

MAX
PLANCK

MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

WWW.MPI.NL

how speaking modelling control of speaking rate fast is like running

Joe Rodd



Doctoral thesis

How speaking fast is like running

modelling control of speaking rate

Joe Rodd

This research was conducted at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands and the Centre for Language Studies, Radboud University, with financial support from the Netherlands Organization for Scientific Research (NWO Gravitation Grant, 024.001.006) awarded to the Language in Interaction consortium.

Copyright © Joe Rodd, 2020

ISBN: 978-94-92910-14-1

Printed and bound by Ipskamp Drukkers, b.v.

**How speaking fast is like running:
modelling control of speaking rate**

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus prof. dr. J.H.J.M van Krieken,

volgens besluit van het college van de decanen

in het openbaar te verdedigen op donderdag 24 september 2020

om 10.30 uur precies

door

Jonathan James Eyre Rodd

geboren op 15 februari 1991

te Southampton (Verenigd Koninkrijk)

Promotoren:

Prof. dr. Mirjam T.C. Ernestus

Prof. dr. Antje S. Meyer

Copromotoren:

Dr. Hans Rutger Bosker

Max Planck Institute for Psycholinguistics

Dr. Louis F.M. ten Bosch

Manuscriptcommissie:

Prof. dr. J. Paula M. Fikkert

Prof. dr. Hugo Quené

Universiteit Utrecht

Prof. dr. Falk Huettig

Dr. Audrey Bürki

Universität Potsdam, Duitsland

Dr. Stefan L. Frank

Contents

1	Introduction	7
1.1	Speaking rate control is essential for communication	7
1.2	A working model of speech production	8
1.3	The research in this thesis	9
1.4	Outline and research questions	11
2	Planning unit timings from acoustics and articulation	15
2.1	Background	16
2.2	Study aims	19
2.3	Speech materials	21
2.4	Planning unit timing from articulatory measurements	22
2.5	Planning unit timing from the acoustic signal	25
2.6	Results and discussion	26
2.7	Conclusion	29
3	POnSS: efficient and accurate segmentation	31
3.1	Introduction	32
3.2	POnSS	35
3.3	Baseline manual segmentation method	39
3.4	Assessing the reliability of transcription data	40
3.5	Analysis 1: Reliability of modalities	42
3.6	Analysis 2: Efficiency of modalities	46
3.7	Discussion	48
4	PiNCeR corpus	51
4.1	Background	52
4.2	Methods	53
4.3	Confirmatory analysis: word duration	59
5	Simulating speaking rate control	63
5.1	Introduction	64
5.2	Serial order in speech production and the Dell et al. (1997) model	69
5.3	Mechanics of the model	73
5.4	Speech corpus	76
5.5	Training and testing the computational model (strand 1)	78
5.6	How do regimes relate to each other? (strand 2)	97
5.7	General discussion	103
5.8	Conclusion	111

6	Asymmetric switch costs between speaking rates	113
6.1	Introduction	114
6.2	This study	115
6.3	Methods	118
6.4	Results	123
6.5	Discussion	129
6.6	Conclusion	132
7	General discussion	133
7.1	Summary of contributions and findings	133
7.2	What is a gait?	135
7.3	Speaking rate control within gaits	140
7.4	A working theory of mechanisms of speaking rate control	141
7.5	Implications of gaits for phenomena adjacent to speaking rate control	143
7.6	Conclusion	146
	References	149
	Appendix: elicitation materials	167
	Nederlandse samenvatting	171
	English summary	175
	Curriculum vitae	179
	Publications	181
	Acknowledgements	183
	MPI Series in Psycholinguistics	187

1 Introduction

Everyone knows someone who speaks particularly fast, or especially slowly. Different language varieties have different typical speaking rates, as do different individuals who speak the same language variety (e.g. Verhoeven et al., 2004). Speaking rates are also highly variable within individuals (Quené, 2008; Jacewicz et al., 2010; Miller et al., 1984). Speaking rate, as a component of speaking style, is strongly correlated with the communicative situation: when the situation demands clearer speech, speakers tend to slow down accordingly (e.g. Bosker & Cooke, 2018; Cooke et al., 2014; Hazan & Baker, 2011). This happens both automatically as a response to environmental noise (Lombard, 1911; van Summers et al., 1988), and voluntarily, for instance on request when speaking to a language learner. Voluntary modulation of our speaking rate away from our most comfortable speed is experienced as effortful, implying that top-down executive control is required to adjust how fast we speak.

This thesis addresses the question of how speakers tweak the cognitive processes that prepare speech to adjust their speaking rate voluntarily. Aside from examining a previously under-explored aspect of speech production, accounting for how speakers modulate their speaking rate contributes to our understanding of speech production more generally. Firstly, an account of how speakers ‘deliberately’ vary speaking rate has the potential to usefully constrain theories of speech production, which would need to be compatible with this ability. Secondly, understanding which variation is deliberate and which arises as a result of stochastic error in the process of planning speech can offer insights into the nature of the process itself (Bürki, 2018). Thirdly, accounting for modulation of the speech production system may clarify the interface between executive control and speech production (Miyake et al., 2000; Rietbergen et al., 2018; Jongman et al., 2015).

1.1 Speaking rate control is essential for communication

Rather than being mere stylistic variation, variation in speaking rate is important for speech comprehension, which in turn means that accurate control of how fast we speak is essential for successful communication. This is because

variation in the speech signal prevents listeners from relying on simple mappings between signal and meaning. Instead, listeners combine local and contextual cues to arrive probabilistically at the correct meaning (e.g. Martin, 2016). In other words listeners ‘normalise out’ the variation (Johnson, 2008), using the context as an additional cue to the intended meaning. Speakers must therefore ensure that combination of the variants that they produce and the contextual speaking rate will lead to the intended message being understood by the listener.

In the temporal domain, the use of this context is well established, using ambiguous recordings of words, for instance, a recording that is manipulated to be half way between Dutch long *taak* “task” and short *tak* “branch” in vowel quality and duration of the vowel. Listeners who hear the exact same production of an ambiguous word embedded into a slow sentence perceive the short alternative of the pair, *tak*; when the same ambiguous word is embedded into an otherwise fast sentence, they perceive the long alternative, *taak* (e.g. Maslowski et al., 2019b). This effect can even cause entire words to disappear: in a slow context, speakers hear ‘a dollar twenty’ when presented with exactly the same ambiguous recording that leads them to perceive ‘a dollar or twenty’ if surrounded by faster speech (Dilley & Pitt, 2010).

Speaking rate and other speech variation is conditioned by the communicative situation (Hazan & Baker, 2011), suggesting that speakers ‘design’ the speech that they produce to help the listener retrieve maximal meaning from the signal (Lindblom, 1990). Thus, variation in speaking rate is an informative property of the system, from which listeners can extract cues about the intended meaning and about paralinguistic information, such as prosodic phrasing, information structure, and relative importance of constituents (Poupier, 2012).

1.2 A working model of speech production

To explore how speaking rate is controlled, it is necessary to establish a theoretical framework to build upon. The research in this thesis is predicated on a working model, which is sketched in Figure 1.1. This working model divides the work of speech production into three phases, consistent with the classical conceptualisation of speech production as a modular, feed-forward processing system (e.g. Dell & O’Seaghdha, 1992; Levelt et al., 1999; Levelt, 1989; Stemberger, 1985). After a meaning representation has been selected (‘conceptualisation’), the lexical

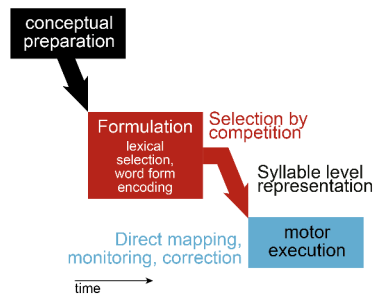


Figure 1.1: A working model of speech production, proposing three distinct phases: a conceptual preparation phase; formulation, a lexical selection and word form encoding phase, involving selection by competition; and a motor execution phase involving mapping, monitoring and correcting.

selection stage begins, where abstract representations of words that best correspond to the conceptual message are selected. Processes of word form encoding then construct detailed word form representations. Lexical selection and word form encoding together can be considered as a formulation phase. The operations of the formulation phase involve competitive selection. Once a word form representation is selected, a motor execution phase is entered, where movement commands for the articulatory apparatus (e.g., the tongue, lips, vocal chords) are calculated, carried out, and monitored (Guenther, 2016a; Tourville & Guenther, 2011). Because speakers typically plan as late as possible, rather than storing a pre-planned utterance in working memory (e.g., Damian & Dumay, 2007; Kello et al., 2000; Levelt, 1989; Levelt et al., 1999), the formulation system must keep up with the desired rate of articulation, requiring modulation of its operation to maintain synchronisation. In contrast to formulation, there is no competition in the execution phase, which instead involves mapping and monitoring.

1.3 The research in this thesis

Current theories of speech production do not account for the ability to modulate the process of producing speech, nor explain how variation might emerge and how it might be conditioned. This thesis has the broad aim of explaining how speaking rate may be controlled. Previous research on control of speaking rate has primarily examined how the operation of the execution phase of speech production varies with altered speaking rate, that is, how the mechanics of ar-

ticulation are modulated in order to fit the same planning units into less time, or stretch them to occupy more time. Important topics have been articulator movement (Adams et al., 1993; Gay, 1981; Kuehn & Moll, 1976; Ostry & Munhall, 1985; van Brenk et al., 2013), gestural timing (Byrd & Tan, 1996; Tjaden & Weismer, 1998), and how articulatory gestures within the syllable are coordinated relative to each other (Tilsen, 2014; Tilsen, 2016). In contrast, how the formulation phase might be controlled to achieve different speaking rates is less well studied, with focus instead on strategies to comply with imposed response deadlines (Lupker et al., 1997; Kello & Plaut, 2000; Kello & Plaut, 2003; Kello, 2004). In this thesis, a theoretical model is constructed that is compatible with the consensus view of speech production (Levelt, 1989; Dell et al., 1997; Dell & O'Seaghdha, 1992; Levelt et al., 1999; Stemberger, 1985; Roelofs, 2008; Hickok, 2014). Controlled experimentation to elicit speech at various known, stable speaking rates is combined with computational modelling to test for hypothetical control strategies given the theoretical model.

The gaits hypothesis

The control mechanisms engaged to regulate speaking rate at the level of utterance planning and preparation are largely unknown. This invites us to look to how control is exerted on other cognitive-biological systems and hypothesise by analogy to these systems. Inspiration is taken from the human and animal locomotion system, which, like speech, can operate at any of a continuous range of speeds, but since it is easily observed is well explored and understood. In animals with legs, qualitatively different gaits are adopted to achieve different speeds of movement: humans walk, skip or run, while animals with four legs have more possibilities: galloping, trotting, cantering, etc. Different gaits differ in the cycle of the limbs: in walking, at least one foot is on the ground at all times, whilst in running, both feet are raised from the ground simultaneously for part of the cycle (Minetti, 1998; Alexander, 1989). Legged animals can achieve a continuous range of movement speeds, but not all gaits are appropriate, or even feasible, at all speeds. This means that animals must switch between gaits to achieve different speeds.

Alongside hard limits on feasibility of certain gaits at certain speeds, the selection of locomotive gaits is tightly linked to their relative efficiency. Each gait has a 'sweet spot' speed, at the approximate centre of the range of speeds achievable

with that gait, where exertion (ml O₂ consumed to move 1 metre) is minimised (Hoyt & Taylor, 1981, their Figure 2). These sweet spot speeds are preferred (Pennycuik, 1975) over less efficient speeds.

The link between gaits and efficiency has previously inspired locomotive gaits to be used as a metaphor for different, equally optimal coordination modes of the execution phase of speech production. For instance, Pouplier (2012) proposed that the local context would influence which of various possible articulatory coordination modes would be selected, when the perceptual outcome was equivalent for the listener. In this thesis, I make a related, but different proposal: that there are qualitatively different configurations of the formulation component that resemble the gaits of locomotion. Each formulation gait yields word forms that are subtly distinct in their internal timing properties, capturing durational differences between speech at broad ranges of speaking rate.

If there were no gaits, the speech planning apparatus might have a single configuration which is gradually up- or down-regulated in response to changes in required speaking rate, resembling a simple gain knob. This is the case for non-speech motor tasks where temporal precision is required, in both gross motor movements (Wright & Meyer, 1983), and fine movement requiring extensive coordination, such as piano playing (Bella & Palmer, 2011).

1.4 Outline and research questions

The research in this book concerns how speakers adjust their speaking rate at the level of formulation, or more briefly: **‘how do speakers control their speaking rate?’**. My aim is to test the gaits hypothesis against the alternative hypothesis of absence of gaits.

Chapters 2 and 3 are methodological chapters, and describe and validate analysis tools that were developed to prepare the data for the other chapters.

Chapter 2 addresses the research question **‘how can planning unit onset and offset times be identified from the acoustic speech signal?’**. To be able to model speech timing, it is essential to know when each syllable-level ‘planning unit’ begins and ends. As a purely psychological construct, it is of course impossible to observe the beginning and ends of planning units. It is argued that, with thoughtful selection of words included in the experimental materials, it is possible to get close to identifying the onsets and offsets of planning units di-

rectly from the acoustic signal, which brings practical benefits in comparison to approaches that track articulator motion directly. This approach is validated by comparison to articulographic data.

Chapter 3 addresses the research question **‘how can speech most efficiently be segmented to the word level?’**. The chapter describes a speech segmentation system, POnSS, that was developed to allow efficient word-level segmentation of the speech materials elicited in the various experiments. POnSS combines automatic processes and human input. POnSS is compared to conventional hand segmentation with Praat, to validate its relative reliability and efficiency.

The primary aim of **Chapter 4** is to introduce an experiment and a corpus of speech data that were used in Chapter 5. The chapter has the secondary aim of addressing the research question **‘do explicit instructions to avoid pausing influence the syllable durations of elicited speech?’**. In the experiment, speakers had to name pictures, in Dutch, at three pre-determined speaking rates. The pictures were arranged around a ‘clock-face’, and a dot jumped clockwise from picture to picture to indicate which picture was to be named when, and thus specify the required speaking rate. The data were segmented with POnSS, and the analysis technique introduced in Chapter 2 was used to identify the onsets and offsets of syllable-level planning units.

Chapter 5 addresses the research question **‘do speakers switch between qualitatively distinct configurations (“gaits”) of the speech production system to control their speaking rate, or is rate control achieved purely by up or down regulation of the speech planning system?’**. Our model of speech production, EPONA, is introduced. There are two strands running through the chapter: the first strand describes the construction and implementation of the model. While doing so, this strand evaluates and compares variations on possible representations of the temporal structure of words in the frame node at the heart of the metrical stream. The second strand addresses the empirical question of how speaking rate is regulated, given the model. The model was used to simulate the speech data gathered in Chapter 4. An evolution-inspired optimisation algorithm was used to find values for the parameters of the model that resulted in simulated speech durations that most resembled those observed in Chapter 4. The multi-dimensional space formed by considering each of the parameters of the model as a dimension is an approximation of the cognitive space of the human speaker. That means that patterns found in the optimal pa-

parameter values for the three speaking rates in the model can be used to draw tentative conclusions about the modulation of the configuration of the human speech production system. Several different analyses were used to check for patterns consistent with the gait hypothesis.

Chapter 6 builds on Chapter 5 by addressing the research question ‘**which gaits are engaged to achieve fast, medium and slow speaking rates?**’. An experiment is described that aimed to test the key conclusion of Chapter 5 more directly. In the experiment, speakers named pictures from a clock-face display, similar to Chapter 4. This time, speakers were trained to speak at the three pre-determined speaking rates before the experiment. They then had to maintain the speaking rates themselves. The required speaking rate was indicated by the colour of a frame placed around the picture display. At an unpredictable moment during the trial, the colour of the frame changed, indicating that the speaker should adjust their speaking rate. We expected that differences would emerge in how quickly it would be possible to switch between different pairs of speaking rates: faster adjustment would indicate that less cognitive reconfiguration was required to make the switch between the relevant rates. A difference in the rate of switching between slow and medium rates on the one hand and fast and medium rates on the other would be compatible with the gaits-of-speech account from Chapter 5, implying, depending on the direction of the effect, that medium and either fast or slow are achieved by one gait, and the remaining rate by another. To test this, the speech materials were segmented using POnSS (Chapter 3), then statistical modelling was used to look for differences in adaptation speed.

Finally, **Chapter 7** summarises and discusses the results of the methodological and experimental chapters, discusses how the results support the gait hypothesis, and elaborates on the utility and implications of the EPONA model of speech production. The consequences of the presence of gaits in the speech production system are discussed, as are future research directions that I think would prove fruitful.

2 Deriving the onset and offset times of planning units from acoustic and articulatory measurements

Many psycholinguistic models of speech sequence planning make claims about the onset and offset times of planning units, such as words, syllables, and phonemes. These predictions typically go untested, however, since psycholinguists have assumed that the temporal dynamics of the speech signal is a poor index of the temporal dynamics of the underlying speech planning process. This chapter argues that this problem is tractable, and presents and validates two simple metrics that derive planning unit onset and offset times from the acoustic signal and articulatographic data.

This chapter was adapted from Rodd, J., Bosker, H. R., ten Bosch, L., & Ernestus, M. (2019b). Deriving the onset and offset times of planning units from acoustic and articulatory measurements. *The Journal of the Acoustical Society of America*, 145(2), EL161–EL167. <https://doi.org/10.1121/1.5089456>. Code is available at <https://git.io/fh8EM>.

2.1 Background

Typically, the inverse mapping between the acoustic signal and the articulator configuration is characterised as highly non-linear and one-to-many, in that many speech sounds can be produced by multiple configurations of the vocal tract (e.g. Lindblom, 1983). This assumed intractability complicated the evaluation of psycholinguistic models of speech planning, specifically claims about the implementation of abstract linguistic planning units by speech motor programs.

While it is the case that speakers can make use of alternative vocal tract configurations to achieve speech sounds when articulatory freedom is constrained (Lindblom et al., 1977), or to reduce the required movement from the previous configuration (e.g. Boyce & Espy-Wilson, 1997), the opacity of the correspondences between acoustics, articulation, and the dynamics of higher planning processes may be overestimated (Hogden et al., 1996). This paper posits that the problem is tractable, and proposes methods to characterise the dynamics of higher planning processes from the acoustic signal or from tracked articulator movements. Thus, the testing of previously untestable predictions of psycholinguistic models is facilitated.

2.1.1 Acoustic change largely reflects articulatory change

Despite assumptions to the contrary, in practice, the inverse mapping from the acoustic signal to articulatory configurations can be defined in an appropriate way to predict articulatory configurations from the acoustic signal, within a certain tolerance for deviations in the articulatory domain. For speech sounds that intrinsically consist of multiple acoustic events (such as diphthongs, plosives), the mapping results in an estimated trajectory in articulatory space. For a subset of stable speech sounds, ‘codebooks’ of articulatory configurations associated with acoustic outcomes can be compiled (e.g. Hogden et al., 1996). Moreover, machine learning approaches that can make use of contextual information and sufficiently large corpora of training data have proven successful in predicting articulatory configuration from the acoustic signal with no constraints on speech materials (e.g. Richmond, 2006; Illa & Ghosh, 2018; Uria et al., 2011).

Relatedly, it holds that when the vocal tract is in a stable configuration, the acoustic output is also stable, and that when the acoustic output is changing, the vocal tract configuration must also be changing. This observation has been

exploited in blind speech segmentation, where frame-by-frame changes in the acoustic spectrum are tracked, and peaks in spectral change are detected. These peaks correspond to perceptually relevant phone boundaries (e.g. Dusan & Rabiner, 2006; Hoang & Wang, 2015; ten Bosch & Cranen, 2007). These approaches are intended to automate the preparation of corpora to test speech recognition systems, and assume that segments are concatenated without overlap, making these algorithms unsuited for the retrieval of onset and offset times of overlapping planning units predicted by psycholinguistic models. They can, however, serve as inspiration for the development of new techniques to retrieve planning unit dynamics.

Note that although changes in the acoustic signal must reflect changes in the articulatory configuration, it does not follow that when the vocal tract configuration is changing, the acoustic signal always changes with it, since for many speech sounds, the precise positioning of non-critical articulators is unimportant (such as tongue position during the realization of /m/).

2.1.2 The mapping between planning units and acoustics and articulation

A class of psycholinguistic speech production models (which we will term phoneme-based models) characterise the units that mediate between formulation (lexical access and phonological encoding) and execution (speech motor programming and articulation itself) as phonemes, or sequences of phonemes, such as syllables, demi-syllables, or whole words (e.g. Levelt, 1989; Levelt et al., 1999; Dell & O'Seaghdha, 1992; Tourville & Guenther, 2011). Phoneme-based models also conceptualise the execution process as an obedient servant of formulation (e.g. Levelt, 1989; Levelt et al., 1999; Dell & O'Seaghdha, 1992; Tourville & Guenther, 2011), which entails that the observable movements of the articulators and the resulting speech acoustics are inherently a consequence of planning units in formulation becoming active and subsequently being deactivated. That the dynamics of the activation of planning units directly influences the articulatory configuration and thereby the acoustic output seems plausible in the light of findings that competing representations in the formulation phase exert some influence on fine detail in articulation (e.g. Goldrick & Blumstein, 2006).

The DIVA model (Tourville & Guenther, 2011) operationalises the planning units by defining them in terms of upper and lower bounds for articulator positions, and upper and lower bounds of the expected auditory outcome in terms of

fundamental frequency and formants. Planning units typically overlap in time, and all simultaneously active planning units exert influence on both the articulatory configuration and speech acoustics directly via the feedforward route. They also influence articulation and acoustics indirectly by shaping the expected acoustic and somatosensory outcomes, which in turn lead to corrective feedback.

The temporal overlap of adjacent planning units (at the output stage of phoneme-based psycholinguistic speech planning models) results in local coarticulation in the overt speech. Equivalently, low level pre-activation (priming) of upcoming planning units and incomplete deactivation of preceding planning units result in longer-range coarticulation in the overt speech.

The retrieval of planning units from articulatory measurements has previously been attempted by Steiner and Richmond (2009), who developed an analysis-by-resynthesis approach that reconstructs a gestural score from electromagnetic articulography (EMA) data in terms of vocalic and consonantal gestures for the VocalTractLab (VTL) synthesizer (Birkholz et al., 2007). This representation differs somewhat from that inherent to phoneme-based models, in that vowel and consonants are treated as fundamentally distinct units of representation on distinct tiers of the gestural score, while phoneme-based models instead predict a chain of potentially overlapping planning units of the same class, on the same tier.

Vaz et al. (2016) described an algorithm to retrieve underlying structure from multivariate time series data, and tested it on vocal tract constriction distances measured from real-time MRI vocal tract data. The algorithm was able to construct an inventory of gestures from the data, and an activation time series for each of these gestures, which are collectively analogous to a gestural score in the articulatory phonology (AP) framework. AP diverges from phoneme-based production models in that the planning units it supposes are not phonemes or sequences of phonemes, but rather articulatory gestures defining articulatory events, such as opening of the glottal aperture, or the creation of a labial closure (Browman & Goldstein, 1992), which cannot easily be translated into phonemes.

The direct retrieval of the timings of planning units from the acoustic signal has been attempted by Nam et al. (2012), again with an analysis-by-synthesis approach, and similarly rooted in the articulatory phonology (AP) framework. Their procedure involves constructing a task dynamic gestural score (encoding

the speech to be produced in terms of degrees of constriction at different positions in the vocal tract) from an orthographic transcription of the speech. Then, the TADA model (Nam et al., 2004; Saltzman & Munhall, 1989) is used to predict time-varying vocal tract dimensions from the gestural score, which is then synthesized to produce a speech signal. Next, dynamic time warping (DTW) is applied between the synthesized and natural speech signals. This involves stretching and compressing the synthesized speech signal in the temporal dimension, to improve the temporal alignment with the natural speech signal. The result of the DTW is a warping scale, which can then be applied to the gestural score, yielding a warped gestural score from which activation and deactivation times of individual gestures can be established.

Aside from requiring potentially difficult to acquire articulatory measurements (EMA in the case of Nam et al. (2012), real time MRI in the case of Vaz et al. (2016)), these procedures that construct multivariate gestural scores cannot readily be applied to phoneme-based models of speech production, since the gestures are not consistent with or easily mapped to the planning units hypothesised by phoneme-based models of lexical access and multi-word processes of speech production (e.g. Levelt, 1989; Levelt et al., 1999; Dell & O'Seaghdha, 1992; Bohland et al., 2010). An additional concern is that the process leaves the researcher relatively unconstrained in the construction of the gestural score for a given utterance, either directly or through their parameterization of the linguistic model.

2.2 Study aims

This study aims to provide a means to estimate the onset and offset times of phoneme-based planning units (such as words, syllables or phonemes) from recorded speech materials. The tight temporal locking between formulation and execution processes in speech production (e.g. Goldrick & Blumstein, 2006) suggests that reconstructing the activation dynamics of planning units from measurements of articulator movement is feasible. That the inverse mapping between acoustics and articulation is transparent enough to construct codebooks describing the mapping implies that reconstructing the activation dynamics of planning units from the acoustic signal should also be feasible for a constrained repertoire of speech sounds.

We propose two approaches to retrieve planning unit onset and offset times from speech materials; from the acoustic signal, and from EMA data. We compare the outcomes of the two techniques, to establish that recovering planning unit onset and offset times from the acoustic signal is broadly equivalent to recovering planning unit timing from articulatographic data.

The first metric uses fleshpoint position data gathered by electromagnetic articulography, and begins by deriving upper and lower bounds for each fleshpoint position for each segment from corpus data. Subsequently, a multi-dimensional, time-varying target for a multi-segmental speech sequence is constructed, the temporal parameters of which are adjusted to achieve a good fit to the observed data.

The second is a metric that exploits the acoustic signal directly with no need to record articulator motion, but constrains the speech sounds that can be evaluated. This metric depends on the claim that acoustic instability mirrors articulatory instability, which in turn reflects simultaneous activation of multiple planning units.

Neither metric is predicated on any specific theoretical treatment of speech production, aside from the assumption that planning units are phonemes or sequences of phonemes, and the parameterization of both metrics is wholly data-driven. For the experimental psycholinguist, a metric that can be collected from the acoustic signal alone is clearly preferable, since that reduces the burden of data collection on both researcher and participant, and makes recording of electrophysiological or other measures during speech production possible because no articulatographic data needs to be collected.

The two metrics were tested on acoustic and articulatory data for the same vowel-consonant sequences, taken from the electromagnetic articulography subset of the mngu0 corpus (Richmond et al., 2011), where monophthongs transitioned into continuant consonants. The choice of this limited subset was driven by the need to use segments that were acoustically stable during realization, for the acoustic metric. Comparing the performance of the metrics against a ‘gold standard’ baseline annotation of the onsets and offsets of speech planning units is clearly impossible, given that any hand annotation of speech planning unit onsets and offsets would inherently be largely arbitrary and noisy.

2.3 Speech materials

The EMA subset of the mngu0 corpus (Richmond et al., 2011) was used, which consists of TIMIT sentences read by a single male speaker of British English. EMA sensors were placed on the lower and upper lips, at the tongue tip, blade and dorsum and on the lower incisors (to track jaw motion). A further sensor was placed on the upper incisors to serve as a reference for the others. For technical details relating to the data collection and preparation see Richmond et al. (2011).

2.3.1 Post processing and annotation

From the 1263 sentences of the mngu0 corpus, vowel - consonant (VC) sequences of interest were identified, where a monophthong transitioned into a continuant consonant. The sequences of interest were all one of the following: /am/, /aʃ/, /av/, /ɪʃ/, /ɪv/, /im/, /iv/, /ʌm/, /is/, /ʌs/, /ɒn/. Note that in the context of phoneme-based speech planning models, where no distinction is made between planning units for different classes of phonemes, there is no reason to suppose that sequences of a different composition (CVs, or CCs, for instance) would behave any differently from the VCs tested here. This means that the predictions of phoneme-based speech planning models can effectively be tested by this reduced set of sequences. This yielded 775 sequences of interest, which were identified based on the forced aligned transcriptions available in the corpus. Analysis intervals from the temporal center of the forced aligned vowel to the temporal center of the forced-aligned consonant were defined (see Figure 2.1(a)). The analysis interval served as a landmark to identify the planning unit transitions found; so the precision of the start and end points of the interval was not critical, as long as the transition between the planning units was included.

In the EMA data, lateral movement was discarded, yielding articulator positions on the mid-sagittal plane only. To facilitate annotation, the remaining two dimensional data was rotated independently for each sensor by means of principal component analysis, so that PC1 captured the most informative direction of movement for that sensor, which in all cases was the open-close dimension. Since PC2 is orthogonal to PC1, it captured forward-backward movement of each sensor. Then, manual annotation was undertaken (by the first author) to identify articulatory stable periods of each segment for use in the preparation of the targets used in the articulatory metric. In the manual annotation procedure,

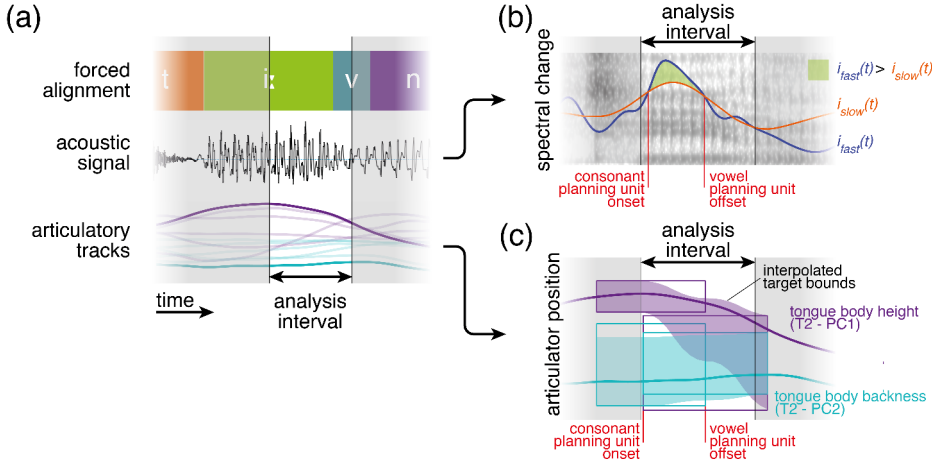


Figure 2.1: An example analysis. (a) An analysis interval is defined that stretches from the temporal center of the forced aligned vowel to the center of the forced aligned consonant. (b) The acoustic metric. Lines show $I_{slow}(t)$ and $I_{fast}(t)$, the Gaussian smoothed, interpolated spectral distance functions used to identify the acoustically evident planning unit overlap during the analysis interval. $I_{slow}(t)$ has a 90 ms kernel, $I_{fast}(t)$ has a 30 ms kernel. Shading identifies periods of atypically fast acoustic change (where $I_{fast}(t) > I_{slow}(t)$), from which the onset of the consonant planning unit and the offset of the vowel planning unit are derived. (c) The articulatory metric. The heavy lines indicate the recorded movement of the tongue body sensor, in the open-close dimension (PC1) and the forward-backward dimension (PC2). The outlined boxes indicate the segmental targets, the shading indicates the interpolated sequence level target.

movement tracks in PC1-PC2 dimensions were displayed on a graphical interface, in which the periods of stability associated with the vowel and continuant consonant could be highlighted. The articulatory configuration was considered stable if there was little to no change (assessed visually) in several sensors. Since the targets were defined in terms of 95% highest density intervals (see section 2.4.1), some noise in this annotation procedure was acceptable.

2.4 Planning unit timing from articulatory measurements

The articulatory metric approaches the identification of planning unit onset and offset times from EMA data by essentially inverting the motor control process: reconstructing a multidimensional articulatory target that could have lead to the recorded movements during a vowel-consonant sequence. This was done sepa-

rately for each vowel-consonant transition token, using a parameter optimization routine which adjusted the onset and offset times of the segment targets to construct a target that fitted the recorded movements well.

2.4.1 Establishing segment targets

First, separate segmental targets are established for the vowels and for the consonants, defined in terms of upper and lower bounds for the positions for each fleshpoint (lower jaw, upper and lower lips, tongue tip, blade and dorsum) on the two dimensions (principal components) of the mid-sagittal plane. These maxima and minima are derived from the distribution of sensor positions during the hand-annotated stable periods of those segments in the corpus, irrespective of context, by extracting the 95% highest density interval(s). When the positioning of a fleshpoint is of crucial importance to the identity of the segment, the positioning of that fleshpoint varies little between realizations, and the target is therefore narrow (e.g. the positioning of the tongue tip in /s/). When the positioning of a fleshpoint is only marginally relevant for the identity of the segment, the target is broad (e.g. the positioning of the tongue back in /v/), since there is lots of variability in the source data.

2.4.2 Combining segmental targets to form a sequence target

The sequence targets were constructed by temporally-overlapping the vowel and consonant targets. Figure 2.1(c) depicts an example of the construction of the targets, for the sequence /i:v/, showing the target bounds for each segment as boxes (purple for PC1, blue for PC2), for the tongue body sensor. The segmental targets are fixed at the outer edges, such that the vowel target begins at the hand-annotated onset of vowel stability, and the consonant target ends at the hand-annotated offset of consonant stability. The other two temporal parameters, the offset of the vowel target and the onset of the consonant target are free parameters that can be optimised.

The upper bound of the sequence target is calculated as an exponential moving average (with a window of 20 ms) of the upper bounds of the segmental targets over time. This means that for time points when only the vowel target is engaged, the upper bound is equal to the upper bound of the vowel target. When both segmental targets are engaged, however, the upper bound switches

smoothly from following the upper bound of the vowel target to reflecting the average upper bound of both targets. Once the vowel target is disengaged, the upper bound again smoothly shifts to reflect the upper bound of the consonant target. The lower bound of the target is derived in the same way.

2.4.3 Parameter optimization

For each analysis interval, an independent parameter optimization routine is conducted. Two parameters, the onset time of the consonant target and the offset time of the vowel target, are optimised with the BOBYQA algorithm (Powell, 2009; Ypma et al., 2018).

To evaluate how well a sequence target defined by a pair of consonant target onset and vowel target offset times fitted the observed movements, the proportion of time points where the recorded sensor positions are outside the bounds of the multidimensional target is counted. This proportion is used as a score to be minimised during the parameter optimization process.

For each realization, 200 starting points for these parameters are tried, sampled from normal distributions ($SD = 25$ ms) centered around the annotated end of vowel stability (this is the center-point of the starting distributions for the consonant onset parameter) and the annotated beginning of consonant stability (this is the center-point of the starting distributions for the vowel offset parameter). A search space constraint ensures that the algorithm only considers solutions where the overlap between the segment targets is greater than 0. Having multiple starting points allows us to assess how consistently the algorithm selects the best performing parameter sets, and offers more protection from premature convergence to local minima. To select a single vowel offset time and a single consonant onset time from the distributions that resulted from the 200 initializations, a two-dimensional distribution is estimated from the resulting parameters, where the dimensions are the vowel offset time parameter and consonant onset time parameter. The distribution is weighted by one minus the score achieved in each attempt, so as to weight the best performing solutions most heavily, and the peak is identified. The coordinates of this peak define the planning unit onset and offset times.

2.5 Planning unit timing from the acoustic signal

The acoustic metric quantifies the rate of change in the acoustic signal (the spectral change). Local peaks in this signal identify periods where the speech acoustics, and therefore the underlying vocal tract configuration, are changing. At the transition between two planning units, this change is due to the interaction of the two overlapping planning units, and the duration of the instability is equated with the duration of the overlap. We term this overlap ‘acoustically evident planning unit overlap’. To be able to establish the onset and offset of instability, a method is required to transform a continuous signal into a categorical one: to distinguish acoustic stability from instability. This is done by overlaying two different smoothings of this signal; a ‘fast’ smooth that captures local changes in the signal, and a ‘slow’ smooth that captures longer trends. We identify periods when the ‘fast’ smooth exceeds the ‘slow’ smooth as unstable, and other periods as stable. The onset of the second planning unit is equated with the start of such a period of instability. The offset of the first planning unit is equated with the end of that same period of instability. This is illustrated in Figure 2.1(b).

2.5.1 Step 1: quantifying acoustic change

To identify the period of overlap, the MFCC vectors (mel frequency cepstral coefficient; 25 ms analysis frame length, samples every 10 ms) for the analysis intervals (with a margin of 40 ms before and after) are extracted using the HTK front end (Young et al., 2006). MFCC vectors may be seen as a numeric representation of the spectral content of the speech signal during a short (25 ms) window, and are one of the best spectro-temporal representations of speech acoustics. From each frame to the next, the Euclidean distance in MFCC space was calculated as follows, where j is the index of the MFCC coefficient and t is the index of the frame:

$$D_{spec} = \sqrt{\sum_{j=0}^{12} (\text{MFCC}_{j_t} - \text{MFCC}_{j_{t+1}})^2} \quad (2.1)$$

This gives $D_{spec}(t)$, a spectral distance function quantifying the degree of spectral change evident in the acoustic signal, sampled every 10 ms.

2.5.2 Step 2: identifying periods of fast acoustic change

This spectral distance function is smoothed twice, once with a 30 ms wide Gaussian kernel, yielding $D_{fast}(t)$, which captures relatively fast changes in the spectral distance function; and once with a 90 ms wide Gaussian kernel, yielding $D_{slow}(t)$, which captures longer term trends in the function.

Spline interpolation (every 0.1 ms) is then applied to these functions in order to improve temporal resolution, yielding $I_{fast}(t)$ and $I_{slow}(t)$. The two interpolated functions are overlaid, and parts of the signal in each analysis interval where $I_{fast}(t)$ is larger than $I_{slow}(t)$ are identified as candidate overlaps (in Figure 2.1(b) shown as green shading). Where $I_{fast}(t)$ exceeds $I_{slow}(t)$, atypically fast acoustic change is occurring: acoustically evident planning unit overlap. It is possible that there are multiple periods where $I_{fast}(t)$ exceeds $I_{slow}(t)$, however, typically one period is longer and the associated peak is larger. Therefore, a heuristic is engaged to select precisely one period per analysis window: the duration of each of these periods is calculated. Periods that cross the boundaries of the analysis interval (into the margins) are discarded. When an analysis interval still contains multiple periods, all but the longest candidate are discarded. This yields precisely one period of acoustically evident planning unit overlap per analysis interval. The onset of the remaining period of overlap (where $I_{fast}(t)$ becomes larger than $I_{slow}(t)$) yields the onset of the consonant planning unit. The offset of the overlap (where $I_{fast}(t)$ becomes smaller than $I_{slow}(t)$) yields the offset of the vowel planning unit.

This procedure was refined by testing various kernel widths and interpolations via a grid search, in which the parameters that resulted in the highest spectral change peak were selected.

R scripts implementing the two metrics and the data preprocessing method are available from <https://git.io/fh8EM>.

2.6 Results and discussion

2.6.1 Validity of the metrics

Figure 2.2 shows the onsets and offsets of planning units (event times) as predicted by the articulatory (x-axis) and acoustic metrics (y-axis). All event times are relative to the forced-aligned offset of the consonant segment, meaning that

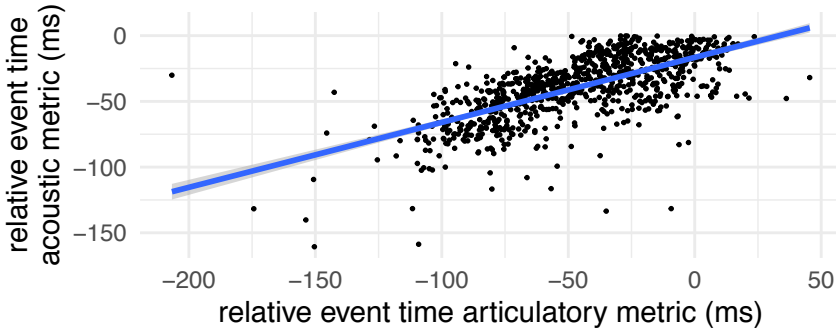


Figure 2.2: The correlation between planning unit onset and offset times, derived from the articulatory (x-axis) and acoustic metrics (y-axis). All event times are relative to the forced-aligned offset of the consonant segment, meaning that times less than 0 are to be expected.

times less than 0 are to be expected. An r^2 of 0.447 was calculated between the event times derived by the two metrics. This moderately high correlation between the predictions of the two metrics indicates that they both capture the same underlying dynamic process of planning unit activation.

The intercept of -10.64 indicates that the acoustic metric systematically predicts earlier event times than the articulatory metric does. This is approximately half the width of the 25 ms analysis window employed in the acoustic metric, which suggests that this anticipation may be an artifact of the spectral analysis inherent to the acoustic metric.

2.6.2 Reliability

The metrics were evaluated by comparing the planning unit onset and offset times predicted by each metric. Because the two metrics are so divergent in the modality of the data used and the approach used to derive event times from the data, we interpreted the finding that the two metrics predicted comparable event times as evidence that they are both indexing the onset and offset times of planning units. This is of course weaker evidence in support of the validity of a metric than comparison against data capturing the ground truth, but the ground truth is clearly unobtainable for psychological processes such as the activation dynamics of planning units. Comparison against the results obtained by Nam et al. (2012) is also problematic given the AP theoretical framing inherent to their procedure.

2.6.3 Applicability and ecological validity

The articulatory metric is in principle equally suited to examining transitions between any pair of segments where there is at least a short period of articulatory stability in each segment, including stops. Of course, given the metric-comparison approach we took to evaluate the performance of the two metrics, the articulatory metric was only tested on materials also suitable for the acoustic metric.

The acoustic metric is inherently limited to identifying planning unit onset and offset times at transitions between a subset of segment types involving at least a short period of articulatory stability and incomplete obstruction of the airflow: monophthong vowels, nasals and continuant fricatives. Nevertheless, for the experimental psycholinguist, the convenience of the acoustic-only recording may well outweigh the disadvantage of constrained material selection.

Both metrics share the inherent assumption that the onsets of all the movements or gestures involved in the production of a phoneme are synchronised. This assumption is inherent to the class of phoneme-based models, which form the mainstream in psycholinguistic models of higher speech planning. Adhering to it was necessary to achieve this paper's goal of making it possible to test and refine phoneme-based models by relating activation dynamics to the speech signal. Models based on a multivariate gestural score may achieve better fits to the data given that they are not constrained by this synchronicity assumption.

The metrics were developed and tested using the mngu0 corpus (Richmond et al., 2011), which contains a large quantity of English data from a single speaker, rather than smaller quantities of data from multiple speakers available in other corpora (e.g. the Wisconsin x-ray microbeam database, Westbury et al., 1990). The mngu0 corpus was selected because we sought to have a large number of realizations of each segment to reliably compute the static segment targets for the articulatory metric. It remains to be seen how the articulatory metric would perform given a smaller dataset from which to derive target boundaries. A requirement for a large speaker-specific dataset would be disadvantageous in the context of experimental psycholinguistics, where it is typically desirable to test multiple speakers on a small set of materials, though recent success in using a generalised background model and a speaker-specific adaptive model in acoustic-to-articulatory inversion (Illa & Ghosh, 2018) offers hope that a comparable approach could work for this metric too.

