LINGUISTICA URALICA XLVIII 2012 3 doi:10.3176/lu.2012.3.08

MICHAEL L. O'DELL (Tampere), *TOMMI NIEMINEN* (Joensuu),
MIETTA LENNES (Helsinki)

MODELING TURN-TAKING RHYTHMS WITH OSCILLATORS

Abstract. Our aim in this paper is to explore ways of modeling the distribution of pause durations in conversation using oscillator models (Wilson, Wilson 2006), and to consider how these models might be integrated into our Coupled Oscillator Model of speech timing (COM (O'Dell, Lennes, Werner, Nieminen 2007; O'Dell, Lennes, Nieminen 2008; O'Dell, Nieminen 2009)).

Keywords: Finnish, speaker synchronization, turn-taking, pause duration, Coupled Oscillator Model.

1. Overview

Modeling the durations of conversational pauses has recently attracted some attention (cf. the excellent overview in Heldner, Edlund 2010). M. Wilson and T. P. Wilson (2006) have modeled conversational turn-taking based on coupled oscillators, and Beňuš tested this model against a database of conversational American English (Beňuš 2009; Beňuš, Gravano, Hirschberg 2011). Beňuš's results provided some support for the model, but the support was weak due to small (although significant) correlations, and a lack of predicted phase patterns.

As M. Wilson and T. P. Wilson (2006) pointed out, it is important to gather data on a variety of languages in addition to English. In this paper, we apply Beňuš's analysis to the Finnish Dialogue Corpus (Lennes, Anttila 2002; Lennes 2009) and also consider integrating the Wilson & Wilson model into our own speech timing model, which has hitherto lacked an explicit mechanism for dealing with pausing behavior.

2. Wilson & Wilson model

2.1. Motivation for oscillators

There are several facts about turn-taking behavior in spoken dialogue which M. Wilson and T. P. Wilson (2006) explain using an oscillator model. According to M. Wilson and T. P. Wilson (2006), turn transitions with virtually no gap [< 200 ms] are a common occurrence in ordinary conversation. This is testified to in the Finnish corpus as well: slightly more than a third of the transitional pauses were less than 200 ms in duration (cf. Table 1). According to M. Wilson and T. P. Wilson (2006) and many others, conversational speech also tends to avoid simultaneous starts.[1]

[1] It is often assumed that overlapping speech is avoided in general, although this has been questioned along with the assumption that dialogues actually exhibit clear

Table 1

Number of transitional pauses for a pair of Finnish speakers (speaker 1, speaker 2)

	$1 \rightarrow 2$	$2 \rightarrow 1$	Both
Total	145	174	319
< 200 ms	55 (38%)	54 (31%)	109 (34%)

The reason for this is fairly obvious given that conversation has a real, dialogic function. Simultaneous starts after pause (defined as both speakers initiating speech in less than 200 ms of each other) are relatively rare in our Finnish corpus as well: Approximately 6% of pauses ended in simultaneous starts (cf. Table 2).

Table 2

Number of "simultaneous" starts after pause for a pair of Finnish speakers

	$1 \rightarrow x$	$2 \rightarrow x$	Both
Total	461	409	870
< 200 ms	31 (7%)	22 (5%)	53 (6%)

A fact that is not so obvious is that (according to Wilson, Wilson 2006) pauses tend to be multiples of some unit length of time, which ranged from 80 to 180 msec with an average of 120 msec. (Wilson, Wilson 2006, based on data in Wilson, Zimmerman 1986).[2] This raises the possibility that turn cycle might be related to some other oscillatory cycle in speech, and M. Wilson and T. P. Wilson (2006) suggest possible candidates such as syllable duration, jaw cycles or even the theta rhythm.

2.2. Synchronization and turn cycle

The idea behind synchronization in dialogue is that each participant monitors the speech of the other and tries to keep in synchrony. Arguably such behavior is either a by-product or a prerequisite of speech perception in general.

