'JA, MHM, ICH VERSTEHE DICH' – OSZILLATOR-BASIERTES TIMING MULTIMODALER FEEDBACK-SIGNALE IN SPONTANEN DIALOGEN

Petra Wagner[1], *Benjamin Inden*[2], *Zofia Malisz*[1] *und Ipke Wachsmuth*[2]

[1]*AG Phonetik und Phonologie, Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld*
[2]*AG Wissensbasierte Systeme, Technische Fakultät, Universität Bielefeld*
petra.wagner@uni-bielefeld.de

Kurzfassung: Die Produktion von Feedback-Signalen ist ein wichtiger Bestandteil spontansprachlicher zwischenmenschlicher Interaktion. Hörer signalisieren durch kurze Feedback-Äußerungen wie „ja" oder „mhm" bzw. durch begleitende (nickende) Kopfgesten, dass sie mehr oder weniger aufmerksam zuhören, den Sprechenden mehr oder weniger gut verstehen bzw. mit dem Gesagten übereinstimmen oder nicht [2]. Wir gehen in unserem Projekt davon aus, dass Sprecher ihre Feedback-Äußerungen an Sprechrhythmus und Sprechtempo des Dialogpartners anpassen. Dieser Angleichungsprozess wird häufig als *Entrainment* bezeichnet und wurde auf verschiedenen sprachlicher Interaktion beobachtet [7]. Für die Modellierung rhythmischen Entrainments bieten sich Oszillatoren an, wie sie bereits für Modelle der Wahrnehmung komplexer Rhythmen in Musik, aber auch für die Modellierung der zeitlichen Koordination von Sprecherwechseln in Dialogen angewandt wurden [9, 21]. Die Oszillatoren passen sich dabei in Phase und/oder Periode an ein Eingabesignal (hier: das Sprachsignal des Dialogpartners) an. Erste Ergebnisse bestätigen die Möglichkeit der Modellierung von Entrainment mit Hilfe von Oszillatoren für spontansprachliche Daten.

1 Einleitung

Zur besseren Modellierung der Interaktion von Dialogpartnern gehört auch die Produktion zeitlich passender Feedbacksignale. Kurze Feedback-Äußerungen oder sprachbegeleitende Kopfgesten wie Kopfnicken signalisieren, dass die Hörer mehr oder weniger "bei der Sache" sind, mit dem Sprecher übereinstimmen oder nicht, dass sie das Gesagte verstanden haben oder nicht. Es wird vermutet häufig, dass sich das Timing von Feedback an der rhythmisch-prosodischen Struktur der Sprecheräußerungen orientiert und nicht zufällig oder in regelmäßigen Abständen, unabhängig von den Äußerungen des Dialogpartners, produziert wird. Ein natürliches Timing verbessert die wahrgenommene Natürlichkeit von Äußerungen in Dialogsystemen [17]. Bisherige Ansätze zur Modellierung von Feedback-Timing sind oft regelbasiert [20] oder verwenden probabilistische Ansätzen [14].

In unserem Projekt verfolgen wir die Modellierung des Timings von Feedback-Äußerungen mittels Oszillatormodellen, wie sie bereits für das Entrainment musikalischer Rhythmen eingesetzt [9] und für sprachliche Phänomene vorgeschlagen wurden [13]. Ähnlich wie musikalische Rhythmen sind sprachliche Rhythmen hierarchisch organisiert und bilden komplexe rhythmische Muster. Gruppen von betonten und unbetonten Silben bilden *metrische Füße*, Gruppen von Füßen bilden prosodische oder Intonationsphrasen. Zu diesen hierarchischen Ebenen finden sich Analogien in der Lyrik (Silbe, Fuß, Vers) und der Musik (Schlag, Takt, Phrase). Muster ähnlicher Frequenzen spiegeln sich in der Neurokognition wider und spielen eine Rolle bei der

Sprachwahrnehmung [6, 5]. Es gibt Indizien, dass diese rhythmischen Frequenzen eine Schnittstelle der zeitlichen Koordination von Produktions- und Perzeptionsmechanismen verschiedener Modalitäten bilden, da sie sowohl in der Perzeption als auch in der Bewegungskoordination, z.B. bei der Planung von Sprache und sprachbegleitender Gesten, eine Rolle spielen [19]. Trotz der Beobachtung sprachlicher Regelmäßigkeiten qualitativer Natur gestaltet sich Spontanprache weitaus unregelmäßiger als Musik. Außerdem ist sie dynamisch veränderlich, z.B. durch Sprechtempoanpassungen. Nicht-dynamische, auf zeitliche Gleichförmigkeit ausgelegte Ansätze eignen sich daher nicht für die rhythmische Modellierung von Spontansprache. Ein weiterer Grund für den Einsatz von Oszillatormodellen liefern Indizien, dass Sprecher sich ihren Dialogpartnern zeitlich-dynamisch hinsichtlich Phase und Periode anpassen. Entrainmentprozesse dürften daher auch für die Modellierung von Sprecherwechsel (*turn taking*) und bei der Äußerungsplanung insgesamt ein wichtiger Faktor sein [21, 22].

