Proceedings of the Edinburgh Linguistics Department Conference '96 pp. **18–27**

An analysis of between-speaker intervals

Matthew Bull
matthew@ling.ed.ac.uk

Abstract

It has been suggested (e.g. by Couper-Kuhlen) that the timing of turn-taking is coordinated by a perception of isochrony in a speaker's utterance. Next-speakers subconsciously pick up the rhythm of current-speakers' utterances, and time their entry to the floor so that the first prominent syllable in their turn coincides with the 'beat' that was set up beforehand. A pilot study was carried out to test the validity of this hypothesis. Digitised recordings of spontaneous dialogue were marked for prosodic prominences approximating to pitch accents. The intervals between these prominences were measured, both within and between speakers. It was expected that the between-speaker and within-speaker intervals would be approximately equal — or at least between-speaker intervals would approximate to multiples of within-speaker intervals. However, no evidence was found to support this, even when allowing for a relatively loose definition of isochrony. When inter-speaker intervals were analysed according to move type, it was found that there was a relationship between the two, and that some move types involve longer or shorter inter-speaker intervals than others.

1 Introduction

The timing and coordination of turn-taking is a non-trivial process on the part of both speaker and listener in any conversation. The process of deciding on the next speaker is said to follow a set of rules as described by Sacks *et al.* (1974) where a current speaker may select a next speaker. If not, a next speaker may self-select. Failing both these conditions, the option to continue passes back to the current speaker, who may or not continue. However, not all conversation follows these rules, and more recent research by Coates[1] amongst others shows that some types of conversation amongst friends operate according to completely different rules of floor-taking. For the purposes of simplicity, however, it is as well to consider only those two-party conversations in situations where the rules of Sacks *et al.* obviously do apply.

But even in this relatively uncomplicated form of interaction, the timing of turn-taking is extremely precise. It is clearly not the case that a next speaker is only able to time a new turn by waiting till the current speaker's turn is completely over, and then starting

[1]According to the paper given in the Fifth Manchester Postgraduate Conference, 9th March 1996.

the new turn. If this were the case, overlap would never occur. Inter-speaker intervals[2] may often be only 100ms or less (in this study, about a quarter of all intervals were less than 100ms, and nearly a third were less than 200ms). However, evidence from shadowing studies (where subjects are asked to listen to an utterance, and repeat what they hear as soon after they hear it as possible) indicates that typically about 200ms is the shortest response time to speech. And this does not include the planning time involved in normal speech production. It is clear that some degree of prediction of turn closures must occur, and that the start of a new turn can be coordinated with this prediction point rather than with the closure itself. This point is referred to in the literature as the transition relevance place (TRP) (Sacks *et al.* 1974).

Cues to turn closure may vary from the syntactic to the paralinguistic. Duncan (1972) noted six turn-yielding cues. One or more of these cues may be displayed to give a turn-yielding signal. The six cues are: intonation, paralinguistic drawl, body motion, sociocentric sequences (the use of stereotyped expressions), paralinguistic pitch or loudness, and syntax. Further to this, it would seem that gaze may act as a cue to turn-taking. Kendon (1967) found that listeners generally look more at their conversational partners than speakers do, and that the close of a turn may be signalled to some extent by the speaker looking more at the listener — thereby signalling a willingness to switch to the role of listener. Another cue, mentioned by Ford & Thompson (1995), is pragmatic completion. They point out that their judgements of pragmatic completion remain intuitive and provisional, but that it cannot be ignored because it is involved in the prediction of turn completion.

But given that cues exist which allow a speaker to predict when a TRP is imminent, what remains uncertain is what rules, if any, are used by next speakers to choose a moment to start their turn. It seems clear that if speakers delay their start too long this results in an 'uncomfortable silence'. And if excessive overlap occurs this may be construed as impolite. Such limits may be largely socially determined, although it is not at all certain whether different cultures vary greatly in their acceptable limits. Also, the limits for inter-speaker intervals will vary greatly from one context to another. Yet these limits do seem to exist, and the research reported here is part of an attempt to understand their nature. In particular this paper considers a hypothesis which proposes a rhythmic basis for the coordination of turn-taking (Couper-Kuhlen 1993). But also under consideration to some extent is the question of whether turn-taking can be viewed as consisting of inter-speaker intervals which are simple response latencies, or as consisting of intervals which are generated collaboratively (Wilson & Zimmerman 1986).

