Testing the effectiveness of mouse tracking in speech perception

Bachelorarbeit

an der Philosophischen Fakultät der Universität zu Köln
im Fach Linguistik und Phonetik

vorgelegt von

Mathias Stoeber

Osnabrück, 26. August 2019

Prüferin: Prof. Dr. Martine Grice

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich diese Bachelorarbeit selbstständig verfasst und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die Stellen meiner Arbeit, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche unter Angabe der Quelle kenntlich gemacht. Diese Arbeit habe ich in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Ort, Datum	Unterschrift

Contents

List of Figures vii

List of Tables ix

1. Introduction 1
- 1.1. Motivation and structure . . . 1
- 1.2. Discreteness and continuity . . . 2
- 1.3. Towards on-line measures . . . 4

2. The mouse tracking paradigm 11
- 2.1. Introduction . . . 11
- 2.2. Types of measures . . . 13
- 2.3. Trajectory distributions . . . 17
- 2.4. Design decisions . . . 21

3. Implementation 29
- 3.1. Expectations . . . 29
- 3.2. Experiment 1 . . . 30
- 3.3. Experiment 2 . . . 41
- 3.4. Between-experiment comparison . . . 49

4. General discussion 55
- 4.1. Revisiting McMurray et al. (2008) . . . 56
- 4.2. Methodological concerns . . . 58
- 4.3. Conclusion . . . 64

A. Supplementary figures 71

B. Supplementary tables 77

Bibliography 83

List of Figures

1.1. Evidence from reaction times: Pisoni & Tash (1974) 5
1.2. Evidence from eye tracking: McMurray et al. (2008) 8

2.1. Stylised depiction of trajectory curvature 15
2.2. Trajectory prototypes 20

3.1. The main experimental display 31
3.2. Stimulus spectrograms 32
3.3. Experiment 1: Across-subject identification 34
3.4. Experiment 1: Across-subject reaction times 36
3.5. Experiment 1: Across-subject MAD 38
3.6. Experiment 1: Proportions of trajectory types 38
3.7. Experiment 2: Across-subject identification 42
3.8. Experiment 2: Across-subject reaction times 44
3.9. Experiment 2: Across-subject MAD 46
3.10. Experiment 2: Proportions of trajectory types 46
3.11. Velocity profiles for both experiments 49
3.12. Reaction times per stimulus VOT and experiment 50
3.13. MAD per stimulus VOT and experiment 52
3.14. Proportions of trajectory types per experiment 53

A.1. Examples of well-formed and malformed mouse trajectories. 72
A.2. Experiment 1: Trajectory heatmap 73
A.3. Experiment 1: Trajectory types per stimulus VOT 73
A.4. Experiment 2: Trajectory heatmap 74
A.5. Experiment 2: Trajectory types per stimulus VOT 74
A.6. Experiment 1: MAD per relative VOT 75
A.7. Experiment 2: MAD per relative VOT 76

List of Tables

3.1. Experiment 1: Across-subject /pa/-probability 35
3.2. Experiment 1: Across-subject reaction times 37
3.3. Experiment 1: Across-subject MAD 39
3.4. Experiment 1: Trajectory types per stimulus VOT 40
3.5. Experiment 2: Across-subject /pa/-probability 43
3.6. Experiment 2: Across-subject reaction times 45
3.7. Experiment 2: Across-subject MAD 47
3.8. Experiment 2: Trajectory types per stimulus VOT 48

B.1. Stimulus & VOT durations 78
B.2. Experiment 1: Estimated category boundaries 78
B.3. Experiment 2: Estimated category boundaries 79
B.4. Reaction times per experiment 79
B.5. Maximum Absolute Deviations per experiment 80
B.6. Proportions of trajectory types per stimulus and experiment 81

1. Introduction

1.1. Motivation and structure

Research on speech perception has yet to reach an overarching consensus on how listeners manage to take in a multitude of properties of the phonetic signal which are manifested along its continuous physical dimensions, and then discretely map them onto phonological units like phonemes. Investigating this question, traditional accounts have proposed that perceptual processes do not make systematic use of, but instead discard all non-distinctive variability in the speech signal in favour of discrete phonological representations, resulting in a phenomenon known as categorical perception (Liberman et al., 1957). Opposed to this variance-reduction view on speech perception, however, more recent work has been successful in procuring evidence for listeners' sensitivity to within-category phonetic variation, demonstrating that solely measuring the discrete outcome of a decision process does not lead to a sufficient understanding of the process itself. Studies simultaneously looking at, for example, reaction times (Pisoni & Tash, 1974) or oculomotor behaviour (McMurray et al., 2008) in addition to the categorisation outcome have been able to show that listeners are sensitive to sub-phonemic variation in the speech signal while still producing categorisation results which are entirely in line with the framework of categorical perception.

Informing the main thrust of the present thesis are the innovative experimental methods with which these types of studies were capable of "unpacking" low-dimensional data from participants' motor behaviour into rich, multi-dimensional data, thereby helping to pry open the window into the dynamics of the cognitive processes connected to that behaviour. In focusing on mouse tracking, a relatively recent experimental technique, I report on an implementation of this method as a tool for research into the perception of speech by way of two exploratory mouse tracking experiments.

To that end, the first chapter offers a brief overview of the phenomenon of categorical perception of speech, as well as of findings that suggest the existence of within-category sensitivity to the acoustic details of speech sounds. Further, it discusses the distinction

between such paradigms which aim (and are only able) to collect discrete, outcome-based response measures, and those which are additionally capable of recording rich, meaningful data over the entire time course of response behaviour, pointing to the on-line tracking of eye movements as a prominent example.

The second chapter then introduces the mouse tracking method as another representative of such on-line paradigms, with its own idiosyncratic advantages and drawbacks. It provides a look at the multitude of different measures the paradigm allows to be collected, and explores two different theoretical points of view which can be assumed when analysing multi-dimensional data from mouse movements. Lastly, it acknowledges the status of mouse tracking as a nascent methodology for which a well-tested set of best practices has yet to be developed by discussing recent evidence for the role that specific design decisions can play.

Subsequently, chapter three describes the design of the two mouse tracking experiments mentioned above, and reports on the data gathered during their implementation by strictly descriptive assessments of each experiment's results. An equally descriptive comparison of the (purely methodological) between-experiment manipulation closes this chapter.

The concluding discussion in chapter four then attempts to integrate these experimental outcomes, previous findings, and known conceptual and analytic issues and limitations of the mouse tracking method itself. In doing so, it further endeavours to establish a way forward for future research on these matters.

1.2. Discreteness and continuity

The acoustic speech signal manifests itself along multiple, decidedly continuous physical dimensions like duration, frequency, and intensity. This lack of discrete steps or invariance within the signal exists in stark contrast to the effortlessness and reliability with which listeners are nevertheless able to map the complex combination of those continuous acoustic properties onto discrete linguistic units like phonemes or words. Indeed, categorical linguistic behaviour can readily be observed, which in turn might suggest that cognitive representations of speech are discrete. For example, upon hearing one half of a minimal pair like "beach" and "peach", listeners will hardly ever be under the impression of having heard anything in between these two words.

Past attempts at finding any invariant acoustic cues to phonetic features, which

would hold considerable explanatory power with regards to the inner workings of the underlying cognitive processes involved in the perception of speech, have met with little success (Lindblom, 1996; Holt & Lotto, 2010). Thus, discovering the "so-called invariant cues [that] represent nuggets of discreteness buried in this continuous signal" (McMurray et al., 2008) has become a notorious puzzle for researchers studying speech perception. At the base of this search for phonetic invariance lies the traditional assumption (Liberman et al., 1957) that only a minor portion of the continuously variable detail found in the speech signal "survives" the perceptual processes and thus serves in systematically distinguishing discrete phonological representations, whereas the bulk of the signal, namely all non-distinctive variability, gets discarded as essentially uninformative noise. This variance-reduction type of process has become known as categorical perception.

A classic example for this phenomenon in the realm of speech, suggesting a categorical (phonological) representation of the graded (phonetic) signal, can be found by turning toward the categorical perception of the voicing contrast in stop consonants (e.g. Liberman et al., 1961; Repp, 1984).

Voice onset time Obstruent "voicing" is described as a means to contrast lexical meaning in numerous languages. Voicing as a phonological feature manifests itself phonetically in the form of several continuous acoustic parameters (Lisker, 1986; Toscano & McMurray, 2016). Examples from this group of cues to stop voicing are the duration of the stop closure, the duration of the preceding vowel, or the onset frequency of the first formant. One of the more prominently discussed cues has been voice onset time (Cho & Ladefoged, 1999), which is traditionally defined as the time interval between the release of stop closure and the beginning of glottal pulses belonging to the following vowel (Jessen, 1999). For example, German and English distinguish words such as "beach" and "peach" mainly by differing stop voice onset time (henceforth VOT).

One of the reasons why VOT has been such an enduring object of investigation (e.g. Lisker & Abramson, 1964; Cho & Ladefoged, 1999) is the way listeners seem to perceive speech sounds in which VOT is systematically manipulated: They experience very little difficulty in identifying a stimulus as either a voiced or a voiceless stop when presented with syllables containing gradually differing voice onset times in an otherwise invariant phonetic context. For example, when stimulus VOT is continuously manipulated between prototypical /ba/ and /pa/ sounds, listeners robustly identify

tokens along this continuum as belonging to two distinct categories, with a clearly defined category boundary in between (e.g. Repp, 1984).

Intra-categorical sensitivity Subsequent research into the phenomena of categorical perception, however, has cast doubt on the variance-reduction approach to speech perception, in which listeners exhibit no sensitivity to within-category phonetic detail (for a short overview, see McMurray et al., 2008). For example, Pisoni & Tash (1974) demonstrated that in an identification task, listeners responded faster to prototypical exemplars, whereas tokens near the category boundary produced longer reaction times. Figure 1.1 summarises their results. Evidence of equally graded responses to within-category variation comes from a body of work looking at both selective adaptation effects (Miller et al., 1983; Samuel, 1982) and goodness ratings (Allen & Miller, 2001; Massaro & Cohen, 1983). In these studies, the highest ratings or largest adaptation effects occur for prototypical stimulus exemplars at the ends of a manipulation scale, while stimuli near the category boundary receive the lowest ratings or exhibit the smallest adaptation effects, with a monotonically falling gradient of responses in between. Together, these findings suggest both a gradient nature of phonological representations and that "these graded categories interact with continuous auditory cues during perception" (McMurray et al., 2008), meaning that listeners are, in fact, sensitive to acoustic detail below the category level, i.e. they exhibit sensitivity to sub-phonemic variation.

1.3. Towards on-line measures

In this light, it appears that both assertions are true—on the one hand, listeners do exhibit discrete behaviour when identifying speech stimuli as members of categories, (e.g. they categorise a given phonetic form as either voiced or voiceless)[1]. On the other hand, they nevertheless show gradient sensitivity to intra-categorical information. Prior to any theoretical and/or empirical attempt at deciding which one of these assertions (if any) has to be disregarded as false, one methodological observation stands out as particularly important for such an endeavour: Solely measuring the (discrete) outcome of a decision process is not enough to understand the processes that lead to the decision.

An experimental setup, for example, that only allows for discrete, binary "A or B"

[1] At least they do so in forced-choice-contexts, but see Schouten et al. (2003)

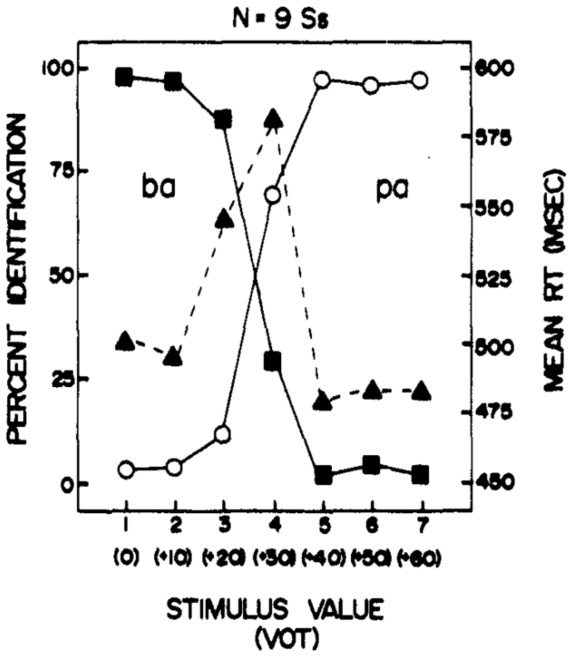


Figure 1.1.: Stimulus identification in percent on the left-hand y-axis, stimulus VOT on the x-axis, and RTs on the right-hand y-axis. Squares used for /ba/-stimuli, circles for /pa/-stimuli, and triangles for RTs of both stimuli. Figure reprinted with permission from Pisoni & Tash (1974).

type participant responses (e.g. by way of button-box keypresses) will at the very least encourage researchers to interpret its resulting data as evidence for matching discrete cognitive categories. In order to circumvent this kind of artificially-imposed, pre-determining, and potentially distorting curtailment of data collection, a necessary first measure has to be the transformation of the hitherto discrete response space into a continuous one which allows for equally continuous responses. An example for this is the aforementioned strategy to employ goodness ratings in addition to categorical response options (or replacing them altogether). Note that the inverse of the problem of artificially restricted response data is not true here: Given the opportunity to exhibit graded responses, participants are still entirely free to behave in an exclusively discrete fashion.

While "goodness" and similar types of ratings do begin to open up the response space to continuity, they still share a substantial drawback with purely discrete response measures: Both are outcome-based measures, i.e. they are derived from responses a participant enters into the record *after* she has made her (task-dependent) decision, and

thus after the cognitive processes that led to her decision have run their course (Spivey, 2007). Once more, a remedy for this disadvantage can be found in the manipulation of the response or measurement dimensions. Where a first step elevated the response space from its previous zero-dimensionality to having one (more or less) continuous dimension[2], a logical next step in the same direction is the addition of the continuous temporal dimension to the data being recorded. As described above (Pisoni & Tash, 1974), this measuring of reaction times can provide a first glimpse of the dynamic nature of the processes that precede and inform a behavioural outcome. Still, this method will only be able to offer a temporally external, *after-the-fact* perspective on those processes, recorded from a vantage point which is very similar to the one of entirely outcome-based or off-line measures.

In order to arrive at truly continuous and simultaneously on-line measures that are capable of tracing the processes leading up to and accompanying response behaviour as they are unfolding, a further increase in the number of data dimensions appears indispensable. Indeed, innovative experimental paradigms have been able to make use of this methodological approach by recording, for example, motor behaviour from limb, head, or eye movements whilst the processes of response selection are ongoing. Practices in this vein can naturally supplement the resulting data with continuous information from up to three spatial (movement) dimensions plus time. It stands to reason that this process of adding measurement dimensions to certain experimental tasks and designs has been instrumental in unpacking previous findings as well as in reaching novel insights into cognitive processes like the ones involved in speech perception.

Eye tracking A widely-used representative of the kind of paradigm able to elicit continuous, on-line measurements from participants' response behaviour is the eye tracking method (see Spivey (2007) for an overview). Its primary utility lies in recording the saccadic movements of participants' eyes as they perform experimental tasks, capturing where they look (fixation place), how long their gaze remains at certain points (fixation duration), and how often they look at certain points relative to others (fixation proportions). By the same token, it allows for the calculation of the probability with which participants look to certain parts of a visual display given a particular stimulus or stimuli. Critically for the assessment of underlying cognitive processes, this means that "eye movement data provide a [...] record of regions of the display that are briefly considered relevant for carrying out whatever experimental task is at hand"

[2]Like, for example, a Likert scale

(Spivey, 2007). It is, of course, entirely possible (and in many cases integral to the core experimental design) to collect other (off-line) data simultaneously within the same setup, e.g. by recording mouse clicks indicating discrete identification decisions.

Gradient sensitivity to VOT A series of experiments from a 2008 study by McMurray et al. utilised the eye tracking method to assess the apparent conflict between the more traditional view of speech perception as categorical perception and the gradiency hypothesis described above by investigating listener sensitivity to VOT manipulations. In their experiments, McMurray et al. equipped their subjects with a head-mounted eye tracking device capable of recording their gaze positions and presented them with a display containing two boxes, either labelled "b" or "p". Subjects were then auditorily exposed to simple consonant-vowel syllables (either /ba/ or /pa/, one syllable per trial). These CV-syllables came from an artificially created VOT continuum. The experimental task consisted of choosing the response box that corresponded with the stimulus heard by clicking on it with the computer mouse.

During this response selection process, McMurray et al. measured the proportion of gaze fixations to the target response box and to the competitor response box, respectively. For example, if a subject indicated having heard a /ba/ syllable, there would still be a number of gaze fixations made to the competing response option /pa/ during the trial. This vacillation between the available response alternatives was taken to be an indication of dynamical competition of partial activations during the response selection process (Spivey, 2007; Magnuson, 2005). Crucially, fixations to the competitor response occurred gradually more frequently as the stimuli were approaching the category boundary, as shown in figure 1.2: Participants looked to the competitor more often when the phonetic stimulus was close to the category boundary, i.e., stimuli that were more ambiguous between /ba/ and /pa/[3]. Thus, they concluded that tracking oculomotor behaviour was capable of unravelling the perceptual processes underlying phoneme identification, and of showing that voicing identification involves graded decision mechanisms.

Limitations Unfortunately, the method of tracking eye movements suffers from some significant drawbacks of its own. On the practical side, the required eye tracking hardware tends to be rather costly (with prices for single tracking devices regularly

[3]This observation held true for some specific portions of the five experiments McMurray et al. conducted, but did not do so for others, see chapter four for a discussion.

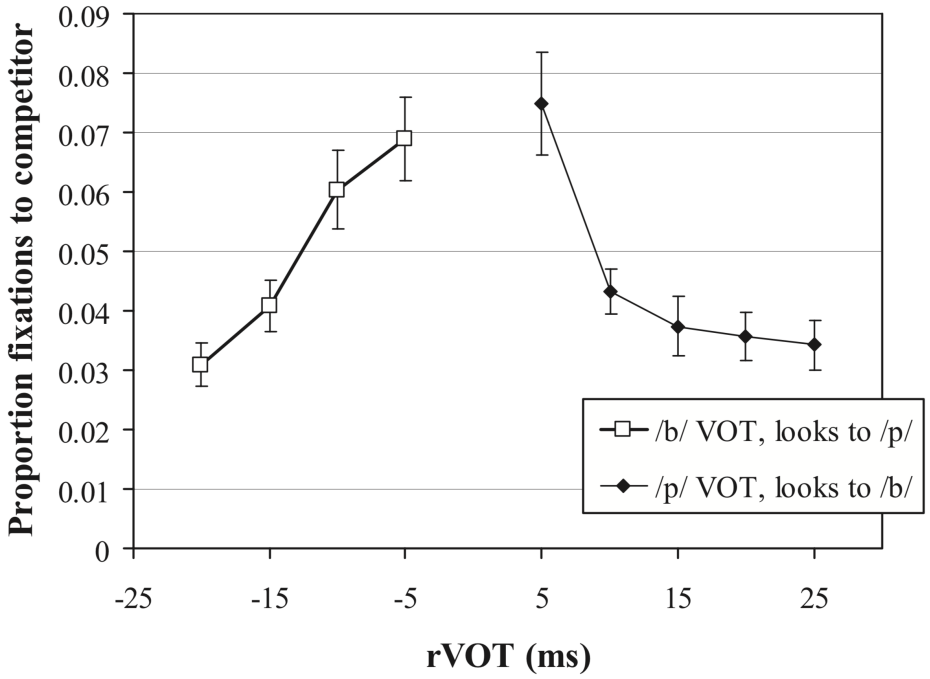


Figure 1.2.: Gaze fixation proportions on the y-axis, and voice onset time (standardised per subject and termed rVOT) on the x-axis. Figure reprinted with permission from McMurray et al. (2008).

in excess of 10.000EUR). Additionally, these devices are not particularly comfortable to wear (in the case of head-mounted trackers), and often, they require subjects to restrict their bodily and/or head movements in order to retain comparability between trials (in the case of table- or monitor-mounted trackers), diminishing the naturalness of the experimental situation from the participants' perspective.