Wir beschränken uns in unseren Modellierungen derzeit auf die Modelle von [10] und [13]. Wir überprüfen zunächst die grundsätzliche Eignung der Oszillatoren für die Modellierung von Entrainment auf spontansprachlichem Datenmaterial, d.h. wie gut können die Oszillatoren die Auftretenszeitpunkte zukünftiger rhythmischer Einheiten – Silben und metrische Füße – trotz ihrer Unregelmäßigkeit vorhersagen. Klar ist, dass eine Anpassung an das Eingangssignal aufgrund der stark veränderlichen Sprachsignale möglichst schnell erfolgen muss. Neben der grundsätzlichen Eignung überprüfen wir die folgenden Fragestellungen:

1. Passen sich Oszillatoren, die auf ein Eingangssignal mit Phasenreset reagieren, schneller an ein Eingabesignal an als Oszillatoren mit einer graduellen Phasenangleichung.

2. Ist eine Oszillatorbank besser geeignet als ein sich in der Phase anpassender einzelner Oszillator? Dies könnte der Fall sein, da die benötigte Zeit für die Anpassung der Periode abhängig vom Umfang der notwendigen Anpassung ist. In einer Oszillatorbank entfällt diese Anpassungsphase, da die Aktivierungszeit eines zeitlich "stimmigen" Oszillators konstant bleibt.

Die Oszillatoren werden zunächst separat auf zwei verschiedenen für *Entrainment*-Prozesse in Frage kommenden prosodischen Ebenen getestet: Silben und metrische Füße. Es wird untersucht, ob *Entrainment* bereits auf nur einer prosodischen Ebene, z.B. durch einen Silben- oder Fußoszillator möglich ist. Zudem wird die langfristige und kurzfristige Anpassung der Oszillatoren durch unterschiedliche Eingabedaten (Konversationen vs. Einzelphrasen) untersucht.

2 Verwendetes Datenmaterial

Als Datenmaterial für die Modellierung und Evaluation des Feedback-Alignments diente ein multimodales, spontansprachliches Datenkorpus [3], in welchem das Hörer-Feedback auf eine Erzählsituation (Urlaubserlebnisse) im Vordergrund der Untersuchung steht. Die Dialogpartner saßen sich gegenüber und konnten natürlich interagieren. Das Audiomaterial wurde auf verschiedenen rhythmisch-prosodischen Ebenen etikettiert: Die Ebene der Sprechsilbe bzw. das Intervall zwischen zwei *Perceptual Centers* oder *p-centers* welche mit dem wahrgenommenen Einsatz eines rhythmischen Schlags zusammenfallen [15]. Diese Silbensequenzen bilden die niedrigste Stufe für ein rhythmisches *Entrainment* bzw. die Eingabe für den angenommenen Silbenoszillator. Die Silbenanlaute wurden automatisch erfasst nach [4, 18] und handkorrigiert. Der zweite, langsamere Oszillator erhält handetikettierte metrische Füße bzw. Interbetonungsintervalle (IBIs) als Eingabesignal, welche durch betonte Silben voneinander abgegrenzt werden [1]. Zusätzlich wurden Interpausenintervalle (IPIs) etikettiert, die in der Regel *Intonationsphrasen* entsprechen. Wir untersuchen in unserem ersten Experiment das kurzfristige, schnelle

