HAZARD REGRESSION FOR MODELING CONVERSATIONAL SILENCE

Michael L. O'Dell[1], Tommi Nieminen[2], Mietta Lennes[3]

[1]University of Tampere, [2]University of Eastern Finland, [3]University of Helsinki
michael.odell@uta.fi, tommi.nieminen@uef.fi, mietta.lennes@helsinki.fi

ABSTRACT

It is often assumed that the participants of a conversation try to avoid simultaneous starts or lengthy silences. For this reason, they may tend to synchronize rhythmically with each other's speech. A model of conversational turn-taking based on the idea of coupled oscillators has been suggested by Wilson & Wilson [1]. However, the model has received only weak empirical support from previous studies where distributions of silence durations have been modeled directly. In the present study, we attempt to detect signs of oscillatory behavior during silence utilizing nonparametric hazard regression. In order to understand the shape of the estimated hazard rates, we postulate a latent stochastic process [2] with end of silence occurring when the process crosses a threshold. This finer-grained approach using Bayesian estimation yields a more detailed picture of synchronization between speakers and a more powerful test of oscillatory behavior.

Keywords: pause, gap, hazard rate, stochastic process, Finnish

1. INTRODUCTION

Wilson & Wilson [1] (hereafter W&W) suggested a turn-taking model based on coupled oscillators. Beňuš [3] tested several predictions of this model against a database of conversational American English, and O'Dell *et al.* [4] used a similar analysis for a Finnish database. In both cases results provided some support for the model, but support was weak due to small correlations, and a mismatch of latencies predicted by W&W.

It is often assumed that speakers (a) avoid starting to speak at the same time, and (b) avoid lengthy silence. These goals are somewhat contradictory, since lengthy silence in itself would diminish the risk of simultaneous starts. Indeed, this is part of the motivation for the W&W model: they postulate that each speaker maintains the syllable oscillations of speech during a following silence in order to stay synchronized and thereby avoid simultaneous starts even without lengthy periods of silence.

But just how useful is oscillatory behavior during silence? During short periods of silence synchronized oscillation could help to avoid simultaneous starts. As silence continues, however, it is obvious that between speaker synchrony will deteriorate. At the same time, it will also be less needed, since the risk of simultaneous starts will diminish in any case as time goes on. Thus, along with W&W, we might expect a trade-off between short silence with oscillations vs. longer asynchronous *lapses*. W&W suggest "Exactly how long it takes for the cycling [...] to break down is unclear, but it is probably a matter of a few seconds." [1].

The present research uses a different statistical technique, Bayesian nonparametric hazard regression, to elucidate these issues. Do syllable rhythms continue during silence? If so, for how long and for how many cycles? How does speaker behavior differ depending on which speaker initiated silence? (In what follows we adhere to the terminology of [5] for the sake of brevity: *pause* refers to within-speaker silence and *gap* refers to between-speaker silence.)

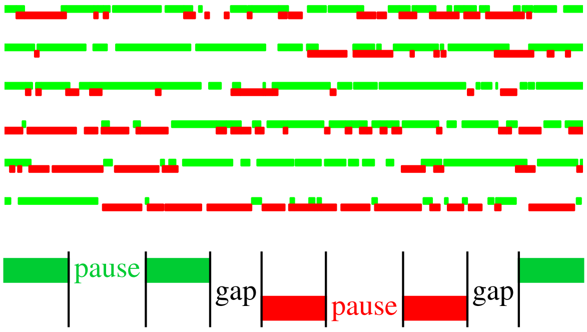
2. CORPUS AND METHODS

2.1. Finnish Dialogue Corpus

The Finnish Dialogue Corpus [6] consists of informal unscripted dialogues with pairs of young Finnish adults recorded in an anechoic room. The participants in each dialogue were close friends and they were allowed to chat freely and unmonitored for a total of 40 to 60 minutes on either given or self-selected topics. Speakers sat a few meters apart facing opposite directions. Each speaker's speech was recorded to a separate channel of a DAT recorder using high-quality headset microphones. The recorded material was then transferred to a computer and sampled at 22050 Hz. The two channels of the stereo files were separated, resulting in one audio file per speaker.

