LEGAL NOTICE

Dynamic Perception-Production Oscillation Model in Human-Machine Communication

Igor Jauk
Artificial Intelligence Group
Faculty of Technology
Bielefeld University, Germany
ijauk@techfak.uni-bielefeld.de

Ipke Wachsmuth
Artificial Intelligence Group
Faculty of Technology
Bielefeld University, Germany
ipke@techfak.uni-bielefeld.de

Petra Wagner
Faculty of Linguistics and Literature
Bielefeld University, Germany
petra.wagner@uni-bielefeld.de

ABSTRACT

The goal of the present article is to introduce a new concept of a perception-production timing model in human-machine communication. The model implements a low-level cognitive timing and coordination mechanism. The basic element of the model is a dynamic oscillator capable of tracking *re-occurring* events in time. The organization of the oscillators in a network is being referred to as the *Dynamic Perception-Production Oscillation Model (DPPOM)*. The DPPOM is largely based on findings in psychological and phonetic experiments on timing in speech perception and production. It consists of two sub-systems, a *perception sub-system* and a *production sub-system*. The perception sub-system accounts for information clustering in an input sequence of events. The production sub-system accounts for speech production *rhythmically entrained* to the input sequence. We propose a system architecture integrating both sub-systems, providing a flexible mechanism for perception-production timing in dialogues. The model's functionality was evaluated in two experiments.

General Terms

Human-Machine Communication

Keywords

Oscillation, Timing, Rhythm, Dynamic systems, Perception, Production, Dialogue Technology

1. INTRODUCTION

Modern real-time dialogue systems, such as embodied conversational agents (e.g Max [2]), have to deal with timing and coordination issues. Some of them need *only* to time turn-taking. Some of them also integrate gestures which have to be timed with their own speech. One way is to explicitly specify the synchronization points of speech and gestures (e.g. [2]). Difficulties arise when the demand is to create a system which is capable of *natural* real-time conversation and behavior. Such a system should integrate all possible aspects of timing in dialogues, as the gestures accompanying speech (or vice-versa), but also the adequate and coordinated interaction of the conversational agent with the human speaker. This interaction could be expressed in terms of *turn-taking* and *back channels*. Wilson & Wilson [6] propose an oscillator model for timing phenomena in conversations and show how their oscillator model can explain various phenomena in speech and dialogues, such as syllable timing and turn-taking.

Turn-taking has also been discussed by Bonaiuto and Thórisson [1]. They implemented a hybrid model composed of the *Ymir Turn-Taking Model (YTTM)* and the *Augmented Competitive Queuing (ACQ)*. Using it, they implemented a conversation between two agents. For their experiments they used a restricted set of only few possible non-verbal and speaking actions. However, real natural-language conversation is much more complex. Turn-taking signals have to be filtered from a set of complex acoustic cues and gestures. Another inflexible aspect in their model is that it only achieves successful turn-taking after a training session. In addition, trained neural networks might not be flexible enough to adopt to the high variance of timing in natural-speech conversations.

2. DYNAMIC PERCEPTION-PRODUCTION OSCILLATION MODEL (DPPOM)

We propose a model which implements cognitive low-level timing processes, the *Dynamic Perception-Production Oscillation Model (DPPOM)*. The DPPOM is based on findings in psychological and phonetic experiments on timing in dialogues, speech perception and production. The general task of the model is to produce reaction signals to a rhythmic input. There are several requirements imposed on such a model. First, it has to be able to adapt to different timing situations in real-time dialogues *dynamically*. Second, it has to be able to filter out the necessary signals in order to time the feedback.

2.1 Basic terms

The incoming speech is being referred to as *system input*. The system input is composed of *re-occurring events*. A cue or a set of cues indicating a certain marked event in a conversation will probably indicate the same event at the next point of time of its occurrence (consider a stressed syllable,

a pitch rise at the end of an echo question or a final lengthen at the end of a phrase). The re-occurrence of certain events in speech is being referred to as *speech rhythm*. Each rhythmic event in speech can be described in two ways: (1) *when* does it occur (the *timing* of the event), and (2) *what* does characterize the event (the *form* of the event). Since we consider acoustic events only, we can express the form in terms of *amplitude*. Amplitude mirrors the relative *prominence* of the events and thus represents the cues which indicate individual events.

A timing model should be able to filter out important events, predict their occurrence in the future, and coordinate an entrained response, regardless of the complexity. With this we identify two tasks. The first task is a *perceptual* task. The model has to perceive the input signal and filter the important events. The second task is a *coordination task*. The model has to coordinate production with perception. The challenge in this demand is the flexibility of the model.