During silence, the ability of the speakers to synchronize is considerably weakened. M. Wilson and T. P. Wilson (2006) conjecture that the speakers maintain a turn cycle which is also synchronized during speech (possibly related to e.g. syllable rhythm) and then continued during pauses. Such behavior is hypothesized to minimize the offset between their conversational turns. Thus, when the current speaker reaches the end of his turn, the current listener may step in with a minimum overlap or gap (when no pause is intended). The participants' oscillators have the same period (when synchronized) but the listener's cycle is counterphased to that of the speaker (Wilson, Wilson 2006, cf. Fig. 1). Because of this counterphasing, the probability of simultaneous starts will be relatively low (Wilson, Wilson 2006).

Note that c o u n t e r p h a s e d describes the situation from the individual participant's point of view: each one oscillates between phases "my turn to start" and "your turn to start" and these phases are opposed. From a system point of view, however, the two oscillators are actually in phase: each one oscillates between phases "1st speaker's turn to start" and "2nd speaker's turn to start" and these phases are in agreement.

---

turn-taking structure at all. In the present work we are not directly concerned with overlapping speech, but we hope to return to this question in the future.

[2] M. Wilson and T. P. Wilson (2006) refer to this unit length of time, or turn cycle period, as S. Confusingly, the earlier article Wilson, Zimmerman 1986 refers to uses S to mean each speaker's slot length, which is half of a total turn cycle. Thus Wilson & Wilson's S equals twice Wilson & Zimmerman's S. In what follows we retain S for the slot length and use R for the period of the turn cycle, so that $R = 2S$.

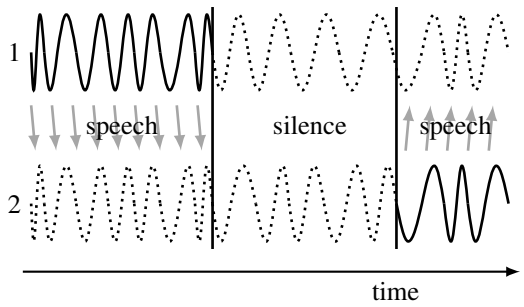

Figure 1. **Speakers synchronize during speech, during pause each speaker oscillates between "my turn to start" and "your turn to start".**

2.3. Empirical testing

Beňuš attempted to test the empirical consequences of the Wilson & Wilson model (Beňuš 2009; Beňuš, Gravano, Hirschberg 2011). If a putative turn cycle is a continuation of some rhythm accessible during speech, the question natually arises as to which of the many possible rhythms is the relevant one. Beňuš considered two possibilities in his analysis of a database of conversational American English: syllable rhythm and pitch accent rhythm.

For empirical testing purposes Beňuš compared two measures derived from the database: l a t e n c y, defined as difference between the end of the chunk [inter-pausal unit] and the beginning of the next chunk and r a t e, represented by average (syllable or accent) duration within each chunk (Beňuš 2009). These measures are illustrated in Fig. 2.

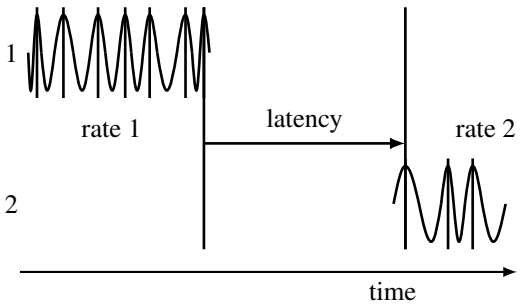

Figure 2. **Schematic diagram of Beňuš's measures.**

Beňuš also makes an overt terminological distinction between synchronization within speaker vs. between speakers: the first one he calls isochrony, the second one entrainment. In both cases, whether isochrony proper or entrainment, the following points hold: a) Rate should be correlated across pause, b) Latency should be correlated with previous rate and c) The latency distribution (normalized by previous rate) should be multimodal, with modes at interval steps.

Results provided some support for the model, but support was weak due to small (although significant) correlations, and a lack of predicted phase patterns.

3. Present study

We set out to apply Beňuš's procedure to Finnish conversational material following in effect Wilson & Wilson's plea for more material from diverse languages. Presently we have studied only one speaker pair, and the results are thus very preliminary, albeit suggestive.

T. P. Wilson and D. H. Zimmerman (1986) estimated S using time series analysis (ARIMA) applied to histograms reinterpreted as a time series. Here we model empir-

ical pause distributions as a mixture of normal distributions (one for each possible turn cycle), imposing various constraints on the means, variances and mixing probabilities. This procedure allows a series of increasingly complex models to be fit to data.