Entrainment, indem Einzelphrasen einer Konversation als Eingabematerial verwendet werden. Als Eingabe für die Oszillatoren wurden IPIs ausgewählt, die aus mindestens zwei metrischen Füßen oder IBIs bestehen. Phraseninitiale unbetonte Silben sowie phrasenfinale Silben wurden aus den Datensätzen entfernt, da die Dauern dieser Einheiten stark durch die Nähe der Phrasengrenze beeinflusst werden – diese Effekte werden aber durch die Oszillatoren nicht modelliert. Das Datenmaterial besteht aus flüssig gesprochenen, spontansprachlichen deutschen Äußerungen ohne Pausen. Die mittlere Silbendauer im Datenmaterial entspricht 125ms, die mittlere IBI-Dauer (= metrische Füße) entspricht 365ms. Pro Konversation wurden 69 Phrasen als Eingaben für verschiedene Oszillatoren verwendet. Die Zeitpunkte der jeweiligen Silbenanlaute bzw. der metrischen Füße dienten als Eingabepuls. Zusätzlich wurde für eine Kontrollmenge bestehend aus regelmäßigen Phrasen generiert, in denen die gleichförmigen Eingabepulse der mittleren Frequenz der Pulse in der korrespondierenden spontansprachlichen Phrase entsprechen. Neben der Untersuchung des Entrainment innerhalb einer Phrase könnten sich verschiedene Modelle hinsichtlich ihrer Anpassungsfähigkeit über längere Zeiträume, also gesamten Konversationen, unterscheiden. Diese Daten enthalten zudem mehr Irregularitäten und könnten somit auch die Robustheit verschiedener Modelle untersuchen. Daher verwenden wir in einem zweiten Experiment als zusätzliches Eingabematerial Daten einer gesamten Konversation (siehe auch [12]) und vergleichen die Ergebnisse beider Experimente.

3 Oszillator-basierte *Entrainment*-Modelle

3.1 Phasenanpassungsoszillator (PAO)

Dieses Oszillatormodell ist eines von mehreren Modellen, welche ursprünglich von [10] für das Entrainment musikalischer Rhythmen vorgeschlagen wurden. Die Phase dieses Oszillators ist definiert als $\phi(t) = \frac{t-t_x}{p}$, wobei t_x der Zeitpunkt des letzten Ereignisses (in der Eingabe oder hinsichtlich der Erwartung des Oszillators) darstellt, und p die Oszillatorperiode ist. Die Phase wird auf 0.0 zurückgesetzt (= Phasenreset), wenn sie den Wert 1.0 erreicht. Die Ausgabe des Oszillators wird als Periodenfunktion $o(t) = 1 + \tanh(\gamma(\cos(2\pi\phi(t)) - 1))$ modelliert, wobei der *Output Gain* Parameter γ die Schärfe der Aktivierungsgipfel kontrolliert. Der Oszillator hat drei Adaptationsregel, die vom Eingabesignal $s(t)$ sowie von den Lernraten η_1, η_2, η_3 abhängen. Die erste Regel erzielt eine Anpassung der Phase:

$$\Delta t_x = \eta_1 s(t) \frac{p}{2\pi} \text{sech}^2(\gamma(\cos(2\pi\phi(t)) - 1)) \sin(2\pi\phi(t))$$

Die zweite Regel passt die Periode an:

$$\Delta p = \eta_2 s(t) \frac{p}{2\pi} \text{sech}^2(\gamma(\cos(2\pi\phi(t)) - 1)) \sin(2\pi\phi(t))$$

Die dritte adaptiert eine Schätzung Ω der Eingabevariabilität:

$$\Delta\Omega = \eta_3 s(t) \text{sech}^2(\gamma(\cos(2\pi\phi(t)) - 1)) \\ (\cos(2\pi\phi(t)) + 2\gamma(o(t) - 1)\sin^2(2\pi\phi(t)))$$

Diese Schätzung bestimmt die Breite des rezeptiven Feldes τ des Oszillators, d.h. die Länge eines Zeitfensters um seine maximale Aktivierung, innerhalb dessen er sich stark an Eingangssignale anpasst: $\tau = \tau_{max} + 0.5(\tau_{min} - \tau_{max})(1 + \tanh\Omega)$. Der Anstieg der Ausgabe (Output gain) verhält sich invers zur Breite des rezeptiven Felds: $\gamma = \frac{-0.416}{\cos(2\pi\tau)-1}$. Mit verringerter Eingabevariabilität schrumpft das rezeptive Feld, die Ausgabegipfel werden schärfer, gibt es hingegen mehr Eingabevariabilität, wächst das rezeptive Feld, die Ausgabegipfel werden flacher. Der

Ausgabewert $o(t)$ wird multipliziert mit einem Konfidenzwert $c = c_{max} + 0.5(c_{min} - c_{max})(1 + \tanh\Omega)$.[1] Da wir davon ausgeben, dass sich die Silbenperioden im Bereich $[0.1, 0.25]$, und Perioden metrischer Füße im Bereich $[0.2, 0.5]$ bewegen, initialisieren wir die Perioden der Silben- und Fußoszillatoren auf den Mittelwert dieser Bereiche, 0.175 und 0.35.