2.2 General architecture

Fig. 1 shows a system overview of the *Dynamic Perception-Production Oscillation Model (DPPOM)*.

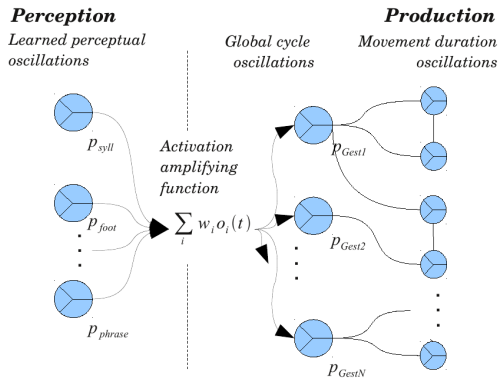

Figure 1: *DPPOM overview.*

The DPPOM is composed of two sub-systems, the *perception* and the *production* sub-system. The system split is motivated by findings which reveal at least two timing components, a perceptual and a motor control component (e.g. Steinbüchel et al [5]). The basic component is an oscillator model. In our implementation we use the oscillator model proposed by Large [3].

Each sub-system implements distinct layers of oscillators organized in a special way. The communication between layers is considered to be a matter of activation. A single oscillator in each layer produces a periodic pulse with a period p and an activation c. The activation is implemented as oscillator *confidence* (Large [3]). Let c_{max} and c_{min} define the confidence range, then the confidence is defined as:

$$c = c_{min} + 0.5(c_{max} - c_{min})(1 + tanh\Omega) \quad (1)$$

where Ω is a control parameter. It is being adjusted during entrainment and translates the confidence value in the defined range. The output confidence of an oscillator can be limited in order to limit its activation capabilities. On the other hand we introduce a threshold level of activation, such that the oscillators might need summed activation in order to be driven. Summed activation results from simultaneous firing in a real-time network. It can be approximated as:

$$\vec{o'} = \sum_i w_i \vec{o_i} \quad (2)$$

where $\vec{o_i}$ is the oscillator output of the oscillator i and w_i the individual oscillator weight. Weighting would provide different coupling strength between oscillators.

We can define the oscillator model in terms of system theory:

$$o(t') = S\{s(t)\} \quad (3)$$

Since t' is not necessary equal to t, both, the input signal, and the system are time variant. Thus, the oscillator output $o(t')$ considers the modification of time t with respect to the input sequence. Considering equations 2 and 3 we can define our system as:

$$\vec{y}(t'') = S\{\sum_i w_i o_i(t')\} \quad (4)$$

where $\vec{y}$ is the output of a vector of oscillators in the output layer. Let γ be a reverse term of the receptive window width of the oscillator, then according to the output definition proposed by Large [3] we define the DPPOM output as:

$$\vec{y}(t'') = S\{\sum_i w_i[1 + tanh(\gamma(cos(2\pi\Phi(t)) - 1)]_i\} \quad (5)$$

It might be a conceptual question, whether the task of the system is to keep the point of time t'' close to the point of time t, or to insert some irregularity. In the next sections we take a closer look at the individual sub-systems.

2.3 Perception sub–system

The perception sub-system processes the input stimuli. It identifies the important events in the input train. The important events in some way are more prominent than others. They indicate places where a reaction or feedback is more probable. We assume, that the important events are organized in a *metrical* structure. The meter defines the re-occurrence of such events. In the Fig. 1 the oscillators, which track the important events, are organized in the perception sub-system layer. They have different periods, according to the periods of the events they track. In the figure we give exemplary oscillators for syllables, feet, and phrases. In this case these oscillators are motivated phonetically. Their periods can be considered as learned during language acquisition.

2.4 Production sub–system

The production sub-system coordinates any kind of DPPOM's output. The output is defined as entrained reaction to the input. In general, the output is given by motor coordination. We can sub-divide motor coordination in two groups, articulatory gestures, and all the other gestures.

Each motor oscillator controls a distinct motor task. We have to configure the motor oscillators concerning the demands of the individual motor tasks, e.g. if we want to time arm movements it might be significantly distinct from timing of head nodding. Another crucial point in the configuration is the concerning of the events, which activate the

reaction of oscillators. Such an important event can be a phrase boundary. Now, we can wish to react to every important occurring event, or only to few of them. This can be controlled by distinct coupling ratios of the motor oscillators and the perceptual system response. We also can construct complex response timing patches by introducing an intermediate motor layer and calculating complex coupling ratios.