Wilson & Zimmerman note three principle strategies which have been proposed as a model of turn-taking: stochastic models, signalling models, and sequential-production models (they do not include the rhythmic hypothesis in their account).

The stochastic approach is essentially a Markov or other probabilistic process, where

[2]Note that throughout this paper I use the term 'inter-speaker interval' to refer to the period where neither conversationalist is talking. Inter-turn intervals are a type of inter-speaker interval where the current speaker has not selected the next speaker, and where silence cannot be described as falling within the turn of either speaker. This distinction has not been accounted for in this study, but should be in future research because there may well be important timing differences between the two cases.

the conversation is made up of different states such as someone speaking, or silence, or both parties speaking. The model then describes the probabilities of transitions from one state to another (e.g. Jaffe & Feldstein 1970). The signalling approach assumes that conversation is organised purely according to a series of cues which signal the imminent closure of a turn, as in the approach outlined earlier. The sequential-production model is the only interactive one of the three, and involves conversationalists actively collaborating to organise turn-taking (Sacks *et al.* 1974).

Wilson & Zimmerman separate the stochastic and the signalling models from the sequential-production model. They involve the treatment of inter-speaker intervals as basically a simple response latency on the part of the next speaker. The prediction from this is that an analysis of intervals would indicate a monotonic decline in the frequency of intervals as their duration increases. Wilson & Zimmerman claim that the sequential-production model differs in that it predicts a periodic pattern to the durations of intervals. This prediction arises because of the turn-taking rules of Sacks *et al.* (1974), where the option to take the floor may pass between current speaker and next speaker until one of them actually takes the floor. This process is based in real time, and the interval during which each conversationalist has the option to take the conversational floor is termed the *slot.* Wilson & Zimmerman claimed, if somewhat tentatively, to have found evidence for the existence of a slot duration somewhere between 40 and 90ms.

The rhythmic hypothesis also approaches the coordination of turn-taking from an interactional perspective, and for different reasons from those motivating the sequential-production model it also predicts a degree of periodicity to durations of inter-speaker intervals.[3] The rhythmic hypothesis proposes that in conversation perceptual isochrony may occur. That is, even where there is no measurable or exact isochronous pattern to the stresses in speech, listeners nevertheless perceive the stresses to be isochronous. The validity of this claim is uncertain, and there does not appear to have been any systematic test carried out on it other than in restricted cases such as word lists (e.g. P-center research; see Donovan & Darwin 1979; Darwin & Donovan 1980). Rather, it seems to have been assumed to be the case. In fact, it is not at all clear which level of the prosodic hierarchy applies in perceptual isochrony, although the foot, or rhythm group (Cruttenden 1986) has been used traditionally in isochrony research. Couper-Kuhlen (1993) claims that different levels may be used according to the tempo of the speech.

Assuming that perceptual isochrony applies, the next consideration is how this relates to turn-taking. According to Couper-Kuhlen (1993) the rhythmic beat (the term ‘rhythmic beat’ will be used to describe a perceived isochronic pattern of stresses) in a current speaker’s turn may be continued by the next speaker, such that an extrapolation of the beat coincides with the first beat of the next turn. This would represent an idealised form of turn-taking with no sense of ‘awkwardness’.

The problem in testing this hypothesis is that it relies on essentially subjective impressions of what counts as isochronous, and what counts as a ‘fluid’ transition. Nevertheless,

[3] However, it is not totally impossible that the periodicity predicted in both cases is in fact caused by similar processes.

a study (Bull 1995) was carried out to assess subjective listener judgements concerning the naturalness of turn transitions in relation to the rhythm of an utterance. The results from this did not indicate any ability to discriminate on the basis of rhythm. But the study suffered from attempting to quantify what is apparently not quantifiable beyond subjective experience, or through any relatively simple analysis of speech. As with the determination of stress, its intangibility does not deny its existence — it simply makes it extremely difficult to test adequately.

