Rapid entrainment to spontaneous speech: A comparison of oscillator models

Benjamin Inden
Faculty of Technology
Bielefeld University
binden@techfak.uni-bielefeld.de

Zofia Malisz
Faculty of Linguistics and Literary Studies
Bielefeld University

Petra Wagner
Faculty of Linguistics and Literary Studies
Bielefeld University

Ipke Wachsmuth
Faculty of Technology
Bielefeld University

Abstract

Oscillator models may be used for modeling synchrony between gestures and speech, or timing of backchanneling and turn-taking in dialogues. We find support for the hypothesis that oscillator networks can better predict rhythmic events on the syllable and foot level than single oscillators, but we do not find support for the hypothesis that phase resetting oscillators perform better that phase adapting oscillators. Overall, oscillators can be used to predict rhythmic events in speech, but higher level information needs to be integrated into such models to reach a satisfactory performance.
Keywords: speech rhythm, entrainment

Introduction

Spontaneous speech, like music, exhibits temporal regularities, but these cannot be captured by simple descriptions. Rhythm, i.e., hierarchical structured temporal regularities, is widely believed to be an important principle for understanding both music and speech (Large, 2008; Cummins & Port, 1998). Regularity of timing greatly contributes to speech perception and understanding. Regular sequences of, e.g., interstress intervals in speech or tone sequences in music speed up perception by facilitating meaningful grouping and contrast within a very rich acoustic signal. The role of rhythmic expectancies both in speech and music perception has been the basis of Dynamic Attending Theory (Jones, 1990) and more specific phonological models such as PolySP (Hawkins, 2003). Humans can even perceive rhythms in music that do not directly correspond to any frequency found in a spectral analysis of the signals (Large, 2008). Similarly, the timing of speech production is coordinated via rhythmic principles. The same principles govern the neuro-physiological dynamics of all motor behavior. On the syllable level, the vocalic pulse represents the basic timing coordination of the articulatory system (Browman & Goldstein, 1992).

When measuring brain activity, one can easily find a number of prominent frequencies. Some of them are in the range of typical speech units: the theta band (3-12 Hz) corresponds to the typical duration of a syllable (100-300 ms), whereas delta band oscillations (0.5-3 Hz) correspond to typical lengths of prosodic and metrical units. Many kinds of oscillators have the property of entraining to an externally provided periodic signal, i.e., they become phase-locked to the signal. Therefore, it seems plausible that neural oscillators might play a role in the production and perception of speech by synchronizing certain systems with the speech signal (Buzsáki & Draguhn, 2004; Ghitza & Greenberg, 2009). Previous research has also found that gestures may be synchronized with speech rhythms (Condon, 1986; Tuite, 1993; Wachsmuth, 1999; Loehr, 2007). Furthermore, listeners can become entrained to a speaker's rhythm, which helps them to provide backchanneling or take turns in a dialogue at a suitable moment in time (Wilson & Wilson, 2005).

A number of oscillator models have been proposed that can entrain to musical rhythms. Here we focus on models proposed by Large and McAuley (Large, 1994; McAuley, 1995). These have been shown to achieve entrainment to input signals not just in a period ratio of 1:1, but also in more complex ratios, which makes coupling between several levels of speech rhythm possible. Furthermore, oscillator banks have been shown to reproduce empirical findings about human perception of rhythm: they can resonate at frequencies that are not present in the input signals, but perceived by human subjects, too (Large, 2008; Large, Almonte, & Velasco, 2010).

We are interested in whether these oscillator models originally built for modeling music perception can also be used to model human entrainment to less regular speech signals. In general, we believe in the necessity to couple oscillators for different levels of the rhythmic hierarchy, so ultimately we will include this in the models discussed in this article. However, here we focus on the question what particular features might make an oscillator model more capable of correctly predicting syllable and foot onset times when considered separately from the other levels of the rhythmic hierarchy. After all, speech is less regular than music, so adapta-

tion to input signals should be very fast. Therefore, we will compare two previously proposed oscillator models and then make a number of changes to one of them to see whether prediction performance improves. In particular, we examine the following hypotheses: First, oscillators that reset their phase upon arrival of an input signal may be faster than those that adapt their phase gradually. Second, it may be better to have a bank of oscillators tuned to different frequencies than to have a single oscillator that adapts its period. This would be because period adaptation time is dependent on the amount of change necessary whereas in banks of oscillators, the time for a differently tuned oscillator to become activated is constant with regards to the amount of frequency change. To the degree these hypotheses turn out to be supported by the data, they can inform future modeling of human entrainment to speech rhythms. Besides addressing these two hypotheses, our experiments also show how much can be learned at all by oscillator models without considering the hierarchical organization of speech, i.e., from a pure low-level approach.