Models of pause duration distributions

Constant expected duration "no effects model" $E(\text{dur}) = \mu$
Cyclical expected duration "Wilson & Zimmerman model" $E(\text{dur}) = nR$ or $(n - 1/2)R$
Variable turn cycle "Wilson & Wilson model" $E(\text{dur}) = nR(t)$ or $(n - 1/2)R(t)$, $R(t)$ depends on previous speech
Multiple hierarchical cycles "COM model" $E(\text{dur}) = c_1 + c_2 n_2 + \cdots + c_k n_k$

A generic graph for these models is shown in Fig. 3. In this figure Z_i is the measured duration, μ_i is the expected duration and σ_i^2 is the duration variance for the ith pause. Expected duration is a function of ni , the number of silent turn cycles ($\mu_i = R1 + (n_i - 1)R$, where R is the period of one cycle, and R_1 is the duration of the first cycle). Two parameters, β_0 and β, are included to allow the variance σ_i^2 to increase slightly as n increases ($\ln \sigma_i^2 = \beta_0 + \beta_{ni}$). The probability of n turn cycles is modeled as a geometric distribution with probability p_0 of success.

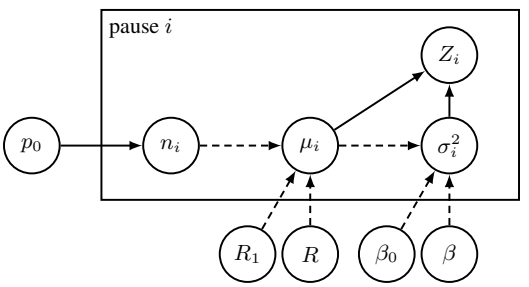

Figure 3. **Graph of statistical model.**

Bayesian inference of periodicity can be based on the ratio of total variance to within cycle variance for the first two cycles (say $\phi = \sigma^2_{\text{total}}/\sigma^2_{\text{within}}$). When this ratio is smaller than two the cyclic structure of the mixture distribution is not apparent, so we use the posterior probability $\Pr(\phi < 2)$ to indicate the significance of periodicity. An almost equivalent alternative which is easier to assess visually is to compare the cycle period (R) with the sum of standard deviations for the first two modes ($\sigma_1 + \sigma_2$, cf. Fig. 4): Periodicity can be considered significant when $R \gg \sigma_1 + \sigma_2$.

3.1. Cyclical expected duration

Beňuš did not look directly at the raw latency distributions in his data for signs of periodicity (and thus did not attempt to estimate S as Wilson and Zimmerman (1986) did), but normalized latency duration using syllable (or accent) rate of the previous chunk. Before proceeding to the Wilson & Wilson model, however, we start with the simpler Wilson & Zimmerman model to see whether a clear periodicity in the pause duration distribution can be discerned and whether it agrees

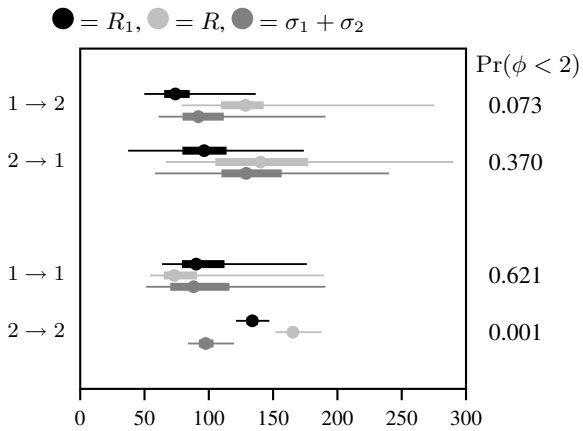

Figure 4. **Estimated parameters (ms) for the Cyclical model.**

with Wilson & Zimmerman's estimate of S with a range from 40 to 90 ms with a mean of 60.00 ms (Wilson, Zimmerman 1986).