The aim of the current study was to make an instrumental analysis of inter-speaker intervals by comparing the inter-stress intervals both within and between speakers. According to the theory, the perception of the duration of the between-speaker stress interval (the between interval) and the duration of the within-speaker stress intervals (the within interval) should be equal — or failing this, the between interval should be some multiple of the within interval. It is not certain what the upper limit of this multiple should be, but it would seem unreasonable to allow anything above three times the 'within' interval. Naturally, the qualms noted above about the testability of the hypothesis cannot be disregarded. But they may be accounted for to a large extent assuming some degree of latitude around which intervals may differ from one another and yet still be perceived as isochronous.

2 Method

624 exchanges (two turns separated by a speaker switch) were chosen from the HCRC Map Task Corpus (Anderson *et al.* 1991). Of these only 343 were of any use for testing the rhythmic hypothesis, since many exchanges consisted of turns which were one-word utterances and which therefore had no rhythmic pattern. Note that exchanges were included when only the second turn was a one-word utterance, since the hypothesis would predict that the stressed syllable in such utterances would fall on the beat. Prosodic prominences approximating to pitch accents were marked in these exchanges. In some instances it was possible to test subjective judgements against F0 pitch tracks, although often the pitch tracks did not necessarily fit these subjective judgements. In these cases, listener judgements took precedence owing to the perceptual groundings of the hypothesis. Also, problems were encountered because of the relatively flat intonation of many of the speakers in the recordings.

A further discrimination was made in selecting the data, which I have called *fluency*. This is primarily a subjective measure which accounts for those exchanges where a second turn occurred only after many seconds delay, and accompanied by obvious linguistic markers of hesitation such as 'erm'. Such exchanges are non-fluent and were not included in the analysis reported here, because in these cases the causes of the interval were potentially too uncertain.

Inter-stress intervals were calculated from the durations between the prominences (the within intervals). Inter-stress intervals across speaker boundaries were also calculated (the between intervals). Additionally, inter-speaker intervals were measured. That is, the duration from the end of the first turn to the beginning of the second turn was measured. While this had no direct bearing on an analysis of the rhythmic hypothesis, it was a worthwhile measurement to make for a more general analysis of patterns of durations across turn boundaries. Finally, the type of move involved at each turn was noted, so that some preliminary account of the effects of context could be made.

3 Results

Figure 1 shows the ratios of between intervals to within intervals, expressed as percentages of the total number of intervals measured. Thus exchanges where the between to within ratios fall in the range 1.01 - 1.10 make up 6.3% of the total number of exchanges. The rhythmic hypothesis makes the prediction that the ratios of between intervals to within intervals should approximate to integers — primarily ratios of 1:1 and 2:1. 6.3% of ratios fall within 1.01 - 1.10, and 3.5% fall within 2.01 - 2.10.

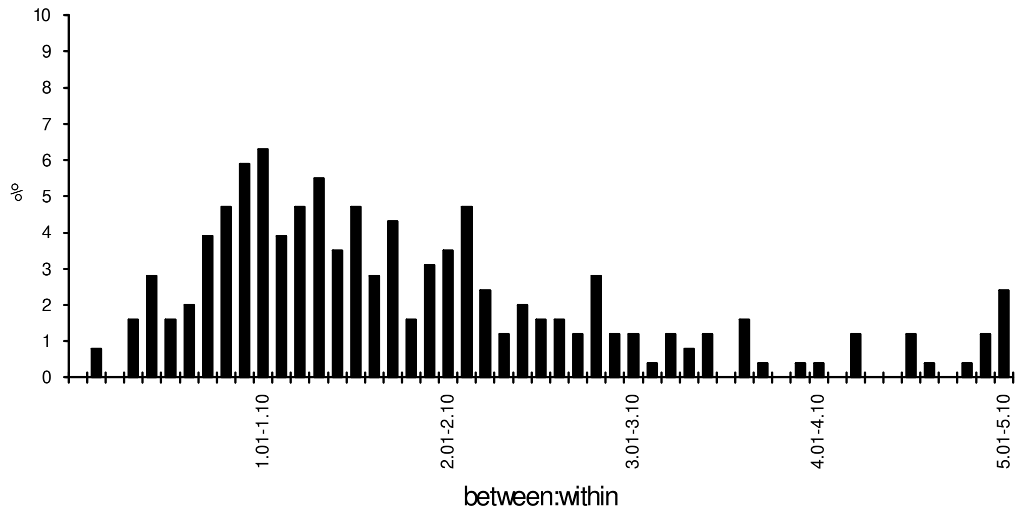
25.5% of the total ratios fall within the range 0.81 - 1.21, and 15.3% fall within the range 1.81 - 2.21. Therefore, 40.8% of all exchanges considered here consist of between interval to within interval ratios which fall inside a $\pm 20\%$ window of either the 1:1 or the 2:1 ratio.

