EXEMPLARMODELLE DER SPRACHPRODUKTION UND IHRE RELEVANZ FÜR DIE SPRACHSYNTHESE

Bernd Möbius

FR 4.7, Phonetik, Universität des Saarlandes
moebius@coli.uni-saarland.de

Kurzfassung: Sprachsynthese nach der Methode der *Unit Selection* wählt aus einem Korpus akustische Einheiten aus, die nach Verkettung eine gegebene Zieläußerung optimal repräsentieren. Bei der Auswahl werden die akustischen Einheiten danach bewertet, ob sie gute Repräsentanten der jeweiligen Zieleinheit sind und sich zugleich gut mit den benachbarten Einheiten verketten lassen. Wissenschaftshistorisch gesehen ist die *Unit Selection*–Methode nicht durch Modelle der menschlichen Sprachproduktion inspiriert. Exemplarbasierte Modelle der Sprachproduktion, die in den letzten Jahren zunehmend Beachtung gefunden haben, legen jedoch eine Analogie zwischen Modellen der Sprachproduktion und Verfahren der Sprachsynthese nahe. Dieser Beitrag widmet sich der Frage, ob diese Analogie oberflächlicher Art ist oder ob die komputationelle Modellierung von Prozessen der menschlichen Sprachverarbeitung die Implementierung sprachtechnologischer Verfahren informieren kann.

1 Einleitung

Korpusbasierte Sprachsyntheseverfahren orientieren sich nicht primär an den Prozessen der menschlichen Sprachproduktion in dem Sinn, dass sie eine Modellierung der Sprachproduktion anstreben. Sie arbeiten im Gegenteil mit dem Endprodukt dieser Prozesse, dem natürlichen Sprachsignal, dessen Bausteine – üblicherweise beschrieben durch traditionelle strukturelle Einheiten (Laute, Diphone, Silben usw.) – zu neuen, synthetischen Zieläußerungen rekombiniert und resequenziert werden.

Die Exemplartheorie wurde ursprünglich in der Psychologie als ein allgemeines Modell für Wahrnehmung und Kategorisierung eingeführt [11, 18] und ist erst in jüngerer Zeit zur Erforschung von Sprachperzeption [8, 9] und -produktion [19, 28, 30] aufgegriffen worden. Die zentrale Annahme ist hier, dass sprachliche Stimuli als vollspezifizierte Exemplare im Langzeitgedächtnis abgelegt werden, die als Referenzen für die Kategorisierung neuer Stimuli bei der Sprachwahrnehmung und als Ziele bei der Sprachproduktion dienen können. Die Repräsentation der Exemplare enthält phonetische Details, Informationen über den sprachlichen Kontext und den Sprecher und über die kommunikative Situation. Die Exemplarrepräsentation sprachlicher Kategorien wird ständig aktualisiert und ist sensitiv gegenüber Effekten der Auftretenshäufigkeit, die für diverse linguistische Domänen und phonetische Parameter dokumentiert worden sind [3, 4, 5, 10, 16]. Die Varianz der Realisierungen sprachlicher Kategorien drückt sich in der statistischen Verteilung ihrer Exemplare entlang zahlreicher Dimensionen aus.

Eine Herausforderung für Exemplarmodelle ist es zu erklären, wie Exemplare sprachlicher Kategorien auf unterschiedlichen strukturellen Beschreibungsebenen (z.B. Laute, Silben, Wörter) miteinander interagieren und wie sich Exemplare kleinerer Einheiten zu solchen größerer Ein-

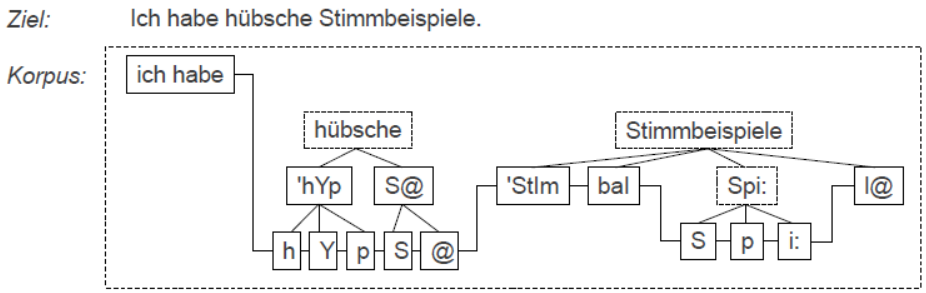

Abbildung 1 - Non-Uniform Unit Selection: Zur Abdeckung der Zieläußerung werden Einheiten im Korpus entlang einer linguistischen Hierarchie gesucht. Wenn auf einer höheren Ebene (z.B. Wort) keine Kandidaten gefunden werden, wird die Suche auf einer niedrigeren Ebene (z.B. Silbe) fortgesetzt.