The Data

The speech data comes from a corpus of spontaneous dialogue in German where one dialogue partner told a holiday story and the other was instructed to listen actively (Buschmeier, Malisz, Włodarczak, Kopp, & Wagner, 2011). The corpus was collected for the purposes of modeling entrainment in dialogue, multimodal behavior of the listener, i.e., feedback signals, head and manual gesture, as well as the prosody of the storyteller. The latter objective is addressed in the present paper.

Audiovisual recordings were made in a sound-treated studio. Participants were positioned approximately three meters apart to minimize crosstalk. Close talking high-quality headset microphones were used. The signal properties were annotated in Praat (Boersma & Weenink, 2012). Careful annotation of the acoustic signal enables to approximate emergent rhythmic phenomena (Gibbon & Fernandes, 2005). To represent the syllabic oscillator hypothesized for speech production, we first semi-automatically extracted vowel onsets from the data (Cummins & Port, 1998; Barbosa, 2006). Secondly, experts annotated rhythmic feet, representing the slower stress oscillator, where each prominent syllable is a pulse on that level. We also annotated interpausal units (IPUs) with a criterion that only minimally perceptible interruptions in the flow of speech were marked (not all acoustic pauses).

For the present simulations, two conversations (henceforth *dataset 1* and *dataset 2*) were used. Phrases (IPUs) consisting of at least two feet events were selected. Any phrase initial unstressed vowel events (anacruses) were excluded as well as the phrase final vowel event. The trimmings were done to exclude any extra lengthening at the end of phrase and extra irregularity at the beginning of phrase that typically signal a boundary in German. The resulting phrases consist of fluent, spontaneous, uninterrupted speech with a minimal phrase length of one second. The mean duration of a syllable-sized intervocalic interval was 125 msec in this material and 365 msec for the foot. 69 phrases each from dataset 1 and dataset 2 were provided as input to the different oscillator models, i.e., the resulting onset times for each vowel or foot event served as the input pulse.

For each conversation, a control set of regular phrases was created by generating completely regular pulses with frequencies equal to the mean frequencies of events in the corresponding individual phrases from the conversation data.

Models of entrainment

Phase adaptation oscillator (PAO)

This oscillator model is one of several similar models originally proposed by Large for entrainment to musical rhythms (Large, 1994). The phase of this oscillator is defined as $\phi(t) = \frac{t-t_x}{p}$, where t_x is the time of the last event (in the input or according to the oscillator's expectation) and p is the period of the oscillator. The phase is reset to 0.0 when it reaches 1.0. The output of the oscillator is modeled as a periodic function $o(t) = 1 + \tanh(\gamma(\cos(2\pi\phi(t)) - 1))$, where the output gain parameter γ controls the sharpness of the activity peaks. The oscillator has three adaptation rules that depend on the input signal $s(t)$ as well as learning rates η_1, η_2, η_3. The first rule in effect adapts the phase:

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

The second adapts the period:

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

The third adapts an estimate Ω of input variability:

