der auch im Bonn-Tempo-Korpus [16] verwendet wurde. Er stammt aus der Erzählung *Selbs Betrug* von B. Schlink (1994, S. 242) und umfasst insgesamt 76 Silben in 4 Haupt- und 3 Nebensätzen. Um vor der Erstellung eines umfangreicheren Bausteininventars für die Sprachsynthese zu verifizieren, dass die Sprecherin den oben genannten Kriterien tatsächlich genügt und sie insbesondere in der Lage ist, bei Bedarf auch bei hoher Sprechgeschwindigkeit deutlich zu sprechen, wurden 3 weitere schnelle Versionen mit intendiert zunehmender Sprechgeschwindigkeit erstellt. Dabei wurde die Sprecherin aufgefordert, bei jeder Version bewusst den Artikulationsaufwand zu erhöhen und besonders deutlich zu sprechen. Für die Evaluation des schnellen Sprechtempos lagen folglich insgesamt 6 Instanzen in 2 Deutlichkeitsstufen und jeweils 3 unterschiedlich hohen Sprechgeschwindigkeiten vor.

4.1 Akustische Evaluation

Ein erster Schritt war die Untersuchung der akustischen Eigenschaften der verschiedenen schnellen Versionen. Zentral war hierbei die Frage, ob anhand der akustischen Eigenschaften der Sprachsignale bereits deutlich werden würde, dass unerwünschte Effekte wie zu starke Koartikulation und Reduktion von der Sprecherin in schneller und bewusst deutlicher Sprache tatsächlich vermieden werden können. Hierzu wurden einzelne Ausschnitte (siehe Abb. 1a und 1b) der verschiedenen Aufnahmen betrachtet, bei denen ein hohes Maß an Reduktions- und Koartikulationsphänomenen zu erwarten war.

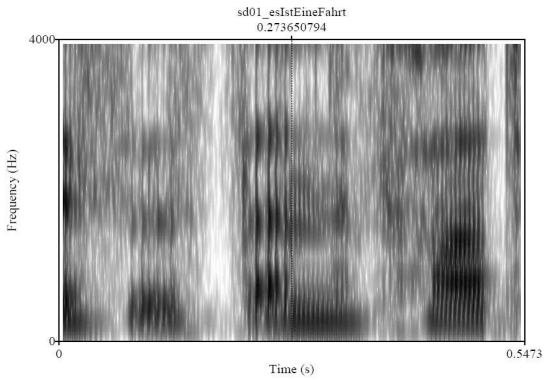

Abbildung 1a: Spektrogramm einer deutlichen Version des Abschnitts „Es ist eine Fahrt“.

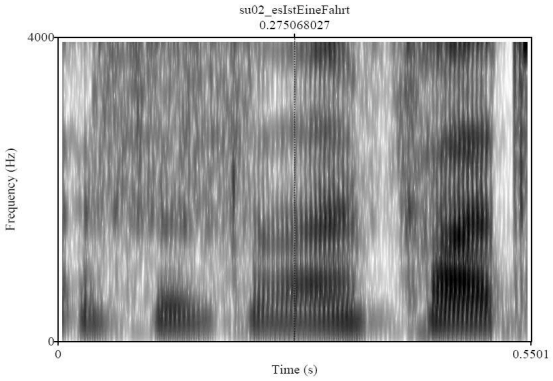

Abbildung 1b: Spektrogramm einer eher undeutlichen Version des Abschnitts „Es ist eine Fahrt“.

Im Einzelnen wurden folgende Phänomene betrachtet:

- Verkürzung und Reduktion von Vokalen
- Elision des Schwa-Lautes
- Silbifizierung von Konsonanten, i.d.R. bedingt durch vorhergehende Schwa-Elision
- Assimilation benachbarter Konsonanten
- Unvollständiger Verschluss und unvollständige Verschlusslösung bei Plosiven
- Abnahme des durch die Voice Onset Time (der Zeit zwischen Einsetzen der Stimmbandschwingung und Verschlusslösung) bedingten Unterschieds zwischen stimmlosen und stimmhaften Konsonanten
- Verminderung der Intensität bei Frikativen
- Abnahme der Anzahl betonter Silben
- Verminderung der Anzahl und Intensität von Phrasengrenzen
- Flacherer Grundfrequenzverlauf

Es zeigte sich, dass diese Phänomene gemäß der H&H-Theorie in den bewusst deutlicher artikulierten Versionen tatsächlich seltener auftraten als in den eher undeutlichen Versionen.

4.2 Perzeptive Evaluation

Für die perzeptive Evaluation wurde aus den ausgewählten Signalabschnitten mit Hilfe der Sprachevaluierungssoftware Praat [18] ein Perzeptionsexperiment erstellt, das aus insgesamt 9 Teilexperimenten bestand. In jedem Teilexperiment wurde die jeweilige Instanz desselben Ausschnitts der unterschiedlich schnell und deutlich gesprochenen Versionen in einer paarweisen Gegenüberstellung (siehe Abb. 2) sowohl von phonetisch erfahrenen (n = 10) als auch von phonetisch unerfahrenen (n = 13) Hörern verschiedener Altersklassen bezüglich ihrer Deutlichkeit beurteilt. Es war davon auszugehen, dass die bewusst deutlicher artikulierten Versionen bei ähnlicher Sprechgeschwindigkeit über alle Ausschnitte als besser bewertet werden würden.

