INTERACTION PHONOLOGY – RHYTHMIC CO-ORDINATION AS SCAFFOLD FOR COMMUNICATIVE ALIGNMENT

Petra Wagner

Bielefeld University, Bielefeld, Germany

petra.wagner@uni-bielefeld.de

1. INTRODUCTION

Interaction Phonology (Wagner et al., 2013) postulates a process of rhythmic co-ordination based on entrainment processes which provide the temporal scaffold for higher order adaptation among interlocutors in critical situations, and hence, improves communication. 10 years after the publication of our framework, the time is more than ripe for its first evaluation and a thorough re-assessment. To achieve this, I will first give an overview of the general assumptions and motivations underlying Interaction Phonology, and then describe its mechanism as a logistic, attention-guiding component in a model of speech processing in interaction. I will then derive a set of model predictions, and evaluate them based on a thorough review of more recent empirical studies. In a last step, I will slightly modify our original model of Interaction Phonology (cf. Figure 1, for an overview of the original model; cf. Figure 2, for the adapted version), and list desiderata for its further testing in the future.

2. A SKETCH OF INTERACTION PHONOLOGY

When two or more people communicate, they agree on a shared language system, with the ultimate goal to enable a common understanding with the help of an interactionally grounded, shared symbolic representation. However, assumptions about the shared symbol inventory may differ. For instance, whether you refer to certain vegetables as 'potatoes', 'spuds', 'solanum tuberosum', or 'root vegetables', may depend on your individual assessment of the situation, individual preference, spoken variety or linguistic context. It is likely that speakers will therefore negotiate the conditions of usage of a particular term, to clarify reference, or to signal mutual cooperativeness and perspective taking in a process called grounding (Clarke and Brennan, 1991). During this process, it is not sufficient to agree on a shared inventory of symbols and grammatical constraints (e.g., "English"), because the way that abstract symbols are realized in the speech signal may differ, due to different speaking styles, varieties, registers, or external factors such as cognitive distraction or various types of 'noise'. For this reason, sub-symbolic phonetic convergence has been claimed to be closely linked to the phenomenon of symbolic alignment, i.e., the tendency of interlocutors to agree on a shared or similar inventory (Pickering and Garrod, 2004).

So, agreeing on speaking the same 'language' has something in common with two people agreeing on dancing a waltz. While the dance move sequences that qualify as 'symbolic' figures of a waltz may be clear to both dancers, the velocity, amplitude and detail of the pertaining movement trajectories need

to be precisely negotiated, helped by an external pacemaker in the form of the rhythm of the accompanying music. In speech-based communication, it is likewise not sufficient to agree on an abstract set of phonemes, lexemes, and syntactic structures. Rather, speakers need to agree on a fine-grained execution of the shared movement patterns within their individual motor systems, to achieve pronunciations that are mutually understood, e.g., similar to the relative timing of articulators as expressed within Articulatory Phonology (Browman & Goldstein, 1992).

So far, researchers have accumulated plenty of evidence for sub-symbolic co-ordination processes taking place between interlocutors: Speakers align their pronunciation patterns, speech tempo and prosody (Bosch et al., 2005; Gessinger et al., 2021; Levitan and Hirschberg, 2011; Pardo, 2006, Lewandowski, 2011, inter alia), and occasionally even their conversational laughter (Ludusan and Wagner, 2022), but also on higher-order levels of linguistic organization such as lexical choice, syntactic structures, or referential gestures (Brennan and Clarke, 1996; Bergmann and Kopp, 2012, inter alia). However, most studies find a lot of individual variation both in rhythmic-prosodic entrainment and higher-level linguistic alignment. Still, a key assumption of mechanistic accounts of interpersonal alignment (Pickering and Garrod, 2004; Pickering and Garrod, 2007) is that sub-symbolic, rhythmic-prosodic entrainment fosters symbolic alignment and hence, comprehension, on higher levels of grammatic organization.