3.2 Phasenresetoszillator (PRO)

Dieser Oszillator wurde ursprünglich von [13] für die Wahrnehmung von Sprache vorgeschlagen, und von [16] für Anwendungen der Mensch-Maschine Kommunikation modifiziert. Seine Ausgaben ist wie beim PAO eine Periodenfunktion, welche hinsichtlich einer Schärfung der Ausgabegipfel modifiziert wird. Zusätzlich wird ein Term für die exponentielle Dämpfung definiert:

$$o(t) = \left(\frac{1+\cos(2\pi\phi(t)}{2}\right)^{(1-\Omega(n))\gamma_{min}+\Omega(n)\gamma_{max}} \exp(-\frac{\beta t_x}{p_{ini}})$$

Die Phase $\phi(t)$ bleibt im Bereich $[-0.5, 0.5]$, im Falle einer Eingabe gibt es einen Reset nach 0.0. Synchronizität $\Omega(n) = (1-\varepsilon)\Omega(n-1) + \varepsilon(1-2|\phi^r(n)|)$ wird jedes Mal überprüft, wenn ein Eingabesignal erfolgt: $\phi^r(n)$ ist die Oszillatorphase zum Zeitpunkt des Reset und $\varepsilon = 0.2$ ist ein Parameter, welcher den gegenwärtigen Impuls gegenüber der gespeicherten Synchronizität füherer Impulse gewichtet. $\gamma_{min} = 1$ und $\gamma_{max} = 5$ schränken den Umfang der Ausgabeschärfung ein, welche von der gemessenen Synchronizität mit der Folge von Eingangsimpulsen abhängt. Der letzte Term der Ausgabegleichung dämpft die Exponentialität der Ausgabe für den Fall, dass keine Eingabe erfolgt. $\beta = 0.5$ ist die Dämpfungsrate, p_{ini} ist die initiale Periode des Oszillators, und t_x die seit der letzten Eingabe vergangene Zeit. Die Periode wird angepasst durch $\Delta p = \alpha \Delta t P M \frac{p}{2}$, wobei $\alpha = 1$ die Entrainmentrate ist, der Periodenkoppelungsterm $P = \phi^r(n)(1-\Omega(n))$ hängt von der Synchronizität, von der Phase beim letzten Reset, sowie von der Funktion der Impulsantwort ab $M = \frac{1}{1+\exp(-\Gamma(i^r(n)\exp(-\Theta t)-0.5))}$ (bei einem Zuwachs der Impulsantwort $\Gamma = 1000$. Der Impulsantwort-Bias $\Theta = 2$) stellt sicher, das beinahe sämtliche Anpassungen direkt nach der Eingabe erfolgen. Wie im PAO-Modell initialisieren wir die Perioden der Silben- und Fußoszillatoren auf 0.175 bzw. 0.35. Sämtliche weiteren Parameter entnehmen wir der Literatur.

3.3 Phasenresetoszillatornetzwerk (PRN)

Wir verwenden ein Netzwerk bestehend aus 20 parallelen Oszillatoren, die ähnlich dem PRO-Modell aufgebaut sind. Einziger Unterschied besteht darin, dass die Ausgabe nicht im Fall fehlender Eingangssignale gedämpft wird, sondern dann, wenn der Oszillator sich nicht synchron zur Folge der Eingangssignale verhält.

$$o_i(t) = \exp(c_d(1-\sigma_i(t)))(\frac{1+\cos(2\pi\phi(t))}{2})^{c_s}$$

Die Konstante $c_s = 20$ bestimmt die Schärfe des Oszillatorausgabesignals (je mehr Oszillatoren wir im Netzwerk für einen gegebenen Frequenzumfang haben, umso höher dürfte diese Konstante sein, um die eine Unschärfe der Netzwerkausgabe zu reduzieren), wobei $c_d = -20$ bestimmt, wie stark die Oszillatorausgabe abhängig von seiner jeweiligen Asynchronizität

[1] Für weitere Erklärungen verweisen wir auf die angegebene Literatur. Die folgenden Parametereinstellungen wurden ebenfalls der Literatur entnommen: $\eta_1 = 1.0$, $\eta_2 = 0.3$, $\eta_3 = 0.3$, $\tau_{min} = 0.02$, $\tau_{max} = 0.5$, $c_{min} = 0.0$, $c_{max} = 1.0$.

