EINATMUNGSGERÄUSCHE VOR SYNTHETISCH ERZEUGTEN SÄTZEN – EINE PILOTSTUDIE

Jürgen Trouvain und Bernd Möbius

Universität des Saarlandes, Computerlinguistik und Phonetik
trouvain|moebius [at] coli.uni-saarland.de

Abstract: Die vorliegende Pilotstudie untersucht, ob das Einfügen kaum wahrnehmbarer Einatmungsgeräusche die Präferenz synthetischer Sprache positiv beeinflusst. Während in synthetischer Sprache Einatmungsgeräusche üblicherweise nicht vorkommen, sind sie in menschlicher Sprache mehr oder minder gut hörbar, werden aber meist "überhört". Experimente mit Formantsynthese zeigen, dass synthetisch erzeugte Sätze, denen Einatmungsgeräusche vorangestellt werden, besser memoriert werden als solche ohne [6]. Im hier berichteten Experiment werden mit konkatenativer Synthese erzeugte Telefonnummern verwendet. Ein Stimulus bestand aus zwei Kopien derselben Telefonnummer, wobei jeweils einer Kopie ein Einatmungsgeräusch vorangestellt wurde. Die elf Versuchspersonen hatten zu entscheiden, ob sie die erste oder die zweite Telefonnummer bevorzugen. Die Ergebnisse zeigen zwar für wenige Hörer die erwartete Präferenz der Version mit Einatmungsgeräusch, für einen Hörer allerdings eine negative Einstellung. Die Mehrheit zeigt weder eine Präferenz noch eine bewusstes Wahrnehmen dieser kurzen Vokalisierung. Die Hypothese, dass synthetisch erzeugte Äußerungen durch Voranstellen hörbarer Einatmungsgeräusche positiver wahrgenommen werden als solche ohne, konnte durch das vorgestellte Hörexperiment nicht im angenommenen Umfang bestätigt werden. Es kann nicht ausgeschlossen werden, dass die fehlende Kongruenz zwischen dem Sprecher der Synthese und dem "Sprecher" des Atmungsgeräusches sowie die fehlende Variation der Einatmungsgeräusche zu diesem Befund beigetragen haben. Die vorliegende Pilotstudie zeigt bezüglich der Steigerung der "Natürlichkeit" synthetischer Sprache zum einen, wie komplex sich die Beziehung zwischen Modellierung und Analyse des natürlichen Vorbilds verhält, zum anderen deutet die Studie aber auch den potenziellen Nutzen, aber auch das Risiko dieser Modellierung an.

1 Einführung

Ziel der vorliegenden Studie ist, zu überprüfen, ob kaum wahrnehmbare Einatmungsgeräusche die Präferenz synthetischer Sprache positiv beeinflussen.

Einer von vielen Unterschieden zwischen synthetischer und menschlicher Sprache besteht in der Ab- bzw. Anwesenheit von Einatmungsgeräuschen. Oftmals geht menschlichen Äußerungen ein durch verstärktes Einatmen bedingtes Geräusch voraus. Auch wenn diese Geräuschlaute mehr oder minder gut hörbar sind und ihre spektrale Energie entsprechend im Sprachsignal beobachtbar ist, so scheinen Hörer diesem selten Beachtung zu schenken. Annotationen spontansprachlicher Korpora spiegeln dieses "Überhören" wieder [5].

Einatmungsgeräusche können aber auch wesentlicher Bestandteil der Äußerung sein, bei der ein "Überhören" eigentlich ausgeschlossen ist. Dies ist bei affektiven Äußerungen zum Ausdruck von Überraschung und Erschrecken der Fall (z.B. [13]). Daher werden in den wenigen Ansätzen, Spontansprache für Synthese zu modellieren, auch Atemgeräusche [4] verwendet, insbesondere für expressive Sprachsynthese [7].

