EXPERIMENTE ZUR WAHRNEHMUNG GEZIELT DEGRADIERTER SYNTHETISCHER SPRACHE

Eva Lasarcyk[1]*, Heiner Drenhaus*[2] *und Bernd Möbius*[1]
Universität des Saarlandes, FR 4.7, [1]*Phonetik /* [2]*Psycholinguistik*
Email: {evaly,drenhaus,moebius}@coli.uni-saarland.de

Kurzfassung: In dieser Studie wurde die Wahrnehmung natürlicher und gezielt manipulierter und in ihrer Qualität degradierter synthetischer Sprache untersucht. Die auditive Bewertung bezog sich auf natürliche Sprachaufnahmen sowie auf Stimuli, die das Sprachsynthesesystem auf der Basis derselben Stimme generierte. Die synthetischen Stimuli wurden hinsichtlich der Häufigkeit und Güte von Konkatenationen manipuliert. Es ergab sich ein signifikanter Unterschied zwischen den Stimulusqualitäten. Leicht degradierte Stimuli erhielten, wie intendiert, schlechtere Bewertungen als die Baseline-Stimuli und bessere Bewertungen als die stark degradierten Stimuli. Objektive Distanzmaße wurden ebenfalls berechnet und erwiesen sich als kongruent mit den subjektiven Bewertungen.

1 Einführung

Die Motivation für die vorliegende Studie leitet sich aus Ergebnissen einer Untersuchung mit Hilfe der funktionellen Kernspintomographie (*functional Magnetic Resonance Imaging*, fMRI) ab, die darauf hindeuten, dass bestimmte Areale im mittleren frontalen Gyrus (MFG) nahe der präzentralen Furche der linken Gehirnhemisphäre natürliche und synthetische sprachliche Stimuli unterschiedlich verarbeitet [3]: "A significant dissociation between natural and synthetic speech was observed only in the left dorsolateral precentral sulcus, being more active for natural than synthetic speech. The left precentral sulcus focus may be related to differences in the degree to which mnemonic processes are engaged by the intrinsically more familiar natural speech [...]" (p. 383). Diese Hirnareale sind bei der Wahrnehmung natürlichsprachlicher Stimuli signifikant aktiver als bei der Wahrnehmung synthetischer Sprachstimuli. Unsere Interpretation dieser Ergebnisse ist, dass zumindest die im Experiment von Benson et al. verwendete Synthesemethode (Formantsynthese) ungeeignet ist, beim Hörer eine Assoziation von Sprechereigenschaften auf der Basis des Sprachsignals auszulösen, die bei der Perzeption natürlicher Sprache jedoch unweigerlich erfolgt. Wir postulieren, dass dieser differenzielle Effekt eine Ursache für die Einschätzung der Formantsynthese, und vermutlich auch der mit anderen Methoden erzeugten synthetischen Sprache, als unnatürlich ist. Eine zentrale Forschungsfrage ist daher, ob sich mit den aktuellen Syntheseverfahren eine Sprachqualität erzielen lässt, die vom Gehirn des Hörers insoweit als der natürlichen Sprache adäquat wahrgenommen wird, dass sie ähnliche Assoziationen von Sprecher- und Stimmmerkmalen auslösen kann.

Um den Raum verfügbarer Synthesemethoden und Sprachqualitäten abzudecken, wird derzeit ein Inventar von Stimuli mit Hilfe unterschiedlicher Methoden der Sprachsynthese erstellt. Die Methoden umfassen Formantsynthese, Unit-Selection-Synthese sowie statistisch-parametrische Synthese. Die für die Einschätzung der Natürlichkeit relevanten akustischen Dimensionen betreffen die Repräsentation von Stimmmerkmalen (natürlich vs. modelliert vs. generiert) sowie die Häufigkeit und Güte der Verkettung akustischer Synthesebausteine (natürlich vs. glatt-modelliert vs. unterschiedlich starke Diskontinuitäten im Zeitsignal und im Spektrum).

