PROSODIC PREDICTORS OF TEMPORAL STRUCTURE IN CONVERSATIONAL TURN-TAKING

Kathryn Franich[1,2]

[1]Harvard University and [2]University of Delaware
kfranich@fas.harvard.edu

ABSTRACT

Fluid conversation depends on conversation partners' ability to make predictions about one another's speech in order to forecast turn ends and prepare upcoming turns. One model used to explain this process of temporal prediction is the coupled oscillator model of turn-taking [1]. A generalization that the model captures is the relative scarcity of interruption in turn-taking, as it predicts partners' turns should be counter-phased to one another, with minimal pause time between turns. However, in naturalistic conversation, turns are often delayed, rather than occurring in perfect succession. We hypothesize that these delays are not of arbitrary duration, but are structured in their timing, just as between turns with immediate transitions. We demonstrate that relative timing of prosodic events occurring at *turn ends* is key to modelling pause duration between turns, providing evidence that inter-turn pauses exist in a temporal trading relation with the final syllable and prosodic word of immediately preceding turn.

Keywords: Turn-taking, prosody, conversation, speech timing.

1. INTRODUCTION

Research on timing in conversational turn-taking supports the idea that transitions between turns generally take place so rapidly that conversation partners must be able to make *predictions* about turn ends in order to time their own turns appropriately. For example, Levinson & Torreira [2] point out that the average time to prepare and execute an utterance is around 600 ms, but the average pause time between speaker turns is much shorter, closer to 200 ms. The authors take this to mean that speakers must be projecting the end of their partner's turn well before it actually ends. Research has shown that several cues are predictors of turn ends, including falling or rising intonation patterns, segment or syllable lengthening, lower intensity, lexical cues, syntactic structure, and utterance completion, among others [3]-[9]; however, the specific way in which conversation partners time their turns in the context of these cues remains unclear. One influential model that has been proposed to capture conversational patterns is the coupled oscillator model of turn-taking developed by Wilson & Wilson [1]. The model proposes that turns are timed to cycles of *readiness* based on speakers' syllable rate. Readiness is at its highest towards the end of the syllable and at its lowest in the middle of a syllable. Counter-phasing of listener-speaker syllable oscillations can be used to explain the fact that speakers rarely interrupt one another, rather waiting until their conversation partner is finished speaking to initiate their own speech. The fact that pause durations between turns tend to be relatively short—on the order of 100-300 ms—can be explained by a high frequency syllable oscillator which governs turn-taking.

1.1. Rhythmicity, turn latency, and turn ends

Despite a trend toward immediate turn transitions, turn latencies often exceed the duration of a single syllable, sometimes by quite a lot. Given that speakers need to be able to make turn-taking predictions even when their conversation partner delays a response, we might hypothesize that delayed turns still display a temporal structure that is consistent with the coupled oscillator model. One possibility is that longer turn latencies constitute several cycles of the syllable-level oscillator proposed by Wilson & Wilson. However, in languages like English where there is relatively high variability in the timing of syllables [10]-[12], the timing of syllables alone may not allow for precise predictions. Prior research on speech timing in turn-taking has found that variability in timing between stressed syllables is reduced at turn transitions between conversation partners, consistent with the idea of a greater degree of foot-based isochrony at turn transitions [13]. Greater isochrony at a turn end may help the listener to develop more fine-grained predictions about speech timing of their conversation partner, and hence to plan their own speech in a more consistent way. Additional research by Shattuck-Hufnagel and Turk [14] has identified the phrase-final syllable and the main stress syllable of the final prosodic word of a phrase as sites for final lengthening, indicating that this process—which also occurs at turn ends—is not limited to a single prosodic event. Thus, it may be that larger turn-final

constituents, such as the final prosodic word or foot, would be more stably timed events around which to plan turns. In this paper, we explore how well the timing of various turn-final prosodic structures can predict latency of a conversation partner's turn, with the goal of modelling in greater detail the temporal planning process of turn-taking.

2. METHOD

Five pairs (10 subjects total) of English speakers from the northeast and mid-Atlantic regions of the United States participated in a game of "Twenty Questions" in which participants took turns thinking of a person who was mutually known to the pair and having their partner ask yes/no questions until they could identify the person. In order to ensure that participants would have a sufficient number of mutual acquaintances to engage in the task, pairs of participants who already knew one another well were recruited. Pairs included three married couples (one male and one female each), one pair of female friends, and one pair of sisters (avg. age = 45 years). Participants within pairs had all known one another for at least five years.