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

This estimate in turn determines the receptive field width τ of the oscillator, i.e., the width of a window in time around its maximal activation where it is highly adaptive to input signals: $\tau = \tau_{max} + 0.5(\tau_{min} - \tau_{max})(1 + \tanh\Omega)$. The output gain is inversely related to the receptive field width: $\gamma = \frac{-0.416}{\cos(2\pi\tau)-1}$. So if there is less input variability, the receptive field shrinks, and the output peaks are sharper, whereas if there is more input variability, the receptive field grows, and the output peaks are softer. Finally, the output value $o(t)$ is multiplied by a confidence value $c = c_{max} + 0.5(c_{min} - c_{max})(1 + \tanh\Omega)$. Further explanations about the motivation behind these choices, and the behavior of the oscillator, can be found in the literature. The following parameter settings were also taken from the literature: $\eta_1 = 1.0$, $\eta_2 = 0.3$, $\eta_3 = 0.3$, $\tau_{min} = 0.02$, $\tau_{max} = 0.5$, $c_{min} = 0.0$, $c_{max} = 1.0$. Because we expect syllable periods to be in the range $[0.1, 0.25]$, and feet periods in the range $[0.2, 0.5]$, we set the initial periods of the period and feet oscillators to the middle of these ranges, i.e. 0.175 and 0.35.

Phase reset oscillator (PRO)

This oscillator has been originally proposed by McAuley for the perception of music, and modified by Nerlich in the context of human-machine interaction (McAuley, 1995; Nerlich, 1998). Its output, like that of the PAO, is a periodic function modified to modulate the sharpness of the output peaks, together with a term for exponential decay of the output:

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

The phase $\phi(t)$ is always kept in the range $[-0.5, 0.5]$, and reset to 0.0 when an input event arrives. The synchrony $\Omega(n) = (1-\varepsilon)\Omega(n-1) + \varepsilon(1-2|\phi^r(n)|)$ is measured every time an input event arrives: $\phi^r(n)$ is the phase of the oscillator at the reset, and $\varepsilon = 0.2$ is a parameter that weights the current impulse against the memory of earlier synchrony with input events. $\gamma_{min} = 1$ and $\gamma_{max} = 5$ constrain the range of output sharpening that is dependent on measured synchrony with the train of input events. The final term in the output equation dampens the output exponentially when no input arrives. $\beta = 0.5$ is the decay rate, p_{ini} the initial period of the oscillator, and $t_x$the time since the last input event.

The period is adapted using $\Delta p = \alpha \Delta t P M \frac{p}{2}$, where $\alpha = 1$ is the entrainment rate, the period coupling term $P = \phi^r(n)(1-\Omega(n))$ is dependent on the synchrony and on the phase at the last reset, and the impulse response function $M = \frac{1}{1+\exp(-\Gamma(i^r(n)\exp(-\Theta t)-0.5))}$ (with impulse response gain $\Gamma = 1000$ and impulse response bias $\Theta = 2$) ensures that almost all adaptation is done shortly after an input event. Like in the PAO model, we set initial periods of the period and feet oscillators to 0.175 and 0.35, while all other parameters are taken from the literature.

Phase reset oscillator network (PRN)

We use a network of 20 parallel oscillators that are similar to the PRO model. However, we let the output decay not when no input arrives as before, but when the individual oscillator is not synchronous with the train of input signals:

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

The constant $c_s = 20$ determines the sharpness of the oscillator output signal (the more oscillators we have in the network for a given frequency range, the higher this constant should be to reduce blurring of the network output), while $c_d = -20$ determines how much the oscillator output decays depending on its asynchrony. The synchrony is measured each time an input event arrives using $\sigma_i(t_r) = (1-c_p)\sigma_i(t_{r-1}) + c_p(1-\exp(c_e\phi(t_r)^2)$, where $c_p = 0.2$ is a constant that weights the current impulse against the memory of earlier synchrony with input events just like in the PRO model, and $c_e = -200$ determines how much prediction error is still considered synchronous. Using an exponential term here instead of a piecewise linear term as in the PRO model ensures that only a few oscillators will consider themselves synchronous with the input signal, which again reduces blurring of the network output. Period adaptation is not used in the PRN model, but phase reset works just like in the PRO model.

The initial periods are logarithmically distributed in the range of $[0.1, 0.25]$ for syllables, and $[0.2, 0.5]$ for feet. There is an additional network output unit with a sigmoid output function $n(t) = 1/(1+\exp(-\sum_i o_i(t-1)))$, where the $o_i(t)$ are the outputs of the individual oscillators. This variant of the model is called PRN1. In the variant called PRN2, the output unit is also connected to the network input. After an input event, its output remains zero until the sum of its input has a positive slope. Because there may be high oscillator outputs immediately after an input event that could disrupt this behavior, an absolute refractory period of 5 simulation steps after an input event is enforced unconditionally.