Posterior distributions for R_1, R and $\sigma_1 + \sigma_2$ are shown in Fig. 4 for the four conditions: switches from speaker one to speaker two (1 → 2), switches from speaker two to one (2 → 1), speaker one internal pausing (1 → 1) and speaker two internal pausing (2 → 2). Raw distributions of pause durations in the four conditions are shown in Fig. 5. Also shown in this figure superimposed on the raw distributions are the median posterior fits for the mixture model.

The Wilson & Zimmerman model predicts that for between speaker pauses (which contain an even number of slot lengths S), pause duration will be $2kS$, k = 0, 1, 2, . . . , so that $R_1 \approx R$. On the other hand for within speaker pauses (which contain an odd number of slot lengths), the pause duration will be $(2k + 1)S$, so that $R_1 \approx R/2$.

In our data only the within speaker pauses for speaker 2 (2 → 2) showed a significant periodic structure (although condition 1 → 2 was also close to significance; see Figs. 5 and 4). The posterior mean for R for 2 → 2 was 165 ms with a 95 % credible interval of 152—187 ms, which agrees well with Wilson and Zimmerman's estimates, remembering that $R = 2S$. For the within speaker condition the Wilson & Zimmerman model predicts $R_1 \approx R/2$. As shown in Fig. 4, R_1 (posterior median 134 ms) is reliably less than R (posterior median 165 ms), but much greater than $R/2$. This could indicate that the first cycle is slower, or that the two halves of the turn cycle (say S' and S'', so that $R_1 = S'$, $R = S' + S''$) are not necessarily equal (with $S' > S''$ for speaker 2).

Another interesting feature for the 2 → 2 pauses is that there appears to be a second local maximum in the vicinity of 0.6 to 0.8 s (fourth and fifth bump, cf. Fig. 5). This might indicate the existence of two simultaneous rhythms during pause.

In general, what are the chances of this type of test succeeding? Assuming the turn cycle during pause is a continuation of the syllable cycle during speech, the distribution of durations during speech provides a compar- ison for judging whether quasiperiodicity could be de- tected even in an ideal case. To put this another way,if we were not sure that speech was composed of syllables, could this be deduced given only the total durations of various units (such as stress groups)? For the present data at least, applying the above statistical procedure to inter-pause groups indicated that periodicity due to recurring syllables during speech is entirely masked by the variability in syllable rate. If pauses are indeed composed of "silent syllables", and if silent syllable rate is as variable as normal syllable rate, then the same may hold for pauses, obscuring the cyclic nature of pausing due to cycle period variation. Of course this cannot be construed as evidence f o r periodicity during pauses, but lack of clear multimodality in the duration distributions does not provide strong evidence against it either. A possible way forward is to look for additional covariates which correlate with the variable turn cycle period.