The question remains as to whether these results are significant enough to support the rhythmic hypothesis. If nothing else, one might argue that a noticeable feature of Figure 1 is the bias of results toward a ratio of approximately 1:1, and that this alone would suggest some rhythmic effect. But it is still uncertain whether this is significant, and whether this bias is coincidental. Figure 2 shows a scatter plot of between intervals against within intervals.

Clearly, there is no significant correlation of the between intervals against the within intervals (R = 0.050, p = 0.427). There does appear to be some degree of clustering of intervals below about 1000ms for within intervals, and below about 1800ms for between intervals. But there is very little relationship between the two other than this. Even if there were, one would at least expect a gradient of 1 to the regression line. This was not the case here ($y = 882.9 + 0.0831x$).

Figure 3 shows the distribution of inter-speaker intervals, each range of intervals expressed as percentages of the total number of exchanges. As with Figure 1, we have a bell-shaped distribution, biased toward one end of the curve. As such, relatively short inter-speaker intervals appear to be more common than longer ones as well as negative ones (that is, cases of overlap).

Finally, consideration was made of move type. For current purposes a detailed description of the different types of move will not be given, although the reader is referred to Carletta *et al.* (1995). The main point here is that exchanges involving different move types were found to have different average inter-speaker intervals. Figure 4 shows the mean inter-speaker intervals for each exchange, categorised according to the move type of the second move in each exchange. These results are only preliminary, however, because of the small sample of data and thus the small number of cases of some of the move types.



n = 254
mean = 1.81
SD = 1.37

Figure 1: *Ratio of between to within intervals.*

4 Conclusions and discussion

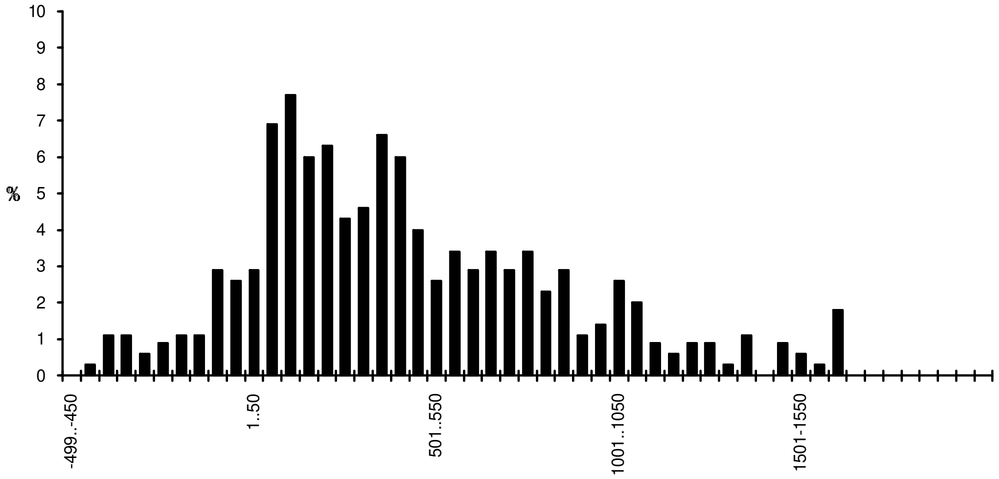
The results offer no evidence in favour of the rhythmic hypothesis. Between intervals and within intervals did not appear to be correlated to any significant degree. The observation that the ratio of between intervals to within intervals was biased toward 1:1 did not therefore appear to result from any rhythmic effect. Rather, it resulted from a tendency for between and within intervals to cluster within certain ranges (as shown in Figure 2) which are coincidentally similar. The results obtained from Figure 3 showed that inter-speaker intervals do tend to be more frequent at lower durations, gradually becoming less frequent as duration increases. Such results are consistent with other studies (see for example Wilson & Zimmerman 1986).

