ENTRAINMENT AS A BASIS FOR CO-ORDINATED ACTIONS IN SPEECH

Richard Ogden, Sarah Hawkins

University of York, University of Cambridge
richard.ogden@york.ac.uk, sh110@cam.ac.uk

ABSTRACT

This paper asks how rhythmicity is used to manage speaker transition in spontaneous talk and how temporal alignment helps to achieve interactional alignment. 56 Question + Answer (Q+A) pairs were analysed. 44 (79%) Qs ended rhythmically: in their last few accented syllables, f0 prominences were quasi-periodic. Of the *As* to these rhythmic *Qs*, 32 (73%) began with the same periodicity as the *Q*. *A*s with non-rhythmic entry into 'turn space' set up by a rhythmic Q were sequentially and interactionally complex. Rhythmic *A* entries included accented syllables, in-breaths, clicks and nods, suggesting 'embodied' rather than solely 'linguistic' temporal entrainment. Interactional alignment thus seems to exploit temporal entrainment in the vicinity of turn boundaries, like that established for musicians.

Keywords: entrainment, rhythm, question-answer

1. INTRODUCTION

1.1 Entrainment as a basis of co-ordinated action

Talking and making music require participants to actively co-construct their interaction in time. Joint music-making demands tight timing and compatible use of e.g. genre, pitch and loudness. Likewise for conversation. Participants in conversation time and shape both their verbal and non-verbal contributions. Temporal and prosodic relations between adjacent turns determine how the second speaker's turn is interpreted in relation to the first, as a socially and structurally preferred or dispreferred response [17, 22]. They allow expression and perception of shared intentionality. Fundamental to synchronised timing is the concept of temporal entrainment. Entrainment is basic to coordinated music-making [2] but its role in conversation is less clear, presumably because consistent rhythm, if present at all in conversational speech, is much less obvious than in most music.

However, we showed [7] that interactants seem to entrain to one another *over short periods* in spontaneous conversation as well as in music-making, with gesture as well as sound. That work did not distinguish types of utterance, nor explore parameters influencing behaviour. This paper extends the question about entrainment to conversations when no one is making music, and compares more rigorously instances of rhythmicity vs. non-rhythmicity in a frequent and well-documented interactional structure, question-answer (*Q+A*) pairs. Using new data, we explore whether and if so how entrainment is achieved in everyday talk, and how temporal alignment in *Q+A* structures functions in terms of interactional alignment. We relate our findings to the wider issue of coordinated social interaction in general.

1.2 Questions and answers in conversation

Questions and answers constitute an adjacency pair [20]. In an adjacency pair, the first pair part, 1PP, (here: a *Q*) projects a second pair part, 2PP, (here: an *A*) which is pragmatically and syntactically fitted to the 1PP. Thus *As* stand in particular relation to *Qs* in several ways [17, 25, 26]. One of the most important interactional parameters is alignment: does the responsive action, *A*, treat the initiating action, *Q*, as a *Q*? If so, the response *aligns* with the *Q*. In 1) and 2) below, *A* aligns with *Q*. The design of *A* provides evidence for this independent of the phonetics. In 1), the yes-no *Q* gets a *yes*; in 2) *where* is responded to with a place name (*Girton*). Both *As* recycle the *Q's* morphosyntax e.g. *are you/I am*. In contrast, the *A* in 3) is not aligned: there is no *yes*/*no* in response to theyes-no *Q*, and no material from the *Q* is recycled.

1) are you enjoying your place now ‖ yeah I am, it's great
2) where is that ‖ it's near Girton
3) was that here as well ‖ I walked in & saw the cameras

In addition to these formal properties, relative timing is critical. Delay in producing a 2PP is treated by interactants as displaying a problem; well-fitted, preferred 2PPs typically start within a particular time-slot relative to the end of the 1PP [25]. Thus the *Q+A* structure serves as a useful test-bed for examining how turns are coordinated in time.