Four speaker pairs were analyzed for the present study. Each speaker's utterances were orthographically transcribed and silence (including short hesitations), words, syllables and morae were annotated for each speaker using Praat [7].

Figure 1: Turn chart for part of speaker pair F6/F7 dialogue, and schematic showing types of silence.



2.2. Hazard regression for silence

The hazard function $\lambda(t)$ in survival analysis indicates the risk of an event (end of silence in the present case) occurring at time t, given that the event has not occurred earlier. This differs from the more familiar probability density function, which is not conditional on (lack of) previous occurrence.

Estimating the hazard rate from empirical duration data provides an alternative to empirical density estimation which is more sensitive to potential oscillatory behavior. Hazard functions can be estimated in several ways. Here we use the Linear Dependent Dirichlet Process Mixture of Survival models [8], which is a form of Bayesian nonparametric hazard modeling allowing the inclusion of covariates, such as syllable rate for the speech preceding silence. All hazard functions and statistics were computed with the DPpackage function *LDDPsurvival* [9] in **R** [10].

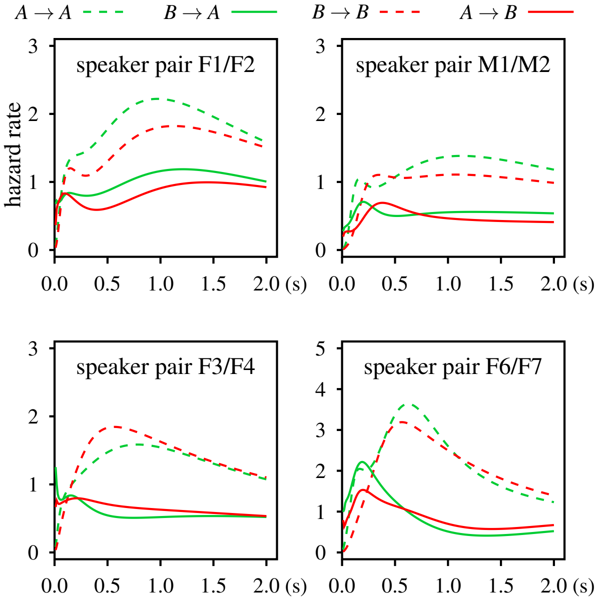
In estimating the risk of a speaker starting to speak it is important to take into account the cases when the other speaker starts instead. For instance, if speaker B starts speaking after silence, we don't know how long speaker A would have continued to wait before initiating speech, but the observed silence does provide a lower bound. This is called *right censoring* in survival analysis.

3. RESULTS

3.1. Hazard functions without covariates

Hazard rates estimated without covariates are plotted in Fig. 2. Hazard rates for pause (dashed in Fig. 2) start at zero, while hazard rates for gap (solid) are positive even at time zero, meaning there is some chance of (near) simultaneous turn switching. In fact, in the between speaker condition, a "negative gap" or overlapping speech is a possibility (cf. [5]).

Figure 2: Hazard functions without covariates (posterior pointwise mean; note different scale for pair F6/F7).



Pause risk is thus initially smaller than gap risk, but as silence continues the situation reverses and pause risk becomes about twice as large as gap risk. It is worth noting that for speakers in the same situation (following a given speaker: dashed line paired with opposite color solid line) there is always only one crossover point (at least within the 2 s period examined here), ranging from about 60 ms to about 290 ms. Given the oscillatory hypothesis we would expect the hazard functions to exhibit a periodic pattern leading to several crossover points.