Participants wore head-mounted microphones and were video and audio recorded in a room in the Phonetics and Phonology lab at the University of Delaware playing the game while sitting in chairs and facing one another. Partners took turns thinking of a person and guessing. Each pair played the game for approximately 15 minutes, during which time an average of 80 turns, or 'inter-pausal units' [3] (40 per person) were elicited. An example exchange is in (1).

(1) Partner 1: *Does this person live in Bóston?*
Partner 2: *Yes.*
Partner 1: *Is it a fámily member?*
Partner 2: *No.*

2.1. Data coding

Questions were coded in Praat [15] on a series of TextGrid tiers with intervals marked for phrase duration, duration between the onset of the final pitch-accented word (marked with acute accents in (1)) and the end of the phrase, duration of the final word, duration of the final foot, and duration of the final syllable, as well as duration of the pause between the end of the final word of the question and the beginning of the 'yes' or 'no' response. Response to the question was also coded, though data for turns of both responses is pooled in the present work due to sample size. In order to control for response structure as much as possible, data were trimmed so that only turns with responses of 'yes' or 'no' were included. Turns that exceeded 2 standard deviations in duration from the mean were excluded from the analysis, as many of these responses involved clear uncertainty on the part of the partner answering. Exchanges containing a disfluency on the part of either speaker were also excluded. The final dataset contained 370 turns.

3. RESULTS

A linear mixed effects model was used to evaluate how well inter-turn pause duration was predicted by the various prosodic variables, including 1) duration of the preceding inter-pausal unit (the question, which was usually a single phrase, e.g. "Is he a family member?"), 2) time from final pitch accent of the question to the end of the phrase, 3) duration of the phrase-final foot, and 4) duration of the phrase-final syllable. Two-way interactions between all variables and speech rate (syllables per second) were also included, and number of syllables in the phrase was also included as a fixed effect. A by-subject random slope for speech rate was included in the model. All continuous predictors were mean-centered.

As expected based on previous work [16] there was a significant negative relationship between pause duration and speech rate (β=-100.21, t=-4.32; p<0.001) and a positive relationship between pause duration and number of syllables in the phrase (β=-75.378, t=3.35; p<0.01). Among the prosodic variables, only the duration of the question-final word predicted inter-turn pause duration to a significant degree. Despite the overall trend toward shorter pauses at higher speech rates, there was a *negative* relationship between final word duration and inter-turn pause duration (β=-35.73; t=-2.10; p<0.05). In other words, as the duration of the final word increased, pause duration decreased (Fig. 1).

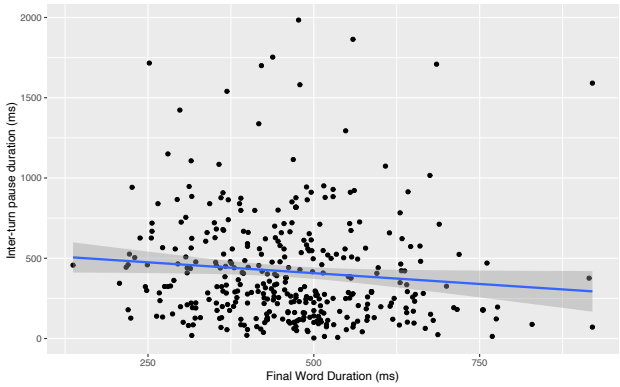

Figure 1: Duration of question-final word negatively predicts inter-turn pause duration

This result is consistent with a complementary (or 'trading') relationship between the final word and inter-turn pause, and may suggest that pauses constitute part of a larger planning unit which also includes the phrase-final prosodic word. Such a

pattern is reminiscent of findings on within-subject pause durations between sentences [18,19], where pause duration has been found to be in a complementary relationship with duration of phrase-final feet. Specifically, Fant & Kruckenberg report a multimodal distribution in pause durations in Swedish, such that the duration of a pause plus final lengthening equate to integer multiples of subjects' average foot (or inter-stress interval) durations.

Looking more closely at by-subject distributions of inter-turn pause latencies in our data (Fig. 2), many participants also displayed more than a single peak in latency—indeed, a Shapiro-Wilk normality test revealed that the data were not normally distributed (p<.001). Individual subject means for pause duration were also highly variable, ranging from 190 ms to 550 ms. Pairs of partners tended to have similar means—both subjects in Pair 3 (P3), in particular, had shorter latencies than the rest of the participants—though partners did not always pattern together, as demonstrated, for example, by P1.