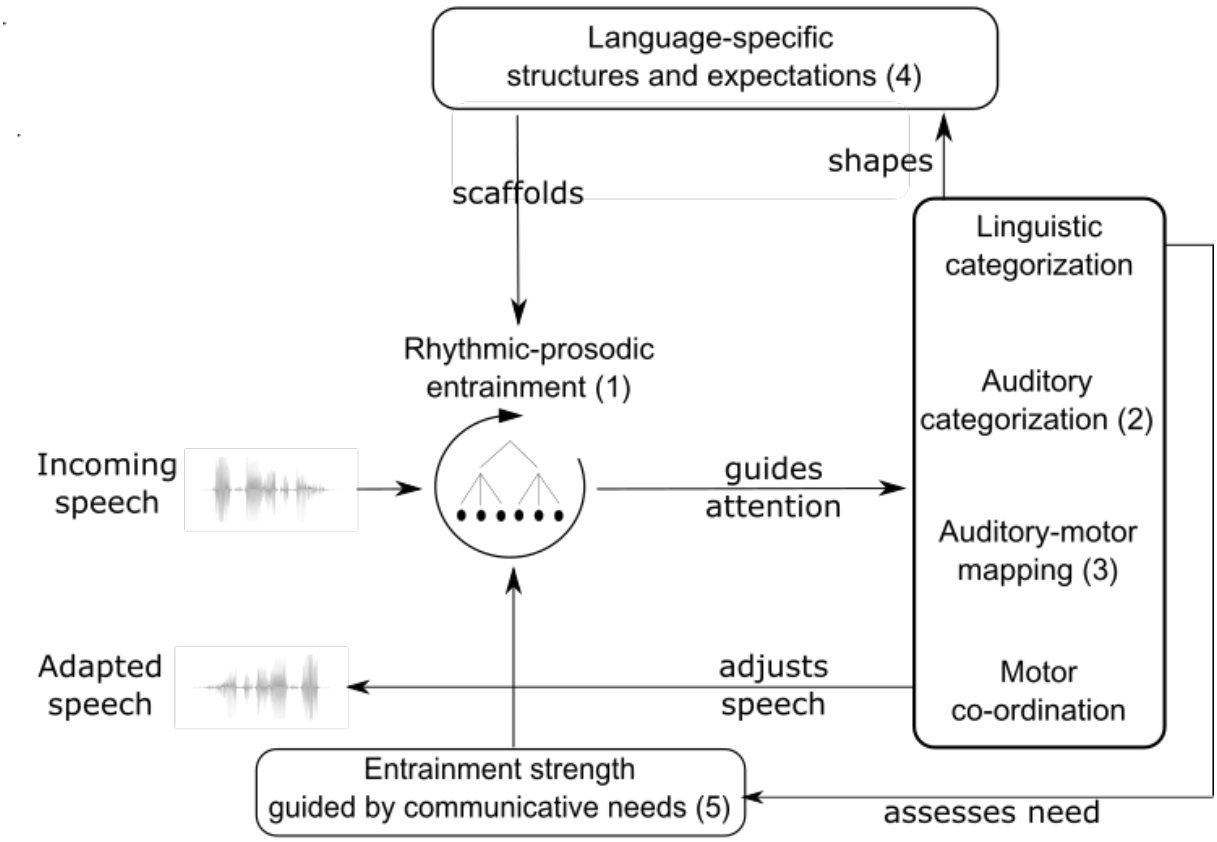

FIGURE 1. AN OVERVIEW OF THE PROCESSES AND STRUCTURES INVOLVED IN INTERACTION PHONOLOGY

The diagram depicts the processes in a listener, who entrains to the rhythmic patterns of speech, based on the expectations inherent in their language competence. The level of rhythmic prosodic entrainment can be strengthened in difficult communicative situations. That way, the listeners' attention is guided to higher order linguistic aspects connected to the rhythmic structures thus enhanced. This attentional process may alter the way that rhythmic-prosodic structures are connected to higher order linguistic patterns, but also intensify the level of entrainment with an interlocutor. Taken together, these processes are expected to aid mutual understanding, particularly in 'difficult' situations. The model relies on a set of modules, some of which are part of the speaker's grammar. These encompass (1) an entrainment module, (2) an auditory analysis guided by it, which is also linked to (3) motor patterns, which automatically lead to convergence in speech production as an automatic bi-product of entrainment, a set of (4) linguistic structures and expectations as part of a speaker's grammar, which are linked to the levels of entrainment via their corresponding levels of prosodic organization, and (5) a monitoring of communication relevance, which estimates the need for entrainment (informed by the auditory and linguistic analysis), and adjusts the level of entrainment by modulating the coupling strength.

To this day, speech processing architectures lack a unified account of whether and how any interaction between sub-symbolic and symbolic adaptation is actually achieved. In Wagner et al. (2013), we therefore argued for Interaction Phonology as a logistic, attention-guiding component that enables interlocutors to co-ordinate their articulatory movements on a low signal-level by a process of temporal entrainment. That way, listeners may guide their attention to crucial aspects of phonetic detail (Ghitza and Greenberg, 2009; Ghitza, 2012; Giraud and Poeppel, 2012, Chapters 22 and 23, this volume) that will permit an easier access to higher-level linguistic information. As a consequence, symbolic alignment should be fostered by temporal co-ordination in an automatic, bottom-up fashion. We use the term entrainment in a narrow sense (Obleser and Kayser, 2019), where it describes a dynamic process of physically coupled oscillatory systems, which adapt their cycles both in period and phase, thereby ultimately achieving a fixed phase relationship. Humans are capable of interpersonal entrainment without an external, isochronous pacemaker, e.g., when spontaneously synchronizing their clapping behavior in enthusiastic applause by period doubling (Néda et al., 2000), or when speaking in synchrony (Cummins, 2009). Strikingly, humans have shown to synchronize their brain activities, strengthened by shared engagement and joint activity (Dikker et al., 2021).

In Interaction Phonology, the rhythmic properties of a language play a crucial role in this entrainment process. It has been noted that speech lacks the isochrony or regularity necessary for entrainment (Cummins, 2012). However, it may occasionally show fixed phase relationships or a high degree of regularity, at least in highly formalized speaking styles such as poetry (Wagner, 2012), which may lend itself to rhythmic entrainment, even though we do not yet understand the exact mechanism behind this. While absolute co-ordination cannot be meaningfully expected between interlocutors at all times, there is some evidence in favor of entrainment: Across several languages, overlapping speech shows a preference for speakers being in phase with the interlocutor's syllabic speech stream (Włodarczak et al. 2012). In line with entrainment models of attention (Lakatos et al., 2008; Large and Jones, 1999), Interaction Phonology postulates that this process helps listeners gain access to language-specific phonological and higher-level properties of the utterance spoken.

Furthermore, we argued that rhythmic entrainment is a necessary prerequisite for the automaticity and swiftness of representational alignment in human interaction. While not excluding the possibility of a reductionist account of the phenomena described, we do not think it necessary for now to subscribe to this idea. Still, we argued for an inter-speaker co-ordination mechanism as being fundamental not only for speech perception, but for communicative interaction, i.e., the permanent

active attuning to one another. Interaction Phonology can be preliminarily defined as taking care of the co-ordinative interactive processes that are strongly built on rhythmic-phonological structures.

Interaction Phonology furthermore postulates that there are universal and language-specific structures, on which co-ordination takes place. In particular, it predicts that the rhythmic-prosodic organization of a language constrains the levels of temporal co-ordination between interlocutors. For a lack of better knowledge, these are assumed to be identical to the language-specific levels of prosodic organization (syllables, feet, prosodic phrases) and their internal metrical organization (Jun, 2005). In other words, according to Interaction Phonology, the temporal co-ordination between interlocutors who speak varieties with a similar rhythmic-prosodic organization should be comparatively easy. However, Interaction Phonology also postulates that the mechanisms of temporal co-ordination are to some degree universal, based on syllabic structures that are grouped into larger units such as phrases or similar. Even though their regularity, function and organization within the prosodic hierarchy may differ across languages, there is some space for rhythmic co-ordination even when interlocutors cannot rely on a large set of common temporal mechanisms that may serve as anchors to higher-level linguistic organization. An example would be an L2 listener's strategic reliance on prosodic universals as well as language-specific prosodic cues as indicators of lexical stress, which often is a useful approach to segment a speech stream into words (Endress & Hauser, 2010; Tyler & Cutler, 2009; Ordin & Nespor, 2013). The idea of rhythmic entrainment as a universal co-ordinative process underlying linguistic organization has received further support by developmental studies that described movement synchronization between neonates and their caregivers (Condon, 1974; Jaffe et al., 2001), where a baby's acquisition may be helped by anchoring into the universal prosodic properties of speech, to pave the way for higher-order symbolic alignment (Chapters 34, 35, 36, 38, 40, this volume). In fact, neonates are born with an ability to use prosodic strategies independently of segmental properties to identify word boundaries in their early language acquisition process (Fló et al., 2019). An early alignment to the rhythmic prosodic detail of a caregiver's movements may therefore be a generally useful strategy in language acquisition. However, as prosodic and phonetic alignment has shown to be to some degree voluntary, situation specific, and is less strong in populations with Autism Spectrum Disorder (Schweitzer and Lewandowski, 2014; Schweitzer et al., 2017; Wynn et al., 2018), Interaction Phonology allows for the modulation of underlying entrainment processes. That is, if conversational needs for mutual understanding and grounding are high, it predicts that entrainment can be willingly strengthened, thereby actively supporting mutual comprehension and conversational grounding.