In einem Pilotexperiment, dessen Ergebnisse in dem vorliegenden Aufsatz präsentiert werden, wurde die Wahrnehmung natürlicher und gezielt manipulierter und in ihrer Qualität degradierter synthetischer Sprache untersucht. Wir gingen der Frage nach, ob die Hörer bestimmte akustische Manipulationen von Stimuli bewusst wahrnehmen können. Die Stimuli bestanden aus natürlichen Sprachaufnahmen sowie aus Stimuli, die das Sprachsynthesesystem auf der Basis derselben Stimme generierte. Die synthetischen Stimuli wurden hinsichtlich der Häufigkeit der Konkatenationsstellen und der Stärke der dort auftretenden Diskontinuitäten manipuliert. Es ergab sich ein signifikanter Unterschied zwischen den Stimulusqualitäten. Leicht degradierte Stimuli erhielten, wie intendiert, schlechtere Bewertungen als die Baseline-Stimuli und bessere Bewertungen als die stark degradierten Stimuli. Objektive Distanzmaße wurden ebenfalls berechnet und erwiesen sich als kongruent mit den subjektiven Bewertungen.

2 Synthesemethode

2.1 Basisstimme

Als Basisstimme für die experimentellen Stimuli wurde die männliche BITS3 Unit-Selection-Stimme des Sprachsynthesesystems MARYTTS [6] verwendet. Die Wahl fiel aus mehreren Gründen auf diese Stimme. Zum einen erlaubt es die Unit-Selection-Methode, die Anzahl der Einheiten und somit der Verkettungsstellen sowie die Stärke der Diskontinuität an der Verkettungsstelle auf algorithmischer Ebene zu beeinflussen. Auf diese Weise können gezielt lokale Sprünge im synthetischen Sprachsignal eingeführt werden. Zum anderen enthält das lautsprachliche Korpus, das zur Konstruktion der Synthesestimme diente, eine hinreichend große Zahl vollständiger und strukturell für das Experiment geeigneter Sätze. Die Originalaufnahmen im Korpus wurden im Experiment als Referenz (“Baseline”) verwendet, um die natürliche Stimme und die auf ihr beruhenden synthetischen Versionen miteinander vergleichen zu können.

2.2 Manipulation der Synthesequalität

Der Einheitenauswahlalgorithmus von MARYTTS kann durch verschiedene Parameter über Konfigurationsdateien beeinflusst werden. Der zentrale Parameter für das vorliegende Experiment, WTC (*weight of target cost function*), beeinflusst das relative Gewicht der Zielkosten gegenüber den Verkettungskosten. Die Zielkosten steigen, wenn die symbolischen oder akustischen Merkmale eines Einheitenkandidaten von der gewünschten Zielspezifikation abweichen. Die Verkettungskosten steigen, wenn zwei benachbarte Einheiten eine Diskontinuität an der Verkettungsstelle hervorrufen. Der Standardwert für WTC in MARYTTS ist 0,7; wird der Wert erhöht, wodurch die Zielspezifikation mehr Gewicht erhält, müssen Kompromisse in der Anzahl der Verkettungsstellen und in der Güte der Verkettung hingenommen werden. Dieser Effekt entsteht, weil die stärker gewichtete Anforderung an die Erfüllung der Zielvorgaben zu einer schlechteren Passgenauigkeit über die Verkettungsstellen hinweg führt.

Im Experiment wurden die leicht degradierten Stimuli mit WTC=0,9 und die stark degradierten Stimuli mit WTC=5,0 generiert. Diese Werte wurden empirisch in einem Vortest ermittelt. Jenseits des Wertes von 0,9 beginnt die Aussprache- und Stimmqualität der synthetischen Äußerungen wahrnehmbar nachzulassen, und ab einem Wert von 5,0 wird die Synthese in der Regel unverständlich. In einem webbasierten Vortest mit 16 Hörern wurde verifiziert, dass selbst die stark degradierten experimentellen Stimuli noch verständlich war. Fast alle Zielwörter wurden korrekt identifiziert, jedoch wurde in sehr wenigen Fällen die Wortform falsch erkannt oder das Zielwort mit einem ähnlich klingenden Wort verwechselt. Die MOS-Werte (*Mean Opinion Scores*) für Verständlichkeit lagen zwischen 1,25 und 3,63 (Tabelle 1).