2.7 Conclusion

This paper presented two techniques to identify planning unit onsets and offsets from articulographic and acoustic data in the context of phoneme-based models of speech production. The first metric requires articulographic recording, but imposes less constraint on speech material selection. The second metric exploits the acoustic signal directly, with no need to record articulator motion, but constrains the speech sounds that can be evaluated. This metric depends on the claim that acoustic instability mirrors articulatory instability, which in turn reflects simultaneous activation of multiple planning units. The two metrics are agnostic to the duration of planning units (syllables, demi-syllables, phonemes, entire words), and make minimal assumptions about precisely what is encoded by the planning unit, other than that upper and lower bounds for articulatory positions are encoded. A moderately high correlation between the event times predicted by the two metrics indicates that they capture the same underlying dynamic process of planning unit activation. This correlation means in turn that temporal predictions arising from phoneme-based psycholinguistic models of speech planning can be tested using the acoustic signal without the need to collect articulographic data.

3 A tool for efficient and accurate segmentation of speech data: announcing POnSS

Despite advances in automatic speech recognition (ASR), human input is still essential to produce research-grade segmentations of speech data. Conventional approaches to manual segmentation are very labour-intensive. We introduce POnSS, a browser-based system that is specialized for the task of segmenting the onsets and offsets of words, that combines aspects of ASR with limited human input. In developing POnSS, we identified several sub-tasks of segmentation, and implemented each of these as separate interfaces for the annotators to interact with, to streamline their task as much as possible. We evaluated segmentations made with POnSS against a baseline of segmentations of the same data made conventionally in Praat. We observed that POnSS achieved comparable reliability to segmentation using Praat, but required 23% less annotator time investment. Because of its greater efficiency without sacrificing reliability, POnSS represents a distinct methodological advance for the segmentation of speech data.

3.1 Introduction

In many speech-based disciplines, the availability of adequately segmented and transcribed speech corpora is essential for designing and benchmarking computational models of speech processing and for sharpening theories of speech production and perception. Many of the speech databases available to date (e.g., via The Language Archive, 2019; European Language Resources Association, 2019; Linguistic Data Consortium, 2019) have been (at least partly) enriched with a verbatim word-level and/or a phonetic transcription.

Speech transcription concerns the generation of a verbatim textual record of speech. The related process of segmentation concerns additionally determining when the transcribed words and segments occur in a speech recording. This article primarily addresses segmentation. Constructing transcriptions and segmentations typically involves three challenges. The first challenge is to take into account the purpose of the segmentation for determining the desired granularity level for the segmentation units. Due to fine phonetic details (Hawkins, 2003) and reduction phenomena (Ernestus & Warner, 2011), word-based transcriptions are much easier and faster to construct than high quality finer-grained faithful phonetic segmentations. Rough, errorful transcription may be sufficient for text query-based services, and may be quickly constructed. Segmentation of varying degrees of accuracy may be required for rich diarisation of meetings, or for the adaptation of acoustic models in automatic speech recognition (ASR). Language research represents a highly niche segmentation usage case, with its own specific requirements and constraints.

The second challenge is the construction of the segmentation itself. This is not a trivial task. One may perform segmentation by hand or apply an automatic speech segmentation system, or a combination of these. Over the last decades, several tools have been developed to ease this task (see, e.g., van Bael et al., 2007; Lecouteux et al., 2012). In general, there is a clear trade-off between the invested time on the one hand and the quality of the resulting segmentation on the other (Rietveld, Ernestus, et al., 2004).

The third challenge is the validation of the segmentation. Manual or automatic segmentations may be validated in terms of their resemblance to each other, or to another “expert-based” hand-crafted reference segmentation. Alternatively, they may be assessed by using e.g. the inter-rater or inter-system

agreement as objective function. However, since symbolic segmentation cannot fully represent the subtle phonetic details in speech, the status of a “reference” segmentation as a single reference for the quality of other segmentations might be questionable a priori. In addition, the validation procedure will largely depend on the purpose. For example, verbatim ‘summary’ transcriptions of meetings may be of sufficient quality to serve a service based on text queries, but still far from sufficient for the development or adaptation of acoustic models in ASR systems.

In this paper we focus on the construction of segmentations at the word level, given a large collection of speech recordings. Several linguistic research tools are available for semi-manually segmenting, annotating or labelling speech corpora. Tools may combine multiple functionalities such as speech recognition, speaker identification, and diarisation to provide real-time and/or offline transcription of audio recorded in various conditions. Based on ASR approaches (e.g. Young et al., 2006; Povey et al., 2011), segmentation and transcription can be done automatically or semi-automatically. We will use the term ‘forced alignment’ to refer to automatic segmentation of speech data using ASR where a transcription already exists, and the term ‘recognition’ to refer to generation of a segmentation without a pre-existing transcription. The quality of automatically generated segmentations depends on the acoustic quality of the recordings (presence of background noise, interference from speakers, echo etc.) and the degree of match between input speech signal and the speech material used for training the ASR (dialects, accents, age, speaking style, mood etc.). Several tools (e.g., the DART tool, Weisser, 2016) are able to identify speech acts automatically, provide multiple interactive annotation functions, and allow special tools for those features that require post-processing. Praat (Boersma & Weenink, 2019) allows the user to manually segment and transcribe speech corpora using different tiers. EMU (Winkelmann et al., 2017) offers similar segmentation and transcribing possibilities as Praat, but in a web interface and in combination with a sophisticated database to store and manage speech data, segmentations and annotations. Despite the availability of these tools, the creation and checking of a segmented and transcribed speech corpus is still a considerable effort.

The recent advent of deep learning techniques, together with improved computational power and availability of data, has lead to significant improvements in the performance of ASR systems. Despite these substantial improvements in

their quality and practicality, fully automatic approaches to the segmentation of speech data for research purposes is still faced with challenging issues (Hannun et al., 2014), especially for under-represented languages (e.g. Bhati et al., 2019) and in case of more complex types of speech (pathological speech, multi-speaker recordings, recordings in adverse listening conditions, disfluent, highly reduced spontaneous speech). The aim of segmentation is often different in different research domains: the goals of the researcher in segmenting a speech dataset (precise information about the timing of features of speech) is somewhat (but increasingly) at odds with the big-data oriented requirements of modern commercial ASR research (Jurafsky & Martin, 2008; for zero-resourced languages there are alternatives, see e.g. Prasad et al., 2019). Furthermore, as long as completely automatic approaches are unable to deliver the reliability that researchers seek, human intervention will remain essential. A serious drawback of human intervention is its repetitive and time consuming character, putting it at risk of poor task execution, and therefore unreliable data.

In this article, we discuss POnSS (Pipeline for Online Speech Segmentation), a system we have created and used for segmentation work for a number of recent studies involving large-scale segmentation (Rodd, Bosker, ten Bosch, et al., 2019a, see Chapter 4; Rodd et al., 2020, see Chapter 5; Rodd et al., under review, see Chapter 6). With POnSS, we sought to improve the efficiency of the word segmentation task for human annotators. The aim of POnSS differs from, for instance, EMU (Winkelmann et al., 2017) in that we focus on optimising a single task that takes a large amount of annotator time, rather than developing a fully featured speech data management system.

POnSS achieves its efficiency through combining forced alignment with manual checks and correction, an easy to use browser interface and, most innovatively, through subdividing the manual component of the overall task into sub-tasks and distributing them at the level of individual word recordings over annotators. To our knowledge, this task subdivision approach has not been tried before. In constructing POnSS, aside from segmenting our own datasets, our aim was to provide a practical implementation of a distributed, subdivided segmentation system, as well as to evaluate the reliability and efficiency of such an approach. We perform this evaluation in comparison to a conventional segmentation of the same data, performed using TextGrids in the phonetics software Praat (Boersma & Weenink, 2019), after forced-alignment bootstrapping.

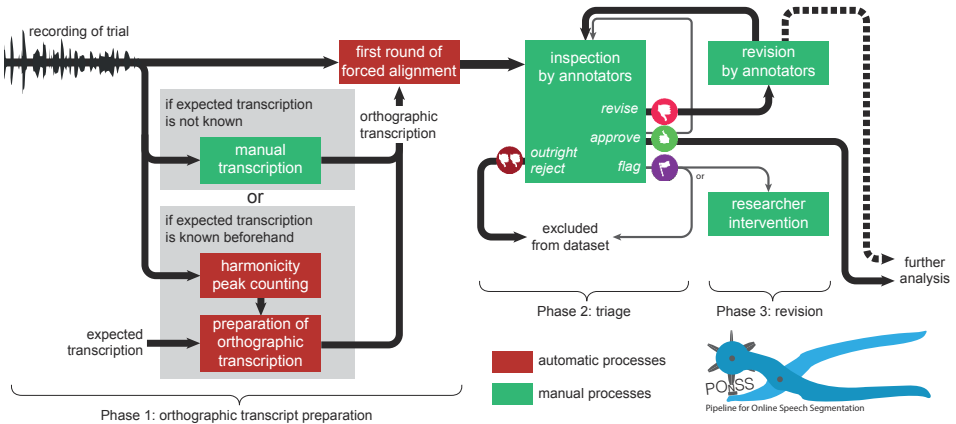


Figure 3.1: A diagrammatic representation of the annotation process. See the text for full details.

The data that we use in the evaluation of POnSS come from Experiment 2 of the PiNCeR corpus (Rodd, Bosker, ten Bosch, et al., 2019a, see Chapter 4). In that experiment, 13 speakers had to name pre-familiarised Dutch '(C)CV.CVC words (e.g., *snavel* ['sna:vəl] “beak”) from line drawings displayed in groups of 8 arranged on a ‘clock face’. A cursor moved clockwise from picture to picture to indicate at which of three trained rates (fast, medium and slow) participants were required to name the pictures. Each trial of the experiment was recorded separately. The task was relatively difficult, meaning that speakers omitted or mispronounced words in many trials. On average, trials contained 6.39 correctly pronounced words that were ultimately analysed, the modal number of included words was 7. Applying POnSS to the PiNCeR data provides a test-case where the words to be produced were known in advance, but not reliably present, a particularly difficult case for forced alignment. This is in contrast to data where it is not reliably known what will be said. POnSS can be useful for this latter type as well, but with a few adjustments, as explained below.

3.2 POnSS

POnSS is a multi-step acoustic analysis and forced alignment pipeline to segment speech materials, intended to be used by a panel of phonetically trained annotators, with each annotator seeing a partially-overlapping part of the dataset. This pipeline is illustrated diagrammatically in Figure 3.1. POnSS divides the work of

speech segmentation into three broad phases; orthographic transcript preparation, triage, and retrimming, each stage combining both manual and automatic processes. The manual processes are standalone, and each unit of work is small, meaning that annotators can themselves choose which of the tasks they do, and for how long, as long as there are materials available to be worked on.

3.2.1 Phase 1: orthographic transcript preparation

The first phase of POnSS is the preparation of an orthographic transcription. POnSS includes both a manual procedure for when the exact word sequence is not known, and a fully automatic procedure for when an expected trial transcription is known beforehand.

Manual transcription

For datasets and experiments where speakers may be particularly errorful in their speech, or where no specific expected transcription exists, POnSS includes a module that facilitates the full manual transcription of the speech data. This approach was used to transcribe the data from Chapter 6, where the vocabulary of possible words was known, but we expected the speakers to make many errors given the longer trials. We expected these frequent errors to make a transcription based on the picture sequence insufficiently reliable for forced alignment. First, silence/pause detection divides the trial recordings into audio chunks with a duration of minimally 5 seconds and maximally 30 seconds. These chunks are inserted into the database.

Annotators use a browser interface (Figure 3.2, left panel) to transcribe each chunk individually, orthographically. Annotators are asked to use real word forms, also in the cases where speakers use reduced pronunciation variants. When the experiment involves a constrained vocabulary of words that can appear, the interface is able to suggest word completions as annotators type, which reduces the number of required keystrokes.

Harmonicity-aided automatic procedure

For datasets where the expected ordering of words is known, POnSS offers a fully automatic transcription generation procedure. This begins with the analysis of the harmonicity (autocorrelation method, default settings) of the trial

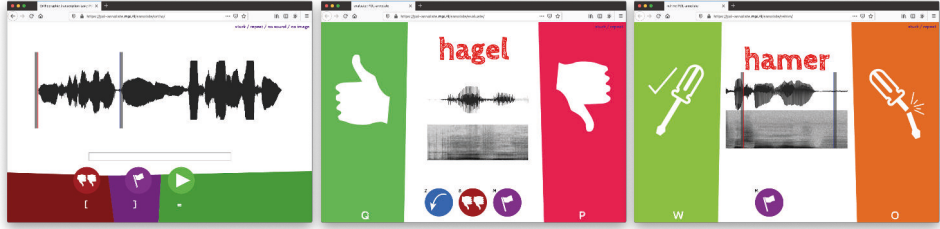


Figure 3.2: Screenshots of the browser interfaces for the orthographic transcription (left), triage (middle) and retrimming tasks (right) in POnSS.

recordings, using Praat (Boersma & Weenink, 2019). In the analysis of the PiNCeR dataset, each harmonicity peak is assumed to correspond to one vowel in the recording, allowing the number of disyllabic words actually produced to be estimated. This may serve as a check for the degree of match between audio and prompted text. In the PiNCeR dataset, in which speakers were asked to pronounce eight disyllabic words in a sequence at varying speaking rates, we observed that when speakers produced fewer than the full eight words, words occurring later in the sequence were much more frequently omitted than earlier ones. Based on this observation, the peak counts were used to produce candidate orthographic transcriptions for use in the forced alignment procedure. In the case of the PiNCeR data, if fifteen or sixteen harmonicity peaks were detected (indicating equally many syllables), all eight words were included in the transcription. If thirteen or fourteen peaks were detected, the first seven words were included, and so on. This is done with the aim of achieving better forced alignment results than simply forced aligning against the orthographic transcription including all eight words would do.

Forced alignment

Once an orthographic transcription is available of a trial or chunk, forced alignment is performed using the application programming interface (API) to web-MAUS (Schiel, 2015), which offers good quality forced alignment for Dutch and other languages using HTK (Young et al., 2006). The resulting word onsets and offsets are used to cut out the audio chunks related to individual words from the longer trial audio recording. We term the resulting labelled chunks of audio ‘word candidates’, since we cannot yet be sure of the accuracy of the segmentation.

3.2.2 Phase 2: triage

In the triage phase, annotators use a browser interface in which each word candidate is presented individually, displaying the transcription, waveform, and spectrogram. The spectrogram and waveform include a ‘shoulder’ of adjacent material either side of the word candidate, made translucent in the waveform (See Figure 3.2, middle panel). The audio plays immediately on loading, and can be replayed as often as required by pressing the tab key or clicking on the waveform. Words are selected randomly from the stack of word candidates that still need to be triaged. The annotator’s task is to choose from one of four options:

- (1) Mark the word candidate as correctly annotated; our annotators were instructed to decide whether the “complete word is isolated, with no extraneous material included”. (thumbs up in middle panel of Figure 3.2)
- (2) Mark the word candidate as requiring further attention in the retrimming phase. (thumbs down)
- (3) Discard the word candidate because it contains non-speech, for instance environmental noise or a cough. (double thumbs down)
- (4) Mark the word candidate as requiring manual intervention, for instance because a speech error (such as a mispronunciation or naming a different word) was made. In our case, these words were also excluded, but POnSS can collect them for later intervention by the researcher. (flag)

Each of these options is associated with a button in the browser interface and associated with a specific key. As soon as a decision is made, the interface automatically proceeds to the next word candidate.

Depending on decisions made by the researchers, word candidates that are marked as good are either returned to the ‘stack’ to be checked again until the word candidate has been approved by a defined quorum of annotators, or removed from the stack and enter the dataset. In our case, we set a target that 20% of the word candidates should be triaged more than once. Which word candidates that passed the triage were revisited was decided randomly.

3.2.3 Phase 3: retrimming

In the retrimming phase, the onset and offset boundaries of the fraction of word candidates that were marked by annotators as requiring retrimming are ad-

justed. Again, a browser interface was used (Figure 3.2, right panel). The label, spectrogram and waveform of each word candidate are again presented on screen. This time, the annotator drags the onset and offset boundaries with the mouse to correct the segmentation. They have three options:

- (1) Report that they successfully corrected the segmentation (screwdriver with check mark in right panel of Figure 3.2)
- (2) Request that the word candidate should return, with more margin (snapped screwdriver)
- (3) Mark the word candidate as requiring manual intervention, for instance because a speech error was made. In our case, these words were also excluded, but POnSS can collect them for later intervention by the researcher (flag)

Depending on researcher-controlled settings, word candidates that annotators report as successfully corrected can be returned to the triage ‘stack’ to be double checked, or they can be removed from the stack and enter the dataset.

3.2.4 Computational implementation

Most components of POnSS are implemented in Python as a web application using the Django framework (Holovaty & Kaplan-Moss, 2009). The interfaces themselves are implemented using HTML, CSS and JavaScript. In-progress segmentation data, along with all meta-data about the annotators’ interaction with the system are stored in a PostgreSQL database.

Although POnSS at present has its own PostgreSQL back-end, elements of the pipeline and the orthographic transcription, triaging and retrimming task interfaces could be relatively easily coupled to another speech data management system, such as EMU-SDMS (Winkelmann et al., 2017).

All code implementing POnSS is available at <https://git.io/Jexj3>, along with the supplementary materials.

3.3 Baseline manual segmentation method

We designed a baseline task that is typical for the type of segmentation projects that are conducted for production data in psycholinguistics (for instance, Zormpa

et al., 2019; Sjerps et al., 2019), combining forced-alignment and Praat TextGrid annotation.

We selected a sample of 468 trial recordings from Experiment 2 of the PiNCeR corpus (Rodd, Bosker, ten Bosch, et al., 2019a, see Chapter 4) that were balanced for speaking rate and speaker. These trial recordings were forced-aligned using webMAUS (Schiel, 2015), based on the expected word productions, and Praat TextGrids were prepared with the forced-alignment result. A panel of seven trained annotators, all of whom were native speakers of Dutch, were asked to correct the MAUS transcriptions of all of the trials in the sample in Praat. They were employed as research assistants and worked on this project as part of their paid work. In contrast to typical practice, where only one annotator looks at each recording, in this case, all seven annotators looked at all 468 trials. A script in Praat selected an audio file and the corresponding preprocessed TextGrid and opened both. Annotators were asked to check the boundaries for word onset and offset and move them if necessary, and check the labelling of the words. Annotators clicked on a “continue” button to save the adjusted TextGrid and load data for the next trial.

3.4 Assessing the reliability of transcription data

Like all human-derived data generation processes, speech segmentation / annotation procedures are liable to various kinds of unreliability. Although one intuitively understands what it means for data to be reliable, formalising this into a working definition is less straightforward. A frequent definition is that reliability is ‘the consistency with which a measure assesses a given trait’ (e.g. Bartko & Carpenter, 1976), framing reliability as synonymous with reproducibility. In the domain of speech segmentation, this definition implies we should be assessing how consistent annotators are in the boundary time stamps that they assign. This could be operationalised within annotators working on the same dataset multiple times (as a kind of test-retest reliability) or between annotators (as a kind of inter-rater reliability).

Relatively little attention has been given to the concept of reliability in the temporal dimension of speech data annotation, with discussion of (un)reliability primarily focused on the label dimension (e.g. Gut & Bayerl, 2004; Widlöcher & Mathet, 2012; Mathet & Widlöcher, 2011; Yoon et al., 2004; Mathet et al., 2015).

Outside the speech domain, a number of inter-rater reliability coefficients are prevalent (Popping, 1988). Many such coefficients are constructed with the assumption of categorically distinct data, assume precisely two raters, or assume that all raters will look at each case. A few coefficients are proposed as being suitable for continuous data, notably intraclass correlation (ICC; Bartko, 1966) and Krippendorff's alpha (Krippendorff, 1970; Hayes & Krippendorff, 2007). Krippendorff's alpha is broadly applicable to data of different forms, suitable for an arbitrary number of annotators, and tolerant of missing data. An alpha value of 1 indicates perfect reliability, an alpha value of 0 indicates the absence of reliability. Negative alpha values indicate above-chance systematic disagreement. In practice, standardised reliability coefficients have not gained traction in speech research, and it is typical to calculate the percentage of segmentations that fall within some tolerance relative to another annotator's segmentation, or relative to a gold standard segmentation, which may be hard to motivate (Ernestus et al., 2015; Raymond et al., 2002; Kipp et al., 1997).

Because of its broad applicability and comparability, we initially selected Krippendorff's alpha as the metric to be used to evaluate POnSS. We intended to use bootstrap re-sampling to create variance in the coefficient, to allow statistical comparison across samples annotated by the baseline method and by POnSS. However, we found disturbingly little variation in the alpha coefficients that we calculated. To explore this systematically, we set about exploring the properties of Krippendorff's alpha, ICC and 'percentage within tolerance' measures in the context of the baseline annotation data. We did this by adding or removing noise to the individual segmented onset and offset times in the dataset of word segmentations performed with the baseline method, and calculating the coefficients for each 'tweaked' dataset. None of the tested coefficients were able to distinguish between datasets that we had artificially made more or less reliable, with Krippendorff's alpha and ICC essentially exhibiting no variation. These simulations are reported in the supplementary materials.

3.4.1 Distribution fitting approach

Given our conclusion that none of the established reliability metrics offered a sufficiently sensitive way to assess the reliability of our speech segmentation data, we developed an alternative approach based on distribution-fitting. This approach aims to quantify variability by finding the parameters of a model of

the data-generating process that explains the variability in the word boundaries resulting from the segmentation process, rather than deriving a result directly from the outcomes. We consider the distribution of the differences between individual segmented onset and offset times and the median of all onset or offset times recorded for that same word across annotators. This distribution, illustrated in Figure 3.3A, has both a high, narrow peak, and broad tails, and is centred around 0 ms, where there is no difference between an individual segmentation and the median of segmentations of the same material, which led us to fit it as a mixture of overlapping Gaussian distributions.

The model that we fit consists of a mixture of three Gaussian distributions. A Gaussian distribution is defined by two parameters, a central tendency (μ) and a standard deviation (σ). The narrowest Gaussian captures the very best segmentations, where all annotators were in full agreement. This is constrained to σ values between 0.0001 and 2 ms. The second Gaussian captures segmentations that deviated somewhat from the median, constrained to σ values between 1.5 and 8 ms. The third Gaussian captures very poor segmentations, where boundaries were placed a long way from the median, constrained to σ values between 2.5 and 40 ms. The search regions overlap to keep the fitting as data-driven as possible. All three Gaussians have their μ parameter clamped at 0. The relative contribution of each Gaussian to the overall mixture is also parametrised, θ_i is the proportion of the mixture that is contributed by Gaussian i . The θ s must sum to 1.0.

Once the mixture model has been fitted to the data, the resulting σ s, weighted by the θ s, quantify the reliability of the sampled segmentations. These could either be summarised as a weighted sum, or used for inference in a weighted regression, as we do in the next section. Various approaches could be used to fit the mixture model to the data; we used particle swarm optimization to minimize the Kullback-Leibler divergence between the modelled distribution and the sampled distribution.

3.5 Analysis 1: Reliability of modalities

3.5.1 Materials

To construct a dataset to evaluate the performance of POnSS, we combined the data from the baseline manual segmentations described in Section 3.3 and a sub-

set of the word segmentations produced using POnSS for Experiment 2 of the PiNCeR corpus (Rodd, Bosker, ten Bosch, et al., 2019a, see Chapter 4), namely word candidates that had been retrimmed minimally twice (as part of random double work to facilitate this investigation). As far as possible, the same words were used as in the baseline manual task. The panel of 8 paid research assistant annotators who contributed to the POnSS data sample were similar in training and background to those who annotated for the baseline task, and included some of the same research assistants.

For each individual word token, the median word onset time across all annotators and both modalities (POnSS or baseline) was calculated. The same was done for the offset times. For each segmentation, the difference between the segmented onset and offset times and the medians was calculated. A balanced sample was taken for statistical modelling, including 300 onset segmentations and 300 offset segmentations for each modality for each of the speaking rates in the experimental data (fast, medium, slow). This sample is shown in Figure 3.3A. In the distribution for POnSS, there were small peaks at -20 ms and +20 ms. These likely emerged because it was possible to adjust the position of the boundaries during retrimming with the keyboard; pressing shift+left or shift+right moved the boundary 20 ms.

3.5.2 Quantifying differences in reliability

To be able to identify the effects of modality and speaking rate on the fitted sigmas, we prepared a dataset that would allow us to predict the sigmas fitted in the Gaussian mixture model by modality and speaking rate. We constructed subsets of the test dataset that varied in the proportion to which each speaking rate or modality was represented. The proportions were predefined, at approximately 10% to approximately 80%, in steps of 10% for the rate conditions. For the modality conditions, we set the proportions of manual annotation to between 20% and 80%, again in steps of 10%. The different levels were exhaustively combined, meaning that 252 samples were constructed, for instance a sample might contain segmentations that were 30% POnSS segmentations and 70% baseline, 40% from the slow condition, 40% medium and 20% fast; a second sample might be 60% POnSS, 40% baseline, 50% slow, 10% medium and 40% fast.

Next, we performed optimisation to find, for each sample, plausible values for the parameters of the mixture model described in Section 3.4.1. The σ of each

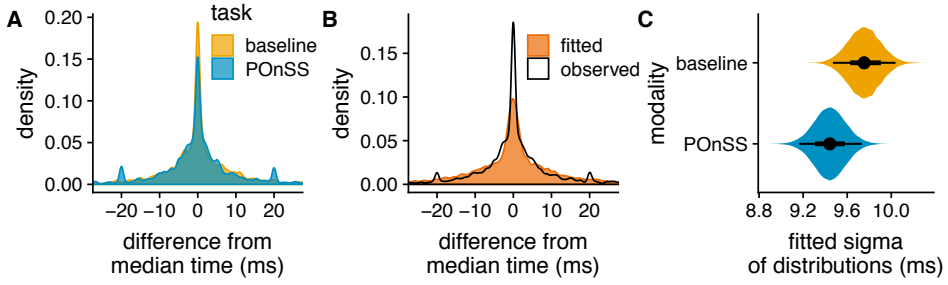


Figure 3.3: Panel A: the observed distributions of the difference between segmented times and the median segmentation for each word, for POnSS and manual annotation modalities (colours). Panel B: an example of the optimized mixture-model fit (orange) to the observed distribution of one of the samples (black line). Panel C: Solid violins show the posteriors of Model 1 for the effect of modality on the sigma, with median (points), 95% HDIs (highest density intervals, thin black lines) and 66% HDIs (thick black lines).

Gaussian was a free parameter, as were the mixing proportions of the Gaussians (θ). The central tendency (μ) was always 0. The quality of the fit was quantified as the mean Kullback-Leibler divergence (KL) between the observed and the fitted distribution, and between the fitted distribution and the observed distribution. Optimisation was performed using the `hydroPSO` implementation of the particle swarm algorithm in R (Zambrano-Bigiarini & Rojas, 2018). 60 particles were simulated for maximally 2000 iterations. The parameter values (θ and σ s) of all 60 particles in the final iteration of the optimisation were recorded, along with the achieved KL for that set of parameter values. In general, good fits are achieved of the fitted distribution to the observed distribution. A sample fit is shown in Figure 3.3B.

3.5.3 Inferential model

We then fitted a Bayesian regression model to quantify the influence of using POnSS rather than the baseline task. This model, and all further statistical models reported, were fitted with the R package `brms` (Bürkner, 2018), allowing us to fit Bayesian mixed-effects models in which the width of the fitted distributions is parametrised. Rather than dealing with binary decisions between significant and not significant, Bayesian regression focuses on quantifying uncertainty about the magnitude of an effect (e.g. Vasishth et al., 2018), so no p -values are reported. Instead, we report the size of the effects we identify, in their rel-

evant units, and where appropriate, standardised for comparability (Cohen's d). All intervals reported are 95% highest density intervals (HDIs).

The model predicted the sigmas fitted in the optimisation phase by the proportions of each modality and each speaking rate represented in the subsets. The interaction between modality and speaking rate was also included. We will refer to this model as Model 1. The model was sampled with the NUTS sampler with 6 chains of 4,000 warm-up and 4,000 test iterations. The model converged for all parameters, as assessed by the Gelman-Rubin diagnostic \hat{R} being within 0.001 of 1.0.

Predictors are included for the proportion of POnSS segmentations, the proportion of segmentations of words from the fast condition and the proportion of segmentations of words from the slow condition. It was not necessary to include the proportion of manual annotations or the proportion of segmentations of words from the medium condition, since these are entirely correlated with the proportion of POnSS segmentations and the sum of the proportion of fast and slow, respectively. This is intuitively comparable to treatment coding of a categorical variable. For each of these linear predictors, a weakly informative prior was specified ($\mu = 0$, $\sigma = 5$). A deviation-coded categorical predictor was included for component (narrow, medium or wide), as were interactions between the categorical and linear predictors. The model fitted a student-t distribution, the σ and ν parameters which were predicted by the component. Regression weights were applied, consisting of the fitted θ values associated with the relevant component, multiplied by $1 -$ the KL score achieved by the fitting. This means that the sigmas of the three mixture components contributed to the main effects in proportion to their weighting, and that the best fits contributed more than worse performing fits. Full details about the model are available in the supplementary materials.

No reliable difference emerged between POnSS and manual segmentations on medium-rate speech: -0.31 ms $[-0.71, 0.084]$, though the central tendency suggests that, had only POnSS segmentations been present in a sample, we would expect to see marginally narrower distributions than in a sample annotated only by the manual method. This effect is depicted in Figure 3.3C. This effect was involved in interactions, such that, with POnSS, reliability was marginally worse in the narrow component: 0.43 ms $[0.14, 0.73]$, in the medium and wide components, the interaction effect was not distinct (medium: -0.0052 ms $[-0.34, 0.34]$;

wide: -0.43 ms [$-1.1, 0.2$]). A figure depicting these interactions is available in the supplementary materials (Figure S9). Had a sample only contained fast speech, we would expect wider distributions: 2.5 ms [$2.1, 2.9$]. No reliable difference emerged between medium and slow speaking rates: 0.051 ms [$-0.36, 0.46$]. There were no reliable interactions between modality and rate (POnSS and fast rate: 0.12 ms [$-0.38, 0.63$]; POnSS and slow rate: -0.33 ms [$-0.79, 0.14$]), suggesting that POnSS is equally reliable across speech that may be assumed to differ in style.

Together, these results indicate that segmentations performed with POnSS are at worst equally as reliable as segmentations performed conventionally using Praat, and potentially slightly better.

3.6 Analysis 2: Efficiency of modalities

To assess the efficiency of POnSS, we calculated how many annotator-hours would be required to yield 5000 segmented words in the baseline modality and using POnSS. In the case of the baseline modality, we assumed that only one annotator would segment each recording. During the manual segmentation, the Praat script recorded the time when each trial recording was opened and saved, meaning that we could calculate the time spent on that trial, and then divide that by the number of segmented words, to result in a duration of annotator time investment per segmented word.

In the POnSS case, we have, for each task, timestamps for the moment the word-candidate was presented to the annotator, and for the moment at which they finished interacting with it. For each original word candidate in the trials from Experiment 2 of the PiNCeR dataset (Rodd, Bosker, ten Bosch, et al., 2019a, see Chapter 4), we identified every interaction with that word recording across both the triage and retrimming tasks. We summed together the time spent triaging and retrimming each word, and then checked to see if that word was accepted into the finished dataset or not.

In an approach akin to bootstrap re-sampling, we subsequently took 1000 samples of 5000 words, and took the sum of the time investment for the words in each sample as the time investment to segment that sample of 5000 words. In the POnSS modality, the sample size was increased iteratively until the sample contained 5000 non-rejected words, to account for the time spent working on words that ultimately were not accepted into the dataset. Around 5% of word

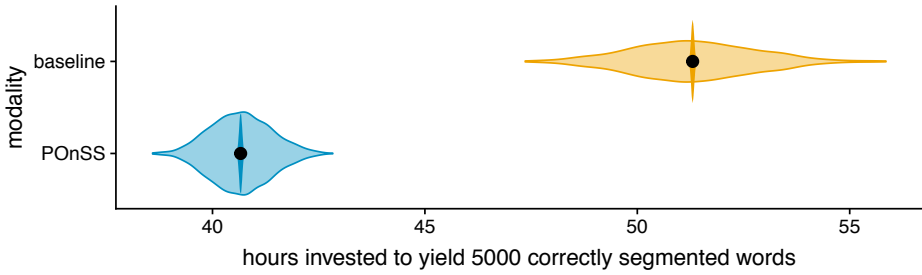


Figure 3.4: Distributions of resampled estimates of time investment required to yield 5000 good words by the two modalities (translucent violins). Overlaid are solid violins showing the posteriors of Model 2 for the effect of modality, with median (points), 95% and 66% HDIs are too narrow to see in the figure.

candidates did not make it into the finished dataset. In the baseline modality, this correction for missing words is already implicitly made, since the time spent working on a trial is divided by the number of resulting good words. The distribution of the time it took to yield 5000 good words is shown in the translucent distributions in Figure 3.4. Note that in this dataset, no manual transcription was required, since we used the harmonicity-aided automatic transcription generation procedure; in the baseline case, MAUS was used, meaning that the analogous part of the task was not used there either.

We fitted these distributions with a Bayesian regression model (Model 2). Like Model 1, Model 2 was sampled with the NUTS sampler with 6 chains of 4,000 warm-up and 4,000 test iterations. The model converged for all parameters, as assessed by the Gelman-Rubin diagnostic \hat{R} being within 0.001 of 1.0. The model predicted the hours invested to yield 5000 correctly segmented words, with a deviation-coded categorical predictor for the modality (1 indicated the baseline method, -1 the POnSS method). A weakly informative prior was set for this predictor, a normal distribution centred at 0 with a σ of 5.44 hours, meaning our expected effect size was 0, with a standard deviation of 1 Cohen’s d . For the model intercept, the prior was a Student- t distribution centred at 45.98, the average of all the data with a σ of 54.42 hours and a ν (degrees of freedom) of 16.33, which are derived by scaling the recommended properties of this prior in `brms` to this dataset.

The distributions of the model coefficients of interest are shown in Figure 3.4, as solid violins. The difference between the two approaches in the time taken to yield 5000 correctly segmented words was very clear (difference between means:

11 hours [11, 11], Cohen’s $d = 2$), such that segmentation using POnSS required much less investment of annotator time than the baseline method.

3.7 Discussion

In this article we introduced POnSS, an online pipeline for the segmentation of speech data. POnSS is optimised for this single task, sacrificing functional flexibility in favour of time/effort efficiency.

We argued that, while fully automatic speech transcription and segmentation is gaining traction, for many purposes human intervention remains essential to ensure data quality in conditions adverse to speech segmentation. A key diagnostic for the quality of a speech segmentation is its reliability, conventionally defined in terms of reproducibility. We explored how two widely employed approaches to measuring reproducibility were sensitive to the kind of variance expected in speech segmentation data. From this analysis, we concluded that neither Krippendorff’s alpha nor simple percentage agreement within a tolerance were ideal ways to assess reliability in speech segmentation data, since they were not sensitive to artificial noise. In their stead, we proposed a reliability-quantification approach based on modelling the underlying error process as a mixture of Gaussian distributions, where the sigmas of the distributions quantify the reliability of the segmentation process.

We then turned to quantifying the consequences of segmenting to the word level with POnSS rather than with a conventional procedure using TextGrids in Praat preceded by naive forced-alignment. We analysed the relative reliability and efficiency of POnSS. These analyses revealed that segmentation with POnSS was approximately equally reliable compared to conventional manual segmentation, and considerably faster. In the reliability analysis, we found that the sigmas fitted to the data segmented by POnSS were comparable to the sigmas fitted to the data segmented conventionally. The efficiency analysis showed that 23% less investment of annotator time was required to yield the same number of acceptable word transcriptions. In the efficiency analysis, the way that we compared the modalities slightly biases against POnSS, since we calculate the time investment based on our practice whereby some word-candidates got triaged and retrimmed multiple times by different annotators, while we assume that under the baseline modality, each word-candidate will only be worked on once.

These findings license the further use of POnSS for segmentation of speech corpora. For the evaluation conducted here, we used data from the PiNCeR corpus (Rodd, Bosker, ten Bosch, et al., 2019a, see Chapter 4). The PiNCeR corpus was a good test case, since it contains experimentally elicited, errorful speech, which is particularly challenging for forced alignment. POnSS also includes a manual transcription component that makes the segmentation of spontaneous speech viable. Read speech resembling, for instance, TIMIT (Garofolo et al., 1993), which forms the basis of the training data for many ASR systems, may be forced aligned well enough to require only minimal checking of a sample to assess the suitability of the segmentation. POnSS could also be trivially adapted to manage this forced alignment and perform this checking.

Aside from reliability and efficiency, a subordinate aim in developing POnSS was to improve the experience of the annotators, who consider segmentation to be the least preferred of the tasks that they perform as research assistants. Anecdotally, the annotators report POnSS to be preferable to work with, compared to conventional segmentation using TextGrids in Praat. This may be due to the colourful visual appearance. Furthermore, with POnSS, the annotator is freed from a number of meta-tasks inherent to conventional segmentation projects, including the necessity to keep track of how far through a project they are and recording this to prevent double work; planning how many trials they can do in the time remaining until the next task begins; and ensuring that their work is saved and archived. Additionally, they have some operational freedom in that they can choose which of the subtasks to perform. Future analysis might examine whether they work longer effective stints with POnSS than with the baseline task. In POnSS it is possible to employ aspects of ‘gamification’, for instance tracking and displaying each individual annotator’s longest streak of triage decisions or retrimmings performed within some time limit to boost motivation, though whether this would come at the cost of reliability would need to be established.

In POnSS, different component tasks of the overall segmentation project are separated out into small, easily explained and understood sub-tasks. This implies that the less taxing triaging task could potentially be adequately performed by entirely untrained annotators, through online crowd-sourcing systems such as Amazon’s Mechanical Turk (Buhrmester et al., 2018), or allowing paid members of an institute’s participant pool to segment data at home at their conve-

nience. This would drastically reduce the wait for the researcher for completed segmentations, and free up trained research assistants for more productive and motivating tasks. Further careful pretesting is required to establish whether crowd-sourced, non-expert triage decisions are of equal quality to expert triage decisions, and to introduce data-quality controls like catch trials with known good answers.

Our aim with POnSS was to provide a practical implementation of a distributed, subdivided segmentation system, to be able to evaluate the efficiency and reliability of such an approach. As such, there are various researcher degrees of freedom, such as the length of chunks in the transcription task and the proportion of word candidates that are triaged and retrimmed multiple times that could influence the reliability and accuracy of the resulting segmentations. Optimal settings for these researcher degrees of freedom need to be explored more fully with various annotator populations and speech data types, which may allow further improvement on the efficiency benefit relative to Praat TextGrids reported here.

The test dataset that we used to evaluate POnSS was evaluated to the word level, and we used other techniques to perform sub-word level analyses (Rodd, Bosker, ten Bosch, et al., 2019b, see Chapter 2). However, there is no a priori reason to think that POnSS would not also perform comparably to conventional segmentation on phoneme- or syllable-level segmentations.

In conclusion, POnSS offers reliable segmentation of speech materials to the word level, in an appealing form that makes efficient use of human input by combining human decisions with forced alignment.

4 PiNCeR: a corpus of cued-rate multiple picture naming in Dutch

PiNCeR is a corpus of speech recordings from Dutch speakers who named pictures at different speaking rates. Participants named pre-familiarised '(C)CV.CVC words (e.g., *snavel* ['sna:.vəl] “beak”) from line drawings displayed in groups of 8 arranged on a ‘clock face’. A cursor moved clockwise from picture to picture to indicate at which of three trained rates (fast, medium and slow) participants were required to name the pictures. Annotation was performed using the POnSS tool (Rodd et al., in press, see Chapter 3), where manual and automatic segmentation is combined to yield accurate word onsets and offsets. To detect the onset and offset times of syllables within words, we identified excursions of above-average acoustic instability between the vowel of the initial syllable and the first consonant of the second syllable (Rodd, Bosker, ten Bosch, et al., 2019b, see Chapter 2). This approach was licensed by careful control of segmental content in the target words to maximise correspondence between acoustics and articulation. The PiNCeR corpus was intended for use in modelling control of speaking rate (Rodd et al., 2020, see Chapter 5), but may be of interest for other purposes. Trial-level recordings from two related experiments are made available for 25 participants (12 for Experiment 1, 13 for Experiment 2), along with the onset and offset times of the words and the syllables.

This chapter was adapted from Rodd, J., Bosker, H. R., ten Bosch, L., Ernestus, M., & Meyer, A. S. (2019a). *PiNCeR: A corpus of cued-rate multiple picture naming in Dutch*. *PsyArXiv*. <https://doi.org/10.31234/osf.io/wyc6h>

The speech materials for 25 participants, consisting of trial-level recordings, along with the onset and offset times of the words and the syllables in csv format and as R data format are archived at the Language Archive, and available on request from <https://hdl.handle.net/1839/7c210d30-bb55-4cbe-9eeb-baf18570460c>

4.1 Background

This paper presents the PiNCeR (Picture Naming at Cued Rates) corpus, which was collected to serve as a dataset for the modelling of cognitive control of speech rate (Rodd et al., 2020, Chapter 5), and may also be of interest for investigation of phonetic variation as a consequence of speech rate change. The corpus contains productions of experimentally elicited disyllabic Dutch words at three predetermined speaking rates, and temporal annotations of word and syllable onsets and offsets. This paper also documents the procedures used in the preparation of the corpus, notably a distributed annotation system (POnSS) that allowed us to efficiently annotate the corpus (Rodd et al., in press, Chapter 3) and a metric that allowed us to detect syllable onset and offset times from the acoustic signal (Rodd, Bosker, ten Bosch, et al., 2019b, Chapter 2).

Two multiple picture naming experiments were conducted, in which the required speaking rate (fast, medium or slow) was indicated with a cueing dot that jumped from picture to picture on a display with 8 pictures. Since the corpus was intended to be used to model cognitive aspects of the preparation of speech, a task that engaged all phases of speech planning before articulation was desirable. Picture naming is the gold standard task for eliciting single word productions, ensuring that all planning phases need to be completed.

In Experiment 1, speakers were explicitly instructed to avoid pausing between words, and instead to adjust their speaking rate by adjusting the duration of the words. In this fashion, we attempted to ensure that we would elicit variation in the way individual words were articulated, rather than variation in the usage of pauses. In Experiment 2, this instruction was not given, to ensure that differences in strategy adopted in the slow speaking rate were the result of speaker-intrinsic processes rather than purely an effect of task.

A Bayesian mixed effects regression analysis was run to characterise the durations of words in the different speaking rates, to assess their compliance with the required rate, and to verify whether the different instructions given to participants in each experiment resulted in different word durations, which would indicate different task strategies.

4.2 Methods

4.2.1 Experiment 1

The speech was elicited with a multiple picture naming task, forcing speakers to complete all planning phases before articulation of each picture name could begin. Different sets of eight pre-familiarised line drawings were displayed in each trial, in an arrangement reminiscent of a clock face (c.f. Meyer et al., 2012). A cursor indicated which picture was to be named, moving in a clockwise direction from picture to picture at three predetermined, participant-independent rates: fast, medium, and slow.

Participants

Native Dutch speaking participants ($N = 12$, two males, ten females, $M_{age} = 22$ years) with normal hearing and normal or corrected-to-normal vision were recruited from the participant pool of the Max Planck Institute for Psycholinguistics, with informed consent as approved by the Ethics Committee of the Social Sciences Faculty of Radboud University (Project Code: ECSW2014-1003-196).