1.3 Pikes in conversation and music-making

To treat timing as a multimodal property of speech, music and gesture, we need a temporally-precise measure applicable to all three domains. We thus sought a simple, proven measure of rhythmicity that is applicable across modalities, ties together rather than distinguishes musical and speech domains, and can be reliably applied to gesture as well. Most work on speech timing focuses on correlates of rhythmic

type (stress vs. syllable timing) and rhythm metrics [1]. An early work on rhythm in interaction [3] estimated Perceptual (P-)centres by marking onset of periodicity in accented syllables, and counted as rhythmic those intervals which varied in duration by up to 30%. However, besides 30% seeming a lax criterion, it is problematic to estimate P-centres from an acoustic signal. While onset properties seem best able to predict P-centre location in both speech [28] and music [4, 5, 29, 30], there is currently no P-centre model that can be applied reliably to acoustic events in either domain, let alone to gesture, for which the concept has not been explored. Gestural work seems more promising. Loehr [11, 12] noted that f0 peaks on accented syllables, eye blinks, and peaks of gestures, tend to be coordinated in time and to co-occur among interactants. He proposed the term *pike* (π) to refer to a point of maximal physical activity. Loehr did not explore the question of whether πs function as audible and visible resources for the co-ordination of activities in interaction, but work on speech with music suggests they can [7]. The present work extends Loehr's by examining temporal details in adjacent turns more closely.

2. METHODS

2.1 Participants and recording procedure

Data come from five same-sex pairs of friends (two pairs female) aged 18-31, available from a larger set [7]. All were university educated, native speakers of stress-timed English (Southern British and Scottish).

Each pair was recorded doing a structured set of music-making and other activities together, designed to facilitate cooperative interaction, see [7]. Pairs sat in a recording studio at a round table, at an angle of about 120° to each other. Recordings used 4 digital video cameras and 5 microphones, including two close-talking head-mounted microphones. Signals were synchronised *re* one camera to a maximum error of 40 ms for video and 20.83 μs for audio.

2.2 Materials and labelling

The data come from the parts of the interaction where participants were talking but not manipulating other objects or playing music. There were 56 *Q+A* pairs—the commonest adjacency pair in these data.

Speech was labelled into Praat textgrids without video. For each speaker, words were segmented and f0 prominences in ToBi H and L accented (*) and boundary (% were annotated. Eye gaze and various types of gesture were also labelled but intervals between pikes reported here use f0 prominences unless explicitly mentioned (e.g. re Example 2 below).

2.4 Inter-turn temporal organisation

We use the following terms. *Qs* are **rhythmic** or **arhythmic**. *Qs* with at least 3 πs were classed as rhythmic when the intervals between adjacent πs differed by no more than ±15% and/or when there was a percept of rhythmicity as judged by 3 expert listeners. Arhythmic *Qs* display no such periodicity of πs. After a rhythmic *Q*, **rhythmic entry into the turn space occurs** when the first π of the *A*, π_1, comes in on the beat established by the *Q's* πs. Rhythmic *A* entry may be early relative to *Q* (π_1 of *A* co-occurs with a π of *Q*), or on-beat (π_1 of *A* falls on the next projected pulse after *Q*), or late (π_1 of *A* is on a beat projected by the pulse established in *Q*, but after one or more silent pulses). In **non-rhythmic entry**, π_1 of *A* does not come in on the beat established by *Q* [3]. In *Qs* with two πs, the speaker who produces the *A* can still establish rhythmicity across the turn space if π_1 of *A* comes after a similar interval to the interval between the *Q's* πs.

3. RESULTS

3.1 Rhythmicity in Questions

79% (44) of the *Qs* exhibited rhythmicity, so about 4/5 *Q+A* pairs had the potential for rhythmic entry. Arhythmic *Qs* had irregular intervals between πs; or too few πs to generate a pulse (e.g. *d'they RECognise you*); or no measurable f0 (e.g. breathy, low intensity); or were followed by an expansion in the same turn by the same speaker; or contained perturbations in production such as self-repairs.

