Zu Mustern der Pausengestaltung in natürlicher und synthetischer Lesesprache

Jürgen Trouvain & Bernd Möbius
(Sprachwissenschaft und Sprachtechnologie, Universität des Saarlandes)

Die Qualitätsbeurteilung von Text-to-Speech-Synthese (TTS) beschränkt sich meist auf Einzelsätze. Damit entgeht man zwar der Problematik, wie längere Abschnitte mit mehreren Sätzen methodisch sauber zu bewerten sind, allerdings bleibt dabei auch unklar, wie man längere Abschnitte prosodisch verbessert.

In der vorliegenden Studie haben wir vorgelesene Versionen desselben Textes ("Die drei kleinen Schweinchen", 13 Sätze) bezüglich der darin vorgefundenen Pausen untersucht (Parameter: Anzahl, Dauer, hörbare Atmung und Ort der Pausen). Das Material stammt von 10 zufällig ausgewählten Sprechern des IFCASL-Korpus [1] sowie von den TTS-Systemen MaryTTS [2] (2 verschiedene Stimmen) und 2 Versionen der bei Google benutzten Synthese bei Home [3] bzw. Translate [4].

Die Resultate zeigen, dass wie erwartet alle natürlichen Sprecher Pausen an den Satzgrenzen produzieren und diese nahezu immer mit Atemgeräuschen versehen sind. Diese Atempausen zeigen relativ lange Dauern (ca. 800 ms). Alle Sprecher produzieren auch Pausen innerhalb der Sätze (zwischen 10 und 20). Diese sind zumeist keine Atempausen und haben kürzere Dauern (ca. 300 ms). Pausen innerhalb von Sätzen konnten an 20 verschiedenen Wortgrenzen festgestellt werden, wobei aber nur 7 dieser Stellen von fast allen Sprechern zur Pausierung genutzt werden.

Die TTS-Systeme zeigen wie die natürlichen Sprecher auch Pausen an allen Satzgrenzen. Allerdings sind diese im Kontrast zu den natürlichen Sprechern ohne Atemgeräusch und mit 420 ms bzw. 640 ms beträchtlich kürzer. Bezüglich der Intra-Satz-Pausen differieren beide TTS-Systeme markant von den natürlichen Sprechern: die MaryTTS-Versionen weisen relativ viele Pausen auf. Diese sind im Schnitt auch noch etwas länger als die Zwischen-Satz-Pausen. Erschwerend kommt hinzu, dass manche dieser Pausen an ungewöhnlichen Stellen auftreten. Bei den Google-Versionen gibt es sehr wenige Pausen innerhalb der Sätze. Diese sind extrem kurz (weniger als 100 ms), was zu sehr langen pausenfreien Intervallen führt.

In einer Folgestudie ist geplant, die hier generierten TTS-Versionen mit solchen TTS-Versionen zu vergleichen, die nach natürlichen Vorbildern bezüglich der Pausen manipuliert wurden. Dazu würden längere Pausen an Satzgrenzen (eventuell mit Atemgeräusch) genauso gehören wie angepasste Dauern von Intra-Satz-Pausen. Die Auswahl der geeigneten Pausenstellen hängt dabei sowohl von der syntaktischen Struktur als auch von der Phrasenlänge ab, die nach dem Vorbild von Gee & Grosjean [5] modelliert werden könnte. Basierend auf den aktuellen Befunden kann die Hypothese aufgestellt werden, dass Hörer im Vergleich zu den manipulierten Versionen die existierenden TTS-Versionen als weniger adäquat empfinden und beim Zuhören weniger Informationen im Gedächtnis behalten.

Referenzen

[1] Trouvain et al. 2016. The IFCASL corpus of French and German non-native and native read speech. Proc. 9th Language Resources and Evaluation Conference (LREC), Portorož, pp. 1333-1338.
[2] http://mary.dfki.de/
[3] persönliche Anfrage bei Google London
[4] https://translate.google.com/?hl=de
[5] Gee, J. & Grosjean, F. 1983. Performance structures: A psycholinguistic and linguistic appraisal. Cognitive Psychology 15, pp. 411-458.