Modell	Phrasen als Eingabedaten			Regelmäßige Kontrolldaten		
	Vorhersage an Vokal-einsätzen	Vorhersage zu anderen Zeitpunkten	Differenz	Vorhersage an Vokal-einsätzen	Vorhersage zu anderen Zeitpunkten	Differenz
PAO	0.241±0.005	0.265±0.006	-0.024±0.007	0.593±0.030	0.186±0.010	0.408±0.041
PRO	0.296±0.011	0.327±0.004	-0.031±0.013	0.544±0.033	0.274±0.007	0.270±0.034
PRN1	0.311±0.013	0.273±0.006	0.039±0.010	0.854±0.012	0.257±0.005	0.597±0.014
PRN2	0.311±0.013	0.168±0.006	0.143±0.011	0.854±0.012	0.139±0.005	0.715±0.014

Tabelle 1 - Vorhersage der Vokaleinsätze (Ausgabemittelwert) für verschiedene Osziallatormodelle mit Phrasen als Eingabe.

gedämpft wird. Die Synchronizität wird immer dann gemessen, wenn ein Eingangssignal registriert wird: $\sigma_i(t_r) = (1-c_p)\sigma_i(t_{r-1}) + c_p(1-\exp(c_e\phi(t_r)^2)$, wobei $c_p = 0.2$ eine Konstante ist, welche den gegenwärtigen Impuls mit der gespeicherten Synchronizität früherer Ereignisse (analog zum PRO-Modell) vergleicht. $c_e = -200$ bestimmt hierbei, bis wann ein Vorhersagefehler weiterhin als synchron betrachtet wird. Durch die Verwendung eines exponentialen anstatt eines stückweise linearen Terms (wie im PRO-Modell) wird sichergestellt, dass nur wenige Oszillatoren als synchron zum Eingabesignal betrachtet werden. Dies führt wiederum zu einer Reduktion der Unschärfe der Netzwerkausgabe. Die Periodenadaptation kommt in im PRN-Modell nicht zum Einsatz, der Phasenreset funktioniert identisch zum PRO-Modell.

Die initialen Perioden werden logarithmisch verteilt im Bereich von $[0.1, 0.25]$ für Silben, und im Bereich von $[0.2, 0.5]$ für metrische Füße. Es existiert zusätzlich eine Netzwerkausgabeeinheit mit einer sigmoiden Ausgabefunktion $n(t) = 1/(1+\exp(-\sum_i o_i(t-1)))$, wobei $o_i(t)$ die Ausgaben der individuellen Oszillatoren sind. Diese Variante des Modells nennen wir PRN1. In der Variante PRN2 ist die Ausgabeeinheit zusätzlich mit der Netzwerkeingabe verknüpft. Nach einem Eingabeereignis bleibt die Ausgabe so lange auf null, bis die Summe der Eingabe einen positiven Anstieg erreicht hat. Da viele hohe Oszillatorausgaben direkt nach einem Eingabeereignis dieses Verhalten stören könnten, wird zusätzlich eine absolute Refräkturperiode im Umfang von 5 Simulationsschritten nach jedem Eingabeereignis erzwungen.

4 Ergebnisse und Diskussion

Wir haben verschiedene Eingabedaten (spontansprachliche Einzelphrasen vs. Gesamtkonversationen) und verschiedene Oszillator-basierte *Entrainment*-Modelle miteinander verglichen. Für die Modellevaluation messen wir die mittlere Ausgabeaktivierung des jeweiligen Oszillatormodells in den Fällen, in denen ein Eingabesignal vorliegt, und in den Fällen, in denen kein Eingabesignal vorliegt. Die Differenz beider Mittelwerte bildet unser Performanzmaß für das jeweilige Modell (vgl. [8]). Anders als in früheren Analysen [12] zeigt sich kein klarer Vorteil von PAO gegenüber PRO. Wenn die Oszillatoren auf Phrasen als Eingaben arbeiten, liegt ihre Performanz außerdem im Bereich der Zufallswahrscheinlichkeit oder sogar darunter (siehe exemplarisch Tabelle 1 und 2 für einen Datensatz basierend auf Phrasen). Die auf Oszillatorbanken basierenden Modelle verbessern die Performanz gegenüber einzelnen Oszillatoren signifikant über Zufallsniveau. Eine weitere signifikante Verbesserung konnte auf beinahe sämtlichen Daten erzielt werden, indem die Refräkturperiodenregel hinzugefügt wurde. Beispiele für die Ausgabetrajektorien der verschiedenen Oszillatormodelle sind in Abbildung 1 zu sehen.