3.2 Rhythmicity in turn space

Figure 1: No. of each type of *A* entry after rhythmic *Qs*.

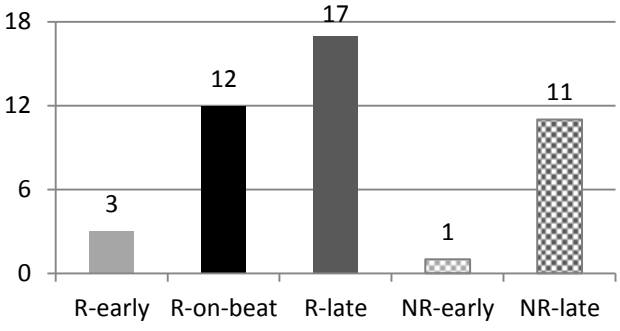


Fig. 1 shows the number of *As* with rhythmic (R) vs. non-rhythmic (NR) entry into turn spaces created by *Qs* classed as rhythmic. Rhythmic *A* onsets are by far the most common: 32 (73%) in total, 12 on-beat with no delay (R-on-beat), & 17 on-beat but having missed one or more beats (R-late). This compares with only 11 *A* entries classed as late and off-beat (NR-late). All 4 early entries (3 R, 1 NR) occurred

in the 'transition space' [8, 22] when it was clear the *Q* would soon end; illustrated by Example 3 below.

3.3 Analysis in interactional terms

Different types of entry into the turn space reflect the relation of the *A* to the *Q*: preferred *As* match the syntax and lexis of the *Q* [17], and tend to have tighter temporal relations with *Q* than dispreferred *As*. Yet our data indicate phonetically more complex ways for talkers to enter the turn space than those noted by [3]. In particular, non-verbal but vocal material may coincide with the projected next π.

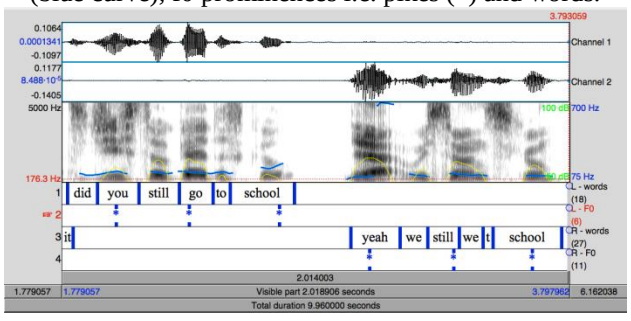
We illustrate with representative examples. In each example, each pike is marked π. Intervals between πs are in seconds. Orthographic transcriptions use conventions based on GAT [21]: accent-bearing syllables (* in ToBI) are aligned with π, their vowels being in capitals. Where the intervals are regular enough to establish a pulse, this is marked on the last line with ^. When a pulse is established, a next speaker can use it to place a next π (or other speech event) which is heard as on-beat with the rhythm set up by the prior talker.

3.3.1 Rhythmic entry into the turn space

Example 1. On-beat rhythmic entry (no delay)

pike		π	π	π			
interval			0.29	0.36			
L	did yOU still		gO to	schOOl?			
pike					π	π	π
interval					0.36	0.33	0.31
R					yEAh?. we	stIll went	-schOOl- to
pulse				^	^	^	^

Figure 2. Example 1's waveforms, spectrogram, f0 track (blue curve), f0 prominences i.e. pikes (*) and words.



Example 1 shows an *A* that comes in on the beat established by the last π of the *Q*. Fig. 2 shows the acoustics and labels. The first π falls on *you*. At this point R can recognise from the syntax of the turn so far that L is producing a *Q*. The second π is on *go*, with an interval of 0.29 s. The third π is on *school*, 0.36 s later, setting up a weak pulse. The pulse established in the *Q* provides R with a time slot to which to align his answer.