Materials

Twelve disyllabic Dutch concrete nouns with stress on the first syllable were selected as target words for the production experiment. The first syllable was always open, and the second syllable was always closed (C(C)V.CVC, where C = consonant, V = vowel). Vowels were always monophthongs and consonants were never stops. This means that we selected only segments where the articulators do not move during the production of the segment, in contrast to diphthongs or stop consonants, where changing articulatory configuration during the segment is inherent to the segment identity. This is required for the derivation of the onset and offset times of syllables within words. Additionally, words with an ambisyllabic consonant were excluded. This yields words such as *snavel* ['s-na:.vəl] “beak”, *vriezer* ['vri:.zər] “freezer” and *wafel* ['wa:fəl] “waffle”. In addition, twelve similar filler words were selected. A full list is provided in the Appendix on page 167.

Lists of 70 sets of eight words were pseudo-randomly created from the vocabulary of filler and target words, one set for each of the 70 trials in each rate

condition. A different list was used for each participant in each rate condition. Within each set, no word appeared more than once. Within each list, the number of times each word was used was matched as closely as possible (average s.d. in used lists: 0.4769, minimum frequency 26 words per 560, maximum frequency 28 words per 560), as was the frequency with which each word appeared in each of the five ‘target’ positions on the clock face (average s.d. in used lists: 2.107), and the frequency of each pair of words co-occurring in a set (an analogue of transition probability, average s.d. in used lists: 2.294).

For each word, a line drawing was either taken from the Snodgrass and Vanderwart (1980) picture library, or prepared in the same style.

Experimental procedure

Participants were tested individually in a sound attenuated booth. Stimulus presentation, eye-tracker synchronisation and audio recording were controlled by Presentation software (Version 16.5; Neurobehavioral Systems, Berkeley, CA, USA). A Sennheiser ME64 directional microphone was used to record the participants’ speech at a sampling rate of 48 kHz.

The session began with familiarisation of the pictures and their names, by means of (1) a printed card and (2) naming of the pictures as they were displayed individually on screen, in a pseudo-randomised order with two repetitions of each picture. The experimenter immediately gave the correct name when the participant named a picture incorrectly. After the structure of the experiment was described (three blocks, each at a different rate condition, in a random order), the participant was instructed to “*name the exact picture that the marker indicates*”. They were instructed to achieve slow speech rates by slowing down their speech, not by producing longer pauses in between words. In this fashion, we attempted to ensure that we would elicit variation in the way individual words were articulated, rather than variation in the usage of pauses. Instructions were presented on screen. Six practice trials at the medium rate then followed, after which the remote eye-tracker (Eyelink 1000 in remote mode; SR Research, Ottawa, ON, Canada) was prepared and calibrated with a standard 9-point calibration procedure. The gaze position measurements, originally collected with future computational simulation work in mind, are not discussed in this article.

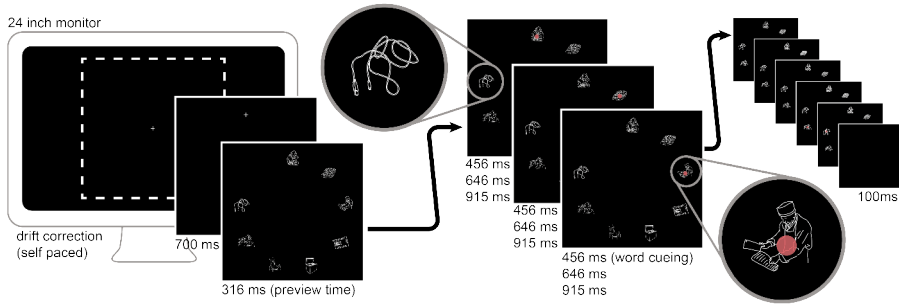


Figure 4.1: The trial sequence. The trial began with a drift-correction fixation cross (self-paced). A fixation cross was then presented at the location of the first picture for 700 ms, followed by 316 ms preview time. Cueing then began: each word was cued for 456, 646 or 915 ms by overlaying a translucent red dot on the relevant picture. The trial concluded with a blank display for 100 ms.

A block of seventy trials was presented for each rate condition, followed by a short break. The order of the three rate blocks presented in the experimental session was counterbalanced across participants.

The trial structure is illustrated in Figure 4.1. Before each trial, the participant performed a self-paced drift-correction procedure for the eye tracking measurements. After successful drift-correction, a fixation cross was presented at the location of the first picture (“12 o’clock”) for a duration of 700 ms. Then, the pictures appeared without the cursor, and were presented for 316 ms of ‘preview time’, to allow the participant to prepare for naming.

The pictures were displayed in sets of eight, in a clock-face arrangement with 9 positions. Positions 2 to 6 were occupied by target pictures. The first, seventh and eighth positions were occupied by filler pictures, since these positions were expected to be particularly susceptible to listing intonation. The ninth position (at “10 o’clock”) was always left empty to visually reinforce the beginning and end of the sequence of pictures. This arrangement is illustrated in 4.1. The whole display fitted into an area of 780 x 780 pixels. Each picture was scaled such that it would occupy an area of 90 x 90 pixels.

Once the preview time had elapsed, a cursor was overlaid on each picture in turn for the duration appropriate to the rate condition; fast, medium or slow. The cursor was a translucent red circle with a diameter of 20 pixels, which appeared in the centre of each picture whilst that picture was to be named. The cursor jumped from picture to picture, starting with the topmost and proceed-

ing in a clockwise direction. After all the pictures had been cued, the pictures and the cursor disappeared and a blank screen was presented for 100 ms, after which the drift correction procedure for the next trial started immediately.

The three cueing rates tested were 456 ms/word (2.19 Hz, fast condition), 646 ms/word (1.54 Hz, medium condition) and 915 ms/word (1.09 Hz, slow condition). These rates were derived from the non-cued speaking rates realised by three further participants in a small pre-test (research assistants with a similar background to the participants tested in the main study). This pre-test data was also used to establish an appropriate length of ‘preview time’ to allow the participants to prepare for the naming task.

4.2.2 Experiment 2

The second elicitation experiment was identical to the first experiment, except participants were given no explicit instruction to avoid pausing. For Experiment 2, 13 further participants were tested (two males, eleven females, $M_{age} = 22$ years), recruited from the same pool of native Dutch speakers as the participants tested in Experiment 1, under the same ethics approval.

4.2.3 Word boundary finding

The extent of the speech data collected (5,250 trials, yielding up to 26,250 target words and 15,750 filler words if no errors were made) precluded fully manual annotation. A fully automatic annotation was also not possible since the nature of the task resulted in many hesitations, omissions and deviations from the canonical productions. Instead, a multi-step acoustic analysis and forced alignment pipeline, POnSS (Pipeline for Online Speech Segmentation), was used to create automatic transcriptions of the speech materials, which were then adjusted as necessary by a panel of ten phonetically trained annotators, including research assistants and the first author. POnSS was developed and validated by Rodd et al. (in press, Chapter 3). For completeness, we also describe it here. Use of the pipeline results in equivalently reliable transcriptions compared to conventional annotation with Praat software, with greater annotator comfort and greater time efficiency.

The POnSS pipeline is illustrated for an example word in Figure 4.2. First, the harmonicity (autocorrelation method, default settings) of the trial recordings

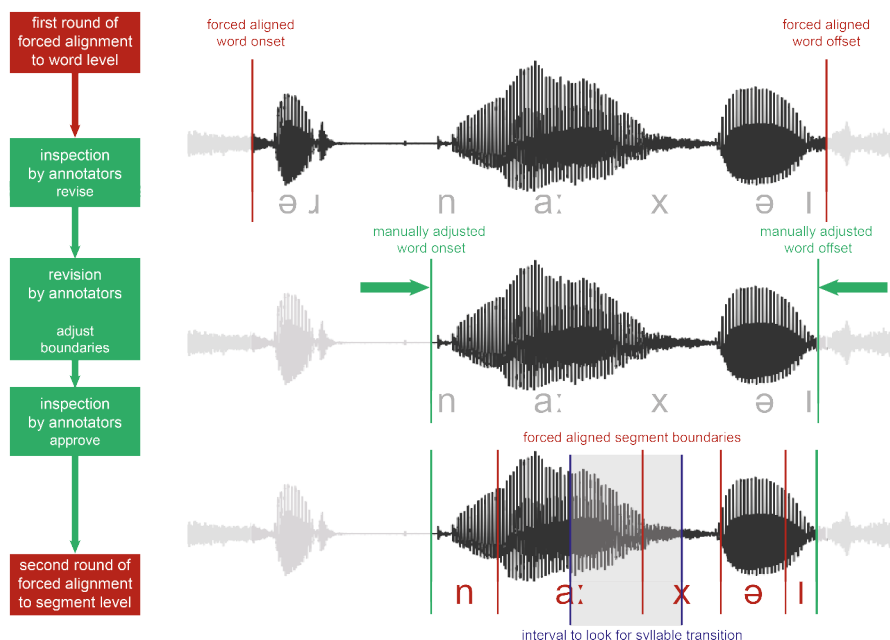


Figure 4.2: An example of the pipeline for annotating the word *nagel* [ˈnaːxəl] “fingernail”. First, an the initial forced alignment run identifies candidate word boundaries. These are inspected by a human annotator. In this case, they are wrong, so the word is marked as needing revision. Later, the same or another annotator adjusts the boundaries. The revised word is checked again, and approved. Then, forced alignment is applied to the single word recording, to identify segment boundaries. An interval is defined, spanning from the centre of the vowel of the first syllable to the centre of the first consonant of the second syllable, as identified by the segment level forced alignment. This interval is used to direct the search for the syllable transition, using the metric developed by Rodd, Bosker, ten Bosch, et al. (2019b, Chapter 2).

was analysed using Praat software (Version 6.0.18, Boersma & Weenink, 2015). Each harmonicity peak can be assumed to correspond to one vowel in the recording, allowing the number of disyllabic words produced (i.e. not omitted) to be estimated. We observed that when speakers produced fewer than the full eight words, the words occurring later in the sequence were much more frequently omitted than earlier ones. Based on this observation, the peak counts were used to produce candidate orthographic transcriptions for the forced alignment. If fifteen or sixteen harmonicity peaks were detected (indicating sixteen syllables), all eight words were included in the transcription. If there were fourteen or fifteen, the first seven words were included, and so on. This was done with the aim of achieving better forced alignment results than simply forced aligning against the ‘script’ including all eight words would have done. From these candidate orthographic transcriptions, forced alignment to the word-level was performed using the MAUS software (Schiel, 2015), which offers good quality forced alignment for Dutch using HTK (Young et al., 2006).

A specially constructed web application using the Django framework (Holvaty & Kaplan-Moss, 2009) was used by the annotators to screen out words that had been poorly aligned or labelled by MAUS and therefore needed revision. Each annotation was presented individually with the waveform and spectrogram of the relevant audio. Annotators could listen to the audio as many times as they wished. For each of the 23,218 annotations produced by MAUS, they decided whether the complete word was isolated, with no material from surrounding words included. If that was not the case (because, for example, some part of the word was missing, or part of the following word was included), they flagged the annotation as requiring further attention. They also had the option to discard annotations containing non-speech or speech errors, 1,095 annotations were discarded for this reason.

The annotations that were flagged by any one of the annotators, but were not outright discarded (81.5% of 5,400 words from the fast rate; 64.5% of 7,864 words from the medium rate; 45.6% of 7,812 words from the slow rate) were subsequently re-trimmed by other annotators from the panel. This was done by dragging word boundaries on a visual display of the waveform and spectrogram.

4.2.4 Automatic syllable boundary finding

After annotation, syllable onset and offset times were derived using the automatic metric developed and validated by Rodd, Bosker, ten Bosch, et al. (2019b), using an analysis interval spanning from the centre of the vowel of the first syllable to the centre of the first consonant of the second syllable, as identified by the phone-level forced alignment. Syllable planning units tend to overlap, so a method was required to identify the onsets and offsets of syllables where they overlap with neighbouring syllables. The dynamics of the acoustics of speech broadly reflect the dynamics of the articulation that produces it: when the configuration of the articulators is stable, the acoustic signal is also stable. It was therefore possible to identify periods of articulatory stability from the acoustic signal, and periods of transition. We interpreted the period of acoustic transition (the *acoustically evident planning unit overlap*) as coterminous with the period of planning unit overlap, allowing us to identify the onsets and offsets of planning units from the acoustic signal. A similar approach was adopted by Hoang and Wang (2015) to identify phone transitions.

4.3 Confirmatory analysis: word duration

To confirm that participants were indeed performing the task as we expected, that is, primarily modulating speaking rate rather than merely adjusting pause durations, we first examined overall word durations.

A Bayesian mixed effects model was constructed using the *brms* R package (Bürkner, 2018; R Development Core Team, 2008; Stan Development Team, 2018) to model the log-transformed word duration. We used the log-transformed word duration in order to reduce skewness in the distribution.

Dummy coded fixed effects of cueing rate (categorical predictor, dummy coded with medium rate on the intercept) and experiment (categorical predictor, dummy coded with Experiment 1 on the intercept) were included in the model, along with the interaction of cueing rate by experiment. Random intercepts were included for speaker. Random slopes were included for the log-transformed trial-level residual rate for each speaker-cueing rate combination, grouped by experiment. The trial-level residual rate was calculated as the difference between the realised speaking rate in a trial (the total contiguous speaking time divided by the number of words produced) and the target rate (the duration for which each

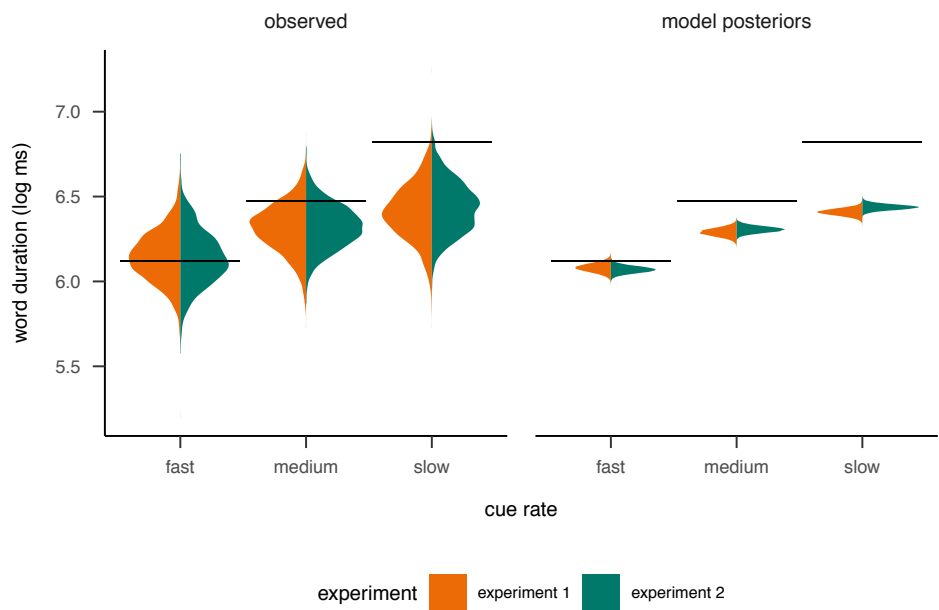


Figure 4.3: Left panel: word duration (log ms) as a function of cued speaking rate, plotted as violins (the width of the violin shows the distribution of values). The colour indicates which experiment the data come from. Black horizontal lines indicate the target speaking rates. Right panel: the model posterior distributions for the mean, shown as violins.

word was cued with the cursor). For the cueing rate predictor, low-informative priors were set centred at each target speaking rate, with a σ of 3.4 log ms (equivalent to 30 linear ms). For the effect of experiment, a normally distributed low-informative prior was set, centred at 0 with a σ of 3.4 log ms (equivalent to 30 linear ms, roughly five times the noise prevalent in the annotation task; Rodd et al., in press). The model was well converged (assessed by the Gelman-Rubin diagnostic \hat{R} , effective number of samples and visual inspection of traceplots) after running eight chains of 3,000 warm-up and 3,000 critical iterations.

The observed durations of the words produced in each condition are presented in the left panel in Figure 4.3, measured from the onsets and offsets established by the annotation procedure described in section 2.4. The posterior distributions for the means of each speaking rate in each experiment are shown in the right panel, along with HDI (highest density interval) covering 95% of the posterior.

The results of the Bayesian mixed effects model are summarised in Tables 4.1 and 4.2. The model confirmed that, in both experiments, speakers produced shorter words in the fast condition than in the medium condition, and longer words in the slow condition than in the medium with large effects (Cohen's d minimally 0.687, maximally 1.366). Since there were large differences in word duration between each rate condition, we concluded that the participants produced different speaking rates for each cueing condition. However, in all cases, the speakers produced words somewhat shorter than the target rate. This effect is smallest for the fast cueing condition and largest for the slow cueing condition. This arises because the target rates assume continuous production without pauses between words. This suggests that speakers were, even when explicitly asked to try to modulate their word duration, also modulating pause duration to comply with the cued speaking rate.

The model also confirmed that there was no effect of experiment, since all 95% credible intervals included 0, and all effect sizes were small (Cohen's d minimally 0.061, maximally 0.168), and all 95% credible intervals overlapped with a ROPE (region of practical equivalence; Kruschke, 2018) defined to include all effects smaller than 15ms, a reasonable estimate of the degree of noise prevalent in annotation data (Rodd et al., in press). This means that the word durations measured from speakers instructed to try to avoid pausing between words did not differ from those who did not receive this instruction.

Table 4.1: Results of the Bayesian mixed effects model for comparisons of realised word duration by cued rate, within experiments.

experiment	comparison	estimate	CI	Cohen's d
experiment 1	medium → fast	-0.210	[-0.172, -0.247]	-1.213
experiment 1	medium → slow	0.119	[0.156, 0.081]	0.687
experiment 2	medium → fast	-0.236	[-0.199, -0.275]	-1.366
experiment 2	medium → slow	0.132	[0.169, 0.093]	0.761

Table 4.2: Results of the Bayesian mixed effects model for comparisons of realised word duration by experiments, within cued rates.

cued rate	comparison	estimate	CI	Cohen's d	ROPE %
medium	exp. 1 → exp. 2	0.016	[0.055, -0.022]	0.093	99.49%
fast	exp. 1 → exp. 2	-0.010	[0.032, -0.054]	-0.061	99.36%
slow	exp. 1 → exp. 2	0.029	[0.069, -0.01]	0.168	97.6%

5 Control of speaking rate is achieved by switching between qualitatively distinct cognitive ‘gaits’: Evidence from simulation

That speakers can vary their speaking rate is evident, but how they accomplish this has hardly been studied. Consider this analogy: when walking, speed can be continuously increased, within limits, but to speed up further, humans must run. Are there multiple qualitatively distinct speech ‘gaits’ that resemble walking and running? Or is control achieved by continuous modulation of a single gait? This study investigates these possibilities through simulations of a new connectionist computational model of the cognitive process of speech production, EPONA, that borrows from Dell, Burger, and Svec’s model (1997, *Psychol. Rev.* 104(1), 123). The model has parameters that can be adjusted to fit the temporal characteristics of speech at different speaking rates. We trained the model on a corpus of disyllabic Dutch words produced at different speaking rates. During training, different clusters of parameter values (regimes) were identified for different speaking rates. In a one gait system, the regimes used to achieve fast and slow speech are qualitatively similar, but quantitatively different. In a multiple gait system, there is no linear relationship between the parameter settings associated with each gait, resulting in an abrupt shift in parameter values to move from speaking slowly to speaking fast. After training, the model achieved good fits in all three speaking rates. The parameter settings associated with each speaking rate were not linearly related, suggesting the presence of cognitive gaits. Thus, we provide the first computationally explicit account of the ability to modulate the speech production system to achieve different speaking styles.

This chapter was adapted from Rodd, J., Bosker, H. R., Ernestus, M., Alday, P. M., Meyer, A. S., & ten Bosch, L. (2020). Control of speaking rate is achieved by switching between qualitatively distinct cognitive ‘gaits’: Evidence from simulation. *Psychological Review*, 127(2), 281–304. <https://doi.org/10.1037/rev0000172>.

Code and supplementary materials are available at <https://osf.io/3mqgu/>

5.1 Introduction

Speaking is a uniquely human behaviour. It is by nature temporal: concepts and ideas are encoded as a stream of rapidly fluctuating sound, and the correct ordering and duration of the components is of crucial importance for intelligibility and conveying meaning. At the same time, there is great variability in the timing of speech sounds: different speakers have different habitual speech rates, and individual speakers can vary their speech rate from situation to situation, and even within utterances in the same conversation (e.g. Miller et al., 1984; Quené, 2008). A portion of this variation presumably arises to accommodate different communicative situations: speakers may slow down to provide listeners with sufficient time to extract the necessary details from the acoustic signal (e.g. Lindblom, 1990; Bosker & Cooke, 2018; Cooke et al., 2014). Alternatively, they may speed up, for instance to convey more content in the same period of time. Listeners use speech rate information in shaping their perception (Maslowski et al., 2019a; Kaufeld et al., 2020; Dilley & Pitt, 2010), making control of speech rate an essential communicative skill.

The fact that humans have control over the rate at which they speak means that they are capable of adjusting the cognitive apparatus that plans speech, from the selection of words to the tightly coordinated movements of the articulators of the vocal tract. Understanding how speech planning is controlled can give us insights into how the apparatus itself works. Given the large degree of speaker-controlled variability in speech, identifying the mechanisms of control over speech planning is also important in its own right. In the present study, we examine the control processes speakers may engage to achieve different speech rates.

Speech production is classically characterised as a modular, feed-forward processing system (e.g. Dell & O'Seaghdha, 1992; Levelt et al., 1999; Levelt, 1989; Stemberger, 1985). After a meaning representation has been selected ('conceptualisation'), the lexical selection stage begins, where abstract representations of words that best correspond to the conceptual message are selected. Processes of word form encoding then construct detailed word form representations. These stages together can be considered as a formulation phase. Once a word form representation is selected, a motor execution phase is entered, where movement commands for the articulatory apparatus (e.g., the tongue, lips, vocal chords) are

calculated, carried out, and monitored (Guenther, 2016a; Tourville & Guenther, 2011). Because speakers typically plan as late as possible, rather than storing a pre-planned utterance in working memory (e.g., Damian & Dumay, 2007; Kello et al., 2000; Levelt, 1989; Levelt et al., 1999), the formulation system must keep up with the desired rate of articulation, requiring modulation of its operation to maintain synchronisation.

5.1.1 'Gaits' in speech production

In a working model of the production system with formulation and execution phases, adjustment in speaking rate results from adjusting the state of the formulation system; to speak slowly we shift to a regime that results in slow speech and to speak fast we shift to a regime that causes speech to emerge more quickly. How are these regimes related to each other? How does the regime invoked to produce slow speech differ from the regime invoked to produce medium rate speech?

The control mechanisms engaged to regulate speaking rate at the level of utterance planning and preparation are largely unknown. A more concrete and readily observable system that operates at a continuously varying range of speeds is that of human and animal locomotion. In humans, walking and running gaits are adopted to achieve movement at different speeds. The movement patterns of walking and running are qualitatively different; in walking, at least one foot is on the ground at all times, whilst in running, both feet are raised from the ground simultaneously for part of the cycle (Minetti, 1998; Alexander, 1989). A continuous range of movement speeds can be achieved by firstly increasing the speed of walking, and then switching to a running gait to speed up further. Alongside hard limits on feasibility of certain gaits at certain speeds, the selection of locomotive gaits is tightly linked to their relative efficiency. In horses, which typically have walking, trotting, and galloping gaits, each gait has a clear 'sweet spot' speed, at the approximate centre of the range of speeds achievable with that gait, where exertion (ml O₂ consumed to move 1 metre) is minimised (Hoyt & Taylor, 1981, their Figure 2). Horses and migratory animals select these speeds preferentially (Pennycuik, 1975), and avoid the inefficient speeds in the shoulder of each gait. This feature of gaited systems previously inspired speech researchers working at the level of articulatory movements, who link qualitatively different mechanical realisations of speech movements to their relative

efficiency to achieve a required standard of intelligibility (e.g. Pouplier, 2012), as predicted by the hyper- and hypo- articulation theory (Lindblom, 1990).

Pouplier (2012) related the metabolic equivalence of the optima of the locomotive gaits to speaking, conceptualising the gaits of speech as equally optimal coordination modes, suitable for different contexts. This holds well for the execution phase of speech production, which incorporates motor planning and articulation, where there are ‘many roads to Rome’: different gestural coordination configurations, which are chosen between according to local context, can lead to acoustic outcomes that are equivalent for the listener. For instance, speakers can make use of alternative vocal tract configurations to achieve speech sounds when articulatory freedom is constrained (Lindblom et al., 1977). Immediately adjacent speech sounds also condition the selection of alternative articulatory configurations, so as to minimize the articulator movement required (e.g. Boyce & Espy-Wilson, 1997). This reconfiguration can be thought of as analogous to switching between gaits in locomotion.

More global contextual factors such as prosody and speech rate can also lead to gestural reconfiguration in the execution component, for instance in coda consonant resyllabification, whereby a consonant may be realised in a way more similar to an onset consonant (Scobbie & Pouplier, 2010) in rate-scaling experiments. Similarly, anti-phase synchronisation of gestures tends to reconfigure to in-phase synchronisation as rate increases (Kelso et al., 1986); for instance in West Andalusian Spanish, Parrell (2012) finds that speakers shift from anti-phase oral-glottal coordination in sequences like [‘ka.^hta] from /casta/, “*caste*” (with preaspiration before the [t]) to in-phase coordination [‘ka.t^ha], by making the tongue articulation of the /t/ earlier so it occurs at the same time as the glottal opening.

Alternatively, the speech planning apparatus might be purely linearly up- or down-regulated in response to changes in required speaking rate. This is the case for motor tasks where temporal precision is required, in both gross motor movements (Wright & Meyer, 1983), and fine movement requiring extensive coordination, such as piano playing (Bella & Palmer, 2011).

5.1.2 Approach adopted in this study

We extend the metaphor of gaitedness to the *psychological* system of speaking rate control. We ask whether there are multiple cognitive gaits in speech plan-

ning that resemble locomotive gaits. Without a choice of gaits, the cognitive regimes adopted to achieve different speaking rates would be similar in nature, but only quantitatively different. In other words, the difference between the regimes required to produce slow and medium speech would be similar to the difference between the regimes required to produce medium and fast speech. This is akin to only having one gait, which can be sped up or slowed down linearly. Alternatively, with multiple gaits of speech planning, the regimes would differ from each other in a non-linear way, with a qualitative difference between, for instance, the regimes adopted for slow rates (walk-speaking) and the regimes adopted for fast rates (run-speaking).

We address the question of how speakers control speech rate. More concretely, we aimed to ascertain how the cognitive regimes that are associated with each speaking rate relate to each other, to assess if multiple gaits might be present. To do this, we constructed a family of computationally implemented connectionist models of the formulation phase of speech planning (strand 1), and explored how each model variant could be optimised to mimic the temporal properties of natural word productions taken from a speech corpus elicited at different cued speaking rates. We then evaluated the performance of the optimised model variants. This process allowed us to identify optimal model parameter settings associated with producing speech at a given rate, which provide a window onto the arrangement of the regimes of the underlying cognitive systems (strand 2).

Computational model (strand 1)

A computational model of the speech planning system provides a psycholinguistic sandbox to explore how the regimes adopted to achieve speech at different speaking rates relate to each other. We propose such a computational model, EPONA. EPONA has parameters that determine its behaviour (controlling features such as rate of activation spreading, rate of activation decay, and connection weightings). These parameters can be optimised to cause the model to optimally fit speech data produced at different speaking rates. The sets of parameter values chosen by the model for each rate condition mirror the regimes of the cognitive system that the model emulates. More concretely, we adopted an optimisation procedure which identified the parameter values required to fit the distributions of three durational features measured from elicited productions of

disyllabic words: first syllable durations, second syllable durations and overlap durations. The distributions of these durational features together form a ‘fingerprint’ of the regime of the speech production system engaged to achieve that speaking rate. This process was repeated for three different speaking rates: fast, medium and slow.

The theoretical model that we selected as inspiration for EPONA is that of Dell et al. (1997). The model is a good starting point since it captures the ability to produce sequences of elements from a hierarchical structure. The model separates the encoding of the segmental content of the word from the encoding of the metrical structure (the ordering and timing of the segmental content, and supra-segmental content such as word stress). EPONA inherits this property.

How do regimes relate to each other? (strand 2)

The parameters of the EPONA model can be thought to represent the regimes of the cognitive system that underlie natural speech production at different rates. The different regimes of the system exist as locations in a multi-dimensional ‘parameter space’, where the parameters form the dimensions.

With a sample of three speaking rates, and assuming that each rate is associated with a single regime, there are five logical possibilities for how the regimes might be arranged with respect to each other. (1) The cognitive system has a single gait, and different speaking rates are achieved by continuous adjustment of this single gait. This is akin to only walking, but walking at three different speeds. (2) The cognitive system has three gaits, one for each speaking rate. These three gaits are qualitatively different, like walking, trotting, and galloping in horse locomotion. The cognitive system has two gaits, grouping the medium speaking rate with either the slow rate (3) or the fast rate (4). Finally, (5) The cognitive system has two gaits, a habitual gait adopted for the medium speaking rate, and an exceptional gait adopted for slow and fast speaking rates. This fifth option supposes that there is a default gait for the most frequently used speed, and that a fall-back ‘all purpose’ gait is adopted for other rates.

In the single-gait scenario, the three regimes would be arranged along a single axis in parameter space. In a multiple gait scenario, the three regimes would be arranged in a triangle in parameter space. Each side of the triangle is potentially the axis of a gait to which two regimes belong. For each axis, if both speaking rate regimes belong to the same gait, we would expect a continuous, linear vari-

ation in the predictions of models fitted at points along the axis. If, however, the two regimes belong to different gaits, we would expect to see a non-linearity at some point along the axis, indicating a shift from the area of parameter space associated with one gait to the area of parameter space associated with the other. To distinguish between the single and multiple gait scenarios, we examined the results of the optimisation procedure undertaken in strand 1 to identify the arrangement of the regimes in parameter space. To distinguish between various two- and three-gait scenarios, we fitted additional models at points along the axes between the three regimes, and assessed the predicted ‘fingerprints’ for (non)-linearity by means of Bayesian statistical modelling.

In Section 5.2, we will discuss previous approaches to modelling serial ordering in speech production. The mechanics of the proposed model are discussed in Section 5.3. We then present the corpus of speech data that we test against, in Section 5.4. We then turn to the methods and results applied to answer the research questions of each of the strands in turn in Section 5.5 (strand 1) and Section 5.6 (strand 2).

5.2 Serial order in speech production and the Dell et al. (1997) model

The core task of the formulation process is to ensure that after a lexical concept becomes active at the conceptual-formulation frontier, the gestural scores required to produce it become active at the frontier between formulation and motor execution. In this article, we will follow Levelt et al. (1999) and Tourville and Guenther (2011) in assuming that the gestural score representation encodes the relative onset and offset times of abstract gestures (comparable with the gestures described by e.g. Browman & Goldstein, 1992) of a single syllable, and that this representation is shared by formulation and motor execution to allow activation to spread. In the execution component, a more concrete motor plan and auditory and somatosensory expectations are retrieved for this gestural score (Tourville & Guenther, 2011; Guenther, 2016b).

A naive connectionist model of this process might assume direct connections from each word node to the relevant syllable nodes. Asking such a model to predict the temporal organisation of a multisyllabic word such as the Dutch word *snavel* /'sna:.vəl/ ‘beak’, however, will fail: /'sna:/ and /vəl/ will become active

simultaneously. A successful model therefore needs to account for serial order; the fact that sequences of speech sounds are overwhelmingly often produced in the correct order (one or two errors per 1,000 words; Garnham et al., 1981), despite the subunits of each word presumably being activated from a single word-level parent node.

It is not trivial to construct a model that, in response to activation in a single parent node, can activate and then deactivate child elements in a sequence in turn. In the speech production domain, the most prominent model to deal with serial ordering is that of Dell, Burger and Svec (1997, hereafter the DBS model). Dell et al. enumerate the requirements of serial ordering: preparation of the future, activation of the present and suppression of the past. That is, an ideal model should (1) prime upcoming syllables, (2) activate them at the correct time and (3) deactivate them once they have been produced.

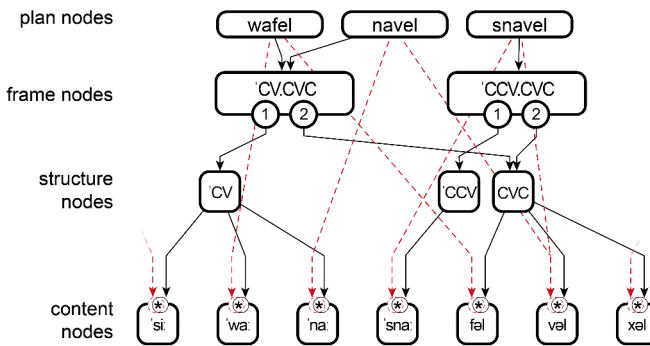


Figure 5.1: An instance of the EPONA model containing the nodes necessary to produce the Dutch disyllabic words *wafel* ['wa:.fəl] 'waffle', *navel* ['na:.vəl] 'navel' and *snavel* ['sna:.vəl] 'beak'. The segmental route is shown with red dashed connections. At the top level, there is a unique plan node for each word. Frame nodes are shared between words with the same metrical structure (*wafel* and *navel* both have a 'CV.CVC structure, so are connected to the same frame node). Each frame node has multiple output ports (here numbered 1 and 2), one associated with each child element of the sequence. Each port is connected to a structure node. In turn, each structure node is connected to all content nodes representing syllables with the relevant metrical structure. Structure nodes and content nodes are also shared between words. Multiplication in the content nodes (represented by asterisks) ensures that only syllables receiving input from both routes become active.

An example instantiation of the EPONA model capable of producing three Dutch disyllabic words is illustrated in Figure 5.1. The word-level input 'plan

nodes' are shown at the top of the model. At the bottom of the model are the syllable-level gestural score 'content nodes'. In between, there are two top-down routes along which activation can flow. The first route connects the plan nodes directly to the content nodes (shown with dashed red lines in Figure 5.1). The connections of this route are responsible for encoding the segmental content of the word, so we term it the 'segmental route'. The second route is responsible for maintaining correct serial order of syllables and encoding the metrical structure of the words by means of a frame node, which represents the word-level metrical structure, so we term it the 'metrical route'. The concept of separating the planning of segmental content and metrical structure into separate streams and employing a frame to enforce serial order is well established in framed-based psycholinguistic models of the production system (Bock, 1982; Dell, 1986; Garrett, 1976; Levelt, 1989; MacKay, 1972; Shattuck-Hufnagel, 1979; Stemberger, 1991). Note that throughout this article, C indicates a consonant, V indicates a vowel, ' indicates the syllable with primary stress, while . indicates the syllable boundary.

Frame-based models have two key advantages compared to models without them. Firstly, because they separate information about sequential ordering from segmental information, they can explain the ordering of novel sequences without additional learning: if the correct frame and the correct content are known, previous separate experience with the frame and the content can be combined to produce the sequence correctly. Secondly, they account for the observation that errors where sub-elements are misordered within a sequence are overwhelmingly outnumbered by errors where elements from the same position in the sequence exchange ('caterpillar' → 'patterkiller') or are copied between adjacent sequences. A model without frames would predict much more frequent misorderings of the elements within a sequence than is observed (Boomer & Laver, 1968; MacKay, 1970; Vousden et al., 2000; Vousden & Maylor, 2006).

The metrical route is shown in Figure 5.1 with solid black arrows. Aside from the frame node, there are structure nodes, which are connected to all content nodes sharing a metrical structure at the syllable level. The first connection in the metrical route passes activation from the plan node to the relevant frame node. The frame node has an output port for each syllable in the word, so in our case, two ports. The first port is connected to a structure node for the metrical shape of the first syllable of the word. The second port is connected to a struc-

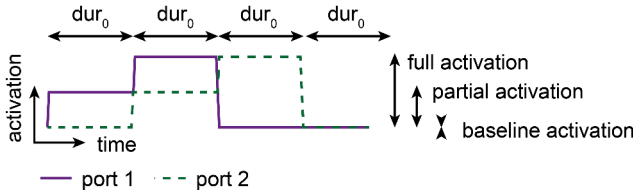


Figure 5.2: The activation patterns produced by the frame node for port 1 (purple, solid) and port 2 (green, dashed).

ture node for the metrical shape of the second syllable of the word. A mechanism within the frame node ensures the activation initially flows primarily from the first port, and subsequently from the second port; we will address the nature of this mechanism and the activation flows it generates shortly. The structure nodes therefore receive activation asynchronously: first the structure node representing the shape of the first syllable becomes active, and then the structure node representing the shape of the second syllable. The structure nodes spread their activation to all the content nodes that share that structure. In the content nodes, the incoming activation from the metrical route is multiplied by the incoming activation from the segmental route, meaning that non-zero activation must be received from both streams for the content node to become activated. The activation in the content nodes can be considered to be the output of the DBS model.

We will now turn to the frame node, which generates activation streams for each syllable in response to receiving activation from the word node above it. The DBS model is agnostic regarding the precise nature of the serial order mechanism employed in the frame node. Rather than including a pure-connectionist mechanism such as a competitive queue in the frame node (e.g. Hurlstone et al., 2014), Dell et al. (1997) construct a transparent model that exhibits serial-order behaviour. This has the advantage of simplicity and interpretability.

In the EPONA, the frame nodes directly produce parametrically defined activation patterns for each of the ports after they receive activation from the plan node. The ports can produce activation at three (parametrically defined) activation levels: baseline activation, partial activation, and full activation.

Activation is produced at these levels in a specific order (depicted in Figure 5.2). Before word onset, both ports produce baseline activation. The activation pattern for the first port (solid lines) begins with a period of partial acti-

vation, then a period of full activation, then baseline activation. The activation pattern for the second port (dashed) begins with baseline activation, then partial activation, then full activation, then baseline activation. The second port is therefore producing the same pattern as the first port, but delayed by the duration of one period. The partial activation level is proposed by Dell et al. (1997) as a means to prime the ‘future’ (the next content to be produced). The full activation level is associated with activating the ‘present’ (the content currently being produced). The baseline activation state serves as the baseline for ports connected to items that have not yet been produced, and is also associated with deactivating the ‘past’ (content that has already been produced).

5.3 Mechanics of the model

Dell et al. (1997) describe a mechanism that accounts for serial order behaviour in speech production. They used the model to predict probabilities of speech errors. Error probabilities were calculated directly from predicted activation levels. To do so, it was not necessary to extract precise onset and offset times from the model. Rather than examining errors, we seek to understand how speakers adjust their speaking rate in correct utterances. To do so, we propose EPONA, a model that borrows its conception and underlying connectionist architecture from DBS. EPONA is able to predict the onset and offset times of syllable level planning units, and to model differences between speaking rates. EPONA differs from DBS in the specification of the timing behaviour of the frame node, and extends it to add a rudimentary operationalisation of the execution component. EPONA is implemented computationally, and is tested with speech timing data, rather than speech error proportions.

5.3.1 Timing in the frame node

The DBS model assumes that all the periods of the activation patterns associated with the ports of the frame node have equal duration. A model with this assumption is sufficient for the prediction of the rate of serial order errors, but it is improbable that such a model will be successful in fitting the relative onset and offset times of syllables in real speech, where the durations of syllables in a word are rarely equal (varying as a product of, among other things, the number of segments, the specific segments involved, the stress status of the syllable).

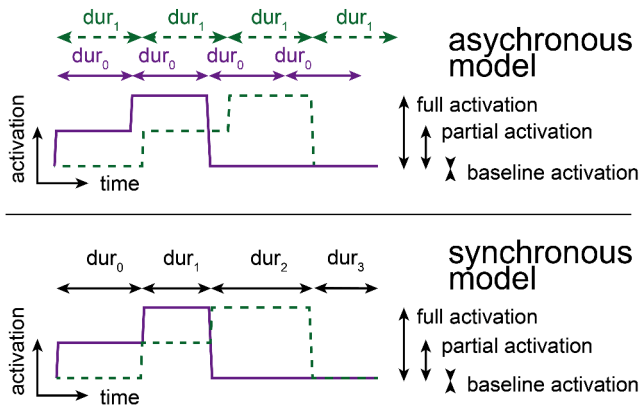


Figure 5.3: The activation patterns produced by the frame node for port 1 (purple, solid) and port 2 (green, dashed) in the asynchronous and synchronous model variants. The duration of each step in the activation patterns is controlled by various parameters, depending on the model variant (such as dur_0 , see text for full details).

ble, and phonological processes such as final lengthening: Booij, 1995; Sloatweg, 1988; Cambier-Langeveld et al., 1997). There are (at least) two ways that this constraint could be relaxed to allow the duration of full activation on each port to differ (and thus the overt production of each syllable), which should make the frame node more effective in encoding the metrical properties of the word shape it represents. These possibilities are described in the remainder of this section. We construct variants of EPONA consistent with each possibility.

The present implementation of EPONA produces only disyllabic words, but the mechanisms described here could be adapted to produce more syllables. In the following descriptions, we again assume a model producing disyllabic words, and refer to two frame node output ports, though, of course, frame nodes encoding the metrical structure of words with more syllables are also possible, where further ports would be required.

Asynchronous model

The first option to relax the equal duration constraint is to allow the durations of the periods of the activation pattern associated with each output port to differ. Thus, under this variant, the two ports are potentially out of sync relative to each other after word onset, because one parameter controls the durations of the activation periods output by the first port, and the other parameter controls

the durations of the output of the second port. An example of a possible set of frame output patterns produced by this variant is depicted in the upper cell in Figure 5.3. This variant requires two parameters: dur_0 and dur_1 . These control the duration in ticks of all phases of the output of port 1 and port 2 respectively.

Synchronous model

Alternatively, synchronisation between the activation patterns could be maintained, such that when port one is outputting full activation, port two is outputting partial activation, but allowing the durations of each pair of steps to differ. This means that both ports always switch activation level at the same moment, but the amount of time that elapses between these switching events may vary. An example of a possible set of frame output patterns produced by this variant is shown in the lower cell in Figure 5.3. This variant has four duration parameters: dur_0 , dur_1 , dur_2 , and dur_3 , defining the duration of four phases that occur simultaneously in both output patterns – that is, the parameters all have influence on the activation patterns emitted from both ports. The parameter dur_0 defines the duration of the first phase, where port 1 outputs partial activation and port 2 outputs baseline activation. The duration of the second phase, where port 1 outputs full activation and port 2 outputs partial activation is specified by dur_1 . The duration of the third phase, where port 1 outputs baseline activation and port 2 outputs full activation is defined by dur_2 . The duration of the final phase, where both ports output baseline activation, is defined by dur_3 .

Control model

We also constructed a control model variant that retains the timing structure described by Dell et al. (1997). The model variant performed poorly relative to the asynchronous and synchronous model variants, as expected. Full details about the control model variant are available in the online supplementary materials.