Zielwort	#korrekt	Alternative	MOS (S.D.)
öffnet	15	wühlt (1)	1,25 (0,58)
öffnet-2	16		1,31 (0,60)
bestimmen	16		1,38 (0,50)
kamen	16		1,38 (0,50)
steigen	16		1,38 (0,62)
benutzte	16		1,63 (0,62)
besticht	16		1,75 (0,77)
kostet-2	16		1,81 (0,54)
vertreten	16		1,81 (0,54)
flüstert-2	16		1,81 (0,83)
traten-2	15	rannten (1)	1,88 (0,72)
bedeutete	16		1,94 (0,68)
traten	16		1,94 (1,00)
besuchte	16		2,00 (0,63)
kommen	16		2,00 (0,73)
vertreten-2	16		2,06 (0,68)
berichtete	16		2,06 (0,77)
drohen	15	droht (1)	2,06 (0,77)
streichelt-2	16		2,06 (0,77)
kommen-2	16		2,13 (0,72)
bleiben-2	16		2,19 (0,66)
kritisieren	16		2,19 (0,75)
bleiben	16		2,31 (0,60)
berät-2	16		2,31 (0,79)
spielten	16		2,38 (0,89)
veränderte	16		2,44 (0,73)
flüstert	16		2,50 (0,82)
fordert	16		2,50 (0,89)
kombinieren	15	komplettierten (1)	2,50 (0,97)
besuchte-2	16		2,56 (0,63)
reagieren	15	reagiert (1)	2,56 (0,73)
warnte-2	16		2,56 (0,73)
kostet	16		2,56 (0,81)
berät	16		2,69 (0,87)
warnte	15	fragte (1)	2,75 (0,45)
streichelt	16		2,75 (0,68)
endeten	16		2,75 (0,93)
interagiert	16		2,88 (0,96)
erläuterte	16		2,94 (1,06)
repräsentierte	15	präsentierte (1)	3,25 (0,58)
arrangierte	16		3,25 (0,86)
rühmen	14	rühmten (2)	3,63 (0,62)

Tabelle 1 - Ergebnisse des Vortests: Zielwörter, Anzahl der korrekten Antworten, alternative Antworten, MOS-Werte (Durchschnitt und Standardabweichung) für Verständlichkeit. "*Verb*-2" steht für das zweite Auftreten des Verbs in einem anderen Trägersatz.

Nr.	Text	Struktur
1	Die Parlaments-Vizepräsidentin warnte die Kommissionsmitglieder eindringlich.	NP V NP
2	Björn besuchte die deutsch-tschechische Kunstausstellung.	NP V NP
3	Der Lehrer benutzte den Projektor im Unterricht.	NP V PP
...	...	...

Tabelle 2 - Sprachmaterial für das Perzeptionsexperiment (Sätze 1–3 von 42).

3 Stimulusmaterial

3.1 Sprachmaterial

Das Sprachmaterial für das Perzeptionsexperiment bestand aus 42 Sätzen, die die gleiche syntaktische Struktur aufweisen: Nominalphrase, Verb, Nominal- und/oder Präpositionalphrase (NP V NP/PP; siehe Tabelle 2).

3.1.1 Stimuluskonstruktion

Die Baseline-Stimuli wurden direkt aus dem Sprachsynthesekorpus entnommen. Diese Originalaufnahmen wurden abgesehen von einer globalen Lautstärkeanpassung auf 65 dB nicht manipuliert. Synthetische Versionen der Originale wurden mit MARYTTS unter Verwendung unterschiedlicher Einstellungen des WTC-Parameters (0,9 vs. 5,0) generiert. In den experimentellen Stimuli sollten nur die Zielwörter (Verben) qualitativ degradiert werden. Daher wurden zunächst die vollständigen Sätze mit den jeweiligen Parametereinstellungen synthetisiert, um die gewünschten prosodischen Eigenschaften der Zielwörter zu erhalten. Anschließend wurden die Zielwörter zwischen ihrer ersten und letzten Einheitengrenze extrahiert, an die Lautstärke in der Originalaufnahme angepasst und in die Trägersätze eingefügt. Auf diese Weise wurde ein Inventar von Stimuli erzeugt, in dem sich die manipulierten Stimuli nur in der akustischen Qualität des jeweiligen Zielwortes (Verbs) von den Baseline-Stimuli unterschieden.