Die einzelnen Signalausschnitte waren so gewählt worden, dass sie trotz der in ihnen enthaltenen hohen Anzahl an Lauten oder Lautkombinationen, für die ein größeres Maß an Koartikulation und Reduktion zu erwarten war, inhaltlich noch verständlich waren. Um möglicherweise dennoch auftretende Verständnisprobleme zu vermeiden, wurde der Inhalt des jeweiligen Ausschnitts am Anfang des entsprechenden Teilexperiments in Schriftform angezeigt. Außerdem hatten die Versuchspersonen die Möglichkeit, sich das Gesagte durch Anklicken des Wiederholen-Buttons (siehe Abb. 2) bis zu 5 Mal anzuhören.

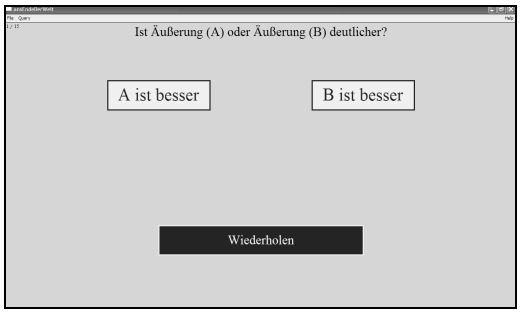

Abbildung 2: Grafische Oberfläche des Perzeptionsexperiments mit Wahl der deutlicheren Version.

Den Probanden wurden in den 9 Teilexperimenten jeweils 15 Stimuli = insgesamt 135 Stimuli dargeboten. Das Experiment wurde in ruhiger Umgebung mit Kopfhörern durchgeführt. Die Versuchspersonen wurden angewiesen, von dem zu hörenden Stimuluspaar die Version auszuwählen, welche sie besser verstehen würden bzw. welche ihnen deutlicher ausgesprochen schien.

Da die intendierten Sprechgeschwindigkeiten variierten, wurde für die Auswertung zunächst das exakte Sprechtempo jeder einzelnen Version in Silben pro Sekunde bestimmt (siehe Tab. 1). Anschließend wurde für je ein Paar ähnlich schneller deutlicher und undeutlicher Versionen das arithmetische Mittel der Sprechgeschwindigkeiten gebildet (ibid.).

Version	Silben/Sekunde	arithm. Mittel
deutlich03	7,25	7,30
undeutlich01	7,35	7,30
deutlich01	7,53	7,69
undeutlich03	7,85	7,69
deutlich02	8,26	8,32
undeutlich02	8,38	8,32

Tabelle 1: Exakte Sprechgeschwindigkeit und arithmetisches Mittel ähnlich schneller Versionen.

Jeder Signalausschnitt, der im paarweisen Vergleich als besser beurteilt wurde, erhielt einen Punkt. Um die unterschiedlichen Sprechgeschwindigkeiten aufeinander abbilden zu können, wurde die erhaltene Gesamtpunktzahl durch das jeweilige exakte Sprechtempo dividiert und anschließend mit dem vorher bestimmten arithmetischen Mittel multipliziert, so dass sich für die verschieden schnellen Versionen mit der skalierten Anzahl der Punkte ein vergleichbarer Wert ergab, der die Beurteilung der Deutlichkeit in Abhängigkeit vom Sprechtempo wiedergab. Anhand der grafischen Darstellung der Ergebnisse (siehe Abb. 3) wird bereits deutlich, dass die bewusst deutlicher gesprochenen Versionen signifikant (Chi-Test, $p < 0.001$) besser abschneiden als die undeutlicher artikulierten Versionen.

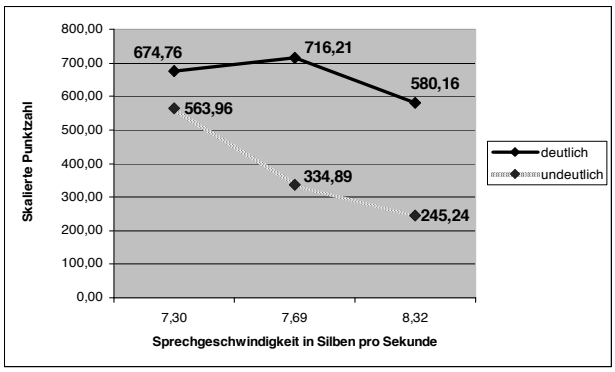

Abbildung 3: Punktzahl, abgebildet auf die jeweilige gemittelte Sprechgeschwindigkeit.