On the conceptual side, proponents of the method have to concede that by recording the saccades of eye movements, which are of a ballistic nature and happen only around three to four times per second (Magnuson, 2005), the desired continuous data cannot truly be obtained. Instead, analyses of eye tracking data looking at this "semicontinuous record of eye position, alternating between steady fixations of 300–400 ms and fast, ballistic saccades of 20–40 ms" (Spivey et al., 2005) must necessarily involve "averaging 'categorical' data (steady fixations of one object or another over time) to produce 'continuous' functions. Thus, [they] can only approximate continuous central tendencies of group data" (ibid.). There is, however, at least one alternative process tracing method capable of transcending the problems described here, and thereby of delivering authentically continuous, on-line response data in an affordable manner: the tracking

of hand or arm movements by way of an ordinary computer mouse.

2. The mouse tracking paradigm

2.1. Introduction

The mouse tracking paradigm affords a way of observing continuous details from the time course of participants' response selection processes which is comparable to the eye tracking method described in the previous chapter. Instead of obtaining data from gaze fixations to competing response options, the hand-mouse movement trajectory towards the response location is taken as the origin of most dependent variables. Taking after a seminal study (Spivey et al., 2005), numerous investigations from different fields have already proven that fine grained aspects of cognitive processes can be explored by use of this paradigm. Areas of interest have covered a range of fields, including social categorisation, decision making, and impulse control (for two recent reviews, see Freeman (2018) and Stillman et al. (2018)), as well as questions concerning linguistic processing, for example regarding spoken word recognition (Spivey et al., 2005), syntactic parsing (Farmer et al., 2007), and pragmatic inferences (Roettger & Stoeber, 2017; Roettger & Franke, 2019).

The experimental design that has been used most frequently in these types of studies (Freeman, 2018) instructs participants looking at a computer monitor to use a computer mouse in order to click a start button located at the mid-point of the lower display edge, which (more or less directly or indirectly) triggers the current trial's stimulus to be presented (graphically, orthographically, auditorily, or in a combination thereof). The main experimental task, then, is to move the mouse cursor to one of two rectangular response buttons located in the upper left and right corners of the display, each of which contain visual information on the kind of response they represent (e.g. letters, words, or graphics).[1] During the entirety of the resulting mouse movements, the stream of cursor coordinates (and, by extension, the details of the actual motions of the hand and arm), are recorded along two spatial dimensions plus time (x-coordinates, y-coordinates,

[1] See figure 3.1 for an example

timestamps). Usually, the experimental setup and task are made to work together in a manner designed to ensure the simultaneous occurrence of the hand movement and the time period in which the cognitive processes informing response selection are thought to take place (Freeman et al., 2011).

Advantages and limitations The mouse tracking method thus seems fit to enable the recording of on-line measurements covering the time course of decision processes leading up to response outcomes. In its ability to deliver continuous measurements, it is clearly superior to eye tracking, as a standard mouse tracking setup employing consumer-grade hardware can exceed a sampling rate of one hundred data points per second (100Hz; cf. eye tracking: 3-4Hz), thereby "proving to be a temporally fine-grained measure by which participants' tentative commitments to various choice alternatives can be tracked continuously over hundreds of milliseconds" (Freeman, 2018). Additionally (and especially when seen as an alternative to eye tracking),

- the implementation of the paradigm is simple and inexpensive, as computer mice are ubiquitous, very low-cost office tools
- participants are generally already quite familiar with mice and use them intuitively
- software facilitating the collection and the analysis of multi-dimensional mouse tracking data is available in multiple packages, some of which are free and open source (Kieslich & Henninger, 2017; Mathôt et al., 2011); cf. Freeman & Ambady (2010)

However, as mouse tracking is too recent a method to have been exhaustively evaluated and validated, it may seem prudent to regard inferential findings from studies employing the method with an appropriate measure of reservation. A consensus on best practices has not yet formed, and the endeavour to lay important methodological groundwork has only begun even more recently (Fischer & Hartmann, 2014; Hehman et al., 2015; Kieslich et al., 2019a).

Moreover, the movements of the eyes and hands differ in significant ways, as "mouse movements are largely under conscious control, whereas people are unaware of saccades unless they explicitly monitor them" (Magnuson, 2005), an observation which potentially diminishes the immediacy of the hypothesised link between cognitive processes and motor behaviour for hand tracking data (Song & Nakayama, 2009; Cisek & Kalaska, 2005).

Lastly, there is another methodological area of interest relevant to the validity of mouse tracking data as evidence for details of cognitive processes: the technological signal chain between the mouse and the components of the computer running the experimental software. These components include hard- and software, such as the central processing unit, the display providing graphical feedback to the participant, the operating system, or drivers for human interface devices such as the mouse. Any stage of this chain is in principle able to introduce idiosyncratic delays when transferring the digitised motion signal. At the very least with regards to studies which aim to look at the temporal (micro-)structure of response dynamics, this means that there can be combinations of hard- and software components that may inadvertently introduce unexpected (and, as a possible consequence, unnoticed) bias into the collected data. There is, however, some evidence for the existence of combinations that make the computer mouse "reliable enough to be considered as an acquisition device for the analysis of human movement velocity signals" (O'Reilly & Plamondon, 2011).

2.2. Types of measures

Central to the utility of the mouse tracking paradigm is the assumption that hand/cursor movements can be interpreted as indicators of the grade of response option activation, and of how the respective activation levels change over the time course of response selection. Thus, the paradigm produces two primary varieties of evidence (Stillman et al., 2018). It can provide insight into

1. the magnitude of the conflict or competition between response options
2. the unfolding of this competition across time, up to and including its resolution

The mouse tracking literature disagrees, however, about the type of theoretical model most capable of accurately explaining hand trajectories, with the principal contenders being stage-based (or dual-systems), and dynamical systems approaches (see Magnuson (2005), Spivey (2007), and Hehman et al. (2015)). The prediction of stage-based models would see a split across the mouse trajectories of an experiment, with one sub-group exhibiting no strong indication for competing response activations (meaning relatively straight trajectory lines), and another sub-group exhibiting a strong, early commitment to one response option (by system one), followed by a corrective override (by system two), resulting in more angular trajectories containing

prominent directional shifts. Dynamical systems models, on the other hand, expect that the temporally extended conflict of simultaneous activations vying for selection gets resolved gradually before eventually collapsing onto the chosen option, a process that should result in graded trajectory curvature across all trials (Kieslich et al., 2019a). The following section will present an overview[2] of mouse tracking measures, and thus show how the aforementioned two primary categories of measures can be subdivided further.

Curvature-based measures The most commonly used measures in the mouse tracking literature (Kieslich et al., 2019a; Stillman et al., 2018) are based on trajectory curvature, which is thought to be indicative of the relative attraction to the target response option and the distractor response option, respectively, and thereby capable of quantifying the amount of conflict or competition between these options. This means, for instance, that the more excursion towards the distractor a trajectory exhibits, the greater the competition between the response options is taken to be. Hence, curvature-based measures fall into the first primary category of measures; they assess the magnitude of the competition that is present.

Although trajectories can in principle—depending on the goals of the analysis—be subdivided into temporally separate regions, each possessing its own curvature, the most common practice is to calculate curvature-based measures on complete, single trajectories. While this analytic approach has in many cases proven fruitful for comparing differences in response conflict, it should be noted that it aggregates and condenses each two-dimensional, richly informative trajectory shape into one number, while also disregarding its temporal components. In this process, relevant information contained in the overall motion signal may potentially be missed (Kieslich et al., 2019b).

Three variants of the most commonly used measures of trajectory curvature are each based on the amount of movement deviation from an idealised straight line connecting the starting point of the trajectory to its end point (i.e. the chosen response option):[3]

- The *Maximum (Absolute) Deviation* (MAD or MD) denotes "the length of a perpendicular line between the idealized straight-line trajectory and farthest point from that straight line in the observed trajectory" (Hehman et al., 2015).

[2]This overview is not intended to be exhaustive, but nevertheless provides descriptions of a majority of the measures that are currently used

[3]While these measures differ slightly from each other conceptually, they are highly correlated in practice (Stillman et al., 2018)

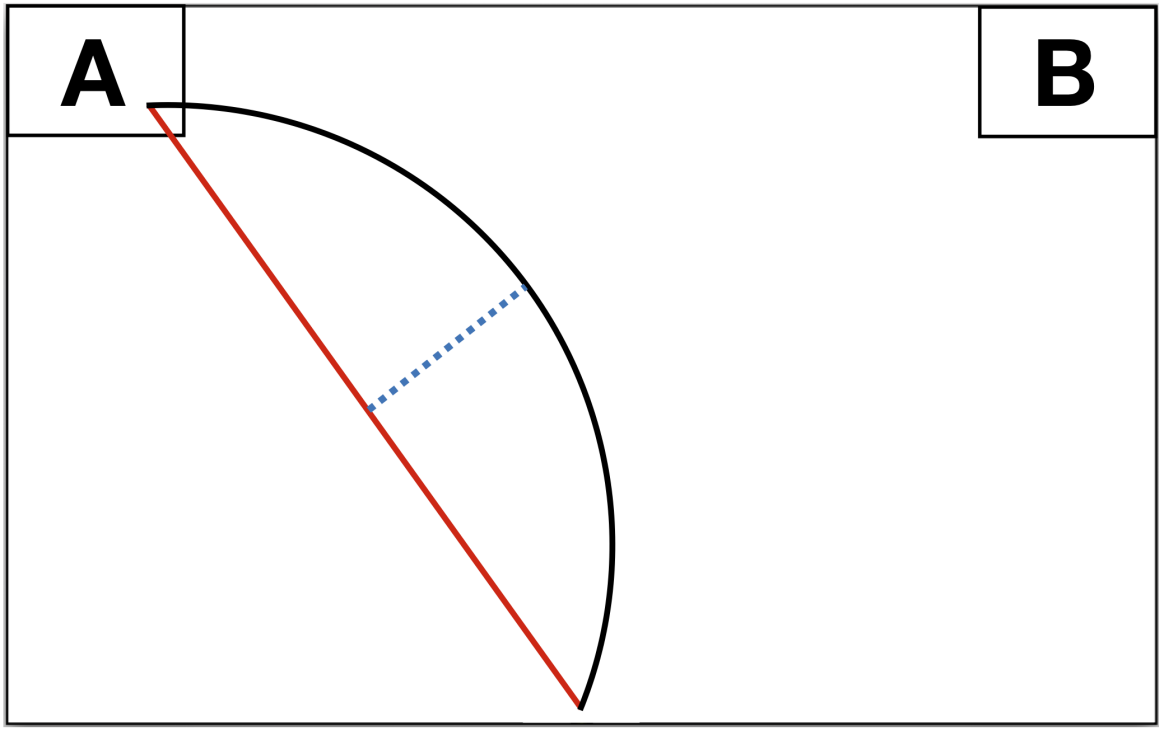


Figure 2.1.: Stylised depiction of trajectory curvature. The black line represents a curved trajectory deviating from the idealised straight line (in red). The dotted blue line represents the Maximum Absolute Deviation. Response options "A" and "B" as rectangular buttons in the upper corners of the screen

- The *Average Deviation* (AD) is equal to the mean distance between the idealised straight line and the actual trajectory across its length.
- The *Area Under the Curve* (AUC) represents the geometric area between the idealised straight line and the actual trajectory across its length.

Another measure of trajectory deviation has been termed *x-neg*, which describes the longest line perpendicular to the y-axis (at the middle of the screen) that connects the y-axis to the observed trajectory, i.e. it measures how far a mouse movement veered into the target's or the competitor's hemi-space, respectively.

Measures of temporal evolution In contrast to deviation-based analyses, the following measures inspect time-dependent spatial features along the course of a trajectory as well as spatiotemporal ratios. They therefore belong to the second primary category of measures, assessing the evolution of conflict or competition across time.

The first group of this type of measure is concerned with determining points in time at which certain events take place within a mouse trajectory. They report

- the point in time at which MAD is reached
- the point in time at which movement reaches its maximum within-trial velocity value
- the point in time at which movement reaches its maximum within-trial acceleration value
- the *Turn Towards Target* (TTT), defined as "the latest point in time at which the trajectory did not head towards the target" (Roettger & Franke, 2019)

These measures can be of particular interest for speech research, as they are not only able to provide information about the unfolding of a decision influenced by competing response options, but are also capable of taking the temporal structure of the experimental stimuli into account, which is potentially relevant for any design employing auditory stimuli from speech. Since they may allow inferences about the times at which different (e.g. sub-segmental, segmental, supra-segmental) attributes of the (speech) stimulus are integrated into the mouse movement, they have also been referred to as *information integration times* (Stillman et al., 2018).

Another type of these measures describes durations and ratios of spatial and temporal features, such as

- *Peak velocity*, the maximum within-trial velocity value
- *Peak acceleration*, the maximum within-trial acceleration value
- *Initiation time*, the time it took for the participant to initiate their movement of the mouse
- *Reaction time* (RT) or *Movement time* (MT), the time it took for the participant to reach the chosen response[4]

Lastly, *movement angles* are another trajectory measure which can be used as temporally informative (in spite of their geometric nature): Any two subsequent sections of a mouse trajectory (which can be of arbitrary length) may be compared by inspecting the angle they form between them, which can in turn indicate changes in trajectory curvature as a function of time.

[4] RT is, of course, hardly exclusive to mouse tracking, but it can prove to be additionally informative especially in analytic comparison to other (e.g. more mouse-specific) measures

Measures of complexity Also called measures of uncertainty, of unpredictability, of spatial disorder, or of entropy (Stillman et al., 2018; Hehman et al., 2015), this class of mouse tracking measures assesses the relative lack of smoothness of mouse trajectories, i.e. their unpredictable directional fluctuations en route to the chosen response.

- *Sample entropy* is the "negative natural logarithm of the conditional probability that [...] sequences similar for m points remain similar at the next point" (Richman & Moorman, 2000), and is thus used to measure the degree of irregularity and unpredictability of the trajectory's horizontal movement component (since, in most studies, the response options are separated along the horizontal axis).
- A simpler measure of movement consistency can be arrived at by looking at *x-flips*, which denote the total number of times the mouse movement reversed its horizontal direction during a trial.
- Similarly, *x-reversals* are defined as the total number of times the mouse cursor crossed the y-axis (at the middle of the display), changing from one hemi-space into the other (Kieslich et al., 2019b).

If applied to the whole of a trajectory, these measures can be ascribed to the first category, since in such a case, they provide information on the global amount of conflict within the trajectory. In comparison to other measures belonging to this category, however, complexity measures are better able to remain informative even if portions of a trajectory are analysed separately and then compared to each other. This, in turn, means that they can rather easily be made to provide insights on the time course of mouse movements,[5] which would place them into the second category of measures.

2.3. Trajectory distributions

The usage of software which provides appropriate processing functions (e.g. Kieslich & Henninger, 2017) can make the calculation of most of the aforementioned mouse tracking measures from trajectory data a relatively trivial task. Once these measures have been obtained, it may seem equally trivial to conduct common statistical analyses by comparing groups of these measures to each other. The next paragraphs will outline how this approach may not function as straightforwardly for sets of mouse trajectories, and how it can be modified in order to make them interpretable.

[5] E.g. by investigating the point in time after which a trajectory steadily loses complexity, possibly indicating the conclusion of the underlying decision processes

Artefacts of aggregation Analyses of trajectory curvature are often performed by comparing aggregated, between-condition indices (e.g. MAD, AD, AUC) in order to assess whether the amount of excursion towards response options (and therefore the amount of conflict between them) exhibits systematic differences. It is important for researchers who intend to follow this approach to keep in mind that the resulting curvature comparisons will be based on multiple successive steps of information aggregation:

1. Curvature calculation: The individual trajectory shape is collapsed onto a singular data point.
2. Within-subject aggregation: The curvature values are aggregated across all trials from each individual participant.
3. Between-condition aggregation: The curvature values are aggregated per condition.

While the curvature indices obtained from this process may make a first comparison of average trajectory excursion in each experimental condition feasible, these aggregates "do not necessarily represent the shape of the individual trajectories well" (Kieslich et al., 2019b), having lost the bulk of information about their respective contours. It might, for instance, be the case that the trajectories collected in a study largely belong to two distinct types of trajectory shape, with one type being relatively straight paths to the chosen response, and another type heading towards the non-chosen option for a large portion of its path before reversing direction and ultimately settling on the chosen option (a kind of trajectory distribution which would correspond strongly to the predictions of a stage-based model of cognitive processing). Aggregating mouse trajectories from this distribution in the manner described above would lead to a representation of the data suggesting continuously graded curves and, for example, the existence of a gradual attraction to the competing response (e.g. Spivey et al., 2005). Accordingly, this representation might then be interpreted as evidence in favour of a dynamical systems approach despite its real nature as an artefact of aggregation.

In order to avoid this kind of aggregation effect which at best limits the meaningfulness of the measures used, and at worst renders them misleading, mouse tracking studies have traditionally made use of bimodality tests such as calculating the *bimodality coefficient* for the curvature indices (e.g. MAD) of a given distribution of trajectories, or employing *Hartigan's dip statistic* (Hehman et al., 2015). The purpose of these analytic tools is the assessment of the homogeneity of the distribution at hand, on the grounds of the expectation that a set of genuinely graded curves should exhibit a unimodal

distribution, while a mixed set of straight and strongly curved trajectories (as described above) should exhibit a bimodal distribution (Pfister et al., 2013). Having come upon a bimodal distribution of trajectories, researchers would have to refrain from interpreting the aggregate-level data as evidence for any *gradient* participant sensitivity to the attributes of the response options or the experimental stimuli, respectively.

Trajectory prototypes Even though analyses of bimodality have regularly been used for assessing the nature of trajectory distributions (Freeman, 2018; Hehman et al., 2015), they can ultimately only make a binary decision: whether that distribution is a unimodal or a bimodal one. With this constraint in mind, more recent efforts have chosen a more holistic approach to the characteristics of mouse movement data by forgoing the aggregate-level comparison of curvature indices altogether, substituting a method which aims to "take into account the complete shape of each trajectory by using every point of the trajectory" (Kieslich et al., 2019b).

One of these newly proposed methods designed to retain variability in trajectory types is the mapping of observed trajectories onto trajectory prototypes (Kieslich et al., 2019b; Wulff et al., 2019). Two such prototypical trajectories can be found in the distribution example described above, in which two distinct types of trajectory (one straight, one strongly curved) appeared. After having performed a meta-analysis of forty published mouse- and hand-tracking studies, Wulff et al. found that "the majority of datasets consist of multiple types of trajectories", including both the straight and the strongly curved prototypes (Wulff et al., 2018). Figure 2.2 depicts five common prototypes (as labelled by Kieslich et al., 2019a):

- *Straight*: The mouse trajectory does not deviate from the most direct path to the chosen response option.
- *Curved*: The trajectory exhibits a modest amount of curvature towards the non-chosen option.
- *Continuous change of mind* or CCOM: The trajectory noticeably veers into competitor space / towards the non-chosen option, but ultimately settles on the chosen option.
- *Discrete change of mind* or DCOM: The non-chosen option is actually reached before an abrupt directional shift leads the trajectory to the chosen option.
- *Double change of mind* or DCOM2: First, the chosen option is reached, then the non-chosen option is reached, then the chosen option is reached once more.