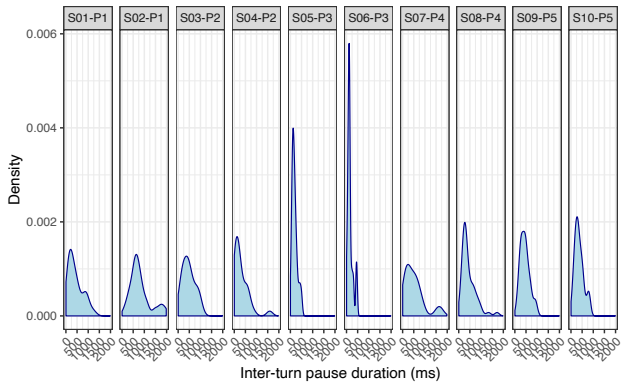

Figure 2: By-subject density plots of pause duration

Given these patterns of variation in turn latency and the potential for a complementary relationship between inter-turn pause (ITP) duration and question-final word duration, we hypothesized, similar to Fant & Kruckenberg, that inter-turn intervals might be more efficiently modelled as a *proportion* of the duration of the question-final word and the inter-turn pause interval added together (henceforth 'word+pause'). An equation deriving this measure (labelled Prop_Word) is provided in (2).

(2) Prop_Word = ITP Dur / (Final Word Dur + ITP Dur)

Specifically, we hypothesized that the duration of inter-turn pause intervals would exist in a quantal relationship to the word+pause duration, such that pause durations would cluster around lower order fractions, e.g. 1/3 or 1/2, of the word+pause interval.

For comparison, we calculated similar values for question-final syllable+pause ('Prop_Syllable'), question-final foot+pause ('Prop_Foot'), and final pitch accent to question end+pause ('Prop_PA'). Given the multimodal nature of the data, we fit the data to mixtures of Gaussian distributions with different number of components, using a parametric bootstrap of log-likelihood ratio statistics to evaluate the optimal number of components for the data for each measure. For all variables, bootstrap results revealed 2 components as the optimal number, indicating that pause proportions overall showed a bimodal distribution in the data.

To better understand the role of different prosodic structures in conditioning inter-turn pause durations, we fit the data for each prosodic pause proportion measure to three different 2-component models. Following a procedure from [20], in the first model (Mod1), means, mixing proportions, and standard deviations were chosen automatically in order to maximize model fit to the data. In the second model (Mod2), means were set to .25 and .5, to explore the goodness of fit of a model where pause duration was 1/4 or 1/2 of the word+pause duration.[1] In the third (Mod3), means were set to .33 and .67 to explore the goodness of fit of a model where pause duration was 1/3 or 2/3 of word+pause duration. Differences in log likelihood were then compared between the unrestricted model and the two more restricted models to evaluate which of the latter two models better captured patterns in our data. The same process was repeated for other prosodic variables; results for Models 1-3 are presented in Table 1.

	Log Likelihood				
	Mod1 (Unr.)	Mod2 (.25,.5)/ (.5,.75)	Mod3 (.33,.67)	Diff Mod1 vs. Mod2	Diff Mod 1 vs. Mod 3
Prop_ Word	102.02	99.67	94.38	2.34	7.64
Prop_ Syll	71.92	71.35	59.99	0.57	11.93
Prop_ Foot	98.23	95.41	89.08	2.82	9.15
Prop_ PA	89.25	68.66	66.48	20.59	22.77

Table 1: Log likelihoods of three mixture models fit to proportional measures of four prosodic variables

For all prosodic variables, the difference between Model 2 and Model 1 was smallest, indicating Model 2 was a better fit to the data than Model 3. Mod2 turned out to provide a very close fit to the Prop_Syllable data, with a difference in log likelihood to the best fit model of only 0.57. Mod2 also provided a close fit to the data for both Prop_Word and Prop_Foot, with differences in log likelihood from the best fit model at 2.34 and 2.82, respectively. Mod2 for Prop_PA was a considerably

worse fit to the data. Results for Mod2 for Prop_Syllable and Prop_Word are plotted in Figure 3. Histograms represent raw data, and gray density curves represent model-estimated components.

