# Is there an optimal window size for a moving window analysis of pitch entrainment?

May Pik Yu Chan[1], Meredith Tamminga[1]

[1]*University of Pennsylvania (USA)*
pikyu@sas.upenn.edu, tamminga@ling.upenn.edu

**1 Introduction.** Conversational entrainment broadly refers to the adaptation of speech characteristics between interlocutors to become more (or less) similar to one another, which shapes the dynamics of speech variation and may be a mechanism for the propagation of sound change. [1] reviews the ever-growing range of work on this phenomenon and propose a systematic entrainment typology. In their terms, our current study focuses on **static global synchrony**, which refers to whether interlocutors fluctuate together in variability of speech features (here, pitch) “across any time scale greater than adjacent turns” [1].

We adopt a moving window approach to static global synchrony, following e.g. [2, 3]. As [1] points out, past work has measured static global synchrony over windows of widely varying sizes; however, there has been little systematic inquiry on what constitutes an ideal window size for analysis. On one hand, window sizes that are too small may be insensitive to entrainment behaviour that requires time beyond a turn domain to appear. On the other hand, window sizes that are too large may get too close to the overall average of the conversation and wash out effects of observable synchrony. Methodologically, we might therefore wish to find the minimum window size that maximizes consistency in the conclusions drawn. Thus, we systematically compare window sizes to ask whether different methodological decisions would lead to consistent conclusions about which dyads exhibit weaker or stronger pitch synchrony. However, the question of window size optimization is not merely methodological; it also speaks to questions about the true time scale on which entrainment behavior takes place and whether that time scale might vary across individuals, dyads, or interactions.

**2 Methods.** We analyzed recordings from fifty four same-gender (47 female, 7 male) existing friendship pairs (108 participants) from Philadelphia that spoke English as a native language. The pairs of friends engaged in dyadic conversations lasting 30 minutes in a laboratory room without a researcher present. Recordings were transcribed and forced-aligned using FAVE [4]. F0 information was extracted on a frame by frame basis from Praat based on the autocorrelation method.

The HYBRID method as outlined in [2] was adopted, which is an utterance sensitive approach to the time aligned moving average method of synchronic convergence [3]. In this method, the summarized speech characteristic (in our case, median F0) of both speakers within a given window length until the end of the speakers' utterance is measured for each step. We test 20 different settings ranging from a step size spanning 5s to 100s, each 5 seconds apart, with the window length being 2.2 times the length of the step size, which is comparable to [2]’s settings of a window length of 110s, and a step size that spans 50s. In other words, for every step size of a given length (e.g. 5 seconds), we took the median F0 from a series of overlapping windows of a given length (e.g. 22 seconds), but extending to the end of the utterance(s) that began within the window boundary. Here we label step/window settings using the step size in seconds. In this study, utterances were separated by silent portions or interruptions (e.g. laughter, noise) of a speakers' speech as defined by the forced aligner. Upon converting the series of median F0 values for each speaker into semitones with a baseline of 100 Hz, a Spearman's $\rho$ value was calculated for each dyad, for every given window length setting. This was used as a proxy for the overall magnitude of synchrony across the conversation for each dyad.

We then compared the rankings of dyad synchrony (the Spearman's $\rho$) for each given window length setting with every other step size setting using Kendall's $\tau$. These results tell us whether the rankings of high vs low synchrony dyads are consistent across different window length settings.

**3 Results.** Results of the matrix of Kendall's $\tau$ values are shown in Figure 1. Brighter yellow/orange regions show that the pair of window length settings resulted in similar dyadic rankings of synchrony. In other words, those pairs of window length settings led to more consistent results regarding which dyads synchronized more. Darker purple regions, conversely, indicate that the pair of window lengths resulted in different conclusions about which dyads synchronize most.

Figure 1 shows that smaller window lengths result in dark purple regions on the heatmap, suggesting that dyad synchrony rankings are quite unstable at small window lengths (especially below 25s). After around step size 45s however, the general $\tau$ values increase, suggesting a relatively stable ranking between conversations at higher step sizes. For example, when comparing 40s to 100s steps (τ = 0.48), there is less stability in the dyad rankings than comparing 45s to 100s steps (τ = 0.56).