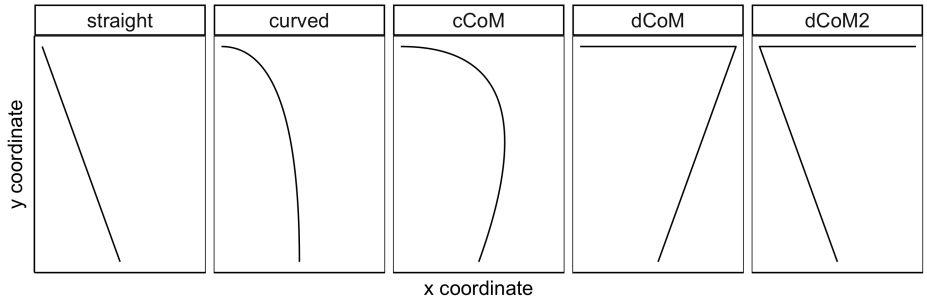


Figure 2.2.: Prototypical mouse trajectories. cCoM = continuous change of mind, dCoM = discrete change of mind, dCoM2 = double change of mind. Figure adapted from Kieslich et al., 2019a.

Computationally mapping multi-trajectory data onto prototypes requires those prototypes to be defined in advance (Wulff et al., 2019). In order to accomplish this, a sensible starting point for most mouse tracking studies may be to make use of the set of prototypes above, followed by a visual inspection of the grouped mappings to ascertain that as large a portion of the trajectories as possible is accurately represented by the prototypes. Should the prototypes fail to sufficiently capture the observed trajectories, a modified set of prototypes may then be defined which incorporates the idiosyncratic types from the dataset at hand.[6] This strategy does not in principle differ from an algorithmic clustering approach, in which the number of clusters ultimately has to be set manually, too (Friedman et al., 2009). Moreover, prototype matching renders the grouped trajectory sets comparable across conditions by keeping the parameters of the set of clusters consistent, as opposed to the variable clusters resulting from k-means or hierarchical clustering algorithms (Wulff et al., 2019; Friedman et al., 2009).

Interpreting multimodality Hence, a more appropriate way of assessing sets of movement trajectories may be a bottom-up method similar to the one described above, enabling researchers to account not only for the possibility of unimodal and bimodal, but also for that of multimodal distributions consisting of multiple trajectory types. Critically, when dealing with a trajectory distribution which exhibits any form of multimodality,

- the processes of condensing and aggregating (curvature) values may conceal meaningful variability in trajectory types

[6]Note that this process may include reducing the number of prototypes to as low as a singular one (in the case of having observed a truly homogenous, graded distribution of trajectories)

- statistical summaries regarding central tendencies will not readily be applicable[7]
- applying common continuous statistics to, for example, aggregated MAD values may violate statistical assumptions of normality

As the two most common underlying notions informing mouse tracking analyses have been the conceptions of cognitive processes as either *discretely stepwise* or *dynamically continuous*, it has to be noted that the existence of a set of types in a given distribution of trajectories may pose a challenge for researchers aiming to interpret mouse tracking data in terms of either of these notions. It is not only a well-defined bimodal distribution of straight and extremely curved trajectories that will be problematic for a dynamical systems approach, but potentially any size of a set of trajectory types, as long as they capture the underlying observations sufficiently well, since such categorical data do not appear appropriate as grounds for conclusions of continuity. In turn, a non-homogenous distribution of trajectories may turn out problematic for stage-based approaches after all, for instance if it is entirely made up of discrete types which nevertheless all lack the characteristic angles resulting from abrupt directional shifts as exhibited by DCOM trajectories.

Future research investigating the link between cognitive processes and trajectories from motor output may discover another point of complication for dual-systems theorists, who might readily accept angular trajectories as evidence for stepwise processing: As of yet, it is unclear if those trajectories actually result from discretely successive processes or alternatively, "considering that cognitive processes reside in the near-continuous communication of neurons in the brain" (Wulff et al., 2019),[8] from a discontinuous mapping of continuous cognitive processes onto movement.

2.4. Design decisions

To date, we possess only limited knowledge about the potential effects that details of the experimental setup can exert on mouse tracking studies. Similarly, evidence-based methodological standards or best practices do not exist yet, leading to a wide variety of setups and combinations of design choices in the literature (Scherbaum & Kieslich,

[7] See Wulff et al. (2019) for recommendations on how to statistically attempt detection of systematic differences in multimodal trajectory sets (e.g. analysing types as categorical variables, or performing ordinal logistic regression on ordered clusters)

[8] Also see Cisek & Kalaska (2005), Song & Nakayama (2009), and Spivey (2007)

2018). The validity and interpretability of data gathered by use of the mouse tracking paradigm depend, however, on a sound understanding of the intricacies of the way in which these data are elicited, not least with regards to underlying theories about cognitive processes like speech perception. Design factors might influence different aspects of the resulting mouse trajectories, e.g. their velocity profiles, their curvature, their individual shapes, the appearance of trajectory types, or the distribution of such types. Recent research investigating low-level design choices (Kieslich et al., 2019a; Scherbaum & Kieslich, 2018) has proposed that at least some of these choices are indeed capable of influencing trajectories and their distributions. The following paragraphs will provide an overview of three possible methodological issues of this kind.

Response selection mode One of the design decisions that have to be made when setting up mouse tracking experiments is concerned with the manner in which participants make their choice of response option known. In general, computer mice afford two ways of implementing this: Participants can either click on a response button, or they can simply move the cursor onto the button without clicking it. In the experimental software[9], either one of these actions can be configured to record their decision, and thus to end the current trial.

Although there might exist other ways in which the different modes of response selection could affect trajectories and their distributions, one of them appears to stand out. Recall the two kinds of extreme DCOM trajectories (figure 2.2): If the experimental design does not require a click of the mouse to indicate the final choice of response, but instead immediately accepts as that final choice the button which the cursor simply touches upon first, then the experiment will fail to record the second half of those trajectories. Consequently, they will appear as trajectories of another type (straight) in the analysis stage, falsely adding to the set of authentically straight trajectories in the resulting distribution, while "missing" from the set of DCOM trajectories.[10] By the same token, the following expectations should also hold true:

- In a design that does require a mouse click, larger aggregated curvature values should appear, since all extreme trajectories will get captured as such.
- The likelihood with which bimodality tests will classify a distribution as bimodal should be higher for designs which require a click.

[9] E.g. OpenSesame (Mathôt et al., 2011)

[10] Which, depending on the specific definition of the DCOM type, might still come to exist in the data, made up of extreme trajectories that happened to not quite reach the response area

Indeed, the findings of Kieslich et al. (2019a) are in accordance with these expectations, as their trajectory data

- exhibited stronger average curvature when a click was required
- contained fewer straight trajectories and more of the DCOM types when a click was required
- was classified as bimodal when a click was required
- was classified as unimodal when no click was required

This means that "a theoretically peripheral design aspect" (Kieslich et al., 2019a) can have the power to bias multiple kinds of analyses and interpretations that researchers may want to derive from mouse tracking data. Studies solely interested in, for instance, the comparison of aggregated curvature values may equally be led astray as those adopting a prototype-matching or clustering approach, and so may any other type of analysis of trajectory distributions. The fact that modifying a simple design feature can also change how bimodality tests classify a trajectory distribution should be of particular relevance for researchers who are interested in the nature of cognitive processes, since unimodal distributions of mouse trajectories have often been interpreted as evidence in favour of dynamic models, and bimodal distributions as evidence for stage-based models.

Movement initiation parameters If mouse tracking is to be used effectively as a method to capture on-line response behaviour whilst decision processes are ongoing, the experimental design has to ensure that the response movements coincide with those underlying cognitive processes. To this end, it has been recommended "to instruct participants to start moving their hands at the beginning of a trial, before the decision-related cognitive process is completed" (Fischer & Hartmann, 2014). However, the issuing of verbal or written instructions to participants is not the only way in which researchers may attempt to elicit early movement initiation (Scherbaum & Kieslich, 2018). It is also possible to impose a "maximum initiation time" (an interval which begins with the click on the start button), after which the experimental software either cuts short the current trial or displays a warning asking the participant to start their movement earlier in the coming trials. Alternatively, researchers may choose to employ a dynamic starting procedure, in which stimulus presentation does not occur until the mouse cursor has crossed an invisible horizontal boundary on its way upwards. Equally feasible are combinations of all three methods.

It should be noted that, the specifics of the chosen method notwithstanding, the issue of timely movement initiation appears particularly relevant for researchers interested in

- responses to experimental tasks which are known to elicit very short reaction times, i.e. are not cognitively demanding
- responses to very short auditory stimuli
- responses to longer auditory stimuli, where task-critical information is conveyed early on in the acoustic signal

In these kinds of cases, it stands to reason that crucial portions of the underlying cognitive processes of interest will have come and gone especially quickly, and so will the opportunity to capture any motor output that might show traces of their influence. All three of these scenarios may well pertain to mouse tracking studies in speech research.

Naturally, it would also be possible to implement neither of the above techniques, thus enabling the decision processes to run their course fully before the response movement commences.[11] This type of design could reasonably be expected to produce a greater number of prototypically straight trajectories. Accordingly, the averaged curvature values calculated from the resulting trajectory distribution should be expected to be lower than the ones resulting from a design which encouraged early movement by any of the means mentioned above.

Once more, these expectations are in line with the findings from Kieslich et al. (2019a), who observed

- greater proportions of prototypically straight trajectories when no measures were taken to promote early movement initiation
- comparable numbers of straight and curved types when an initiation time threshold was implemented
- a majority of curved trajectories when the starting procedure was dynamic

Thus, the parameters of movement initiation appear able to shape mouse trajectory distributions in a similarly meaningful way as does the mode of response selection. Studies refraining from the use of techniques designed to ensure that decision processes are reflected in the mouse movement in their entirety might not only lose critical information, but they might also interpret the resulting large number of straight

[11] E.g. for studies in which initiation time is the variable of interest

trajectories as "belonging to a low-conflict decision in a dual-system model" (Kieslich et al., 2019a). Conversely, studies employing a dynamic starting procedure might take the relatively large number of curved trajectories to be indicative of continuously dynamic underlying processes.

Sensitivity and acceleration settings The mouse tracking paradigm takes advantage of a commonplace input device, forgoing the specialised hardware-software combinations which may be encountered when using dedicated research instruments. As a consequence, some important technical parameters depend on the idiosyncratic features of the experimental computer, for example, the specifics of the mouse hardware and of the way its output is translated into on-screen cursor movement. Luckily, one defining factor of this translation process is user-facing, namely the settings concerning the *sensitivity* and the *acceleration* of the mouse pointer on the level of the computer operating system. These settings thus form another design choice mouse tracking studies should consider.

The sensitivity setting modulates how the physical displacement of the mouse (commonly reported in dots per inch) is converted to the displacement of the graphical mouse pointer on the screen (in pixels). Higher sensitivity values lead to greater pointer displacement for equal physical movements, which means that less effort is required from users (participants) in order to reach any and all areas of the screen, while it gets more difficult for them to perform precise pointer movements.

The acceleration setting determines whether and how the mouse sensitivity may be dynamically adjusted by the operating system depending on either the velocity of the mouse or the duration of its being in motion. Activated, this setting means that there will be dynamically rising sensitivity values during sustained mouse movements and/or accelerating mouse movements, i.e. increasingly small movements of the hand or wrist will suffice to displace the mouse pointer greatly, again at the expense of precision. The series of threshold values which makes up the *acceleration curve* determines the extent to which this setting is then automatically applied to mouse movement values.

Together, these settings can considerably affect participants' interaction with the experiment. For instance, it would be possible to encourage early mouse movement by any of the methods described in the previous section, while at at the same time having entered a high mouse sensitivity value. For sufficiently high values, this design might result in cursor movements which reliably hit the upper display edge (maybe even before any stimulus presentation has begun). The resulting trajectories should be expected to be exceedingly similar to one another, each exhibiting one square angle instead

of any curvature. Likewise, extremely low sensitivity values might be tantamount to using an outcome-based design, as the resulting sluggish mouse movements might prohibit any fine-grained reflections of cognitive processes from being detectable in the accompanying trajectories. In order to be able to arrive at appropriate sensitivity and acceleration values between these extremes, the following recommendations ought to be considered:

- Mouse tracking studies employing speech stimuli (or other auditory stimuli) should take into account the respective durations of those stimulus sounds. While it might be desirable to have stimulus playback start only when the mouse movement is already underway (e.g. by employing a dynamic movement initiation procedure), the mouse settings should be such that the cursor will not have reached the upper display edge before stimulus playback ends.
- Similarly, mouse tracking studies might furthermore want to consider employing pilot studies to generate baseline reaction time values for the experimental task(s) at hand. At least in conventional mouse tracking designs, the mouse settings should always allow participants to fully conclude their cognitive processes before the moving mouse pointer reaches a display edge, in order to maximise the meaningfulness of the resulting distribution of trajectories.
- Moreover, the fact that most participants are well-versed in the usage of a computer mouse also means that they likely expect a certain mapping of their hand movements to on-screen pointer location. Any manipulation of the mouse settings could therefore be expected to feel unfamiliar to them, thereby heightening their sense of using a technical device in an experimental setting. A sensible recommendation for future studies could be to allocate a certain amount of time for each subject to (re-)familiarise themselves with the way the mouse pointer reacts to their motions before the experiment starts.
- With regards to theories linking movements of the arms and hands to cognitive processes, it seems vital to consider the fact that any mouse acceleration setting differing from an entirely flattened acceleration curve constitutes a non-linear mapping between the physical and graphical displacements involved.[12] To date, the role which such a non-linear mapping might play for any linking hypotheses regarding the validity of mouse trajectories as substitutes for indexical pointing gestures is unknown (Kieslich et al., 2019a).

[12]The default settings of all major operating systems are non-linear in this regard

- Also, it has been argued that, in order to "capture cognitive effects in the trajectory measures", mouse tracking researchers should "lower the default speed of the mouse to a reasonable range" (while also disabling mouse acceleration) (Fischer & Hartmann, 2014). The logic informing this recommendation runs parallel to that regarding movement initiation parameters: The primary goal is to "envelop" any underlying decision processes with motions of the mouse, lest any portion of said processes miss their chance to leave a mark on those motions. The smooth hand movements which could be expected to result from the settings Fischer & Hartmann suggest should be more desirable in these terms than the potentially rougher, jerkier motions that might otherwise ensue. Additionally, slowed mouse settings should better ensure that the hand movements are still ongoing during the posterior parts of the underlying processes.

On the one hand, stage-based accounts of decision evolution would predict that higher sensitivity and acceleration values should result in a greater proportion of the DCOM types of trajectories, since any potential corrective interference by secondary processing stages would get driven into increasingly later spatial sections of the trajectories. Seeing as greater DCOM proportions can easily lead to greater curvature values on the aggregate level (e.g. MAD), it appears likely that continuity-based theories, on the other hand, would infer from this kind of result that speedier mouse settings lead to larger (captured) effects (Fischer & Hartmann, 2014). Thus, both accounts agree with the findings of Kieslich et al. (2019a), who observed greater proportions of DCOM trajectories alongside greater aggregate-level MAD values for default (i.e. fast and accelerated) mouse settings.

In sum, the power of seemingly innocuous design decisions to influence the ways in which participants interact with a given mouse tracking experiment as well as the resulting trajectory distributions should not be underestimated. Moreover, researchers interested in drawing conclusions about the validity of theoretical models fundamental to processes such as (speech) perception and categorisation should take care to always assess trajectory data in the context of those design factors. Still, in order to make evidence gathered by use of the mouse tracking paradigm more reliably interpretable, much additional methodological work will yet have to be undertaken, for example testing separately the effects of sensitivity and acceleration settings, testing different intensities of these settings for different experimental tasks, or examining possible interactions with e.g. different movement initiation parameters.

3. Implementation

3.1. Expectations

In order to make first inroads with regards to the use of the mouse tracking paradigm as an experimental technique for speech perception research, two exploratory experiments were conducted. Beyond this methodological primary focus, the experimental design closely followed the eye tracking study of McMurray et al. (2008) as described in section 1.3, in principle attempting a replication of their findings as a proof-of-concept for mouse tracking in speech research.[1] To reiterate, McMurray et al. claimed to have found evidence for graded intra-categorical (i.e. sub-phonemic) sensitivity to acoustic detail in that they observed greater proportions of eye fixations to the competing response option (/b/ vs. /p/) the closer the corresponding stimulus VOT got to the category boundary between those voiced and voiceless stops.

Comparable to eye tracking measures, the mouse tracking paradigm makes concrete predictions with regards to phoneme identification. If the dynamics of hand movement patterns should indeed reflect intra-categorical sensitivity to VOT values, a three-fold replication of McMurray et al.'s eye movement data should be expected:

- Outcome-based: a replication of classical categorical perception findings, where stimuli are reliably identified as belonging to one of two categories, with little to no within-category variation
- Online-based: a replication of previous reaction time findings, with longer reaction times for more ambiguous stimuli (i.e. stimuli approaching the category boundary)
- In mouse tracking terms, the hypothesised graded increase in competition between the response options (for increasingly ambiguous stimuli) should manifest itself in an equally graded increase in the excursion of the mouse movement towards

[1] In particular, McMurray et al.'s second experiment was the template for the present study, with the main difference being the substitution of stimuli from natural speech for McMurray et al.'s synthetically produced stimuli

the competitor response option on the way to the chosen response.

The gradiency hypothesis as presented by McMurray et al., which posits as the source for this kind of graded motor output the existence of underlying cognitive mechanisms which are themselves of a gradual nature (as opposed to being discrete or stage-based), also predicts that the resulting trajectory distributions should principally consist of gradedly curved trajectories, with sufficiently few DCOM-type trajectories as to preclude the possibility of distorting aggregation effects arising from a bimodal distribution (see section 2.3).

The following sections will describe both mouse tracking experiments, and will report on their results in a descriptive manner. The exploratory nature and small sample size of the experiments, which were primarily conducted to test the fundamental technical feasibility of using the paradigm for speech perception research, prohibit any inferential assessment of the resulting data. Any new hypotheses that might be generated in examining these data would need to be independently tested in experimental designs with sufficient statistical power. This equally applies to both the individual experiment reports (sections 3.2 and 3.3) and the between-experiment comparison (section 3.4).

3.2. Experiment 1

Method

Design The experiment was conducted at the Universität zu Köln, Germany. Participants were seated at a desk in front of a MacBook Air (1.6 GHz Intel Core i5) with a display resolution of 1440 × 900 pixels. They used a Logitech B100 USB mouse in order to control the mouse pointer. The principal experimental task as well as the graphical design closely followed the prototypical two-alternative forced choice mouse tracking design described in section 2.1: Participants had to choose from two visually presented response options upon hearing pre-recorded speech files containing either of the CV syllables /ba/ and /pa/, which differed only in VOT duration. Figure 3.1 shows the main experimental display seen during each trial. Hoping to mitigate potential handedness effects, participants were assigned by lot to one of two groups which differed with regards to the on-screen placement of the response buttons: One group always saw the “ba”-button in the upper left corner, while the other group always saw the “pa”-button in that corner. The buttons were 110 pixels wide and 80 pixels high.