Aufgrund der Ergebnisse der perzeptiven Evaluation wird somit ebenso wie anhand der akustischen Untersuchung deutlich, dass es der ausgewählten Sprecherin möglich ist, unerwünschte Phänomene wie Koartikulation und Reduktion in bewusst deutlich artikulierter schnell gesprochener Sprache zu vermeiden und sie somit als Sprecherin für das Aufsprechen eines schnell gesprochenen, möglichst deutlichen Bausteininventars für die Unit-Selection basierte Sprachsynthese geeignet ist.

5 Ausblick

Die hier vorgestellte Sprecherevaluation ist Teil eines Projekts zur Modellierung hoher Sprechgeschwindigkeit in der Unit-Selection basierten Sprachsynthese. Um für die Modellierung verschiedener Sprechstile – insbesondere für schnell gesprochene Sprache – ein möglichst hochwertiges Syntheseinventar erstellen zu können, wurde unter Vorgabe bestimmter Kriterien ein geeigneter Sprecher gesucht. Nach der Durchführung einer Vorauswahl sollte die tatsächliche Eignung des Sprechers zum Aufsprechen eines schnell gesprochenen Bausteininventars mittels eines Perzeptionstests sowie anhand akustischer Merkmale der Sprachaufnahmen evaluiert und bestätigt werden.

Da sich die ausgewählte Sprecherin tatsächlich als für diese Aufgabe geeignet erwiesen hat, wird als nächstes ein Korpus in schneller Sprache aufgenommen werden, welches anschließend als Bausteininventar für die Erzeugung synthetischer schnell gesprochener Sprache in unserem Unit-Selection basierten Sprachsynthesesystem BOSS [19] verwendet und weiter evaluiert werden wird.

Literatur

[1] Moers, D.; Wagner, P.; Breuer, S. (2007): 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.

[2] Fellbaum, K. (1996): Einsatz der Sprachsynthese im Behindertenbereich. In *Fortschritte der Akustik. DAGA'96*. Oldenburg : DEGA. S. 78 – 81.

[3] Moos, A.; Trouvain, J. (2007): Comprehension of Ultra-Fast Speech – Blind vs. „Normally Hearing" Persons. In *Proceedings ICPhS XVI*. Saarbrücken. S. 677 – 684.

[4] Goldman-Eisler, F. (1961): The significance of changes in the rate of articulation. *Language and Speech*. Vol. 4, S. 171 – 174.

[5] Daniloff, R.G.; Hammarberg, R.E. (1973): On defining coarticulation. *Journal of Phonetics*. Vol. 1, S. 239 – 248.

[6] Kohler, K.J. (1990): Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In *Hardcastle, W.J.; Marchal, A. (eds.): Speech Production and Speech Modelling*. Dordrecht : Kluwer. S. 69 – 92.

[7] Peterson, G.E.; Lehiste, I. (1960): Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*. Vol. 32, S. 693 – 703.

[8] van Son, R. J. J. H.; Pols, L. C. W. (1996): An acoustic profile of consonant reduction. In *Proceedings ICSLP*. Philadelphia. S. 1529 – 1532.

[9] Gopal, H.S. (1990): Effects of speaking rate on the behaviour of tense and lax vowel durations. *Journal of Phonetics*. Vol. 18, S. 497 – 518.

[10] Crystal, T.H.; House, A.S. (1990): Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*.. Vol. 88, S. 101 – 112.

[11] Monaghan, A. (2001): An Auditory Analysis of the Prosody of Fast and Slow Speech Styles in English, Dutch and German. In *Keller, E.; Bailly, G.; Monaghan, A. et al. (eds.): Improvements in Speech Synthesis*. Chichester. S. 204 – 217.

[12] Greisbach, R. (1992): Reading aloud at maximal speed. *Speech Communication*. Vol. 11, S. 469 – 473.

[13] Breuer, S.; Abresch, J. (2004): Phoxsy: Multi-phone Segments for Unit Selection Speech Synthesis. In *Proceedings ICSLP*. Jeju.

[14] Lindblom, B. (1990): Explaining phonetic variation: A sketch of the H&H-Theory. In *Hardcastle, W.J.; Marchal, A.: Speech Production and Speech Modelling*. Dordrecht: Kluwer. S. 403 – 439.

[15] Maus, V. (2004): Zur Frage der Eignung von Sprechern als künstliche 'Stimme' in der konkatenativen Sprachsynthese. *Unveröffentlichte Magisterarbeit*. Rheinische Friedrich-Wilhelms-Universität Bonn.

[16] Dellwo, V.; Wagner, P. (2003): Relations between language rhythm and speech rate. In *Proceedings ICPhS*. Barcelona. S. 471 – 474.

[17] Janse, E. (2003): Production and Perception of Fast Speech. *Dissertation*. Universiteit Utrecht.

[18] Boersma, P.; Weenink, D.: Praat: Doing phonetics by computer. http://www.fon.hum.uva.nl/praat/ (20.06.2008).

[19] Klabbers, E.; Stöber, K.; Veldhuis, R.; Wagner, P.; Breuer, S. (2001): Speech synthesis development made easy: The Bonn Open Synthesis System. In *Proceedings of Eurospeech*. Aalborg.