Für die Kombination von Atmungslauten und formantsynthetisch erzeugten Sätzen haben Whalen et al. [6] herausgefunden, dass Sätze mit vorangegangenem Atmungssegment eine bessere Erinnerungsleistung bei Zuhörern zur Folge haben als solche ohne (bzw. solche mit einem nicht-menschlichen Geräusch wie Laubrascheln). Ungeklärt bleibt allerdings, ob synthetisch erzeugte Sätze *mit* hörbarer Einatmung auch zu einer Bevorzugung gegenüber solchen Sätzen *ohne* Einatmung führen. Ebenfalls ungeklärt ist, ob das hörbare Einatmen dem Hörer auch bewusst wird und er aus diesem Grund eine bestimmte Version bevorzugt.

Da in [6] ein Bewusstwerden des Einatmungsgeräuschs nicht erwähnt wurde und sich dies mit dem oben erwähnten Annotationsgebrauch deckt, ist unsere erste Hypothese, dass in unserer Pilotstudie niemand das Einatmen bewusst wahrnimmt. In Übereinstimmung mit [6] ist die zweite Hypothese, dass synthetisierte Sätze mit Einatmungslaut solchen Sätzen ohne bevorzugt werden.

2 Methode

2.1 Erzeugung der Stimuli

Insgesamt wurden zehn Telefonnummern vom Muster 0421 – 369 – 2781 mit deutschen Stimmen zweier Sprachsynthesen erzeugt: Festival [2] und Mary [3]. Es wurde darauf geachtet, dass die prosodische Phrasierung derjenigen natürlicher Sprache ähnelt [1]. Dazu wurden auch nachträglich die Pausendauern angepasst.