Wenn als Eingaben Konversationen verwendet werden, zeigt sich eine Verbesserung der Einze-

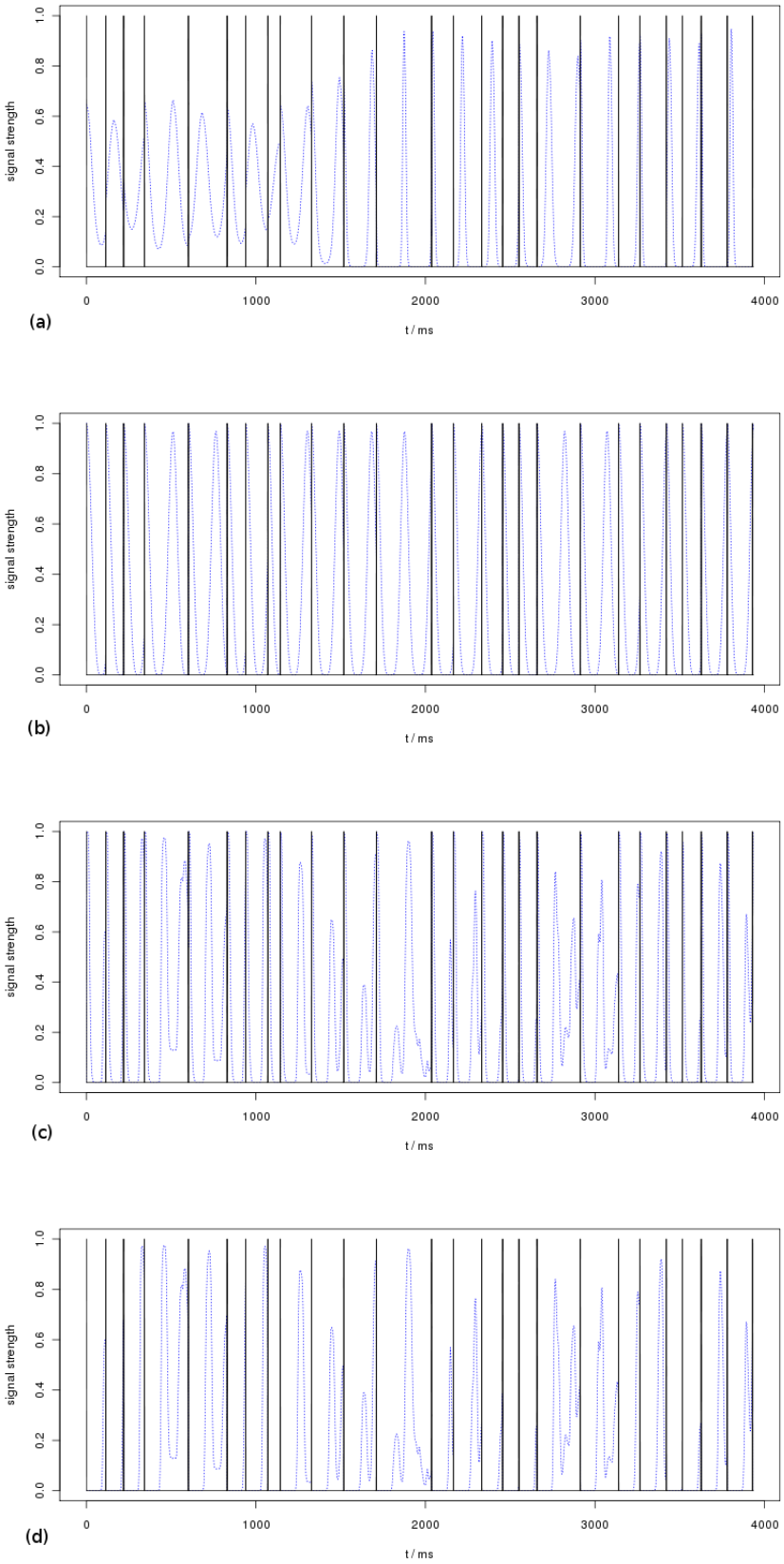

Abbildung 1 - Beispiele für die Ausgabetrajektorien (blau) der verschiedenen Oszillatormodelle und die Vokaleinsätze (schwarz) für Silben der deutschen Phrase "... eine Urlaubsreise mit meiner Familie, also ich war mit meiner Schwester und meiner Mutter dort.". Dargestellt sind das (a) PAO-Modell, das (b) PRO-Modell, das (c) PRN1-Modell sowie das (d) PRN2-Modell.