3. The Mechanism of Interaction Phonology

Interaction Phonology postulates that the incoming speech signal is subject to a process of rhythmic-prosodic analysis that

- guides the listener's attention to the fine phonetic detail of the speech signal that may be of particular relevance for a given language, which coincide with crucial boundaries of higher-level linguistic organization, and therefore facilitate their prompt identification. For now, we believe that the levels of entrainment are identical to the levels of organization in the prosodic hierarchies of the different languages. It is possible, that this language-specific co-ordination does not constitute an independent level of a language's grammar, but rather is a by-product of its morphosyntactic or phonological organization.

- is driven by a process of rhythmic entrainment, modulated according to communicative needs such as overall the level of 'noise', and informed by linguistic analyses of the ongoing interaction. Apart from objectively present external noise, this may also relate to the overall level of distraction, or the relevance of successful communication.
- is adaptive with respect to its level of entrainment, or coupling strength; these adaptations are strongly guided by the rhythmic-prosodic patterns of the language chosen to communicate, but may also be subject to long-term entrainment between interlocutors, if these (initially) speak different languages or varieties.
- leads to an adaptation in speech production with respect to tempo and rhythmic modulation via perception-production coupling, and hence, an improved attention to detail on the listener's side and representational alignment in (adapted) speech production.

These analyses are organized within various model components, which are described in detail in Table 1, and are indicated by their respective numbers in Figure 1.

TABLE 1. MAIN STRUCTURES AND PROCESSES OF INTERACTION PHONOLOGY.

Model component	*Description*
1	Entrainment module, guiding listener's attention to points in the incoming speech signal which are crucially linked to higher-order linguistic organization.
2	Entrained, or "guided" auditory analysis of incoming speech input, which interfaces with subsequent linguistic analysis of input
3	Motor patterns mapped to incoming acoustic analysis, automatically leading to adapted speech output
4	Set of linguistic structures and expectations as part of a speaker's grammar, which are linked to the levels of entrainment via their corresponding levels of prosodic organization, and which correspond to attractors in entrainment
5	A communication relevance monitor, which assesses the situation and ongoing communication (and with this, the need for entrainment), which may adapt the strength of necessary entrainment depending on present noise and the necessity of communicative success

So far, Interaction Phonology has not yet a spelled out connection or interface with existing models of speech production and perception. However, most of these models miss a link between symbolic and sub-symbolic processing, and Interaction Phonology may help improving our understanding of this interface. Given its focus on communication, Interaction Phonology can only be meaningfully integrated with architectures that account for both perception and production.

4. PREDICTIONS OF INTERACTION PHONOLOGY

Here, we spell out a set of testable predictions by Interaction Phonology. The predictions are chosen as they all test crucial aspects of the model. Interaction Phonology makes predictions beyond this list, especially with regards to prosodic universals and language-specific constraints. Also given its current lack of formality and underspecification, it should be clear that this list is currently incomplete and lacks formal rigor.

Prediction 1: Speech rate adaptation should improve speech perception in similar communication settings.

Prediction 2: Entrainment should be visible across the levels of the prosodic hierarchy, in a language-specific fashion.

Prediction 3: The level of entrainment should be situation-specific, and vary within individuals across different situations.

Prediction 4: If rhythmic entrainment occurs, it should automatically result in symbolic alignment.

Prediction 1 falls out of the model, as the model postulates a positive effect of entrainment on speech perception by its guiding the listener's attention to relevant phonetic detail using the entrainment module (cf. Figure 1: component 1). However, it needs to be taken into account that the model also predicts entrainment for those communicative situations in which perception may be impeded by various types of noise. Therefore, it is crucial for testing Prediction 1 that speech perception and entrainment are measured across similar settings, without added cognitive load or external noise. Speech rate is chosen mostly as a test (in favor of other potential features of rhythmic-prosodic entrainment) as there exist a considerable amount of empirical research on it. Prediction 2 falls out of the assumed link between levels of entrainment and language-specific structures (cf., Figure 1: component 4). That is, Interaction Phonology expects a certain language-dependence with respect to the levels of entrainment that mirror the prosodic organization of the involved languages or varieties. Going back to speech rate entrainment, depending on the rhythmic-prosodic structure of the language to be entrained to, speech rate adaptation may concentrate on morae, syllables, prosodic feet, prosodic words, or even phrasal structures. Prediction 3 is derived from the Interaction Phonology's assumption that entrainment is to some degree deliberate, and strategically chosen by interlocutors rather than a fully automatized process that will always be enabled (Figure 1: component 5). In other words, Interaction Phonology predicts the level of entrainment to a certain speech rate to be stronger in challenging communicative situations. Prediction 4 falls out of the assumed automatic link between sub-symbolic co-ordination and symbolic alignment (cf., Figure 1: connection between components 1 and 2). Here, the control mechanism that enables entrainment, automatically takes into account higher-level similarities. If these two fail to be coupled, this would be a challenge for our control mechanism, and would point to a strongly strategic symbolic alignment which is not necessarily coupled to sub-symbolic, motor-level processes of articulation. In other words, an entrainment to speech rate ought to be also visible in the usage of more similar words, or syntagmatic structures.

5. Evaluating Interaction Phonology

Next, it will be determined whether more recent empirical research is in line with the assumptions and predictions of Interaction Phonology, or falsifies (aspects of) it. Where no research results lend themselves to model evaluation, suggestions for future studies will be made, in order to better understand Interaction Phonology's flaws, limitations as well as strengths. The analysis will concentrate on the 4 main predictions of Interaction Phonology that has been spelled out above.

5.1. Prediction 1: Speech Rate Adaptation helps Speech Perception

In incremental, online speech perception, listeners need to simultaneously pay attention to several levels of linguistic organization. The ability to do this may be enhanced by the different time scales

underlying the spell-out of these levels (phones, syllables, words, phrases), which can be entrained to cortical rhythms working on similar time scales (Ghitza and Greenberg, 2009; Ghitza, 2012; Chapter 2, this volume). Much work around rhythmic entrainment during perception has concentrated on attentional selection, which ought to focus on crucial parts of the speech signal, e.g., the initializations of syllables. There is converging evidence that some form of temporal entrainment indeed helps selectively attending to the incoming speech stream of a particular speaker among several concurrent speakers (Obleser and Kayser, 2019). Also, neural entrainment processes have shown to (somewhat) aid speech perception and sentence comprehension (Lamekina and Meyer, 2022; Riecke et al., 2019; Wilsch et al., 2018; Zoefel et al., 2018).