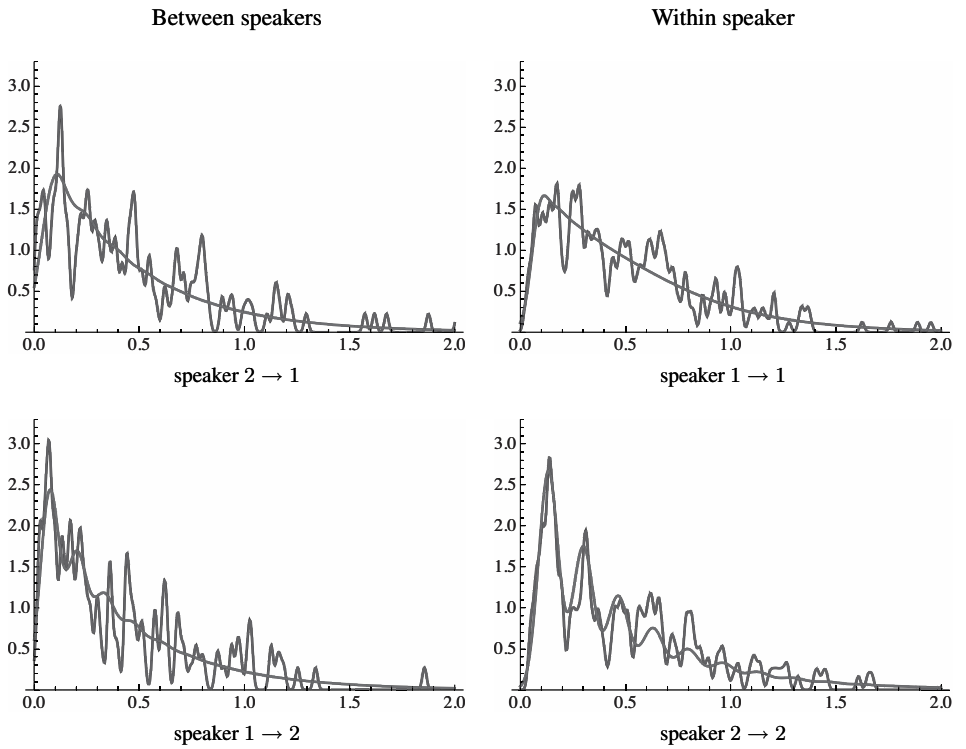

Figure 5. **Distributions for pause durations (s).**

3.2. Variable turn cycle

M. Wilson and T. P. Wilson (2006) hypothesized that turn cycle is a continuation of syllable cycle during speech. If this is the case, we would expect the turn cycle period to vary with syllable rate, rather than being constant (for each speaker or speaker pair). Following Beňuš's lead, we attempt to assess the possible relevance of syllable rate preceding a pause.

In the ideal situation, a scatterplot of pause duration against previous syllable rate would look something like Fig. 6: Pauses with an equal number of "silent syllables" (say k) form slanting stripes because slot length duration (S(t)) is tightly clustered around average syllable duration of the previous chunk. A stripe pattern of some kind should be evident even if syllable duration has a nonlinear relation to pause duration.

In such a case it is obvious that ignoring syllable rate will radically obscure the periodic pattern. On the other hand, dividing pause duration by the average

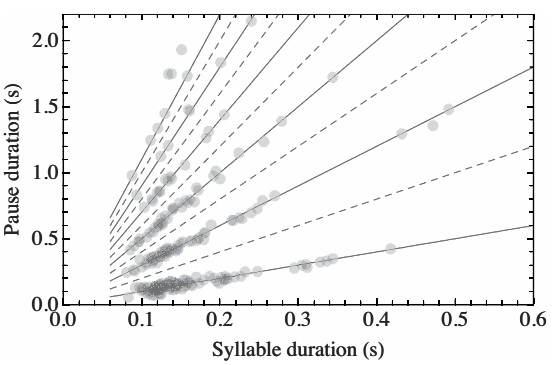

Figure 6. **Ideal scattergram of within speaker pauses.**

syllable duration (say $\hat{S}(t)$) similar to Beňuš's normalization procedure, gives an index ($I = 2kS(t)/\hat{S}(t) \approx 2k$ or $I = (2k + 1)S(t)/\hat{S}(t) \approx 2k + 1$) which should have an empirical distribution with clear modes at integer values (even for between speaker pauses, odd for within speaker pauses), given that $S(t) \approx \hat{S}(t)$.

For the present data, averaging syllable duration over the entire previous chunk as Beňuš did, produced the scatterplots shown in Fig. 7 for the four conditions. To aid the eye, in both Fig. 6 and Fig. 7 lines have been added indicating where pause duration equals an integer times syllable duration, solid for odd and dashed for even integers.