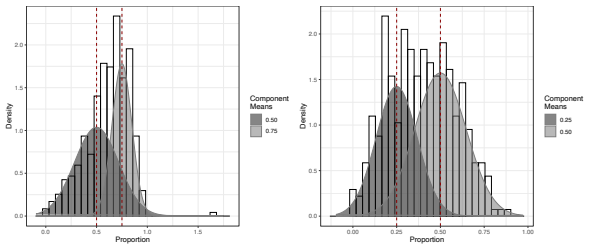

Figure 3: Results of Mod2 fit to Proportion_Syllable and Proportion_Word measures

In sum, it appears that inter-turn pause durations can be modelled efficiently as a proportion of a single unit spanning the pause and either the turn-final syllable or turn-final word. Specifically, it seems the pause tends to occupy either one-quarter or one-half of the duration spanning the final prosodic word and the pause, and one half or three-quarters of the duration spanning the final syllable and the pause.

5. DISCUSSION

Our results show some patterns which are consistent with Wilson & Wilson's coupled oscillator model of turn-taking, including that inter-turn pause duration was found to shorten overall as speech rate (measured in syllables per second) of the preceding utterance increased. This is in line with the idea of a syllable-level oscillator which regulates both speech rate and pause duration. The picture is complicated by the fact that the duration of the phrase-final prosodic word is negatively related to pause duration. One explanation for this pattern, in line with Wilson & Wilson's proposal, is that listeners plan their next turn based on the average syllable rate of their interlocutor's speech. As they anticipate the end of their partner's turn, they plan their pause based on a certain number of silent cycles of the syllable-level oscillator, which operates at this average frequency. This state of affairs predicts that, in the presence of more phrase-final lengthening (leading the oscillation rate of the partner's speech to be slowed relative to the average for the utterance), the inter-turn pause duration will be relatively shorter, since lengthening will cause the phrase-final word to extend further into the timing window projected by the listener for the pause. Under this account, the listener is essentially ignoring the presence of phrase-final lengthening for the purposes of pause planning.

The results of our mixture modelling procedure would suggest that pause duration is much more structured in its timing relative to phrase-final prosodic constituents, however. Specifically, our results suggest that pause time between turns tends to cluster around lower-order fractions of temporal units comprising turn-final prosodic constituents and the pause itself. Specifically, proportions tend to cluster around ¼, ½, and ¾ of the duration from the start of a turn-final syllable+pause or prosodic word+pause.

These results suggest that listeners are more sensitive to the durations of phrase-final constituents themselves, rather than simply ignoring the lengthening that applies to these constituents when planning pause duration. If this is the case, an account of the negative relationship between phrase-final word duration and pause duration could be that listeners plan for a certain number of oscillatory 'beats' based on the duration of the final prosodic word: if the word is shorter, they incorporate more silent beats, and if it is longer, they incorporate fewer. This may reflect the need for less turn-planning time where the phrase-final word is longer and affords the listener more syllables during which to plan their own speech. This proposal is in line with findings from Griffin [21], who shows that speakers utilize the time during articulation of their own speech to plan an upcoming word: in uttering pairs of nouns, for example, less silent planning time is needed at the start of the utterance if the first noun has a greater number of syllables during which the participant can plan articulation of the second noun. As previously mentioned, evidence suggests that listeners must plan their own turns during conversation while their interlocutor is finishing speaking [2,22]; if the listener's processing of the final word is mostly accomplished during the first syllable [23,24], then less attention needs to be paid to the remainder of the word, and this time can be used for planning the upcoming utterance. Of course, some pauses in our corpus were very long, and these longer durations are not likely to reflect solely lexical processing time. In these cases, we propose that participants are incorporating additional oscillatory 'beats' in order to consider their response to their partner's question, but still answer it within a predictable timeframe.

Our findings are in line with work by Fant and Kruckenberg [18] for speaker-internal pause timing, where the authors found evidence for complementary timing between phrase-final prosodic constituents and within-speaker pause durations, as well as a proportional relationship between pause durations and prosodic constituents like the foot. This work highlighted the importance of pauses in the temporal planning of speech. Our findings add to this body of research, and suggest that prosodic constraints on pausing behavior at the level of the individual are also active at the interpersonal level during conversational turn-taking.