heiten verbinden. Die Interaktion solcher Einheiten auf unterschiedlichen Ebenen ist eine wichtige Charakteristik der Sprache. Sie ist auch Kern der Strategie der *Non-Uniform Unit Selection* in der Sprachsynthese [20, 22, 24, 26].

2 Sprachproduktion und Sprachsynthese

In den folgenden Abschnitten wird zunächst die als *Unit Selection* bekannte Synthesemethode mit ihren für die Zwecke dieses Artikels relevanten Eigenschaften charakterisiert. Anschließend werden zwei komputationelle Modelle der Sprachproduktion vorgestellt, die im theoretischen und konzeptionellen Rahmen der Exemplartheorie entwickelt wurden und für die Weiterentwicklung der Sprachsynthese potenziell interessante Konzepte und Ergebnisse bieten.

2.1 *Unit Selection*–Synthese

Sprachsynthese nach der Methode der *Unit Selection* [6, 25] wählt aus einem Korpus akustische Einheiten aus, die nach erfolgter Verkettung eine gegebene Zieläußerung optimal repräsentieren. Bei der Auswahl werden die akustischen Einheiten danach bewertet, ob sie gute Repräsentanten der jeweiligen Zieleinheit sind und sich zugleich gut mit den benachbarten Einheiten verketten lassen. Bei der Berechnung der Zielkosten werden akustische, phonetische und linguistische Eigenschaften des Kandidaten selbst sowie die des Kontextes einbezogen. Die Verkettungskosten basieren auf der akustischen Diskontinuität, die bei der Verkettung von je zwei Einheiten entsteht.

Viele *Unit Selection*–Systeme verwenden (sub)segmentale Einheiten (Demiphone, Phone, Diphone) als Basiseinheit. Längere Sequenzen, z.B. Silben, Wörter oder ganze Phrasen, werden bei der Auswahl insofern indirekt begünstigt, als im Korpus bereits adjazente Laute keine Verkettungskosten verursachen. Einige Ansätze zielen auf eine unmittelbarere Auswahl größerer Bausteine ab, indem sie das Korpus entlang einer hierarchischen linguistischen Repräsentation von oben nach unten durchsuchen. Wenn auf einer höheren Ebene (z.B. Wort) keine Kandidaten gefunden werden, wird die Suche auf einer niedrigeren Ebene (z.B. Silbe) fortgesetzt. Die Zielspezifikationen werden ebenenspezifisch formuliert, und die Verkettungskosten zwischen zwei längeren Einheiten entsprechen den Kosten der Verkettung der beiden Lautsegmente an der Verkettungsstelle. Solche Verfahren werden häufig als *Non-Uniform Unit Selection* bezeichnet [20, 22, 24, 26] (Abb. 1). Sie sind am besten für eingeschränkte Anwendungsdomänen geeignet, in denen viele längere Einheiten bereits im Korpus enthalten sind, und sind weniger effizient

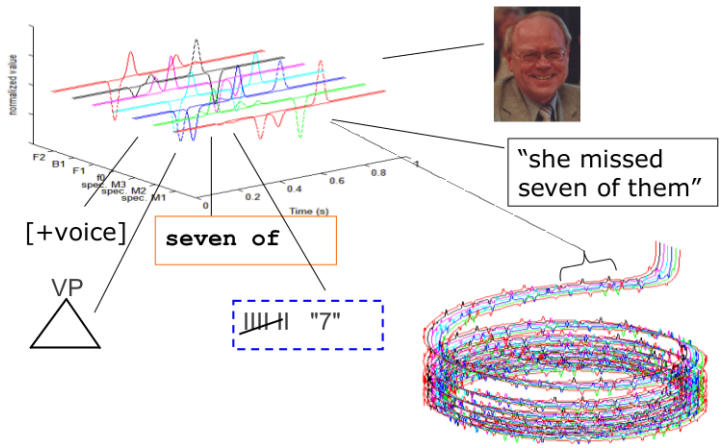

Abbildung 2 - Context Sequence Model: Exemplare sind entlang vieler akustischer, linguistischer und außersprachlicher Dimensionen spezifiziert und bilden ein episodisches Gedächtnis in Form einer langen Sequenz analysierter sprachlicher Ereignisse [28].

in unbeschränkten Domänen, in denen häufig auf die kleinsten, (sub)segmentalen Einheiten zurückgegriffen werden muss.