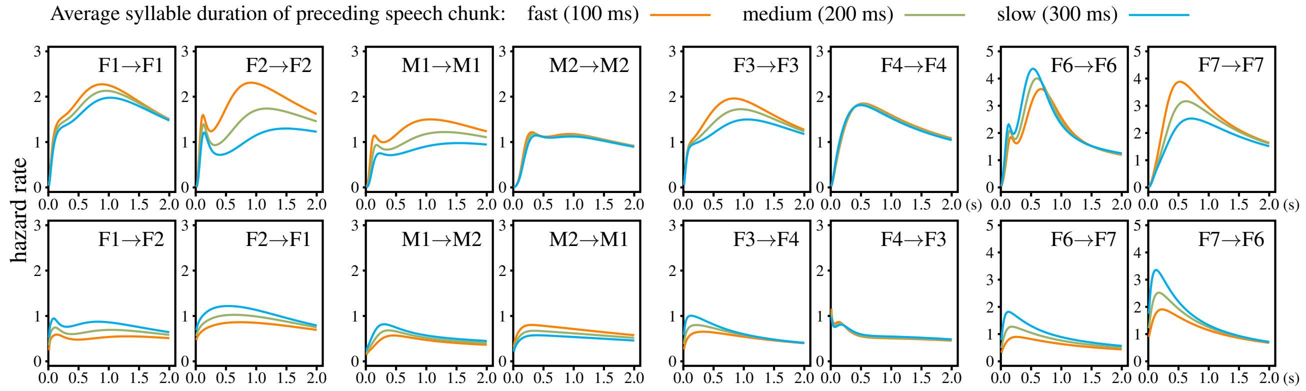
The hazard functions do, however show some evidence of alternating structure. All curves exhibit a general rising-falling pattern and many also have an additional early local maximum located well before 500 ms.

3.2. Hazard functions with syllable rate as covariate

Since W&W postulate that syllable rhythm continues into silence, it is of interest to investigate the possible influence of observable syllable rate on the hazard function of following silence. It is also conceiveable that periodicity has been obscured in the hazard functions for Fig. 2 by pooling cases with differing syllable rates. The results of hazard regression with syllable rate as a covariate are shown in Fig. 3. Following [3, 4], syllable rate was obtained by dividing the duration of the previous chunk of speech by the number syllables it contains.

Speakers do appear to be sensitive to the syllable

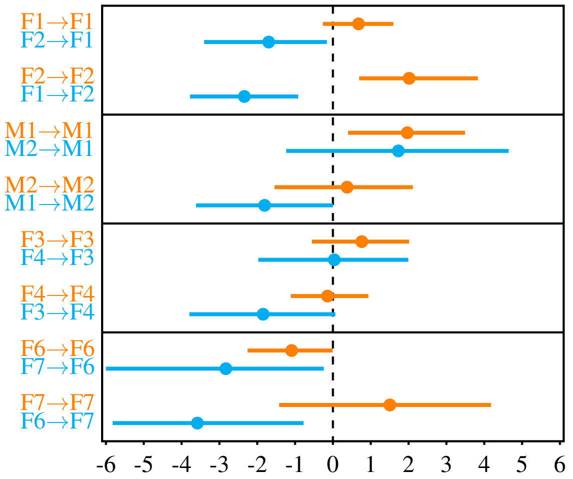
Figure 3: Hazard functions (posterior pointwise mean) with syllable rate of preceding speech as covariate.



rate of previous speech in both conditions. Faster rate generally *increases* risk of starting to speak again (pause, top row in Fig. 3). Faster rate *decreases* risk of starting to speak after gap (bottom row in Fig. 3). These general effects are minimal or even slightly reversed after speakers M2 and F4 (and F6 for pause), a fact which may be due to individual speaker differences, or merely a consequence of the imprecise estimate of syllable rate used. Significance of the effects can be approximately assessed in Fig. 4, which shows the posterior 95 % highest density intervals (HDI) for overall hazard level by syllable rate.