Das Stimulusinventar umfasste somit 42 Sätze jeweils in drei akustischen Versionen: Baseline (Referenzqualität), leicht degradiert und stark degradiert; insgesamt 126 Sätze. Die Stimulusliste wurde durch die doppelte Anzahl von Sätzen ergänzt, die nach ähnlichen Kriterien wie die Referenzsätze aus dem Sprachsynthesekorpus ausgewählt wurden, aber nicht Gegenstand der Auswertung waren.

3.1.2 Akustische Eigenschaften

Infolge der verwendeten Synthesestrategie wiesen die qualitativ degradierten Zielwörter wie beabsichtigt suboptimale akustische Transitionen an den Verkettungsstellen der einzelnen Einheiten auf. Das Ausmaß der Degradation war eine Funktion der Anzahl der Bausteine und somit Konkatenationen sowie der Stärke der Diskontinuitäten. Je nach phonologischer Länge des Zielwortes variierte die Anzahl der Verkettungsstellen zwischen 2–5 (WTC=0,9, leicht degradiert) und 5–13 (WTC=5,0, stark degradiert) (Tabelle 3).

Als zusätzliche Information über die akustische Qualität der Stimuli berechneten wir Werte für *Mel Cepstral Distortion* (MCD) und *Perceptual Evaluation of Speech Quality* (PESQ). MCD wird häufig im Kontext der Sprachsynthese als ein objektives Maß verwendet, um den akustischen Unterschied zwischen synthetisierten und natürlichen Stimuli quantitativ auszudrücken [7]. Je kleiner die MCD-Werte, desto näher kommt die synthetische Sprache dem Ideal, akustische Eigenschaften der natürlichen Sprache zu reproduzieren. Die über die gesamte

	#Konkat.			Δ MCD			PSEQ		
Satz	Ref	WTC09	WTC50	WTC09	WTC50	Diff	WTC09	WTC50	Diff
1	0	3	6	1,027	1,490	0,464	**3,648**	2,787	0,861
2	0	3	6	0,976	1,337	0,361	3,226	2,511	0,715
3	0	3	7	0,744	1,249	0,504	2,484	2,119	0,365
4	0	3	7	0,837	1,267	0,429	2,585	2,001	0,584
5	0	2	7	1,566	1,430	-0,136	2,089	1,976	0,113
6	0	3	9	1,228	1,462	0,233	1,889	1,856	0,033
7	0	3	10	0,807	0,876	0,068	2,262	2,007	0,255
8	0	3	9	1,143	1,229	0,085	3,332	2,792	0,540
9	0	4	13	0,847	0,990	0,144	2,170	2,268	-0,098
10	0	2	6	1,180	1,437	0,257	2,590	2,727	-0,137
11	0	2	6	1,188	1,794	0,606	3,196	2,715	0,481
12	0	2	7	1,003	1,073	0,069	3,130	2,529	0,601
13	0	2	5	0,970	1,032	0,062	3,451	3,226	0,225
14	0	3	8	1,342	1,358	0,016	2,459	2,249	0,210
15	0	2	9	0,871	0,937	0,066	2,867	2,345	0,522
16	0	3	8	1,196	1,213	0,017	2,598	2,874	-0,276
17	0	3	8	0,659	0,829	0,171	3,049	2,416	0,633
18	0	3	8	0,647	0,815	0,168	3,006	2,491	0,515
19	0	2	5	0,874	1,280	0,406	2,789	2,465	0,324
20	0	3	6	**1,587**	1,639	0,052	2,415	2,157	0,258
21	0	3	5	1,299	1,312	0,013	3,242	3,423	-0,181
22	0	3	6	0,780	0,803	0,023	3,190	3,285	-0,095
23	0	3	6	0,941	0,914	-0,028	3,327	3,398	-0,071
24	0	2	5	1,338	1,157	-0,180	3,533	3,294	0,239
25	0	3	5	0,855	1,136	0,281	3,285	3,107	0,178
26	0	3	7	0,812	0,700	-0,112	2,309	2,510	-0,201
27	0	4	7	0,819	**0,626**	-0,193	2,567	2,550	0,017
28	0	2	6	0,842	1,101	0,259	2,278	2,365	-0,087
29	0	3	9	0,960	1,224	0,264	2,372	2,788	-0,416
30	0	3	9	0,845	1,127	0,282	2,252	2,910	-0,658
31	0	2	8	1,101	1,092	-0,009	2,695	2,491	0,204
32	0	3	10	0,867	1,218	0,351	2,946	2,678	0,268
33	0	4	10	0,917	1,136	0,219	2,649	2,237	0,412
34	0	2	5	1,332	1,450	0,118	2,486	2,231	0,255
35	0	3	5	1,489	1,481	-0,007	2,776	2,321	0,455
36	0	2	6	1,425	1,553	0,128	2,695	2,627	0,068
37	0	2	6	1,479	1,574	0,095	2,965	2,472	0,493
38	0	2	5	1,173	1,803	0,630	3,056	2,627	0,429
39	0	2	6	0,992	0,920	-0,071	2,477	2,840	-0,363
40	0	2	8	0,898	0,990	0,093	3,478	2,798	0,680
41	0	3	8	0,935	1,183	0,248	3,117	2,501	0,616
42	0	5	11	1,299	1,327	0,028	**1,724**	1,899	-0,175
$\bar{x}$	0	2,74	7,21	1,050	1,204	0,154	2,777	2,568	0,209