Results

In experiments presented elsewhere (Malisz, Inden, Wachsmuth, & Wagner, 2012), we fed event signals from the whole conversation into PAO and PRO models and measured their internal phases when an input signal arrived. In those experiments, we found a significant advantage of the PRO model over the PAO model. By contrast, here we feed data from individual phrases separately into the oscillators and measure their average output activation when an input signal arrives. We also measure average output activation when no input signal arrives and take the difference between the averages as a performance measure. As Tables 1 to 4 show, there is no significant advantage of the PRO model over the PAO model in this case. Furthermore, both are at or below random level on most of the real data sets.

As Tables 1 to 4 also show, using a bank of oscillators like PRN1 is a significant improvement over using a single oscillator (and is significantly above random level). When adding the refractory period rule to the oscillator network, performance further improves significantly for almost all datasets.

The output trajectories of the different oscillator models for an example phrase can be seen in Fig. 1.

Discussion

Our experiments do provide some support to the hypothesis that oscillator networks may be better suited to speech data than single oscillators that adapt their period. Such insights can inform modeling of human rapid entrainment to spontaneous speech. However, the experiments provide no support for the hypothesis that phase resetting oscillators like the McAuley oscillator are better suited to the rather irregular speech data than phase adapting oscillators like the Large oscillator. This might be because performance of both models is so close to chance level when used in that way. More than anything else, these results show that the level of prediction performance that can be reached by considering just one level of speech rhythm is rather low regardless of the used oscillator models.

oscillator model	phrase data			regular control data		
	prediction at vowel onset	prediction at other times	difference	prediction at vowel onset	prediction at other times	difference
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

Table 1: Prediction of vowel onsets (mean oscillator output) for different oscillator models and dataset 1.

	phrase data			regular control data		
	prediction at foot event	prediction at other times	difference	prediction at foot event	prediction at other times	difference
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

Table 2: Prediction of foot events (mean oscillator output) for different oscillator models and dataset 1.

oscillator model	phrase data			regular control data		
	prediction at vowel onset	prediction at other times	difference	prediction at vowel onset	prediction at other times	difference
PAO	0.246±0.006	0.261±0.005	-0.015±0.007	0.689±0.023	0.146±0.008	0.543±0.030
PRO	0.295±0.011	0.323±0.003	-0.027±0.012	0.627±0.027	0.254±0.005	0.372±0.026
PRN1	0.329±0.013	0.273±0.005	0.056±0.010	0.879±0.006	0.252±0.003	0.628±0.008
PRN2	0.329±0.013	0.169±0.005	0.160±0.010	0.879±0.006	0.136±0.003	0.743±0.007

Table 3: Prediction of vowel onsets (mean oscillator output) for different oscillator models and dataset 2.

	phrase data			regular control data		
	prediction at foot event	prediction at other times	difference	prediction at foot event	prediction at other times	difference
PAO	0.307±0.011	0.293±0.004	0.015±0.013	0.344±0.019	0.281±0.006	0.063±0.024
PRO	0.376±0.020	0.358±0.005	0.018±0.020	0.470±0.026	0.372±0.006	0.098±0.029
PRN1	0.255±0.021	0.236±0.016	0.019±0.012	0.558±0.031	0.245±0.014	0.313±0.024
PRN2	0.255±0.021	0.172±0.013	0.083±0.014	0.558±0.031	0.172±0.012	0.386±0.026

Table 4: Prediction of foot events (mean oscillator output) for different oscillator models and dataset 2.