Taking syllable rate into account did not reveal more periodicity. In fact, rather than affecting the hazard time scale as predicted by the hypothesis of syllable rhythm continuing into silence, syllable rate appears to influence the overall hazard level. Moreover the effect is generally in opposite directions for pause vs. gap, which can be taken to mean that it is more an indication of willingness to give up the turn: faster speech means the speaker is less likely to give up the turn, while the other speaker, sensitive to this aspect, is more willing to wait.

Figure 4: Estimated effect of syllable rate (posterior mean & 95 % HDI)



4. DISCUSSION

4.1. Parametric models

While *nonparametric* modeling provides a general view with few assumptions, *parametric* modeling may help to understand the underlying processes. Several types of parametric model are available (differing in the form of the hazard function and how it is modified by covariates), two of the most popular being *accelerated life* models and *proportional hazards* models. Accelerated life (or “time stretching”) models cannot be ruled out for pauses (ie. faster speech simply makes pause time go more quickly), but are not plausible for gaps. Proportional hazards (or Cox) models are models in which covariates multiply overall hazard level rather than time course. This type of model appears to be more plausible.

4.2. Stochastic process model for silence

One technique for understanding the shape of a hazard function is to postulate an underlying stochastic process which triggers the observed event in question when the process crosses a threshold [2]. The time until a threshold is crossed by a stochastic process is known as the *first passage time* (or sometimes first hitting time). In the present context this can be equated with silence duration. Interpretation in terms of an underlying process also potentially facilitates the interpretation of covariates, which rather than influencing the hazard of an event directly, can be thought of as influencing the underlying process.

Base hazard rates for our pause and gap data have the general character of rising then falling. This is to say that at the beginning of silence the tendency to resume speaking first grows, but later gradually falls so that the longer silence continues the less likely it is to end.

As discussed in [2], this is a fairly general situation for first passage times in a wide range of stochastic processes: starting points relatively close to the threshold generate a decreasing hazard, relatively distant starting points generate an increasing hazard, while intermediate distances lead to a rising falling hazard function.

While several theoretical stochastic processes can lead to a hazard rate with the requisite property, given an appropriate distribution of starting points, the standard Wiener process (continuous time analogue of a random walk; also known as Brownian motion) may be considered a canonical case.

In the present setting, the Wiener process interpretation implies a continuously fluctuating variable representing inclination to start speaking. A speaker initiates speech when this inclination reaches a threshold level for the first time (first passage time or first hitting time). Including possible drift towards the threshold, the Wiener process can be characterized as

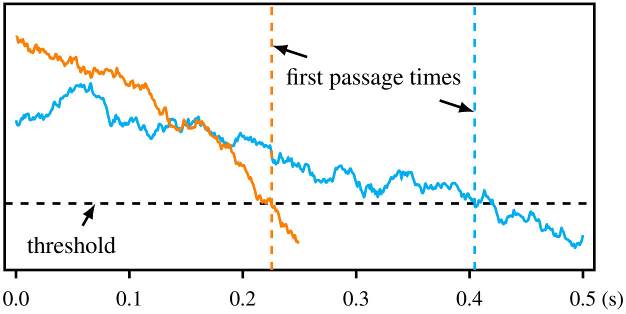
$$(1)\quad S(t) = c - \mu t + W(t)$$

where $S(t)$ is the fluctuation, $c = S(0)$ the starting point (level at time zero), μ the drift and $W(t)$ a standard Wiener process. Two examples of such processes are shown in Fig. 5, one starting closer to threshold with smaller drift, the other starting farther from threshold but with greater drift.

Several modifications of the basic Wiener process are possible to account for bimodality in the hazard function [11]. All of these modifications represent a combination of two unimodal processes.