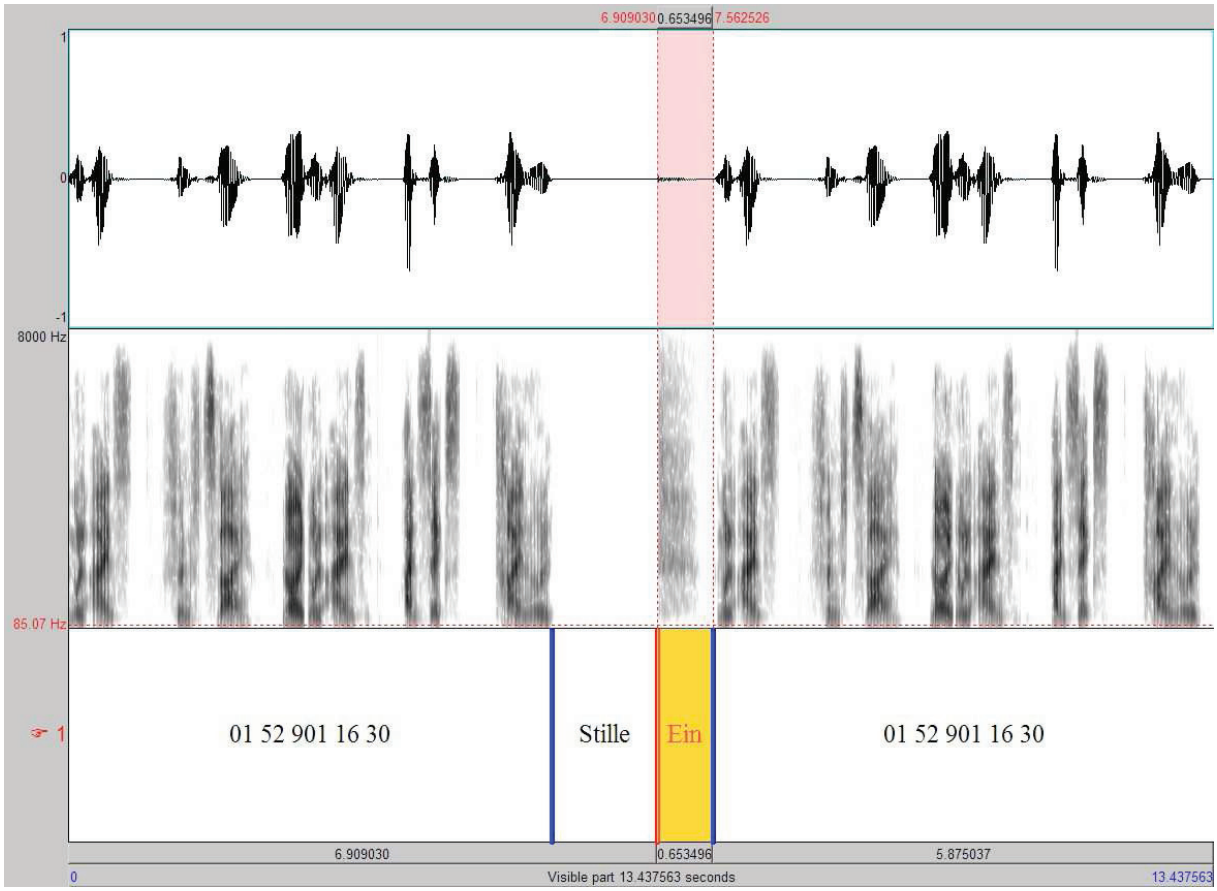

Abbildung 1 – Sprachsignal mit Wellenform und Spektrogramm eines verwendeten Stimulus beginnend mit der Telefonnummer 0152 – 901 – 16 30 (erzeugt mit der Sprachsynthese Festival [2]), gefolgt von einer Sekunde Stille. Die zweite Hälfte des Stiumulus beginnt mit dem

Einatmugsgeräusch und konkatenierter Stille von 200 ms (eingefärbt), dem wiederum eine Kopie der Telefonnummer folgt.

Das verwendete Einatmungsgeräusch stammte aus der Spontansprache eines natürlichen Sprechers, der aber nicht identisch ist mit einem der "Spender" der synthetischen Stimmen. Ein Stimulus bestand aus zwei Kopien derselben Telefonnummer (durch jeweils 1 Sekunde Stille voneinander getrennt). Einer der beiden Nummern wurde das ausgewählte Einatmungsgeräusch (Dauer: 480 ms plus 200 ms Stille) vorangestellt (vgl. Abbildung 1). Jedes Telefonnummernpaar kam zwei Mal vor: einmal mit der hörbaren Atmung vor der ersten Nummer, beim zweiten Mal vor der zweiten Nummer. Jeder der so erzeugten 20 Stimuli wurde doppelt vorgespielt, so dass insgesamt 40 Stimuli zu bewerten waren.

2.2 Versuchspersonen

Als Versuchspersonen (Vpn) fungierten 11 Studenten (davon 3 weiblich), die via Kopfhörer die Stimuli in randomisierter Reihenfolge vom Experimentleiter in einem ruhigen Büro vorgespielt bekamen. Nach jedem Stimulus mussten die Vpn entscheiden, ob sie die Version A oder B bevorzugten. Nach zwanzig Stimuli wurde eine kleine Pause eingelegt. Insgesamt dauerte diese Testphase ca. 12 Minuten.

Nach dem Test wurde jede Vp gefragt, ob sie einen Unterschied zwischen den beiden Versionen eines Paares gehört haben. Falls die Vp Einatmung als Unterscheidungsmerkmal genannt hat, wurde weiterhin gefragt, ob und in welche Richtung diese Wahrnehmung ihre Präferenz gelenkt hat.

3 Ergebnisse

Die Mehrzahl der Vpn (7 von 11) nahm das Einatmungsgeräusch *nicht* wahr (siehe Abbildung 2). Diese Vpn nahmen aber bei einem erneuten Abhören nach dem Test dieses Geräusch überrascht zur Kenntnis. Als mögliche Unterschiede zwischen den bis auf das Geräusch identischen Versionen wurden Variationen in Lautstärke, Intonation, Dauer und Pausenlänge angegeben. Nur bei einer dieser sieben Vpn war eine leichte Tendenz zu der Bevorzugung der Nummern mit vorangegangenem Atmungsgeräusch zu beobachten, die Ergebnisse der anderen sechs Vpn bewegen sich auf Zufallsniveau.

