Sprachsynthese mit prosodischen Phrasenbausteinen

Jürgen Trouvain[1,2] & Thomas Blug[1]
[1]*Institut für Phonetik, Universität des Saarlandes, Saarbrücken;* [2]*Phonetik-Büro Trouvain, Saarbrücken*
{trouvain;blug}@coli.uni-sb.de

Einleitung

Viele automatisierte Ansagen in Dialogsystemen und Auskunftsdiensten werden meist nicht durch eine vollständige Text-to-Speech-Synthese (TTS) erzeugt, sondern durch ganze aufgenommene Sätze (Voice-Prompts) oder durch die Zusammenfügung einzelner Wörter oder Satzteile ("Satzteilsynthese"). TTS-Synthese hat den Vorteil, jeden Text wiedergeben zu können, erreicht bei vielen Benutzern aber nicht die erforderliche Akzeptanz bezüglich Verständlichkeit und Natürlichkeit. Diesbezüglich sind Voice-Prompts besser geeignet, sie bieten aber keine Flexibilität bei der Textgenerierung. Eine Satzteilsynthese dagegen ist anpassungsfähiger, ihr fehlt allerdings die Natürlichkeit von Voice-Prompts.

Daher ist es günstig eine Satzteilsynthese mit einer Natürlichkeit ähnlich der von Voice-Prompts anzustreben. Beschreibungen solcher Arten von Sprachausgabe sind eher selten zu finden (siehe aber [1,2,4]). Eine phonetische Analyse einiger Beispiele aus alltäglichen Anwendungen hat ergeben, dass sehr häufig eine suboptimale *prosodische* Realisierung für die eher mittelmäßige Qualität der Satzteilsynthese verantwortlich ist.

Satzteilsynthese vs. Prosodische Phrasensynthese

An der Satzteilsynthese und ihren akustischen Bausteinen sind Auffälligkeiten auf folgenden prosodischen Ebenen zu vermerken:

- Rhythmus (arhythmisch und "abgehackt")
- Pausierung (der Phrasierungsstruktur nicht angemessen)
- Intonation (Position in prosodischer Phrase und Intonationskontur unberücksichtigt)
- Koartikulation (Konflikte auf lautlicher Kontextebene)
- Intensität (Schwankungen zw. akustischen Bausteinen)

Es bietet sich daher an, prosodische Phrasen statt den rein textbasierten Einheiten wie "Satzteile" oder Einzelwörter für die gesprochenen Aufnahmen als Grundeinheit zu verwenden.

Typischerweise liegen die Formulierungen, die in gesprochene Sprache umgesetzt werden sollen, als Output der Textgenerierung in zwei verschiedenen Klassen vor: zum einen als Trägersätze ("Satzschablonen") und zum anderen als Elemente, die variabel in diese Trägersätze eingesetzt werden können ("Füllsel", die in Beispiel (1) in abweichender `Schrift` stehen):

(1) Am `Tag` um `Zeit` gibt es keine Verbindung nach `Ort`.

Die Art der prosodischen Repräsentation bei der *Satzteilsynthese* besteht in einer 1:1-Umsetzung der Schablonen-Füllsel-Nahtstellen, so dass akustische Bausteine wie in Beispiel (2) aufgenommen und aneinander gefügt werden (hier durch eckige Klammern markiert):

(2) [Am `Tag`][um][`Zeit`][gibt es keine Verbindung nach][`Ort`].

Verwendet man dagegen prosodische angemessenere Repräsentationen, so sollten Stellen, an denen Pausen erwartet werden können, mit Schnittstellen zusammenfallen. Im Beispielsatz würde die Anzahl der Verkettungsbausteine um zwei reduziert werden:

(3) [Am `Tag`][um `Zeit`][gibt es keine Verbindung][nach `Ort`].