If we consider a simple arm movement we can figure out at least two oscillation processes. The first process defines the duration of the actual movement, reflecting the inertia of the arm. This process is important since other motor tasks eventually have to be coordinated with the arm movement (consider simultaneous movement and speech). The second oscillation process is the one which coordinates the global movement cycle, i.e. it defines the re-occurrence of the movements. If we define a motor event to occur at each important input event, then maybe we do not need the global cycle coordination. It will be done intrinsically by the perceptual layer and can be controlled via the activation. However, if we define more complex ratios, then we also have to define the global cycle coordination.

3. PERCEPTION SUB-SYSTEM TEST

In the first experiment we used a German rap song, the first minute annotated on the syllable level. We believe that rap music is rhythmically highly structured and thus could be a good first approach to test our model. Since we model a perceptual system it is reasonable to implement perceptually motivated tracking. Thus we calculated *perceptual centers (P-centers)* of syllables, according to Marcus [4].

We ran a set of three oscillators organized in a parallel fashion, on a syllable, "foot" and "phrase" level, respectively. A "foot" as defined to be the distance between two stressed syllables, thus effectively we tracked *stressed* syllables on this level. A "phrase" is defined to be the distance between two *accentuated* syllables due to the rhythm of the song. We assigned an amplitude of 1 to all unstressed syllables, of 2 to all stressed syllables, of 3 to all accentuated syllables and of 0.5 to all non-speech units (like breathing pauses). We initiated the oscillator periods with mean duration values of 174 ms, 458 ms, and 1868 ms, on the syllable, "foot", and "phrase" levels respectively. The oscillators started firing and entraining when they received the first pulse on their level. We also configured oscillator's adaptation thresholds according to the amplitude of the input syllables, thus the oscillators only entrained to pulses of their own levels.

Derived from the period values, the oscillators coupled with ratios of *1:1* for the syllable level to the input sequence, *2:5* for the "foot" to the syllable/input level, and *1:10* for the "phrase" to the syllable/input levels. We can observe these ratios on Arnold's tongue coupling diagram [3]. The *1:1* ratio is the most stable one, *2:5* and *1:10* are rather unstable. According to the Arnold's tongue diagrams we can adjust the maximal τ values (maximal receptive window size) in order to provide more stability for certain coupling ratios. We should choose a high τ_{max} value for the *1:1* ratio and rather low or middle τ_{max} values for the *2:5* and *1:10* ratios. We ran the oscillators with τ_{max} values of 0.05, 0.1, 0.2 and 0.5 in order to figure out the best configuration.

Table 1: RMS and autocorrelation values for Euclidean distances in Experiment 1.

τ_{max}	Syllable	"Foot"	"Phrase"
	RMS | CORR	RMS | CORR	RMS | CORR
0.05	33.67 | 0.87	34.14 | 0.13	28.22 | -0.34
0.1	38.56 | 0.86	47.65 | 0.88	28.02 | -0.33
0.2	44.58 | 0.73	29.62 | -0.43	27.24 | -0.41
0.5	24.49 | 0.36	67.30 | 0.99	42.20 | 0.50

For the evaluation we calculated Euclidean distances of tracked pulse times to the pulse times in the input sequence. The values describe the temporal deviation of individual events. Thus they represent the error which commits the oscillator. We calculated root mean square values (RMS) of the Euclidean distances for all τ_{max} values comparing them with the original sequence on all levels. We also calculated autocorrelation values for all Euclidean distances using Pearson coefficients. Successful entrainment should cause smaller distances. Thus, small or negative correlations indicate better accuracy. Close-to-zero correlations would occur in the case of a perfect match if $RMS = 0$ or in the case of a constant mismatch if $RMS \neq 0$.

3.1 Results

Table 1 shows the RMS and autocorrelation results for all levels and all τ_{max} values. The best accuracy on the syllable level, as expected, shows the oscillator with the largest τ_{max} value. For the "Foot" level the best accuracy shows the oscillator with $\tau = 0.2$; and for the "Phrase" level the oscillators with low and middle τ_{max} values.

3.2 Discussion

The results show the general capability of the oscillators to track rhythmic speech input on distinct levels. To assure best accuracy it is crucial to adopt the maximal receptive window size τ_{max} to the expected oscillator coupling ratio.

4. DPPOM IN MAX

For the second experiment we combined both sub-systems to the DPPOM. We used the same rap input as in Experiment 1. We also conserved the same configuration for the perception layer. For the production layer, we defined two gestural actions, head nodding and arm movement. The actions were timed to certain important events in the input sequence. Head nodding is a faster and smaller movement, thus we timed it with certain stressed syllables. Arm movement was timed to certain accentuated syllables.

Figure 2: *Conversational agent Max.*

The movements were performed by the conversational agent Max [2], see Fig. 2. The DPPOM values were translated to *MURML*, an XML representation language for speech and gesture generation in conversational agents [2].