It could be argued that the rhythmic hypothesis is a perceptual model, and as such detailed instrumental analyses of inter-stress intervals are not necessarily meaningful. There are two problems with this. First, perceptual studies run the risk of being subject to too many contextual variables. There is moreover little support for the rhythmic hypothesis from perceptual studies. In Bull (1995) it was found that subjects were unable to decide consistently whether a given inter-speaker interval was any more or less 'natural' than any other, within certain extremes.

The second problem is that without some form of instrumental support, the perceptually-based rhythmic hypothesis cannot be used meaningfully to generate a predictive model of the timing of turn-taking. It seems reasonable for a certain degree of manoeuvre to be allowed in the analysis, such that between and within intervals would not need to be exactly

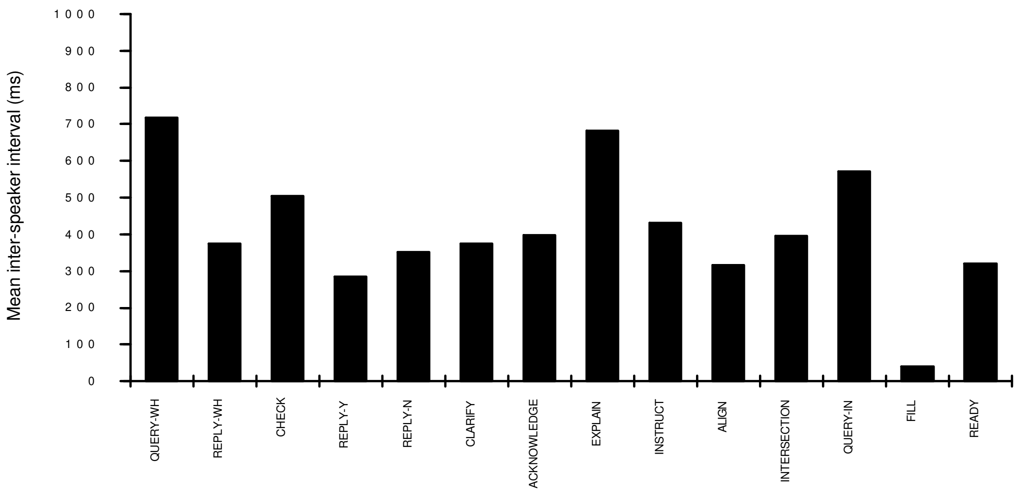
```
         .+.....+.....+.....+.....+.....+.....+.....+.....+.....+.....+......
   5000  +                                                                  +
         -                                                                  -
         -                                                                  -
         -                                                                  -
         -                                                                  -
         -                                                                  -
         -                                                                  -
   4000  +                                                                  +
         -                                                                  -
         -               1                                                  -
         -                                                                  -
         -                                                                  -
         -        1                                                         -
         -                                                                  -
b  3000  +                                                                  +
e        -                                                                  -
t        -                   1                                              -
w        -                                                                  -
e        -                                                                  -
e        -             1                                                    -
n        -                1    1  1                                         -
   2000  +                     11         1         1                       +
         -                   1 1     1                                      -
         -         11    1 1                     1                          -
         -          11  31 1  1 1                  1              1         -
         -        11 21 21    2          1                                  -
         -      1 11  1211 1 2 1 11      1    1                             -
         -       1  112 1  1121 1 2 2  12   1     11                1       -
   1000  +         1113233213 1 1113     11  1       1                      Y
         Y       1  3 331 22   2111 11  1 1 1                               -
         -     2   11534211 2 2 1332  11    11        1                     -
         -       13  2543 142  251  11    1        1                        -
         -       12222113 113 1      1  1  1                       1        -
         -        1   211 1  2   311       1 1                              -
         -                  1     1          1                              -
   0.00  +                     1                                            +
         M+.....+.....+.....+.....+.....+.....+.....+.....+.....+.....+......
               200.        600.        1000        1400        1800
         0.00        400.        800.        1200        1600        2000
N =  254
R = .050                               within interval (ms)
P = .427
                                    MEAN    S.D.
--REGRESSION LINE--  -RES.MS-    X  626.27  304.08
Y= 882.85 +.08314*X    255463    Y  934.92  505.07
```