Der Minderheit der Vpn (4 von 11) fiel das Einatmungsgeräusch im Laufe des Tests bewusst auf. Für zwei Vpn führte diese Bewusstmachung dazu, die Nummern mit hörbarer Einatmung zu präferieren, bei den zwei anderen Vpn war dies aber – entgegen der anfänglich geäußerten Erwartung – nicht der Fall. Entweder wurde angegeben, dass man "mal so, mal so" entschied oder dass man sich bewusst *gegen* die Nummer mit dem Geräusch entschieden habe (Vp 5). Bei diesen beiden Vpn konnte im Gegensatz zu allen anderen Vpn eine eindeutige Ablehnung des Atemgeräuschs in Kombination mit einer der beiden synthetischen Stimmen festgestellt werden. Allerdings sind in dieser Beziehung beide Vpn offenbar unterschiedlicher Ansicht, welche der beiden Synthesen als unpassend zum natürlichen Atmungsgeräusch gilt. Alle anderen Vpn zeigten keine bemerkenswerte Bevorzugung oder Ablehnung einer der beiden Synthesestimmen.

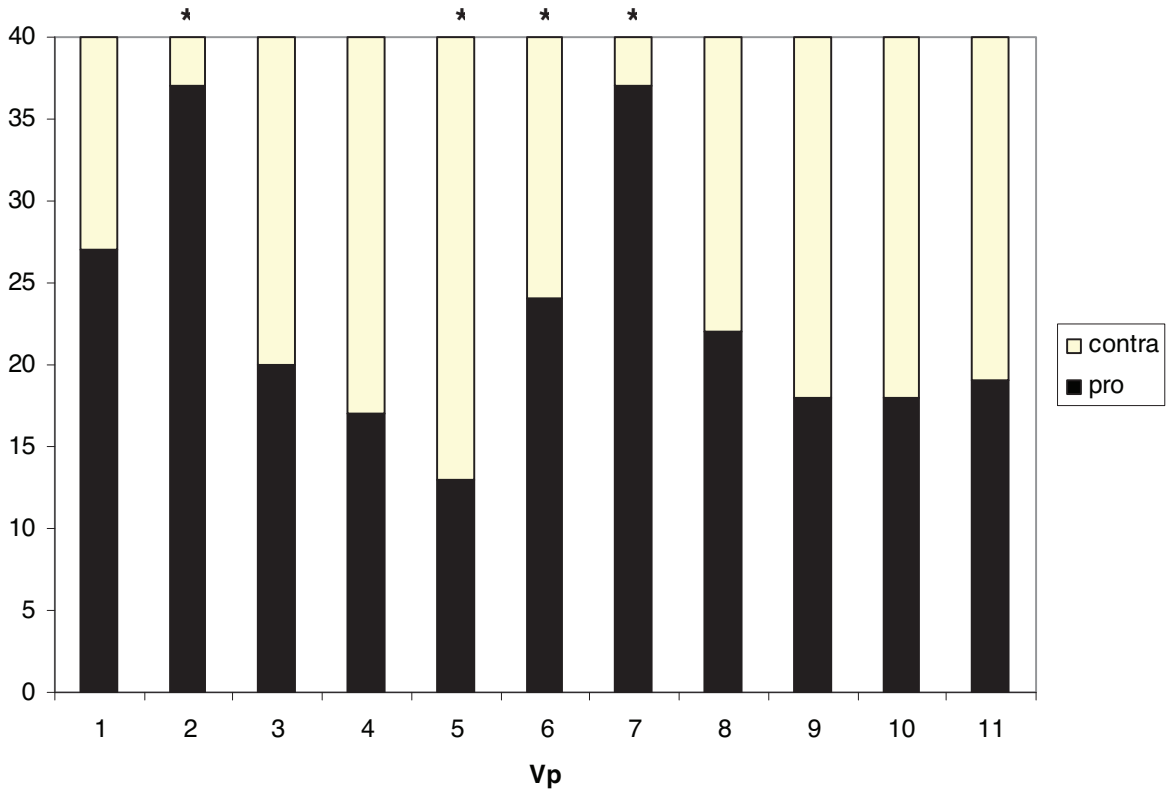

Abbildung 2 – Präferenzen der 11 Vpn: Telefonnummer mit Einatmungsgeräusch (dunkel) vs. ohne Einatmungsgeräusch (hell). Das Zufallsniveau liegt bei 20. Die 4 Vpn, denen die Einatmungsgeräusche bewusst geworden sind, sind mit Sternchen markiert.