However, as speech tempos change dynamically in ongoing speech within the same speaker (Quené, 2008), and speech is not isochronous like music (Cummins, 2012), for entrainment to be a successful tool for enhancing speech perception, listeners need to be able to swiftly adapt to these speech tempo changes. Speech rate convergence in production is a phenomenon largely supported by empirical research, appearing in both in monological priming tasks (Jungers and Hupp, 2009) and conversations (Cohen-Priva et al., 2017; Schultz et al., 2016; Fuscone et al., 2021). For perception, Dilley and Pitt (2010) presented first evidence for listeners' indeed quickly adapting to the speech tempo of an incoming speech signal, leading them to perceptually insert additional syllables/words into a speech stream that was locally produced slowly, e.g., "leisure time" was perceived as "leisure or time". This effect, which they term LRE (lexical rate effect) is restricted to speech processing and does not generalize to tone perception (Pitt et al., 2016), but can be built up over longer stretches of time, thereby generating the expectations that drive selective attention (Baese-Berk et al., 2014, chapter 23, this volume). Bosker (2017) showed in a series of experiments that it is the (isochronous) speech rate prior to a target that creates an anticipatory effect on perception. He interprets this as evidence for an underlying neural entrainment mechanism at play, which is not tied to the speech mode. What is not yet resolved is the question of whether entrainment is restricted to speech processing, or is a general monitoring and adaptation device. The studies reported here that have examined an impact of rhythmic entrainment on speech perception have done so in highly controlled laboratory settings. Thus, it can at least be said that in such contexts, an adaptation to speech tempo can be traced and appears to have a positive impact on speech perception. However, it still remains unclear how entrainment can actually be achieved given the intrinsic non-isochrony present in speech signals. Here, Bosker (2017), Obleser and Kayser (2019) and Meyer et al. (2020) claim that the – at best – pseudo-rhythmic acoustic properties of speech are sufficient to induce an entrainment mechanism which may lend itself to higher-order synchronicities in more abstract linguistic domains. For the moment, one can only speculate that highly adaptive (neural) oscillators with a fast reset should also be able to achieve a rapid entrainment to dynamically changing rhythms (Inden et al., 2012).

5.2. Prediction 2: Entrainment Should Be Visible across Language-Specific Levels of the Prosodic Hierarchy.

Building on the conspicuous similarities between the multi-timescales of cortical and speech rhythms (Ghitza and Greenberg, 2009; Ghitza, 2012), Interaction Phonology postulates that rhythmic entrainment should pertain to various time scales, and these time scales should reflect the rhythmic structure of the language(s) spoken. In particular, this should lead to language-specific rhythmic entrainment, as languages, language varieties or speaking styles differ with respect to their prosodic organization, and this should be reflected in the long-term abilities of entrainment. For example, languages may differ vastly with respect to the length and complexity of syllable-sized units (Zec,

2007). Interaction Phonology now predicts that speakers of languages with a higher degree of phonotactic complexity and variability are either more flexible in entraining to syllable streams, or maybe make less use of syllable-level entrainment, as it is more often doomed to fail. Also, languages differ with respect to their higher order prosodic organization, and may use different patterns of metrical organization (Jun, 2005). Such differences should also show in language-selective rhythmic entrainment.

Currently, empirical evidence indeed points towards rhythmic entrainment being active on different time scales: the LRE (see above) has been shown to also apply for syllable-level speaking rate as well as rhythm, indicating a certain degree of higher order entrainment on the foot or word level, where listeners modulate their perception based on whether they expect a stressed or unstressed syllable (Morrill et al., 2014). Furthermore, the effect has shown to be additive, and listeners are more attentive when several rhythmic boundaries co-occur. However, despite considerable work on entrainment to pulse and higher-order meter in music (cf. the overview in Fitch, 2013), and despite a long tradition in research to hypothesize about similar processing mechanisms being at play in music and speech processing and organization (e.g., Lehrdahl and Jackendoff, 1983; Wagner, 2008; chapters 9, 31–33, this volume), very little is actually known about language specific entrainment, and most evidence remains speculative. Some similarities between music and speech perception can be drawn from finger tapping studies, a sensorimotor synchronization paradigm that is well-established in music rhythm perception research (Repp, 2005; Repp and Su, 2013). Finger tapping to music rhythms has been shown to help improve music time perception, similar to the perceptual benefits of entraining to speech rhythm (Manning and Schutz, 2013). In speech perception tasks, finger tapping duration and intensity has likewise been shown to be sensitive to rhythmic structure linked to linguistic organization such as syllable onsets, lexical stress, sentence stress or accent (Parrell et al., 2014, Rathcke et al., 2021). Another paradigm called speech cycling investigated rhythmic entrainment of repetitive short phrases to external high and low tones, and found cross-linguistic similarities in patterning speech to these external tones for Japanese and English, despite their different syllable structures (Tajima and Port, 2003). While these studies point to a common sensorimotor entrainment mechanism, it should be noted that tapping in real-time to an incoming speech signal is extremely difficult to do, and listening to or reproducing repetitive single phrases resembles music rather than speech processing (Anbari et al., 2013). An alternative methodological paradigm, in which listeners tapped a perceived rhythmic structure directly after perceiving an utterance, revealed an ability of listeners to encode language-specific rhythmic-prosodic features in tapping patterns as well, and showed a stronger reliance on acoustic-prosodic features as compared to non-motor prosody perception tasks (Wagner et al., 2019, Bruggemann et al., 2022). However, it is yet unclear whether the results of tapping are indicative of sensorimotor entrainment proper, or simply fall out of a general analysis of linguistic structure, integrating linguistic, acoustic-phonetic and sensorimotor cues. Similar problems exist with studies on L2-acquisition which show that rhythmic priming has a beneficial effect on learning the target prosody in an L2, as they either rely on multimodal reproduction tasks, or use musical (rather than speech) rhythms as priming materials (e.g., Baills and Prieto, 2021, Wang et al., 2016) – neither is conclusive as to whether it really is rhythmic entrainment to speech that leads to the positive effects on mastering an L2 prosody. Overall, we have some empirical evidence pointing into the direction that rhythmic entrainment has long-term consequences, leading to long-term rhythmic expectations which result in an improved rhythmic entrainment performance in an L1 as compared to an L2, and which may result in better abilities of learning an L2 prosody in speakers with a high degree in rhythmic training. However, clear-cut evidence for this prediction of Interaction Phonology is still lacking.