R's answer starts with *yeah*, whose π occurs 0.36 s after L's last pike—the same duration as L's last π interval. Although the pulse in L's turn is imprecise, R seems to orient to it and use it to time his *A* onset. The *A* onset occurs on the next beat. It comes in 2 parts, a confirmation + a Turn Constructional Unit (TCU) which recycles much of the syntax and lexis of the *Q*. The πs in *A* come at intervals of roughly 0.33 s, so they maintain the pulse established at the end of the *Q*. In this *Q*+*A* pair, then, interactionally the *A* is syntactically and lexically fitted, and it delivers a confirming and aligning response [25, 26]. Our analysis shows that, typically for such pairs, the temporal organisation of the turn transition is tight.

Example 2. On-beat rhythmic gestural entry

pike	π	π			
interval		0.38			
L	sAme	pEople?			
pike			π	π	π
interval			~0.38	~0.38	~0.38
R					yeah
			NOD	NOD	NOD
pulse			^	^	^

Responses need not be verbal or even vocal. Example 2 shows confirmation done with head nods. Here L's *Q* only has one interval, but R's *A* has πs at roughly the same interval (the figures are estimates due to video frame length), so in this case the pulse is established by the *A* rather than by the *Q*. πs in this case are manifest through a physical action other than speech; this example shows that πs are not just relevant to speech, and can be used to rhythmically coordinate adjacent actions in time.

Example 3. Early rhythmic entry, pulse to 3rd position

pike	π	π	π		π
interval		0.37	0.321		0.734/0.376
L	it was a SUnday	wAsn't	it		yEAh
pike			π	π	π
interval			0.301	0.377	0.381
R			yEAh it	wAs	yEAh
pulse			^	^	^

Example 3 shows an early entry after a *Q* whose ending was predictable, and also that a pulse established in one turn can be maintained beyond the *Q*+*A* pair into the third position [20], which in this case is a place for L to confirm and ratify R's *A* with *yeah*. At this point, L and R display (in Conversation Analytic terms) a shared understanding, and their talk is temporally aligned. In this and other examples, sustained rhythmic entrainment can be thought of as a device to display social or interpersonal alignment.

Example 4. Late verbal rhythmic entry; complex *A*

pike	π	π						
interval		0.33						
R	[hOw was your	`mEEting.						
pike					π	π	π	π
interval					0,99	0,33	0,48	0,43
L	[good			it was	`gOOd; it was very	`lO:ng; it	stARted at e-	lEven
pulse			^	^	^	^		^

Example 4 shows a late rhythmic entry into the turn space. Here, the first two πs of *A* are on-beat (*good… long*), but *A's* expansion, which provides an account for *long*, does not maintain the pulse established across the transition space. The account started with this TCU provides a more complex *A*.

Finally, a difference between our data and the literature [3] on rhythmicity in turn-taking space for English is that some rhythmic entries are rhythmic due not to the timing of an accented syllable, but to the timing of a non-linguistic sound preceding any talk. Example 5 illustrates such a novel finding.

Example 5. Late, non-verbal on-beat entry

pike	π	π	π					
interval		0.329	0.344					
L	**hOw did your**	**mEEting**	**gO;**					
pike					π	π	π	
interval				0.632	0.796	0.974	0.474	
R				**CLICK**	**It (0.5)**	**wEnt all**	**rIGHt (0.5)**	
pulse			^	^	^			

The *Q* has πs with an interval of about 0.33 s, setting up a pulse. After a silent beat, R produces an on-beat click; but the *A* πs do not maintain the pulse. The first TCU of the *A* (shown) is a low-key assessment which prefaces a longer telling by R. Sequentially, the relation of the *A* to the *Q* is complex, which is reflected in the loss of the pulse.

More generally, though π_1 of the *A* may not be aligned rhythmically with *Q*, the start of an *A* is often prefaced with on-beat pre-turn material such as in-breaths, clicks, *um*, etc. These project 'incipient speakership' [13] without yet taking a turn. They display an orientation to the temporal and rhythmic structure established in the *Q*, even when the turn would be classed as non-rhythmic in [3]'s terms because the first *A* π is not on-beat. Such cases suggest rhythmic entrainment of embodied processes.