Figure 2: *Scatter plot of between intervals against within intervals.*



n = 349
mean = 404ms
SD = 421ms

Figure 3: *Distribution of inter-speaker intervals.*



n = 349
mean = 404ms
SD = 421ms

Figure 4: *Inter-speaker intervals by move type.*

the same, but could differ by, say, $\pm 20\%$ of each other. Even if this is accounted for, the results in Figure 1 do not cluster around ratios of 1:1 or 2:2 sufficiently for string evidence of a rhythmic effect. And if one were to increase the size of the window, the usefulness of the model is further diminished.

This does not mean that rhythmic structure plays no role in turn-taking. It may be that rhythm aids prediction of TRP location, because if TRP falls on rhythmic beats then it would be possible to predict when one might occur (although factors such as syntax, pragmatics, intonation, and gesture would be needed to predict if a TRP is actually imminent). It is also quite possible that the tempo of the speech surrounding any given exchange may to some extent be correlated to inter-speaker intervals. But any role played by rhythm would appear to be small, particularly compared to context and sociolinguistic concerns.

One alternative hypothesis would follow lines similar to that suggested by Wilson & Zimmerman (1986). An important difference would be that the slot length suggested by them would be much longer — something possibly of the order of several hundred milliseconds, although presumably this would vary greatly according to context. There may be situations where next-speakers would have to take the floor within one hundred milliseconds of a TRP in the current-speaker's turn before the floor is lost. The situation is still further complicated because of the distinction between local and global context changes. Figure 4 shows that in the same general linguistic and social context (same global context), different move types typically have different inter-speaker intervals (different local context). It follows that if these differences are in any sense conventionalised that current-speakers would be willing to tolerate varying slot lengths before resuming the conversational floor. Therefore, future research would look in greater detail at cases where the floor is offered but not taken. Move type and gaze would also have to be taken into account to a greater extent.

References

Anderson, A.H., M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H.S. Thompson & R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech* 34 (4): 351–366.

Bull, M. 1995. An appraisal of rhythm as a coordinator of turn-taking. *Proceedings of the XIIIth International Congress of Phonetic Sciences Vol 3*. pp.480–483, Stockholm.

Carletta, J., I. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon & A. Anderson. 1995. The coding of dialogue structure in a corpus. In J. A. Andernach, S. P. van de Burgt & G. F. van der Hoeven (eds). *Proceedings of the Ninth Twente Workshop on Language Technology: corpus-based approaches to dialogue modelling*. Universiteit Twente pp.25–34.

Couper-Kuhlen, E. 1993. *English Speech Rhythm: form and function in everyday verbal interactions.* Amsterdam: John Benjamins.

Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

Darwin, C.J. & A. Donovan. 1980. Perceptual studies of speech rhythm: isochrony and intonation. In *Spoken Language Generation and Understanding. Proceedings of the NATO Advanced Study Institute*. Dordrecht: D. Reidel.

Donovan, A. & C.J. Darwin. 1979. The perceived rhythm of speech. *Ninth International Congress of Phonetic Sciences*, Vol. 1 pp.268–274.

Duncan, S. 1972. Some signals and rules for taking speaking turns in conversation. *Journal of Personality and Social Psychology* 23: 283–292.

Ford, C.E. & S.A. Thompson. 1995. Interactional Units in Conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E.A. Schegloff & S.A. Thompson (eds). *Interaction and grammar*. Cambridge: Cambridge University Press.

Jaffe, J. & S. Feldstein. 1970. *Rhythms of Dialog*. New York: Academic.

Kendon, A. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica* 26: 22–63.

Sacks, H., E.A. Schegloff & G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation, *Language* 50: 696–735.

Wilson, T.P. & D.H. Zimmerman. 1986. The structure of silence between turns in two-party conversation. *Discourse Processes* 9: 375–390.