Die Algorithmen der Einheitenauswahl in *Unit Selection*–Systemen sind nicht an Modellen der Sprachproduktion und nur bedingt, nämlich durch auditorische Skalierung akustischer Parameter, an Erfordernissen der Sprachwahrnehmung orientiert.

2.2 Exemplarbasierte Sprachproduktion

In einem Exemplarmodell der Sprachproduktion dienen Akkumulationen von Exemplaren, die die Zielkategorie repräsentieren, als Produktionsziele [19, 23]. Zur Konstruktion der Produktionsziele werden diejenigen Exemplare herangezogen, die der Zieleinheit im gegebenen Kontext optimal entsprechen. Bei Kategorien mit hoher Auftretenshäufigkeit ist es wahrscheinlich, dass zahlreiche passende Exemplare zur Verfügung stehen. Das daraus gebildete Produktionsziel und das auf dieser Grundlage neu produzierte Exemplar ist somit früheren Realisierungen ähnlich. Die empirisch beobachtete höhere Variabilität hochfrequenter Kategorien im Vergleich zu seltenen Kategorien [23] ist darauf zurückzuführen, dass Exemplare von Kategorien mit hoher Auftretenshäufigkeit in sehr vielen unterschiedlichen Kontexten vorkommen, was die Entstehung von Kategorien mit großer, aber systematischer Variabilität begünstigt.

Im folgenden sollen zwei implementierte Exemplarmodelle vorgestellt werden, die eine komputationelle Simulation von Aspekten der Sprachproduktion erlauben.

Im **Context Sequence Model** (CSM) [28, 29] werden akustische Ziele der Sprachproduktion durch die Auswahl von Einheiten aus einem Gedächtnisspeicher bestimmt. Der Speicher enthält eine große Anzahl früher wahrgenommener (oder produzierter) sprachlicher Einheiten, die sowohl abstrakt (phonologisch) indiziert als auch entlang vieler akustischer Dimensionen spezifiziert sind (Abb. 2). Die Signale im Gedächtnisspeicher entsprechen langen Sequenzen kontinuierlicher Sprache, so dass individuelle Sprachlaute immer in einem größeren Kontext auftreten. Eine zentrale Eigenschaft des Modells besteht darin, dass die Auswahl von Exemplaren für die Produktion auf einer Bewertung der Ähnlichkeit zwischen dem Kontext, in dem sie ursprünglich aufgetreten waren, und dem aktuellen Produktionskontext beruht.

Simulationen der Sprachproduktion mit realistischen akustischen Daten zeigen, dass die opti-

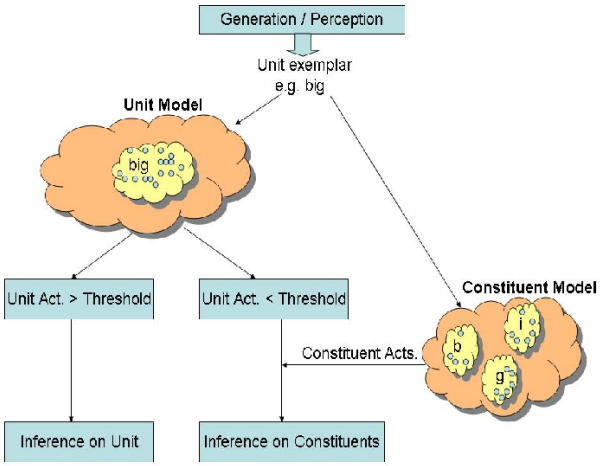

Abbildung 3 - Multilevel Exemplar Model: Interaktion zwischen Einheiten auf unterschiedlichen linguistischen Ebenen; hier: Verarbeitung von Silben als Einheit bzw. als Sequenz der sie konstituierenden Einzellaute, in Abhängigkeit von der Aktivierungsstärke oder Auftretenshäufigkeit [30].