3.3.2 Non-rhythmic entry into the turn space

Example 6. Late, non-rhythmic (off-beat) entry

pike		π	π					
interval			0.259					
R	**was it nIce**		**fOOd**					
pike							π	π
interval							1.16	0.64
L							**yEAh it was quite**	**pOsh**
pulse				^	^	^	^	

As with non-rhythmic entry into the turn space tend to convey dispreferred actions and to be sequentially more complex, e.g. by correcting a presupposition of the *Q* across more than one TCU. In Example 6, the assessment in *A* is indirectly about the quality of the food; the rest of the *A* (not shown) is about the restaurant. The first π of the *A* comes in late, and not on beat. The *A* continues across more TCUs.

Examples 5 and 6, with π_1 of *A* not rhythmically aligned with *Q*, display alignment (*A* treats *Q* as a question) but not affiliation (*A* treats *Q* as in some way problematic) [24]. But we class Example 5 as rhythmic because it displays embodied entrainment.

4. DISCUSSION

The work described here shows 1) that pikes provide a unified account of rhythmicity across modalities, and 2) that rhythmicity is not a feature that English 'has' or 'does not have'. Isochronous rhythm is certainly possible in English conversation, where participants produce stretches of talk with clear rhythmical beats. But the analysis suggests that rhythmicity is a *locally* available resource which handles the contingencies of interacting in time and facilitates turn-taking. Mostly, this is done through speech; but we show evidence that gesture and non-verbal sounds preparatory to speech (like in-breaths and clicks) can also work this way. This suggests, then, that timing is sensitive to interactional function and sequential position, i.e. it is not a monolithic and single system, but rather something which occurs meaningfully and systematically over short stretches of speech. It is self-evidently also embodied and not specific to conversation, being fundamental to joint music-making [2] and no doubt other types of cooperative action, from dancing to joint use of tools like saws. Outside of music, rhythmicity is often not relevant: in conversation, it is critical at some points in interaction, but less significant at other points.

These data offer new support for the view that reduced temporal variability facilitates joint action [27]. They accord with neuroscientific evidence that local phase adjustments enhance periodicity during increased attention [9, 10, 18] and that brain activity synchronizes during social interaction [6, 23]. Three implications are: 1) musical aspects of speech allow successful conduct of conversation—the words *per se* may be less crucial except to aid prediction of rhythm [14]; 2) so temporal entrainment may serve a general function in human communication; 3) the acoustic cues that allow people to predict each other's behaviour, and to coordinate their actions over extended time periods, are relatively local, and possibly confined to phrase endings and beginnings where attention to the interaction itself is critical.

An interactional account shows that rhythmicity is not simply a 'private' matter for individuals; it is a shared resource for interactants, who can generate a pulse which is used to synchronise activities, and can be either followed or broken with social and interactional consequences. This speaks to the need for a grammar which is dynamic and which is a shared resource between participants: built not so much on a speech chain model, as on a model of socially shared cognition [cf. 15, 16, 19, 31].

5. REFERENCES

[1] Arvaniti, A. 2012. The usefulness of metrics in the quantification of speech rhythm. *J Phon* 40, 351-373.

[2] Clayton, M., Sager, R., Will, U. 2005. In time with the music: The concept of entrainment and its significance for ethnomusicology. *ESEM counterpoint* 1, 1-82.

[3] Couper-Kuhlen, E. 1993. *English Speech Rhythm. Form and Function in Everyday Verbal Interaction*. Amsterdam: Benjamins.

[4] Gordon, J.W. 1987. The perceptual attack time of musical tones. *JASA* 82, 88-105.

[5] Gordon, P.C. 1997. Coherence masking protection in speech sounds: The role of formant synchrony. *Percept Psychophys* 59, 232-242.

[6] Hasson, U., Ghazanafar, A.A., Galantucci, B., *et al.* 2012. Brain-to-brain coupling: a mechanism for creating and sharing a social world. *TICS* 16, 114-121.