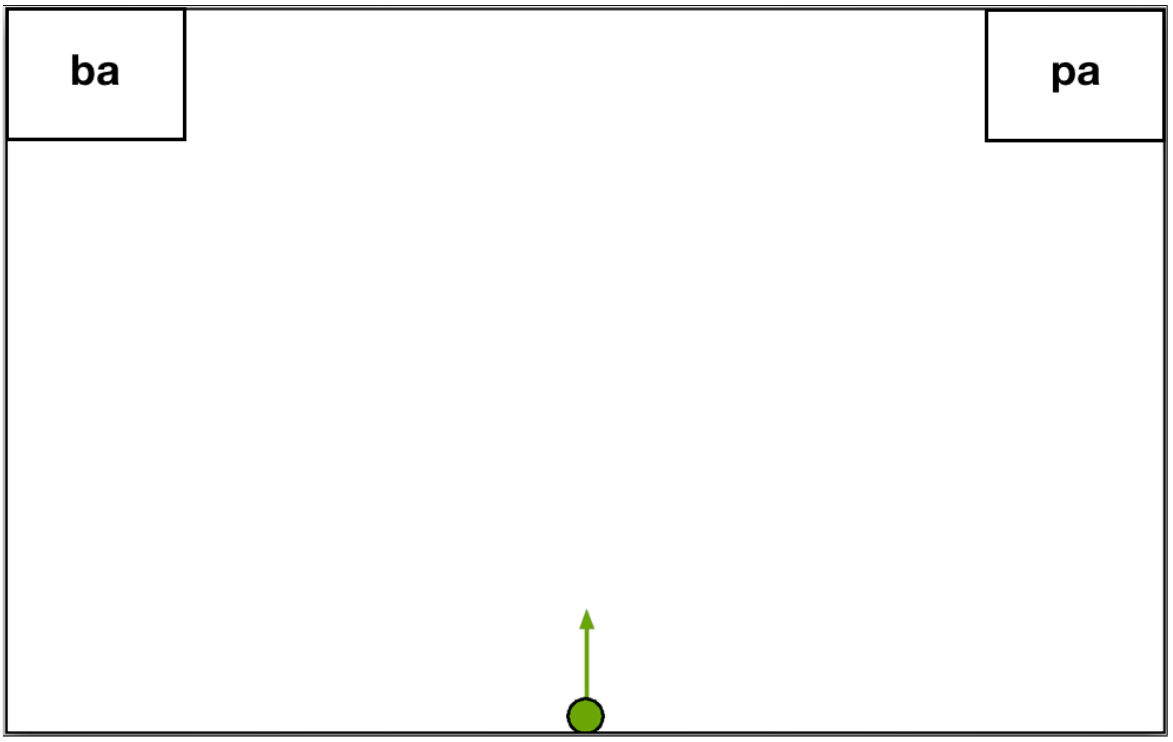


Figure 3.1.: The main experimental display

Written instructions asked the participants to begin each trial by clicking on the green circle on the bottom edge of the display, and to immediately commence the movement of the mouse away from themselves, i.e. to effect a vertical displacement of the mouse pointer towards the upper display edge. They were additionally asked to keep the movement constant, not to perform any backtracking or loops, and to respond as quickly and as accurately as possible. Stimulus playback was initiated dynamically the moment the mouse pointer crossed an invisible horizontal boundary line located at 338 pixels from the bottom edge of the display (roughly corresponding to the tip of the green arrow seen in figure 3.1). The experiment did not require participants to click on the response buttons in order to indicate their choice. Instead, their response was recorded as soon as the mouse pointer reached one of the response button rectangles, thereby ending the current trial. Mouse pointer acceleration was disabled (i.e. linearised) and pointer sensitivity was slowed down to 75% of the macOS default value via the CursorSense software (Plentycom Systems, 2016).

Each of the participants in the study was exposed to a total of 117 stimuli, separated into a training block consisting of 27 trials (9 VOT steps $\times$ 3 repetitions) and a test

block consisting of 90 trials (9 VOT steps × 10 repetitions). This resulted in 900 total test trials to be assessed in the analysis.

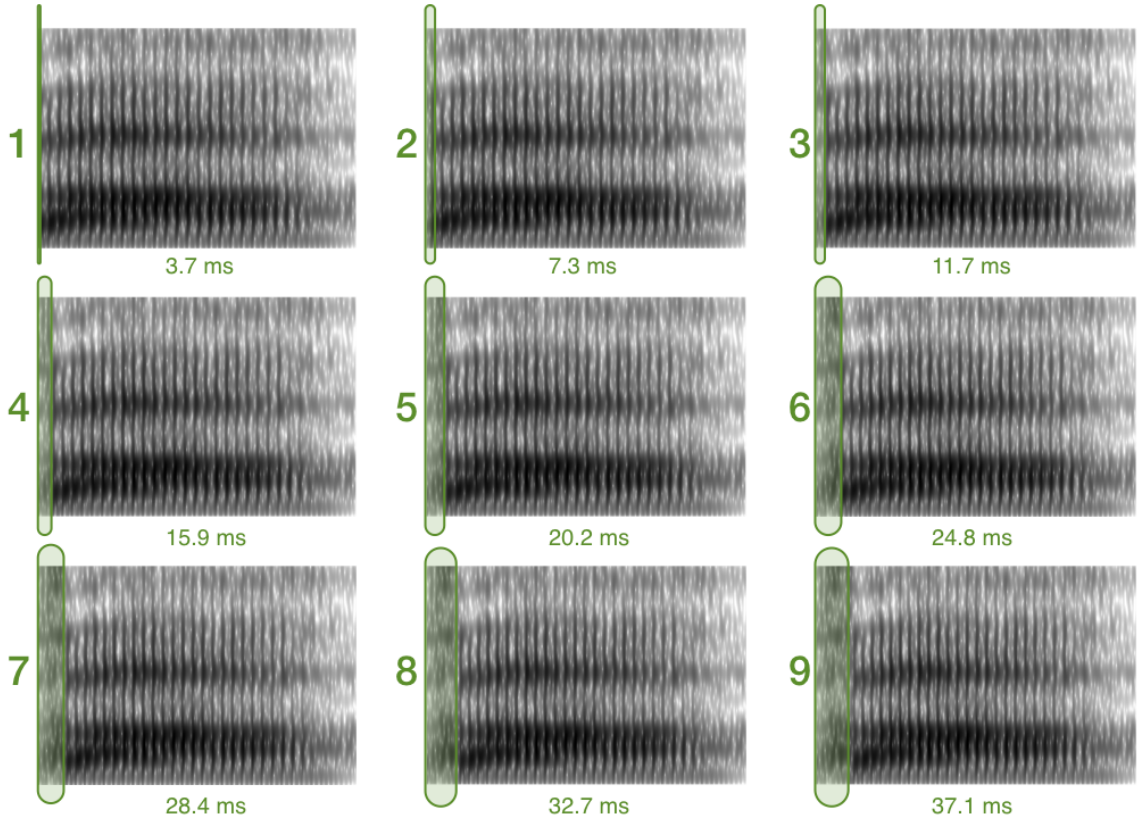


Figure 3.2.: Spectrograms of the nine stimuli, with the progressive VOT splicing highlighted

Stimuli The stimuli were produced by a 34-year-old male native speaker of Standard German. They were recorded in a quiet room through a Røde NT1-A condenser microphone plugged into an Edirol UA-25 audio interface equipped with an AKM AK5381 A/D converter. Recording parameters were set to a 44.1kHz sample rate and 16-bit depth. For both CV syllables /ba/ and /pa/, 15 productions each were recorded and then normalised across tokens. The endpoint tokens for the VOT continuum that was to be created were selected using three selection criteria prioritised in the following order:

1. Best match with the respective prototypical VOTs for voiced and voiceless German

obstruents[2]
2. Best match on vowel pitch
3. Best match on stop formant frequencies

The VOT continuum was then created in *Praat* (Boersma & Weenink, 2016) by removing larger and larger portions of the aspiration phase from the voiceless /pa/ token and inserting them into the immediately post-burst phase of the voiced /ba/ token (Andruski et al., 1994). Following McMurray et al. (2008), a continuum consisting of nine VOT steps was chosen, which in this case resulted in stimuli durations ranging from 3.7 to 37.1 ms, with a calculated VOT step size of 4.17 ms. The necessary resetting of the individual splicing points onto zero crossings of the waveform resulted in slightly irregular actual step sizes and corresponding stimulus durations (see table B.1 in appendix B). Spectrograms of all nine stimuli are shown in figure 3.2.

Participants Ten native speakers of German (5 female, 5 male), participated in this experiment. All ten were trained phoneticians working at the I*f*L Phonetik, Universität zu Köln. All of them reported being right-handed and had normal or corrected-to-normal vision.

Results

Data processing All calculations on the data were performed in the statistical programming language *R* (R Core Team, 2018), using the integrated development environment *RStudio* (RStudio Team, 2015).

A check of the actual sampling frequencies that occurred across all 900 test trials revealed that 99.7% of data points had been sampled at exactly the desired frequency of 100 Hz set in the experimental software. Small deviations from this desired frequency (like the residual 0.3% found here) are impossible to avoid due to the complex interplay of the hard- and software stages involved in the signal chain.

Trajectories exhibiting any movement reversals on the y-axis, including loops, were excluded from the analysis. Moreover, trajectories that had reached any border of the screen for any amount of time were excluded, since they exhibited artificially flat

[2]In order to arrive at these prototypical VOT values, VOT durations from both Jessen's review of studies measuring German obstruent VOTs and his own production experiment (Jessen, 1999) were aggregated and averaged

contour portions. This resulted in exclusion of 8.1% of the data. See figure A.1 in appendix A for examples of well-formed and malformed trajectories, respectively.

Where appropriate, the trajectories were transformed into a more easily analysable state by normalising them spatially and/or temporally. Time-normalising trajectories, for instance, renders trajectories which differ in duration comparable by interpolating them "so that each is represented by the same number of positions [...] separated by a (within-trial) constant time interval" (Kieslich et al., 2019b).

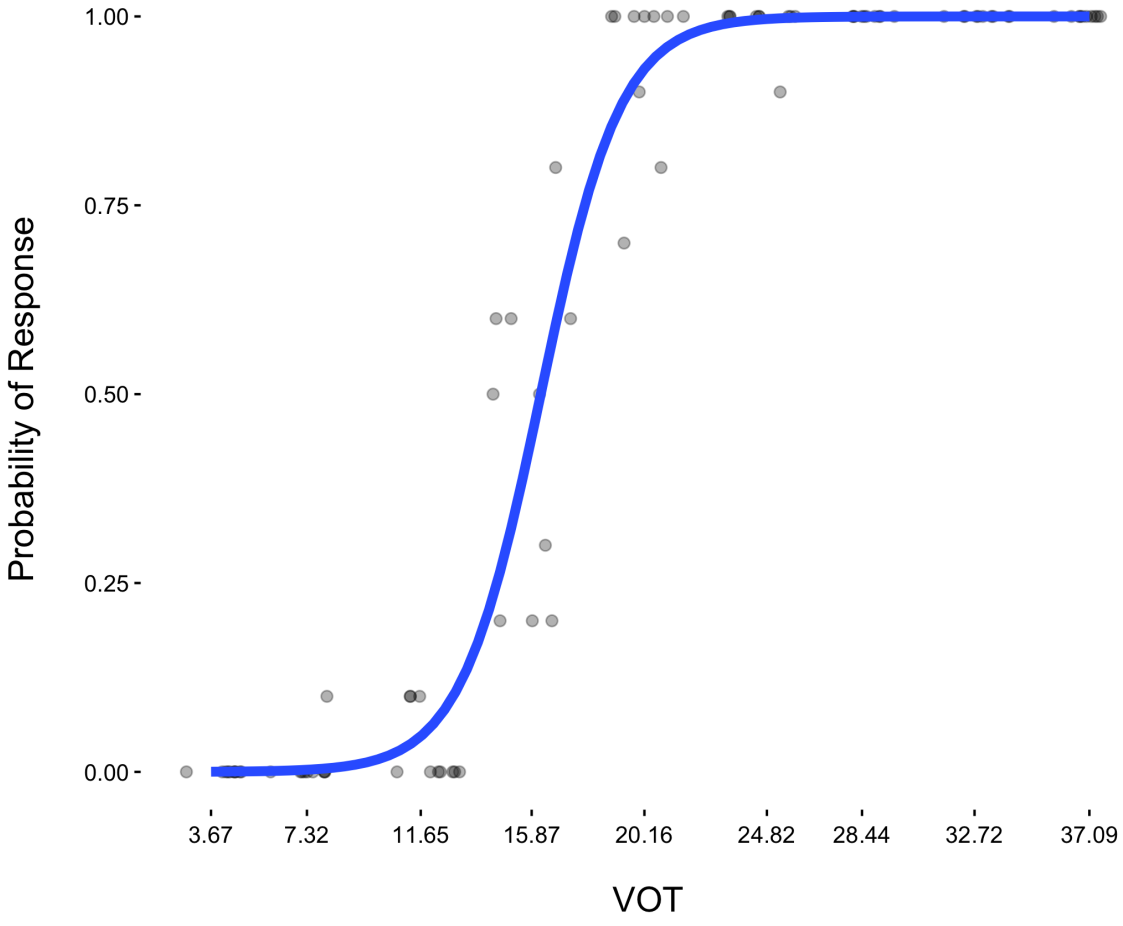


Figure 3.3.: Experiment 1: The modelled identification function across participants. Estimated identification proportions on the y-axis, stimulus VOT on the x-axis (with the actual stimulus VOTs as axis ticks)

Identification In order to assess participants' mouse movement responses in the context of the category boundaries shaping their perception of stop voicing, their

average category boundary was determined. To do so, a logistic regression model was fit to the participants' response data. The regression plot in figure 3.3 shows voice onset time on the x-axis labelled with the actual VOT continuum steps. The points depict the mean proportions of responses per subject, and the regression line shows the modelled probability of a voiceless-/pa/ response on the y-axis. It is apparent that participants exhibited response behaviour in line with categorical perception, as there is only a narrow VOT range that resulted in ambiguous responses across trials, and plateaus on either side of the VOT spectrum where identification was almost perfect, resulting in a rather steep sigmoidal identification curve. Table 3.1 shows the modelled across-subject probability of a /pa/-response, given the specific stimulus VOTs.

VOT (ms)	P(/pa/)	SD
3.67	0.00	0.00
7.32	0.01	0.03
11.65	0.03	0.05
15.87	0.45	0.21
20.16	0.94	0.11
24.82	0.99	0.03
28.44	1.00	0.00
32.72	1.00	0.00
37.09	1.00	0.00

Table 3.1.: Experiment 1: The modelled across-subject probability of a /pa/-response, given the specific stimulus VOTs

Should the present experimental design be capable of eliciting comparable data from a larger sample, such a replication of classic outcome-based results would validate the stimuli created by use of the progressive splicing method. The mean proportion of /ba/-responses was calculated to be 39.78% ± 3.43% (see table B.2 in appendix B for the individual category boundaries and response proportions). This proportion appears to indicate that the continuum VOT range chosen here may constitute a valid, but sub-optimal compromise between prototypical German VOT values as found in the literature (Jessen, 1999), and aiming to design a VOT range with an average 50% of responses falling on either side of it. The mean across-subject category boundary was calculated to be separating the continuum at 16.19 ms ± 1.15 ms.

Trajectories resulting from trials in which the subject heard a stimulus on the voiced side of their individual category boundary, but nevertheless selected the voiceless response option (or vice versa) were treated as erroneous responses and excluded from the analysis (McMurray et al., 2008). This resulted in exclusion of an additional 5.3% of the data.

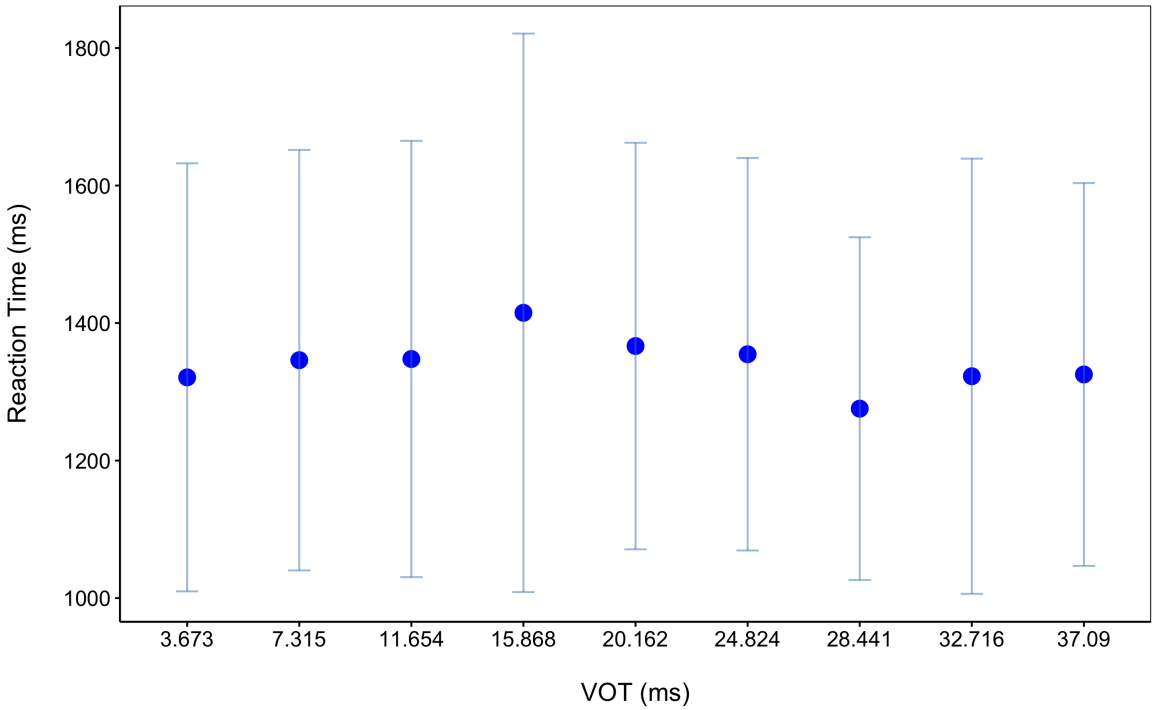


Figure 3.4.: Experiment 1: The mean across-subject reaction time values (in ms) along the y-axis, with stimulus VOT (in ms) on the x-axis. Error bars show standard deviations

Reaction time When averaged across subjects and all trials, the amount of time participants needed to complete the experimental task of identifying the speech stimuli by completing a selecting mouse movement[3] exhibited an ambiguous pattern, as figure 3.4 shows. Reaction time values ranged from 1275.55 ms ± 249.23 ms for the antepenultimate continuum step up to 1414.94 ms ± 406.1 ms for the fourth step, the latter being the stimulus closest to the across-subject category boundary. While the "voiced" side of the continuum (the first four steps) did elicit reaction times which rose monotonically for stimuli approaching the boundary, the difference in RT between

[3]Specifically, reaction time was defined as the time period between the click on the start button and the arrival at one of the response boxes

steps two and three was exceedingly small (1.68 ms). On the "voiceless" side of the continuum, reaction times fell monotonically up until the seventh stimulus step, before rising again for the last two steps. Table 3.2 provides a summary of these values.