5.3. Prediction 3: The Level of Entrainment Should Be Situation-Specific, and Vary Within Individuals across Different Situations

Interaction Phonology postulates that interlocutors make a deliberate (though not necessarily conscious) choice in whether they employ prosodic entrainment, or not, and it is expected that entrainment should be selectively activated in challenging communicative situations, in which the benefits of attention can be exploited best. As the vast majority of studies have been performed in laboratory conditions, often relying on short, highly controlled utterances that show little resemblance with everyday interactions, this has not been investigated in ecologically valid conditions. However, some first approaches do exist:

In a study that looked at rate-dependent adaptive listening in quiet and noisy conditions, Reinisch and Busker (2022) could show that listeners dynamically adapt at a low-level to challenging contexts, and can identify noisy target items more easily when these are preceded by coherently noisy signals. There is also increasing evidence that the selective attention to an individual speaker in a multi-party listening condition decreases entrainment to the ignored voices (e.g., Ding and Simon, 2013; Fuglsang et al., 2017). This points towards the level of entrainment being to some degree adjustable according to situation-specific needs.

Several studies investigated the impact of acoustic manipulation (vocoded speech) on the level of entrainment, hypothesizing that vocoding would be detrimental to speech quality and therefore trigger higher entrainment. Peele et al. (2012, 2013) find evidence for neural entrainment to speech being higher when it is vocoded (more difficult to comprehend). While this may point into a direction of selective entrainment in the case of speech that is difficult to process, Baltzell et al. (2017) showed that vocoded speech preceded by natural speech primes also aided the comprehension of vocoded speech. This is in line with findings on synchronous speech, where synchronization is not influenced by intelligibility, but by rhythmic cues (Cummins, 2009).

Rather than manipulating the acoustics of their stimuli, Song and Iverson (2018) and Iverson et al. (2018) tested the influence of overall intelligibility on neural entrainment by comparing the performance of L1 and L2 listeners when hearing L1 or L2 speech. Their results point to patterns of stronger neural entrainment when listening to the less familiar (L1 for L2 listeners, L2 for L1 listeners), and hence, more challenging variety. However, the idea that any challenges to the ongoing communication success lead to an automatic increase in entrainment appears to be overly simplistic: a study by Hjortkjær et al. (2020) indicates that a higher working memory load actively decreases the level of neural entrainment, and also Abel and Babel (2017) show lower phonetic convergence under high cognitive load. Interestingly, this effect was present both for a more difficult task as well as an increase in acoustic noise that had been tested for not being detrimental to speech intelligibility. These results indicate that entrainment needs cognitive resources by itself, possibly to uphold selective attention.

5.4. Prediction 4: If Entrainment Occurs, it Should Automatically Result in Symbolic Alignment Between Interlocutors

By now, there is a long research tradition that has demonstrated adaptation between communication partners both on fine phonetic detail such as speech tempo, pause duration, intonation or segmental articulation as well as more abstract linguistic representations such as the lexicon, syntactic structures, or referential gestures (cf. 2). What is unclear is whether low-level co-ordination on fine phonetic

details indeed quasi-automatically triggers an agreement on higher-order linguistic concepts, as predicted by Interaction Phonology, and in line with models of interpersonal adaptation that link production and perception (Pickering and Garrod, 2004; 2007). However, clear-cut evidence for this idea appears to be difficult to come by, despite the undeniable benefit found for listening entrainment in speech perception (see above). Krivokapic (2013) suggest that speech rate convergence between dialogue partners correlates with their alignment of variety-specific rhythmic patterns (Indian English and American English), indicating a certain automaticity in convergence across low-level and higher-level rhythmic prosodic organization. Alternatively, this could be explained by falling out of inter-speaker entrainment in speech rate, as duration indicates both speech rate as well as rhythmic organization. One of the few studies that found evidence for a communicative benefit (beyond intelligibility) of speech rate adaptation is Manson et al. (2013), who reported an increase in cooperation between interlocutors when they also aligned in speech rate. Similarly, Lubold and Pon-Barry (2014) report an increase in perceived rapport. These results may point to a higher degree of conversational grounding in situations where rhythmic-prosodic entrainment is evident, and may indicate a mechanistic link between low-level speech rate adaptation to higher-order linguistic processing. However, other interpersonal factors such as mutual likeability were not affected by an increase in entrainment (Manson et al., 2013), which is further evidence that the underlying mechanism may be specialized to linguistic processing.

However, a clear-cut effect of entrainment on symbolic-linguistic alignment appears to be difficult to prove: Weise and Levitan (2017) fail to find evidence for a link between acoustic-prosodic and symbolic alignment, while Rahimi et al. (2017) suggest that entrainment across different levels of linguistic organization can occur. Generally, most studies report a high degree of individual variation in entrainment, which seems to be at least to some degree driven by personal and interpersonal factors, e.g. mutual likability, perceived attractiveness, gender as well as the power dynamics between interlocutors (Babel, 2012; Pardo, 2012; Michalsky and Schoormann, 2017; Schweitzer and Lewandowski, 2014; Schweitzer et al., 2017; Reichel et al., 2018).

6. Conclusion: An adapted sketch of Interaction Phonology

For some key predictions of Interaction Phonology, empirical evidence is growing stronger. In particular, we see that rhythmic entrainment does take place in speech tempo adaptations, has a positive effect on intelligibility. While cross-linguistic studies on entrainment are very rare, there is evidence that it connects to the rhythmic-prosodic structure of individual languages, thereby probably also enhancing higher-level comprehension. Another key assumption of Interaction Phonology is that rhythmic-prosodic entrainment can be adjusted based on situative needs. Here, we indeed saw that listeners do adapt their entrainment to individual voices or increase entrainment in challenging listening situations. However, entrainment is not increased independently of the type of communicative challenge. Contrary to our prediction, working memory load seems to decrease entrainment, indicating that entrainment comes with a certain cognitive load of its own. Here, more research is necessary to better understand which type of situation triggers or decreases its effectiveness. Despite the positive effect entrainment has on intelligibility, when explicitly linking it to higher-level linguistic organization, there appears to be no automaticity in rhythmic-prosodic entrainment and higher-order symbolic alignment between interlocutors. At best, researchers find that this connection is not ruled out. In our account of Interaction Phonology, this connection is therefore removed for the time being, and the link between rhythmic prosodic entrainment and

symbolic linguistic organization is limited it to grammatical aspects of sound structure (phonology). From there, a connection to higher-order linguistic organization can made as part of listening comprehension, but the connection to further symbolic entrainment needs to be questioned. For now, the results leave a question mark as to the exact nature of the interface between sound-related and higher-order linguistic processing in Interaction Phonology. As to the auditory-motor mapping, which predicts automatic convergence of rhythmic-prosodic patterns in speech-based interaction, it is left as optional in the model, as most data shows high individual variation in speaker's level of converging prosodic patterns, even though it seems to be to some degree automatized for speech tempo. Here, further empirical work is needed to highlight the level of automaticity or deliberate control, and how it covaries with other speaker traits, their level of neurocognitive alignment, or situative factors. Overall, it can be concluded that the model of Interaction Phonology can still be helpful to further inform psycholinguistic models of speech processing, to extend them to models of communicative interaction, and to improve and clarify the interface of symbolic and sub-symbolic processing in the models. Our adapted sketch of Interaction Phonology is illustrated in Figure 2.