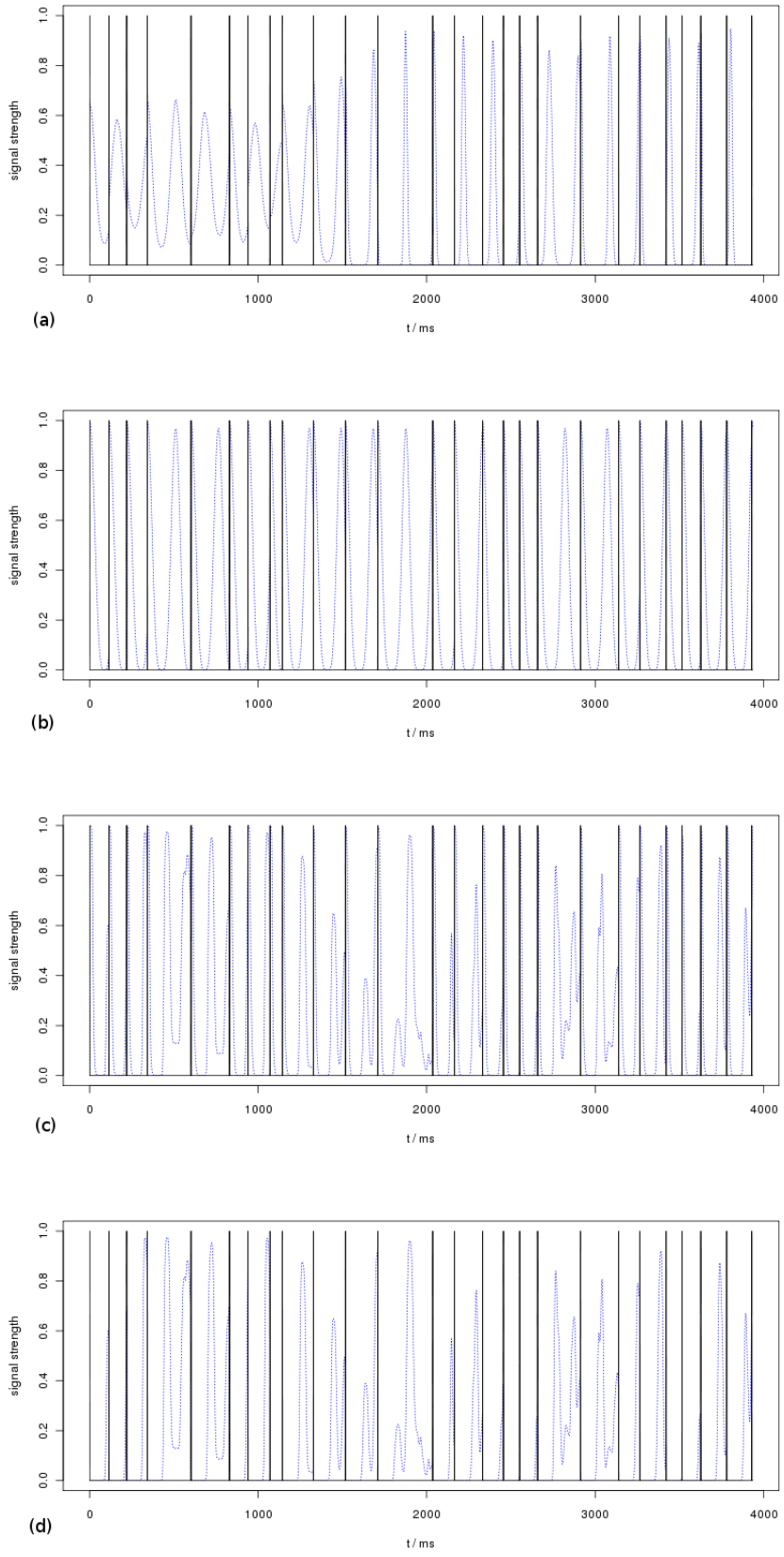

Figure 1: Example output trajectories (blue) and vowel onsets (black) for syllables in the German phrase "... eine Urlaubsreise mit meiner Familie, also ich war mit meiner Schwester und meiner Mutter dort." ("... a vacation trip with my family, that is, I was there with my sister and my mother.") (a) PAO model, (b) PRO model, (c) PRN1 model, (d) PRN2 model.

The parameters used for the oscillator models seem to be reasonable and have been found by looking at the literature (PAO and PRO models) or preliminary experiments (PRN model). However, it cannot be totally excluded that other parameter settings will lead to higher performance. Therefore, we searched the space of the most important parameters of the PAO and PRO models for better performance on a randomly selected subset of the data for one conversation using evolutionary algorithms (De Jong, 2006) (details and results not shown here). While the evolutionary algorithm found different parameter settings that performed better on the training set, the subsequent performance on the complete set of data was only marginally better in most cases, and did not change any of the previously mentioned conclusions.