Tabelle 3 - Akustische Eigenschaften der Stimuli: Anzahl der Konkatenationen im Zielwort, Mel Cepstral Distortion zwischen Synthese und Referenz, Perceptual Evaluation of Speech Quality. Ref = Referenz (natürlich), WTC09 = leicht degradiert, WTC50 = stark degradiert. Fett: Max/Min-Werte.

Äußerung berechnete durchschnittliche Euklidische Distanz zwischen leicht degradierten Stimuli und ihren natürlichen Pendants sollte daher geringer sein als die zwischen stark generierten und natürlichen Stimuli. In der Tat sind die MCD-Werte der WTC09-Stimuli im Schnitt kleiner als die der WTC50-Stimuli (Tabelle 3).

PESQ ist ein Standardverfahren zur automatischen Bewertung der Sprachqualität, insbesondere im Kontext der Telefonie [4]. Die Methode zielt auf ein objektives Qualitätsmaß ab, das die Ergebnisse subjektiver Qualitätseinschätzungen, z.B. MOS-Bewertungen, widerspiegelt. Höhere PSEQ-Werte entsprechen einer besseren, niedrige Werte einer geringeren Sprachqualität. Erwartungsgemäß sind die jeweils über den gesamten Stimulus berechneten PSEQ-Werte der WTC09-Stimuli im Schnitt größer als die der WTC50-Stimuli (Tabelle 3). Die Interpretation der MCD- und PSEQ-Werte wird im Abschnitt 4.2 noch einmal aufgegriffen.

3.2 Datenerhebung und Analyse

Das Sprachmaterial wurde auf sechs Listen aufgeteilt und unterschiedlichen Gruppen von Studierenden der Universität des Saarlandes in ruhigen Seminarräumen dargeboten. Die Teilnehmer wurden gebeten, alle Stimuli intuitiv hinsichtlich ihrer akustischen Qualität zu beurteilen. Es wurde darauf hingewiesen, dass es keine 'falschen' Antworten gibt. Die Fragestellung lautete: "Wie gut klingt für Sie der Satz als Ganzes?" Die Bewertung erfolgte auf einer Skala von 1 ("sehr schlecht") bis 6 ("sehr gut").

Während einer Eingewöhnungsphase wurden sechs Sätze als Beispiele dargeboten, um die Bandbreite unterschiedlicher Qualitäten zu illustrieren. Diese Sätze traten nicht im eigentlichen Test auf, und ihre Bewertung wurde nicht in die Analyse einbezogen. Im Experiment selbst bewertete jede Versuchsperson 120 Stimuli, von denen zwei Drittel Ablenkungsstimuli waren. An dem Experiment nahmen 91 deutsche Muttersprachler teil. Jede der sechs Listen wurde von mindestens zehn Probanden bewertet.

Zur Auswertung der Hörurteile wurden gemischte Modelle (*linear mixed effects models*) mit festgelegten Achsenabschnitten mit Hilfe des Pakets *lme4* [2] in R [5] erstellt. Proband und Stimulus wurden als Zufallsvariablen behandelt, Stimulusqualität als unabhängiger Faktor (3 Ausprägungen: Baseline, leicht degradiert, stark degradiert), und die Bewertung der Stimulusqualität (1 = "sehr schlecht" bis 6 = "sehr gut") als abhängige Variable.