4 Diskussion

Unsere Hypothese, dass synthetisch erzeugte Äußerungen durch Voranstellen hörbarer Einatmungsgeräusche positiver wahrgenommen werden als solche ohne, konnte durch das vorgestellte Hörexperiment nicht – beziehungsweise nicht im angenommenen Umfang – bestätigt werden. Es kann nicht ausgeschlossen werden, dass die fehlende Kongruenz zwischen dem Sprecher der Synthese und dem "Sprecher" des Atmungsgeräusches sowie die fehlende Variation der Einatmungsgeräusche zu diesem Befund beigetragen haben.

4.1 Kongruenz der Sprecher in der Synthese

Der eindeutigen Bevorzugung der zusätzlich manipulierten Versionen bei zwei Vpn steht die (partielle) Ablehnung durch andere Vpn gegenüber. Hier scheinen die im Atmungselement enthaltenen Informationen über den Sprecher nicht ausreichend kongruent zu denen der beiden synthetischen Stimmen zu sein.

Bei der Studie von Whalen et al. [6] wurde dieser Aspekt insofern berückischtigt, dass sich die Vokaltrakteigenschaften des "Atmungssprechers" mit dem Sprechermodell für die Formantsynthese ähneln. Bei der Weiterführung der Testreihe muss daher der Aspekt der Sprecherkongruenz unbedingt beachtet werden.

4.2 Variation der Einatmungsgeräusche

Studien über die so häufig produzierten Einatmungsgeräusche zeigen durchaus eine beachtliche inter-individuelle Variation, aber auch eine bemerkenswerte intra-individuelle Variation. Lauf [8] beispielsweise zeigt, dass sich beim lauten Lesen die Sprecher bezüglich der Dauer und Intensität ihrer in Sprechpausen produzierten Einatmungsgeräusche sehr unterschiedlich verhalten können. Dieser Befund wird durch forensisch orientierte Arbeiten (z.B. [9]) sowie durch Studien mit Spontansprache bestätigt [11].

Die Whalen-Studie [6] benutzt drei verschiedene Atmungsgeräusche, die sich in erster Linie in ihrer Dauer unterscheiden: von ca. 600 ms bis über 700 ms, jeweils gefolgt von 300 ms Stille. In der vorliegenden Pilotstudie wurde nur ein einziges Geräusch benutzt, wobei Geräusch und darauffolgende Stille kürzer waren.

Neben individualtypischen Eigenschaften spielt auch die Kontrolle und Planung linguistischer Einheiten eine Rolle bei unterschiedlich ausgeprägten Einatmungslauten. Die Sprechplanung langer Phrasen bedingt eine größere Einatmungstiefe mit darauffolgendem intensiveren und/oder längerem Einatmungsgeräusch im Vergleich zur Planung kurzer Phrasen [10, 11].

Auf der paralinguistischen Ebene spielen die Produktion und Perzeption von Einatmungsgeräuschen bei den bereits genannten Affektäußerungen wie zum Ausdruck von Schreck eine zentrale Rolle [13]. Aber hörbare Einatmung findet auch als Marker der Höflichkeit Verwendung, z.B. im Koreanischen [12].

4.3 Unbewusste Verarbeitung hörbarer Einatmungsgeräusche

Das in synthetischer Sprache eingesetzte Einatmungeräusch wurde mehrheitlich unbewusst verarbeitet. Möglicherweise werden die meisten Einatmungsgeräusche generell nicht bewusst verarbeitet. Daher würde dieser Befund zu der relativ geringen Konsistenz bei der Annotation natürlichsprachlicher Daten [5] passen. Auch in der Whalen-Studie [6] wird ein bewusstes Wahrnehmen der Einatmung durch die Vpn nicht berichtet (was eine gelegentliche Wahrnehmung aber nicht ausschließt).