5.3.2 Execution component

To calculate the onsets and offsets of each syllable, we need to connect a model of the execution phase of speech production to the formulation phase. Our conception of the execution phase is straightforward; we assume that the duration

of strong activation of a syllable output node is linearly related to the duration of articulation of that syllable (c.f. Tourville & Guenther, 2011). To identify strong activation, we compare the activation of each syllable node over time to a syllable-specific threshold. Syllable-specific thresholds are used to account for variability in the magnitude of activation change in each syllable position. When the activation first exceeds this threshold, we consider syllable production to start, and when it decreases below the threshold again, we consider syllable production to stop. This procedure is fully specified in Section 5.5.1, and is functionally equivalent to assuming that execution faithfully reproduces the temporal dynamics of formulation, and that continuing activation from formulation is necessary during articulation.

5.3.3 Computational implementation

The EPONA model is programmed in Python 3, using the NetworkX library (Hagberg et al., 2008, version 1.11), in which nodes and connections between them are defined and the spread of activation from node to node can be computed as a function of time. The optimisation and learning of the model is also programmed in Python, using the Platypus library (Hadka, 2017, version as of April 2017).

5.4 Speech corpus

The model requires speech data to compare against. In this case, speech data were taken from the PiNCeR corpus gathered by Rodd, Bosker, ten Bosch, et al. (2019a, Chapter 4), which contains speech recordings and is annotated for word and syllable onset and offset times in ('CV.CVC and 'CCV.CVC) disyllabic Dutch words. The speech was elicited by means of cued picture naming, whereby twelve speakers named pre-familiarised line drawings presented in sets of eight on a 'clock face' display. The words that were elicited are provided in the online supplementary materials. The picture to be named was indicated by a cueing dot, which moved clockwise from picture to picture, at slow (915 ms/word, 1.09 Hz), medium (646 ms/word, 1.56 Hz) and fast (456 ms/word, 2.19 Hz) rates. These speaking rates were selected on the basis of a pilot experiment where speakers were not cued, but instead encouraged to speed up or slow down as much as they could. These rates fall within the range of rates measured in the Switchboard

corpus of spontaneous speech (Greenberg et al., 2003), but are all slower than the median rate in that corpus, and are slower than an estimate of mean rate for Dutch speakers of similar demographics (Quené, 2008). This is likely because the picture naming task, which included only middle-to-low frequent concrete nouns, was relatively hard compared to conversational speech, which includes many closed class words that are fast to plan.

The word onset and offset times were obtained by a multi-step process. First, forced alignment using MAUS (Schiel, 2015) was applied to each trial (set of eight pictures). The resulting word boundaries were subsequently checked by a panel of experienced annotators, who evaluated whether the segmentation was accurate or not. Finally, the panel of annotators adjusted the boundaries of words that were marked as inaccurate in the previous step. Since the words were disyllabic, the onset of the first syllable and the onset of the word were simultaneous, and the offset of the second syllable and the offset of the word were simultaneous. To detect the onset of the second syllable, and the offset of the first syllable, a metric was employed to quantify the stability of the acoustic signal. Heightened acoustic instability was equated with temporal overlap between the gestural score encoding the first syllable and the gestural score encoding the second syllable. For further details about this metric, see Rodd, Bosker, ten Bosch, et al. (2019b, Chapter 2).

The corpus contains 4,023, 3,575, and 2,627 word tokens for the slow, medium, and fast rate conditions, respectively. The size of the corpus sections differ primarily due to more frequent speaker error and less successful forced alignment in the faster conditions. However, within each speaking rate section, the remaining tokens were evenly distributed across the target words, and the proportion of 'CV.CVC versus 'CCV.CVC words was comparable between the corpus sections (29.7%, 29.9%, 30.6% 'CV.CVC words for fast, medium and slow rates, respectively).¹

The distributions of the first and second syllables and the overlap between them are shown for each cueing rate condition in Figure 5.4. The rate conditions differed significantly on all of these metrics (Rodd, Bosker, ten Bosch, et al., 2019a). Also, the durations of the second syllable are consistently longer than the durations of the first syllable, which is to be expected given the metrical struc-

¹Note that statistical testing to confirm whether or not the corpus sections differed is not appropriate, since the sets of words here are closed populations, rather than samples from some larger population (Sassenhagen & Alday, 2016).

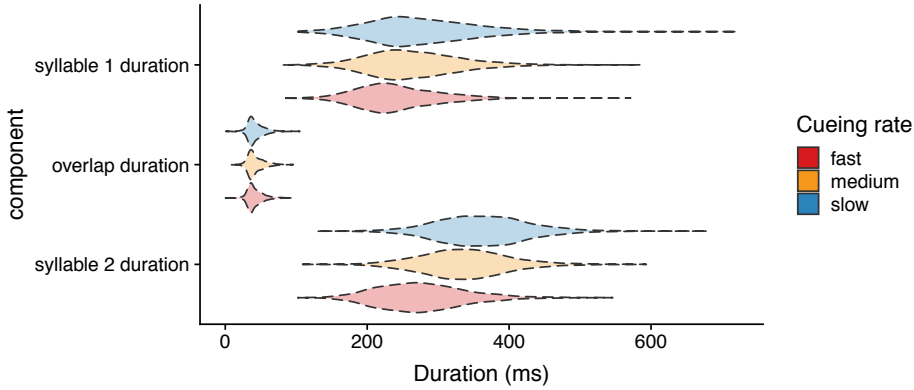


Figure 5.4: The distributions (violins) of the durations measured in the PiNCeR corpus, separated by rate condition. These form the three ‘fingerprint’ distributions that the model seeks to mimic.

ture was consistently either 'CV.CVC or 'CCV.CVC, and that an utterance-final lengthening process in Dutch tends to counteract reduction of unstressed final syllables in utterance final position (Booij, 1995; Sloodweg, 1988).

5.5 Training and testing the computational model (strand 1)

Strand 1 concerns the construction of a family of computationally implemented connectionist models of the formulation phase of speech planning, optimisation of the model variants to mimic temporal properties of natural speech production, and evaluation of the performance of the model variants.

5.5.1 Methods: evaluating the performance of model variants

Our aim in strand 2 was to apply simulation of the EPONA to reveal how the cognitive system underpinning speech production can be modulated to achieve speech at different speaking rates. To achieve this, we require the model to simulate the performance of human speakers using different rates in strand 1. However, it is not straightforward to evaluate how well a model simulates human speech production.

We consider the set of distributions of first and second syllable duration and overlap duration in each rate condition of the PiNCeR corpus as a ‘fingerprint’ of the speech production system operating at that speaking rate (see Figure 5.4). Together, the fingerprint distributions capture more about the regimes of the

speech production system than only the average durations of the syllables and the overlap between them would do, because the variation present in the durations is a product of variability inherent to the production system operating in a given regime. Since we are not concerned with individual differences between participants, but, instead with characterising the regimes of the speech production system more broadly, we collapse across the 12 speakers when constructing the fingerprint distributions. The distributions of the data in the corpus shown in the violins in Figure 5.4 are therefore identical to the fingerprint distributions used to fit the models. Model optimisation is then conducted independently for each speaking rate.

Optimisation procedure

The Platypus (Hadka, 2017) implementation of the NSGAIII (Deb & Jain, 2014) algorithm was used to find the best parameter values in each speaking rate for each model architecture. The fitting procedure is depicted in Figure 5.5. The optimiser must find a set of parameter values that produce a prediction that is a good fit for all three fingerprint distributions simultaneously. In line with the optimisation literature, we will term such a set of parameter values a *solution*. Since the model produces a distribution for each of the three fingerprint distributions, we obtain three estimates of fit quality for each solution tested: one for each distribution. In the optimisation literature, such a quality estimate that is to be maximised or minimised is termed an *objective*. We obtain independent estimates of fit quality, in the form of the Kullback-Leibler (KL) divergence for each objective.

The KL divergence is a commonly used measure of the dissimilarity of two distributions, where a lower KL divergence indicates more similar distributions. By definition, its magnitude is dependent on the variability of the observed distribution. In our case, the variability of the observed duration distributions differs substantially between the three objectives. This means that the scales of the KL divergences calculated for each of the three objectives are not directly arithmetically comparable. We have no theoretical reason to prefer that the model concentrate on learning to fit one of the objectives ahead of the others, but simply summing (or averaging) the KL divergences would place undue weight on one of the objectives. We must therefore consider all three objectives together. Such an optimisation problem with multiple independent estimates

of fit quality (or objectives) that cannot be straightforwardly collapsed is known as a multi-objective problem. Typically, there is no single solution that is optimal for all objectives: solutions that work well for one objective may be poor for another. Instead, the optimisation algorithm aims to identify the solutions that are *Pareto efficient*, that is, the fit that they achieve for one objective cannot be improved upon without worsening the fit for one of the other objectives. This set of Pareto efficient solutions is termed the *Pareto front*.

Alongside the complication of multiple objectives, our models also have multiple free parameters to be optimised (between 11 and 14 depending on model variant; a full listing of parameters is available in the online supplementary materials), and are computationally expensive (time consuming) to evaluate because we simulate activation spreading through the network for each and every solution, and require multiple repetitions to simulate the fingerprint distributions. A complex error landscape with more than a handful of free parameters can prove difficult to search effectively; a classical method such as grid search, where evenly spaced points in the parameter space are sampled, requires prohibitively many model evaluations to get good coverage, and still runs the risk of missing good solutions between the sampled points. We suspected that our parameter space might be quite complex, containing multiple clusters of good solutions in each rate condition. For these reasons, we selected NSGAIII. NSGAIII belongs to a class of optimisation algorithms that accumulate knowledge about the search space over time (in multiple ‘generations’ of learning ‘agents’). This means that the search can become gradually more focused on promising regions of the space. NSGAIII combines the ability to solve multi-objective problems with active preservation of diversity in the solutions it retains from generation to generation, making it suitable to search a space with many local minima. Other search methods such as Particle Swarm Optimisation have a tendency to converge early: that is, they are poor at exploring spaces where there are local minima (plausible solutions that are good, but not as good as the best solution in the space) at different positions (Kennedy, 2011; Peer et al., 2003).

In the remainder of this section, we will discuss the workings of the evolutionary algorithm in more detail, and then describe the procedure for evaluating the fit of the models, and the procedure employed to train the models.

- 1) spawn 464 agents (generation 1) from normal distributions centred on plausible fitting coefficient settings.

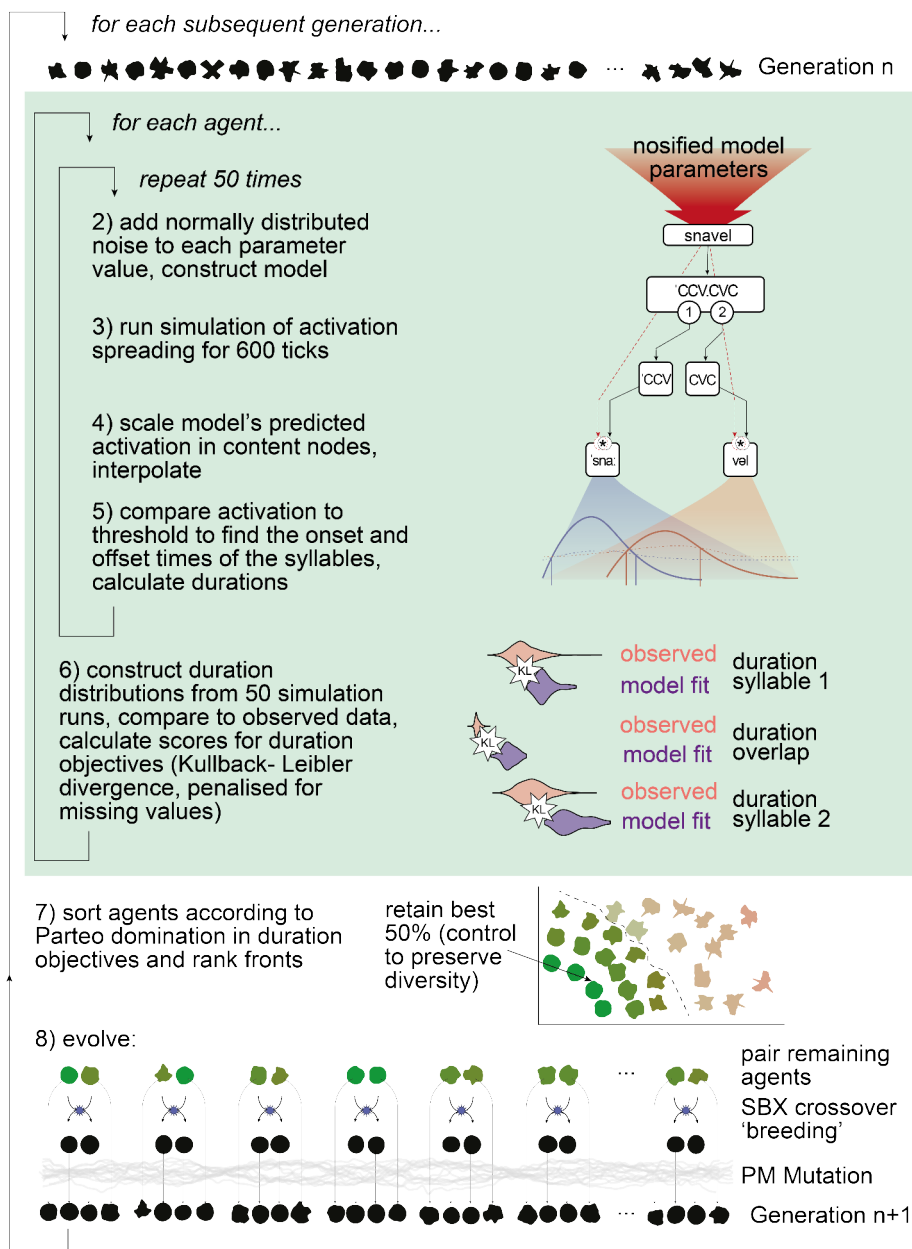


Figure 5.5: A diagrammatic representation of the fitting process. See the text for full details.

Evolutionary algorithm

The NSGAIII (Deb & Jain, 2014) algorithm mimics evolution in biology. The evolutionary process begins by spawning a population of agents. An agent is a carrier of a ‘genome’ (set of parameter μ and σ values, which define the central tendency and spread of distributions associated with each model parameter) that interacts with other agents to explore the parameter space. In each generation, the genome of the agent is varied somewhat by processes of mutation. Therefore each agent tests a different solution in each generation.

At the start of the optimisation procedure, we spawn 464 agents, a population size recommended by the Platypus package based on the number of free parameters of the model variant with the most free parameters (Hadka, 2017). For the first generation, the parameter μ values for each agent (that agent’s genome) are sampled from relatively broad normal distributions centred around values that we identified in pre-testing as producing plausible activation sequences (step 1 in Figure 5.5).

The model is then evaluated using the parameter μ and σ values associated with each agent for that generation, resulting in a fitness score for each fingerprint distribution for that solution. The simulation of the model and the procedure for evaluating a solution are described in Section 5.5.1; steps 2 to 6 in Figure 5.5. The fitness scores are Kullback-Leibler divergences between the observed and predicted fingerprint distributions.

Once all agents in the generation are evaluated the Pareto optimal solutions are selected. Formally, a solution b can be said to dominate another solution a (denoted $a \prec b$) if it has a lower score on at least one objective whilst not having a higher score than a on any objective. The Pareto front is therefore the set of solutions that are not dominated by any other solution. The solutions of this first ‘Pareto front’ are assigned a rank of 0. From the remaining unranked population, a new set of solutions that are Pareto optimal in the smaller population are identified, and assigned a rank of 1. This procedure is repeated to find subsequent fronts, with the agents in the third front being assigned a rank of 2, and so on, until all agents are ranked.

The agents are then entered into selection ‘tournaments’, in which two agents are randomly drawn from the population, compared, and the agent with the lower rank is retained. The losing agent is discarded from the population and no longer contributes to future generations. Further tournaments are performed

until all agents have competed once (step 7 in Figure 5.5). This has the advantageous effect that all agents from the best rank will be retained, all agents from the worst rank will be excluded, and that agents from the ranks in between have a gradually decreasing chance of being retained. This, along with a further mechanism to preserve agents in under-represented parts of the parameter space (Deb & Jain, 2014, p. 582), means that the retained agents represent, broadly, the best half of the initial population, but that, simultaneously, variability is maintained, which ensures that the optimisation procedure searches the ‘bumpy’ parameter space effectively.

Then, the evolution stage begins (step 8). The remaining agents are randomly paired up and recombined to make offspring by the Simulated Binary Crossover operator (Deb et al., 2007; Deb & Agrawal, 1995), which simulates the mixing of two genomes in sexual reproduction. For each pair of parents, for each value in the set of parameter μ or σ , a polynomial probability distribution is constructed around each parent value. Two sets of child values are then sampled from the mixture distribution (Deb & Agrawal, 1995). This results in child agents that combine traits from each parent agent. The parents and the children together form the population for the next generation of evaluation, competition and recombination, after having been subjected to further random mutation by the Polynomial Mutation operator, where a perturbation is sampled for each parameter μ or σ value from a polynomial distribution centred at zero (Deb & Agrawal, 1995; Deb & Goyal, 1996). Because of this mutation step, specific solutions are usually not repeated in subsequent generations, and the overall fitness of a next generation may be worse than a previous generation, but in general the optimisation procedure will result in improved scores over time. In our implementation, 5000 generations were run (see Section 5.5.1 for more details).

Evaluation of a solution

The process of evaluating the set of parameter μ and σ values associated with an agent is illustrated in Figure 5.5 (in the green box), and described in detail below. The aim of this evaluation procedure is to assess how well each set of parameter μ and σ values mimics the observed fingerprint distributions. This requires us to construct a distribution of each of these variables.

To construct the predicted distributions, we run the model 50 times with each set of parameter μ and σ values. In each of the 50 repetitions, a small amount

of noise is added to each parameter μ value, sampled from a normal distribution centered at 0, the standard deviation of which is defined by the parameter σ value. These noisified model parameters are used to construct an instance of the model variant to be tested, with node properties and connection weights defined by the model parameters (step 2; see also the online supplementary materials).

The model keeps time internally using a unit that is 9 ms long, a ‘tick’; activations are recalculated once per tick. This value was arrived at by pretesting with models where the number of ms that each tick represents was learnt along with the other parameters. In the simulations reported here, the duration (in ms) of a tick was held constant across word productions. A unit of this order of magnitude is convenient because it allows sufficiently detailed sampling (e.g. the shortest segments are still represented by several ticks) but allows faster computation than a shorter tick length (c.f. typical window shift of 10 ms in MFCC measurements, Young et al., 2006).

Each model is run for 600 ticks (that is, we calculated the activations in the network 600 times) which amounts to 5400 ms, a duration long enough for the word to be produced and the activation of all nodes in the network to return to baseline, whatever the speech rate condition.

Activation of the plan node always occurs after 4 ticks, and persists for 28 ticks at a constant activation level determined by a model parameter. After 28 ticks, the activation in the plan node decays, at a decay rate determined by a model parameter. These values were also arrived at during pretesting, where these parameters were allowed to vary. Holding these parameters constant across conditions ensures that the differences between speech rates emerge in the nodes contained in our model, rather than resulting from higher level processes that we assume to be responsible for activating the plan nodes. The activation in the plan node spreads through the nodes of the network, finally reaching the content nodes (step 3). The time courses of the activation in the content nodes of the model are extracted, and the resulting time courses are linearly interpolated every 0.1 ticks (step 4), yielding time courses k_{t_s} , over a range of times t , for each syllable s .

Next, we need to establish the times where we suppose that the activation in the content nodes is sufficient to result in production of the syllable. We do this by comparing the interpolated activation time course k_{t_s} against a separate threshold θ_s for each syllable s . The threshold for each syllable is calculated

as the sum of a constant which is the same for all syllables, and a weighted exponential moving average of previous activations in the relevant content node. This means that the threshold gradually increases in response to activation in the content node, mimicking short term adaptation to the activation.

To calculate the threshold, we need to calculate the moving average activation. We calculate the average over a Gaussian kernel. Firstly, a weighting factor α is calculated, to cause the moving average activation to operate over a span of 9 ticks (90 observations with one observation every 0.1 ticks). The moving average activation m_{t_s} at a given time t for a given syllable s is then calculated recursively from the activation time series k_{t_s} :

$$m_{t_s} = \begin{cases} k_t, & t = 1; \\ \alpha k_{t_s} + (1 - \alpha) m_{t-1}, & t > 1. \end{cases} \quad (5.1)$$

$$\alpha = \frac{2}{90 + 1} = 0.022$$

Then, the threshold θ_{t_s} is calculated as the sum of the offset u , which is a model parameter, ‘threshold_constant’, and the moving average activation m_{t_s} , multiplied by a weighting ($c = 0.1$, for all conditions):

$$\theta_t = u + c m_t \quad (5.2)$$

The moment when the activation in the first syllable content node exceeds its threshold is taken as the onset word production, and the time when the activation falls below the threshold again is taken as the offset of the first syllable. The moment that the activation in the second syllable content node exceeds its threshold is taken as the onset of the second syllable, and the time when the activation falls below its threshold is taken as the offset of word production (step 5). In some instances, the model may predict multiple periods or activation for a syllable, or no activation at all. In cases where there is not precisely one period of activation above the threshold for each of the two syllables, no onset or offset times are recorded for that repetition. This suggests that the set of parameters is not very robust, and is excessively sensitive to the subtle changes introduced by the noisification, and should be dispreferred by the optimisation algorithm.

From the syllable-level onset and offset times, the three objectives can be calculated for each repetition: syllable 1 duration, between-syllable overlap, and syllable 2 duration. The durations from each of the 50 repetitions (n_{reps}) are

collected and a predicted distribution is constructed (step 6). To score the quality of the fit achieved by the values of the parameters, the observed fingerprint distributions p are compared to the predicted distributions q , for each fingerprint duration objective obj (step 7). The predicted and observed distributions are first binned (bin width 8 ms, from -200 ms to 1000 ms relative to simulation onset, 150 bins, n_{bins}), and a constant floor value ϵ of 1×10^{-13} is added to the count in each bin. The count in each bin is then divided by the sum of the counts in all the bins:

$$\begin{aligned} p_{obj_b} &= \frac{\text{count predicted}_{obj_b} + \epsilon}{\sum_{b=1}^{n_{bins}} \text{count predicted}_{obj_b} + \epsilon} \\ q_{obj_b} &= \frac{\text{count observed}_{obj_b} + \epsilon}{\sum_{b=1}^{n_{bins}} \text{count observed}_{obj_b} + \epsilon} \end{aligned} \quad (5.3)$$

Then, the Kullback Leibler divergence is calculated:

$$KL(p_{obj}, q_{obj}) = \sum_{i=1}^{n_{bins}} p_{obj_i} \times \log_2 \left(\frac{p_{obj_i}}{q_{obj_i}} \right) \quad (5.4)$$

where p is the observed distribution and q is the predicted distribution. $KL(p_{obj}, q_{obj})$ is taken as the score for the objective obj .

In cases where not all of the 50 simulation repetitions resulted in a duration (because the onsets and offsets of the syllables stray outside the period of the binning, because the activation time series never crosses the threshold, or because the activation times series crosses the threshold multiple times), the score was penalised by multiplying the KL by 50 (the number of repetitions, n_{reps}) divided by the number of values present. This penalisation is intended to favour solutions that are more stable; i.e. all 50 repetitions predicted exactly one period of activation for each syllable:

$$\text{missing values}(p, q) = n_{reps} - \sum_{b=1}^{n_{bins}} \text{count}(p, q)_b \quad (5.5)$$

$$\text{score}_{obj} = \begin{cases} KL(p_{obj}, q_{obj}) \times \frac{n_{reps}}{n_{reps} - \text{missing values}(p, q)}, & \text{missing values} < n_{reps}; \\ KL(p_{obj}, q_{obj}) \times n_{reps} \times 1.2, & \text{otherwise.} \end{cases} \quad (5.6)$$

Learning procedure

To test the models, two phases of optimisation were conducted for each model variant for each rate condition. During the first 100 generations of the optimisation procedure, some of the parameters are clamped; that is, the algorithm does not adjust them. This phase can be thought of as a rough initial search of a dimensionally-reduced subset of the parameter space. After this phase, the clamping of these parameters is released, and all the parameters are fine tuned to optimise the model's output. A full listing of the parameters, indicating which are clamped during the first 100 generations, is available in the online supplementary materials. The optimisation procedure is run for another 900 generations. During the first 1000 generations, the σ associated with each parameter is linearly related to the parameter μ value ($\sigma = 0.08 \times \mu$), following the observation of a linear relationship between the centre and the spread of the distribution in, for instance, response times (Luce, 1986; Wagenmakers & Brown, 2007).

After the 1000th generation, clamping is applied to most of the parameter μ values, such that they no longer undergo changes during the evolution and mutation phases of the NSGAI algorithm (see the table in the online supplementary materials for full details), whilst the parameter σ values are released, and therefore learnt independently. Starting with the 1500th generation, this is reversed, and the σ values are clamped and μ values are learnt. Starting with the 2000th generation, σ values are again released from clamping, and μ values are clamped. From the 2500th generation, no clamping is applied. The learning procedure is stopped after the 5000th generation (This arrangement is depicted graphically in the shading in Figures 5.6 and 5.7).

This multi-phase approach is an attempt to speed up the overall search for a well performing parameter set, by allowing quick rejection of unpromising areas of the parameter space during the first 100 generations, and successively finer-grained searching in the subsequent phases. The long total run, of 5000 generations, ensured that the optimisation process was sufficiently converged to make valid model comparisons.

5.5.2 Explicit test of the advantage of non-linearity

It is also possible to vary the fitting procedure to more directly assess the hypothesis of the presence of gaits manifested as qualitatively different regimes in

the parameter space. This ‘linear constraint’ model is functionally identical to the asynchronous model variant, but is optimised in a different manner, to force the parameter values found during the optimisation routine to be linearly related. Instead of conducting an independent optimisation run for each rate condition, the parameters of the linear constraint model associated with all three speaking rates are optimised together via a meta-model. This meta-model has parameters for the slope of a line for each of the model parameters, as well as an intercept parameter for each speaking rate. From these slopes and intercepts, parameter values for each speaking rate are derived, and passed to instantiations of the asynchronous model variant for each speaking rate. The *KL* scores for each metric are gathered from the submodels, and together form the nine objectives (syllable 1 duration, syllable 2 duration and overlap duration for each of the three speaking rates) of the multi-objective optimisation routine. For clarity and conciseness, the results obtained from this additional model variant are reported along with those of the other model variants in Section 5.5.3, where the model variant is referred to as the ‘asynchronous model variant with linearity constraint’.

5.5.3 Results: model performance

Conventionally, statistical comparison of models for the purpose of model selection takes into account the number of parameters (degrees of freedom) that each model has; assigning models a ‘handicap’ per extra degree of freedom to identify the model that strikes the best balance between quality of fit and parsimony (Akaike, 1974). In a framework where a model predicts variance, it is fairly clear how one would go about doing this. In our case, however, the models predict the three fingerprint distributions, which we evaluate on the basis of the Kullback Leibler divergence between the model and the observed fingerprint distributions, rather than predicting values for each observation, from which likelihood-based metrics might be calculated. This makes it difficult to select a plausible handicap with which to penalise the model performance without adding further simulations.

A typical approach to assess the performance of different variants of a model is to directly compare their ability to fit the data after learning, by seeing how well the target function is satisfied by each trained variant. In our case, this is not possible because of the multi-objective nature of the problem. Recall that the

model optimisation process results in a Kullback-Leibler score for each of the target distributions for each solution and that these scores are not mathematically comparable across the three objectives without unduly favouring one objective above another. We therefore needed to take a different approach to ascertain how well the different model variants learned, and how well they ultimately performed after training, that would not arithmetically collapse the Kullback-Leibler scores. To assess learning over time, we adopt a metric in terms of Pareto dominance. To assess final performance, we adopt a regression approach.

Learning trajectories

To characterise the learning trajectory of each run, we identified the Pareto front in each generation cumulatively. This means that, for each generation, we looked for solutions in that generation and all generations before it that were Pareto optimal. We used loess-fitting (Cleveland & Devlin, 1988) to identify the trend in the score for each objective function in each rate condition. These loess-fits are shown in Figure 5.6, where we can observe, very broadly speaking, that for all three model variants, the most progress is made in finding solutions that improve the fit in the overlap duration objective. Much more restrained progress is made on improving the fit of the syllable duration objectives. The asynchronous model appears to perform moderately better than the synchronous variant on the syllable duration objectives, while the variant with the linearity constraint never achieves scores as good as the other two variants, with the notable exception of the syllable 1 duration objective, which performs comparably to or slightly better than the other other model variants.

Convergence

If the model is learning, the quality of the Pareto front will improve with each generation. Conventionally, convergence in the optimisation multiobjective problems is assessed with the hypervolume indicator (Zitzler et al., 2007), which calculates the volume of the dominated space between a reference point and the Pareto front. The hypervolume indicator for our optimisation runs, normalised to have a value between 0.0 and 1.0, is presented in the upper panels of Figure 5.7. The value of the normalised indicator increases as the volume of the dominated

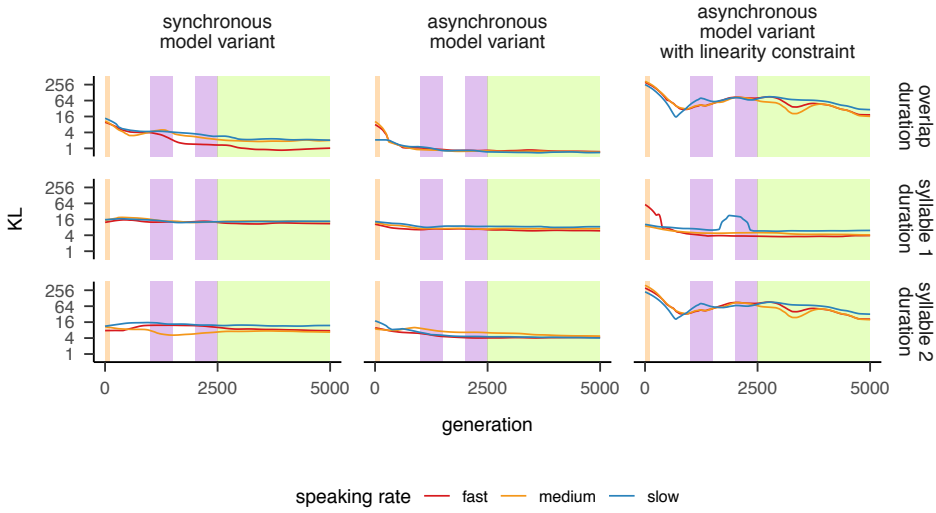


Figure 5.6: Loess-fits of the Kullback Leibler scores (y-axis, log-transformed scale, lower values indicate better performance) of the solutions in the Pareto front in each generation (x-axis), for the three rate conditions (line colours), the three objective functions (rows) and three model variants (columns). The shading indicates the optimisation phases of the model, orange is the phase where only the μ component of a subset of the parameters was adjusted by the optimiser, white indicates that the μ component of all parameters was adjusted by the optimiser, purple indicates that the σ component of all parameters was adjusted by the optimiser, and green indicates that both μ and σ components of all parameters were adjusted.

space increases. Convergence is evidenced by stabilisation of the indicator at a value close to 1.0.

Although simple to interpret and widely applied, the hypervolume indicator has the disadvantage of arithmetically combining the values of the objective functions into a single fit quality metric. This is undesirable for our KL objective functions (see page 88). We therefore calculated a second indicator of model convergence, which assesses the change in the composition of the Pareto front after each generation.

When the model finds a new solution that is nondominated, this solution joins the Pareto front. Sometimes, this solution falls between two others, improving the coverage of the Pareto front, but not improving the fitness of the Pareto front in general. Other times, the solution dominates a solution or several solutions that were in the Pareto front in the previous generation. These dominated solutions are ‘relegated’ from the Pareto front. Since we are primarily interested in finding optimal parameters to fit the observed data, and only secondarily interested in increasing the size of the Pareto front, we want a metric that is sensitive to the second type of new solution. Therefore, rather than counting new solutions, we count the number of solutions that are relegated from the Pareto front (c.f. Martí et al., 2009). When the optimiser has converged, no relegation events will be observed. The lower panels of Figure 5.7 show loess fits of the proportion of former Pareto front members that are relegated in each generation.

Both the hypervolume indicator and the relegation count metrics indicate stability after around 3000 generations, leading us to conclude that the optimisers are sufficiently converged by the end of the 5000 generations tested.

Statistically testing model variant performance

In order to evaluate the performance of the different model variants, we need to identify and statistically test differences in the KL scores achieved by the Pareto front solutions of each of the model variants. Simultaneously, we need to disregard variation in the KL scores as a function of objective, since KL scores for the various objectives are not directly arithmetically comparable because of differences in the observed distributions, as previously discussed. The same holds for comparing models fitting different rate conditions, between which there are also differences in the variability of the observed distributions.

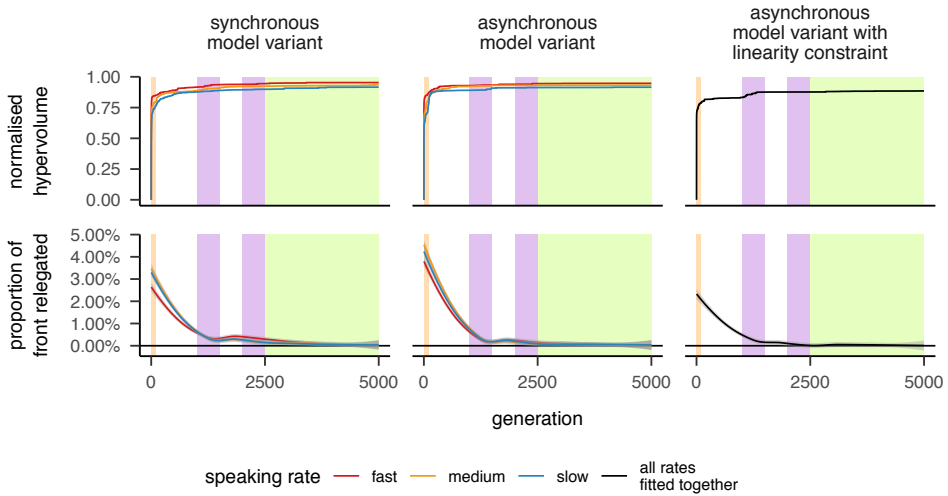


Figure 5.7: Upper panels: The normalised hypervolume indicator (y-axis) during the 5000 generations of the optimisation run (x-axis), for the three model variants (columns). Stabilisation of the normalised hypervolume indicator at a value close to 1.0 indicates successful convergence. For the synchronous and asynchronous model variants, coloured lines indicate the speech rate condition being optimised. Lower panels: the proportion of former front members relegated from the front in each generation. See the caption of Figure 5.6 for the meaning of the shading.

Instead of averaging scores across objectives, linear regression with categorical predictors for model variant, rate condition and objective can be used to isolate the effect on the KL score attributable to model variant, independent of rate condition and objective. This leads to a regression model with the following structure (Wilkinson-Rogers notation, 1973):

$$KL \sim \text{model variant} * \text{rate condition} * \text{objective} \quad (5.7)$$

This is a model predicting KL with categorical predictors for *model variant*, *rate condition* and *objective*, and all interactions between the levels of those categorical predictors.

The KL scores were bootstrap re-sampled to introduce variation required to perform regression modelling. The bootstrapped distributions of the KL scores are shown in the first three panels of Figure 5.8. We took 2,000 samples with replacement of sets of syllable 1 duration, syllable 2 duration, and overlap duration values from the observed dataset. For each of these samples, we calculated the KL s between the re-sampled observed distributions and the model's predicted distributions. The resulting bootstrapped KL s were then log transformed and z-normalised. The log transformation was necessary to de-skew the KL s, which obey a log distribution.

The regression model fitted the data quite well, achieving an adjusted R^2 value of 0.76. The fits of the regression model for the main effect of model variant are shown in the fourth panel of Figure 5.8, as black dots. The full table of model coefficients is provided in the online supplementary materials.

Relative to the asynchronous model variant, the synchronous model variant performed significantly worse ($\beta = 0.55$, $SE = 0.0083$, $t = 66***$, $d = 0.52$).

As discussed earlier in this section, it is not possible to draw meaningful conclusions from the significance of the main effects of rate condition or objective; these were included to enable us to use the regression model to avoid arithmetically comparing KL scores calculated with different observed distributions and therefore different scales.

Are predicted fingerprint durations plausible?

It is also informative to assess the performance of the model variants qualitatively, by directly examining their success in emulating the target distributions.

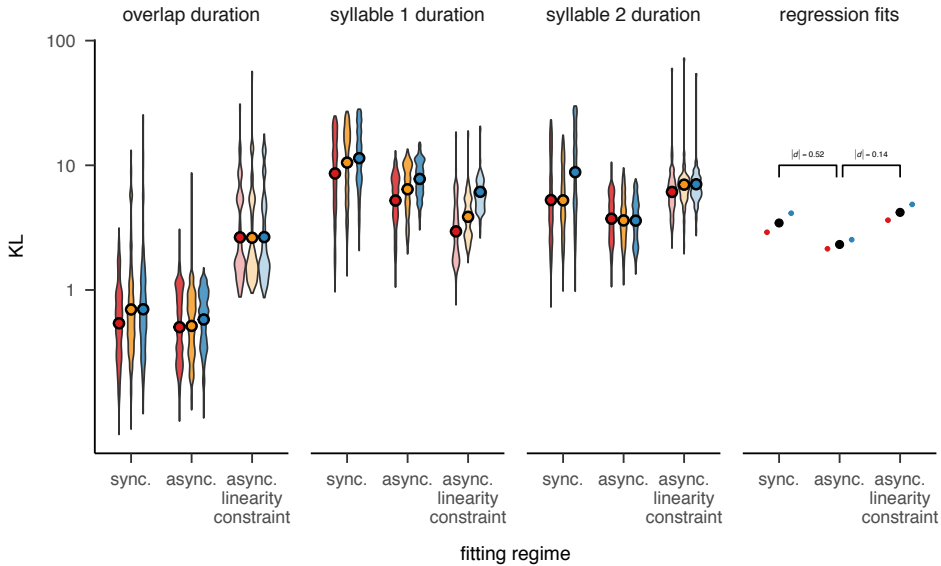


Figure 5.8: First three panels: the bootstrapped distributions (violins) of the KL scores (y-axis, smaller is better, log scale) achieved by the 0-ranked agents (the Pareto front) for each model variant (x-axis, *sync.*: synchronous model variant, *async.*: asynchronous model variant and *async. linearity constraint*: asynchronous model variant with linearity constraint, see Section 5.5.2 for full details) in each speaking rate condition (fill colours), in each objective (panels). The coloured dots indicate the model fits for the three-way interaction term in the regression model. Fourth panel: the fits of the model variant term from the regression model (main effect shown as black dots, fits of rate condition: model variant interaction in smaller coloured dots). 95% confidence intervals are omitted because they are too small to be visible. Significant differences in the main effect are indicated. The main effect of model variant is plain to see; the asynchronous model variant performs significantly better (achieves lower KL scores) than the synchronous model variant. The asynchronous model variant without the linearity constraint outperforms the asynchronous model variant with the linearity constraint.

In Figure 5.9, we show the distributions resulting from combining the duration distributions predicted by each member of the Pareto front of each run as solid violins. These are compared against the target distributions measured from the corpus (translucent violins with dashed edges).

For all three model variants, relatively good fits are achieved to the syllable 2 duration distribution, with the asynchronous model variant arguably mimicking the precise shape of the distribution somewhat better than the synchronous model variant and the asynchronous model variant with linearity constraint. In fitting the syllable 1 duration distribution, the synchronous model variant produces a bimodal distribution, rather than the unimodal distribution in the observed data, and also fails to fit the central tendency well. The asynchronous variant performs better, although the distributions it predicts are slightly too narrow. The asynchronous variant with linearity constraint predicts syllable 1 duration distribution very well. In fitting the overlap duration distribution, the asynchronous model variant performs best, fitting the central tendency well but overestimating the spread of the distribution somewhat. The asynchronous model variant with the linearity constraint predicts a slightly wider unimodal distribution. The synchronous model variant again predicts a bimodal distribution where one mode matches the density peak of the observed distribution.

It should be noted that the vast majority of simulation papers in this domain report only central tendencies. The distributional fits that we achieve seem acceptably good in (qualitative) comparison with the few psychological modelling studies that we found that did fit distributions (Wiecki & Frank, 2013, Figure 4; Engbert et al., 2005, Figure 10).

5.5.4 Summary of strand 1

In strand 1 of this study, we introduced EPONA, a new model inspired by the DBS model, that was successful in predicting the temporal structure of disyllabic word production. EPONA provides the first computationally explicit connectionist account of speakers' ability to modulate the speech production system to achieve different speaking rate.

The methods that we used to train and evaluate the variants of the model were also novel. We adopted a framework whereby the model predicts distributions of three objectives, which were measured from the PiNCeR corpus of elicited speech (Rodd, Bosker, ten Bosch, et al., 2019a): the duration of the first syllable, the du-

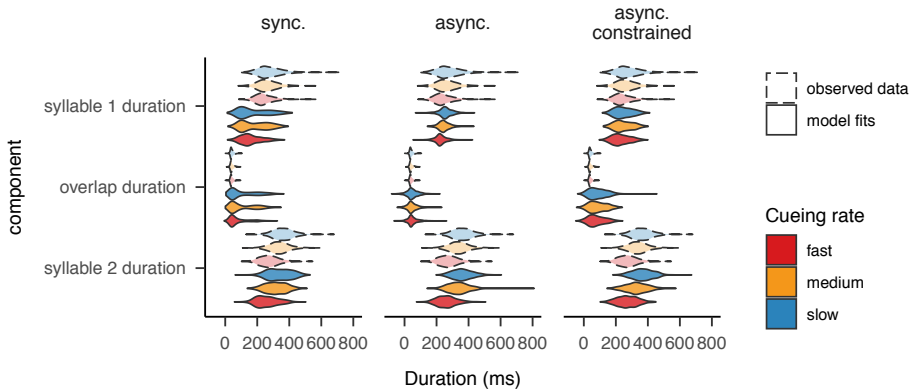


Figure 5.9: The duration (x-axis) distributions (filled violins) predicted by three models variants (facets) at the three rate conditions (colours) for each of the three target distributions (y-axis), compared against the observed distributions (translucent violins with dashed edges).

ration of the inter-syllable overlap, and the duration of the second syllable. We assumed that the central tendency and the variability of these distributions together reflect the characteristics of the underlying cognitive system. This means that during the training process, the models learned to resemble the underlying cognitive system.

Training proceeded using an evolutionary algorithm that optimised the parameter values so as to minimise the Kullback-Leibler divergence scores associated with each objective distribution. The success of the evolutionary algorithm in learning parameter values that fitted the objective distributions for each model variant is an index of how well suited that model variant is as a model of the formulation phase of speech production.

Alongside the asynchronous and synchronous model variants, we introduced a third model variant that was fitted using a different optimization regime. This allowed us to directly test the prediction of a single gait system, where all three speaking rates are linearly related in parameter space. This model is discussed further in strand 2.

The asynchronous model variant without the linearity constraint performed best on the quantitative criteria we set and offered the most plausible predicted fingerprint durations. We therefore perform further analyses for strand 2 only on the asynchronous model variant.