VOT (ms)	Mean (ms)	SD (ms)
3.673	1321.06	311.25
7.315	1346.06	305.76
11.654	1347.74	317.25
15.868	1414.94	406.10
20.162	1366.59	295.60
24.824	1354.67	285.39
28.441	1275.55	249.23
32.716	1322.70	316.45
37.09	1325.27	278.38

Table 3.2.: Experiment 1: Across-subject mean reaction times per step of the VOT continuum

Curvature In order to quantify the nature of the expected differences in trajectory excursion (as stated in section 3.1), a focus on Maximum Absolute Deviation as a typical measure of curvature was decided upon. Across-subject MAD was calculated for each stimulus step, with figure 3.5 providing a graphical representation of the results. MAD values (measured in pixels) ranged from 203.4 px ± 49.35 px for the first step of the VOT continuum up to 233.43 px ± 77.28 px for the third step, while the fourth step (closest to the across-subject category boundary) elicited a very similar 233.11 px ± 75.2 px of maximum deviation on average. Other rises in MAD occurred within the "voiceless" side of the continuum, where deviation reached 225.69 px ± 78.69 px for the seventh step, as well as 220.51 px ± 52.8 px for the ninth step. Notably, curvature values for both sides of the stimulus continuum do not correspond well to the expectation of monotonically rising movement excursion as the stimulus VOTs approach the category boundary from either the "voiced" or the "voiceless" side of the continuum. See table 3.3 for a summary of these results.

Trajectory distribution As a first measure in the assessment of the distribution of mouse trajectories, the overall bimodality coefficient was calculated by standardising

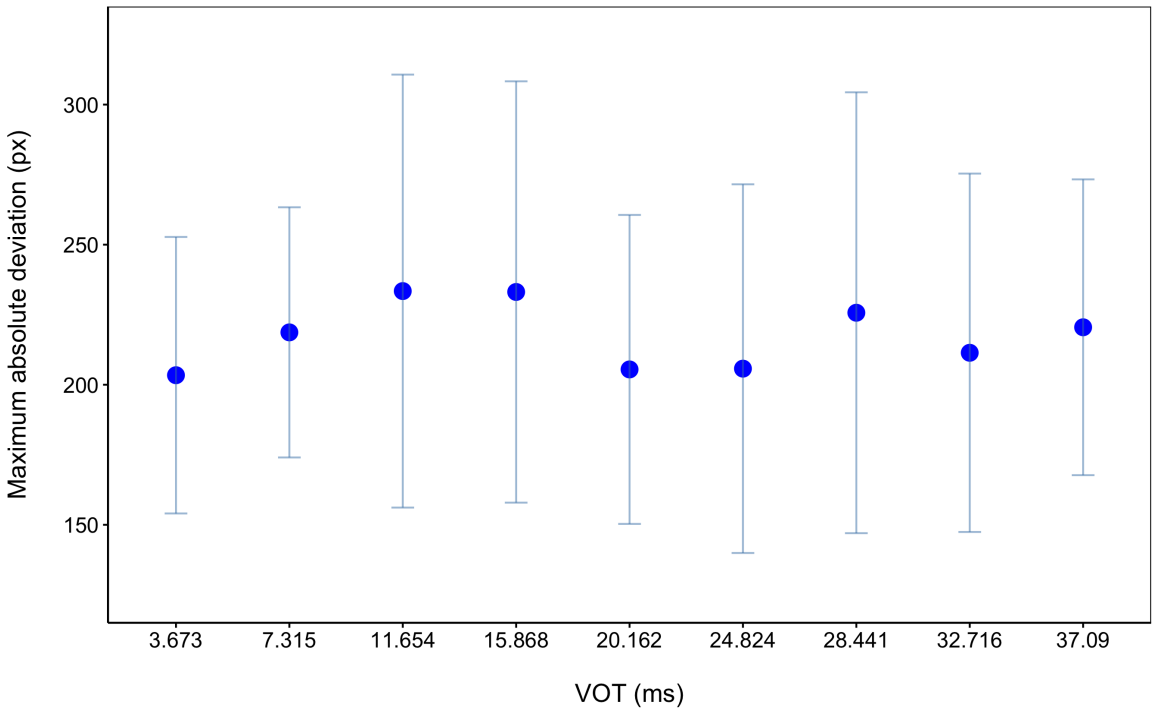


Figure 3.5.: Experiment 1: The mean across-subject MAD values (in px) along the y-axis, with stimulus VOT (in ms) on the x-axis. Error bars show standard deviations

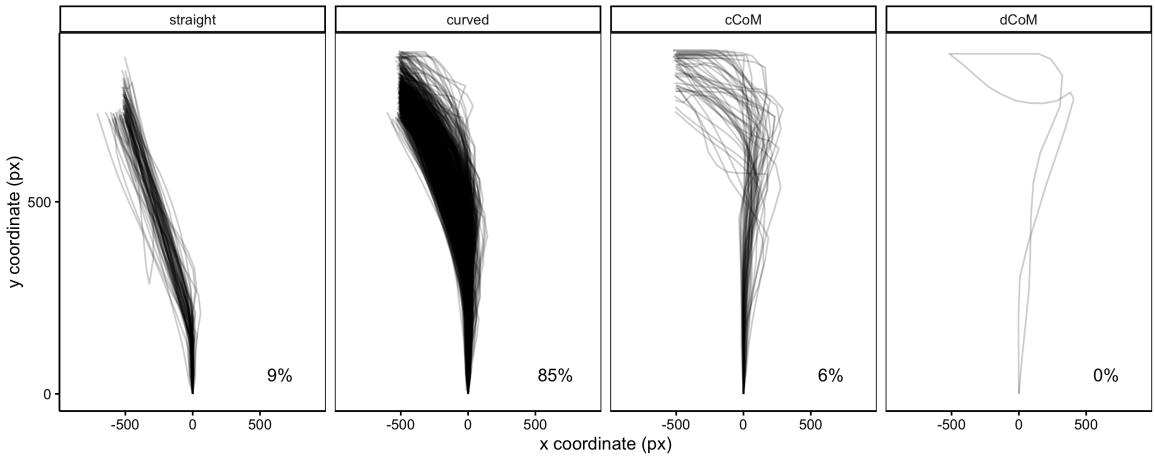


Figure 3.6.: Experiment 1: Proportions of trajectory shapes as mapped to a set of four prototypes

VOT (ms)	Mean (px)	SD (px)
3.673	203.40	49.35
7.315	218.69	44.65
11.654	233.43	77.28
15.868	233.11	75.20
20.162	205.46	55.15
24.824	205.74	65.80
28.441	225.69	78.69
32.716	211.40	63.97
37.09	220.51	52.80

Table 3.3.: Experiment 1: Across-subject mean MAD per step of the VOT continuum

the MAD values per subject and then computing the coefficient across all trials (following typical analyses found in the literature, (e.g. Freeman & Ambady (2010), Spivey et al. (2005), and Kieslich et al. (2019a)). The bimodality coefficient was found to be 0.475, smaller than the value of 0.555, which is generally taken to be the cutoff point after which distributions are assumed to be bimodal (Pfister et al., 2013).

An overall visual inspection of the trajectories by way of a trajectory heatmap also did not reveal any bimodality (or multimodality) in the distribution (see figure A.2 in appendix A). In order to gain more fine-grained insight into the nature of the trajectory distribution, a set of four trajectory prototypes similar to the ones shown in section 2.3 was constructed. The observed trajectories were then mapped onto these prototypes as shown in figure 3.6, revealing that the majority (85%) of trials resulted in mouse movements of the "curved" type, while relatively small proportions of the trajectories fit the "straight" and the "continuous change of mind" types (9% and 6%, respectively). There were only two trajectories belonging to the "discrete change of mind" or DCOM type, resulting in a rounded percentage of 0%. Visually, the computed mapping seemed to have captured the individual trajectory shapes well, with no obvious candidates for an additional prototype to be included in a possible reiteration of the matching procedure.

Any type of "change of mind" trajectories might be taken as evidence for increased conflict between response options in contrast to low-conflict "straight" trajectories. A by-stimulus comparison of the mapped trajectories did not reveal any systematic

VOT (ms)	straight	curved	cCoM	dCoM
3.673	9.6 %	89.4 %	1.1 %	0.0 %
7.315	8.7 %	85.9 %	5.4 %	0.0 %
11.654	8.7 %	82.6 %	8.7 %	0.0 %
15.868	9.7 %	83.9 %	6.5 %	0.0 %
20.162	7.1 %	87.1 %	5.9 %	0.0 %
24.824	11.2 %	83.1 %	5.6 %	0.0 %
28.441	9.0 %	82.0 %	7.9 %	1.1 %
32.716	11.9 %	83.3 %	4.8 %	0.0 %
37.09	9.3 %	83.5 %	6.2 %	1.0 %

Table 3.4.: Experiment 1: Proportions of trajectory types per stimulus VOT. Respective percentages may not sum to exactly 100% due to rounding

increase in proportions of the CCOM (or DCOM) types for stimulus VOT durations closer to the across-subject category boundary, as can be seen in table 3.4 (for a visual representation, see figure A.3 in appendix A). The CCOM type occurred most frequently for the third continuum step (8.7%), and least frequently for the stimulus with the shortest VOT (1.1%), with nonlinear trends of percentages on either side of the voicing spectrum.

Discussion

While the outcome-based results from this experiment corresponded well to the expectations of categorical perception, with distinct phoneme categories lacking a salient area of ambiguous overlap, they were less consistent with previous findings regarding sub-phonemic sensitivity based on reaction times. Although the longest reaction times did occur when participants heard stops with the VOT that was closest to their category boundary, the overall reaction time pattern did not straightforwardly agree with the expected pattern of linear trends on both sides of the VOT continuum. Crucially, the mouse movements did exhibit even less of a discernible pattern in the terms of the gradiency hypothesis put forth by McMurray et al. (2008), since their overall curvature (as represented by the Maximum Absolute Deviation) did not peak exclusively on the boundary-adjacent stimulus, while also showing a decidedly non-linear sequence of values for the "voiceless" side of the stimulus spectrum. The distribution of trajectories

as a whole, however, did conform with the expectations put forth above, as it did not exhibit any signs of bimodality, and as there was a clear majority of smoothly curved trajectories, with relatively few trajectories matching either of the "change of mind" prototypes.

If a study with an appropriate sample size were to confirm the apparent lack of systematic differences with regards to stimulus VOT within a large proportion of continuously curved mouse trajectories (as found in the present exploratory setting), such a failure to replicate the findings of McMurray et al. (2008) would require clarification. Seeing as the mouse tracking paradigm is still a developing experimental technique, it would then appear prudent to search for methodological explanations for the diverging outcome first. One possible point of application for such an approach could be the mouse pointer settings used in this experiment: Assuming that the underlying decision mechanisms indeed exhibit subtle nuances in reaction to differences in VOT on the millisecond scale, it is conceivable that the combination of reduced mouse sensitivity and a flattened acceleration curve forced participants to perform movements of the arms and hands which were simply too slow and sluggish to remain capable of reflecting any such detail. With this in mind, a follow-up experiment was devised, as described in the section below.

3.3. Experiment 2

Method

Design and stimuli The second experiment used the default mouse settings of macOS (i.e. "normal" sensitivity and enabled acceleration). No further design alterations were introduced. The stimuli used in this experiment were identical to the ones described for the first experiment.

Participants Ten native speakers of German (5 female, 5 male, all trained phoneticians working at the I*f*L Phonetik, Universität zu Köln) participated in this experiment. Nine participants reported being right-handed, while one indicated a preference for the left hand. Six of the ten participants had already partaken in the first experiment of the present study. All ten had normal or corrected-to-normal vision.

Results

Data processing An examination of the actual rate with which mouse movement data was sampled once more revealed that 99.7% of all data points had been sampled at the desired frequency of 100 Hz. After visual inspection of the individual trajectories, 4.4% of the data were excluded as malformed trajectories. Space- and time-normalisations of the trajectory data were performed as required by the analysis.

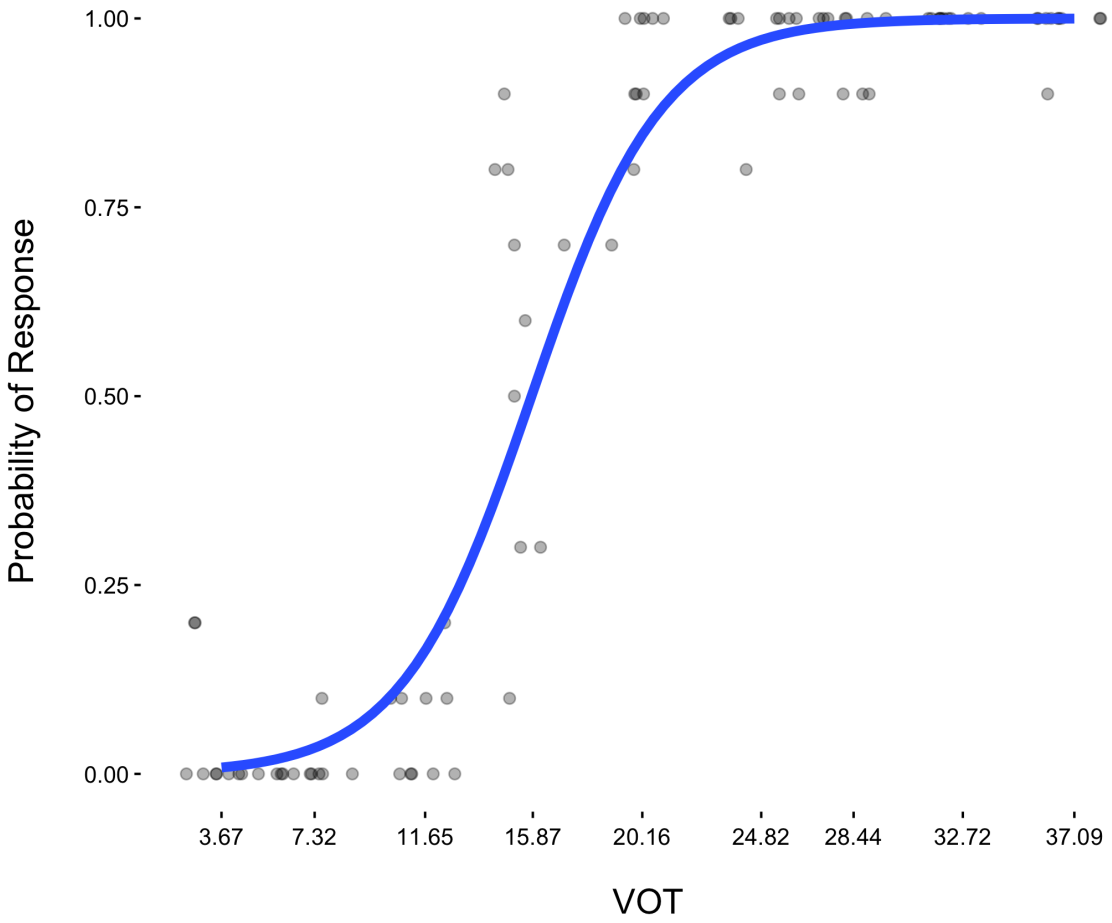


Figure 3.7.: Experiment 2: The modelled identification function across participants. Estimated identification proportions on the y-axis, stimulus VOT on the x-axis (with the actual stimulus VOTs as axis ticks)

Identification Again, a logistic regression model was fit to the participants' response data. Figure 3.7 depicts the across-subject identification (VOT on the x-axis, with

the points showing the mean proportions of responses per subject, and the regression line showing the modelled probability of a /pa/-response on the y-axis). While the resulting sigmoidal curve did not exhibit quite as narrow a range where identification was ambiguous when compared to the first experiment, it still appears clear that—viewed on an outcome-based level—participants' behaviour with regards to categorical perception of stop voicing was consistent with the expectations as noted in section 3.1. Table 3.5 summarises the modelled across-subject probability of a /pa/-response, given the specific stimulus VOTs.

The mean proportion of /ba/-responses was calculated to be 38.67% ± 4.53%, again indicating that the chosen VOT continuum may be skewed in favour of its voiceless portion (see table B.3 in appendix B for the individual category boundaries and response proportions). Dependent on the per-subject boundaries, an additional 7.1% of trials were excluded as individually wrong responses. For this second experiment, the mean across-subject category boundary was calculated to be separating the VOT continuum at 15.89 ms ± 1.71 ms.

VOT (ms)	P(/pa/)	SD
3.67	0.04	0.08
7.32	0.01	0.03
11.65	0.06	0.07
15.87	0.57	0.26
20.16	0.92	0.10
24.82	0.96	0.07
28.44	0.97	0.05
32.72	1.00	0.00
37.09	0.99	0.03

Table 3.5.: Experiment 2: The modelled across-subject probability of a /pa/-response, given the specific stimulus VOTs

Reaction time The mean across-subject reaction times in the second experiment exhibited less of a discernible pattern than did those of the previous experiment, as figure 3.8 shows. Mean RT values ranged from 974.46 ms ± 277.02 ms for the second continuum step up to 1046.59 ms ± 289.62 ms for the fourth step, which again

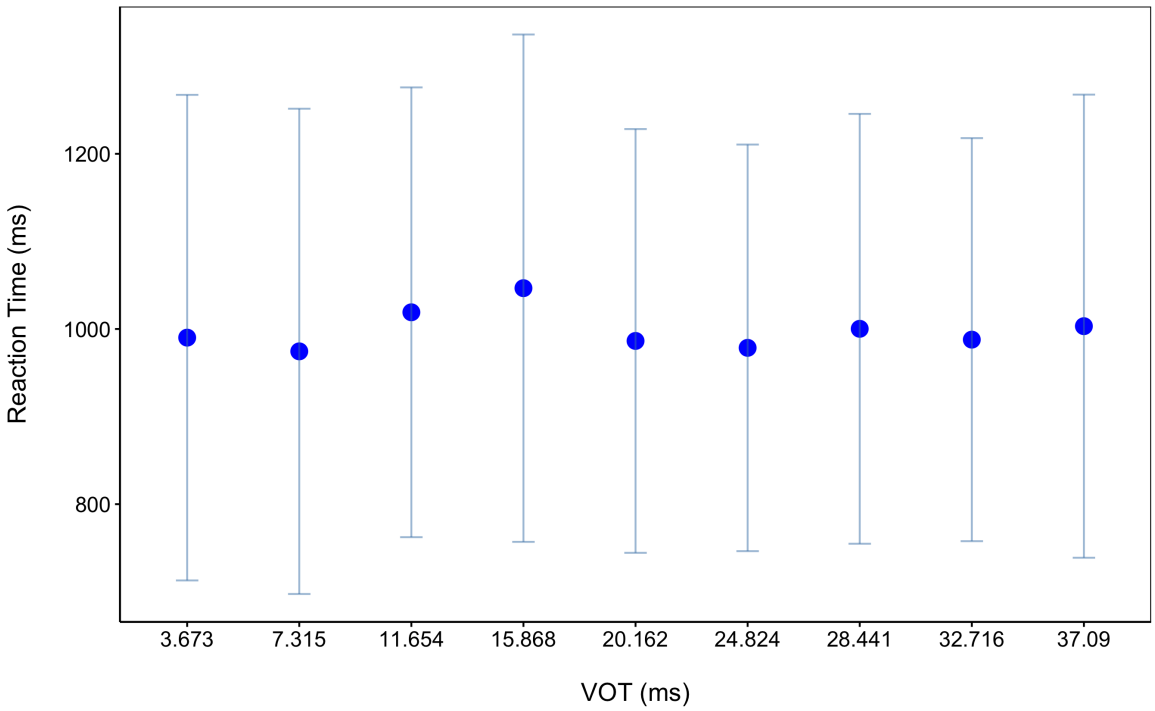


Figure 3.8.: Experiment 2: The mean across-subject reaction time values (in ms) along the y-axis, with stimulus VOT (in ms) on the x-axis. Error bars show standard deviations

corresponded to the stimulus closest to the across-subject category boundary. The "voiced" side of the continuum did not elicit monotonically rising RT values as stimulus VOT approached the boundary, since on average, the reaction time for the second step was shorter than that for the first step. Reaction times on the "voiceless" side of the continuum fell for the first two steps after the boundary, but subsequently rose twice more, on the seventh and ninth steps, respectively. Table 3.6 shows a summary of these values.