Modell	Phrasen als Eingabedaten			Regelmäßige Kontrolldaten		
	Vorhersage am metrischen Füßen	Vorhersage zu anderen Zeitpunkten	Differenz	Vorhersage am metrischen Füßen	Vorhersage zu anderen Zeitpunkten	Differenz
PAO	0.260±0.010	0.288±0.004	-0.028±0.013	0.474±0.029	0.230±0.008	0.244±0.036
PRO	0.316±0.017	0.333±0.004	-0.018±0.018	0.561±0.028	0.322±0.005	0.239±0.030
PRN1	0.356±0.015	0.318±0.006	0.038±0.014	0.714±0.022	0.305±0.006	0.409±0.023
PRN2	0.356±0.015	0.207±0.008	0.149±0.015	0.714±0.022	0.187±0.007	0.527±0.022

Tabelle 2 - Vorhersage der metrischen Füße (Ausgabemittelwert) für verschiedene Oszillatormodelle mit Phrasen als Eingabe.

Modell	Vorhersagen für Vokaleinsätze			Vorhersagen für metrische Füße		
	Vorhersage für Vokal-einsätze	Vorhersage zu anderen Zeitpunkten	Differenz	Vorhersage an metrischen Füßen	Vorhersage zu anderen Zeitpunkten	Differenz
PAO	0.126±0.007	0.120±0.001	0.007±0.007	0.268±0.011	0.193±0.001	0.071±0.011
PRO	0.304±0.009	0.293±0.001	0.011±0.011	0.299±0.014	0.295±0.002	0.004±0.014
PRN1	0.262±0.007	0.227±0.001	0.011±0.007	0.333±0.015	0.259±0.001	0.074±0.012
PRN2	0.262±0.007	0.153±0.001	0.109±0.007	0.333±0.012	0.181±0.001	0.152±0.012

Tabelle 3 - Vorhersage der Vokaleinsätze und metrische Füße (Ausgabemittelwert) für verschiedene Osziallatormodelle mit einer Konversation als Eingabe.

loszillatoren – ihre Performanz liegt teilweise oberhalb des Zufallsniveaus. Es kann allerdings wiederum kein Vorteil für eines der beiden Modelle PAO oder PRO ausgemacht werden. Ferner zeigt sich, dass die Einzeloszillatoren auch bei langfristigem *Entrainment* nicht robuster sind als die Oszillatorbanken. Für die Oszillatorbanken kann allerdings keine klare Verbesserung erzielt werden (siehe Tabelle 3).

Zusammenfassend zeigen die Experimente, dass sich Netzwerke basierend auf 20 parallelen Oszillatoren mit Phasenreset *oder* gradueller Phasenanpassung ähnlich gut eignen, um sich an ein spontansprachliches Eingangssignal rhythmisch anzupassen. Oszillatorbanken sind Modellen einzelner Oszillatoren klar überlegen, sowohl im Bereich der schnellen Anpassung auf der Ebene von Einzelphrasen als auch auf längeren, ganze Konversationen umfassenden Eingabedaten.

5 Schlussfolgerung und Ausblick

Unsere bisherigen Experimente konnten zeigen, dass oszillatorbasierte *Entrainment*-Modelle generell in der Lage sind, aus spontansprachlichen Eingabesignalen die Ereigniszeitpunkte zukünftiger prosodischer Ereignisse auf verschiedenen prosodischen Ebenen vorherzusagen. Dies ist angesichts der starken rhythmischen Schwankungen spontansprachlicher Daten (im Vergleich zur Musik) ein vielversprechendes Ergebnis. Oszillatorbanken haben sich in diesen

Experimenten als geeignet gezeigt, sich schnell an ein Eingangssignal anzupassen. Einzelne Oszillatoren benötigen längere Eingabedaten (hier: ganze Konversationen), um *Entrainment*-Prozesse zu modellieren.

Derzeit arbeiten wir an einer Kopplung von Oszillatormodellen, um die komplexe Interaktion unterschiedlicher prosodischer Ebenen, aber auch das Feedbackverhalten in Sprechpausen besser berücksichtigen zu können. Basierend auf früheren Arbeiten [11] wird außerdem ein Prototyp eines künstlichen Agenten entwickelt, welcher auf der Basis sprachlichen Inputs multimodale Feedbacksignale (Äußerungen, Kopfnicken oder manuelle *Beat*-Gesten) generiert. Darüber hinaus arbeiten wir an verbesserten Evaluationsmodellen für die Oszillatorantwort, welche auch psychoakustische Erkenntnisse, z.B. wahrnehmbare Daueruntерschiedsschwellen mit berücksichtigen.