male Auswahl von kontextadäquaten Einheiten auf Lautebene die Berücksichtigung eines linken und rechten Kontextes von etwa 1 Sekunde, zentriert um den Ziellaut, erfordert. Die Ergebnisse legen außerdem die Interpretation nahe, dass die kontextabhängige Produktion auf der Lautebene einige Effekte der Auftretenshäufigkeit bedingt, die bislang als Effekte auf Silben-, Wort- und anderen höheren Ebenen der sprachlichen Organisation betrachtet wurden [28]. Das CSM sieht keine separaten Ebenen für Exemplare von Lauten, Silben, Wörtern oder Phrasen vor. Das episodische Exemplargedächtnis ist einfach eine lange Sequenz analysierter sprachlicher Ereignisse.

Das CSM zeigt, wie die Sprachproduktion unmittelbar durch phonetisches Wissen beeinflusst wird. Das Modell ist jedoch unvollständig, da die Simulationen nur die Akustik des lautlichen Kontextes berücksichtigen, ohne die zugrunde liegenden Artikulationsprozesse einzubeziehen. Damit ist keineswegs impliziert, dass phonetisches Wissen vorrangig akustisch oder auditorisch ist. Im Gegenteil, das Modell sollte dahingehend erweitert werden, dass die Simulationen auch artikulatorische Prozesse und Beschränkungen bei der Sprachproduktion integrieren [7].

Das **Multilevel Exemplar Model** (MLM) [30, 31] wurde entwickelt, um die Interaktion zwischen Einheiten auf unterschiedlichen hierarchischen Ebenen linguistischer Domänen zu untersuchen. Es erlaubt eine Simulation experimenteller Daten, die bislang für zwei Phänomene durchgeführt wurde: die Variabilität von Silbendauern in der gesprochenen Sprache sowie eine exemplarbasierte Fundierung von Grammatikalitätsurteilen in der Syntax [21]. In der phonetischen Domäne wurde das MLM zur Modellierung von Effekten der Auftretenshäufigkeit auf Laut- und Silbenenebene verwendet. Produktionsstudien zeigen, dass häufige Silben variabler sind als seltene Silben [23]. Dieser Effekt der Auftretenshäufigkeit wird von exemplarbasierten Simulationen im Rahmen des MLM bestätigt [30].

Experimentelle Ergebnisse anhand großer lautsprachlicher Korpora legen die Interpretation nahe, dass die Variabilität von Silbendauern eine Funktion der Variabilität der involvierten Lautdauern ist – allerdings nur in Silben mit geringer Auftretenshäufigkeit. Hingegen konnte die Variabilität der Dauern häufiger Silben nicht direkt aus den Dauern der Einzellaute vorhergesagt werden [23]. Die Ergebnisse der MLM-basierten Simulationen untermauern die Hypothese, dass häufige Silben in der Sprachproduktion als eigenständige Einheiten verarbeitet werden,

während seltene Silben durch sequenzielle Verarbeitung der sie konstituierenden Laute produziert werden [30, 31] (Abb. 3). Diese Ergebnisse sind mit den Konzepten des *Mental Syllabary* [15], einer Komponente des psycholinguistischen Sprachproduktionsmodells von Levelt und Kollegen [12, 13, 14], und des *Dual Pathway* mit Wettbewerb zwischen Einheitensequenzen auf mehreren Ebenen [32] kompatibel.

Das MLM kann außerdem mit einer Version des CSM in Einklang gebracht werden, in der Intervalle der Exemplargedächtnis-Sequenz bezüglich sprachlicher Kategorien annotiert sind, wobei die Intervalle überlappen und ineinander eingebettet sein können.