6. ACKNOWLEDGEMENTS

Thank you to the study participants, to Karee Garvin for helpful feedback and insights, and Nicole Taylor for assistance with data collection. This work was supported by National Science Foundation Linguistics Program Grant No. BCS-2018003 (PI: Kathryn Franich). The National Science Foundation does not necessarily endorse the ideas and claims in this paper. All errors are the researcher's own.

7. REFERENCES

[1] Wilson, M. and Wilson, T.P. 2005. An oscillator model of the timing in turn-taking? *Psych. Bul. & Rev.* 12(6), 957–968.

[2] Levinson, S.C. and Torreira, F. 2015. Timing in turn-taking and its implications for processing models of language. *Front. Psych.*, 6, 731.

[3] Gravano, A. and Hirschberg, J. Turn-taking cues in task-oriented dialogue. 2011. *Comp. Speech & Lang*, 25, 601–634

[4] Corps, R.E., Crossley, A., Gambi, C., and Pickering, M.J. 2018. Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition*, 175, 77–95,

[5] Ford, C.E., and Thompson, S.A. 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the projection of turn completion. In: Schegloff, E.A. and Thompson, S.A. (eds), *Interaction and Grammar*. Cambridge: Cambridge University Press, 135-184.

[6] Ferrer, L., Shriberg, E., and Stolcke, A. 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2003. Proceedings. (ICASSP '03).

[7] M. Atterer, T. Baumann, and D. Schlangen. 2008. Towards incremental end-of-utterance detection in dialogue systems. In *Proceedings of Coling*, Manchester, UK.

[8] Cutler, A. and Pearson, M. 1986. On the analysis of prosodic turn-taking cues. In Johns-Lewis, C. (ed), *Intonation and discourse*. London: Croom Helm, 139–155

[9] De Ruiter, J.P., Mitterer, H., and Enfield, N.J. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82, 515–535.

[10] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh University Press.

[11] Dauer, R.M. 1983. Stress-timing and syllable-timing re-analysed. *J. Phon.,* 11 51−62.

[12] Grabe, E. and Low, E.L. 2002. Durational variability in speech and the Rhythm Class Hypothesis. In: Gussenhoven, C. and Warner, N. (eds), *LabPhon 7*. Berlin: Mouton de Gruyter.

[13] Mooney, S. and Sullivan, G.C. 2015. Investigating an acoustic measure of perceived isochrony in conversation: Preliminary notes on the role of rhythm in turn transitions. *Selected Papers from New Ways of Analyzing Variation (NWAV)* 43, 129–135.

[14] Turk, A.E. and Shattuck-Hufnagel, S. 2007. Multiple targets of phrase-final lengthening in American English words. *J. Phon.* 35, 4, 445–472,

[15] Boersma, P. and Weenink, D. (2022*). Praat: doing phonetics by computer* [Computer program]. Version 6.3.03. http://www.praat.org/

[16] Schegloff, E.A. 1996. Turn organization: one intersection of grammar and interaction. In: Ochs, E., Schegloff, E.A. and Thompson, S.A. *Interaction and Grammar*. Cambridge: Cambridge University Press, 52–133.

[17] O'Dell, M.L., and Nieminen, T. 1999. Coupled oscillator model of speech rhythm. In: Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., and Bailey, A.C. (eds), *Proceedings of the XIVth ICPhS*, 2, 1075−1078.

[18] Fant, G. and Kruckenberg, A. 1996. On the quantal nature of speech timing. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA.

[19] Lea, W.A. *Trends in Speech Recognition*. Prentice Hall, Inc., 1980.

[20] Franich, K. 2021. Metrical prominence asymmetries in Medumba, a Grassfields Bantu language. *Language*, 97(2):365-402.

[21] Griffin, Z.M. 2003. A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psych. Bul. & Rev.* 10, 603-609.

[22] Sacks, H., Schegloff, E., and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking in conversation. *Language* 50, 696-735.

[23] Cutler, A. & Carter, D.M, (1987). The predominance of strong initial syllables in the English vocabulary. *Comp, Sp. Lang.*, 2, 133-142.

[24] Culler, A, & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Exp. Psy.: Hum. Perc. Perf.,* 14, 113-121.

[1] For final syllables, a slightly different model (Mod2a) was constructed to evaluate the fit of a model with component means at .5 and .75, since syllable duration is relatively shorter than our other measures and therefore pause duration may have constituted a relatively larger proportion of the syllable+pause interval.