5.6 How do regimes relate to each other? (strand 2)

To explore how executive control might be exerted on the EPONA model to achieve different speech rates, and thereby assess whether different rates are achieved by shifting between multiple qualitatively different ‘gaits’ of speech production, we need to compare the best parameter values identified by the optimiser for each speaking rate condition. We can think of the solutions as positions in a multidimensional space where each parameter of the model is mapped to one dimension. In such a space, the Euclidean distance between a pair of locations in parameter space represents the difference between solutions.

Note that we have assumed that only one regime exists for each speaking rate, while of course several distinct configurations may have emerged to account for the temporal structure of speech at a given rate. We tested for this possibility by performing k-means clustering on the parameter values associated with each speaking rate. The clustering did not support multiple regimes in any of the rates; see the online supplementary materials for full details.

5.6.1 How are regimes arranged relative to each other?

Method

Having identified the best solutions for each rate, we consider how the regimes adopted for each rate relate to the regimes adopted for the other rates. To do this, we perform principal component analysis (PCA), which involves projecting the 12 parameters on which the speaking rate regimes vary onto principal components (PCs). The procedure loads as much variance as possible onto each component in turn, whilst ensuring that each component is orthogonal to the preceding PCs. A full listing of the parameters is provided in the online supplementary materials. PC1 (the first PC) accounted for 30.0% of the variance, PC2 accounted for 11.6% of the variance, PC3 for 8.6% and PC4 for 5.3%. The loadings of the parameters onto the PCs are listed in the online supplementary materials.

Results

Figure 5.10 shows the spread of solutions across the rate conditions in the first and second principal components. Note that since this is a projection of multiple dimensions into two, much variation is not visible, and points that appear adja-

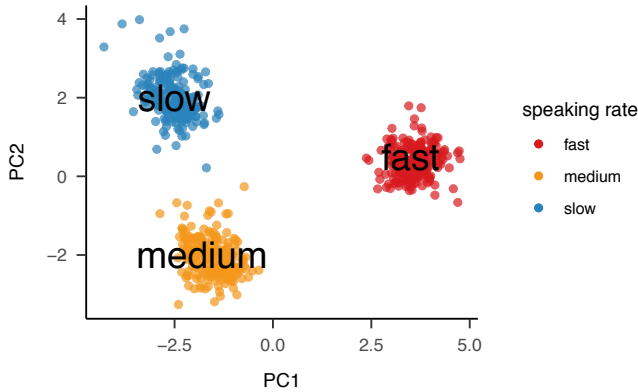


Figure 5.10: the Pareto optimal solutions identified for the fast (red), medium (green), and slow (blue) rate conditions, plotted for PC1 (x-axis) and PC2 (y-axis).

cent on the PC1-PC2 plane depicted may be quite distant on other dimensions. For this reason, it is not certain that medium and slow are closer together than medium and fast, or slow and fast, although it appears so on the PC1-PC2 plane. The optima associated with the three rates (fast in red, medium in green, and slow in blue tones) occupy broadly different areas of the PC1-PC2 plane. On this plane, the clusters of solutions of the three conditions are well separated, and the spread of the solutions in the three conditions is broadly comparable.

The spatial organisation of the rate conditions on the PC1-PC2 plane is clearly not axial in nature, ruling out the single gait account. This is in line with the observation in strand 1 that an asynchronous model variant constrained to only consider linear arrangements of the rates in parameter space performed worse than the asynchronous variant without this constraint. Instead, the gaits are arranged as a triangle, supporting a multiple gait interpretation. Decelerating from the medium speaking rate to the slow speaking rate involves increasing PC2 while slightly decreasing PC1. Accelerating from the medium speaking rate to the fast rate involves increasing both PC1 and PC2.

5.6.2 Which regimes belong to which gaits?

Extrapolating fingerprint durations between rate centres

The previous finding suggests that there is more than a single gait adopted by speakers to control their speaking rate. The parameter optimisation analysis

cannot, however, allow us to assess which, if any, of the three regimes belong to the same ‘gait’. To assess that, we conducted a further exploratory analysis.

We calculated the mean position of each speaking rate regime in parameter space. These means form the ‘reference’ points. Between each pair of reference points, we interpolated 5 equally spaced points along a straight line (axis) through parameter space. Additionally, we extrapolated two extra points on each of these axes beyond the reference points. We therefore have axes from fast to slow, from fast to medium, and from medium to slow, that intersect at the reference points. The arrangement is illustrated the upper panel of Figure 5.11.

We then took the parameter values associated with the location of each point, and constructed and ran new instances of the asynchronous model with these parameter settings, to predict the distributions of the three ‘fingerprint’ durations. Just as in the optimisation procedure, the parameters were noisified, and 50 runs were conducted (see Figure 5.5 and accompanying text for more details). These durations, along with the word duration are indicated in the raincloud plots in Panel C of Figure 5.11, and normalised in Panels D and E.

In Section 5.1.2, we identified five possible mappings of the speaking rate regimes onto one to three gaits (single gait, three gaits, slow is distinct while fast and medium are mapped to the same gait, fast is distinct, medium is distinct). These possible mappings are depicted diagrammatically in Panel B of Figure 5.11. We directly modeled and compared the plausibility of these five hypothetical mappings. If a pair of speaking rate regimes belong to the same gait, we would expect the fingerprint distributions of the interpolated points between them to follow a linear trend, and that all the interpolated points would result in plausible fingerprint distributions. If, however, the regimes belong to different gaits, we would expect to see a non-linearity at some point along the axis, indicating a shift from areas of parameter space associated with one gait to areas of parameter space associated with the other gait, possibly with an area of ‘unproductive’ parameter space in between where non-plausible fingerprint distributions are predicted.

We tested the presence of linearity in the axes through statistical modelling of the simulated durations depicted in the lowers panel of Figure 5.11. We conducted both Bayesian (MCMC sampling) and non-Bayesian analyses using linear regression models and generalised additive models (GAMs, Wood, 2017). Both types of model were multivariate, in that they fitted the simulated durations

of the three axes simultaneously in a single model. The results of the two approaches were comparable. For brevity, only the Bayesian analysis is reported here. The GAM analysis is reported in the online supplementary materials.

Bayesian linear switchpoint regression

For each axis of the extrapolated fingerprint duration data, we regressed the normalised fingerprint durations by the number of the step along the axis. The Bayesian models allow us to identify the locations in parameter space of the switchpoints along the axes, and additionally exploited variation in the distribution along the length of the axes.

Axes could be modelled with either a ‘uniform’ linear fit, or a ‘switching’ fit that permitted non-linearity. The uniform fit predicted normalised duration (both μ and σ) by the step number, with distinct slope and intercept parameters for each component \times axis combination for μ and σ . The switching fit split the axis into two halves at a fixed switchpoint, and fitted a separate regression with separate parameters for each half. For each axis, different fixed switchpoints were tested, namely between steps 4 and 5; between steps 5 and 6; between steps 6 and 7; or between steps 7 and 8. This means that different numbers of models were required for each mapping, ranging from 1 model for the ‘no gaits’ mapping, to 64 models for the three ‘distinct gaits’ mapping (4^3). A Student t distribution was used as the likelihood. This has heavier tails than a normal distribution, meaning that it is a form of robust regression and can better accomodate heteroskedasticity. For all slope and intercept parameters, mild $N(0, 1)$ priors were applied, which makes the assumption that most effects are smaller than Cohen’s $d = 1$ and nearly all effects are smaller than Cohen’s $d = 2$. The fit resulting from the model fitting the ‘fast is special’ mapping is depicted in Panel E of Figure 5.11, by way of example.

For each model, 8 chains of 8000 samples (of which 4000 warm up) were sampled by NUTS in RStan (Stan Development Team, 2018, version 2.18.2). No convergence issues, assessed by the Gelman-Rubin diagnostic \hat{R} , effective number of samples and visual inspection of traceplots, were noted for any of the models. Full details of the Bayesian linear switchpoint analysis are available in the online supplementary materials.

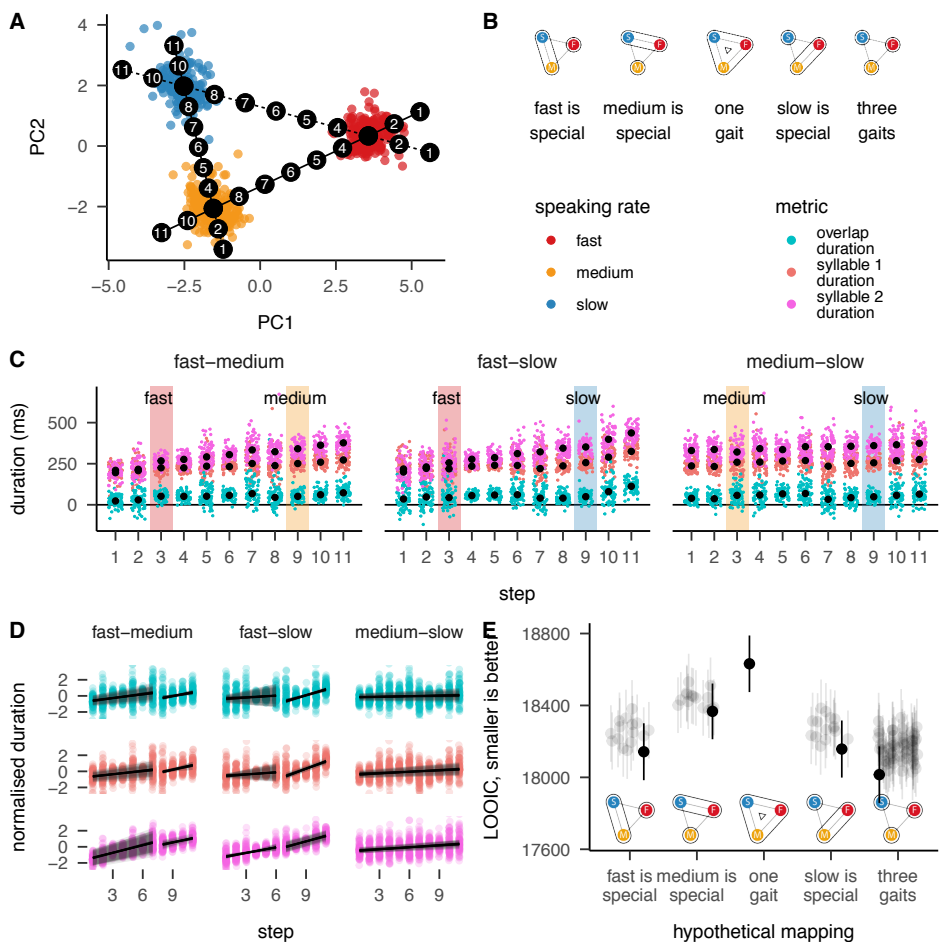


Figure 5.11: Panel A: The extrapolated axes projected onto the PC1 (x-axis) – PC2 (y-axis) plane through parameter space. Each black point indicates a location at which the model was run and fingerprint distributions were calculated. Behind, the optimal solutions identified by the optimisation procedure are shown (See Figure 5.10 for details). Panel B: the hypothetical mappings of rates to gaits, represented diagrammatically, enclosing lines indicate rates achieved by the same gait. Panel C: the distributions of the durations (y-axis), of the overlap, syllable 1 and syllable 2 (colours), shown as rainclouds at each step (x-axis) of the three axes (panels). Black points indicate the median values. Shading indicates the reference points where the axes intersect. Panel D: example fit of the Bayesian linear switchpoint models for the ‘fast is special’ mapping for the three axes (columns) and fingerprint component distributions (rows, colours). Panel E: point estimates and standard errors of the quality of fit of the Bayesian linear switchpoint regression models for each mapping, quantified by an information criterion calculated by leave-one-out cross validation. For each mapping, the models other than the best performing are plotted more lightly.

Results

Panel E of Figure 5.11 presents the model comparison results of both the Bayesian linear switchpoint models. We compare models on information criteria, which aim to quantify the explanatory power of the models in terms of the amount of information lost, while at the same time penalising model complexity to avoid over fitting. Specifically, we calculate an information criterion by leave-one-out cross validation (the LOOIC, Vehtari et al., 2017).

The ‘one gait’ mapping performs notably worse than the other models, achieving higher LOOIC values. That this model performs worst is a useful sanity check, since the earlier findings of worse performance in the linearly constrained model variant, and the triangular arrangement of the rates in parameter space for the unconstrained model variant should have ruled this possibility out. Next comes the ‘medium is special’ mapping. This mapping predicted a distinct gait for medium speech, and a fall-back gait engaged to produce other speaking rates. Such a configuration might emerge as a consequence of speakers producing speech almost always around a specific habitual rate, which would become more practiced. The remaining mappings perform the best. The LOOIC estimation for the Bayesian linear switchpoint models additionally allows us to quantify the uncertainty about the point estimates of model fit. In panel E of Figure 5.11, lines extending from the points indicate the standard error around the LOOIC estimate. For the three best performing mappings, the standard error ranges around the point estimates are extensively overlapping, meaning that we cannot with confidence claim support for any of the three mappings ahead of the other two.

5.6.3 Summary of strand 2

In strand 2 of this study, we explored how cognitive control might be exerted on the parameters of EPONA to model speech produced at different rates. Different settings of the model parameters can be conflated with different regimes of the cognitive system underlying natural speech production. We examined how the regimes related to each other, hypothesising that there might be ‘gaits’ in the speech production system that speakers switch between to achieve different speaking rates. Five hypothetical mappings of rate regimes onto gaits were considered.

We found evidence that different speaking rates were achieved by distinct parameter values, and that these were arranged in a triangle in parameter space, rather than along a straight line. The triangular arrangement rules out a mapping whereby a single gait is quantitatively modulated to achieve different speaking rates. With the aim of distinguishing between the remaining mappings, we conducted further statistical modelling. This modelling ruled out one further account, namely the medium-is-special mapping, but did not allow us to distinguish between the three remaining mappings. It therefore remains an open question whether slow and medium speech is achieved by one gait and fast by another (the ‘fast is special’ account), whether slow speech is achieved by one gait and fast and medium by another (the ‘slow is special’ account), or whether all three rates are achieved by qualitatively distinct gaits (the ‘three gait’ account). Nevertheless, the findings of strand 2 provide strong evidence for a model of speech production control whereby speakers shift between different gaits to achieve different speaking rates.

5.7 General discussion

This study had two aims. In strand 1, we sought to establish EPONA, a new model inspired by the DBS model that would predict the duration of syllables and the duration of the overlap between them, and thereby characterise the configuration of the speech production system at different speaking rates. Subordinate to this aim, we sought to explore how the temporal properties of a word could best be encoded in the frame node.

In strand 2, we explored how cognitive control might be exerted on the parameters of EPONA to model speech produced at different rates. Different settings of the model parameters can be seen as corresponding to different regimes of the cognitive system underlying natural speech production. We sought to examine how the regimes relate to each other, hypothesising that there might be ‘gaits’ in the speech production system that speakers switch between to achieve different speaking rates.

5.7.1 Computational model (strand 1)

The evolutionary algorithm learned distinct parameter settings for each speaking rate for the three model variants, though the quality of the predictions made

by the trained models varied. Linear regression analyses revealed significant differences in performance between the model variants, and effect size analysis allowed us to quantify the extent of the performance differences, demonstrating a distinct performance advantage for the asynchronous model variant ahead of the control and synchronous model variants.

A salient difference between the model variants is that the control and synchronous models exhibit bimodal distributions in their fitting of the overlap duration and syllable 1 duration (see Figure 5.9). In contrast, the asynchronous variant predicts uni-modal distributions for these objectives. It is noteworthy that the modelled syllable 1 duration and overlap duration distributions resemble each other in their overall shape. In examining the duration distributions independently for a sample of the front members (a figure showing these is included as the online supplementary materials), it was plain that the bi-modality of the combined distribution arises because some solutions predict distributions that contribute to the first ‘bump’ of the bimodal distribution, and others predict distributions that contribute to the second. This result suggests that the control and synchronous model variants were not successful in finding a parameter set that solved both the serial order problem and fitted the distributions of the objectives adequately.

The observed distributions for overlap duration for all three speaking rate conditions exhibited notably less spread than the observed distributions for the two syllable duration targets. None of the model variants were particularly good at predicting the spread of the overlap, instead showing excessive spread.

Although the fits achieved by the model are satisfactory for the purposes of our strand 2 investigation, some aspects of EPONA could potentially be revised to broaden its utility. Firstly, the model at present is only capable of producing disyllabic words. Extending the model to produce a variety of word lengths would be relatively trivial, and would potentially allow us to explore questions regarding the extent of the gestural score, i.e., are whole syllables encoded, or instead smaller segmental or demi-syllabic level chunks; or larger chunks at the level of phonological words or entire intonational phrases? Secondly, the current implementation of EPONA produces one word at a time, and cannot capture the interactions between previous and upcoming words, and between target words and competitors in the lexicon, although there is no reason why this could not be

implemented as a network of interconnected EPONA ‘columns’. How that might work is discussed further in Section 7.2.2.

Modelling considerations

The EPONA model follows many speech production models of the 20th century by implementing a strict separation between the formulation and execution phases (e.g. Dell & O’Seaghdha, 1992; Levelt, 1989; Levelt et al., 1999; Stemmerger, 1985). The execution phase of the model is also in its conception *ballistic*, meaning that once activation arrives at the formulation-execution frontier and speech articulation begins, the gestural score will be played out without regard to what happens in the formulation phase after the onset of production.

Recent work has demonstrated that formulation and execution processes are not entirely discrete. Lexical competitors have been found to influence the details of articulation of target words (e.g. Goldrick & Blumstein, 2006; McMillan & Corley, 2010), whilst the articulation of slip errors has been found to differ from canonical productions of the same form (e.g., “pig” erroneously produced as [big] differs from canonical “big” in voice onset time; Goldrick et al., 2016). Relatedly, contextual predictability and frequency predict the extent to which words are reduced by shortening the word duration and eliding segments (e.g. Pluymaekers et al., 2005; Bell et al., 2009). That errors and contextual priorities that arise during formulation propagate into the domain of execution has been taken as evidence in favour of cascading activation, that is, partially active ‘competitor’ units from the formulation phase activate the corresponding articulatory plans.

A fully ballistic, cascading system would require no control on the execution phase over and above the control exerted on the formulation phase. This is of course attractive, but implausible; at the very least, a mechanism is required to allow the interruption of erroneous productions (Levelt, 1983). Alternatively, it is possible that the dynamics of the planning system after the onset of articulation also influence ongoing articulation. Fink et al. (2018) set out to test the assumption of a ballistic execution component, measuring response latency and word duration in sequential picture naming tasks designed to introduce semantic interference. If the production system is ballistic, effects of semantic interference on response latency (an index of planning) and word duration (an index of articulation) should be positively correlated since variation in both metrics

arises from the same process. A ballistic process cannot, however, account for effects of semantic context on duration over and above the effects correlated with the effects on latency. Fink et al. (2018) found consistent coupling of articulation and planning, compatible with the ballistic account, but also some evidence of interaction effects, suggesting that ongoing planning can exert moderate influence on execution after the onset of articulation.

Although EPONA as presented here does not explicitly model for cascading activation and has no mechanism to predict the articulatory outcome of simultaneous activation of multiple articulatory plans, it contains no features that are incompatible with the cascade concept. Similarly, the model could be considered non-ballistic, in that sustained activation of the syllable gestural score is required to cause articulation of the required syllable. A more elaborate model of the execution phase might predict the articulatory outcomes of simultaneous activation of competitors (for instance, in the VOT of stops, as investigated by Goldrick et al., 2016), and of changes in the activation dynamics of the output nodes of the formulation network after word onset.

We followed Dell et al. (1997) in favouring a simple and interpretable model that explains the underlying psychological processes of speech production at a functional level, rather than striving for any semblance of neurobiological plausibility. The predefined activation patterns that the frame node produces on each of the ports are crucial to ensuring the correct ordering of syllable units is achieved, and have a large influence on the timing of syllable production. In general, the requirements to (1) prime upcoming units, (2) activate them at the correct time and (3) deactivate them once they have been produced is referred to as the serial order problem. Dell, Burger and Svec's (1997) approach to resolving the serial order problem using predefined activation patterns is functional and minimal.

It is, however, also possible to achieve correct serial ordering using only components from the standard connectionist toolbox. In this respect, a promising approach is competitive queueing (Grossberg, 1978; Houghton, 1990), which employs a two-layer sub-network to maintain serial order. The first layer is a planning layer, where all nodes for all the elements in a sequence become active in parallel, with their relative activation encoding the order of realisation (a primacy gradient). The nodes of the planning layer project onto the same number of nodes in the second, competitive choice layer, where inhibitory connections

ensure that only the activation of the most active node at any given time is transmitted to the output nodes, and a switch-off mechanism ensures that successfully produced items are inhibited, allowing subsequent items to be produced (see Hurlstone et al., 2014, for an extensive review). It would be fruitful to evaluate a model that employed competitive queueing in the frame node. This would remove the need for the implausible stepped activation patterns in the frame node.

The activation function (that is, the function that computes the activation of a node from the activation arriving at it through connections, also known as a transfer function) in the model is strictly linear. Sigmoid activation functions such as tanh (Harm & Seidenberg, 1999) or soft-max (Chang, 2002; Chang et al., 2006) are employed in several more recent models where competition between nodes at the same level is modelled. It is, however, unlikely that a different choice of activation function would have made a large difference to the outcomes of this study, since our model does not simulate between-node competition. In a model with competitive queueing, a non-linear activation function might prove advantageous.

5.7.2 How do regimes relate to each other? (strand 2)

Since the asynchronous model variant performed significantly better than either the control or synchronous model variants, we performed analyses in parameter space only for this variant. The following discussion refers therefore to the asynchronous model variant only.

The speaking rate regimes identified in this investigation can be compared along two dimensions; firstly, in terms of the parameter values that the model engages to achieve each targeted speaking rate (comparison in parameter space), and secondly in terms of the predicted fingerprint durations (comparison in prediction space).

Which, if any, gaits are present?

To distinguish between single-gait and multiple-gait scenarios, we examined the arrangement of the regimes in parameter space. We predicted that in the single-gait scenario, the three regimes would be arranged sequentially along an axis in parameter space. In a multiple gait scenario, the three regimes would be ar-

ranged in a triangle in parameter space. The arrangement of the optima on the PC1-PC2 plane was clearly non-axial (see Figure 5.10). Our results therefore indicate that cognitive regimes adopted to achieve different speaking rates are arranged in a manner that is incompatible with a single-gait system.

It could however, still be the case that, although the optimisation routine had settled on a non-linear arrangement of rates, a linear arrangement could have been able to fit the data adequately. A further asynchronous model variant was fitted to test this, where the arrangement of the rates in parameter space was constrained to be linear or axial (see Section 5.5.2). This model fitted the data less well than the unconstrained model, reinforcing our conclusion that multiple gaits are present.

Having established that the single gait configuration was unlikely given the data, we moved on to comparing the regimes in prediction space. Aside from all rates being produced by one gait, there are four further possible mappings of rates onto gaits: three gaits; slow is distinct while fast and medium are mapped to the same gait; fast is distinct; medium is distinct.

The plausibility of these mappings could be teased apart by examining the extent of non-linearity in the predicted distributions of models fitted with parameter values taken from the spaces *between* the centres identified in the evolutionary optimisation. We performed statistical fitting to test for (non-)linearity along the axes linking the centre points of each rate, and compared the quality of fit of models instantiating the five possible mappings. We used Bayesian linear switchpoint models, which are able to fit variation in the spread of the distribution, and allowed us to quantify certainty at all stages of modelling, including model comparison.

This statistical modelling allowed us to directly test the plausibility of the five mappings. The one-gait mapping was rejected, consistent with the triangular arrangement of the rates in parameter space and the rejection of the model variant with the linearity constraint in the optimisation paradigm. Support for the ‘medium is special’ mapping was limited. Although the ‘three gaits’ mapping had numerically the best fit, the statistical modelling was unable to distinguish between this mapping and the ‘fast is special’ and ‘slow is special’ accounts. This means that all three mappings are plausible models of the cognitive reality, given the present dataset and modelling approach. While we believe that the statistical modelling is sufficiently sensitive to evaluate the plausibility of the mappings, it

is of course dependent on the data provided by the simulations. These data may be insufficient in two ways. Firstly, they consist only of predicted distributions of the three fingerprint durations, which may not be rich enough a representation of the acoustic reality to highlight subtle differences in linearity between the speaking rates. Secondly, the variability that was valuable in the parameter optimisation paradigm for the reconstruction of the distributions to be compared with the observed distributions may have proved counterproductive for the statistical modelling we conducted.

Further experimental work is required to clarify the nature of the mapping of speaking rates to gaits, possibly testing more than three speaking rates in a denser sampling.

The consequences of the presence of gaits for models of speech production and perception

Our concept of different ‘gaits’, each encompassing qualitatively similar regimes in the formulation component of the speech production system, represents a theoretical step forward that makes predictions that may be fruitfully explored in future modelling and empirical work, building on the conception of gaitedness at the execution level.

Although this study concerned speaking rate variation and demonstrated the presence of cognitive gaits to achieve different speaking rates, it is plausible to think of switching between qualitatively different parameter regimes as a more general mechanism to deliberately modulate the acoustic and temporal properties of speech to suit various communicative situations (Lindblom, 1990; Lindblom et al., 1991).

Natural speech produced by any one speaker varies in many more ways than along a single dimension of speaking rate, in effect adopting what has often been called different *registers* or *speaking styles* (Hirschberg, 2000). It has been observed that speakers transform the acoustics of their speech to enhance its intelligibility for their interlocutor, or in response to the reverberance or background noise of their environment (Cooke et al., 2014). Prepared speech, such as reading aloud, varies from spontaneous speech (e.g. Furui, 2003). Typically, these speaking styles have been thought of (or at least treated as) categorically distinct, driven perhaps by the methodologies used to elicit the speech during

experiments and corpus gathering, or to categorise the situations in which the speech arose in generalist corpora (Hirschberg, 2000).

Although acoustic differences emerge between speech categorised according to these situational categories, knowing that such differences exist says little about how speakers modulate the speech formulation and execution mechanisms to achieve that variation. This is because it remains unknown to what extent the speech planning system engages categorically distinct regimes to achieve different speaking styles, and whether these researcher-imposed situational labels bear any resemblance to the underlying cognitive categories.

If different speaking styles are achieved by switching between qualitatively different gaits of the speech planning system, we would expect there to be observable clustering in the acoustic characteristics of speech across the range of speech variability, reflecting the categorical shifts between cognitive gaits. Two recent findings suggest that speaking style variation may be at least to some extent categorical. The first concerns reduced pronunciation variants, that is, pronunciations of words where acoustic cues, segments, and sometimes entire syllables are omitted, generally when words are highly predictable and in informal spontaneous speaking situations (e.g. Ernestus & Warner, 2011; Ernestus et al., 2015), for example the realisation of American English “yesterday”, the canonical form of which is /jɛstəreɪ/, as [jɛʃeɪ]. Reduction of this type is one of the ways in which acoustic differences between speaking styles surface and can be quantified. Hanique et al. (2013) found evidence that both categorical and gradient processes were simultaneously responsible for an instance of schwa deletion in Dutch.

The second concerns the retrieval of speaking style labels through machine-learning techniques. Bentum et al. (2019) employed a language modelling and dimensionality reduction approach to characterise word choice and co-occurrence across the speaking styles in the orthographic transcriptions of a corpus of Dutch speech containing many different speaking styles (Oostdijk, 2000). Many of the speaking styles labelled in the corpus emerged as distinct clusters, whilst other groups of speaking styles merged to form a single cluster. Again, this hints that, underlyingly, speaking styles differ categorically from each other on various dimensions.

The finding of gaitedness in speech production has consequences for models of speech perception. If the speech produced by speakers varies qualitatively be-

tween gaits, then listeners might also be expected to adopt different processing strategies to make the most of the cues available in the speech signal associated with a specific gait. If that were the case, we might expect to see gaits in speech perception to mirror those in speech production.

5.8 Conclusion

We proposed that to achieve different speaking rates, the speech planning system adopts different configurations, or regimes. Since speakers are able to voluntarily adjust their speaking rate, they must have a control mechanism that enables them to shift from regime to regime. Describing the way in which these regimes are arranged relative to each other in parameter space is highly informative for understanding the nature of the control mechanism that is engaged to shift between regimes, and how control might be exerted on speech production in general. We hypothesised that speech rate control might be achieved by shifting between different, qualitatively distinct ‘gaits’ of the speech production mechanism. Alternatively, different speaking rates might be achieved by continuous adjustment within a single rate.

We set ourselves the task of distinguishing these hypotheses. We developed EPONA, a model inspired by the influential DBS model (Dell et al., 1997), to predict the distributions of syllable and syllable-overlap durations that characterise speech production in a specific speaking rate regime. By optimising the parameters of this model to fit each of three rate conditions independently, we identified optimal parameter settings for each speaking rate, which we conflate with the dimensions of the regime-space of the underlying cognitive system. By examining the arrangement of the parameter optima of the model, we could infer the arrangement of the underlying cognitive system. The model optima resembled a triangle (Figure 5.10), rejecting the idea that the regimes of the speech production system all belong to a single qualitatively consistent gait. By fitting further models where linearity in parameter space was enforced, we provided further evidence ruling out a single-gait account.

6 Asymmetric switch costs between speaking rates: Experimental evidence for ‘gaits’ of speech planning

A recent model of speech production, EPONA, proposes that speakers adopt different configurations of the cognitive speech formulation system to achieve different speaking rates. It characterises these configurations as analogous to the qualitatively distinct gaits (walking, running) adopted in locomotion. Critically, it is assumed that a relatively small set of gaits are required to cover the range of possible speaking rates, and that switching gait is more effortful than modulating speaking rate within a gait. This study tested whether we could find empirical evidence for gaits in speech production by assessing whether switching between one set of rates is more difficult than switching between other sets of rates. In a multiple picture naming task, speakers were required to begin speaking at one of three pre-trained rates. During the trial, the required speaking rate changed. We quantified (1) speakers’ success in achieving the rates and (2) how quick they were to switch from the initial rate to the new rate. A Bayesian analysis showed that speakers were slower to shift between fast and medium speaking rates than they are to switch between slow and medium speaking rates. This is consistent with the presence of a ‘run’ gait for the fast speaking rate, and a ‘walk’ gait for slow and medium speech. We discuss the implications of the finding in the context of the EPONA model.

This chapter was adapted from Rodd, J., Bosker, H. R., Ernestus, M., ten Bosch, L., & Meyer, A. S. (under review). Asymmetric switch costs between speaking rates: Evidence for gaits of speech planning.

Code and supplementary materials are available at <https://osf.io/ruqze/>

6.1 Introduction

Speech is hugely variable, in both the spectral and temporal domains. A salient feature in the temporal domain is variation in speaking rate. Different speakers have different habitual speaking rates, and individual speakers can vary their speaking rate from situation to situation, and even within utterances in the same conversation (e.g. Miller et al., 1984; Quené, 2008). The fact that humans have control over the rate at which they speak means that they are capable of adjusting the cognitive apparatus that plans speech, from the selection of words to the tightly coordinated movements of the articulators of the vocal tract. But how do speakers control their speaking rate? This article sets out to explore how control is exerted on the formulation phase of speech production, testing the hypothesis that speakers switch between ‘gaits’ for different speaking rates, analogous to walking and running gaits in locomotion.

The process of speech production can be divided into two architecturally distinct phases: *formulation* including the processes of lemma retrieval, morphological encoding and phonological encoding stages; and *execution* involving phonetic encoding and motor control. These phases are qualitatively distinct in the operations conducted: the formulation phase involves competition between representations, and incorporates threshold mechanisms to gate the flow of activation. The execution phase, by contrast, involves no competitive selection, instead mapping the abstract representations of the formulation domain onto the more concrete plans of the motor domain (Levelt, 1989). In particular, the execution phase can be thought of as a ‘faithful servant’ of the formulation phase. This is compatible with various models of speech (motor) planning (Dell & O’Searghdha, 1992; Levelt et al., 1999; Stemmerger, 1985; Tourville & Guenther, 2011; Parrell et al., 2019). Although the two phases are architecturally very different, they must closely cooperate, not least to remain synchronised and thus make an extensive buffer unnecessary. This cooperation may be thought to occur by direct spreading of activation from the output level representations of the formulation phase into the input level representations of the execution phase, consistent with the finding that competing representations in the formulation phase exert influence on articulation (e.g. Goldrick & Blumstein, 2006). The tight coupling between formulation and execution also means that both phases must

cooperate to achieve the stylistic outcome that best supports the communicative goal (Lindblom, 1990; Pouplier, 2012).

Speech rate control might be thought to be achieved purely through linear up- or down-regulation of the cognitive speech planning apparatus. Alternatively, the speech planning system might be reconfigured into qualitatively distinct configurations to suit different ranges of speaking rates. The gaits of the locomotion system have been proven an appealing analogy to apply to characterise this reconfiguration account (c.f. Pouplier, 2012; Rodd et al., 2020), since both locomotion and speech are motor behaviours operating at a continuously varying range of speeds. While gaits have proven a fruitful analogy for control of the execution phase of speech planning (e.g. Pouplier, 2012), the analogy can be taken further and applied to the control of the formulation phase as well. In the formulation domain, speaking gaits reflect configurations of the cognitive apparatus supporting higher level processes of speech production, like retrieval of word forms, their phonological encoding and the timely activation of lower-level planning units.

The EPONA model (Rodd et al., 2020, Chapter 5) is a connectionist model that proposes a mechanism that can explain how temporal aspects of variation emerge in the formulation phase of speech production. Gaits may arise in the control mechanisms of EPONA, but the model does not pre-suppose them. EPONA is inspired by the DBS model (Dell et al., 1997). At its core, it assumes a layer of frame nodes that capture temporal properties of words, encoding when syllable-level gestural scores should be produced. In EPONA's computational implementation, activation spreading and thereby the timing of word production is controlled by 13 parameters.

6.2 This study

In the simulation with the EPONA model conducted by Rodd et al. (2020, Chapter 5), the model's 13 parameters were optimised to fit the temporal properties of speech elicited in a multiple picture naming task at fast, medium and slow speaking rates (Rodd, Bosker, ten Bosch, et al., 2019a, Chapter 4), where the speaking rate was indicated by a red cueing dot that indicated when each picture was to be named. Each speaking rate was optimised independently, to minimise the divergence between the simulated and observed distribution of three measurable

features: the duration of the first syllable, the duration of the second syllable and the duration of the between-syllable overlap. The parameters that fitted each rate best differed qualitatively, and a model variant that was constrained to only consider linear regulation of the speaking rate fit comparatively poorly. This indicated that, to achieve the best fit to the observed data, the model would switch between different configurations, or gaits, to achieve different speaking rates.

Whilst Rodd et al. (2020) were able to discount linear modulation in favour of a gaited account, the results of the simulation were not sufficient to distinguish between all five logically possible mappings of the three tested speaking rates onto gaits. Three mappings were equally plausible, given the simulation data:

1. **three-gait mapping:** The cognitive system uses three gaits to achieve the sampled speaking rates, one for each speaking rate. These three gaits are qualitatively different, like walking, trotting, and galloping in horse locomotion.
2. **fast-is-special mapping:** The cognitive system uses two gaits, grouping the medium speaking rate with the slow rate.
3. **slow-is-special mapping:** The cognitive system uses two gaits, grouping the medium speaking rate with the fast rate.

The present study has two aims. Firstly, we aim to provide additional evidence independent of the simulation study that distinguishes between a gaited and a wholly linear control of formulation. Secondly, we aim to build on the conclusions of the simulation study to evaluate the plausibility of the three remaining mappings of speaking rates onto gaits.

In the current experiment, speakers named pictures from a visual display with pictures arranged around a ‘clock face’, at one of three speaking rates (fast, medium, slow) that they had previously been trained to achieve reliably. Unlike the experiment used to elicit the data for the previous simulation study (Rodd, Bosker, ten Bosch, et al., 2019a, Chapter 4), no cueing dot was used for the test phase, because an external rate stimulus might interfere with the effect we hoped to measure. Instead, speakers had to learn to maintain the speaking rate themselves. The rate that they were required to speak at was indicated by a coloured frame around the clock face display. During each trial, the rate they were required to speak at changed, indicated by a change in the colour of the

frame. We term the rate before the switch the ‘initial rate’ and the rate after the switch the ‘subsequent rate’. Switches occurred along the fast↔medium axis (from fast to medium or from medium to fast), and along the slow↔medium axis (from slow to medium or from medium to slow). A two-step Bayesian statistical analysis was used to quantify how quickly speakers were able to adjust their speaking rate. The model predicted this by the axis along which they had to switch (fast↔medium vs. slow↔medium), and whether the switch involved acceleration or deceleration. The model also accounted for variability attributable to the measured difference between the realised initial and subsequent rates.

If, to switch from a initial speaking rate to an subsequent speaking rate, speakers have to reconfigure the production system by switching gaits, we assume that they will be slower to do so, relative to switching from a initial rate to an subsequent rate that can both be achieved with the same gait. We base this assumption on findings from the switch costs literature that switching between disparate tasks is more difficult than switching between similar ones (Arrington et al., 2003; Taube-Schiff & Segalowitz, 2005). The first hypothesis that we test relates to the first aim of the study: we hypothesise that there will be an effect of axis. This finding (regardless of the direction of the effect) would be consistent with the concept of gaits, and contrary to wholly linear modulation of speaking rate.

The second aim of the study is addressed by making more specific predictions about the direction of the effect of axis. The direction of the effect can distinguish between the three remaining mappings. The fast-is-special mapping predicts quicker switching on the slow↔medium axis than on the fast↔medium axis, because the latter crosses a gait boundary, while the former does not. The slow-is-special mapping predicts the opposite: that it should be quicker to switch along the fast↔medium axis than along the slow↔medium axis. The three gaits mapping predicts no effect of axis: since both axes involve crossing a gait boundary, it should be equally difficult (or easy) to switch between either pair of speaking rates. A non-gaited, purely linear model of speaking rate control also predicts no effect of axis, since if there are no gaits, it should be equally difficult (or easy) to switch between either pair of speaking rates, but note that this possibility was unlikely given the results of Rodd et al. (2020).

6.3 Methods

6.3.1 Participants

Healthy native Dutch speakers with no hearing or language impairments and uncorrected normal vision were recruited from the Radboud University community to take part in the experiment ($N = 18$, $M_{age} = 23.4$). All participated with informed consent. The study was approved by the Ethics Committee of the Social Sciences faculty of Radboud University (project code: ECSW2014-1003-196). Participants came for a single 1.5 hour session, and were paid for their time.

6.3.2 Materials

For comparability with Rodd et al. (2020, Chapter 5), we used a subset of the same materials Rodd, Bosker, ten Bosch, et al. (2019a, Chapter 4). Twelve disyllabic Dutch concrete nouns with stress on the first syllable were selected, for instance *snavel* ['sna:.vəl] “beak”, *vriezer* ['vri:.zər] “freezer” and *wafel* ['wa:.fəl] “waffle”, and line drawings were prepared for each word. A full list is provided in the supplementary materials.

6.3.3 Experimental procedure

Participants were tested individually in a sound attenuated booth. An eye-tracker (Eyelink 1000 in desktop configuration with forehead stabiliser; SR Research, Ottawa, ON, Canada) recorded right eye gaze position.

Stimulus presentation, eye-tracker synchronisation and audio recording were controlled by Presentation software (Version 16.5; Neurobehavioral Systems, Berkeley, CA, USA). A Sennheiser ME64 directional microphone was used to record the participants' speech at a sampling rate of 48 kHz.

Sampled speaking rates

The three speaking rates used by Rodd, Bosker, ten Bosch, et al. (2019a) (fast, medium, slow) were also used in the present study. The medium and fast rates were selected by identifying comfortable and maximal rates in a non-cued pretest. The slow rate was selected to be equally distant from the medium rate as the fast rate, in log ms. The rates are summarised in Table 6.1.

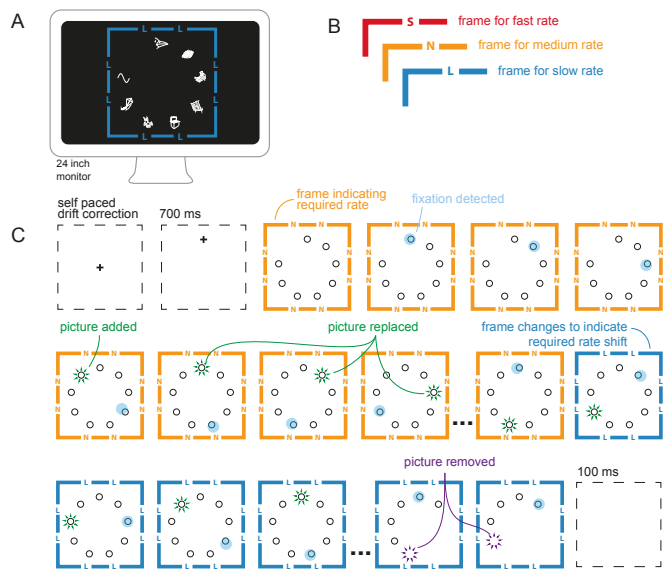


Figure 6.1: Panel A: a general depiction of the experimental display. Panel B: colour mapping for frame indicating cueing rate, letters on the frame also indicate the cueing rate (S = *snel* ‘fast’, red; N = *normaal* ‘medium’, yellow; L = *langzaam* ‘slow’, blue). Panel C: a diagrammatic depiction of the sequence of a trial in the test phase of the experiment, showing how pictures are added or replaced (green ‘starburst’ pattern) in response to fixations (blue dots).

Table 6.1: The target speaking rates used in the experiment

	word duration (ms)	word rate (Hz)	syllable rate (Hz)	syllable rate (log Hz)
fast	456	2.19	4.39	1.48
medium	646	1.55	3.10	1.13
slow	915	1.09	2.19	0.78

Picture familiarisation

The session began with familiarisation of the pictures and their names, by means of (1) a printed card which the participant was invited to study whilst the experimenter set up the experiment, and (2) naming of the pictures as they were displayed individually on screen, in a pseudo-randomised order with two repetitions of each picture. The experimenter immediately gave the correct name when the participant named a picture incorrectly.

Clock face display

In all training and test phases of the experiment, the same visual display was used to instruct the participant which pictures to name, and to indicate the required speaking rate. The display was derived from that used in Rodd, Bosker, ten Bosch, et al. (2019a). The pictures were displayed, white on black, in a clock-face arrangement with 9 positions (Figure 6.1A). Initially, position 9 at 11 o'clock was left empty, to visually reinforce the beginning of the sequence of pictures at 12 o'clock. The clock-face fitted into an area of 780 x 780 pixels. Each picture was scaled such that it would occupy an area of 90 x 90 pixels. The clock-face was surrounded by a coloured square frame (940 x 940 pixels, 20 pixels wide). The colour of the frame indicated the rate that was required from the participant (Figure 6.1B), along with letters that intersected the frame. During the trial, the pictures opposite the picture presently being fixated were replaced so that the participant could make multiple circuits of the display (Figure 6.1C). In this way, the trial could in principle continue indefinitely. On the last circuit of the trial, the pictures were removed rather than being updated. In all experimental phases, lists were constructed such that words occurred equally frequently and that the frequency of all word-to-word transitions was approximately balanced.