Bimodal heterogeneity (mixture of two inverse Gaussian distributions, cf. e.g. [12]) assumes that each silence belongs to one of two different types. If the two types could be identified, each would exhibit a unimodal hazard function. Perhaps some of the silences were initiated as genuine turn-ends (cf. the notion of *transition-relevance place* [13]), whereas other cases were not intended as giving up the floor. It is reasonable to expect these two types to be characterised by stochastic processes with different parameters (especially for pauses), and this could lead to bimodal hazard functions.

Figure 5: Examples of Wiener processes



A *two stage process* starts in one state, but after reaching an intermediate threshold jumps to a second state with different values for some parameters.

Two simultaneous processes (with different properties) during pause with speech starting when either one of the processes reaches threshold. In the present case this could be interpreted as a single fast 'turn-taking' or wait cycle accompanied by a slower background process (for instance thinking of something new to say). The relative importance of the two components would apparently vary for different speakers (and/or situations).

5. CONCLUSIONS

Hazard function modeling promises to be a powerful tool for understanding pauses, gaps and rhythmic tendencies and deserves to be developed further. Looking at the estimated hazard functions for conversational silence categorized in various ways allows us to answer some of our questions with more confidence.

Several speakers exhibit fluctuating hazard functions, for pauses as well as gaps, and this may be taken as evidence for a silent turn-cycle of the kind postulated by W&W. However, even though this turn-cycle is approximately syllable sized, it cannot be easily characterised as a continuation of the syllable rhythm of the preceding speech. Preceding syllable rate does have an effect, but it appears to affect the intensity of the waiting process rather than determining cycle duration, and generally in opposite directions for pauses and gaps, possibly indicating that effects are less related to continuity of preceding rhythms and more related to anticipating turns.

If fluctuating hazard functions are taken as evidence for a silent turn-cycle, this oscillatory behavior during silence would appear to be very short lived. Speakers appear to take at most one such cycle rather than continuing for several seconds as conjectured by W&W.

In future research more covariates (finer classification of surrounding speech) need to be collected, e.g. whether silence interrupts a word or phrase. Also more attention needs to be paid to speech/silence transitions (voiced or voiceless frication, audible clicks) as well as sounds during pauses and gaps (such as laughter or breathing).

6. REFERENCES

[1] Wilson, M., Wilson, T. P. 2006. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review* 12, 957–968.

[2] Aalen, O. O., Gjessing, H. K. 2001. Understanding the shape of the hazard rate: A process point of view. *Statistical Science* 16 (1), 1–22.

[3] Beňuš, Š. 2009. Are we 'in sync': Turn-taking in collaborative dialogues. *Proc. 10th INTERSPEECH* PLACE!. 2167–2170.

[4] O'Dell, M., Nieminen, T., Lennes, M. 2012. Modeling turn-taking rhythms with oscillators. *Linguistica Uralica* 48 (3), 218–227.

[5] Heldner, M., Edlund, J., 2010. Pauses, gaps and overlaps in conversations. *J. Phon.* 38, 555–568.

[6] Lennes, M. 2009. Segmental features in spontaneous and read-aloud Finnish. In: de Silva, V., Ullakonoja, R., (eds), *Phonetics of Russian and Finnish*. Peter Lang 145–166.

[7] Boersma, P., Weenink, D. Praat: doing phonetics by computer. http://www.praat.org/

[8] De Iorio, M., Johnson, W. O., Müller, P., Rosner, G. L. 2009. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* 65 (3), 762–771.

[9] A. Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., Rosner, G. L. 2011. DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software* 40 (5).

[10] R Core Team. R: A Language and Environment for Statistical Computing. http://www.R-project.org

[11] Denrell, J., Shapira, Z. 2009. Performance sampling and bimodal duration dependence. *Journal of Mathematical Sociology* 33 (1), 38–63.

[12] Smith, C. E., Lánský, P. 1994. A reliability application of a mixture of inverse Gaussian distributions. *Applied Stochastic Models and Data Analysis* 10 (1), 61–69.

[13] Sacks, H., Schegloff, E. A., Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50 (4), 696–735.