Table 2: RMS, RMSD and autocorrelation values in Experiment 2.

	Head nodding	Arm movement
RMS (EUCL)	9.84	23.22
RMSD (EUCL)	9.54	18.67
CORR (EUCL)	-0.25	0.05
RMS (PER ORIG)	40.26	62.23
RMSD (PER ORIG)	15.83	33.57
RMS (PER SYS)	40.28	62.35
RMSD (PER SYS)	11.65	10.40

We used a global cycle coordination layer. The movement duration was kept stable for both movements in order to simplify the evaluation of DPPOM. The head nodding needed 1 second to be completed, the arm movement 2 seconds. The head oscillator was initiated with the period of 1500 ms, coupling ratio of approx. *1:10*. The arm oscillator was initiated with the period of 4000 ms, coupling ratio of approx. *1:20*. Both ratios are rather unstable. Thus we chose small τ_{max} values: 0.2 for the head oscillator and 0.05 for the arm oscillator. The activation (confidence) output of the oscillators is ranged between 0 and 1. We defined activation thresholds for both motor oscillators to be 0.5.

The given task of the motor system is not to track each occurring event of some kind but rather "pick out" few important events. Thus it is reasonable to compare only the RMS for these important events. In fact, we as modelers do not decide which events are important, rather we decide which *kind* of events is important. If the chosen events are well chosen or not we can see in the grade of realistic behavior of the conversational agent.

4.1 Results

Table 2 shows the statistical results of the second experiment. The first three lines show the RMS, RMSD and autocorrelation values for the Euclidean distance between the original pulses and the corresponding system answer. For head nodding the values are relatively low. The autocorrelation indicates improvement in accuracy throughout tracking activity. The values for the arm oscillator show higher accuracy errors. The autocorrelation value is very low, thus the error is produced consistently.

The lower four lines show the RMS and RMSD values for the intrinsic period development for the sequence of original pulses and the system answer, respectively. We see that the head oscillator was able to reproduce the mean period and also the period variance of the input sequence. The arm oscillator reproduced well the mean period, but failed in the variance. Thus the arm oscillator averaged the input.

4.2 Discussion

The system was able to create adequate response in order to produce feedback signals at special events in a rap input sequence. The head oscillator showed better accuracy than the arm oscillator. However, the arm movement takes longer time. Thus a tolerance range in time can be defined. To be able to make use of the tolerance range there is need of motor oscillators, which control the actual movements. Such oscillators would react to the deviation of the answer from the original signal and adopt their period (= the time in which a movement has to be completed), accelerating or decelerating the movement speed. The information of deviation can be encoded in the activation level.

5. CONCLUSIONS

In the present study we introduced a new model for coordination of motor tasks as a response to rhythmic input (DPPOM). We introduced two sub-systems of the model, the perceptual and the production sub-systems. We evaluated the perceptual sub-system on a rap input. In this evaluation we introduced a simple use of input amplitude. We used it to entrain the oscillators only when an activation threshold was reached. Afterwards we used a complete DPPOM system in order to coordinate gestural response to the rap input in the conversational agent Max. The results showed realistic timed reactions to the rap input.

We also have revealed important lines of future investigation. First, it is important to study a flexible integration of amplitude in the entrainment and maybe in the activation of the oscillators. The amplitude provides an important source of information about the *form* of the events received. Especially the phonetic source information has to be investigated in order to identify the most effective clues for important events in the input speech. Second, it is important to investigate the relationships between the perceptual and the production layers. It is also important to study the interaction between the global cycle coordination and the coordination of the individual movements in the motor planning. Machine learning techniques could be used to combine both lines of investigation using statistical training approaches.

6. ACKNOWLEDGEMENTS

This work is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center SFB 673.

7. REFERENCES

[1] J. Bonaiuto and K. Thórisson. Towards a neurocognitive model of turn taking in multimodal dialog. In M. L. I. Wachsmuth and G. Knoblich, editors, *Embodied communication in humans and machines*, pages 451–483, New York, 2008. Oxford University Press.

[2] S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Journal of Computer Animation and Virtual Worlds*, 15:39–52, 2004.

[3] E. W. Large. *Dynamic representation of musical structure*. The Ohio State University, Ohio, 1994.

[4] S. M. Marcus. Acoustic determinants of perceptual center (p-center) location. *Perception & Psychophysics*, 30(3):247–256, 1981.

[5] N. von Steinbüchel, M. Wittmann, and E. Pöppel. Timing in perceptual and motor tasks after disturbances of the brain. In M. Pastor and J. Artieda, editors, *Time, internal clocks and movement*, pages 281–304. Elsevier Science B.V., 1996.

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