Die Konkatenationseinheiten entsprechen prosodischen Phrasen, die auch intonatorisch und rhythmisch bei der Aufnahme und auch bei der Zusammenfügung besser kontrolliert werden können.

Die "Füllsel" können in verschiedenen "Satzschablonen" und damit auch in verschiedenen lautlichen und prosodischen Kontexten vorkommen. Bei Erstellung des Aufnahmeskripts wird festgelegt, welche Signalbausteine für die Konkatenation vorliegen müssen. Dazu wird überprüft, welche Signalbausteine mehrfach verwendbar sind. Durch den Einsatz ein und desselben Signalbausteins in mehreren "Satzschablonen" wird die Größe des Aufnahmeskripts und die Anzahl der benötigten Signalbausteine erheblich reduziert. Zwar erhöht sich die Anzahl und auch die Länge der zur Verfügung stellenden Bausteine bei der prosodischen Phrasensynthese gegenüber der Satzteilsynthese geringfügig; doch die zu erwartende Qualitätssteigerung auf Grund höherer Natürlichkeit rechtfertigt diesen relativen Mehraufwand. Durch diese Methode bleibt die Anzahl der Phrasen und somit auch der Schnittstellen entweder gleich oder wird sogar reduziert.

Hörtest

Um die Natürlichkeit einer prosodisch fundierten Phrasensynthese zu überprüfen, wurde ein Präferenztest durchgeführt. Als Anwendung diente ein Dialogsystem, in dem die Sprachausgabe mit den drei genannten Methoden generiert wurde. Die Hypothese war, dass Voice-Prompting besser als prosodische Phrasensynthese und diese besser als "herkömmliche" Satzteilsynthese bewertet wird.

Methode

15 deutschsprachige Versuchspersonen mussten in einer Büroumgebung mit einem fingierten Dialogsystem drei Dialoge führen, wie sie auch per Telefon oder per Navigationsgerät im Auto in der Realität stattfinden. Der Experimentleiter lieferte die jeweiligen Systemäußerungen per Computer über handelsübliche PC-Lautsprecher. Vor jedem Dialog wurde den Versuchspersonen mitgeteilt, welche Information erfragt werden soll. Jeder Dialog bestand von Seiten des Systems aus vier Gespräch-Turns: Begrüßung, Rückfrage zur gewünschten Auskunft, erneute Rückfrage zur gewünschten Auskunft, komplette gewünschte Auskunft mit Nachfrage zu weiteren Auskunftswünschen.

Die drei Dialoge unterschieden sich zum einen bezüglich des Wortlautes im Text (aber nicht der Textsorte) und zum anderen hinsichtlich der Generierungsmethode der Sprachausgabe: Voice-Prompt, prosodische Phrasensynthese, Satzteilsynthese. Alle Stimuli wurden gemäß der oben erläuterten Beschreibung mit ein und derselben Sprecherin aufgenommen. Für die Begrüßung und die abschließende Nachfrage wurden für alle drei Bedingungen Voice-Prompts verwendet. Die drei Information enthaltenden Sätze waren zwischen zwei und 17 Bausteine lang (außer bei Voice-Prompts mit jeweils nur einem Baustein).

Zwei Bedingungen wurden getestet. Unter der einen Bedingung wurden die Versuchspersonen stark abgelenkt. Sie mussten parallel zur Dialogführung eine Aufgabe bewältigen, indem sie auf einem

Blatt Papier einen Weg durch ein Labyrinth finden und aufzeichnen mussten. Unter der anderen Bedingung wurden die Versuchspersonen nicht abgelenkt, sie mussten keine Zusatzaufgabe bearbeiten. Dementsprechend konnten sie sich auf die Dialogführung und damit verbunden auf die Qualität der Sprachausgabe konzentrieren. Um Listeneffekte (z.B. den "recency effect") auszubalancieren, wurden die Versuchspersonen in drei Gruppen mit unterschiedlicher Reihenfolge der Generierungsmethode aufgeteilt. Die Versuchspersonen sollten nach jedem Versuchsdurchgang angeben, welche Generierungsmethode sie als die beste für eine solche Anwendung halten und welche für die schlechteste.