4.4 Positive Beeinflussung synthetischer Sätze durch hörbare Einatmungsgeräusche

Nur bei zwei aus elf Vpn konnte gemäß der Hypothese eine positive Beeinflussung, bei nur einer Vp konnte eine negative Beeinflussung festgestellt werden. Die große Mehrheit blieb unbeeinflusst. Dieses Ergebnis steht im Gegensatz zu [6], wo allerdings ein Recall-Test mit einzelnen Sätzen und kein Präferenztest durchgeführt worden ist.

Sollten weitere Tests ergeben, dass sprecherkongruente Atmungsgeräusche bei synthetischer Sprache *nicht* zu negativen Ergebnissen führen, so besteht durchaus die Möglichkeit, dass bei einigen Hörern eine solche Verwendung zu einer positiveren Einstellung synthetischer erzeugter Sprache führt.

4.5 Fazit

Die vorliegende Pilotstudie zeigt bezüglich der Steigerung der "Natürlichkeit" synthetischer Sprache zum einen, wie komplex sich die Beziehung zwischen Modellierung und Analyse des natürlichen Vorbilds verhält - ein bloßes Abbilden vorgefundener Eigenschaften reicht nicht aus. Zum anderen deutet die Studie aber auch den potenziellen Nutzen, aber auch das Risiko dieser Modellierung an. Zukünftige Studien sollten zum Ziel haben, durch die entsprechende Berücksichtigung der Sprecherkongruenz und der Variation der phonetischen Ausprägung von Atmungsgeräuschen die hier erkannten Risiken auszuschließen sowie in erweiterten Test-Settings den Nutzen zu überprüfen.

Literatur

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

[2] http://www.ims.uni-stuttgart.de/phonetik/synthesis/

[3] http://mary.dfki.de/

[4] Sundaram S. & Narayanan S. 2003. An empirical text transformation method for spontaneous speech synthesizers. Proc. Interspeech, Geneva, pp. 1221-1224.

[5] Trouvain, J. & Truong, K. 2012. Comparing non-verbal vocalisations in conversational speech corpora. Proc. International Workshop on Corpora for Research on Emotion Sentiment & Social Signals, Istanbul, pp. 36-39.

[6] Whalen, D.H., Hoequist, Ch.E. & Sheffert, S. 1995. The effects of breath sounds on the perception of synthetic speech. Journal of the Acoustical Society of America 97, pp. 3147-3153.

[7] Yuan, Ch. & Li, A. 2006. The breath segment in expressive speech. International Symposium on Chinese Spoken Language Processing, Singapore, paper B1.

[8] Lauf, R. 2001. Aspekte der Sprechatmung: Zur Verteilung, Dauer und Struktur von Atemgeräuschen in abgelesenen Texten. In: Braun, A. (Hg.) Beiträge zu Linguistik und Phonetik. Festschrift für Joachim Göschel zum 70. Geburtstag (ZDL Beihefte 118). Stuttgart: Franz Steiner Verlag., pp. 406-420.

[9] Kienast, M. & Glitza, F. 2003. Respiratory sounds as an idiosyncratic feature in speaker recognition. Proceedings 15th International Congress of the Phonetic Sciences (ICPhS), Barcelona, pp. 1607-1610.

[10] Winkworth, A.L., Davis, P.J., Adams, R.A. & Ellis, E. 1995. Breathing patterns during spontaneous speech. Journal of Speech and Hearing Research 38, pp. 124-144.

[11] Fuchs, S., Petrone, C., Krivokapić, J. & Hoole, Ph. 2013. Acoustic and respiratory evidence for utterance planning in German. Journal of Phonetics 41, pp. 29-47.

[12] Winter, B. & Grawunder, S. 2012. The phonetic profile of Korean formal and informal speech registers. Journal of Phonetics 40, pp. 808-815.

[13] Schröder, M., 2003. Experimental study of affect bursts. Speech Communication 40(1-2), pp. 99-116.