VOT (ms)	Mean (ms)	SD (ms)
3.673	990.12	277.20
7.315	974.46	277.02
11.654	1019.06	256.76
15.868	1046.59	289.62
20.162	986.35	241.91
24.824	978.47	232.10
28.441	1000.21	245.39
32.716	987.84	230.09
37.09	1003.21	264.37

Table 3.6.: Experiment 2: Across-subject mean reaction times per step of the VOT continuum

Curvature The Maximum Absolute Deviation for each stimulus step was calculated across subjects and trials. The results are depicted in figure 3.9. For this second experiment, MAD values ranged from 235.04 px ± 40.79 px for the ninth and last continuum step up to 297.56 px ± 88.04 px for the third step, closely followed by the fourth step (adjacent to the category boundary) at 292.74 px ± 63.95 px. Once more, there was no adherence to the expected pattern of excursion values, as there was an early (albeit small) decrease from the first step to the second step on the "voiced" side of the VOT continuum, and another, more pronounced, late rise reaching 281.89 px ± 97.03 px for the third-to-last stimulus step, well into the "voiceless" side of the stimulus spectrum. Table 3.7 provides an overview of these data.

Trajectory distribution With a value of 0.544, the bimodality coefficient computed for the trajectory data from the second experiment turned out to fall short of the

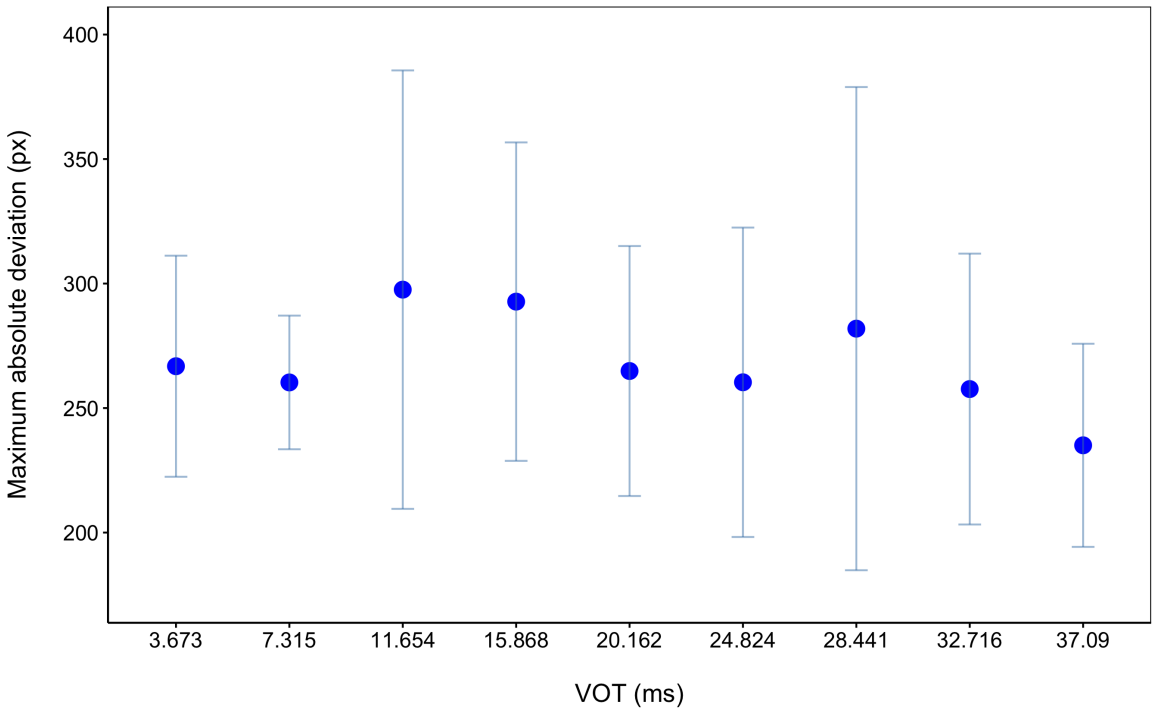


Figure 3.9.: Experiment 2: The mean across-subject MAD values (in px) along the y-axis, with stimulus VOT (in ms) on the x-axis. Error bars show standard deviations

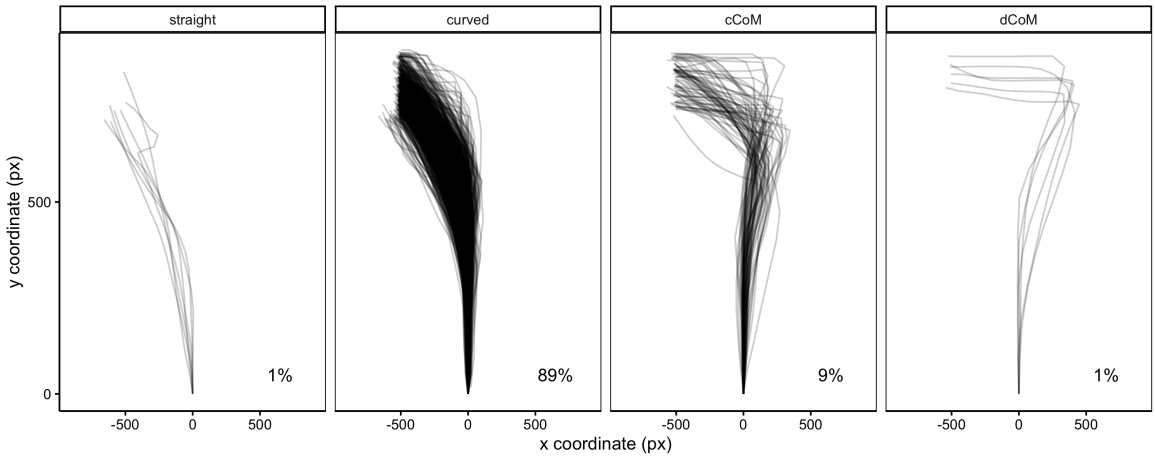


Figure 3.10.: Experiment 2: Proportions of trajectory shapes as mapped to a set of four prototypes

VOT (ms)	Mean (px)	SD (px)
3.673	266.80	44.40
7.315	260.30	26.83
11.654	297.56	88.04
15.868	292.74	63.95
20.162	264.88	50.19
24.824	260.36	62.13
28.441	281.89	97.03
32.716	257.63	54.39
37.09	235.04	40.79

Table 3.7.: Experiment 2: Across-subject mean MAD per step of the VOT continuum

assumed cutoff point for unimodality at 0.555. Upon visual inspection of the trajectory heatmap (see figure A.4 in appendix A), no obvious signs of bi- or multimodality could be detected, although there seemed to be a larger amount of trajectories resembling the "change of mind" types compared to the first experiment. Subsequent matching of the observed trajectories to the same prototypes as used in the first experiment again showed a majority of "curved" trajectories (89%), almost no "straight" and DCOM trajectories (1% each), and a proportion of 9% for the CCOM type. Once more, the prototype mapping process seemed to have captured the actual trajectory shapes well. "Change of mind" trajectories did not occur more often for the more ambiguous stimulus steps, with the highest CCOM proportion (14%) occurring for the third continuum step, and the lowest proportion (3.3%) for the ninth step, with nonlinear trends along both sides of the VOT spectrum, as can be seen in table 3.8 (see figure A.5 in appendix A for a graphical representation).

Discussion

In the second experiment, participants' decision outcomes in this experiment were once more in agreement with the conventional view on categorical perception, showing little within-category identification variance, but favouring a clear between-category distinction. While reaction times exhibited a pattern of values along the VOT spectrum which was not entirely dissimilar from the one found by Pisoni & Tash (1974), including an RT peak occurring for the boundary-adjacent stimulus, the nonlinear trends on both

VOT (ms)	straight	curved	cCoM	dCoM
3.673	1.1 %	89.1 %	9.8 %	0.0 %
7.315	1.1 %	91.2 %	6.6 %	1.1 %
11.654	0.0 %	83.9 %	14.0 %	2.2 %
15.868	3.2 %	83.9 %	12.9 %	0.0 %
20.162	0.0 %	87.8 %	12.2 %	0.0 %
24.824	1.1 %	92.6 %	4.3 %	2.1 %
28.441	1.1 %	85.9 %	12.0 %	1.1 %
32.716	1.1 %	90.5 %	8.4 %	0.0 %
37.09	0.0 %	96.7 %	3.3 %	0.0 %

Table 3.8.: Experiment 2: Proportions of trajectory types per stimulus VOT. Respective percentages may not sum to exactly 100% due to rounding

sides of the voicing spectrum did not match the pattern which was expected for RT. Neither did mouse movement curvature, observed in terms of MAD, track the expected pattern of values, as the largest deviation did not occur for the stimulus closest to the boundary, and the nonlinearity of trends on either side of the VOT continuum was similarly pronounced as that exhibited by reaction times. Nevertheless, the distribution of mouse trajectories consisted largely of smoothly curved shapes, as the gradiency hypothesis predicts. Since there were virtually no "straight" and DCOM type trajectories present, the distribution was found to be effectively made up of only two types, "curved" and CCOM. This fact is reflected in the computed bimodality coefficient (0.544) coming close to the value of 0.555, which is accepted as the threshold for bimodality.

Overall, the data elicited by use of a design which refrained from any manipulation of the mouse settings appeared to be consistent with the previous experiment's results. Notably, another indeterminate series of curvature values emerged, possibly indicating the erroneousness of the working hypothesis after which the artificial deceleration of arm movements was thought to be capable of "masking" any reflections of gradient sensitivity to within-category variation in the motor output. Despite their general similarity, the following section will illustrate a number of ways in which the results from the two experiments differed.

3.4. Between-experiment comparison

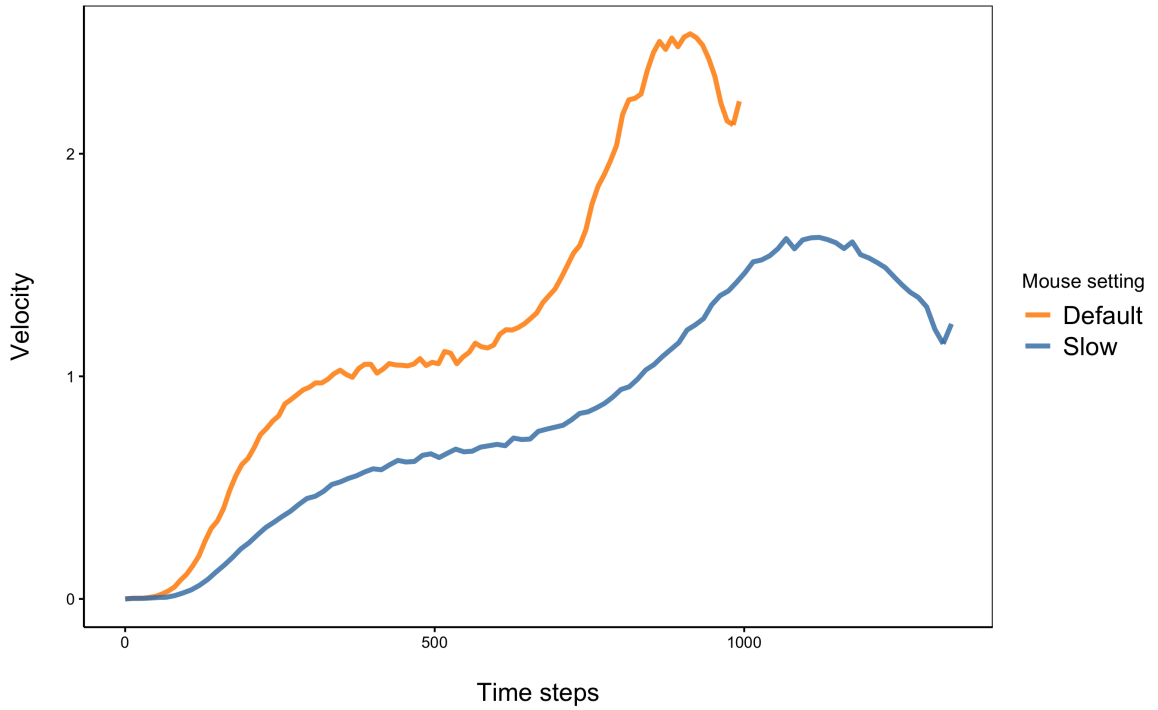


Figure 3.11.: Velocity profiles for both experiments, with time steps on the x-axis (each step corresponding closely to one millisecond), and velocity as the Euclidean distance (in px) travelled by the mouse pointer between two adjacent samples

Velocities and reaction times By design, the mouse movement data from the two experiments conducted for the present thesis were firmly expected to contrast with regards to their respective average velocity. Since the first experiment slowed the mouse pointer considerably and the second experiment did not, participants of the latter should have effected faster pointer movements overall. Figure 3.11 shows that this was indeed the case. Pointer movements in the *Default* mouse setting exhibited higher average velocities along the whole of the trajectories' time course, consequently ending earlier than those from the *Slow* setting. The mean velocity in the *Default* setting[4] was found to have been 1.2 ± 0.86, whereas the mean velocity in the *Slow* setting was 0.84 ± 0.59.

Upon visual examination, the velocity profiles for both experiments showed a distinct

[4]Expressed as the pixel-based Euclidean distance between the mouse pointer coordinates of two adjacent samples

and "typical" (Wulff et al., 2019) pattern of velocity values which quickly rise during the phases of movement initiation and target approach, while this velocity gain decreases for the medial part of the trajectory. For the *Default* setting, this phenomenon was pronounced more strongly than it was for the *Slow* setting, with steeper rises in velocity and narrower portions of transition between the three main phases.

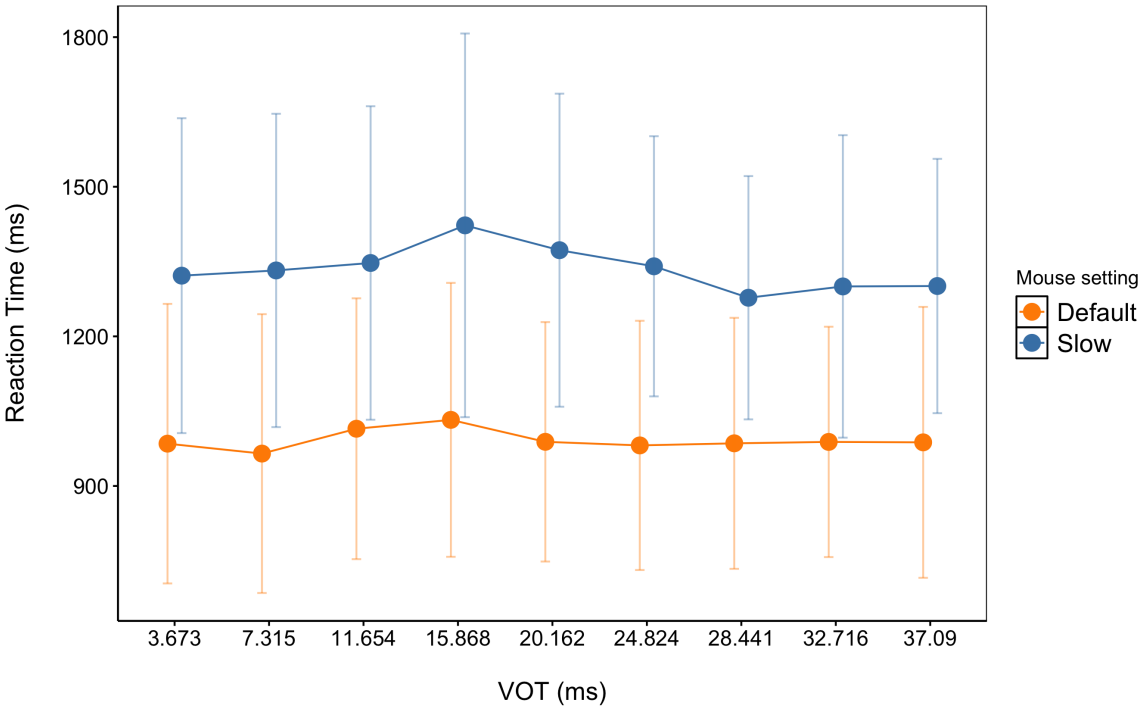


Figure 3.12.: Mean reaction times (in ms) along the y-axis, with stimulus VOT on the x-axis. Error bars show standard deviations. The graphical elements are offset horizontally to prevent overlap

A comparison of the mean reaction times that occurred in the two experiments shows that participants consistently responded faster when the mouse preferences allowed for fast responses, i.e. in the *Default* setting, as can be seen in figure 3.12. The mean overall RT in the *Default* setting was calculated to be 998.48 ms ± 257.16 ms, while participants on average took 1341.62 ms ± 307.27 ms to complete their response in the *Slow* setting (see table B.4 in appendix B for a by-stimulus comparison of reaction times).

Correctness and malformed trajectories With regards to the main experimental task of phoneme identification, a difference in the proportion of individually correct

responses could be observed in the resulting data. To reiterate, responses were classified as incorrect for those trials in which participants heard a stimulus on the "voiced" side of their individual category boundary, but still selected the "voiceless" response option (or vice versa). While participants chose the incorrect option in 4.9% of trials in the *Slow* setting, they responded incorrectly in 6.7% of trials in the *Default* setting. A slight decrease in identification performance such as the one found here may be expected for a faster mouse setting which also accepts as an identification response any movement of the mouse pointer onto either response button, without requiring a click.

By a similar token, a greater number of "malformed" trajectories might be predicted for faster mouse settings, as any directional changes in the mouse movement would get strongly emphasised by the acceleration curve, and participants would have less time to complete their decision process before the mouse pointer reached the upper display edge.[5] The data collected here do not agree with this prediction, however, as there were 11.5% of trajectories which had to be classified as malformed in the *Slow* setting, and only 6.2% of malformed trajectories in the *Default* setting.

Curvature and distribution of types The amount of trajectory excursion toward the non-chosen response was consistently higher when participants could displace the mouse pointer quickly than when the pointer was slowed, as figure 3.13 shows. The overall mean MAD for the *Slow* setting was determined to be 217.49 px ± 62.54 px, while the *Default* setting elicited an averaged MAD of 268.58 px ± 58.64 px (see table B.5 in appendix B for a by-stimulus comparison of deviation values).

While the trajectory distributions of both experiments were classified as unimodal by calculation of the bimodality coefficient, the coefficient was lower for the *Slow* setting (0.475) than it was for the *Default* setting (0.544). A comparison of the proportions of trajectory types in both experiments (see figure 3.14) may aid in clarifying this difference, as it reveals that the "straight" type, which does make up a small portion (8%) of trajectories in the *Slow* setting, is almost completely absent (1%) in the *Default* setting, while at the same time, no salient other types emerge. Instead, as pointed out earlier, the distribution resulting from the *Default* setting is comprised of virtually only two types, which together account for 98% of trajectories (see table B.6 in appendix B for a by-stimulus comparison of type proportions).