Ergebnisse

Wie vermutet zeigen die Ergebnisse in Tabelle 1 klar, dass die Reihenfolge Voice-Prompt > prosodische Phrasensynthese > Satzteilsynthese eindeutig bevorzugt wird. Auch unter starker kognitiver Belastung der Versuchspersonen ist diese Reihenfolge der Präferenz zu beobachten, wenn auch weniger stark ausgeprägt.

Bedingung	*nicht abgelenkt*		*stark abgelenkt*	
Präferenz	*beste*	*schlechteste*	*beste*	*schlechteste*
Voice-Prompt	12	0	8	2
Prosod. Phrasensynthese	2	4	2	4
Satzteilsynthese	1	11	2	6
keine Präferenz	0	0	3	3

Tab. 1: Präferenzen der 15 Vpn unter den zwei Bedingungen.

Unklar bleibt, ob sich dieses Ergebnis auf die verschiedenartige Beanspruchung der Aufmerksamkeit in praktischen Anwendungen (Telefon, Auto, Informationskiosk) übertragen lässt (vgl. [6]).

Vergleich Phrasensynthese vs. Prompts

Um die prosodische Phrasensynthese zu verbessern, ist es angebracht, die Natürlichkeit der Voice-Prompts mit den synthetischen Äußerungen zu vergleichen. Dabei fallen die folgenden Unterschiede bezüglich ihrer prosodischen Ausformung auf. Die synthetischen Versionen zeigen eine geringere Artikulationsgeschwindigkeit auf (durchschnittlich 3,4 Silben/Sekunde gegenüber 4,1 Si/sec). Obwohl in der Phrasensynthese die Dauern der *physikalischen* Pausen mit denen der Voice-Prompts vergleichbar oder sogar kürzer sind, ist die Anzahl der *perzipierten* Pausen höher. Wahrgenommene Pausen können synonym für wahrgenommene Grenzen prosodischer Phrasen verwendet werden. Durch die erhöhte Anzahl der Phrasen wird der Eindruck des langsamen Tempos verstärkt, denn langsamere Rede wird oft durch eine erhöhte Anzahl von prosodischen Phrasen und auch mehr Satzakzenten gekennzeichnet [5]. Beide Charakteristika sind auch in der prosodischen Phrasensynthese vorhanden. Mehr Satzakzente werden durch mehr Tonhöhenbewegungen realisiert, was Grundfrequenzverläufe auch bestätigen. Der Umfang der Grundfrequenz hingegen ist bei beiden Fassungen ähnlich ausgeprägt.

Allgemeine Diskussion

Der Vorteil der prosodischen Phrasensynthese gegenüber der Satzteilsynthese besteht darin, dass prosodisch relevante Abschnitte vorbestimmt, prosodisch angemessen aufgenommen und dann konkateniert werden. Dies findet idealerweise an Stellen statt, an denen am wenigsten Konkatenationsverluste entstehen, nämlich an Pausen. Dieser Vorteil erweist sich aber als Nachteil gegenüber den natürlichen Voice-Prompts, da viele prosodische Phrasen zu einer als unangenehm empfundenen Verlangsamung der sprachlichen Wiedergabe führen.

Um dem Effekt der Verlangsamung entgegenzuwirken, ist es von Vorteil, die Länge der synthetisierten Sätze relativ kurz zu halten. Es sollten möglichst keine komplexen Satzstrukturen verwendet werden. Vielmehr erscheint es sinnvoll, die Informationen in "kleinen" Portionen anzubieten, um den Eindruck des langsamen Tempos zu minimieren. Denn werden die Sätze kürzer gehalten, reduziert sich auch automatisch die Anzahl der Phrasenbausteine pro Satz. Eine solche Portionierung erleichtert dem Anwender die Informationsaufnahme.