Literatur

[1] ABERCROMBIE, D.: *Elements of general phonetics*. Aldine Publishing Corporation, Chicago, 1967.

[2] ALLWOOD, J., J. NIVRE und E. AHLSÉN: *On the semantics and pragmatics of linguistic feedback*. Journal of Semantics, 9:1–26, 1992.

[3] BUSCHMEIER, H., Z. MALISZ, M. WLODARCZAK, S. KOPP und P. WAGNER: *'Are you sure you're paying attention?' – 'Uh-huh' Communicating understanding as a marker of attentiveness*. In: *Proceedings INTERSPEECH 2011*, S. 2057–2060, Florence, Italy, 2011.

[4] CUMMINS, F. und R. PORT: *Rhythmic constraints of stress timing in English*. Journal of Phonetics, 26:145–171, 1998.

[5] GHITZA, O.: *Linking speech perception and neuropsychology: Speech decoding guided by cascaded oscillators locked to the input rhythm*. Frontiers in Psychology, 2:130, 2011.

[6] GHITZA, O. und S. GREENBERG: *On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence*. Phonetica, 66:113–126, 2009.

[7] HIRSCHERG, J.: *Speaking more like you: Entrainment in Conversational Speech*. In: *Proceedings INTERSPEECH 2011 (abstract)*, Florence, Italy, 2011.

[8] INDEN, B., Z. MALISZ, P. WAGNER und I. WACHSMUTH: *Rapid entrainment to spontaneous speech: A comparison of oscillator models*. In: MIYAKE, N., D. PEEBLES und R. COOPER (Hrsg.): *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 2012.

[9] LARGE, E.: *Resonating to musical rhythm*. In: GRONDIN, S. (Hrsg.): *The psychology of time*. West Yorkshire: Emerald, 2008.

[10] LARGE, E. und J. KOLEN: *Resonance and the Perception of Musical Meter*. Connection Science, 6(1):177–208, 1994.

[11] LESSMANN, N., A. KRANSTEDT und I. WACHSMUTH: *Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent Max*. In: *Proceedings of embodied conversational agents workshop. AAMAS '04*, S. 57–64, New York, 2004.

[12] MALISZ, Z., B. INDEN, P. WAGNER und I. WACHSMUTH: *An oscillator based modeling of German spontaneous speech rhythm*. In: *PoRT Workshop*, Glasgow, Scotland, 2012.

[13] MCAULEY, J. D.: *Perception of Time as Phase: Towards and Adaptive Oscillator Model of Rhythmic Processing*. Doktorarbeit, Indiana University, 1995.

[14] MORENCY, L.-P., I. DE KOK und J. GRATCH: *A Probabilistic Multimodal Appraoch for Predicting Listener Backchannels*. Autonomous Agents and Multi-Agent Systems, 1(70-84), 2010.

[15] MORTON, J., S. MARTIN und C. FRANKISH: *Perceptual Centers (P-centers)*. Psychological Review, 83:405–408, 1976.

[16] NERLICH, U.: *Rhythmische Segmentierung sprachlicher Instruktionen in einem Mensch-Maschine-Kommunikations-Szenario*. Diplomarbeit, Technische Fakultät, Universität Bielefeld, 1998.

[17] POPPE, R., K.-P. TRUONG und D. HEYLEN: *Backchannels: Quantity, Type and Timing Matters*. In: *International Conference on Intelligent Virtual Agents, IVA*, S. 15–17, 2011.

[18] SCOTT, S.: *P-centers in Speech: An Acoustic Analysis*. Doktorarbeit, University College London, 1993.

[19] WACHSMUTH, I.: *Communicative Rhythm in Gesture and Speech*. Lecture Notes in Computer Science, 1739:277ff., 1999.

[20] WARD, N. und W. TSUKAHARA: *Prosodic features which cue back-channel responses in English and Japanese*. Journal of Pragmatics, 23:1177–1207, 2000.

[21] WILSON, M. und T. P. WILSON: *An oscillator model of the timing of turn-taking*. Psychonomic Bulletin & Review, 12(6):957–968, 2005.

[22] WŁODARCZAK, M., J. ŠIMKO und P. WAGNER: *Syllable entrainment in overlapped speech*. In: *Proceedings of Speech Prosody 2012*, S. 611–614, Shanghai, China, 2012.