[7] Hawkins, S., Cross, I., Ogden, R. 2013. Communicative interaction in spontaneous music and speech. In: Orwin M., Howes C., Kempson R. (eds), *Language, Music and Interaction.* London: College Publications, 285-329.

[8] Jefferson, G. 1986. Notes on "latency" in overlap onset. *Human Studies* 9, 153-183.

[9] Large, E.W., Jones, M.R. 1999. The dynamics of attending: How people track time-varying events. *Psych Rev* 106, 119-159.

[10] Lindenberger, U., Li, S.-C., Gruber, W., Müller, V. 2009. Brains swinging in concert: cortical phase synchronization while playing guitar. *BMC Neuro* 10, 22.

[11] Loehr, D. 2007. Aspects of rhythm in gesture and speech. *Gesture* 7, 179-214.

[12] Loehr, D. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Lab Phon* 3, 71-89.

[13] Ogden, R. 2013. Clicks and percussives in English conversation. *JIPA* 43, 299-320.

[14] Peelle, J.E., Gross, J., Davis, M.H. 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cort* 23, 1378-1387.

[15] Pickering, M.J., Garrod, S. 2004. Toward a mechanistic psychology of dialogue. *BBS* 27, 169-190.

[16] Pickering, M.J., Garrod, S. 2013. An integrated theory of language production and comprehension. *BBS* 36, 329-392.

[17] Raymond, G. 2003. Grammar and social organisation: Yes/no interrogatives and the structure of responding. *Am Sociol Rev* 68, 939-967.

[18] Sänger, J., Müller, V., Lindenberger, U. 2012. Intra- and interbrain synchronization and network properties when playing guitar in duets. *Front Hum Neuro* 6.

[19] Schegloff, E.A. 1991. Conversation analysis and socially shared cognition. In: Resnick L.B., Levine J.M., Teasley S.D. (eds), *Perspectives on Socially Shared Cognition.* Washington DC: American Psychological Association, 150–171.

[20] Schegloff, E.A. 2007. *Sequence Organisation in Interaction*. Cambridge: Cambridge University Press.

[21] Selting, M., Auer, P., Barth-Weingarten, D., *et al.* 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächs - Online-Zeitschr, Z Verb. Interakt* 10, 353-402.

[22] Sidnell, J. 2010. *Conversation Analysis. An Introduction*. Chichester: Wiley-Blackwell.

[23] Stephens, G.J., Silbert, L.J., Hasson, U. 2010. Speaker-listener neural coupling underlies successful communication. *PNAS* 107, 14425-14430.

[24] Stivers, T. 2008. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Res Lang Soc Interact* 41, 31-57.

[25] Stivers, T., Enfield, N.J., Brown, P., *et al.* 2009. Universals and cultural variation in turn-taking in conversation. *PNAS* 106, 10587-10592.

[26] Stivers, T. 2010. An overview of the question–response system in American English conversation. *J Pragmatics* 42, 2772–2781.

[27] Vesper, C., van der Wel, R.P.R.D., Knoblich, G., Sebanz, N. 2011. Making oneself predictable: Reduced temporal variability facilitates joint action coordination. *Exp Brain Res* 211, 517-530.

[28] Villing, R.C., Repp, B.H., Ward, T.E., Timoney, J.M. 2011. Measuring perceptual centers using the phase correction response. *Atten Percept Psychophys* 73, 1614-1629.

[29] Vos, J., Rasch, R. 1981. The perceptual onset of musical tones. *Percept Psychophys* 29, 323-335.

[30] Vos, P.G., Mates, J., van Kruysbergen, N.W. 1995. The perceptual centre of a stimulus as the cue for synchronization to a metronome: Evidence from asynchronies. *QJEP: HEP* 48A, 1024-1040.

[31] Wilson, M., Wilson, T.P. 2005. An oscillator model of the timing of turn-taking. *Psychon Bull Rev* 12, 957-968.

Supported by British Academy Grant SG120400 and a Cambridge University Newton Trust Small Grant.