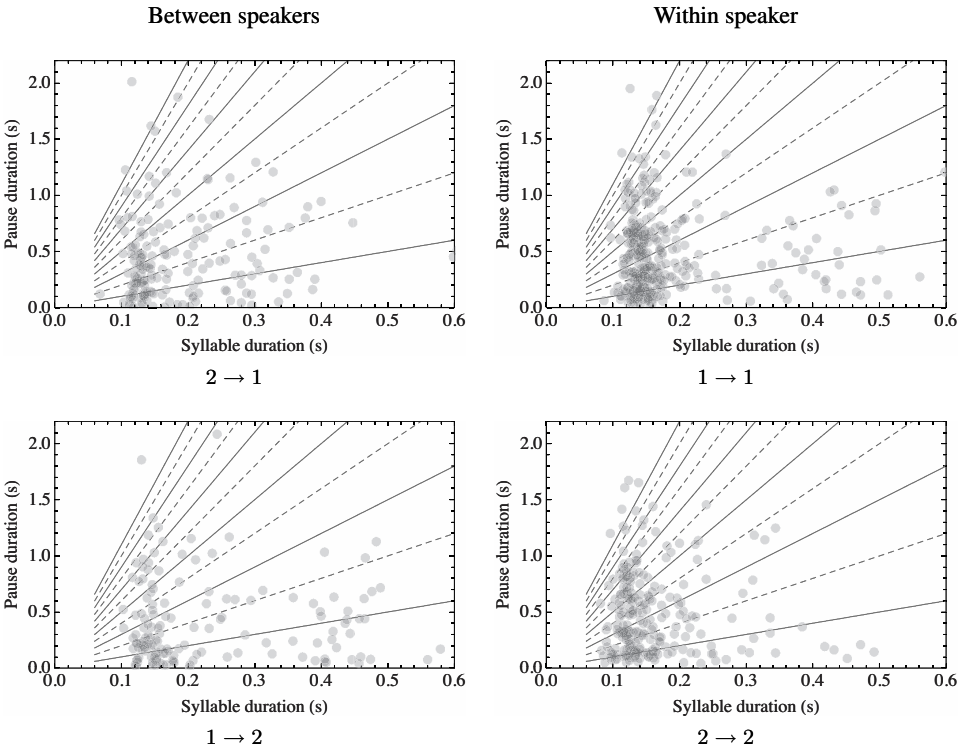

Figure 7. **Pause durations by average syllable duration of previous chunk**.

Fig. 8 shows distributions of pause durations normalized by syllable duration and rescaled to match the unnormalized distributions of Fig. 5 to facilitate comparison. Again, (vertical) lines have been added showing an integer number of syllable durations, solid for odd and dashed for even integers.

Evidence for a possible effect of syllable rate (estimated here by average syllable duration of the preceding chunk) on pause duration is completely lacking in these figures. The scattergrams have no stripes, the normalized duration distributions have no periodic structure. In fact, even the fairly clear periodic structure for the 2 → 2 pauses has been completely obscured in the nor- malized distribution. Looking at 2 → 2 in Fig. 7 we can see why: The periodic stripes are roughly parallel to the syllable duration axis instead of sloping as in the ideal case (Fig. 6). This suggests that the periodic structure of pauses for 2 → 2 is unrelated to the syllable rate of the preceding chunk.

There are various alternative explanations for the failure to observe a rate effect (apart from the conclusion that pausing is not rhythmic in nature). First of all, shortage of data. Thus far we have studied only one speaker pair, and the effect might be quite weak.