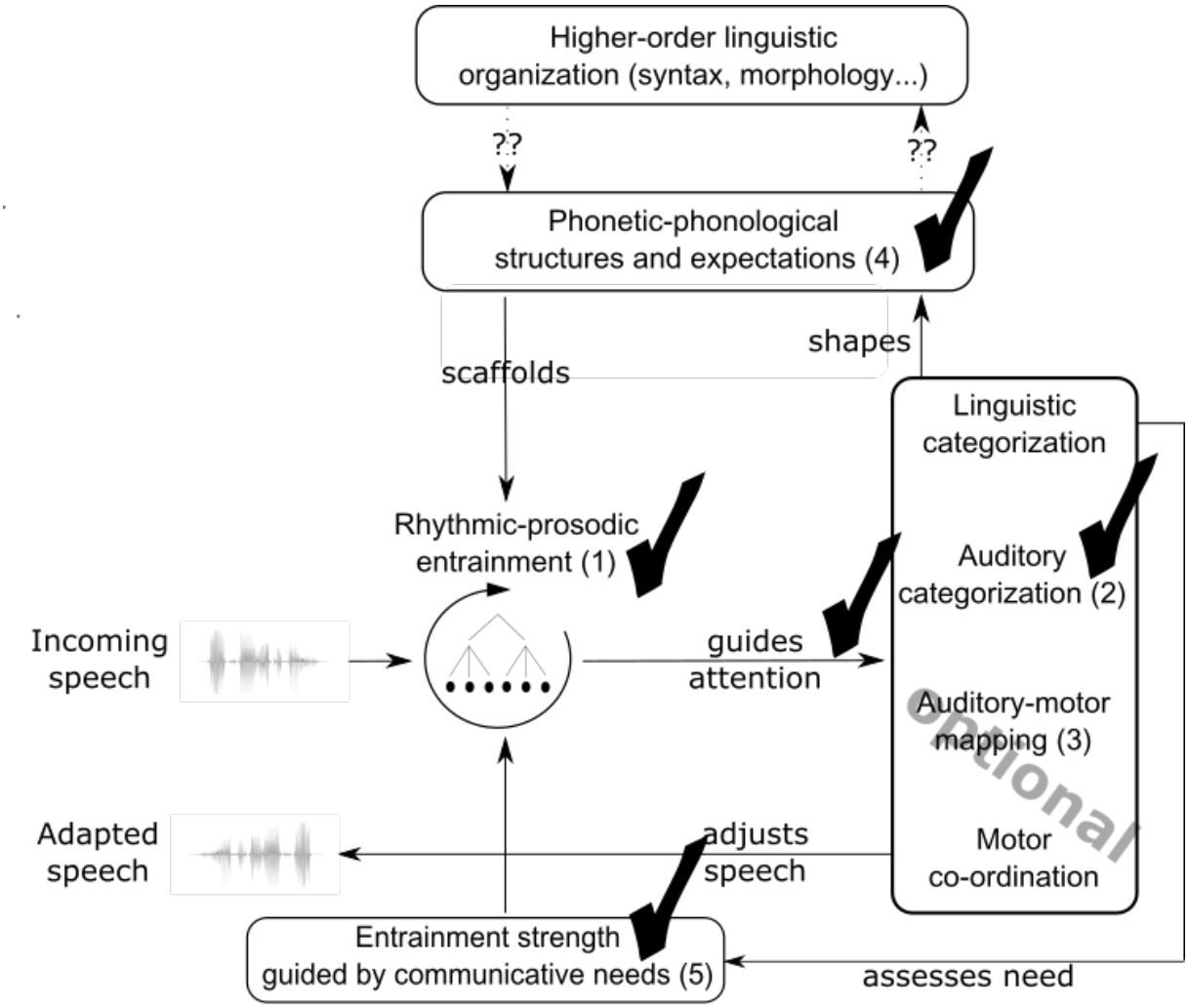

FIGURE 2. AN ADAPTED SKETCH OF INTERACTION PHONOLOGY

Those parts of Interaction Phonology that have received empirical support by are indicated by check marks. Other parts are either commented as optional (auditory-motor mapping and speech adaptation),

or have been modified/extended in line with empirical findings. In particular, the language-specific structures and expectations for which we have evidence to guide rhythmic-prosodic entrainment and to be shaped by it currently are restricted to phonetic-phonological ones. It remains unclear, whether syntactic or lexical adaptations are connected with entrainment processes likewise.

Summary

Interaction Phonology explains symbolic and sub-symbolic inter-speaker adaptations using the mechanism of rhythmic-prosodic entrainment. Many key assumptions (rhythmic entrainment as optional process that helps perception and is linked to grammar) are empirically supported. However, the original model was modified: auditory-motor mapping is optional, entrainment can also be actively decreased under high cognitive load, and the assumed automaticity between entrainment and symbolic alignment is questioned.

Implications

Interaction Phonology provides a testable theoretical framework for evaluating language-specific and language universal hypotheses related to rhythmic entrainment between interlocutors, and their relationship with higher-order alignment of abstract linguistic representations.

Gains

Interaction Phonology provides a theoretical framework that provides the necessary scaffold for enabling an inter-speaker alignment of phonetic-phonological, and potentially also higher-order linguistic representations by a mechanism of rhythmic entrainment. Interaction Phonology extends existing speech processing models with an interface between symbolic and sub-symbolic processing, and integrating them into communication models.

Index terms

speech rhythm, entrainment, interaction, speech rate, intelligibility, synchronization, alignment

7. REFERENCES

Abel, J., & Babel, M. (2017). Cognitive load reduces perceived linguistic convergence between dyads. Language and Speech, 60(3), 479-502.

Anbari, S. A., Włodarczak, M., & Wagner, P. (2013). Rhythmic constraints on read and rapped speech. In Proceedings of the 14th Rhythm Production and Perception Workshop, Birmingham.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. Journal of Phonetics, 40(1), 177–189. https://doi.org/10.1016/j.wocn.2011.09.001.