Training: cued rate learning

In the first training phase, the participants had to practice the speaking rates that they would use in the experiment proper, and become familiar with the paradigm. They started with learning the medium speaking rate. Each trial started with a mock drift correction, where a fixation cross was displayed in the centre of the display. They pressed the space bar to begin the trial. A fixation cross was displayed at the location of the first picture for 700 ms, then the pic-

tures appeared. For 316 ms, the pictures were displayed as ‘preview time’, based on the extra latency before naming the first picture by participants in a non-cued pretest. After the preview time, a red cueing dot appeared over the first picture. Every 456, 646 or 915 ms (see Table 6.1), the dot jumped clockwise to the next picture. As each picture was cued, the ‘opposite’ picture across the clock face was updated if necessary, in preparation for the upcoming circuit. In the cued training phase, each trial contained 12 pictures. After five trials practising the medium rate, participants practised the slow rate for five trials, followed by the fast rate for five trials. Following the cued training, the participant placed their forehead in the stabiliser and the eye-tracker was calibrated with a 13 point sequence.

Training: non-cued rate learning

In the second training phase the participants learnt to produce the required speaking rates accurately, without the support of the cueing dot. Again, they began with the medium rate. The eye tracker recorded the location of the fixations. The precise moment that the participant fixated each picture was recorded. When a new fixation was detected, the ‘opposite’ picture was updated. In this phase, each trial contained 30 pictures. After the trial, the median temporal interval between fixations was calculated. Since speakers in multiple picture naming tend to have tight eye-speech synchrony (Meyer et al., 2012), we could use this interval as a rough measure of speaking rate (word duration) that was immediately available. The median rate was displayed on a graphical interface on a second monitor that was visible only to the experimenter, along with a target word duration range (from 826 to 1014 ms around the target of 915 ms for the slow rate, from 586 to 712 ms around the target of 646 ms for the medium rate, and from 416 to 500 ms around the target of 456 ms for the fast rate). The experimenter compared the displayed median against the target, and informed the participant whether the rate in the previous trial was good, too fast, or too slow.

The training started with two non-cued trials. If the median rate of the second non-cued trial was outside the target range, a cued trial was conducted to remind the participant of the required speaking rate. After the second non-cued trial or the cued reminder trial, three further non-cued trials were conducted. If these were all within the required rate range, the training proceeded to the next

rate. Otherwise, another cued trial and another three non-cued trials were conducted. Therefore, participants performed minimally five and maximally eight non-cued training trials at each speaking rate, and possibly a further one or two cued reminder trials. On average, they performed 7.5, 7.1 and 6.8 non-cued trials for fast, medium and slow speaking rates, respectively. The number of training trials performed was not a criterion for excluding participants.

Test phase and task familiarisation

In the test phase, the required rate was changed part way through the trial (Figure 6.1C). This was communicated to the participant by changing the colour of the frame surrounding the picture display and the letters embedded into it. The switch occurred at the moment of fixation of a predetermined picture for each trial. The switch picture could be the 12th to the 28th, randomly sampled from a uniform distribution; each trial contained 40 pictures. Switches always occurred between the extreme rates (slow and fast) and the medium rate, meaning that the possible switches were slow→medium, medium→slow, fast→medium and medium→fast. Participants were informed that all trials would involve a switch either from or to the medium rate, and completed a familiarisation block with one of each switch condition. In the test phase, blocks were constructed consisting of two trials of each switch condition, with the order of trials within the block shuffled. Between blocks, a reminder cued trial was conducted for each speaking rate. Since the duration of the experimental session was fixed and the number of trials in training phase was variable for each participant, and in some cases, multiple eye-tracker calibration attempts were required, the number of test blocks that each participant completed before the end of the experimental session varied (minimally 5, maximally 10, median 7).

6.3.4 Annotation

The recordings for the test trials were annotated to the word level by hand by a panel of phonetically trained research assistants and the first author. This was done using the distributed browser-based annotation system developed by Rodd et al. (in press, Chapter 3). First, the trial recordings were broken into chunks that were typically a few words long, by splitting at automatically detected silences (chunking). Second, these chunks were orthographically transcribed by

hand in a browser-based interface, which suggested words from the experimental vocabulary as the research assistant typed (orthographic transcription). Third, the chunks were forced-aligned to the word-level using the MAUS system (Schiel, 2015). Fourth, the annotators screened the forced-aligned words for the accuracy of the automatic segmentation (triage). Fifth, the annotators adjusted the segmentation of words that were marked as incorrectly segmented in the triage step (retrimming). Information about speaker and forced-alignment errors is available in the supplementary materials.

6.3.5 Post-processing

To calculate an index of the local speaking rate during the trial, we measured the durations of the onset-onset intervals and offset-offset intervals between pairs of words. Pairs that were interrupted by a hesitation or filled pause were excluded. The onset-onset intervals and offset-offset intervals were transformed to syllable rates in Hz.

6.4 Results

6.4.1 Statistical modelling

All the statistical models reported were fitted with the package `brms` (Bürkner, 2018) in R (version 3.5.2, R Core Team, 2018), allowing us to fit Bayesian mixed-effects and non-linear (stepwise) regression models. Unless otherwise reported, models were sampled with the NUTS sampler with 6 chains of 7,000 warm-up and 7,000 test iterations, with thinning retaining every third iteration. All models converged, as assessed by the Gelman-Rubin diagnostic \hat{R} being within 0.00001 of 1.0. For the models fitted here, we take advantage of possibilities not readily available in frequentist statistical modelling: fitting arbitrary non-linear models for the single-trial models, and using the quantified certainty of the estimates of the trial level models in the meta-analytic model. Rather than dealing in binary decisions between significant and not significant, Bayesian regression focuses on quantifying uncertainty about the magnitude of an effect (e.g. Vasissth et al., 2018), so no p -values are reported. Instead, we report the size of the effects we identify, in their relevant units, and where appropriate, standardised for com-

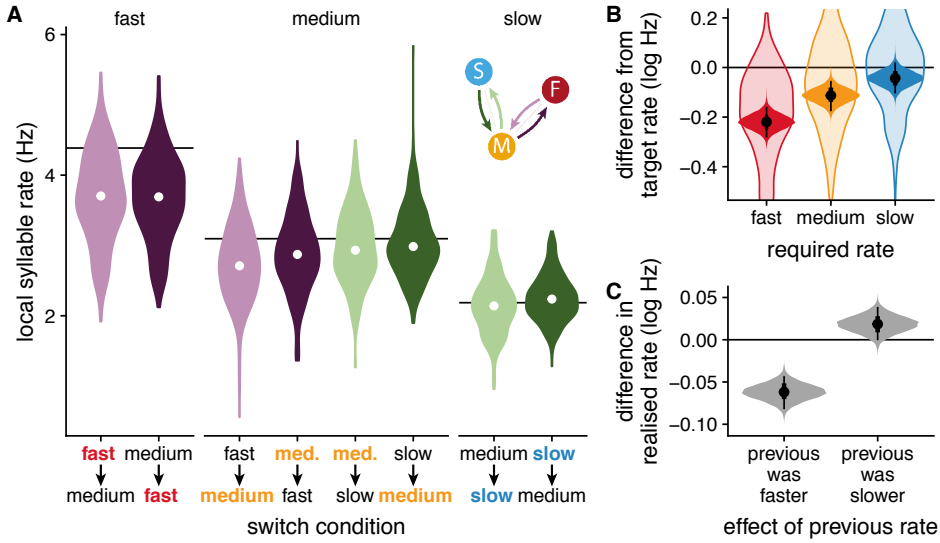


Figure 6.2: Panel A: Violins show the measured local syllable rates in the before- and after-switch analysis windows. The colours and x-axis indicate the different switch conditions. The target rates are shown in facets. There are therefore two violins for each condition, one for the rate in each analysis window. White dots indicate the median. Panel B: translucent violins show the observed differences from the target by the required rate (colours, x-axis). Solid violins show the model posteriors of the mean for each condition with median (points), 95% HDIs (thin black lines) and 66% HDIs (thick black lines). Negative values indicate measured rates slower than the target. The y-axis is cropped to highlight the model posteriors, see the supplementary materials for a larger, uncropped version. Panel C: The effect of the previous rate on the realised speaking rate in the subsequent rate. Again, with median, 95% and 66% HDIs. Negative values indicate measured rates slower than the random intercept for the relevant rate.

parability (Cohen's d). All intervals reported are 95% highest density intervals (HDIs).

6.4.2 Rate compliance

To assess how well speakers complied with the required speaking rates, we calculated the average rate in two windows during the trial. The before-switch window contained the six words immediately before the switch. The after-switch window contained the first six complete words beginning 7000 ms following the switch, at which time we assumed that the speaker would be stable in the new speaking rate. The measured local syllable rates are shown as violins in

Figure 6.2A. Each switch condition appears twice, the separate violins show the rates in the before-switch window and after-switch window, respectively.

In general, speakers tended to speak more slowly than required. To characterise this undershooting, we constructed a Bayesian mixed-effects regression model, predicting the difference from the required rate (by subtracting the required speaking rate from the measured speaking rate; negative values indicate rates slower than the target) by the speaking rate condition. Full details about this model are in the supplementary materials. The fits of this model are shown in Figure 6.2B. The model revealed that, in all cases, speakers were slower than required. For fast: $-0.22 \log \text{ Hz}$ $[-0.27, -0.17]$; for medium: $-0.11 \log \text{ Hz}$ $[-0.16, -0.065]$; for slow: $-0.043 \log \text{ Hz}$ $[-0.093, 0.0067]$.

A second striking pattern in the data is that speakers seem to ‘hyper-correct’ when they have to switch rates. That is, a rate is realised slower if the speaker previously had to speak at a relatively faster rate, and a rate is realised faster if the speaker previously had to speak at a relatively slower rate. To quantify this in a statistical model, we created a new categorical variable, ‘previous rate’. All the measurements from before the switch were coded as ‘baseline’. All the measurements from after the switch were coded according to the relative rate before them: if the speaker had to accelerate in that trial, we coded ‘previous was slower’, if the speaker had to decelerate, we coded ‘previous was faster’. A Bayesian mixed-effects regression model predicted the realised speaking rate by the relative difference of the previous rate with random intercepts for the target speaking rate. Full details about this model are available in the supplementary materials. The model revealed a small, asymmetrical contrastive effect: if the initial rate was slower, the subsequent rate was realised faster than baseline, $0.062 \log \text{ Hz}$ $[0.046, 0.078]$, Cohen’s $d = 0.22$; if the initial rate was faster, the subsequent rate was realised marginally slower than baseline, $-0.019 \log \text{ Hz}$ $[-0.035, -0.0024]$, Cohen’s $d = 0.066$. The contrastive effect of the previous rate is depicted in Figure 6.2C.

6.4.3 Speed of speaking rate switch

To assess how quickly speakers were able to adjust to the new speaking rate after the switch, we modelled speakers speaking rate behaviour in a two-stage Bayesian regression procedure, consisting of a non-linear (stepwise) Bayesian

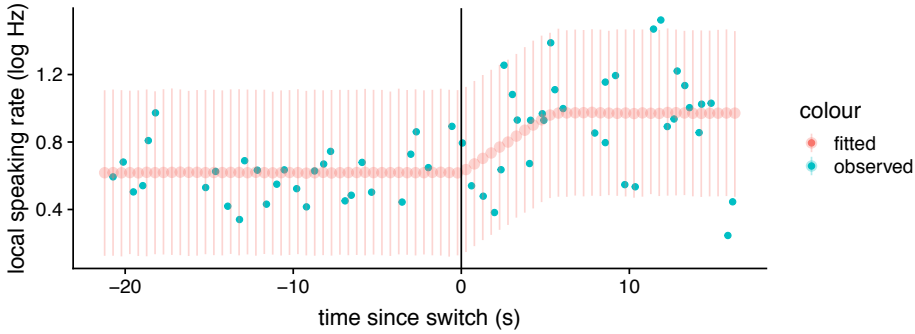


Figure 6.3: An example single trial model for the trial with a switch between slow and medium. Blue dots indicate measured rates, red dots with CIs indicate model fits. The plateau before 0 ms captures the initial rate. After the switch (0 ms) the slope of the model is allowed to vary. An additional parameter determines when the slope ends and the plateau for the subsequent rate begins. See the text for full details.

regression for each trial, followed by a meta-analytic Bayesian regression model to characterise patterns between trials.

Single-trial modelling

The single trial model captures the change in speaking rate during the trial. The dependent variable is the log-transformed speaking rate at the onset and offset of each word measured in Hz, the calculation of which is described in Section 6.3.5. The times in the model are shifted, such that the switch occurs at 0 ms.

The model fit for an example trial is depicted in Figure 6.3. Three parameters are used to fit the data, an *intercept*, a *slope*, and an *offset* time. The model fit is non-linear, and always adopts a broadly sigmoid shape to predict the relationship between time in the trial and the speaking rate, with a plateau before the switch, a sloped part after the switch, and a second plateau later on in the trial.

For all times before 0 ms (the switch), the model predicts the speaking rate with the *intercept* only. After 0 ms, the model predicts the speaking rate as $\text{intercept} + \text{slope} \times \text{time}$. A third parameter determines when the second plateau begins. At times after the value of that parameter, the model predicts the speaking rate as $\text{intercept} + \text{slope} \times \text{offset}$.

Weakly informative priors were set: for the intercept, a normal distribution centred at 1.13 Hz (the required medium rate) with a sigma of 1.13 Hz. For the

slope, a normal distribution centred at 0 with a σ of 0.0005, and for the offset parameter, a normal distribution centred at 5500 ms after switch onset, with a sigma of 500 ms, with a lower-bound set at 1000 ms after switch onset. For each model, 6 chains of 4000 warm-up and 4000 test iterations were run. From these models, the estimate for the *slope* and the error associated with that estimate were extracted.

Meta-analysis

To test the hypotheses of interest, we constructed a meta-analytic model that predicted, for each trial, the estimate for *slope* and the error associated with that estimate (err_{slope}), as determined by the single trial models. The model had a fixed effect predictor, *rate difference* for the difference between the measured speaking rate in the before-switch and after-switch windows in each trial, included to control for the large proportion of the variance in the slope magnitude accounted for variance in the difference between realised initial and subsequent rates (see Section 6.4.2 for more details). For deceleration trials, *slope* and *rate difference* was subtracted from 0, so that *slope* represents the magnitude of the slope, and *rate difference* represents the magnitude of the difference between initial and subsequent rates. Both *slope* and *rate difference* were centred around 0 and standardised for the model fitting, all reported values are back-transformed. Two two-level deviation-coded categorical predictors were included, describing the switch condition relevant to the trial; *axis*, which had levels of fast↔medium (1) and slow↔medium (-1), and *accelerate*, which had levels of accelerate (1) and decelerate (-1). The interaction between *axis* and *accelerate* was also included. Random intercepts were included for *speaker*. In the *brms* dialect of the Wilkinson-Rogers notation, the model formula was:

$$slope|se(err_{slope}) \sim 1 + rate\ difference + axis * accelerate + (1|speaker)$$

The same weakly informative priors were set for all predictors, a normal distribution centred at 0 with a σ of 1.08 hours. For the model intercept, the prior was a Student-*t* distribution centred at 0 with a σ of 10 and a ν of 3.

The distributions of the model coefficients of interest are shown in Figure 6.4. There was a small effect of *axis* (panel A), whereby the fast↔medium axis has

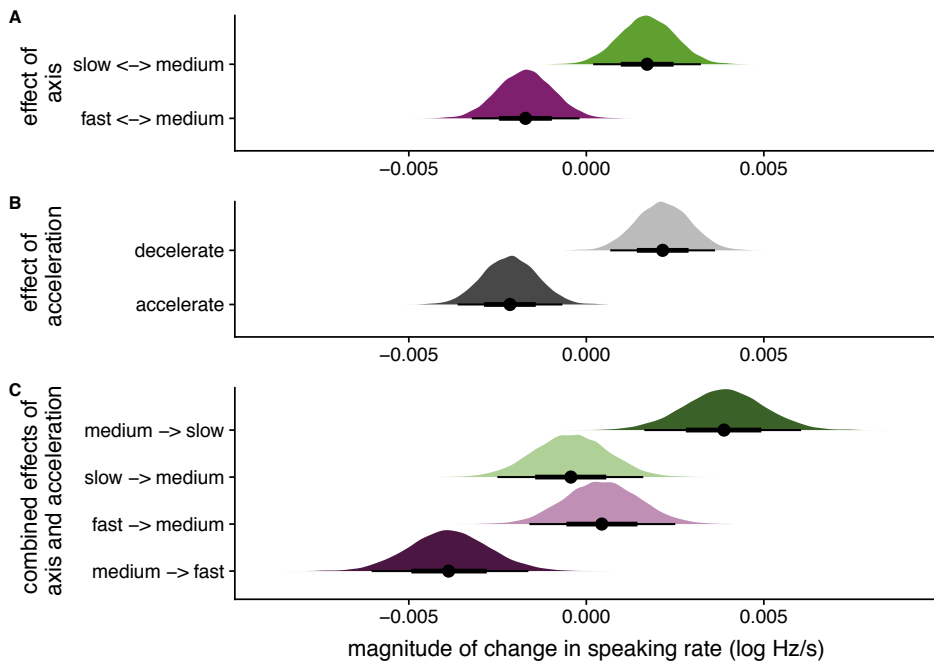


Figure 6.4: The distributions and medians (points), 95% HDIs (thin black lines) and 66% HDIs (thick black lines) of the model coefficients for: (panel A) the main effect of axis; (panel B) the main effect of accelerate and (panel C) the combined effects of axis and accelerate.

steeper slopes than the slow↔medium axis (difference between means: 0.0034 log Hz/s [0.00087, 0.0058], Cohen's $d = 0.26$). This means that speakers found it easier to shift along the slow↔medium axis than along the fast↔medium axis. This effect is equivalent in size to a difference between picture naming RT distributions of 105 ms. See the supplementary materials for full details of this comparison, which is based on data from Zormpa et al. (2019).

There was a larger effect of *acceleration* (panel B), where deceleration is associated with steeper slopes than acceleration (difference between means: 0.0043 log Hz/s [0.0019, 0.0067], Cohen's $d = 0.33$), meaning that, across axes, slowing down is easier than speeding up. This effect is equivalent in size to a difference between picture naming RT distributions of 133 ms. There was a medium effect of the *rate difference* control predictor, 0.0036 log Hz/s slope change per log Hz of rate difference [0.0034, 0.0039], Cohen's $d = 0.74$. The interaction effect was so small as to be meaningless, -0.00041 log Hz/s [-0.0019, 0.0011], Cohen's $d = 0.014$.

6.5 Discussion

This study sought to contribute to our understanding of how speakers control their speaking rate, by testing for an effect predicted by the notion of gaits in the context of the EPONA model of speech production. Specifically, we hypothesised that, if gaits were present, some pairs of speaking rates should be harder to switch between than other pairs of rates, once the difference between the initial and subsequent rates was accounted for. A further goal was to test support between the three remaining mappings of speaking rates onto gaits that Rodd et al. (2020) could not distinguish between, based on the assumption that shifting between rate pairs that involve crossing a gait boundary would be harder than between rate pairs that do not.

We found a clear effect of axis (Cohen's $d = 0.26$). We found shallower slopes for the fast↔medium axis than for the slow↔medium axis, indicating that speakers found it harder to switch between fast and medium speaking rates compared to switching between slow and medium speaking rates. This was interpreted as evidence for gaits in speech production because if no gaits were present, we would have predicted no difference between axes in their slope. Moreover, this specific result supports the fast-is-special mapping, under which the three sampled speaking rates are achieved by two gaits, grouping medium and slow together,

and using a separate gait for fast. This interpretation is based on the assumption that switching between gaits is harder than modulating speaking rate within them. This is in line with the task switching literature, where switching between tasks is more difficult than continuing to perform the same task (Meiran, 2010) and switching between more disparate tasks is harder than switching between tasks that are more similar (Arrington et al., 2003; Taube-Schiff & Segalowitz, 2005). To draw this conclusion, we are assuming that the different gaits represent more disparate tasks sets than the different tasks sets associated with modulating a single gait to be suitable for different speaking rates. One or more of the mechanisms that are proposed to be the underlying cause of switch costs, such as task-set inertia (whereby the previous task requires executive control engagement to suppress) or stimulus set binding (whereby the stimuli become associated with a task set) may explain why speakers find it harder to switch between rates involving crossing a gait boundary.

We also found an effect of acceleration. Slopes were steeper under deceleration than acceleration, meaning speakers found it easier to switch from a relatively faster rate to a relatively slower rate than to do the reverse. There are two appealing possible explanations for this. Firstly, the acceleration effect may result from an *asymmetric switch cost*, whereby it is more costly to switch from a harder task to an easier task than to make the reverse switch. Classical examples are the switch from colour naming to word naming in a Stroop task (e.g. MacDonald et al., 2000), or the switch from naming in a non-dominant language to naming in the dominant one (e.g. Costa et al., 2006). It is proposed that this paradoxical cost is the consequence of greater executive control engagement to keep the easier or dominant task inhibited, while performing the less dominant task. After switching task, this inhibition is slow to dissipate, causing the switch cost. In our case, the fast rate (syllable rate 4.39 Hz) is closest to the median spontaneous speaking rate for Dutch disyllabic words for demographically similar speakers to ours (6.86 Hz, based on modelling by Quené, 2008), suggesting that this most common rate might require the most inhibition, making it hardest to switch to.

That our fast rate was slower than Quené's estimate of a median speaking rate is a constraint inherent to multiple picture naming, which is a demanding task given the absence of predictability of upcoming words. The requirement to train

the participants to produce each rate also limited us in the number of rates that could be sampled.

Secondly, the acceleration effect may be a carry-over effect from engagement of an underlying domain-general rate control mechanism. Loehr et al. (2011) examined pianists' ability to coordinate their performance of complex rhythms to a metronome that slowed or sped up gradually. They also found that acceleration was harder than slowing down. They found better fits to the deviation from the required rhythm with an oscillator model, compared to a linear 'timekeeper' model, implying that the oscillator was the better model of the underlying rate control mechanism. To what extent this account can explain the acceleration effect remains an open question, as does the level of speech planning at which such an oscillator might be engaged.

Finally, we found a 'hypercorrection' effect, such that subsequent rates were more extreme than initial rates. This might be thought to be driven by a general communicative need to boost contrastivity, given the centrality of speech rate normalisation in perception (e.g. Maslowski et al., 2019b).

The presence of gaits, encompassing qualitatively similar configurations in the formulation phase of the speech production system is of consequence for our understanding of speaking rate control, and represents a conceptual foundation to allow further development of theory to link domain-general control mechanisms to speech production models. Although this study concerned speaking rate variation and demonstrated the presence of cognitive gaits to achieve different speaking rates, it is plausible to think of switching between qualitatively different parameter regimes as a more general mechanism to deliberately modulate the acoustic and temporal properties of speech to suit various communicative situations (Lindblom, 1990).

That we conclude that gaits are present, and that a gait boundary lies between fast and medium, but not between medium and slow speaking rates invites a few speculations. It might be supposed that the true system contains three or more gaits, adding a further gait faster than our 'fast' rate, or slower than our 'slow' rate. To test this, an alternative paradigm is required that allows sampling of faster and slower rates than possible with picture naming. Here we present the gait system of an average speaker, but it might be speculated that where gait boundaries fall varies somewhat between speakers, in line with variation in habitual speaking rate. Further, one might speculate that, because of the gaits,

speakers would display preferences in their speaking rates, as is the case in locomotion (Hoyt & Taylor, 1981), though to what extent such preferences might surface in speech, where many more contextual and communicative factors influence speaking style choice is unclear.

6.6 Conclusion

In this study, we examined how speakers switched their speaking rate between three pre-learned rates in a multiple picture naming task, equating slower rate transitions with more difficulty switching. We found that switching between fast and medium speaking rates was harder for speakers than switching between slow and medium rates. We also found that accelerating was harder than decelerating. This study has provided experimental evidence in favour of a conceptualisation of the formulation phase of speech planning whereby speakers switch between qualitatively distinct gaits to control their speaking rate, complementing the computational results of Rodd et al. (2020, Chapter 5), and is consistent with the EPONA model of formulation. Gaits in formulation have wide reaching theoretical implications for models of speech production, and potentially for our view of the mental lexicon in general.

7 General discussion

This thesis examined how the speech production system is voluntarily controlled to achieve different speaking rates. Control of speaking rate is an essential communicative skill, since speakers must design their speech to make it optimally communicative, given context (Hazan & Baker, 2011; Pouplier, 2012).

This question was explored through experiments and simulation of a computational model, EPONA, which was developed as a component of the thesis. This chapter summarises and discusses the results of the preceding chapters, and elaborates on the utility and implications of the EPONA model of speech production and the notion of gaits as a mechanism of speaking rate control.

7.1 Summary of contributions and findings

Chapters 2 and 3 were methodological chapters, and described and validated analysis tools that were developed to prepare the data for the later chapters.

A recurring problem in psycholinguistic modelling of speech is that models operate on units that make sense psychologically, but that are impossible to observe directly (e.g. Levelt, 1989; Levelt et al., 1999; Tourville & Guenther, 2011). This disconnect between model and what is feasible to test experimentally has limited progress in the development of models that describe more complex tasks than the production of monosyllables. In **Chapter 2**, two analysis approaches were developed that identified the onset and offset times of psychologically relevant planning units from electromagnetic articulography (EMA) tracks of articulator movement and from the acoustic speech signal only. These methods improve on previous practice by making it practical to test computationally implemented speech production models on multisyllabic and multi-word speech data, but necessarily constrain the possible speech materials to a subset of speech segments at critical positions in words. Despite using data of different modalities, the outcomes of the two metrics were substantially correlated, validating that they captured the same underlying construct.

The acoustic metric was used to prepare the PiNCeR corpus (Rodd, Bosker, ten Bosch, et al., 2019a, see Chapter 4), which was the dataset modelled by EPONA in Chapter 5. The software implementing these metrics is freely available for other researchers to use.

The quality of the segmentation of speech data that is obtained automatically is in general insufficient for fine-grained phonetic and psycholinguistic analyses. However, conventional manual segmentation of speech data is very time consuming. **Chapter 3** introduced a speech segmentation system, POnSS, that was developed to allow efficient segmentation to the word-level of the speech materials elicited in the various experiments. Compared to conventional, manual

segmentation using Praat TextGrids, POnSS enables much quicker segmentation of speech data to the word-level by dividing tasks that must be performed by humans into small chunks that can be distributed dynamically over a group of annotators, and, where feasible, automating aspects of the process. In a validation experiment, POnSS was shown to be as reliable as conventional segmentation, but about 20% faster. The software implementing POnSS is freely available.

Chapter 4 introduced a pair of related behavioural experiments and a speech corpus containing speech elicited in the experiments, which was then used in Chapter 5. In the experiments, speakers had to name pictures, in Dutch, at one of three pre-determined speaking rates. The pictures were arranged around a ‘clock-face’, and a dot jumped clockwise from picture to picture to indicate which picture was to be named when. The data were segmented with POnSS, and the analysis technique introduced in Chapter 2 was used to identify the onsets and offsets of syllable-level planning units. The resulting PiNCeR corpus contains productions of disyllabic words with known, stable speaking rates from 25 speakers, along with onset and offset times of syllable-level planning units.

Chapter 5 introduced our model of speech production, EPONA. The construction of the model is in itself an important contribution to the field of computational models of speech production; EPONA is the first computationally explicit model of the formulation phase of speech production that describes the temporal unfolding within words. It is also the first frame-based model to have a contemporary computational implementation suitable for fitting to speech data. Chapter 5 reported simulations of the model that aimed to ascertain how speakers adjusted the formulation system to control their speaking rate.

An evolution-inspired optimisation algorithm was used to find the values for the parameters of the model that resulted in simulated speech durations that most resembled those observed in Chapter 4. The multi-dimensional space formed by considering each parameter of the model as a dimension was taken as an approximation of the cognitive space of the human speaker. We found that the parameter values supporting the three speaking rates available in the PiNCeR corpus patterned in a triangle. This result indicated that at least two qualitatively distinct configurations of the speech production system were present, consistent with the hypothesis that speakers achieve rate control by switching within an inventory of such qualitatively distinct configurations, called *gaits* here. Furthermore, in a control analysis, we found that an optimisation run that had been constrained to only consider linear arrangements of the rates in parameter space was clearly outperformed by the unconstrained model, reinforcing our conclusion that there are multiple *gaits* involved in the formulation system. In a further analysis, we attempted to quantify support for different possible mappings of the three speaking rates in the PiNCeR corpus onto two or three *gaits* by statistically modelling the (non-)linearity in the outcome durations along axes between pairs of speaking rates. This analysis was inconclusive, possibly because the representation of the outcome in terms of syllable duration

and overlap duration was insufficiently rich for the predicted non-linearities to be reliably detectable.

Chapter 6 built on the findings of Chapter 5. It described a behavioural experiment that aimed to test the key conclusion of Chapter 5 more directly, without dependence on the computational model. In the experiment, speakers named pictures from a clock-face display, similar to the display used in Chapter 4. This time, speakers were taught to speak at three set speaking rates before the experiment. They then had to maintain the speaking rates themselves. The required speaking rate was indicated by the colour of a frame placed around the picture display. At an unpredictable moment during the trial, the colour of the frame changed, indicating that the speaker should adjust their speaking rate. We expected that speakers would differ in how quickly they would be able to switch between different pairs of speaking rates: faster adjustment would indicate that less cognitive reconfiguration was required to make the switch between the relevant rates. We found that speakers were quicker to switch between slow and medium speaking rates than they were to switch between fast and medium speaking rates. That there was a difference at all was interpreted as consistent with the gaits hypothesis: if no gaits were present, we would have predicted no difference to emerge. Beyond this general confirmation of support for the gaits hypothesis, Chapter 6 also aimed to lend support to one of the mappings of rates onto gaits, which had not been possible based on the analysis in Chapter 5. The finding that it was quicker to switch between slow and medium rates than between fast and medium rates implied a mapping where slow and medium are supported by one gait, and fast is supported by a second, qualitatively distinct gait.

Chapters 5 and 6 provide complimentary evidence from different approaches that speakers switch between qualitatively different configurations of the cognitive system that plans speech to achieve different speaking rates, much as some animals with legs switch between gaits to achieve different movement speeds.

7.2 What is a gait?

While Chapters 5 and 6 make clear that gaits are switched between to achieve gross shifts in speaking rate, what a gait precisely is, remains rather vague. Further research will be required to make progress towards a more concrete operationalisation of what a gait actually is, and what surface features in the speech signal are consistent within gaits. In the next subsections I make some suggestions regarding the nature, purpose, and consequences of gaits in the formulation system.

7.2.1 The link between speaking rate and reduction

A possible reason that gaits may exist is to prepare so-called *reduced pronunciation variants*. Alongside variation in the speaking rate, there is also extensive temporal variation in how words themselves are produced. Words are seldom pronounced in their canonical ‘dictionary’ form. Instead, acoustic cues, segments and sometimes entire syllables are degraded or omitted, generally when words are highly predictable and in informal spontaneous speaking situations (Ernestus & Warner, 2011; Johnson, 2004). For example, American English “yesterday”, the typical full form of which is [jɛstəreɪ], may be reduced to [jɛʃeɪ]. The context of the utterance (speaking style, sentence-level prosody, and in particular speaking rate) has a large bearing on the degree to which reduction phenomena appear. That more than just the immediate phonetic context and the speaking rate influence the degree to which a given word is reduced implies that reduction phenomena are not solely a product of phonetic processes in the execution phase, but must also involve adjustments to the higher-level planning process, i.e., the formulation phase (Ernestus, 2014). Reductions that emerge in formulation might be expected to be more categorical in nature (stepwise reductions in gesture duration or excursion, or binary omission of cues), while reductions arising in execution might be expected to be purely gradient (continuous reduction in gesture duration or physical excursion until a cue is no longer observable). Hanique et al. (2013) found evidence of categoricity in the distribution of the durations of phonemes eroded by a reduction process in Dutch. A further argument that reductions can emerge in formulation is the finding that words are less reduced if the following word is less predictable. This is explained by extra time for the more difficult planning of the upcoming word being ‘bought’ by reducing the word currently being produced comparatively less (e.g. Bell et al., 2009; Pluymaekers et al., 2005).

That reduction processes can arise at the formulation level and that variants can differ from each other categorically raises the question of how different pronunciation variants might be represented in the lexicon. One proposal is that multiple units in the lexicon are required, one for each variant. Such an arrangement is compatible with evidence that the relative frequency of pronunciation variants influences naming times (Bürki et al., 2010) and speed of recognition (Brand & Ernestus, 2018), but is problematic for the dominant theories of planning in speech production, which contend that each word is represented by one unit in the lexicon (e.g. Levelt, 1989; Caramazza, 1997; Levelt et al., 1999). Furthermore, a formal model of how variant pronunciations are connected and organised in the lexicon should also explain the mechanism by which contextual factors might influence the selection of a variant pronunciation.

It seems plausible that the mechanisms that support control of speaking rate in the formulation phase may be closely related to the mechanisms that support categorical reductions. One might hypothesise that the purpose of gait shifting

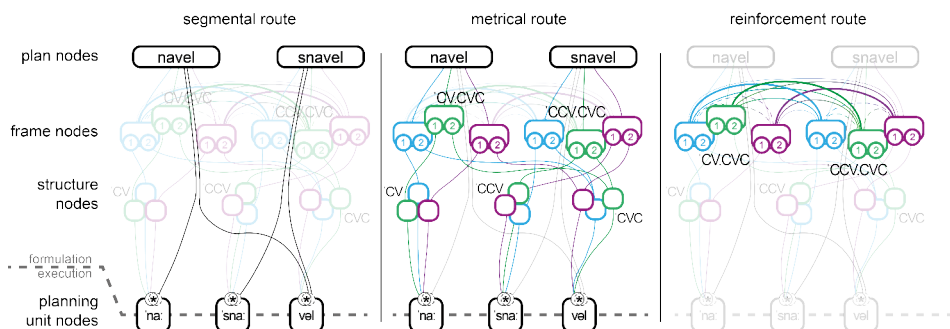


Figure 7.1: A sketch of an EPONA network containing the nodes necessary to produce the Dutch disyllabic words *navel* ['na:.vəl] 'navel' and *snavel* ['sna:.vəl] 'beak', adding the reinforcement route and multiple frame and structure nodes to capture gaited behaviour. Each panel shows one of the routes of the model, the remaining routes are shown in translucent behind. Colours indicate different gait reinforcement networks. See the text for details.

is to prepare speech involving different degrees of reduction, such that by engaging a gait, we are engaging a stratum of stored variants of different words that are similar in their degree of reduction. If that were true, there would be a number of observable effects that we should expect to see, that to my knowledge have not been tested for. Firstly, there should be trends in the frequency of reduced pronunciation variants by speaking rate; some variants should emerge at moderate speaking rates; some, presumably the most heavily reduced, only at the fastest rates. Furthermore, these trends should be non-linear: at different rates supported by the same gait, we should see similar patterns in the relative frequencies of reduced pronunciation variants.

7.2.2 EPONA as a network model

To be able to make testable predictions about the nature of gaits, and thus explore how gait selection would work as a mechanism to control speech rate in multi-word, continuous speech, a fully operationalised model is required. Although EPONA as presented in Chapter 5 is a 'single column' model describing the production of a single word in isolation from other words and without connections to representations associated with other words (see Figure 5.1), the model can be readily extended to form an interactive network model of the selection of word form variants that might form the basis of a model of the production of multi-word utterances.

A network view of EPONA, as illustrated in Figure 7.1, isolates the different parameter settings associated with each gait in a 'variant' frame node for the relevant word shape, which is in turn connected to 'variant' structure nodes en-

coding different temporal realisations of the relevant structure. Each word in the lexicon is connected to all frame nodes suitable to produce that word shape.

We postulate exhaustive excitatory and inhibitory connections between related variants of different frame nodes, and inhibitory connections between variants of the same frame node. This interconnection allows priming activation and suppression that tends to ensure that adjacent words are produced with the same speaking style. We will call these connections between the frame nodes the ‘reinforcement route’, consistent with the segmental and metrical routes.

In the reinforcement route, networks or families of related frame nodes, depicted in Figure 7.1 as different colours of nodes, are connected together by heavily weighted connections. Although three families of frame nodes are depicted in Figure 7.1 for each word form, it is clear that different word shapes will have differing number of frame node variants, reflecting different possibilities for categorical reduction. Some frame node variants will therefore belong to multiple interconnected networks of related frame nodes. A gait in the EPONA model is then the reinforcement network of strongly interconnected frame node variants, which tend to prime each other, and whose priming activation tends to suppress the frame node variants belonging to other gaits. The weightings of the connections in the reinforcement networks are established by associative learning, such that pairs of frame nodes that co-occur in time develop strong excitatory connections, and those that do not co-occur develop inhibitory connections. The excitatory connections within gait networks and inhibitory connections between them have the property of introducing inertia into the system: the default behaviour is always to continue speaking in the same gait. Priming between the nodes is a possible explanation for why switching between rates within a gait should be easier than switching between rates achieved by different gaits, as we find in Chapter 6. This is because executive control would be required to boost the target network and suppress the current network and other competitor networks to achieve switching. This momentary adjustive executive control might be thought of as a ‘gait shifter’. If executive control is invoked to moderate the activation in competing and target gait networks, it might be possible to detect correlations between the various components of executive control ability (Miyake et al., 2000) and individual differences between speakers in their success at modulating their speaking rate.

7.2.3 Gait networks predict ‘sweet spots’

By analogy with locomotive gaits, we might expect speaking gaits to have default rates at which they are most efficient. In human and animal locomotion, the selection of gaits is tightly linked to their relative efficiency. In horses, which typically have walking, trotting, and galloping gaits, each gait has a clear ‘sweet spot’ speed, at the approximate centre of the range of speeds achievable with that gait, where exertion (ml O₂ consumed to move 1 metre) is minimised (Hoyt &

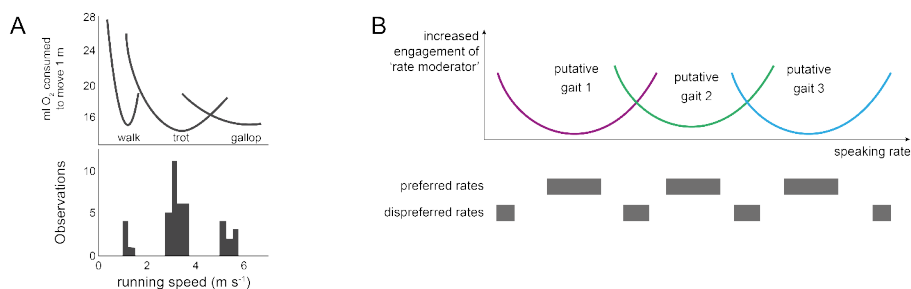


Figure 7.2: Panel A: Adapted from Figure 2 of Hoyt and Taylor (1981). Above: how the oxygen cost to move 1 metre declines to a minimum in each gait, and below, how a free-walking horse selects the speeds in the most efficient range of speeds for each gait. Panel B: A possible system of attractor basins for speaking gaits, showing three putative gaits, and below, indicating which gaits should be preferred and dispreferred in such a system.

Taylor, 1981, their Figure 2, key aspects of which are reproduced in Figure 7.2A). Horses and migratory animals select these speeds preferentially (Pennycuik, 1975), and avoid inefficient speeds in the shoulder of each gait.

In her proposal that qualitatively distinct coordination modes of the speech execution system resemble locomotive gaits, Pouplier (2012) suggested that different modes in execution were each optimal in different contexts, but all equivalent in acoustic outcome for the listener and effortfulness for the speaker. In the case of the cognitive gaits considered in this thesis, a more direct analogy from the locomotion system is possible. Hoyt and Taylor (1981) depict a fit to their data where each gait is a U-shaped curve (Figure 7.2A). The most efficient speed for each gait is in the valley of the curve for that gait. The curves for different gaits intersect, showing speeds where multiple gaits are viable. Such an arrangement can be thought of as a system of attractor basins, where the speeds that are most efficient are the attractors, and speeds that are hard to achieve by any of the gaits are repellers between the basins. This attractor basins concept may apply directly to speaking rates, too. Such an arrangement is sketched in Figure 7.2B, showing three putative gaits, and should be identifiable by testing for preference effects in speaking rates. It is unlikely that preferences in speaking rate are as readily observable as those in horse locomotion; if they were, they would have long been reported. Instead, they are presumably obscured by speaking rate variation arising in satisfaction of communicative aims. It may, however be possible to detect preferences experimentally by quantifying how difficult individual rates are to maintain. Two possible experiments that may be able to do this are described in Box 7.1.

Box 7.1: Experimental approaches to detect rate preferences

Experimental approaches may be able to detect preferences for certain rates by ‘amplifying’ the biases inherent to the cognitive process of speaking. One way to do this would be to ask speakers to read aloud, staying synchronised to a recording of the same text presented over headphones. While the speaker was reading, the recording would drop out, but the speaker would have to continue. Throughout the trial, the speaking rate would be tracked, rather as was done in Chapter 6. Depending on the rate at which they had started, speakers would presumably slow down or speed up to a rate that was easier to achieve, in effect ‘sliding’ down the flanks of the attractor basins towards the sweets spots.

Another possible approach would be to have speakers shadow productions of sentences at known, stable speaking rates, then iteratively shadow their own productions, like a game of ‘broken telephone’ (e.g. Jacoby & McDermott, 2017; Griffiths & Kalish, 2007). Over a number of iterations, their rate would deviate, again either speeding up or slowing down to move in the direction of the sweet spots of the gaits. In contrast to the first experiment, it would be difficult, however, to know whether the biases detected by this experiment arose in comprehension (because participants perceived the speech as slower or faster) or in production.

Either of these experiments would allow the reconstruction of a system of attractor basins as illustrated in Figure 7.2B, and would allow the sampling of many more speaking rates than the experiments employed in Chapters 4 and 6 support.

7.3 Speaking rate control within gaits

This thesis describes the mechanism of gross rate control as switching between gaits. How finer-grained, within-gait rate switching might work has not been explicitly discussed in this thesis. However, if there are relatively few gaits, which is the assumption underlying the experiment in Chapter 6 and the suggestion that gaits represent categorically distinct reduction possibilities, then some further control mechanism must exist to moderate speaking rate within gaits.

One possible fine-grained control mechanism is implied by the alternative hypothesis in Chapter 5, namely that rate control is simply achieved by tweaking the ‘gain’ in the system. An obvious candidate for such a gain manipulation is the general level of activation in the stratum of plan nodes in EPONA; such that more activation results in faster productions. This is broadly consistent with strategies hypothesised to be engaged to control the time course of lexical pro-

cessing (Ratcliff, 1978; Lupker et al., 1997; Kello & Plaut, 2000). I will term this mechanism the ‘rate moderator’.

To describe how the rate moderator can stretch and compress the realisation of speech, we require a more explicit account of the execution phase, such as the DIVA model (Guenther, 2016b; Tourville & Guenther, 2011) or the FACTs model (Parrell et al., 2019). I will take the DIVA model as the starting point for now. An assumption inherent to EPONA and to the working model introduced in Chapter 1 is that the activation of the content node is one and the same as the activation of the associated planning unit in the speech sound map (SSM) that forms the input to the feedforward controller of the DIVA model, consistent with the idea of cascading activation (e.g. Goldrick & Blumstein, 2006). In the feedforward controller of DIVA, the ‘playback’ speed of the motor target is linearly related to the magnitude of the activation of the planning unit: the feedforward controller subtracts the current motor state from the motor target and then multiplies it by a ‘GO’ signal and a constant to calculate the movement commands to the articulators. This means that a higher level of planning unit activation will linearly speed up production, and result in higher velocity of articulator movement. Conversely, a lower level of content node activation will result in a linearly slowed down production, and a lower velocity of articulator movement.