4 Perzeptionsexperiment

4.1 Ergebnisse

Es werden nur die Ergebnisse von dem statistischen Modell berichtet, die die Daten am besten beschreiben (AIC – Akaike Information Criterion). Die statistische Analyse des Perzeptionsexperiments zeigte signifikante Unterschiede zwischen den drei verschiedenen Stimulusqualitäten (natürliche Sprache, leicht degradierte und stark degradierte Synthese). Vergleiche der Bedingungen zeigen, dass jede Stimulusqualität signifikant unterschiedlich von den beiden anderen Qualitäten bewertet wurde: natürlich vs. leicht degradiert, *coefficient* -0,4968 *SE* 0,1068 *t*=-4,65[1]; natürlich vs. stark degradiert, *coefficient* -1,3802 *SE* 0,1856 *t*=-7,44; leicht vs. stark degradiert, *coefficient* -0,8834 *SE* 0,1585 *t*=-5,573. Wie erwartet erhielten die leicht degradierten Stimuli eine durchschnittliche Bewertung, die unterhalb der natürlichsprachlichen Referenzqualität, aber oberhalb der stark degradierten Stimuli liegt (Abbildung 1).

[1] T-Werte >2 oder <-2 zeigen eine Signifikanz auf dem Niveau von 5% an [1].

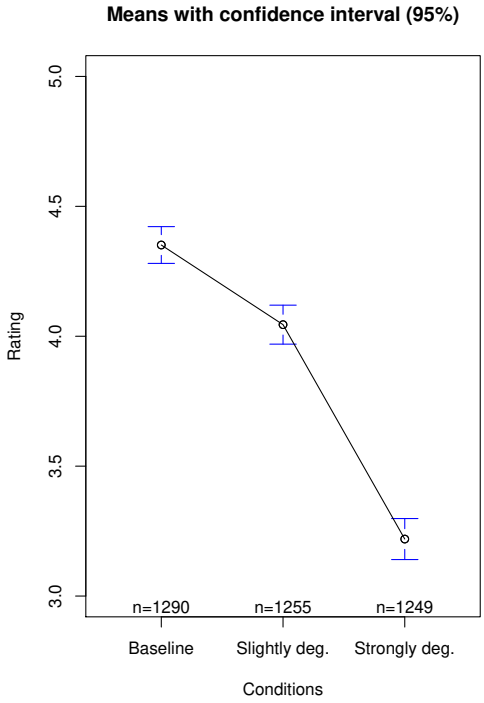

Abbildung 1 - Ergebnisse des Perzeptionsexperiments. Leicht degradierte Stimuli wurden im Durchschnitt signifikant besser bewertet (MOS 4,04, SD 1,35) als stark degradierte Stimuli (MOS 3,22, SD 1,42), jedoch signifikant schlechter als die natürlichsprachlichen Referenzstimuli (MOS 4,35, SD 1,3).

4.2 Vergleich der subjektiven und objektiven Bewertungen

Die subjektiven Bewertungen der natürlichen und synthetisierten Stimuli stützen die Annahme, dass die intendierte Manipulation der akustischen Qualität der Synthese gelungen ist. Die Qualität der leicht degradierten Stimuli wurden von den Teilnehmern des Perzeptionsexperiment hinreichend deutlich und statistisch signifikant geringer beurteilt als die natürlichsprachlichen Originalsignale. Ebenfalls statistisch signifikant, quantitativ jedoch mit erheblichem Abstand, wurde die Qualität der stark degradierten Stimuli wiederum als schlechter bewertet als die der leicht degradierten Stimuli. Dies entspricht der Intention, synthetische Stimuli zu erzeugen, die von wahrnehmbar schlechter Qualität, aber immer noch verständlich sind.

Die objektiven Bewertungen der Stimulusqualität mit den Maßen *Mel Cepstral Distortion* (MCD) und *Perceptual Evaluation of Speech Quality* (PESQ) sind weitgehend mit den subjektiven Beurteilungen kompatibel. Die akustische Distanz der leicht degradierten Stimuli (WTC=0,9) zu den natürlichsprachlichen Referenzsignalen ist objektiv geringer als die der stark degradierten Stimuli (WTC=5,0) (Δ MCD 1,050 vs. 1,204). Auch die PSEQ-Werte der synthetischen Stimuli entsprechen im Durchschnitt den Erwartungen (2,777 vs. 2,568) und decken sich von der Tendenz her mit den subjektiven MOS-Bewertungen.