3 Synergien und ihre Grenzen

Die Exemplartheorie ist teilweise durch die Beobachtung motiviert, dass nicht alle Elemente der Sprache originell und produktiv sind. Zahlreiche sprachliche Bausteine, die wir produzieren und wahrnehmen, sind vorfabrizierte und wiederverwertbare Versatzstücke [2]. Die Herausforderung liegt in der kontextuell adäquaten Auswahl und Resequenzierung der verfügbaren Bausteine. Ein wichtiger Unterschied zwischen einer Konzeptualisierung der Sprachproduktion etwa im Sinn des *Context Sequence Model* und der korpusbasierten konkatenative Synthese liegt darin, dass die Exemplarbasis kontinuierlich durch Verarbeitung neuer Exemplare aktualisiert wird, während das Sprachkorpus des *Unit Selection*–Systems statisch ist. Bei der Sprachsynthese kommt es also darauf an, schon beim Design des Sprachkorpus eine optimale Abdeckung der Zieldomäne – in der Regel der gesamten Zielsprache – anzustreben. Aber auch die Einheitenauswahl zur Laufzeit kann anhand erfolgreicher Simulationen durch Exemplarmodelle modifiziert werden.

Sprachkorpora von *Unit Selection*–Systemen, einmal aufgezeichnet und annotiert, sind statisch. Im Unterschied dazu wird die Exemplarbasis in der menschlichen Sprachverarbeitung kontinuierlich durch Wahrnehmung und Produktion neuer Exemplare aktualisiert. Exemplarmodelle arbeiten außerdem häufig mit einem Konzept von dynamischer Aktivierung: Die Stärke der Aktivierung gespeicherter Exemplare, die ansonsten mit fortschreitender Zeit sinkt, wird durch Verwendung hinreichend ähnlicher Exemplare wieder angehoben. Dies führt dazu, das Exemplare häufiger Kategorien, die in häufigen Kontexten auftreten, als Referenz für wahrgenommene Exemplare und als Ziele für aktuelle Produktionen in großer Zahl zur Verfügung stehen. Seltene Exemplare hingegen sinken allmählich auf ein Niveau, das sie – ohne weitere Modellannahmen – nicht wieder aktivierbar macht. Auf diese Weise werden in der Exemplartheorie Phänomene wie individueller Lautwandel oder gar Kategorienverfall modelliert [19]. Exemplarbasierte Sprachproduktion beruht also auf der Ähnlichkeit, Häufigkeit und Neuigkeit von Exemplaren.

Grundsätzlich ließe sich in der Sprachsynthese ein Mechanismus der dynamischen Aktivierung von Korpusbausteinen auf der Basis der Häufigkeit ihrer Verwendung durchaus implementieren. Zwei Probleme müssten dabei gelöst werden. Zum einen kann die Verwendung immer wieder derselben Bausteine in häufig auftretenden Konstellationen zu einem Grad an Invarianz führen, der in der menschlichen Sprachproduktion aufgrund inhärenter Ungenauigkeiten des Artikulationsprozesses nicht entstehen kann. In exemplarbasierten Produktionsmodellen werden zumeist Zufallsvariationen der Realisierung ausgewählter Einheiten als triviale Implementierung der artikulatorischen Variabilität eingesetzt. In der Synthese würde ein solcher Mechanismus zusätzliche Signalverarbeitungsschritte erfordern, um häufig ausgewählte Bausteine künstlich zu variieren. Damit würde jedoch eine der grundlegenden Motivationen für *Unit Selection*–Synthese, die weitgehende Vermeidung von Signalverarbeitung, konterkariert. Zum anderen müsste sicher gestellt werden, dass selten verwendete Bausteine dennoch weiterhin für seltene Konstellationen

zur Verfügung stehen. Bekanntlich ist die kumulative Wahrscheinlichkeitsmasse sprachlicher Ereignisse, die jedes für sich extrem unwahrscheinlich sind, so hoch, dass seltene Ereignisse nicht generell vernachlässigt werden dürfen [17, 27]. Dies wäre durch die Vermeidung einer Mindestaktivierungsschwelle vermutlich einfach zu erreichen.

Exemplarmodelle nehmen an, dass verarbeitete und gespeicherte Exemplare entlang vieler Dimensionen spezifiziert sind (Abb. 2). Zu diesen Dimensionen gehören akustische Eigenschaften, phonologische Kategorien, linguistische Strukturen (morphologische Struktur, Wortklasse, syntaktische und semantische Information), indexikalische Sprecherinformation bis hin zu Informationen über die kommunikative Situation. Bei weitem nicht alle diese Informationen lassen sich im Sprachsynthesekorpus annotieren, sei es aus praktischen oder theoretischen Gründen. Schwerer wiegt der Umstand, dass – zumindest im klassischen *Text-to-Speech*–Szenario – aus dem Text allein keine Zielspezifikation erstellt werden kann, die indexikalische und außersprachliche Information enthält. Dies gilt zum Teil auch für semantische und pragmatische Informationen.