Future work will include using coupled syllable and foot oscillators, and possibly using evidence for vocal activity rhythms, i.e., cycles in pauses and hesitations in dialogue, to model the structure of the interpausal units (McGarva & Warner, 2003; Merlo & Barbosa, 2010). Ultimately, we aim to use the output from entrained oscillators to control the timing of backchanneling and turn-taking in artificial embodied conversational agents (Kopp, Allwood, Grammer, Ahlsen, & Stockmeier, 2008; Poppe, Truong, Reidsma, & Heylen, 2010).

Acknowledgments This research is kindly supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673.

References

Barbosa, P. A. (2006). *Incursões em torno do ritmo da fala*. Campinas: Pontes.

Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer. version 5.3.04.* Retrieved 30 January 2012, from http://www.praat.org/

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155-180.

Buschmeier, H., Malisz, Z., Włodarczak, M., Kopp, S., & Wagner, P. (2011). 'Are you sure you're paying attention?' – 'Uh-huh'. Communicating understanding as a marker of attentiveness. In *Proceedings of Interspeech 2011* (pp. 2057–2060). Florence, Italy.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, *304*, 1926-1929.

Condon, W. S. (1986). Rhythm in psychological, linguistic and musical processes. In J. Evans & M. Clynes (Eds.), (p. 55-77). Springfield, Ill.: Thomas.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in english. *Journal of Phonetics*, *26*, 145-171.

De Jong, K. A. (2006). *Evolutionary computation — a unified approach*. MIT Press.

Ghitza, O., & Greenberg, S. (2009). 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.

Gibbon, D., & Fernandes, F. R. (2005). Annotation-mining for rhythm model comparison in brazilian portuguese. In *Proceedings of interspeech.*

Hawkins, S. (2003). Roles and representations of systematic phonetic fine detail in speech understanding. *Journal of Phonetics*, *31*, 373-405.

Jones, M. R. (1990). Learning and the development of expectancies: an interactionist approach. *Psychomusicology*, *9*, 193-228.

Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., & Stockmeier, T. (2008). Modeling embodied feedback with virtual humans. In I. Wachsmuth & G. Knoblich (Eds.), *Modeling communication with robots and virtual humans.* Springer-Verlag Berlin Heidelberg.

Large, E. W. (1994). *Dynamic representation of musical structure*. Unpublished doctoral dissertation, The Ohio State University.

Large, E. W. (2008). The psychology of time. In S. Grondin (Ed.), (chap. Resonating to musical rhythm: Theory and experiment). West Yorkshire: Emerald.

Large, E. W., Almonte, F. V., & Velasco, M. J. (2010). A canonical model for gradient frequency neural networks. *Physica D*, *239*, 905-911.

Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture*, *7*, 179-214.

Malisz, Z., Inden, B., Wachsmuth, I., & Wagner, P. (2012). An oscillator based modeling of german spontaneous speech rhythm. In *Perspectives on rhythm and timing workshop.* Glasgow, UK.

McAuley, J. D. (1995). *Perception of time as phase*. Unpublished doctoral dissertation, Indiana University, Bloomington.

McGarva, A. R., & Warner, R. M. (2003). Attraction and social coordination: Mutual entrainment of vocal activity rhythms. *Journal of Psycholinguistic Research*, *32*, 335-354.

Merlo, S., & Barbosa, P. A. (2010). Hesitation phenomena: a dynamical perspective. *Cognitive Processing*, *11*, 251-261.

Nerlich, U. (1998). *Rhythmische Segmentierung sprachlicher Instruktionen in einem Mensch-Maschine-Kommunikations-Szenario*. Unpublished master's thesis, Faculty of Technology, Bielefeld University.

Poppe, R., Truong, K. P., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. In *Proceedings of the intelligent virtual agents conference.*

Tuite, K. (1993). The production of gesture. *Semiotica*, *93*, 83-106.

Wachsmuth, I. (1999). Communicative rhythms in gesture and speech. In *Proceedings of the international gesture workshop on gesture-based communication in human-computer interaction.*

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, *12*, 957-968.