Zusammenfassung und Ausblick

Auch wenn nicht die Natürlichkeit von Voice-Prompts erreicht wird, zeigt eine Phrasensynthese, die prosodische Eigenschaften berücksichtigt und entsprechende Bausteine benutzt, einen entscheidenden Vorteil gegenüber den üblich verwendeten Satzteilsynthesen. Dieser Vorteil kann in vielen Anwendungen zu einer Erhöhung der Akzeptanz "synthetischer" Sprache genutzt werden. Gerade in Domänen, in denen typischerweise ein begrenztes Vokabular verwendet wird, erscheint diese Art der Synthese besonders gut geeignet. Dies gilt beispielsweise für die Bereiche Wettervorhersage, Flug- und Bahnauskunft, oder Sportergebnisdienst.

Die genannten Anwendungsgebiete zeigen sehr häufig stark typisierte prosodische Muster auf, wie die folgenden beiden Beispiele zeigen: Telefonnummern im Deutschen werden hauptsächlich durch drei prosodische Gruppierungsstrategien gekennzeichnet [1]. Fußballergebnisse im (britischen) Englisch werden rhythmisch und intonatorisch so gekennzeichnet, dass das Endergebnis vor der eigentlichen wörtlichen Information antizipiert werden kann [3]. Sobald die prosodischen Strukturen einer Textsorte bekannt sind, lassen sie sich hervorragend mit Hilfe prosodischer Phrasensynthese ausnutzen, was auch andere Studien gezeigt haben [4].

Allerdings gibt es auch Anwendungsbereiche in denen sich die Verwendung prosodischer Phrasensynthese als ungeeignet erweist. Dies gilt für alle Domänen, deren Vokabular sich schnell verändert, so z.B. für Kinoansagen. Alternativ zu allgemeinen TTS-Systemen sind solche Ansätze zu nennen, die eine korpusbasierte "non-uniform unit selection" als Methode verwenden und das Korpus so wählen, dass häufig vorkommende Wörter in vielen prosodischen Varianten vorkommen [2,7]. Eine perzeptuelle Evaluierung der störenden Artefakte durch "unvorbereitete" Wörter und Textpassagen steht allerdings noch aus. Fest steht, dass die Berücksichtigung prosodischer Strukturen sowohl in der korpusbasierten Synthese als auch in der Phrasensynthese zu erheblichen Verbesserungen führt und somit die Akzeptanz der jeweiligen Sprachausgabe steigert.

Referenzen

[1] Baumann, S. & Trouvain, J. (2001): On the prosody of German telephone numbers. *Proc. Eurospeech*, Aalborg, 557-560.

[2] Black, A.W. & Lenzo, K.A. (2000). Limited domain synthesis. *Proc. Intern. Confer. Spoken Lang. Proc. ICSLP* (2), Beijing, 411-414.

[3] Bonnet, G. (1980). A study of intonation in the soccer results. *Journal of Phonetics* **8**, 21-38.

[4] Elam, G. A. & Wayland, S. (1999). Prosody and prompt design in a computer dialog system. *Proc. ESCA Workshop on Dialogue and Prosody*, Veldhoven, 93-97.

[5] Trouvain, J. & Grice, M. (1999): The effect of tempo on prosodic structure. *Proc. Intern. Confer. Phonetic Sciences*, San Francisco, 1067-1070.

[6] Tsimhoni, O. & Green, P. (2001). Listening to natural and synthesized speech while driving: effects on user performance. *International Journal of Speech Technology* **4**, 155-169.

[7] Wagner P., Haas F., Stöber H., Helbig J. (1999): Multilinguale korpusbasierte Sprachsynthese auf der Basis domänenspezifischen Ausgangsmaterials. 10. Konfer. Elektron. Sprachsignalverarb. ESSV Görlitz, 152-159.