Dass selbst eine partielle Berücksichtigung einer um kontextuelle Information erweiterten Zielspezifikation die Synthese verbessern kann, zeigt eine Implementierung des *Multilevel Exemplar Model* zum Zweck der exemplarbasierten Lautdauermodellierung in expressiver Sprachsynthese. Die so generierten Stimuli wurden in Perzeptionstests besser beurteilt als die herkömmlich synthetisierten Stimuli: Die exemplarbasierte Synthese klang eher wie ein echter Sprecher, gab die Sprecherintentionen besser wieder, klang flüssiger und erweckte eher den Eindruck einer konversationellen Sprache [1].

4 Ausblick

Dieser Artikel befasste sich mit Analogien und parallelen Konzepten zwischen Exemplarmodellen der Sprachproduktion und korpusbasierter Synthese und diskutierte mögliche Synergien, aber auch die Grenzen der Synergien. Diese Grenzen liegen zum ersten in der Statik von Sprachsynthesekorpora, zum zweiten in der Dimensionalität der Eigenschaften sprachlicher Einheiten und zum dritten in der Repräsentation des außersprachlichen Kontextes. Simulationen und erste Experimente stützen jedoch die Erwartung, dass exemplarbasierte Modelle von Prozessen der menschlichen Sprachverarbeitung interessante Konzepte für die Weiterentwicklung der Sprachsynthese und generell für die Implementierung sprachtechnologischer Verfahren bieten können.

Literatur

[1] ABOU-ZLEIKHA, M., É. SZÉKELY, P. CAHILL und J. CARSON-BERNDSEN: *Multi-level exemplar-based duration generation for expressive speech synthesis*. In: *Proceedings of Speech Prosody 2012 (Shanghai)*, 2012.

[2] BYBEE, J.: *From usage to grammar: The mind's response to repetition*. Language, 84:529–551, 2006.

[3] BYBEE, J. und J. SCHEIBMAN: *The effect of usage on degrees of constituency: the reduction of don't in English*. Linguistics, 37(4):575–596, 1999.

[4] CARREIRAS, M. A. und M. B. PEREA: *Naming pseudowords in Spanish: Effects of syllable frequency*. Brain and Language, 90:393–400, 2004.

[5] CHOLIN, J., W. J. M. LEVELT und N. O. SCHILLER: *Effects of syllable frequency in speech production*. Cognition, 99(2):205–235, 2006.

[6] CLARK, R. A. J., K. RICHMOND und S. KING: *Multisyn: Open-domain unit selection for the Festival speech synthesis system*. Speech Communication, 49(4):317–330, 2007.

[7] DURAN, D., J. BRUNI, H. SCHÜTZE und G. DOGIL: *Context Sequence Model of speech production enriched with articulatory features*. In: *Proceedings of the 17th International Congress of Phonetic Sciences (Hong Kong)*, S. 615–618, 2011.

[8] GOLDINGER, S. D.: *Words and voices—Perception and production in an episodic lexicon*. In: JOHNSON, K. und J. W. MULLENNIX (Hrsg.): *Talker Variability in Speech Processing*, S. 33–66. Academic Press, San Diego, 1997.

[9] JOHNSON, K.: *Speech perception without speaker normalization: An exemplar model*. In: JOHNSON, K. und J. W. MULLENNIX (Hrsg.): *Talker Variability in Speech Processing*, S. 145–165. Academic Press, San Diego, 1997.

[10] JURAFSKY, D., A. BELL, M. GREGORY und W. D. RAYMOND: *Probabilistic relations between words: Evidence from reduction in lexical production*. In: BYBEE, J. und P. HOPPER (Hrsg.): *Frequency and the Emergence of Linguistic Structure*, S. 229–254. Benjamins, Amsterdam, 2001.

[11] KRUSCHKE, J. K.: *ALCOVE: An exemplar-based connectionist model of category learning*. Psychological Review, 99(1):22–44, 1992.

[12] LEVELT, W. J. M.: *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA, 1989.

[13] LEVELT, W. J. M.: *Producing spoken language: a blueprint of the speaker*. In: BROWN, C. M. und P. HAGOORT (Hrsg.): *The Neurocognition of Language*, S. 83–122. Oxford University Press, Oxford, UK, 1999.