This linearity between speaking rates that are close enough together to be achieved by the same gait should be detectable, as attempted to do in Section 5.6.2 of Chapter 5. This type of linearity is of course in contrast to the classic finding of non-linear scaling between more disparate rates (e.g. Gay, 1981). In the light of the gaits findings presented in this thesis I would argue that the non-linearities detected by Gay (1981) instead represent the consequences of both within-gait and across-gait speech rate modulation.

7.4 A working theory of mechanisms of speaking rate control

In this section, the proposals for mechanisms of speaking rate control are summarised into a working theory, based on the EPONA model as described in Chapter 5. This working theory is illustrated in Figure 7.3A.

I propose that speakers have two main ways to control their speaking rate, which they must use in concert to achieve a continuous range of speaking rates. The first mechanism, the **gait shifter**, is a form of executive control that is engaged momentarily and changes the relative activation of gait ‘reinforcement networks’ of frame nodes. It does so by boosting activation in the network suitable for the target speech rate and suppressing activation in the current network and other competitor networks. The gait shifter is invoked by a variety of higher level processes, such as: voluntary, conscious gait shifts, of the type elicited in Chapter 6 of this thesis; rate shifts associated with automatic responses to en-

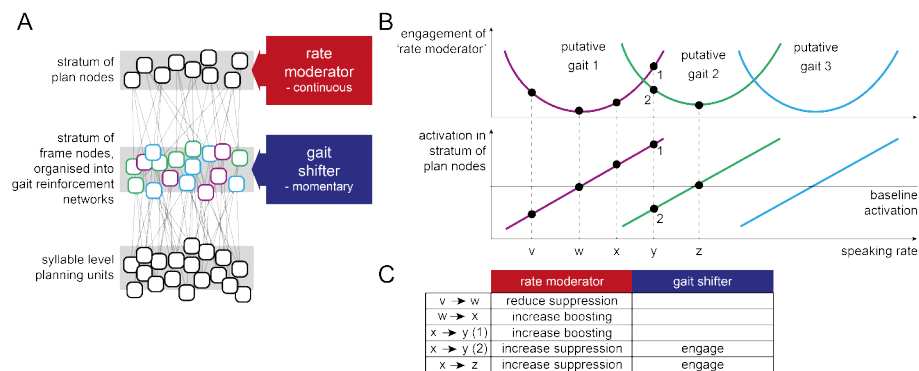


Figure 7.3: Panel A: A working theory of mechanisms of speaking rate control, showing the rate moderator, which continuously modulates the activation of the stratum of planning nodes, and the gait shifter, which momentarily engages to modulate the activation in the gait reinforcement networks of frame nodes. Panel B: A possible system of attractor basins for speaking gaits, showing three putative gaits, and below, the general level of activation (y axis) in the stratum of plan nodes that the rate moderator needs to achieve to move to that rate, a separate line indicates each gait. A number of example speaking rates are indicated (letters). Some rates are achievable using multiple gaits, such as rate *y*. Panel C: the engagement of the rate moderator and gait shifter to achieve different speaking rate shifts.

vironmental noise, as discussed later in Section 7.5.2; and temporary rate shifts undertaken to mark prosodic structure or mark prosodic prominence, as discussed later in Section 7.5.3. The second mechanism, the **rate moderator**, is continuously engaged to a greater or lesser extent, to achieve deviation from the natural, ‘sweet-spot’ rate of a gait, at which its engagement is minimised. To speed up, the rate moderator boosts the activation of the strata of plan nodes; to slow down, the rate moderator suppresses this activation.

Figure 7.3B shows a putative system of gait attractor basins, with five speaking rates labelled, and beneath a depiction of the general level of activation in the stratum of plan nodes. To shift from rate *v* to rate *w*, the rate moderator reduces its suppression of activation in the plan node stratum, which causes the faster unfolding of metrical sequences in the frame nodes and faster playback of the syllable-level planning units. To shift from rate *w* to rate *x*, the rate moderator boosts activation in the plan node stratum, which further speeds up the unfolding of metrical sequences and the playback of planning units. To shift from rate *x* to rate *y*, there are two options: either to use only the rate moderator to further boost activation in the plan node stratum, and thus speaking rate, or to switch gait by momentarily invoking the gain shifter to boost activation in the network for gait 2 and suppress activation in the other gait networks, and simultaneously use the rate moderator to suppress activation in the plan node stratum. Which

option is chosen from each starting point would be a means to diagnose the relative cost of engaging each moderator. To shift from rate x to rate z , however, invoking the gait shifter is the only option, since rate z cannot be achieved by gait 1.

7.5 Implications of gaits for phenomena adjacent to speaking rate control

The findings in this thesis are based on experiments concerning voluntary control of speaking rate, but the notion of gaits as a mechanism of speaking rate control can be readily extended to other phenomena that might be thought to contribute to inter- and intra-speaker rate variation or co-occur with speaking rate variation. This section discusses how gaits might inform thinking about these phenomena and how these phenomena might be integrated into the notion of gaits to work towards a more comprehensive theory of speaking style control.

7.5.1 Inter-speaker variation in gait inventories and habitual speaking rate

In this thesis, it is assumed that all speakers have a similar inventory of gaits for the three rates sampled in the experiments. The participants who are included in the studies are demographically similar; all are native Dutch-speaking members of the Radboud University community, with an average age of 22.7 years, with relatively little variation (s.d. 2.77 years), which led us to assume that they would behave similarly. Furthermore, our sample of three very disparate speaking rates meant that it was unlikely that, even if speakers did vary in precisely which rate ranges were achieved by each gait, we would be able to detect such variation.

To identify the mapping between speaking rates and gaits of individual speakers, which might well prove to vary, different experimental designs are required, whereby many more speaking rates can be sampled than was possible with the picture naming experiments reported in Chapters 4 and 6. Such dense sampling would be possible with either of the experiments described in Box 7.1. Variation in the inventory of gaits might be an explanation for a further highly salient feature of variation in speaking rate, variation in habitual speaking rate.

Part of that variation is not yet understood, and part is conditioned by language and dialect (Quené, 2008), but there are also demographic trends in that variation. Differences in gait inventory may explain both of these. One often reported trend is that older speakers tend to speak more slowly (Duchin & Mysak, 1987; Harnsberger et al., 2008; but c.f. Quené, 2013). In the context of gaits, an appealing explanation for this slowing down would be a change in the gait inventory over the lifetime, such that the fastest gaits are not engaged by older

speakers. Establishing whether older speakers do indeed differ in gait inventory is highly feasible using the experimental techniques suggested in Box 7.1. However, the directionality of such a change would be hard to establish; do older speakers reduce their speaking rate for other reasons and the gait network supporting the fastest rates therefore deteriorate from disuse, or is a change in the gait inventory instead the cause of the change in behaviour?

7.5.2 Variation in speaking rate driven by communicative and environmental context

Voluntarily deciding to speed up or slow down is likely not the most frequent reason that speakers adjust their speaking rate; instead, they do this in response to environmental and communicative context in support of communicative success (e.g. Hazan & Baker, 2011). Manipulating communicative context while maintaining tight experimental control is challenging. Techniques that might be hoped to do so, such as virtual reality (Peeters, 2019), have the potential to elicit changes in speech behaviour in response to features of the virtual interlocutor (e.g. Heyselaar et al., 2017), but implementation of experiments in virtual reality is expensive and time consuming (Casasanto & Jasmin, 2018). The ventriloquist paradigm, which combines pre-recorded speech with a human interlocutor (e.g. Felker et al., 2018) may well prove to be the optimal way to perform such manipulations as the technique matures.

Manipulating environmental context is however much easier. The Lombard effect, whereby speakers slow down in response to noise (Lombard, 1911; van Summers et al., 1988), can be elicited simply by presenting noise binaurally over headphones. This effect is largely automatic (Garnier et al., 2010), making it a good counterpoint to the voluntary rate control examined in this thesis. Alongside the slowing down effect, Lombard speech is also characterised by increased intensity, increased F0, enhanced amplitude modulations in the temporal envelope, and increased power in the higher frequencies of the spectrum (Bosker & Cooke, 2018; Bosker & Cooke, in press; Lu & Cooke, 2008). It is unknown to what extent the speech rate components and vocal effort components of the Lombard effect are independent of each other, or are controlled together. My intuition is that the slowing down component of the Lombard effect may be supported by the gait shifting mechanism described in this thesis. To test this, the switching experiment described in Chapter 6 could be repeated, but rather than training and cueing participants to speak at particular rates, noise would be switched on or off during the trial. The same analysis as in Chapter 6 would reveal how quickly speaking rate was shifted in response to the noise. If the slope coefficients found were comparable to those measured in Chapter 6, it could be concluded that the gait mechanism is also engaged to shift speaking rate in response to environmental noise.

If the vocal effort components of the Lombard effect arise through the same control process, then they should take a similar amount of time to engage (or shut down) as the slowing down in speaking rate, which would emerge as a similar slope coefficient. However, my hunch is that the two aspects of Lombard speech are controlled independently, and would thus have differing slope coefficients as they become engaged in response to noise onset. This result would call into question whether it is valid to characterise Lombard speech as a single phenomenon or as a constellation of speaking style changes that co-occur.

7.5.3 Intonation and prosodic structure

Speaking rate and intonation are very tightly related: speaking rate is typically considered bundled together with intonation in treatments of prosody (Ladd, 2008), and it has been suggested that speech rate variation contributes to the marking of prosodic phrasing (Nooteboom & Eefting, 1994). Furthermore, the realisation of an intonational melody is highly dependent on the speaking rate (Gussenhoven, 2005). Words may be lengthened preceding prosodic phrase boundaries, or when they should be marked as extra prominent (Wightman et al., 1992). Reduction phenomena and prominence are inversely related; such that only the least prominent words get reduced (e.g. Pluymaekers et al., 2005). This suggests that, if (categorical) reductions are determined largely by the currently engaged gait, the relative activation of the other gait networks needs to be modulated momentarily to ensure that the most prominent, most information bearing words are less reduced than other words.

7.5.4 Scope of frame nodes and planning units

The key component of the gait networks described in Section 7.2.2 is the frame node at the heart of the metrical route. This node captures the word-level temporal and metrical structure, mediating the retrieval of syllable-level planning units. That the unit that mediates between the formulation and execution phases is a syllable is a well established assumption made in various models of speech production. For instance, the Levelt et al. (1999) model provides syllabified output. The DIVA model (Guenther, 2016b; Tourville & Guenther, 2011) is agnostic with regards to the length of the planning unit stored in the speech sound map, though syllable level units are privileged in the DIVA literature over segmental or word-level planning units. This ‘mental syllabary’ of stored syllable representations (Levelt & Wheeldon, 1994) is a parsimonious idea, and one that is adopted in many models (Crompton, 1982; but c.f. Goldstein & Fowler, 2003; Levelt et al., 1999; Varley & Whiteside, 2001; Walsh et al., 2010; Whiteside & Varley, 1998).

A key testable prediction of the syllabary concept is that there should be a response time advantage for high frequency syllables compared to low frequency syllables, since the high frequency syllables can be retrieved whilst the low fre-

quency syllables need to be composed. The predicted response time advantage for high frequency syllables is small but well established (in Dutch: Levelt & Wheeldon, 1994; Cholin et al., 2006; Cholin & Levelt, 2009; in Spanish: Domínguez et al., 1993; Perea & Carreiras, 1995; Perea & Carreiras, 1998; Carreiras & Perea, 2004; in French: Brand et al., 2002; Laganaro & Alario, 2006; Perret et al., 2014; in German patients with apraxia of speech and controls: Aichert & Ziegler, 2004; Staiger & Ziegler, 2008; in English: Cholin et al., 2011; Croot et al., 2017). Evidence for the syllabary concept also comes from articulation, where timing differences have been observed between consonant clusters composed of segments from two adjacent syllables (in “bass cap”) or from segments from one syllable (in “a scab” or “mask amp”; Byrd, 1996). A further source of evidence is repetition-suppression in MRI (Peeva et al., 2010), which showed sensitivity to syllabic units in the left ventral premotor cortex, the region associated with the speech sound map (planning units) in the DIVA model (Guenther, 2016b; Guenther et al., 2006). Lesion of this region (among others) has also been implicated in acquired apraxia of speech, a disorder which is thought to reflect impaired retrieval or unpacking of planning units, or impaired feedforward control of the motor system (Maas et al., 2015; Terband et al., 2019; Varley et al., 1999).

Although this evidence suggests that syllables are a real unit of the planning system, it does not exclude that syllable-level units and units at other levels co-occur. A mixture of syllables and other units is consistent with Levelt and Wheeldon’s (1994) mental syllabary, where syllable representations were proposed to be available for the most frequent syllables, which are ‘over-learned’, and online composition would be required for the rest.

Formalised models of the production system that account for variation promise to provide further evidence to address this fundamental question of what the units in the lexicon actually are, since these units must be implicated in rate modulation strategies used by speakers. This might emerge through instantiation of variant frame nodes suitable for different speaking rates and degrees of reduction, as suggested in Section 7.2.2. It should be possible to identify the boundary between over-learned syllables and online composed syllables in how they are manipulated to speed up or slow down. Further modelling with EPONA implementing different assumed granularities of frame nodes and planning units and the gait reinforcement networks may reveal which levels of granularity are best at predicting the speakers’ choice of categorically distinct reduced pronunciation variants.

7.6 Conclusion

The research in this thesis has indicated that, to control the rate of speaking, speakers switch between qualitatively distinct configurations of the cognitive system (i.e., the formulation phase) that plans their speech. This was shown

by simulating speech produced at different speaking rates with a computational model of speech production, EPONA. In some ways, these configurations resemble the gaits that animals with legs adopt to move at different speeds, like walking and running. We therefore termed them the ‘gaits of speech’. Like locomotive gaits, speech gaits are suitable for a range of speeds. In a sample of three speaking rates, we established that relatively fast speech was achieved by one gait, whereas medium and slower speaking rates were achieved by a second gait, by showing that speakers found it easier to switch between medium and slow speaking rates than they did to switch between medium and fast speaking rates.

To reach this conclusion, new analysis methods were developed. The first of these methods facilitates the identification of the onset and offset times of psychologically relevant planning units from the acoustic speech signal. Next, an efficient system for the segmentation of speech data was developed. This allowed us to identify the onsets and offsets of words relatively quickly, to prepare for modelling with EPONA. The EPONA model itself is also innovative, as the first computationally explicit production model that can account for speakers’ ability to voluntarily adjust speaking rate.

Future work may aim to map the inventory of gaits more extensively, which may reveal differences between individuals and explain, for instance, why speakers have different habitual speaking rates. The EPONA model and the notion of gaits may also be extended to account for adjacent phenomena in the speech production system, such as reduction and intonation.

References

- Adams, S. G., Weismer, G., & Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech, Language, and Hearing Research*, 36(1), 41–54.
- Aichert, I., & Ziegler, W. (2004). Syllable frequency and syllable structure in apraxia of speech. *Brain and Language*, 88(1), 148–159. [https://doi.org/10.1016/S0093-934X\(03\)00296-7](https://doi.org/10.1016/S0093-934X(03)00296-7)
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716–723.
- Alexander, R. M. (1989). Optimization and gaits in the locomotion of vertebrates. *Physiological Reviews*, 69(4), 1199–1227. <https://doi.org/10.1152/physrev.1989.69.4.1199>
- Arrington, C. M., Altmann, E. M., & Carr, T. H. (2003). Tasks of a feather flock together: Similarity effects in task switching. *Memory & Cognition*, 31(5), 781–789. <https://doi.org/10.3758/BF03196116>
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19(1), 3–11. <https://doi.org/10.2466/pr0.1966.19.1.3>
- Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease*, 163(5), 307.
- Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1), 92–111. <https://doi.org/10.1016/j.jml.2008.06.003>
- Bella, S. D., & Palmer, C. (2011). Rate Effects on Timing, Key Velocity, and Finger Kinematics in Piano Performance. *PLOS ONE*, 6(6), e20518. <https://doi.org/10.1371/journal.pone.0020518>
- Bentum, M., Ernestus, M., ten Bosch, L., & van den Bosch, A. (2019). Do speech registers differ in the predictability of words? *International Journal of Corpus Linguistics*, 24(1), 98–130.
- Bhati, S., Nayak, S., Murty, K. S. R., & Dehak, N. (2019). Unsupervised Acoustic Segmentation and Clustering Using Siamese Network Embeddings. *Proc. Interspeech 2019*, 2668–2672.
- Birkholz, P., Steiner, I., & Breuer, S. (2007). Control concepts for articulatory speech synthesis. In P. Wagner, J. Abresch, S. Breuer, & W. Hess (Eds.), *Sixth isca tutorial and research workshop on speech synthesis*.
- Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological review*, 89(1), 1–47.

- Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer [computer program]. Retrieved August 15, 2015, from <http://www.fon.hum.uva.nl/praat/>
- Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer [computer program]. <http://www.fon.hum.uva.nl/praat/>
- Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of cognitive neuroscience*, 22(7), 1504–1529.
- Booij, G. (1995). *The phonology of Dutch*. Clarendon Press.
- Boomer, D. S., & Laver, J. D. M. (1968). Slips of the Tongue. *British Journal of Disorders of Communication*, 3(1), 2–12. <https://doi.org/10.3109/13682826809011435>
- Bosker, H. R., & Cooke, M. (2018). Talkers produce more pronounced amplitude modulations when speaking in noise. *The Journal of the Acoustical Society of America*, 143(2), EL121–EL126.
- Bosker, H. R., & Cooke, M. (in press). Enhanced amplitude modulations contribute to the lombard intelligibility benefit: Evidence from the nijmegen corpus of lombard speech. *Journal of The Acoustic Society of America*.
- Boyce, S., & Espy-Wilson, C. Y. (1997). Coarticulatory stability in American English /r/. *The Journal of the Acoustical Society of America*, 101(6), 3741–3753. <https://doi.org/10.1121/1.418333>
- Brand, M., Rey, A., Peereman, R., & Spieler, D. (2002). Naming bisyllabic words: A large scale study. *Abstracts of the Psychonomic Society*, 7, 94.
- Brand, S., & Ernestus, M. (2018). Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: Evidence from native listeners and learners of french. *The Quarterly Journal of Experimental Psychology*, 71(5), 1240–1259.
- Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3-4), 155–180.
- Buhrmester, M., Talaifar, S., & Gosling, S. (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science*, 13(2), 149–154. <https://doi.org/10.1177/1745691617706516>
- Bürki, A. (2018). Variation in the speech signal as a window into the cognitive architecture of language production. *Psychonomic Bulletin & Review*, 25(6), 1973–2004. <https://doi.org/10.3758/s13423-017-1423-4>
- Bürki, A., Ernestus, M., & Frauenfelder, U. H. (2010). Is there only one “fenêtre” in the production lexicon? On-line evidence on the nature of phonological representations of pronunciation variants for French schwa words. *Journal of Memory and Language*, 62(4), 421–437.
- Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>

- Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, 24(2), 209–244.
- Byrd, D., & Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics*, 24(2), 263–282. <https://doi.org/10.1006/jpho.1996.0014>
- Cambier-Langeveld, T., Nespors, M., & Heuven, V. J. v. (1997). The domain of final lengthening in production and perception in Dutch. *EUROSPEECH-1997*, 931–934.
- Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive neuropsychology*, 14(1), 177–208.
- Carreiras, M., & Perea, M. (2004). Naming pseudowords in Spanish: Effects of syllable frequency. *Brain and Language*, 90(1), 393–400.
- Casasanto, D., & Jasmin, K. M. (2018). Virtual reality. In A. M. B. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and neurobiology of language: A practical guide*. Wiley-Blackwell.
- Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, 26(5), 609–651.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review*, 113(2), 234–272.
- Cholin, J., Dell, G. S., & Levelt, W. J. M. (2011). Planning and articulation in incremental word production: Syllable-frequency effects in English. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1), 109.
- Cholin, J., & Levelt, W. J. M. (2009). Effects of syllable preparation and syllable frequency in speech production: Further evidence for syllabic units at a post-lexical level. *Language and cognitive processes*, 24(5), 662–684.
- Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition*, 99(2), 205–235.
- Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403), 596–610.
- Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language*, 28(2), 543–571.
- Costa, A., Santesteban, M., & Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 1057.
- Crompton, A. (1982). Syllables and segments in speech production. *Slips of the tongue and language production*. Walter de Gruyter.
- Croot, K., Lalas, G., Biedermann, B., Rastle, K., Jones, K., & Cholin, J. (2017). Syllable frequency effects in immediate but not delayed syllable naming in English. *Language, Cognition and Neuroscience*, 1–14. <https://doi.org/10.1080/23273798.2017.1284340>

- Damian, M. F., & Dumay, N. (2007). Time pressure and phonological advance planning in spoken production. *Journal of Memory and Language*, 57(2), 195–209. <https://doi.org/10.1016/j.jml.2006.11.001>
- Deb, K., & Agrawal, R. B. (1995). Simulated binary crossover for continuous search space. *Complex systems*, 9(2), 115–148.
- Deb, K., & Goyal, M. (1996). A combined genetic adaptive search (GeneAS) for engineering design. *Computer Science and informatics*, 26, 30–45.
- Deb, K., & Jain, H. (2014). An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Transactions on Evolutionary Computation*, 18(4), 577–601. <https://doi.org/10.1109/TEVC.2013.2281535>
- Deb, K., Sindhya, K., & Okabe, T. (2007). Self-adaptive simulated binary crossover for real-parameter optimization, 1187–1194. <https://doi.org/10.1145/1276958.1277190>
- Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3), 283–321.
- Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological review*, 104(1), 123–147.
- Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42(1–3), 287–314. [https://doi.org/10.1016/0010-0277\(92\)90046-K](https://doi.org/10.1016/0010-0277(92)90046-K)
- Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*.
- Domínguez, A., Cuetos, F., & de Vega, M. (1993). Efectos diferenciales de la frecuencia silábica: Dependencia del tipo de prueba y características de los estímulos. *Estudios de Psicología*, 14(50), 3–31.
- Duchin, S. W., & Mysak, E. D. (1987). Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of communication disorders*, 20(3), 245–257.
- Dusan, S., & Rabiner, L. (2006). On the relation between maximum spectral transition positions and phone boundaries. *Ninth International Conference on Spoken Language Processing*.
- Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A Dynamical Model of Saccade Generation During Reading. *Psychological Review*, 112(4), 777–813. <https://doi.org/10.1037/0033-295X.112.4.777>
- Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, 142, 27–41. <https://doi.org/10.1016/j.lingua.2012.12.006>
- Ernestus, M., Hanique, I., & Verboom, E. (2015). The effect of speech situation on the occurrence of reduced word pronunciation variants. *Journal of Phonetics*, 48, 60–75. <https://doi.org/10.1016/j.wocn.2014.08.001>

- Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants [Editorial]. *Journal of Phonetics*, 39(S1), 253–260. [https://doi.org/10.1016/S0095-4470\(11\)00055-6](https://doi.org/10.1016/S0095-4470(11)00055-6)
- European Language Resources Association. (2019). *Catalogues of the European Language Resources Association*. Retrieved December 18, 2019, from <http://www.elra.info/en/>
- Felker, E., Troncso-Ruiz, A., Ernestus, M., & Broersma, M. (2018). The ventriloquist paradigm: Studying speech processing in conversation with experimental control over phonetic input. *The Journal of the Acoustical Society of America*, 144(4), EL304–EL309.
- Fink, A., Oppenheim, G. M., & Goldrick, M. (2018). Interactions between lexical access and articulation. *Language, Cognition and Neuroscience*, 33(1), 12–24. <https://doi.org/10.1080/23273798.2017.1348529>
- Furui, S. (2003). Recent advances in spontaneous speech recognition and understanding. *ISCA & IEEE workshop on spontaneous speech processing and recognition*.
- Garnham, A., Shillcock, R. C., Brown, G. D., Mill, A. I., & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics*, 19(7-8), 805–818.
- Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of sound immersion and communicative interaction on the lombard effect. *Journal of Speech, Language, and Hearing Research*. [https://doi.org/10.1044/1092-4388\(2009/08-0138\)](https://doi.org/10.1044/1092-4388(2009/08-0138))
- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). *Timit acoustic phonetic continuous speech corpus* (tech. rep.). Linguistic Data Consortium. <https://catalog.ldc.upenn.edu/LDC93S1>
- Garrett, M. (1976). Syntactic processes in language production. In R. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 231–256).
- Gay, T. (1981). Mechanisms in the Control of Speech Rate. *Phonetica*, 38(1-3), 148–158. <https://doi.org/10.1159/000260020>
- Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, 21(6), 649–683. <https://doi.org/10.1080/01690960500181332>
- Goldrick, M., Keshet, J., Gustafson, E., Heller, J., & Needle, J. (2016). Automatic analysis of slips of the tongue: Insights into the cognitive architecture of speech production. *Cognition*, 149, 31–39. <https://doi.org/10.1016/j.cognition.2016.01.002>
- Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. *Phonetics and phonology in language comprehension and production: Differences and similarities*, 159–207.

- Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, 31(3-4), 465–485.
- Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, 31(3), 441–480.
- Grossberg, S. (1978). Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, 17(3), 199–219.
- Guenther, F. H. (2016a). Feedforward Control. *Neural Control of Speech* (pp. 193–220). MIT Press.
- Guenther, F. H. (2016b). *Neural Control of Speech*. MIT Press.
- Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, 96(3), 280–301.
- Gussenhoven, C. (2005). Transcription of Dutch intonation. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 118–145). Oxford University Press.
- Gut, U., & Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. *Speech Prosody 2004*.
- Hadka, D. (2017). Platypus: A Free and open source Python library for multiobjective optimization. Retrieved October 4, 2017, from <https://github.com/Project-Platypus/Platypus>
- Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 11–15.
- Hanique, I., Ernestus, M., & Schuppler, B. (2013). Informal speech processes can be categorical in nature, even if they affect many different words. *The Journal of the Acoustical Society of America*, 133(3), 1644–1655.
- Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). *Deep Speech: Scaling up end-to-end speech recognition* (tech. rep.) [arXiv:1412.5567]. arXiv.
- Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological review*, 106(3), 491.
- Harnsberger, J. D., Shrivastav, R., Brown, W., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of voice*, 22(1), 58–69. <https://doi.org/10.1016/j.jvoice.2006.07.004>
- Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31(3), 373–405. <https://doi.org/10.1016/j.wocn.2003.09.006>

- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1), 77–89. <https://doi.org/10.1080/19312450709336664>
- Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, 130(4), 2139–2152. <https://doi.org/10.1121/1.3623753>
- Heyselaar, E., Hagoort, P., & Segaert, K. (2017). In dialogue with an avatar, language behavior is identical to dialogue with a human partner. *Behavior research methods*, 49(1), 46–60.
- Hickok, G. (2014). Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. *Language, Cognition and Neuroscience*, 29(1), 52–59. <https://doi.org/10.1080/01690965.2013.852907>
- Hirschberg, J. (2000). A corpus-based approach to the study of speaking style. *Prosody: Theory and experiment* (pp. 335–350). Springer.
- Hoang, D.-T., & Wang, H.-C. (2015). Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. *The Journal of the Acoustical Society of America*, 137(2), 797–805.
- Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., & Saltzman, E. (1996). Accurate recovery of articulator positions from acoustics: New conclusions based on human data. *The Journal of the Acoustical Society of America*, 100(3), 1819–1834.
- Holovaty, A., & Kaplan-Moss, J. (2009). *The definitive guide to django: Web development done right*. Apress.
- Houghton, G. (1990). The problem of serial order: A neural network model of sequence learning and recall. *Current research in natural language generation* (pp. 287–319). Academic Press.
- Hoyt, D. F., & Taylor, C. R. (1981). Gait and the energetics of locomotion in horses. *Nature*, 292(5820), 239–240. <https://doi.org/10.1038/292239a0>
- Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological bulletin*, 140(2), 339–373.
- Illa, A., & Ghosh, P. K. (2018). Low resource acoustic-to-articulatory inversion using bi-directional long short term memory. *Proc. Interspeech 2018*, 3122–3126.
- Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, 128(2), 839–850. <https://doi.org/10.1121/1.3459842>
- Jacoby, N., & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, 27(3), 359–370. <https://doi.org/10.1016/j.cub.2016.12.031>

- Johnson, K. (2004). Massive reduction in conversational American English. *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium*, 29–54.
- Johnson, K. (2008). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (p. 363). John Wiley & Sons.
- Jongman, S. R., Roelofs, A., & Meyer, A. S. (2015). Sustained attention in language production: An individual differences investigation. *Quarterly Journal of Experimental Psychology*, 68(4), 710–730. <https://doi.org/10.1080/17470218.2014.964736>
- Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing, 2nd Edition*. Prentice Hall.
- Kaufeld, G., Ravenschlag, A., Meyer, A. S., Martin, A. E., & Bosker, H. R. (2020). Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(3), 549–562.
- Kello, C. T. (2004). Control over the time course of cognition in the tempo-naming task. *Journal of Experimental Psychology: Human Perception and Performance*, 30(5), 942. <https://doi.org/10.1037/0096-1523.30.5.942>
- Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3), 719. <https://doi.org/10.1037/0278-7393.26.3.719>
- Kello, C. T., & Plaut, D. C. (2003). Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory and Language*, 48(1), 207–232. [https://doi.org/10.1016/S0749-596X\(02\)00512-0](https://doi.org/10.1016/S0749-596X(02)00512-0)
- Kello, C. T., Plaut, D. C., & MacWhinney, B. (2000). The task dependence of staged versus cascaded processing: An empirical and computational study of Stroop interference in speech production. *Journal of Experimental Psychology: General*, 129(3), 340–360.
- Kelso, J. A., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, 14(1), 29–59.
- Kennedy, J. (2011). Particle Swarm Optimization. *Encyclopedia of Machine Learning* (pp. 760–766). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_630
- Kipp, A., Wessenick, M.-B., & Schiel, F. (1997). Pronunciation modeling applied to automatic segmentation of spontaneous speech. *Fifth European Conference on Speech Communication and Technology*.
- Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1), 61–70. <https://doi.org/10.1177/001316447003000105>

- Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280. <https://doi.org/10.1177/2515245918771304>
- Kuehn, D. P., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 4(4), 303–320.
- Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.
- Laganaro, M., & Alario, F. X. (2006). On the locus of the syllable frequency effect in speech production. *Journal of Memory and Language*, 55(2), 178–196.
- Lecouteux, B., Linarès, G., & Oger, S. (2012). Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. *Computer Speech & Language*, 26(2), 67–89.
- Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41–104. [https://doi.org/10.1016/0010-0277\(83\)90026-4](https://doi.org/10.1016/0010-0277(83)90026-4)
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT press.
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(01), 1–38.
- Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, 50(1–3), 239–269. [https://doi.org/10.1016/0010-0277\(94\)90030-2](https://doi.org/10.1016/0010-0277(94)90030-2)
- Lindblom, B. (1983). Economy of Speech Gestures. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 217–245). Springer New York. https://doi.org/10.1007/978-1-4613-8202-7_10
- Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H & H Theory. *Speech Production and Speech Modelling* (pp. 403–439). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-2037-8_16
- Lindblom, B., Brownlee, S., Davis, B., & Moon, S.-J. (1991). Speech transforms. *Phonetics and Phonology of Speaking Styles*, 4.1–4.10.
- Lindblom, B., Lubker, J., & Gay, T. (1977). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of The Acoustic Society of America*, 62, S15.
- Linguistic Data Consortium. (2019). *Archives of the Linguistic Data Consortium*. <https://www.ldc.upenn.edu/>
- Loehr, J. D., Large, E. W., & Palmer, C. (2011). Temporal coordination and adaptation to rate change in music performance. *Journal of Experimental Psychology: Human Perception and Performance*, 37(4), 1292–1309. <https://doi.org/10.1037/a0023102>
- Lombard, E. (1911). Le signe de l'elevation de la voix. *Ann. Mal. de L' Oreille et du Larynx*, 101–119.
- Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, 124(5), 3261–3275. <https://doi.org/10.1121/1.2990705>
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

- Lupker, S. J., Brown, P., & Colombo, L. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(3), 570.
- Maas, E., Mailend, M.-L., & Guenther, F. H. (2015). Feedforward and feedback control in apraxia of speech: Effects of noise masking on vowel production. *Journal of Speech, Language, and Hearing Research*, 58(2), 185–200.
- MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, 288(5472), 1835–1838. <https://doi.org/10.1126/science.288.5472.1835>
- MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia*, 8(3), 323–350. [https://doi.org/10.1016/0028-3932\(70\)90078-3](https://doi.org/10.1016/0028-3932(70)90078-3)
- MacKay, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology*, 3(2), 210–227. [https://doi.org/10.1016/0010-0285\(72\)90004-7](https://doi.org/10.1016/0010-0285(72)90004-7)
- Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in psychology*, 7, 120. <https://doi.org/10.3389/fpsyg.2016.00120>
- Martí, L., Garcia, J., Berlanga, A., & Molina, J. M. (2009). An approach to stopping criteria for multi-objective optimization evolutionary algorithms: The MGBM criterion. *2009 IEEE Congress on Evolutionary Computation*, 1263–1270. <https://doi.org/10.1109/CEC.2009.4983090>
- Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019a). How the tracking of habitual rate influences speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(1), 128.
- Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019b). Listeners normalize speech for contextual speech rate even without an explicit recognition task. *The Journal of the Acoustical Society of America*, 146(1), 179–188. <https://doi.org/10.1121/1.5116004>
- Mathet, Y., & Widlöcher, A. (2011). Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. In Lafourcade, Mathieu, Prince, & Violaine (Eds.), *Actes de la 18e Conférence Traitement Automatique des Langues Naturelles (TALN'11)* (pp. 247–258).
- Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3), 437–479. https://doi.org/10.1162/COLI_a_00227
- McMillan, C. T., & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition*, 117(3), 243–260. <https://doi.org/10.1016/j.cognition.2010.08.019>
- Meiran, N. (2010). Task switching: Mechanisms underlying rigid vs. flexible self control. *Self control in society, mind and brain*, 202–220.

- Meyer, A. S., Wheeldon, L., Van der Meulen, F., & Konopka, A. (2012). Effects of speech rate and practice on the allocation of visual attention in multiple object naming. *Frontiers in psychology*, 3.
- Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation Rate and Its Variability in Spontaneous Speech: A Reanalysis and Some Implications. *Phonetica*, 41(4), 215–225. <https://doi.org/10.1159/000261728>
- Minetti, A. E. (1998). The biomechanics of skipping gaits: A third locomotion paradigm? *Proceedings of the Royal Society B: Biological Sciences*, 265(1402), 1227–1235.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive psychology*, 41(1), 49–100.
- Nam, H., Goldstein, L., Saltzman, E., & Byrd, D. (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *The Journal of the Acoustical Society of America*, 115(5), 2430–2430.
- Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., & Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. *The Journal of the Acoustical Society of America*, 132(6), 3980–3989.
- Nooteboom, S. G., & Eefting, W. (1994). Evidence for the adaptive nature of speech on the phrase level and below. *Phonetica*, 51(1-3), 92–98.
- Oostdijk, N. H. J. (2000). Het corpus gesproken Nederlands (the spoken Dutch corpus). *Nederlandse Taalkunde*, 5, 280–284.
- Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *The Journal of the Acoustical Society of America*, 77(2), 640–648.
- Parrell, B. (2012). The role of gestural phasing in Western Andalusian Spanish aspiration. *Journal of Phonetics*, 40(1), 37–45. <https://doi.org/10.1016/j.wocn.2011.08.004>
- Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLOS Computational Biology*, 15(9), e1007321. <https://doi.org/10.1371/journal.pcbi.1007321>
- Peer, E. S., Bergh, F. v. d., & Engelbrecht, A. P. (2003). Using neighbourhoods with the guaranteed convergence PSO. *Proceedings of the 2003 IEEE Swarm Intelligence Symposium, 2003. SIS '03*, 235–242. <https://doi.org/10.1109/SIS.2003.1202274>
- Peeters, D. (2019). Virtual reality: A game-changing method for the language sciences. *Psychonomic bulletin & review*, 26(3), 894–900.
- Peeva, M. G., Guenther, F. H., Tourville, J. A., Nieto-Castanon, A., Anton, J.-L., Nazarian, B., & Alario, F. -X. (2010). Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech produc-

- tion network. *NeuroImage*, 50(2), 626–638. <https://doi.org/10.1016/j.neuroimage.2009.12.065>
- Pennycuik, C. J. (1975). On the running of the gnu (*Connochaetes taurinus*) and other animals. *Journal of Experimental Biology*, 63(3), 775–799.
- Perea, M., & Carreiras, M. (1995). Efectos de frecuencia silábica en tareas de identificación. *Psicológica*, 16(3), 483.
- Perea, M., & Carreiras, M. (1998). Effects of syllable frequency and syllable neighborhood frequency in visual word recognition. *Journal of Experimental Psychology: Human perception and performance*, 24(1), 134.
- Perret, C., Schneider, L., Dayer, G., & Laganaro, M. (2014). Convergences and divergences between neurolinguistic and psycholinguistic data in the study of phonological and phonetic encoding: A parallel investigation of syllable frequency effects in brain-damaged and healthy speakers. *Language, Cognition and Neuroscience*, 29(6), 714–727. <https://doi.org/10.1080/01690965.2012.678368>
- Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory Planning Is Continuous and Sensitive to Informational Redundancy. *Phonetica*, 62(2-4), 146–159. <https://doi.org/10.1159/000090095>
- Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research: Data collection and scaling* (pp. 90–105). MIT Press.
- Poupier, M. (2012). The gaits of speech. In M.-J. Solé & D. Recasens i Vives (Eds.), *The Initiation of Sound Change: Perception, Production, and Social Factors* (pp. 147–164). John Benjamins Publishing.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.
- Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06*, University of Cambridge, Cambridge, 26–46.
- Prasad, M., van Esch, D., Ritchie, S., & Mortensen, J. F. (2019). Building Large-Vocabulary ASR Systems for Languages Without Any Audio Training Data. *Proc. Interspeech 2019*, 271–275.
- 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.
- Quené, H. (2013). Longitudinal trends in speech tempo: The case of queen beatrix. *The Journal of the Acoustical Society of America*, 133(6), EL452–EL457.
- R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>

- R Development Core Team. (2008). *R: A language and environment for statistical computing* [ISBN 3-900051-07-0]. R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org/>
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, 85(2), 59.
- Raymond, W. D., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R., & Hilt, C. (2002). An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. *Seventh International Conference on Spoken Language Processing*.
- Richmond, K. (2006). A trajectory mixture density network for the acoustic-articulatory inversion mapping. *Ninth International Conference on Spoken Language Processing*.
- Richmond, K., Hoole, P., & King, S. (2011). Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. *Twelfth Annual Conference of the International Speech Communication Association*.
- Rietbergen, M., Roelofs, A., den Ouden, H., & Cools, R. (2018). Disentangling cognitive from motor control: Influence of response modality on updating, inhibiting, and shifting. *Acta Psychologica*, 191, 124–130. <https://doi.org/10.1016/j.actpsy.2018.09.008>
- Rietveld, T., Ernestus, M. et al. (2004). Pitfalls in corpus research. *Computers and the Humanities*, 38(4), 343–362.
- Rodd, J., Bosker, H. R., Ernestus, M., Alday, P. M., Meyer, A. S., & ten Bosch, L. (2020). Control of speaking rate is achieved by switching between qualitatively distinct cognitive ‘gaits’: Evidence from simulation. *Psychological Review*, 127(2), 281–304. <https://doi.org/10.1037/rev0000172>
- Rodd, J., Bosker, H. R., Ernestus, M., ten Bosch, L., & Meyer, A. S. (under review). Asymmetric switch costs between speaking rates: Evidence for gaits of speech planning.
- Rodd, J., Bosker, H. R., ten Bosch, L., Ernestus, M., & Meyer, A. S. (2019a). PiNCeR: A corpus of cued-rate multiple picture naming in Dutch. *PsyArXiv*. <https://doi.org/10.31234/osf.io/wyc6h>
- Rodd, J., Bosker, H. R., ten Bosch, L., & Ernestus, M. (2019b). Deriving the onset and offset times of planning units from acoustic and articulatory measurements. *The Journal of the Acoustical Society of America*, 145(2), EL161–EL167. <https://doi.org/10.1121/1.5089456>
- Rodd, J., & Chen, A. (2016). Pitch accents show a Perceptual Magnet Effect: Evidence of internal structure in intonation categories. *Speech Prosody 2016*, 697–701.
- Rodd, J., Decuyper, C., Bosker, H. R., & ten Bosch, L. (in press). A tool for efficient and accurate segmentation of speech data: Announcing POnSS. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-020-01449-6>
- Roelofs, A. (2008). Dynamics of the attentional control of word retrieval: Analyses of response time distributions. *Journal of Experimental Psychology: General*,

- 137(2), 303. Retrieved December 9, 2016, from <http://psycnet.apa.org/journals/xge/137/2/303/>
- Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4), 333–382.
- Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and language*, 162, 42–45.
- Schiel, F. (2015). A statistical model for predicting pronunciation. In M. Wolters, J. Livingstone, B. Beattie, J. Stuart-Smith, & J. Scobbie (Eds.), *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow.
- Scobbie, J. M., & Pouplier, M. (2010). The role of syllable structure in external sandhi: An EPG study of vocalisation and retraction in word-final English /l/. *Journal of Phonetics*, 38(2), 240–259. <https://doi.org/10.1016/j.wocn.2009.10.005>
- Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*, 295–342.
- Sjerps, M. J., Decuyper, C., & Meyer, A. S. (2019). Initiation of utterance planning in response to pre-recorded and “live” utterances. *Quarterly Journal of Experimental Psychology*. <https://doi.org/10.1177/1747021819881265>
- Slootweg, A. (1988). Metrical prominence and syllable duration. *Linguistics in the Netherlands 1988*, 139–138.
- Smorenburg, L., Rodd, J., & Chen, A. (2015). The effect of explicit training on the prosodic production of L2 sarcasm by Dutch learners of English. *Proceedings of the 18th International Congress of Phonetic Sciences*.
- Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, 6(2), 174.
- Staiger, A., & Ziegler, W. (2008). Syllable frequency and syllable structure in the spontaneous speech production of patients with apraxia of speech. *Aphasiology*, 22(11), 1201–1215. <https://doi.org/10.1080/02687030701820584>
- Stan Development Team. (2018). RStan: The R interface to Stan [R package version 2.18.2]. <http://mc-stan.org/>
- Steiner, I., & Richmond, K. (2009). Towards unsupervised articulatory resynthesis of German utterances using EMA data. *Tenth Annual Conference of the International Speech Communication Association*.
- Stemberger, J. P. (1985). An Interactive Activation Model of Language Production. *Progress in the Psychology of Language* (pp. 143–186). Lawrence Erlbaum Associates.
- Stemberger, J. P. (1991). Apparent anti-frequency effects in language production: The Addition Bias and phonological underspecification. *Journal of*

- Memory and Language*, 30(2), 161–185. [https://doi.org/10.1016/0749-596X\(91\)90002-2](https://doi.org/10.1016/0749-596X(91)90002-2)
- Taube-Schiff, M., & Segalowitz, N. (2005). Linguistic Attention Control: Attention Shifting Governed by Grammaticized Elements of Language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 508–519. <https://doi.org/10.1037/0278-7393.31.3.508>
- ten Bosch, L. F. M., & Cranen, B. (2007). A computational model for unsupervised word discovery. *Proc. Interspeech 2007*, 1481–1884.
- Terband, H., Rodd, J., & Maas, E. (2015). Simulations of feedforward and feedback control in apraxia of speech (AOS): Effects of noise masking on vowel production in the DIVA model. *Proceedings of the 18th International Congress of Phonetic Sciences*.
- Terband, H., Rodd, J., & Maas, E. (2019). Testing hypotheses about the underlying deficit of apraxia of speech (AOS) through computational neural modeling with the DIVA model. *International Journal of Speech-Language Pathology*. <https://doi.org/10.1080/17549507.2019.1669711>
- The Language Archive. (2019). *The Language Archive*. <https://tla.mpi.nl>
- Tilsen, S. (2014). Selection and coordination of articulatory gestures in temporally constrained production. *Journal of Phonetics*, 44, 26–46.
- Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics*, 55, 53–77. <https://doi.org/10.1016/j.wocn.2015.11.005>
- Tjaden, K., & Weismer, G. (1998). Speaking-rate-induced variability in f2 trajectories. *Journal of Speech, Language, and Hearing Research*, 41(5), 976–989.
- Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, 26(7), 952–981. <https://doi.org/10.1080/01690960903498424>
- Uria, B., Renals, S., & Richmond, K. (2011). A deep neural network for acoustic-articulatory speech inversion. *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*.
- van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3), 917–928. <https://doi.org/10.1121/1.396660>
- van Bael, C., van den Heuvel, H., & Strik, H. (2007). Validation of phonetic transcriptions in the context of automatic speech recognition. *Language Resources and Evaluation*, 41(2), 129–146.
- van Brenk, F., Terband, H., Van Lieshout, P., Lowit, A., & Maassen, B. (2013). Rate-related kinematic changes in younger and older adults. *Folia Phoniatrica et Logopaedica*, 65(5), 239–247.
- Varley, R., & Whiteside, S. P. (2001). What is the underlying impairment in acquired apraxia of speech? *Aphasiology*, 15(1), 39–84. <https://doi.org/10.1080/02687040042000115>

- Varley, R., Whiteside, S. P., & Luff, H. (1999). Apraxia of speech as a disruption of word-level schemata: Some durational evidence. *Journal of Medical Speech-Language Pathology*, 7(2), 127–132.
- Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of phonetics*, 71, 147–161.
- Vaz, C., Toutios, A., & Narayanan, S. S. (2016). Convex Hull Convolutional Non-Negative Matrix Factorization for Uncovering Temporal Patterns in Multivariate Time-Series Data. *INTERSPEECH*, 963–967.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Verhoeven, J., De Pauw, G., & Kloots, H. (2004). Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, 47(3), 297–308.
- Vousden, J. I., Brown, G. D., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive psychology*, 41(2), 101–175.
- Vousden, J. I., & Maylor, E. A. (2006). Speech errors across the lifespan. *Language and Cognitive Processes*, 21(1-3), 48–77. <https://doi.org/10.1080/01690960400001838>
- Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological review*, 114(3), 830–841.
- Walsh, M., Möbius, B., Wade, T., & Schütze, H. (2010). Multilevel exemplar theory. *Cognitive science*, 34(4), 537–582.
- Weisser, M. (2016). DART—The dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory*, 12(2), 355–388.
- Westbury, J., Milenkovic, P., Weismer, G., & Kent, R. (1990). X-ray microbeam speech production database. *The Journal of the Acoustical Society of America*, 88(S1), S56–S56. <https://doi.org/10.1121/1.2029064>
- Whiteside, S., & Varley, R. (1998). Dual-route phonetic encoding: Some acoustic evidence. *ICSLP*.
- Widlöcher, A., & Mathet, Y. (2012). The Glozz Platform: A Corpus Annotation and Mining Tool. *Proceedings of the 2012 ACM Symposium on Document Engineering*, 171–180. <https://doi.org/10.1145/2361354.2361394>
- Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, 120(2), 329–355. <https://doi.org/10.1037/a0031542>
- Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3), 1707–1717.

- Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3), 392–399. <https://doi.org/10.2307/2346786>
- Winkelmann, R., Harrington, J., & Jänsch, K. (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, 45, 392–410. <https://doi.org/10.1016/j.csl.2017.01.002>
- Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition* (2 edition). Chapman; Hall/CRC.
- Wright, C. E., & Meyer, D. E. (1983). Conditions for a linear speed-accuracy trade-off in aimed movements. *The Quarterly Journal of Experimental Psychology Section A*, 35(2), 279–296. <https://doi.org/10.1080/14640748308402134>
- Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using tobi. *INTERSPEECH*.
- Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2006). The HTK book (for HTK version 3.4). *Cambridge University Engineering Department*, 2(2).
- Ypma, J., Borchers, H. W., & Eddelbuettel, D. (2018). Package ‘nloptr’.
- Zambrano-Bigiarini, M., & Rojas, R. (2018). *HydroPSO: Particle swarm optimisation, with focus on environmental models* [R package version 0.4-1.]. <https://doi.org/10.5281/zenodo.1287350>
- Zitzler, E., Brockhoff, D., & Thiele, L. (2007). The Hypervolume Indicator Revisited: On the Design of Pareto-compliant Indicators Via Weighted Integration. In S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, & T. Murata (Eds.), *Evolutionary Multi-Criterion Optimization* (pp. 862–876). Springer Berlin Heidelberg.
- Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory*, 27(3), 340–352. <https://doi.org/10.1080/09658211.2018.1510966>

Appendix: elicitation materials

Table A.1: Filler words were included in the first, penultimate and last slots of each trial.





orthography	phonetic form	meaning	
gieter	¹ xi.tər	watering can	
kabel	¹ ka:.bəl	cable	
lasser	¹ la.sər	welder	
lichaam	¹ liχ.a:m	body	
molen	¹ mo:.lən	windmill	
monnik	¹ mə.nik	monk	
spiegel	¹ spi.xəl	mirror	
tafel	¹ ta.fəl	table	
trommel	¹ trə.məl	drum	

Table A.1: Filler words were included in the first, penultimate and last slots of each trial. *(continued)*



orthography	phonetic form	meaning	
vinger	¹ za.ŋər	finger	
zanger	¹ za.ŋər	singer	

Table A.2: Target words were included in the second to sixth slot of each trial.








orthography	phonetic form	meaning	
hagel	¹ ha:.xəl	hail	
hamer	¹ ha:.mər	hammer	
havik	¹ ha:.mər	hawk	
nagel	¹ na:.xəl	finger nail	
navel	¹ na:.vəl	navel	
sinus	¹ si.nʊs	sine wave	
slager	¹ sla:.xər	butcher	

Table A.2: Target words were included in the second to sixth slot of each trial.
(continued)

orthography	phonetic form	meaning
snavel	¹ snɑ:.vəl	beak
visum	¹ vi.sum	visa
vlieger	¹ vli.xər	kite
vriezer	¹ vri.zər	freezer
wafel	¹ wa:ˌfəl	waffle
zoemer	¹ zu.mər	alarm

Nederlandse samenvatting

Iedereen kent wel iemand die bijzonder snel spreekt, of juist buitengewoon traag. Naast verschillen tussen sprekers van dezelfde taal, bestaan er ook verschillen tussen talen in hoe snel ze gesproken worden. En zelfs één en dezelfde spreker kan variëren in hoe snel zij of hij spreekt. Zo gaan sprekers bijvoorbeeld trager of sneller spreken, afhankelijk van de communicatieve situatie. Dit gebeurt zowel automatisch, afhankelijk van omgevingsgeluid (zoals in een luidruchtige kroeg), als vrijwillig, bijvoorbeeld in gesprek met iemand met een slecht gehoor, of iemand die de taal nog aan het leren is. Meestal vinden mensen het moeilijk dit soort vrijwillige spraaksnelheidsaanpassingen vol te houden, wat impliceert dat er wat cognitieve inzet nodig is om spraaksnelheid aan te passen.

Om vloeiend te spreken, moeten sprekers abstracte concepten in concrete woorden omzetten, en deze vervolgens weer omzetten in plannen voor hoe de spieren van het spraakkanaal moeten bewegen. Dit hele proces noemen we ‘spraakvoorbereiding’. Huidige theorieën van spraakvoorbereiding hebben geen verklaring voor hoe sprekers vrijwillig hun spraaksnelheid aan zouden passen.

Dit proefschrift haakt in op de vraag hoe sprekers hun spraakvoorbereiding aanpassen om op verschillende snelheden te kunnen spreken. Dit is interessant om drie redenen. Ten eerste, een goede theorie van spraakvoorbereiding moet variatie in spraaksnelheid kunnen verklaren, aangezien variatie zo veel voorkomt. Ten tweede, als we begrijpen hoe spraaksnelheid aangepast wordt, begrijpen we ook beter hoe ‘executive control’-processen (zoals aandacht) op spraakvoorbereiding inwerken. Ten derde, het is nuttig om uit te zoeken welk deel van de variatie in spraaksnelheid opzettelijk is, en welk deel ruis is. Dit draagt bij aan het ontwikkelen van nieuwe theorieën van spraakvoorbereiding.

Dit proefschrift bevat een inleiding, vijf empirische hoofdstukken, en een afsluitende discussie. Hoofdstukken 2, 3 en 4 beschrijven methodes die ik heb ontwikkeld en gebruikt om de data te verzamelen en voor te bereiden. Hoofdstukken 5 en 6 beschrijven simulaties en experimenten, die direct bijdragen aan het beantwoorden van de hoofdvraag.

In **Hoofdstuk 2** beschrijf ik een manier om vast te stellen wanneer, in lopende spraak, lettergrepen precies beginnen en eindigen. Dit doe ik alleen aan de hand van het opgenomen geluid van de spraak. Dit probleem is verrassend moeilijk,

omdat lettergrepen akoestisch gezien in elkaar ‘overlopen’. Ik heb een methode ontwikkeld die kijkt naar hoeveel het akoestische signaal verandert en vervolgens de pieksnelheid van verandering identificeert. Het begin van deze piek nemen we over als het begin van de nieuwe lettergreep. Het eind van de piek laat het eind van de vorige lettergreep zien. De resultaten van deze methode heb ik vergeleken met begin- en eindtijden van lettergrepen die bepaald waren op basis van bewegingsmetingen van de tong en kaak. De uitkomsten van deze twee methodes bleken sterk gecorreleerd te zijn. Dit onderstreept de validiteit van mijn nieuwe akoestische methode.

In **Hoofdstuk 3** ontwikkel ik een oplossing voor een tweede praktisch probleem. Onderzoek naar spraakproductie houdt normaliter in dat de begin- en eindtijden van woorden handmatig vastgesteld worden. Dit wordt handmatig gedaan omdat volautomatische spraakherkenningssoftware onvoldoende betrouwbaar is. Deze handmatige ‘segmentatie’ is een zeer tijdrovende praktijk, en kan ook nog eens foutgevoelig zijn. Onderzoeksassistenten gebruiken software die voor algemene fonetiek bedoeld is om de segmentatie uit te voeren. Deze software is niet geoptimaliseerd voor deze taak. De experimenten in dit proefschrift hebben veel data geleverd, wat er toe leidde dat het wenselijk was om meer gebruik te maken van automatisatie, en om de handmatige procedure efficiënter te laten verlopen.

Mijn oplossing was om een speciaal databanksysteem te bouwen, POnSS, dat de spraak op automatische wijze kon segmenteren. Daarna konden de onderzoeksassistenten inloggen bij een speciale omgeving, waarbij de woorden een voor een afgespeeld werden, samen met een visuele weergave van de bijbehorende geluidsgolven. Dan moesten ze beslissen of de segmentatie voldoende accuraat was. In een tweede omgeving, pasten ze de begin- en eindtijden van woordsegmentaties die afgekeurd werden in de vorige fase aan. De omgevingen zijn te zien in Figuur 3.2, op pagina 37. Ik heb vervolgens geanalyseerd of de segmentaties afkomstig van POnSS even betrouwbaar waren als die van de conventionele procedure. Dit bleek inderdaad zo te zijn. Ik heb ook berekend hoe lang het zou duren om dezelfde hoeveelheid spraakdata op conventionele wijze te segmenteren, in vergelijking tot POnSS. Het bleek dat POnSS 23% sneller was.

In **Hoofdstuk 4** beschrijf ik het experiment dat ik heb gebruikt om de spraakdata mee te eliceren (‘op te nemen’). Deze data heb ik vervolgens gebruikt in de simulaties in Hoofdstuk 5. Sprekers moesten plaatjes benoemen in het Neder-

lands op drie vooraf bepaalde spraaksnelheden (snel, gemiddeld en langzaam). De plaatjes werden weergegeven op een soort wijzerplaat. Er was een rood puntje dat van plaatje naar plaatje met de klok mee sprong, om aan te geven welke plaatje wanneer benoemd moest worden. Op die manier specificeerde ik de spraaksnelheid. Vervolgens liet ik de data segmenteren met behulp van POnSS (Hoofdstuk 3), en zette ik de analyse uit Hoofdstuk 2 in om de begin- en eindtijden van de lettergrepen vast te stellen.

Hoofdstuk 5 richt zich op het beantwoorden van de overkoepelende onderzoeksvraag. Ik introduceer twee mogelijke hypothesen over hoe spraaksnelheid gereguleerd wordt in het cognitieve systeem dat verantwoordelijk is voor spraakvoorbereiding. Eén mogelijkheid is dat snel spreken in essentie hetzelfde is als langzaam spreken, maar slechts versneld op lineaire wijze. Dit zou betekenen dat de snelheidsbeheersing op een gaspedaal zou lijken. Ik noem dit de ‘gaspedaal’-hypothese. Anderzijds zou het kunnen zijn dat sprekers kwalitatief verschillende cognitieve configuraties aannemen voor verschillende snelheden. Dit zou te vergelijken zijn met de verschillende ‘gangen’ die dieren met benen of poten aannemen om voort te bewegen, (zoals paarden, die drie gangen kennen, stap, draf en galop). Het zou kunnen dat sprekers een ‘loop-spreken’ configuratie aannemen om langzaam te praten, maar dat ze ‘ren-spreken’ voor snelle spraak. Ik noem deze mogelijkheid de ‘gangen’-hypothese.

Om uit te zoeken welke hypothese klopt, heb ik een model van spraakvoorbereiding geprogrammeerd dat voortbouwt op bestaande theorieën. Het doel van het model is het nabootsen van het tijds patroon van de spraak van echte sprekers die meededen aan het experiment van Hoofdstuk 4.

Het model heeft parameters die te vergelijken zijn met draaiknoppen. Verschillende standen van deze parameters zorgen voor verschillende voorspellingen. Ik gebruikte kunstmatige intelligentie om standen te vinden die leidden tot een goede nabootsing van de echte spraak. Dit gebeurde met vallen en opstaan: het algoritme probeerde verschillende standen uit, en berekende hoeveel het resultaat van het model op de echte spraak leek. Dit werd duizenden keren herhaald, om tot optimale standen te komen. Dit gebeurde afzonderlijk voor iedere snelheid, wat er toe leidde dat ik verschillende standen vond die optimaal waren voor snelle, gemiddelde en langzame spraak.

Om te kunnen onderscheiden tussen de ‘gaspedaal’- en ‘gangen’-hypothesen, bekeek ik vervolgens de standen van de parameters van het model die opti-

maal waren voor iedere spraaksnelheid. Dit deed ik door ze (als het ware) in een multidimensionale ruimte te tekenen, waarbij de verschillende parameters de dimensies vormden. De gaspedaal-hypothese voorspelt dat de drie spraaksnelheden een rechte lijn vormen binnen de 'parameter-ruimte'. De gangen-hypothese voorspelt daarentegen dat de drie spraaksnelheden in een driehoek zouden zitten. Ik vond een driehoekige opstelling, wat overeen kwam met de gangen-hypothese.

Dit resultaat liet zien dat er verschillende configuraties, of gangen, in het cognitieve systeem bestaan. Maar ik kon nog niet zeggen hoeveel gangen er waren, of welke spraaksnelheden gerealiseerd werden met welke gangen. Om daar achter te komen, voerde ik nog een laatste experiment uit. Dit experiment beschrijf ik in **Hoofdstuk 6**.

In dat experiment gebruikte ik een vergelijkbare wijzerplaat met plaatjes als ik in Hoofdstuk 4 gebruikte. Dit keer moesten de sprekers vooraf de drie snelheden leren, waarna ik het rode puntje dat de spraaksnelheid aangaf weghaalde. Hierna moesten ze de snelheid dus zelf handhaven. De benodigde spraaksnelheid (langzaam, gemiddeld, snel) gaf ik aan door middel van een kleurenbalk rondom de plaatjes. Op een onvoorspelbaar moment tijdens het spreken, veranderde de balk van kleur, als signaal voor de spreker om van spraaksnelheid te wisselen.

Ik verwachtte dat er verschillen zouden voorkomen in hoe snel sprekers zouden kunnen wisselen tussen snelheidsparen. Wanneer de sprekers hun spraaksnelheid heel snel aan de nieuwe snelheid konden aanpassen, dan zou dit laten zien dat ze het relatief makkelijk vonden om tussen de snelheden te wisselen, terwijl langzamere aanpassingen zouden laten zien dat sprekers het relatief moeilijker vonden. Ik vond dat sprekers sneller konden wisselen tussen langzaam en gemiddelde snelheden dan tussen gemiddelde en snelle spraak. Dit komt overeen met de uitkomst van Hoofdstuk 5, en daarnaast laat dit me concluderen dat er twee 'gangen' zijn: één voor langzame en gemiddelde spraaksnelheden, en een tweede voor de snelste spraaksnelheden.

Samenvattend, laten de resultaten van dit proefschrift zien dat het waarschijnlijk is dat spraaksnelheid wordt aangestuurd door te schakelen tussen verschillende configuraties van het cognitieve spraakvoorbereidingsstelsel. Een interessante vervolgstudie zou kunnen laten zien of er nog meer verschillen zijn tussen de verschillende 'gangen'.

English summary

Everyone knows someone who speaks particularly fast, or especially slowly. Different languages and language varieties have different typical speaking rates, as do different individuals who speak the same language variety. Speaking rates are also highly variable within individuals. Speakers slow down or speed up, depending on the communicative situation. This happens both automatically, in response to environmental noise (like in a busy pub), and voluntarily, for instance when speaking to someone who is hard of hearing, or is a learner of the language being spoken. We normally find it hard to maintain this kind of voluntary speech rate change, which implies that extra cognitive effort is required to adjust our speaking rate.

To speak fluently, speakers must turn abstract concepts into concrete words and then into plans for how to move the muscles of the vocal tract. We will call this process ‘speech planning’. The state-of-the art theories of speech planning don’t explain how speakers voluntarily vary the rate at which they speak.

This thesis addresses the question of how speakers tweak speech planning, so that they can speak at different rates. This is interesting for three reasons: Firstly, a good theory of speech planning should account for variation, since it is so prevalent. The account that I propose and the data that we collect will help inform how theories might explain voluntary variation in speaking style. Secondly, understanding how speech planning is adjusted may clarify how ‘executive control’ processes (like attention) interact with speech planning itself. Thirdly, identifying which fraction of the variation in speaking rate is deliberate, and which fraction is just ‘random noise’ may also help in developing new theories of speech planning.

The thesis contains an introduction, five main chapters, and a general discussion. Chapters 2, 3 and 4 describe tools and methods used to collect and prepare the data. Chapters 5 and 6 describe simulations and experiments that were designed to answer the primary research questions.

In **Chapter 2**, I describe a technique to work out when, within words, syllables start and end in running speech, using just the recorded sound of the speech. This is a surprisingly difficult problem, because syllables overlap in time. I developed a method that looks at how much the acoustic signal changes, and finds

peaks in the rate of acoustic change. The beginning of the peak is associated with the start of the new syllable, and the end of the peak is associated with the end of the previous syllable. The results of this method were compared with the syllable times that we found by examining data taken from measurements of tongue and jaw movements. I found that the results of the two techniques were strongly correlated with each other. This showed that my new acoustic method was valid.

In **Chapter 3**, I develop a solution to a second practical problem. Typically, research in speech production involves manually determining when each word begins and ends. This is because automatic speech recognition systems don't work reliably enough. This 'segmentation' is a very time consuming process, which can be error-prone if not done carefully enough. Research assistants use general purpose phonetics software to perform the segmentation, which is not optimised for this specific task. The experiments in this thesis generated lots of speech data, so it was desirable to have a system to make more use of automation, and make the manual intervention more efficient. My solution was to construct a database system, POnSS, that automatically segmented the speech. Research assistants could then log in to a special website, which presented the segmented words one by one. Their task was to listen and look at the waveform depiction of the sound, and decide whether the segmentation was sufficiently good. In a second web interface, they adjusted the boundaries of word segmentations that had been rejected previously. The two interfaces are shown in Figure 3.2, on page 37. I analysed whether the word segmentations produced by POnSS were as reliable as conventional segmentation, and found that they were. I also calculated how long it would take to segment the same set of data conventionally and using POnSS, finding that POnSS was 23% faster than the conventional technique.

In **Chapter 4**, I describe the experiment used to elicit the data that were modelled in Chapter 5. In the experiment, speakers had to name pictures, in Dutch, at three pre-determined speaking rates. The pictures were arranged around a 'clock-face', and a dot jumped clockwise from picture to picture to indicate which picture was to be named when, and thus specify the required speaking rate. The data were segmented with POnSS (Chapter 3), and the analysis technique introduced in Chapter 2 was used to identify the onsets and offsets of syllable-level planning units.

Chapter 5 finally begins to address the main question of the thesis. I introduce two possible hypotheses of how speaking rate might be regulated in the cognitive system responsible for speech planning. One possibility is that speaking fast is essentially the same as speaking slowly, but sped up in a linear fashion. This would mean that control resembles pressing harder on the accelerator pedal. I called this possibility the ‘accelerator hypothesis’. Alternatively, speakers might use qualitatively distinct configurations for different rates. This would resemble the qualitative difference between walking and running gaits that animals with legs adopt to move around: speakers might go into a ‘walk-speaking’ configuration for slow speech, but ‘run-speaking’ configuration for fast speech. I called this possibility the ‘gait hypothesis’.

To test which of these hypotheses was true, I programmed a ‘model’ of speech planning, that extends existing theories. The model aims to imitate the timing of the speech of real speakers as they perform the picture naming task from Chapter 4 at fast, medium and slow rates. The model has parameters (like control knobs). Different settings of these parameters lead to different output. I used machine learning techniques to find values for the parameters that resulted in good imitation of the real speech. This happened by trial and error: the algorithm tried out different settings, evaluated how similar the model’s output was to the data. This was repeated thousands of times to gradually find optimal parameter settings. This was done separately for each rate, so I found different parameter settings that were optimal for slow, medium and fast speaking rates.

To distinguish between the ‘accelerator’ and ‘gait’ hypotheses, I then examined the values of the parameters of the model that were optimal for each rate. I did this by placing them into a multi-dimensional space, where the different parameters of the model were the different dimensions. The accelerator hypothesis predicted that the three speaking rates would be arranged in a straight line through this ‘parameter space’. The gait hypothesis predicted instead that the three rates would be arranged in a triangle. I found a triangular arrangement, consistent with the gait hypothesis.

This result demonstrated that there were different configurations, or gaits, in the cognitive system, but I couldn’t yet say how many gaits there were, or which speaking rates were produced using which gaits. To work that out, I conducted another experiment, described in **Chapter 6**. In this experiment, I used a similar clock-face display with pictures to the one I used in Chapter 4. This time,

speakers were trained to speak at the three predetermined speaking rates before the experiment. Then, I removed the red cueing dot, and the speakers had to maintain the rates themselves. The required speaking rate was indicated by the colour of a frame placed around the picture display. At an unpredictable moment during the trial, the colour of the frame changed, indicating that the speaker should adjust their speaking rate.

I expected that differences would emerge in how quickly it would be possible to switch between different pairs of speaking rates: faster adjustment would indicate that speakers found it easier to switch between the rates involved, whereas slower adjustment would indicate that they found it harder. I found that speakers were quicker to switch between slow and medium rates than they were to switch between fast and medium rates. This reinforced the conclusion of Chapter 5, and additionally allowed me to conclude that there are two gaits: one for slow and medium rates, and a second for the fastest speaking rates.

In general, the results from this thesis show that speaking rate is likely to be controlled by switching between different configurations of the system. Exciting next steps would be to see whether we can see other differences in speech produced using the different gaits, and to explore whether the gait switching mechanism might be generalised to other aspects of speech style.

Curriculum vitae

Joe Rodd (Southampton, United Kingdom, 1991) obtained his bachelor's degree in Linguistics and Spanish from the University of Leeds, United Kingdom, in 2013. This was followed by a master's degree in Linguistics (research) from Utrecht University, in 2015. He was then a research assistant in Utrecht. He began his PhD research at the Centre for Language Studies, Radboud University and Psychology of Language department of the Max Planck Institute for Psycholinguistics in February 2016, which he completed in February 2020, after which he took up a visiting research fellowship at the University of Potsdam, Germany.

Publications

Rodd, J., Bosker, H. R., Ernestus, M., ten Bosch, L., & Meyer, A. S. (under review). Asymmetric switch costs between speaking rates: Evidence for gaits of speech planning

Rodd, J., Decuyper, C., Bosker, H. R., & ten Bosch, L. (in press). A tool for efficient and accurate segmentation of speech data: Announcing POnSS. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-020-01449-6>

Rodd, J., Bosker, H. R., Ernestus, M., Alday, P. M., Meyer, A. S., & ten Bosch, L. (2020). Control of speaking rate is achieved by switching between qualitatively distinct cognitive ‘gaits’: Evidence from simulation. *Psychological Review*, 127(2), 281–304. <https://doi.org/10.1037/rev0000172>

Terband, H., Rodd, J., & Maas, E. (2019). Testing hypotheses about the underlying deficit of apraxia of speech (AOS) through computational neural modeling with the DIVA model. *International Journal of Speech-Language Pathology*. <https://doi.org/10.1080/17549507.2019.1669711>

Rodd, J., Bosker, H. R., ten Bosch, L., Ernestus, M., & Meyer, A. S. (2019a). PiNCeR: A corpus of cued-rate multiple picture naming in Dutch. PsyArXiv. <https://doi.org/10.31234/osf.io/wyc6h>

Rodd, J., Bosker, H. R., ten Bosch, L., & Ernestus, M. (2019b). Deriving the onset and offset times of planning units from acoustic and articulatory measurements. *The Journal of the Acoustical Society of America*, 145(2), EL161–EL167. <https://doi.org/10.1121/1.5089456>

Rodd, J., & Chen, A. (2016). Pitch accents show a Perceptual Magnet Effect: Evidence of internal structure in intonation categories. *Speech Prosody* 2016, 697–701

Smorenburg, L., Rodd, J., & Chen, A. (2015). The effect of explicit training on the prosodic production of L2 sarcasm by Dutch learners of English. *Proceedings of the 18th International Congress of Phonetic Sciences*

Terband, H., Rodd, J., & Maas, E. (2015). Simulations of feedforward and feedback control in apraxia of speech (AOS): Effects of noise masking on vowel production in the DIVA model. *Proceedings of the 18th International Congress of Phonetic Sciences*

Acknowledgements

I don't think I met anyone at all who, on hearing that I was going to have four supervisors, didn't warn me that that might end up in a *right pickle*. I'm not convinced that you thought differently yourselves at the time. Yet this dissertation was finished on time, with four supervisors and a candidate who still all like one another. I think that that shows that we're a very good team, **Hans Rutger, Antje, Mirjam** and **Louis**. Thank you, all of you, for your dedication to the project and unwavering confidence that I was achieving something worthwhile, and for bringing your diverse skills and ideas to train me to be the best scientist you could.

Hans Rutger, thank you for helping me straighten out my half-baked ideas, and for planting seeds (or entire trees) of new ones. Thank you for always being encouraging and positive, even when I was convinced the twenty pages of brightly coloured but poorly labelled graphs I spread over your desk doomed the whole enterprise. **Antje**, thank you for wielding Occam's razor mercilessly when it was called for, and for plugging the logical flaws in my thinking. Thank you for being the right mix of fun and in charge. **Mirjam**, dank je wel dat je aandacht had voor meer dan alleen 'het project'. Of beter gezegd, dank je wel dat je het jouw project maakte om mij goed terecht te laten komen. **Louis**, na een gesprek met jou had ik altijd veel antwoorden. Vaak hoorden ze niet bij de vraag die ik vooraf dacht dat ik wilde stellen, maar dat lag aan mij. Dank je wel dat het altijd alle kanten op kon gaan, want daar wordt een mens een denker van.

Thank you to the manuscript committee, **Paula, Audrey, Hugo, Falk** and **Stefan**, for their speedy but careful reading and evaluation of this thesis. **Audrey**, thank you for giving me the chance to come to Potsdam and think about inspiring questions and take the work in this book in a different direction. Even though Skype was a poor substitute for actually being at Haus 14, and we weren't able to do the data collection we planned, we made the best of the situation.

The Nijmegen Language Bubble was a hugely encouraging and inspiring bubble to be in. Thank you to my colleagues and friends at the CLS Speech Production and Comprehension group: Annika, Aurora, Chen, Ellen, Emily, Esther, Hanno, Katherine, Kimberley, Lieke, Lisa, Lotte, Mark, Martijn, Mirjam, Robert,

Sophie, and Tim. Our group meetings were always friendly and inspiring. Thank you for the insightful discussions about and around research.

I felt pretty instantly at home at PoL. That started with good office mates: **Conny, Jeroen, Eirini, Merel** and **Jieying**. Thank you for many fun times in and out of the office, more than my fair share of coffee and six hour long vrijdagmiddagborrels. Thank you for watering the jungle and for your tolerance of creeping mess. Jeroen, Eirini and Merel, thank you for agreeing to be my paranymphs, despite all of the above. **Ashley, Elliot**, and **Marieke**, thank you for many fun evenings, dinners, afternoons in the sun and, latterly, Zoom-borrels. **Sara** and **Limor**, you belong in this paragraph too. Thank you for great food, terrible-but-great food, inexhaustible gossip, wisdom, nonsense and rapidly depleted wine.

Thank you to the whole PoL crew (Aitor, Alastair, Andrea, Fan, Federica, Florian, Johanne, Markus, Merel, Nina, Renske, Shiri, Sophie, Suzanne, Will, Zeshu) for great lab meetings, unpredictable lunch conversations and fun parties. **Saoradh** thank you for quiet cups of tea at the institute, proper walks and the Ot-tolenghiest dinners in town. **Greta, Miguel**, thank you for making LabPhon in Lisbon so tasty. **Laurel** and **Phillip**, thank you for listening, thinking along, and putting things in perspective. Phillip, thank you for translating what I thought I needed to Google into what I really needed to Google and explaining why, and for explaining Germany better than all the Germans I asked. Thank you for your ideas and support on the *big chapter*. **Caitlin**, thank you for long coffee breaks, fridge smuggling, dinners in Malden and Nijmegen, and for your contribution to Chapter 3. **Amie**, I'm so glad it turned out you didn't have to hate me for coming from Southampton.

A massive thank you to **Annelies** and the student assistants who spent many, many hours segmenting speech data for me: Anne, Carlien, Carlijn, Dennis, Dylan, Esther, Hanna, Inge, Jessica, Milou, Mirte, Moniek, Monique, Nikki, Rosemarije, Sanne, Sjaroes and Zina. You know better than any that 23% faster isn't that much faster... **Carlijn**, you also tested many participants for me for the later chapters. Thank you for carefully and professionally getting on with it, and faultlessly executing difficult experiments.

I am also hugely grateful to the many people who, behind the scenes, helped me with practicalities at the MPI, at CLS, and at the LiI office. **Tobias**, thank you for your patience and flexibility in supporting my esoteric and often half-baked computing and hosting requests, and for your fair and responsive management

of the compute cluster. Thanks also to the rest of TG for making computing and experimenting at the MPI a smooth process, in particular **Johan, Reiner** and **Alex**. Thank you to **Karin** and **Meggie** for magic paper finding and for near-instant answers to other bookish questions.

Being a CLSer at MPI and an MPIer at CLS took some navigating. Thank you to **Evelyn, Thy, Annelies** and the **Operations** team at the MPI, for being on top of the MPI way of doing things, and to **Valerija, Peter, Marc, Nico, Sander, Julia, Carolin** and **Kwan** for keeping an eye on me from the University. More often than not, you were able to help me work out how to ask the right question so I got the answer I wanted. Thank you, **Kevin**, for your support as IMPRS coordinator. **Lisenka**, thank you for your guidance while I was preparing my fellowship applications.

Working on *Kletsoppen* was my favourite distraction at the beginning of my PhD time, thank you in particular **Sharon, Patricia, Caro, Paula**, and **Pim** for giving me creative freedom, and for a great collaboration that led to a very successful series of events.

Aan mijn vrienden in Nederland, maar *niet* in Nijmegen, heb ik ook veel te danken. Jullie zijn een belangrijk tegengewicht geweest. **Cashmyra**, je blijft mijn superheldin met een kater (in beide betekenissen). **Jennifer**, dank je wel voor je nuchtere perspectief op het doen van een PhD. Dank jullie wel voor alle uren in de trein naar Nijmegen, en voor een altijd warme welkom in het verre noorden. **Suzanne** en **Nick**, dank jullie wel voor het in leven houden van onze Utrechtse tijd, exotische alcohol, fantastische feesten, het continuous code-switchen en het exceptional dancing.

Thank you to my friends in the UK, in particular **Susannah, Tom, Lizzy, Flav, Carl, Heather, Duncan, Durand** for believing me that this thesis was going to be worthwhile, but not asking me to explain why too deeply. Thank you to **Emily, Jess, Pippa, Tom** and **Niall** for trekking all the way to ‘Nim-Jim’ to see us, and for hosting Mischa and I at our convenience more than yours.

Thank you to my family. Thanks, **Mum** and **Dad**, for your confidence in me, for your support and encouragement in all the contrary decisions I’ve made to get to this point. I don’t think you thought I’d end up writing a book about psycholinguistics, but I don’t think you thought I wouldn’t either. Thank you **Charlie** and **Maëlle** for making family time fun and a healthy change from working hard. Thank you to my generous and inspiring grandmothers, **Angie**, and **Grandma**.

Thanks to the **Brussels Bakers** for many fun weekends in Brussels, particularly in the period that I felt I wasn't making the progress I wanted to be. "Ge ben ene grote slakken".

Aan mijn Nederlandse familie heb ik ook veel te danken. Bedankt, **Anne-Marie, Jesse, Maartje, Willy** en **Myra** voor jullie warme welkom in de/het familie, voor jullie interesse in mijn vooruitgang met deze/dit project, en voor de/het heel vaak beschikbaar stellen van je auto. Ik ben heel blij dat ik erbij mag horen.

Het laatste bedankje gaat naar jou, **Mischa**. Wellicht had ik dit boek ooit af gekregen zonder jou, maar jouw liefde, steun, geduld en relativerend vermogen maakte me een gelukkige mens tijdens deze PhD-jaren. Je hebt me veel breder leren kijken naar de wereld en veel dieper leren kijken naar hoe het echt zit. Dank je.

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda I. van Turenhout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed class distinction in spoken-word recognition. *Alette Petra Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk J. Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie H. van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Christine Weber*

17. Moving eyes and naming objects. *Femke Frederike van der Meulen*
18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja Helena de Jong*
21. Fixed expressions and the production of idioms. *Simone Annegret Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Danielle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra Martine van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Henriëtte Wilhelmina Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl Jenny Judith Karin Kemps*
29. At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marijtje Elizabeth Debora Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Elisabeth Broersma*
35. Retrieving semantic and syntactic word properties: ERP studies on the time course in language comprehension. *Oliver Müller*

36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*
37. Sensitivity to detailed acoustic information in word recognition. *Keren Batya Shatzman*
38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*
39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*
40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich*
41. Cognitive cladistics and the relativity of spatial cognition. *Daniel Haun*
42. The acquisition of auditory categories. *Martijn Bastiaan Goudbeek*
43. Affix reduction in spoken Dutch: Probabilistic effects in production and perception. *Mark Pluymaekers*
44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Madalla Kooijman*
45. Space and iconicity in German sign language (DGS). *Pamela M. Perniss*
46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*
47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*
48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*
49. Phoneme inventories and patterns of speech sound perception. *Anita Eva Wagner*
50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*
51. A grammar of Savosavo: A Papuan language of the Solomon Islands. *Claudia Ursula Wegener*
52. Prosodic structure in speech production and perception. *Claudia Kuzla*
53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*
54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*

55. Processing the fine temporal structure of spoken words. *Eva Reinisch*
56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*
57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*
58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*
59. Lexical interactions in non-native speech comprehension: Evidence from electroencephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*
60. Processing casual speech in native and non-native language. *Annelie Tuinman*
61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Payton Robinson*
62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*
63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*
64. The meaning and use of ideophones in Siwu. *Mark Dingemans*
65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*
66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*
67. The relevance of early word recognition: Insights from the infant brain. *Caroline Mary Magteld Junge*
68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias Johannes Sjerps*
69. Structuring language: Contributions to the neurocognition of syntax. *Katrien Rachel Segaert*
70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*
71. Gaze behavior in face-to-face interaction. *Federico Rossano*
72. Sign-spatiality in Kata Kolok: How a village sign language of Bali inscribes its signing space. *Connie de Vos*

73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*
74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*
75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*
76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*
77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*
78. Contributions of executive control to individual differences in word production. *Zeshu Shao*
79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*
80. High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*
81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*
82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*
83. The many ways listeners adapt to reductions in casual speech. *Katja Pöllmann*
84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*
85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*
86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*
87. The acquisition of morphophonological alternations across languages. *Helen Buckler*
88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*
89. Rediscovering a forgotten language. *Jiyoun Choi*

90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*
91. Infants' understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*
92. Information structure in Avatime. *Saskia van Putten*
93. Switch reference in Whitesands. *Jeremy Hammond*
94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*
95. Acquisition of spatial language by signing and speaking children: A comparison of Turkish sign language (TID) and Turkish. *Beyza Sumer*
96. An ear for pitch: On the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*
97. Incrementality and Flexibility in Sentence Production. *Maartje van de Velde*
98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*
99. The request system in Italian interaction. *Giovanni Rossi*
100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*
101. Assessing birth language memory in young adoptees. *Wencui Zhou*
102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*
103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*
104. Conversation electrified: The electrophysiology of spoken speech act recognition. *Rósa Signý Gísladóttir*
105. Modelling multimodal language processing. *Alastair Charles Smith*
106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*
107. Situational variation in non-native communication. *Huib Kouwenhoven*
108. Sustained attention in language production. *Suzanne Jongman*

109. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntactic, and orthographic effects. *Malte Viebahn*
110. Nateness, dominance, and the flexibility of listening to spoken language. *Laurence Bruggeman*
111. Semantic specificity of perception verbs in Maniq. *Ewelina Wnuk*
112. On the identification of FOXP2 gene enhancers and their role in brain development. *Martin Becker*
113. Events in language and thought: The case of serial verb constructions in Avatime. *Rebecca Defina*
114. Deciphering common and rare genetic effects on reading ability. *Amaia Carrión Castillo*
115. Music and language comprehension in the brain. *Richard Kunert*
116. Comprehending Comprehension: Insights from neuronal oscillations on the neuronal basis of language. *Nietzsche H.L. Lam*
117. The biology of variation in anatomical brain asymmetries. *Tulio Guadalupe*
118. Language processing in a conversation context. *Lotte Schoot*
119. Achieving mutual understanding in Argentine Sign Language. *Elizabeth Manrique*
120. Talking sense: the behavioural and neural correlates of sound symbolism. *Gwilym Lockwood*
121. Getting under your skin: The role of perspective and simulation of experience in narrative comprehension. *Franziska Hartung*
122. Sensorimotor experience in speech perception. *Will Schuerman*
123. Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond. *Ashley Lewis*
124. Influences on the magnitude of syntactic priming. *Evelien Heyselaar*
125. Lapse organization in interaction. *Elliott Hoey*
126. The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech. *Sophie Brand*
127. The neighbors will tell you what to expect: effects of aging and predictability on language processing. *Cornelia Moers*

128. The role of voice and word order in incremental sentence processing. Studies on sentence production and comprehension in Tagalog and German. *Sebastian Sauppe*
129. Learning from the (un)expected: age and individual differences in statistical learning and perceptual learning in speech. *Thordis Neger*
130. Mental representations of Dutch regular morphologically complex neologisms. *Laura de Vaan*
131. Speech production, perception, and input of simultaneous bilingual preschoolers: Evidence from voice onset time. *Antje Stoehr*
132. A holistic approach to understanding pre-history. *Vishnupriya Kolipakam*
133. Characterization of transcription factors in monogenic disorders of speech and language. *Sara Busquets Estruch*
134. Indirect request comprehension in different contexts. *Johanne Tromp*
135. Envisioning language: An exploration of perceptual processes in language comprehension. *Markus Ostarek*
136. Listening for the WHAT and the HOW: Older adults' processing of semantic and affective information in speech. *Juliane Kirsch*
137. Let the agents do the talking: On the influence of vocal tract anatomy on speech during ontogeny and glossogeny. *Rick Janssen*
138. Age and hearing loss effects on speech processing. *Xaver Koch*
139. Vocabulary knowledge and learning: Individual differences in adult native speakers. *Nina Mainz*
140. The face in face-to-face communication: Signals of understanding and non-understanding. *Paul Hömke*
141. Person reference and interaction in Umpila/Kuuku Ya'u narrative. *Clair Hill*
142. Beyond the language given: The neurobiological infrastructure for pragmatic inferencing. *Jana Bašnáková*
143. From Kawapanan to Shawi: Topics in language variation and change. *Luis Miguel Rojas Berscia*
144. On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions. *Linda Drijvers*

145. Linguistic dual-tasking: Understanding temporal overlap between production and comprehension. *Amie Fairs*
146. The role of exemplars in speech comprehension. *Annika Nijveld*
147. A network of interacting proteins disrupted in language-related disorders. *Elliot Sollis*
148. Fast speech can sound slow: Effects of contextual speech rate on word recognition. *Merel Maslowski*
149. Reasons for every-day activities. *Julija Baranova*
150. Speech planning in dialogue - Psycholinguistic studies of the timing of turn taking. *Mathias Barthel*
151. The role of neural feedback in language unification: How awareness affects combinatorial processing. *Valeria Mongelli*
152. Exploring social biases in language processing. *Sara Iacozza*
153. Vocal learning in the pale spear-nosed bat, *Phyllostomus discolor*. *Ella Lattenkamp*
154. Effect of language contact on speech and gesture: The case of Turkish-Dutch bilinguals in the Netherlands. *Elif Zeynep Azar*
155. Language and society: How social pressures shape grammatical structure *Limor Raviv*
156. The moment in between: Planning speech while listening. *Svetlana-Lito Gerakaki*
157. How speaking fast is like running: Modelling control of speaking rate. *Joe Rodd*