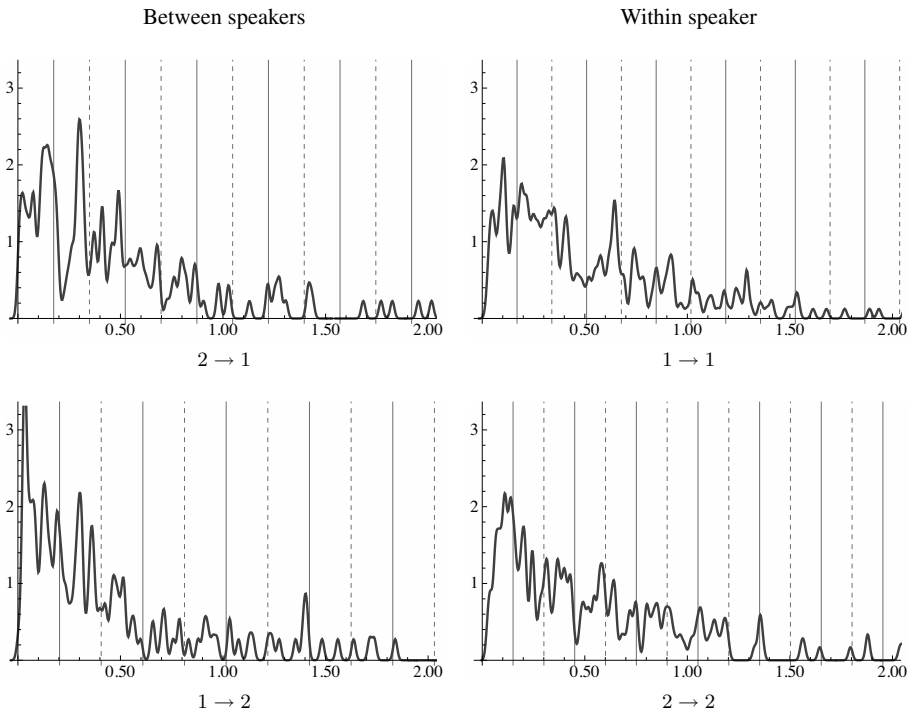

Figure 8. **Normalized pause duration distributions (s).**

Second, perhaps the syllable rate effect is too shortlived to be observed. Speakers may return to a neutral or preferred turn-taking cycle period fairly rapidly as a pause continues, or natural variation in the period may quickly obscure any initial rate related difference at the beginning of pause. It may also be that during pauses speakers maintain a turn-taking oscillator for a few cycles only. After all, as pause duration increases, the chance of a simultaneous start decreases even without any synchronizing mechanism. In either case the Wilson & Wilson model will be inadequate, because while allowing turn cycle to vary from pause to pause, it assumes a constant turn cycle during each pause.

A related issue is the adequacy of the rate estimate itself. It may be possible to obtain a better estimate of dynamic rate, for instance by weighting immediately preceding syllables more, rather than using a straight average over the entire previous chunk. In the future we plan to investigate more sophisticated techniques (such as Gaussian Process regression) for estimating various dynamically varying rates during speech and extrapolating those rates during pause.[3]

Finally, given the hierarchical nature of speech rhythm, some other rhythm might prove more relevant to the turn cycle than syllable rhythm. For instance Beňuš considered recurring accents (phrasal stress rhythm), as well as syllables. For Finnish mora rhythm is another candidate worth investigating.

3.3. Coupled Oscillator Model

The next step in our investigation will be to use the Coupled Oscillator Model (COM (O'Dell, Nieminen 2009)) to allow multiple, dynamically varying rhythms. This step is important also for our goal of incorporating pausing behavior into the COM.

[3] It would also be desirable to include (short) overlap durations as negative pauses in the distributions for turn transitions. This idea was also suggested by M. Heldner and J. Edlund (2010) for a noncyclic (no effects) model.

The Coupled Oscillator Model uses dynamic systems theory to derive a linear regression model for durations (T_1) of various units during speech given the number of synchronized subunits or cycles (n_i) at various levels:

$$T_1 = c_1 + c_2 n_2 + c_3 n_3 + \cdots + c_k n_k, \qquad (1)$$

For instance, our previous analyses of pause group durations in conversational (spontaneous) Finnish speech, allowing for five possible levels, have indicated strong mora and phrasal stress rhythm with possible weaker foot rhythm (O'Dell, Lennes, Werner, Nieminen 2007; O'Dell, Lennes, Nieminen 2008).

Extending the dynamic model to two speakers instead of one is relatively straight forward in principle, since the underlying theory does not require that all oscillators in the system belong to a single speaker. We have, in fact, previously applied the model for analyzing behavior in the so called synchronous speech task, where two speakers read a text out loud together at the same time (O'Dell, Nieminen, Mustanoja 2010; 2011).

A major challenge when modeling the synchronizing behavior of two speakers, however, is how to handle situations such as pauses in which information providing the basis for synchrony is temporarily diminished or absent. One possibility is to introduce stochastic coupling, theidea being that the synchronizing signal between oscillators (and participants) varies as to its reliability, rather than being modeled as exact. The beginning of silence can be taken to be a strong cue as to the phase of the other participant (explaining why subjects typically pause relatively often in the synchronous speech task), but phase uncertainty grows as silence continues.[4]

Such a characterization leads naturally to a distribution of pause durations with expected value corresponding to the equation (1) above. Following Beňuš we might hypothesize, for instance, that each pause contains an integral number of silent stress groups as a continuation of the stress group rhythm of the preceding speech (perhaps with a fixed, preferred number of silent syllables per stress group). Since several levels of rhythm are mutually synchronized in the COM, stress group frequency at the beginning of pause should be estimated not merely on the bases of previous stress groups (whether using a raw average or some other technique), but also taking into account all the relevant interacting rhythms on various hierarchical levels such as mora, syllable, etc.