[5] Assuming they successfully follow the instruction not to interrupt their movement of the mouse

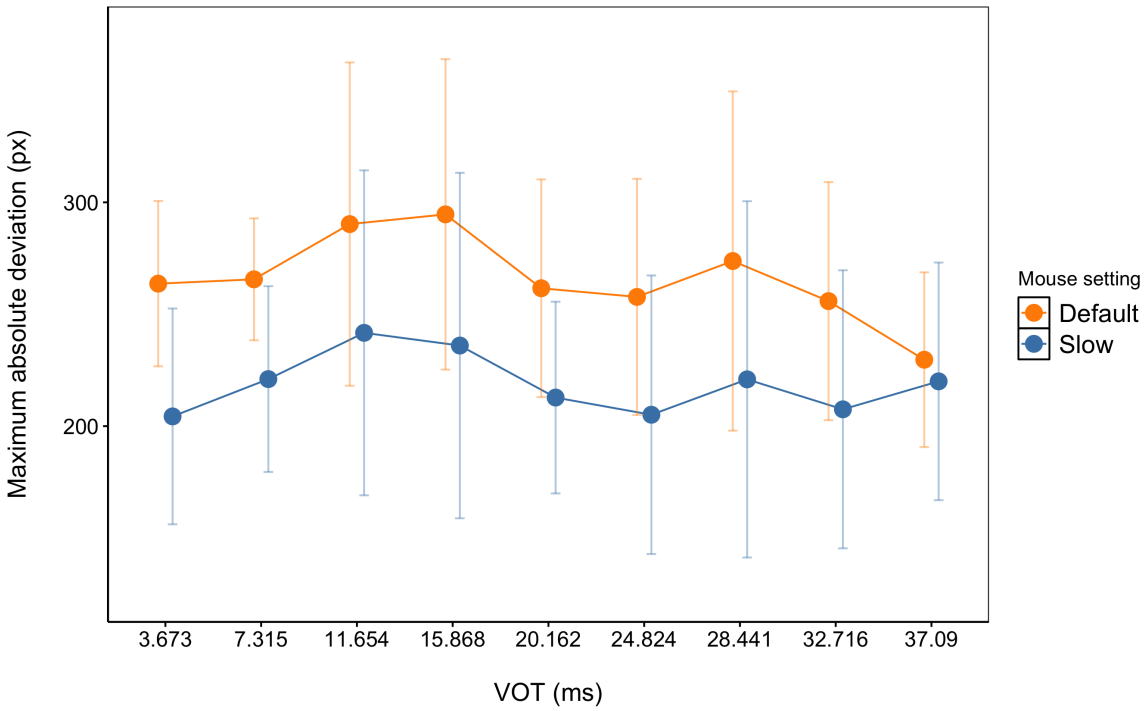


Figure 3.13.: Mean MAD (in px) along the y-axis, with stimulus VOT on the x-axis. Error bars show standard deviations. The graphical elements are offset horizontally to prevent overlap

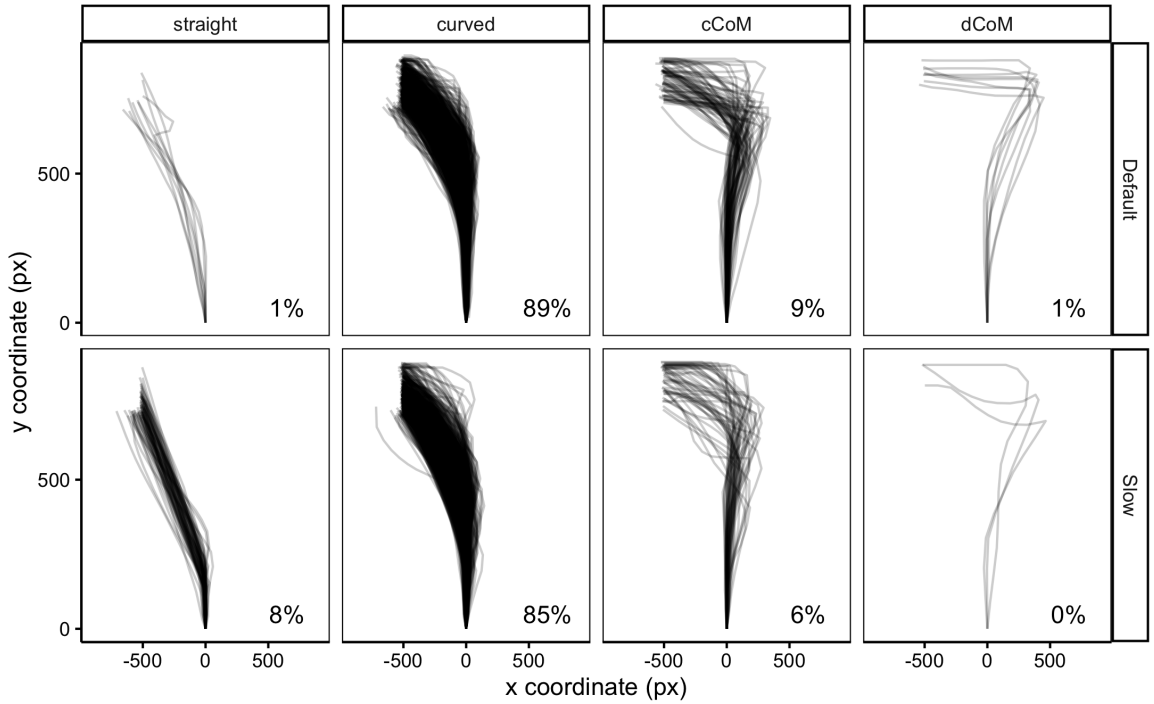


Figure 3.14.: The proportions of classified trajectory types for both experiments

Discussion While it would have been surprising to see no effect of the mouse sensitivity setting on the actual movement velocities, it is (somewhat counterintuitively) not quite as much of a foregone conclusion to have found that reaction times were shorter in the *Default* setting and longer in the *Slow* setting. In the mouse tracking paradigm, reaction time is dependent not only on movement velocity, but also on the trajectory curvature: For any given velocity value, a curved trajectory will take longer to complete than a straight trajectory. The first experiment severely limited mouse pointer speed, and hence it would have taken participants a considerable amount of motor effort in order to arrive at reaction times similar to the ones observed for the second experiment, but nevertheless, the mouse setting did not prohibit such behaviour in principle. Moreover, there a number of studies in which deviation effects were found to be dissociable from reaction times, e.g. by being observed without simultaneously occurring reaction time effects (see Freeman (2018) for an overview). Still, the reduction of mouse pointer sensitivity to 75% of the default value in the present study appears to be tracked closely by the overall mean reaction times, as mean RT in the *Slow* setting reached 74.4% of mean RT in the *Default setting.*

It stands to reason that the combination of a more easily displaced mouse pointer

and being instructed to keep mouse movement swift and constant may have contributed to the fact that participants consistently produced trajectories with stronger curvature in the *Default* setting. At the end of a given amount of completion time of a cognitive decision process, the mouse pointer will have travelled further along the vertical (if so instructed) if mouse sensitivity is high and/or acceleration is enabled, resulting in a greater overall excursion of the trajectory once it is completed. Thus, a component of a trajectory's curvature value always depends on the geometry of the experimental display, any movement instructions participants might have received, and the movement mechanics as modulated by the mouse settings.

The same reasoning may be applicable to the difference in proportion of "straight" trajectories in the two experiments: In the *Default* setting, participants' early, constant, and swift displacement of the mouse pointer quickly comes to preclude the occurrence of this type of trajectory, as the mouse pointer will have already travelled far enough along the vertical to eventually result in a curved trajectory, even when the decision processes are concluded relatively early.

4. General discussion

The present study attempted a replication of findings by McMurray et al. (2008) by making use of the mouse tracking paradigm to collect on-line data from the dynamics of participants' hand motions while they listened to VOT-manipulated stimuli. In a set of eye tracking experiments, McMurray et al. investigated whether listeners exhibit systematically graded sensitivity to acoustic information within their phoneme categories, and obtained evidence in support of this gradiency hypothesis, in turn suggesting that the processes underlying speech perception themselves are of a graded and continuous nature. In contrast to stage-based accounts of cognitive processing, continuity-based approaches such as theirs posit processes of continuously evolving conflict between partial activations of response options as the framework in which decisions—like phoneme identification—are to be understood, and on-line motor output data is taken to be indicative of those partial activations during the conflict resolution. While eye tracking studies such as the one by McMurray et al. are chiefly interested in fixation locations, durations, and proportions, the mouse tracking experiments presented here mainly focused on the relative curvature observed in mouse trajectories on their way to the chosen response.

Within the small samples ($n = 10$) of both exploratory experiments conducted here, no evidence for systematically gradient sensitivity to VOT durations within participants' categories could be observed. Not only did trajectory curvature fail to exhibit a consistent pattern in favour of any such hypothesis, but also did participants' reaction times. Seeing as the identification results from both experiments seem to suggest that the stimuli used here were valid, the negative results concerning RTs are particularly surprising, since they amount to a failure to replicate a well-established effect. Once more, if these current results were to be replicated in a sufficiently powered study, their divergence from the findings of McMurray et al. would necessitate a critical re-examination of that preceding study, as well as of the methodological details of the mouse tracking paradigm as presently implemented. On the basis of the conjecture that a statistically meaningful replication of the present results would indeed be possible,

the following sections will offer such critical assessments.

4.1. Revisiting McMurray et al. (2008)

Relative VOT With the express purpose of limiting the probability of spurious findings in their data, McMurray et al. employed two specific strategies in the data processing stage of their study. Firstly, they used the calculated per-subject category boundaries to standardise the VOT spectrum for each participant, so that the resulting independent variable (termed "relative VOT" or rVOT) had negative values for VOT durations on the individually "voiced" side, and positive values on the individually "voiceless" side of the spectrum:

> "It is important to note that rVOT analyses are more conservative than analyses that ignore the mouse-click category boundary. By using rVOT, variability in category boundary across participants and continua is effectively eliminated as a potential source of gradient effects. This avoids a well-known problem in the analysis of binomial data in which the average of a number of steep logistic functions with different category boundaries is shallower than any of the contributing functions." (McMurray et al., 2008)

Secondly, they expanded on this notion of avoiding artefacts that might suggest continuity in the data where none exists by excluding from their analyses data which resulted from stimulus VOTs adjacent to the individual category boundary:

> "To verify that [the effects of VOT duration on fixation proportions] were not primarily due to tokens that were immediately adjacent to the category boundary, these two tokens were removed and a second set of analyses was conducted using the remaining data (seven steps rather than nine steps)." (ibid.)

The analyses presented in sections 3.2 to 3.4 of the present thesis did not use rVOT as a predictor variable, but instead investigated "raw", non-standardised VOT. Moreover, they refrained from the exclusion of boundary-adjacent tokens. This kind of "anti-conservative" treatment of the data would not have been able to inhibit the appearance of any false indicators of gradedness (as anticipated by McMurray et al.) in the analysis stage. To the negligible extent that such indicators did show up in the

results[1], their meaningfulness should be assessed with some reservation, as they might well represent such artefacts as described by McMurray et al. Indeed, a secondary analysis of across-subject trajectory curvature employing the suggested conservative measures did not alter the principal findings of either mouse tracking experiment (see figures A.6 and A.7 in appendix A for a depiction of mean curvature values of each experiment, using rVOT, and with values resulting from tokens less than 5 ms from the category boundary removed). The present study was, in other words, unable to detect gradient sensitivity to within-category acoustic detail even when no particular precautions designed to avoid gradiency-suggesting artefacts were implemented during the analysis stage.

Limitations of the original evidence In order to examine whether evidence for gradient listener sensitivity could be gathered, McMurray et al. (2008) devised and conducted five experiments in total. These experiments looked at VOT variations in minimal pairs from natural speech (e.g. "beach" and "peach"), and in synthetically produced stop-vowel-syllables. Lexical and phoneme identification tasks were also varied with regards to how many response options were available to participants (two vs. four options). The analyses of each experiment's data were then performed separately for each side of the stimulus spectrum, assessing whether there was a linear trend of rising fixation durations (or fixation proportions) on the "voiced" side of the spectrum, and whether the reverse was true for the "voiceless" side of the spectrum, i.e. whether a linear trend of diminishing values could be observed. Moreover, if (significant) instances of such trends appeared in the analysis stage *before* conservative analytic measures[2] were taken, the authors chose to report those alongside the end results of the respective analyses.

The multitude of partial test results generated by this approach do indeed include a number of indications in favour of the gradiency hypothesis. For instance, there was "clear evidence for gradient effects on the voiceless side of the continuum but not on the voiced side" (McMurray et al., 2008) in their second experiment (which was the template for the empirical effort of the present thesis). Nonetheless, in some cases, the authors found no systematic effect of VOT variation on either side of the stimulus continuum, while in others, their data appeared to exhibit some such effect in principle,

[1] E.g. peak RT appearing for the boundary-adjacent stimulus VOT in the second experiment

[2] As suggested by McMurray et al., i.e. calculation of rVOT and exclusion of immediately boundary-adjacent tokens

but it reached significance only on the voiceless side. Summarising their results as viewed through the lens of "the most stringent test of gradiency [where] the most ambiguous tokens [...] are eliminated", McMurray et al. state that "[o]n the voiced side, the main effect of rVOT and the linear trend analyses did not reach significance for any of the four experiments", while on the voiceless side, their results reliably indicated such effects in two of their five experiments.

Moreover, eye tracking data which are taken to be in support of continuity-based theories of cognitive processes—such as the gradiency hypothesis—should be approached with some caution, as they fail to deliver an authentically continuous window into these processes and thus may be susceptible to the emergence of aggregation artefacts (see section 1.3).

Viewed in this light, the failure to find evidence for gradient sensitivity to subphonemic variation in the present experiments may appear less surprising. Although, in principle, the gradiency hypothesis could still be true, there does not seem to be sufficient evidence yet to be able to uphold its strong formulation, which not only expects *any* indicators of within-category gradient effects, but postulates *monotonic* trends for these effects to be considered "truly gradient" (McMurray et al., 2008).

4.2. Methodological concerns

Movement initiation parameters Both mouse tracking experiments were designed with a dynamic movement initiation procedure. Participants were instructed to initiate the mouse movement early and to keep it constant, and stimulus playback was triggered automatically by cursor movement, in an attempt to ensure that arm/hand movements were ongoing during the processes of perception and identification. This specific design choice may carry relevance for the gradiency hypothesis:

1. As described in section 2.4, this type of design has been shown (Kieslich et al., 2019a) to lead to trajectory distributions with majority proportions of smoothly graded trajectories (as opposed to designs employing other parameters of movement initiation), a finding closely tracked by the data discussed in this thesis.
2. Moreover, Scherbaum & Kieslich (2018) found lower indices of bimodality when movement initiation was designed to be dynamic, observing that the dynamic starting condition showed "a more coherent distribution of movements than the

static condition".
3. They also found significantly greater curvature values when movement initiation was dynamic.

The gradiency hypothesis as formulated by McMurray et al. (2008) makes several assumptions:

1. It assumes that continuously graded responses to continuous acoustic features of speech can be elicited even within participants' (phoneme) categories.
2. By extension, it also assumes the absence of bimodality in any dataset resulting from such responses.
3. Lastly, it assumes that in order to collect such data, experimental paradigms are chosen which do not impede or prohibit the occurrence, or the collection of, graded responses.

A comparison of the above lists reveals that none of the assumptions of the gradiency hypothesis could have been violated by the specific design choice of using a dynamic starting procedure, as unimodal distributions of smoothly curved trajectories and an amplification of curvature values are what distinguishes it from static movement initiation procedures. Even so, the present study was not able to find systematic differences in listeners' responses to sub-phonemic variation.

Response selection mode The fact that almost no DCOM trajectories were found in the present data may well be a result of the chosen response selection mode: Since the design did not require participants to finalise their response by performing a mouse click, all the trajectories resembling the DCOM prototype which participants may have performed simply would not have been entered into the dataset as DCOMs, but rather as "straight" trajectories (see section 2.4). This also means that, should the totality of participants' mouse movements from all trials in actuality have been bimodal (with one mode consisting of DCOMs), the resulting trajectory distribution still would not have shown this bimodality, likely appearing as unimodal instead.

In order to estimate whether the "hidden" occurrence of a bimodal distribution in this case should be deemed likely, it should be taken into account that actual DCOM trajectories might not only have been recorded as "straight" trajectories, but they might also belong to the set of trials excluded from the analysis as "individually incorrect". It is conceivable that participants tried to amend the fact of having arrived

at the wrong response location, but the rest of their corrective trajectory simply did not get recorded because of the response selection mode employed here.[3]

This artificial exclusion of potentially bimodality-inducing trajectory shapes can be seen as another "anti-conservative" measure[4] with regards to the gradiency hypothesis, as it facilitates the emergence of the kind of unimodal distribution containing a large proportion of "curved" trajectories which the hypothesis predicts.

Sensitivity and acceleration settings After the data resulting from the first mouse tracking experiment had been found to exhibit no signs of a gradient sensitivity as McMurray et al. (2008) describe it, a working hypothesis was formulated. This hypothesis stated that the significantly reduced speed of the mouse pointer may have "masked" the effects of such gradient sensitivity in the recorded arm/hand motions by rendering those motions too coarse to be able to reflect fine-grained aspects of the underlying processes of speech perception. However, since the second experiment arrived at similar results while leaving the sensitivity and acceleration settings at their system default values, the working hypothesis appears to have been inadequate. Consequently, it stands to reason that this specific design choice, too, did not on a methodological level prevent the postulated gradient sensitivity from showing up in the mouse tracking data.

Limitations of the present study As stated above, an examination of the data obtained through the mouse tracking experiments conducted for the present thesis must not be considered as a source for any generalising inferences about listener behaviour, since the sample size of both experiments is insufficient to fulfil such a purpose. The potential meaningfulness of these data may further restricted by the fact that all participants were trained phoneticians, who can reasonably be expected to be capable of guessing the investigative direction of an experimental design which exposes its subjects to /ba/ and /pa/ syllables. Even if this assumption is incorrect, a sample made up of individuals of any singular profession will not in any case be as representative of the general population as would be desirable. Moreover, six of the ten participants from the first experiment also took part in the second experiment,

[3]It is thus also possible that the choice of response selection mode shaped the identification results

[4]Note that the response selection mode in question was not chosen for this purpose, but rather in an attempt to minimise noise in the trajectory data. The mechanical act of clicking a spring-loaded mouse button may introduce such noise

similarly diminishing the validity of the resulting data. Both the small sample size and the repeated participation of individual subjects were deemed acceptable for the present purpose of conducting a "proof-of-concept" exploratory study as part of an attempt to devise a feature-complete implementation of a novel paradigm for use in speech perception research.

Activation thresholds Proponents of the eye tracking method have pointed to the special nature of eye movements as a noteworthy advantage for research into decision conflict and the underlying mental processes associated with it. Compared to the majority of other overt movements, which typically involve much larger, more inert parts of the body, eye movements are "extremely fast, quickly corrected, and metabolically cheap" (Spivey, 2007), as they require only a very small amount of spatial attention to get triggered (Kowler et al., 1995), and saccade planning takes a very short amount of time (Matin et al., 1993).[5] The combination of these properties has been argued to render eye movements uniquely sensitive to partially active mental representations:

> " [I]n terms of thresholds for executing motor movement, eye movements have an exceptionally low threshold for being triggered, compared to other movements. [...] Therefore, briefly partially active representations—that might never elicit reaching, speaking, or even internal monologue activity, because they fade before reaching those thresholds—can nonetheless occasionally trigger an eye movement [...]" Spivey (2007)

In other words, a given magnitude of activation belonging to a given mental representation may indeed suffice to trigger some types of associated (task-dependent) responses, while being incapable of triggering other types of responses. Thus, a scale of possible response types emerges, ordered by how easily they will cross the threshold for being triggered. This, in turn, means that in order to successfully obtain any information about the cognitive processes underlying representation activation by way of experimental elicitation of overt motor responses, two conditions have to be satisfied:

1. An appropriate experimental method has to be chosen, eliciting and recording a type of motor output which is located sufficiently low on the triggering scale, considering the hypothetically expected levels of activation.