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. Psychological Science, 25(8), 1546–1553. doi:10.1177/095679761453370

Baills, F., & Prieto, P. (2021). Embodying rhythmic properties of a foreign language through hand-clapping helps children to better pronounce words. Language Teaching Research, 0(0). doi:10.1177/1362168820986716

Baltzell, L. S., Srinivasan, R., & Richards, V. M. (2017). The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. Journal of Neurophysiology, 118(6), 3144-3151.

Bergmann, K., & Kopp, S. (2012). Gestural alignment in natural dialogue. Proceedings of the Annual Meeting of the Cognitive Science Society: Vol. 34. https://escholarship.org/uc/cognitivesciencesociety/34/34

Bosch, L.T., Oostdijk, N. & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. Speech Communication, 50(1–2), 80–86

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. Attention, Perception, & Psychophysics, 79(1), 333-343.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Browman, C.P. & Goldstein, L. (1992). Articulatory phonology: An overview. Phonetica 49(3–4), 155–180.

Bruggeman, A., Schade, L., Włodarczak, M., & Wagner, P. (2022). Beware of the individual: Evaluating prominence perception in spontaneous speech. Speech Prosody 2022: 268-272, doi: 10.21437/SpeechProsody.2022-55

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), Perspectives on socially shared cognition (pp. 127–149). American Psychological Association. doi:10.1037/10096-006

Cohen Priva, U., Edelist, L., & Gleason, E. (2017). Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. Journal of the Acoustical Society of America, 141(5), 2989-2996.

Condon, W.S. (1974). Neonate movement is synchronized with adult speech: Interactional participation and language acquisition. Science, 183, 99–101.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. Journal of Phonetics, 37(1), 16-28.

Cummins, F. (2012). Oscillators and syllables: a cautionary note. Frontiers in Psychology, 3, 364

Dikker, S., Michalareas, G., Oostrik, M., Serafimaki, A., Kahraman, H. M., Struiksma, M. E., & Poeppel, D. (2021). Crowdsourcing neuroscience: inter-brain coupling during face-to-face interactions outside the laboratory. NeuroImage, 227, 117436.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. Psychological Science, 21, 1664–1670.

Ding, N. & Simon, J.Z. (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. Journal of Neuroscience, 33, 5728–5735.

Endress, A., & Hauser, M. (2010). Word segmentation with universal prosodic cues. Cognitive Psychology, 61, 177–199.

Fitch, W. T. (2013). Rhythmic cognition in humans and animals: distinguishing meter and pulse perception. Frontiers in Systems Neuroscience, 7, 68.

Fló, A., Brusini, P., Macagno, F., Nespor, M., Mehler, J. & Ferry, A.L. (2019). Newborns are sensitive to multiple cues for word segmentation in continuous speech. Developmental Science, 22, e12802. doi:10.1111/desc.12802

Fuglsang, S.A., Dau, T. & Hjortkjær, J. (2017) Noise-robust cortical tracking of attended speech in real-world acoustic scenes. Neuroimage, 156, 435–444.

Fuscone, S., Favre, B., & Prevot, L. (2021). Reproducibility in speech rate convergence experiments. Language Resources and Evaluation, 55(3), 817¬–832.

Gessinger, I., Raveh, E., Steiner, I., & Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing. Speech Communication, 127, 43-63

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. Phonetica, 66(1–2), 113-126.

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Frontiers in Psychology, 3,1. doi:10.3389/fpsyg.2012.00238

Giraud, AL., Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience 15, 511–517. doi:10.1038/nn.3063

Hjortkjær, J., Märcher-Rørsted, J., Fuglsang, S. A., & Dau, T. (2020). Cortical oscillations and entrainment in speech processing during working memory load. European Journal of Neuroscience, 51(5), 1279-1289.

Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (2012). Rapid entrainment to spontaneous speech: A comparison of oscillator models. Proceedings of the 34th Annual Conference of the Cognitive Science Society Austin: Vol. 34. https://escholarship.org/uc/item/0c905908

Iverson, P., Song, J., & Bradley, H. (2018). Cortical entrainment to speech under competing-talker conditions: Effects of cognitive load due to presentation rate and task difficulty. The Journal of the Acoustical Society of America, 143(3), 1921–1921.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., Jasnow, M. D., Rochat, P., & Stern, D. N. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. Monographs of the society for research in child development, i-149.

Jun, S. A. (Ed.). (2005). Prosodic typology: The phonology of intonation and phrasing. OUP Oxford.

Jungers, M., & Hupp, J. (2009). Speech priming: Evidence for rate persistence in unscripted speech. Language and Cognitive Processes, 24(4), 611–624.

Krivokapić, Jelena (2013). Rhythm and convergence between speakers of American and Indian English. Laboratory Phonology, 4(1), 39–65. doi:10.1515/lp-2013-0003

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., and Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320, 110–113.

Lamekina, Y., & Meyer, L. (2022). Entrainment to speech prosody influences subsequent sentence comprehension. Language, Cognition and Neuroscience, 38(3), 263–276. doi:10.1080/23273798.2022.2107689

Large, E.W. and Jones, M.R. (1999). The dynamics of attending: How people track time-varying events. Psychological Review, 106(1), 119–159.

Lerdahl, F., & Jackendoff, R. (1983). An overview of hierarchical structure in music. Music Perception: An Interdisciplinary Journal, 1(2), 229-252.

Levitan, R., Hirschberg, J. (2011) Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. Proceedings of INTERSPEECH 2011, 3081-3084. doi: 10.21437/Interspeech.2011-771

Lewandowski, N., & Jilka, M. (2019). Phonetic convergence, language talent, personality and attention. Frontiers in Communication, 4, 18.

Lubold, N., & Pon-Barry, H. (2014, November). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge, 5–12.

Ludusan, B., & Wagner, P. (2022). Laughter entrainment in dyadic interactions: temporal distribution and form. Speech Communication, 136, 42–52.

Manning, F., and Schutz, M. (2013). “Moving to the beat” improves timing perception. Psychonomic Bulletin and Review, 22, 1133–1139. doi: 10.3758/s13423-013-0439-7.

Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. Evolution and Human Behavior, 34(6), 419-426.

Meyer, L., Sun, Y., & Martin, A. E. (2020). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. Language, Cognition and Neuroscience, 35(9), 1089¬–1099.

Michalsky, J., & Schoormann, H. (2017). Pitch Convergence as an Effect of Perceived Attractiveness and Likability. Proceedings of INTERSPEECH: 2253–2256.