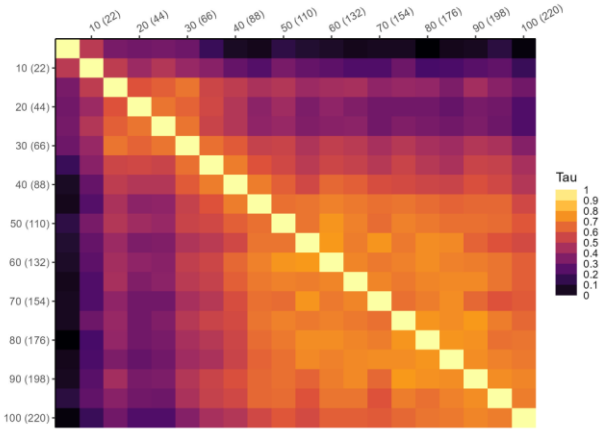

**Fig.1** Heatmap visualizing Kendall's $\tau$ values of dyad synchrony rankings across different step size / window length settings.

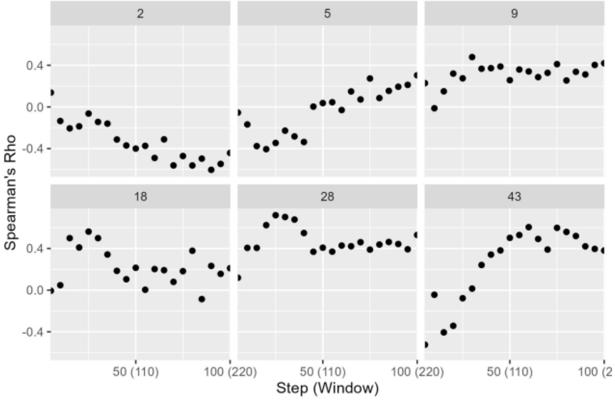

**Fig.2** Spearman's $\rho$ value by window length setting for some sample dyads.

**4 Discussion and Conclusion.** In this study we explored the relationship between different window length settings in an utterance sensitive moving window approach to synchronic conversational entrainment. We found that there is a range of smaller window lengths that lead to relatively unstable rankings between dyads, but once a threshold is passed, the relative rankings of dyads became more comparable at wider window lengths. A theoretical question of why the smaller window lengths have more unstable rankings remains. We outline three possibilities: First, it is possible that the interlocutors are not that fine grained in their levels of entrainment, meaning that the effect of attending to interlocutors' speech characteristics may come with temporal delay or be spread over longer time spans. Therefore, too small of an analysis window may result in too much noise. Second, it may be that there is true micro-temporal entrainment on a dyadic level, but that the optimal window size for detecting this rapid synchrony is contingent on the dyad. In other words, we may be unable to find a one-size-fits-all setting because each pair of speakers entrains to each other on different time windows. This seems possible given Figure 2, which shows the $\rho$ value by window length setting for a sample of dyads. Some dyads appear to have a preferential setting that maximizes $\rho$, while others have more gradient or jittered relationships between settings. A third possibility is that entrainment takes place not on a time-sensitive window, but rather is contingent on discourse-specific units, such as turns, topics or other conversational events. In other words, entrainment might happen at a window length that is dynamic within the conversation, and cannot be prespecified without attention to the content of the interaction.

Taken together, our results suggest that conclusions drawn about dyadic synchrony are sensitive to analysis methods, meaning researchers need to give careful consideration to reasonable settings based on their dataset. Based on our dataset, we might suggest a window length of 99 seconds spaced at 45 second steps as a practical starting point for exploration. Future work comparing different datasets, methods, and speech features may help validate whether this recommendation could be applied widely or if it is context-sensitive.

References

[1] Wynn, C. J., & Borrie, S. A. (2022). Classifying conversational entrainment of speech behavior: An expanded framework and review. *Journal of Phonetics, 94*, 101173.

[2] Bonin, F., Looze, C.D., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., Campbell, N. (2013). Investigating fine temporal dynamics of prosodic and lexical accommodation. *Proc. Interspeech 2013,* (pp. 539-543). Lyon, France.

[3] Kousidis, S., Dorran, D., Wang, Y., Vaughan, B., Cullen, C., Campbell, D., McDonnell, C. & Coyle, E. (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. In *Ninth Annual Conference of the International Speech Communication Association (Interspeech 2008)*, (pp. 1692-1695). Brisbane, Australia.

[4] Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite. http://fave.ling.upenn.edu.