4. Summary

We have begun exploring ways of modeling pause durations in Finnish conversations. Thus far, we have analyzed only one speaker pair but we have developed a general statistical model for testing increasingly complex effects in the gathering material.

The simplest versions of the model do not fit the data (much) better than the "no effects model", but this may yet change as we look at additional speaker pairs and more sophisticated models.

Addresses

Michael O'Dell
University of Tampere
E-mail: michael.odell@uta.fi

Tommi Nieminen
University of Eastern Finland
tommi.nieminen@uef.fi

Mietta Lennes
University of Helsinki
mietta.lennes@helsinki.fi

[4] An interesting finding from our analysis of the synchronous speech task, which may be relevant in the present case, is that while speakers were less synchronized after pause than before, asynchrony did not increase with pause durations greater than approximately 200 ms. This could be taken as further evidence of a silent rhythm during pause.

REFERENCES

Beňuš, Š. 2009, Are We 'in sync'. Turn-Taking in Collaborative Dialogues. — Proceedings of the 10th Interspeech, Brighton, 2167—2170.

Beňuš, Š., Gravano, A., Hirschberg, J. 2011, Pragmatic Aspects of Temporal Accomodation in Turn-Taking. — Journal of Pragmatics 43, 3001—3027.

Heldner, M., Edlund, J. 2010, Pauses, Gaps and Overlaps in Conversations. — Journal of Phonetics 38, 555—568.

Lennes, M. 2009, Segmental Features in Spontaneous and Read-Aloud Finnish. — Phonetics of Russian and Finnish, Frankfurt am Main—Berlin—Bern—Bruxelles—New York—Oxford—Wien, 145—166.

Lennes, M., Anttila, H. 2002, Prosodic Features Associated with the Distribution of Turns in Finnish Informal Dialogues. — Fonetiikan Päivät 2002. The Phonetics Symposium 2002, Espoo, 149—158.

O'Dell, M., Lennes, M., Nieminen, T. 2008, Hierarchical Levels of Rhythm in Conversational Speech. — Speech Prosody 2008. Fourth Conference on Speech Prosody, Campinas, 355—358.

O'Dell, M., Lennes, M., Werner, S., Nieminen, T. 2007, Looking for Rhythms in Conversational Speech. — Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, 1201—1204.

O'Dell, M., Nieminen, T. 2009, Coupled Oscillator Model for Speech Timing Overview and Examples. — Nordic Prosody. Proceedings of the Xth Conference, Helsinki 2008, Frankfurt am Main—Berlin—Bern—Bruxelles—New York—Oxford—Wien. 179—190.

O'Dell, M., Nieminen, T., Mustanoja, L. 2010, Assessing Rhythmic Differences with Synchronous Speech. — Speech Prosody 2010 Conference Proceedings 100141, 1—4.

—— 2011, The Effect of Synchronous Reading on Speech Rhythm [presentation given at Rhythm Perception & Production Workshop 13, Leipzig].

Wilson, M., Wilson, T. P. 2006, An Oscillator Model of the Timing of Turn-Taking. — Psychonomic Bulletin & Review 12, 957—968.

Wilson, T. P., Zimmerman, D. H. 1986, The Structure of Silence between Turns in Two-Party Conversation. — Discourse Processes 9, 375—390.

МАЙКЛ Л. О'ДЕЛЛ (Тампере), *ТОММИ НИЕМИНЕН* (Йоэнсуу), *МИЕТТА ЛЭННЕС* (Хельсинки)

МОДЕЛИРОВАНИЕ РИТМА СМЕНЫ ГОВОРЯЩЕГО С ПОМОЩЬЮ ОСЦИЛЛЯТОРОВ

В статье рассматриваются способы моделирования распределения длительности пауз в разговоре с использованием осциллятора модели (Wilson, Wilson 2006), а также возможности интегрирования этих моделей в наши модели синхронизаций речи, базирующихся на модели соединенных осцилляторов (Coupled Oscillator Model (O'Dell, Lennes, Werner, Nieminen 2007; O'Dell, Lennes, Nieminen 2008; O'Dell, Nieminen 2009)).