Allerdings weicht die objektive Bewertung einiger Stimuli sowohl für MCD als auch für PSEQ von der erwarteten Tendenz ab (negative MCD- oder PSEQ-Werte in Tabelle 3). In Einzelfällen war demnach die intendierte Stärke der Manipulation der Synthesequalität nicht erfolgreich. Dies hängt mit der verwendeten Modifikation des Einheitenauswahlalgorithmus in MARYTTS zusammen: Die verringerte Synthesequalität wurde durch eine veränderte Gewichtung von Zielkosten und Verkettungskosten erzielt. Die Erhöhung des WTC-Parameters bewirkte eine stärker gewichtete Anforderung an die Erfüllung der Zielvorgaben auf Kosten einer

schlechteren Passgenauigkeit über die Verkettungsstellen hinweg. Diese algorithmisch Methode erlaubt nur eine indirekte Kontrolle der Anzahl der Verkettungen und der Stärke der Diskontinuität an den Verkettungsstellen. Der Vorteil gegenüber einer direkten manuellen Manipulation der Einheitensequenz liegt darin, dass die parameterbasierte Modifikation in Echtzeit und in realistischen Anwendungen der Sprachsynthese vorgenommen werden kann.

5 Fazit

In dieser Studie wurde die Wahrnehmung natürlicher und gezielt manipulierter und in ihrer Qualität degradierter synthetischer Sprache untersucht. Wir gingen der Frage nach, ob die Hörer bestimmte akustische Manipulationen von Stimuli bewusst wahrnehmen können. Die Stimuli bestanden aus natürlichen Sprachaufnahmen sowie aus Stimuli, die das Sprachsynthesesystem auf der Basis derselben Stimme generierte. Die synthetischen Stimuli wurden hinsichtlich der Häufigkeit und Güte von Konkatenationen manipuliert. Es ergab sich ein signifikanter Unterschied zwischen den Stimulusqualitäten. Leicht degradierte Stimuli erhielten, wie intendiert, schlechtere Bewertungen als die Baseline-Stimuli und bessere Bewertungen als die stark degradierten Stimuli. Objektive Distanzmaße wurden ebenfalls berechnet und erwiesen sich als kongruent mit den subjektiven Bewertungen. Die Methode ist somit geeignet, synthetisierte Stimuli unterschiedlicher Qualität auf systematische Weise zu erzeugen. Die Stimuli werden aktuell in einer Studie verwendet, in der die unbewusste Reaktion des Gehirns auf unterschiedliche Synthesequalitäten und, mittels ereigniskorrelierter Hirnpotenziale, der Verlauf der Verarbeitungsprozesse in Echtzeit untersucht wird.

Literatur

[1] BAAYEN, R. H.: *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, 2008.

[2] BATES, D. und D. SARKAR: *lme4: Linear mixed-effects models using S4 classes*, 2007.

[3] BENSON, R. R., D. H. WHALEN, M. RICHARDSON, B. SWAINSON, V. P. CLARK, S. LAI und A. M. LIBERMAN: *Parametrically dissociating speech and nonspeech perception in the brain using fMRI*. Brain and Language, 78(3):364–396, 2001.

[4] INTERNATIONAL TELECOMMUNICATION UNION (ITU): *P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. Available at http://www.itu.int/rec/T-REC-P.862/en (accessed Nov 4, 2014).

[5] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. Available at http://www.R-project.org (accessed Feb. 14, 2014).

[6] SCHRÖDER, M. und J. TROUVAIN: *The German text-to-speech synthesis system MARY: A tool for research, development and teaching*. International Journal of Speech Technology, 6:365–377, 2003. MARY available at http://mary.dfki.de/ (accessed Feb 14, 2014).

[7] TODA, T., A. W. BLACK und K. TOKUDA: *Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis*. In: *Proceedings of the 5th ISCA Speech Synthesis Workshop*, S. 31–36, 2004.