[14] LEVELT, W. J. M., A. ROELOFS und A. S. MEYER: *A theory of lexical access in speech production*. Behavioral and Brain Sciences, 22:1–75, 1999.

[15] LEVELT, W. J. M. und L. WHEELDON: *Do speakers have access to a mental syllabary?*. Cognition, 50:239–269, 1994.

[16] LOSIEWICZ, B. L.: *The Effect of Frequency on Linguistic Morphology*. Doktorarbeit, University of Texas, Austin, TX, 1992.

[17] MÖBIUS, B.: *Rare events and closed domains: Two delicate concepts in speech synthesis*. International Journal of Speech Technology, 6(1):57–71, 2003.

[18] NOSOFSKY, R. M.: *Attention, similarity, and the identification-categorization relationship*. Journal of Experimental Psychology: General, 115(1):39–57, 1986.

[19] PIERREHUMBERT, J.: *Exemplar dynamics: Word frequency, lenition and contrast*. In: BYBEE, J. und P. HOPPER (Hrsg.): *Frequency and the Emergence of Linguistic Structure*, S. 137–157. Benjamins, Amsterdam, 2001.

[20] SAGISAKA, Y.: *Speech synthesis by rule using an optimal selection of non-uniform synthesis units*. In: *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (New York, NY)*, S. 679–682, 1988.

[21] SCHÜTZE, H., M. WALSH, B. MÖBIUS und T. WADE: *Towards a unified exemplar-theoretic model of phonetic and syntactic phenomena*. In: *Proceedings of the 29th Meeting of the Cognitive Science Society (CogSci 2007, Nashville, TN)*, S. 1461–1466, 2007.

[22] SCHWEITZER, A., N. BRAUNSCHWEILER, T. KLANKERT, B. MÖBIUS und B. SÄUBERLICH: *Restricted unlimited domain synthesis*. In: *Proceedings of Eurospeech-2003 (Geneva)*, S. 1321–1324, 2003.

[23] SCHWEITZER, A. und B. MÖBIUS: *Exemplar-based production of prosody: Evidence from segment and syllable durations*. In: *Speech Prosody 2004 (Nara, Japan)*, S. 459–462, 2004.

[24] STÖBER, K., P. WAGNER, J. HELBIG, S. KÖSTER, D. STALL, M. THOMAE, J. BLAUERT, W. HESS, R. HOFFMANN und H. MANGOLD: *Speech synthesis using multilevel selection and concatenation of units from large speech corpora*. In: WAHLSTER, W. (Hrsg.): *Verbmobil: Foundations of Speech-to-Speech Translation*, S. 519–534. Springer, Berlin, 2000.

[25] TAYLOR, P.: *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[26] TAYLOR, P. und A. W. BLACK: *Speech synthesis by phonological structure matching*. In: *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, Bd. 2, S. 623–626, 1999.

[27] VAN SANTEN, J. P. H.: *Combinatorial issues in text-to-speech synthesis*. In: *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, Bd. 5, S. 2511–2514, 1997.

[28] WADE, T., G. DOGIL, H. SCHÜTZE, M. WALSH und B. MÖBIUS: *Syllable frequency effects in a context-sensitive segment production model*. Journal of Phonetics, 38(2):905–945, 2010.

[29] WADE, T. und B. MÖBIUS: *Speaking rate effects in a landmark-based phonetic exemplar model*. In: *Proceedings of Interspeech 2007 (Antwerpen)*, S. 402–405, 2007.

[30] WALSH, M., B. MÖBIUS, T. WADE und H. SCHÜTZE: *Multilevel Exemplar Theory*. Cognitive Science, 34:537–582, 2010.

[31] WALSH, M., H. SCHÜTZE, B. MÖBIUS und T. WADE: *Accounting for phonetic and syntactic phenomena in a multi-level competitive interaction model*. In: *ESSLLI Workshop on Exemplar Based Models of Language Acquisition and Use (Dublin)*, S. 22–31, 2007.

[32] WHITESIDE, S. P. und R. A. VARLEY: *Dual-Route Phonetic Encoding: Some Acoustic Evidence*. In: *Proceedings of the 5th International Conference on Spoken Language Processing (Sydney)*, Bd. 7, S. 3155–3158, 1998.