Morrill, T., Dilley, L., McAuley, J. D., & Pitt, M. (2014). Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis. Cognition, 131(1), 69–74. doi:10.1016/j.cognition.2013.12.006

Néda, Z., Ravasz, E., Vicsek, T., Brechet, Y., & Barabási, A. L. (2000). Physics of the rhythmic applause. Physical Review E, 61(6), 6987.

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. Trends in Cognitive Sciences, 23(11), 913-926.

Ordin, M., & Nespor, M. (2013). Transition Probabilities and Different Levels of Prominence in Segmentation. Language Learning, 63, 800-834. doi:10.1111/lang.12024

Pardo, J. (2006). On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America, 119, 2382–2393.

Pardo, J. (2012). Reflections on phonetic convergence: Speech perception does not mirror speech production. Language and Linguistics Compass, 6(12), 753–767.

Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. Journal of Phonetics, 42, 1–11.

Peelle, J., Gross, J., & Davis, M. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23(6),1378–87. doi:10.1093/cercor/bhs118. Epub 2012

Peelle, J., & Davis, M. (2012). Neural oscillations carry speech rhythm through to comprehension. Frontiers in Psychology, 3. doi:10.3389/fpsyg.2012.00320.

Pickering, M.J. and S. Garrod (2004). Towards a mechanistic psychology of dialogue. Behavioural and Brain Sciences, 27, 169–226.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? Trends in Cognitive Sciences, 11(3), 105–110.

Pitt, M. A., Szostak, C., & Dilley, L. (2016). Rate dependent speech processing can be speech-specific: Evidence from the perceptual disappearance of words under changes in context speech rate. Attention, Perception, & Psychophysics, 78(1), 334–345. doi:10.3758/s13414-015-0981-7

Port, R. F. (2003). Meter and speech. Journal of Phonetics, 31(3–4), 599-611.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. The Journal of the Acoustical Society of America, 123(2), 1104–1113.

Rahimi, Z., Kumar, A., Litman, D. J., Paletz, S., & Yu, M. (2017). Entrainment in Multi-Party Spoken Dialogues at Multiple Linguistic Levels. Proceedings of INTERSPEECH: 1696-1700.

Rathcke, T., Lin, C. Y., Falk, S., & Bella, S. D. (2021). Tapping into linguistic rhythm. Laboratory Phonology, 12(1).

Reichel, U. D., Beňuš, Š., & Mády, K. (2018). Entrainment profiles: Comparison by gender, role, and feature set. Speech Communication, 100, 46-57.

Reinisch, E., Bosker, H.R. (2022). Encoding speech rate in challenging listening conditions: White noise and reverberation. Attention, Perception & Psychophysics 84, 2303–2318. doi:10.3758/s13414-022-02554-8

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. Psychonomic Bulletin & Review 12, 969–992. doi:10.3758/BF03206433

Repp, B. H., and Su, Y.-H. (2013). Sensorimotor synchronization: a review of recent research (2006–2012). Psychonomic Bulletin & Review 20, 403–452. doi: 10.3758/s13423-012-0371-2

Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. (2018). Neural Entrainment to Speech Modulates Speech Intelligibility. Current Biology, 28(2), 161–169.e5. doi:10.1016/j.cub.2017.11.033

Schweitzer, A., & Lewandowski, N. (2014). Social factors in convergence of F1 and F2 in spontaneous speech. In Proceedings of the 10th International Seminar on Speech Production, Cologne: 391-394.

Schweitzer, A., Lewandowski, N., & Duran, D. (2017). Social Attractiveness in Dialogs. Proceedings of INTERSPEECH: 2243–2247.

Schultz, B., O'Brien, I., Philipps, N., McFarland, D., Titone, D., & Palmer, C. (2016). Speech rates converge in scripted turn-taking conversations. Applied Psycholinguistics, 37(5), 1201–1220. doi:10.1017/S0142716415000545

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. University of Illinois, Urbana.

Song, J., & Iverson, P. (2018). Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. Cognition, 179(23):163-170. doi: 10.1016/j.cognition.2018.06.001.

Tajima, K., & Port, R. F. (2003). Speech rhythm in English and Japanese. In J. Local, R. Ogden & R. Temple (Eds.), Phonetic Interpretation: Papers in Laboratory Phonology VI (pp. 317–334), CUP.

Tyler, M., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. Journal of Acoustical Society of America, 126, 367–376.

Wagner, P. (2008). The rhythm of language and speech: Constraints, models, metrics and applications. Online manuscript. URN: urn:nbn:de:0070-pub-19168457

Wagner, P. (2013). Meter specific timing and prominence in German poetry and prose. In O. Niebuhr (Ed.), Understanding Prosody: The Role of Context, Function and Communication (pp. 219-236). Berlin, Boston: De Gruyter. doi:10.1515/9783110301465.219

Wagner, P., Malisz, Z., Inden, B., & Wachsmuth, I. (2013). Interaction phonology – a temporal co-ordination component enabling representational alignment within a model of communication. In I. Wachsmuth, J. de Ruiter, P. Jaecks & S. Kopp (Eds.), Alignment in Communication: Towards a New Theory of Communication. (Vol. 6, pp. 109–132). Amsterdam, Philadelphia: John Benjamins.

Wagner, P., Ćwiek, A., & Samlowski, B. (2019). Exploiting the speech-gesture link to capture fine-grained prosodic prominence impressions and listening strategies. Journal of Phonetics, 76, 100911.

Wang H., Mok P., Meng H. (2016). Capitalizing on musical rhythm for prosodic training in computer-aided language learning. Computer Speech and Language, 37, 67–81.

Weise, A., & Levitan, R. (2018). Looking for structure in lexical and acoustic-prosodic entrainment behaviors. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (pp. 297-302): Volume 2.

Wilsch, A., Neuling, T., Obleser, J., & Herrmann, C. S. (2018). Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. NeuroImage, 172, 766–774. doi:10.1016/j.neuroimage.2018.01.038

Włodarczak, M., Šimko, J., & Wagner, P. (2012). Temporal entrainment in overlapped speech: Cross-linguistic study. Proceedings of INTERSPEECH: 615-618.

Wynn, C. J., Borrie, S. A., & Sellers, T. P. (2018). Speech rate entrainment in children and adults with and without autism spectrum disorder. American Journal of Speech-Language Pathology, 27(3), 965-974.)

Zec, D. (2007). The syllable. In P. de Lacy (Ed.), The Cambridge Handbook of Phonology (pp. 161–194). Cambridge CUP.

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. Current Biology, 28(3), 401–408.e5. doi:org/10.1016/j.cub.2017.11.071