[5] Around 200 ms

2. Appropriate stimulus material has to be chosen, capable of eliciting sufficiently high levels of activation, so that the respective triggering threshold may be crossed.

A failure to satisfy the first condition would result in only ever obtaining negative results, since the mismatch in scale between the method and the object of interest would prevent any experimental effects from showing up in the recorded data. A failure to satisfy the second condition would yield similarly negative results, as the mismatch between stimuli and triggering threshold would prevent any effects from occurring on the level of motor behaviour. Crucially, neither case would be capable of making any assertion about the nature or the existence of any underlying cognitive processes, which, in actuality, might well have been elicited and/or modulated by the experimental design, but either simply *were* not captured by the method, or *could* not have been captured by any method looking at motor output, since they did not cross the threshold into that domain.

While it seems intuitively plausible that ocular motion should appear on one of the lower levels of the "ease of triggering" scale, it is not as obvious exactly where other forms of movement, such as movements of the hands and arms, should fall in comparison. With regards to the present failure to replicate the eye tracking results of McMurray et al. (2008), it might thus be reasonable to assume that the mouse tracking method was not sensitive enough to capture whatever spillover-effect the processes of speech perception and phoneme identification may have had on the domain of motor output, while the purely cognitive effects of the VOT spectrum might still have been elicited in the participants. Note that this line of reasoning runs parallel to the "working hypothesis" which was formulated after the first experiment had observed no systematic pattern in trajectory curvature: This hypothesis roughly assumed that the more inert the targeted body parts are, the lower the chance to successfully observe fine-grained reflections of cognitive processing while examining motor output from these body parts. Notably, removing the artificial addition of inertia (i.e. the slowed mouse settings), still did not appear to make the mouse tracking method sensitive enough to capture the effects reported on by McMurray et al. (2008).

In the case of the present study, the replication failure may thus be explained by taking into account the failure to satisfy the first of the conditions outlined above. Nonetheless, it should be considered that this replication failure can, in principle, be explained equally well by assuming that the study failed to satisfy the second condition:

The very short, single-syllable speech stimuli used in the present experiments, and in particular their differences in VOT on the sub-phonemic scale, may simply be incapable of effecting activation levels high enough to cross over into arm movements. If this explanation should be accurate, no amount of cursor speed modification would have sufficed in order to successfully fine-tune the mouse tracking method—there would not have been anything for it to capture in the first place.

Seen in this light, McMurray et al. (2008)'s positive findings may appear immediately plausible: Since for eye movements, activation levels have a much lower threshold to cross than for arm movements, the same or very similar stimuli may succeed in eliciting motor output which systematically corresponds to underlying cognitive processes (satisfaction of the first condition). Naturally, however, this observation does not yet provide any answer as to exactly where the activation threshold for eye movements lies with regards to speech stimuli. In other words, it is not yet clear *in principle* whether the experimental design of McMurray et al. did satisfy the second condition.

Fortunately, there has been some recent work attempting to shed light on the question of how much information is needed to trigger saccades in the type of eye tracking design as employed by McMurray et al.[6] In two eye tracking experiments of their own, Teruya & Kapatsinski (2019) examined the effect that variations in the amount of linguistic information within the speech signal had on the probability with which participants fixated on a corresponding referent. They concluded that the threshold which activation levels of lexical semantic representations have to cross in order to trigger movements of the eyes is not as low as previously believed (e.g. Allopenna et al. (1998) and Tanenhaus et al. (2000)):

> "Our participants appear reluctant to fixate a picture unless the acoustic signal provides evidence for the initial CVC of the picture's name. As a result, participants look at pictures of unrelated distractors as much as they look at pictures of cohort competitors when the cohort competitor shares only the initial C or CV with the target word." (Teruya & Kapatsinski, 2019)

These findings provide yet another explanation for the null results of the present study: Its manipulated VOT durations exist on a level even below that of the CV-syllables used as stimuli in both mouse tracking experiments. However, if manipulations

[6] The "visual world paradigm"

on the CV-level are already insufficient for triggering saccadic eye movements, *sub*-CV differences cannot at all be expected to be capable of systematically influencing arm movements, which are thought to occupy a place higher up on the "triggering scale". Needless to say, these findings also appear to stand in stark contrast to the claims of McMurray et al., whose stimuli had similar properties.

4.3. Conclusion

The preceding sections have attempted to:

1. circumscribe the extent to which the proposal of the "gradiency hypothesis" as supported by the data from McMurray et al. (2008) should be approached with skepticism
2. compile valid explanations for the null result obtained by the mouse tracking experiments conducted for the present thesis
3. make possible a reasoned determination with regards to the effectiveness of mouse tracking as a tool for research into speech perception

Regarding the failed replication of McMurray et al. (2008), it could be shown that the choices made for each of the three major mouse tracking design options[7] have to be considered all but incapable of preventing any gradiency effects from showing up in the resulting trajectory data. Rather, they all seem capable of even facilitating the emergence of such effects. Similarly, the "anti-conservative" manner in which the resulting data were processed and analysed should not be expected to have suppressed the surfacing of gradiency effects. On these grounds, however, it still can not be ruled out that another, untested property of the nascent mouse tracking technique prevented a successful replication on the methodological level.

On the other hand, a critical assessment of the purported evidential strength in the findings of McMurray et al. showed that it is not altogether obvious how the multiple, partial, and variable findings in their study can be consolidated into the positive finding ultimately reported by the authors. Moreover, if further empirical support can be added to Teruya & Kapatsinski's recent conclusions, obtaining results comparable to the ones of McMurray et al. would appear to be exceedingly improbable even by use of the eye tracking paradigm itself.

[7] As discussed at length in section 2.4

It appears that experimental methods aiming to investigate the on-line reflections of manipulated cognitive processes in motor behaviour have to be conscious of two tightly integrated conditions both when designing an experiment as well as in the accompanying analysis stage. It further seems plausible that for each such method, there will be a range of possible stimulus-task-combinations which can be expected to harmonise[8] with that technique. This is the case when stimuli actually do cause sufficiently high activation levels so that *some type of* motor output might get triggered, and simultaneously, the experiment does aim to capture those types of motor output. Conversely, there will be stimulus-task-combinations which do not harmonise with particular experimental techniques. This is the case when the employed stimuli fail to elicit activation levels capable of triggering the type of motor output under observation (while they still might elicit other types).

As a consequence, two more possible explanations for the present study's null result arise. Firstly, assuming the sort of stimulus used in this study and in the one by McMurray et al. is indeed able to elicit activation levels high enough to become capable of triggering any motor responses, the mouse tracking method used here would have failed to satisfy the first condition simply by looking at motor output too high on the "triggering scale", as described above. Secondly, the stimuli used in both studies may in actuality be incapable of sufficiently high activation levels. This option, though it starts as a purely logical one, does turn out to receive empirical support by the findings of Teruya & Kapatsinski. While it may remain true that the present experimental design did not satisfy the first condition, it may *additionally* be the case that it failed to satisfy the second condition, expecting to observe reflections of cognitive processes in motor output, but simultaneously failing to provide stimuli capable of eliciting such reflections.

Unfortunately, this means that the effectiveness of the mouse tracking paradigm as an empirical tool for research into speech perception will likely be severely limited. Very short, sub-phoneme stimuli such as the ones used in the mouse tracking experiments discussed above will fall decidedly short of the CVC activation threshold posited by Teruya & Kapatsinski (and so will comparatively long stimuli such as CV-syllables themselves). Since the threshold they propose refers to the uniquely quick and easily triggered eye movements, however, it is difficult to see how any method investigating motions of the limbs could be successfully implemented for this particular kind of

[8] i.e. deliver meaningful data of the desired kind

research. Hence, the mouse tracking method may be restricted to research into other, higher-order linguistic phenomena, e.g. on the lexical or utterance level.

Future work Even if speech perception research should turn out to be inaccessible to mouse tracking, the present thesis hopes to have succeeded in accurately describing its general advantages. The fact remains that this inexpensive method to gather continuous, on-line data has already been successfully drawn on by multiple studies looking at various properties of (spoken) language, and at least as of yet, there do not appear to be any further methodological road blocks in sight which could prohibit future investigations into linguistic areas of interest. However, the preceding chapters also showed the extent to which any such future endeavours in mouse tracking should take great care to integrate into their experimental designs—as well as into their hypotheses—the profound interdependence of cognitive theories, low-level design choices, and analytic methods, lest they render their data uninterpretable, or worse, predetermine their outcomes.

In section 2.2, a range of possible mouse tracking measures was introduced. Still, there are numerous other measures that can be derived from mouse tracking data, and new analytic methods are currently being developed for, or adapted to, the mouse tracking method, including state space modelling (Calcagnì et al., 2019), decision landscapes (Zgonnikov et al., 2017), or linear discriminant analysis (Maldonado et al., 2018). This abundance of ways in which data from mouse trajectories can hold meaningful information may invite multiple analytic approaches "made in an ad hoc fashion, after having explored several aspects of the data and analytical choices" (Roettger, 2019). However, any such manner of exploiting the associated "researcher degrees of freedom", be it intentionally or not, "increases the likelihood of finding spurious significance" (Hehman et al., 2015) in the data at hand. In the case of the large set of multidimensional measures derivable from mouse movements, this caveat seems especially pertinent already. Still, its importance will only be heightened when the mouse tracking method is used for speech research, which faces "a high number of researcher degrees of freedom due to the inherent multidimensionality of speech behavior" (Roettger, 2019) as it is.

In addition to future linguistic use of the paradigm discussed here, it is vital that further methodological assessment of the paradigm's details be conducted in order to ensure the reliability and interpretability of any data obtained through it. Methodological work in this vein may include separate investigations of the effects

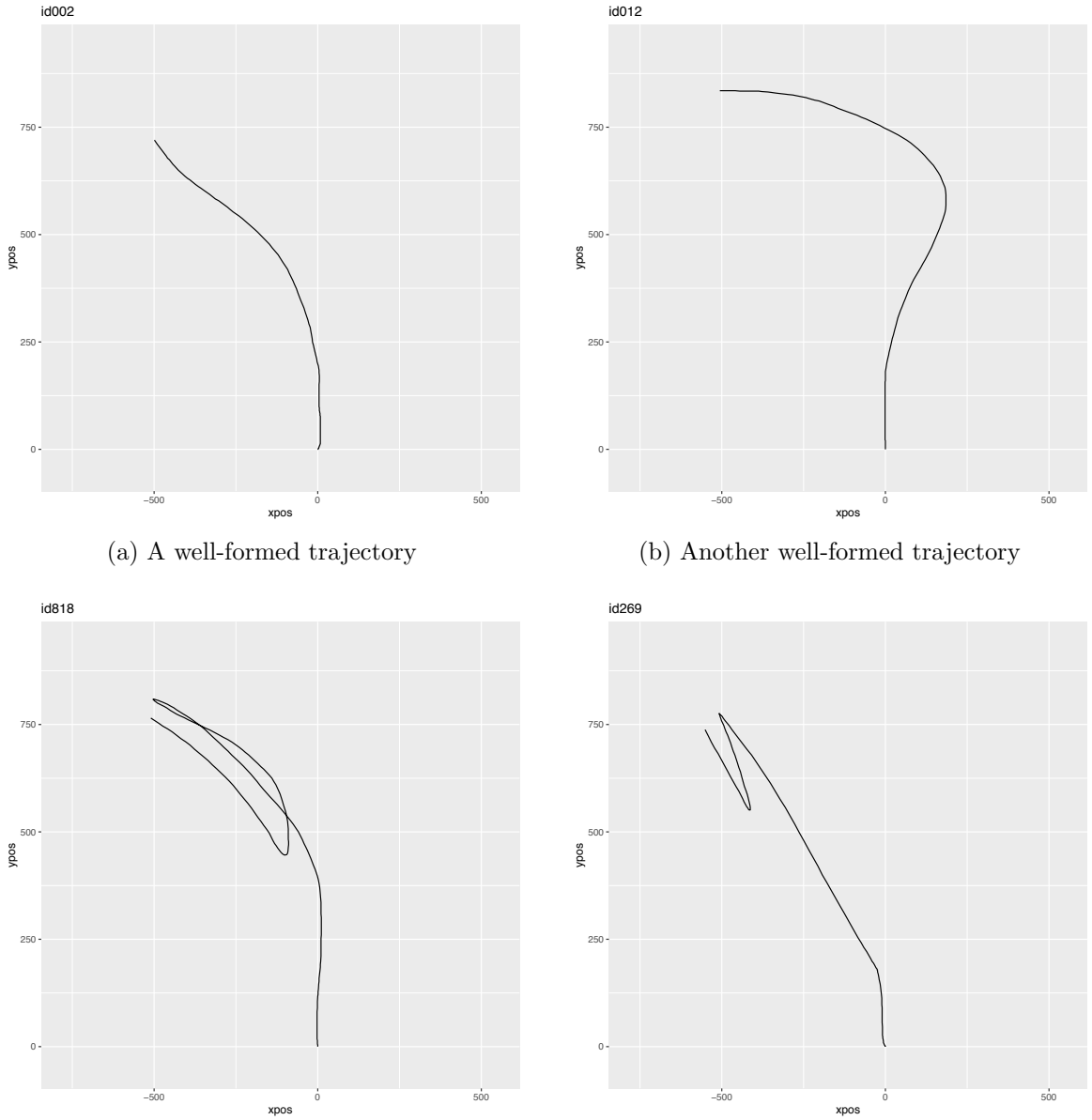
of mouse sensitivity and mouse acceleration, which up until now have only been manipulated in unison. Similarly, any possible effects of display aspect ratios as well as of the geometric properties of the graphical display elements (i.e. visual stimulus placement) have yet to be examined. The same is true for individual differences in mouse usage, including possible effects of handedness. On a more conceptual level, the examination of, for example, sequencing effects on trajectory distributions may be a promising area for future research, and so may the investigation of how task-dependent cognitive load might shape those distributions. It is conceivable, for instance, that a given participant's mouse trajectories will change systematically over the course of an experimental session as a function of how mentally taxing the task is for that participant.

While it is apparent that much clarifying theoretical and methodological work still remains to be accomplished, the mouse tracking paradigm is a promising addition to the speech researcher's toolbox, well-suited to the task of "opening up the response space" that is provided to participants in our experiments in order to gain deeper, more fine-grained, and more meaningful insights into the nature of the cognitive mechanisms at play.

Appendix

Appendix A.

Supplementary figures



(a) A well-formed trajectory

(b) Another well-formed trajectory

(c) A trajectory excluded due to looping

(d) A trajectory excluded due to backtracking

Figure A.1.: Examples of well-formed and malformed mouse trajectories.

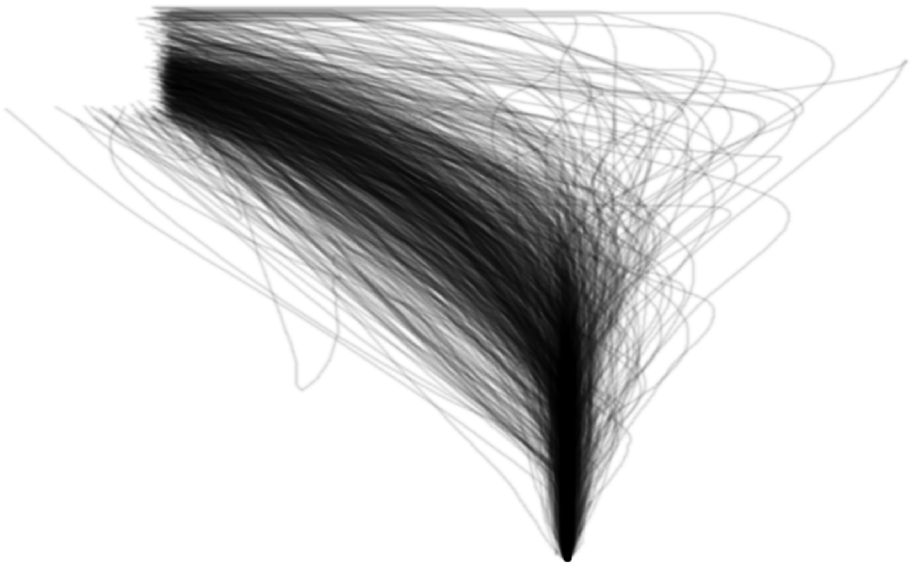


Figure A.2.: Experiment 1: Trajectory heatmap, with all mouse trajectories aligned on common start coordinates and flipped horizontally

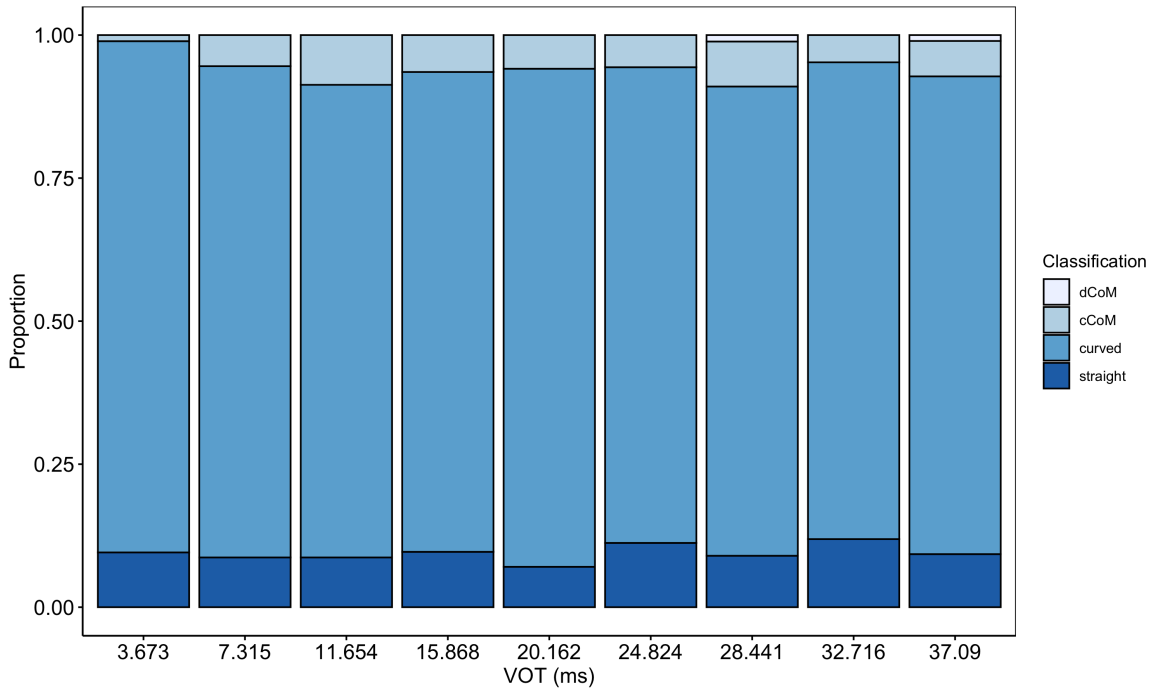


Figure A.3.: Experiment 1: Proportions of the four classified trajectory types on the y-axis, with stimulus VOT on the x-axis

Appendix A. Supplementary figures

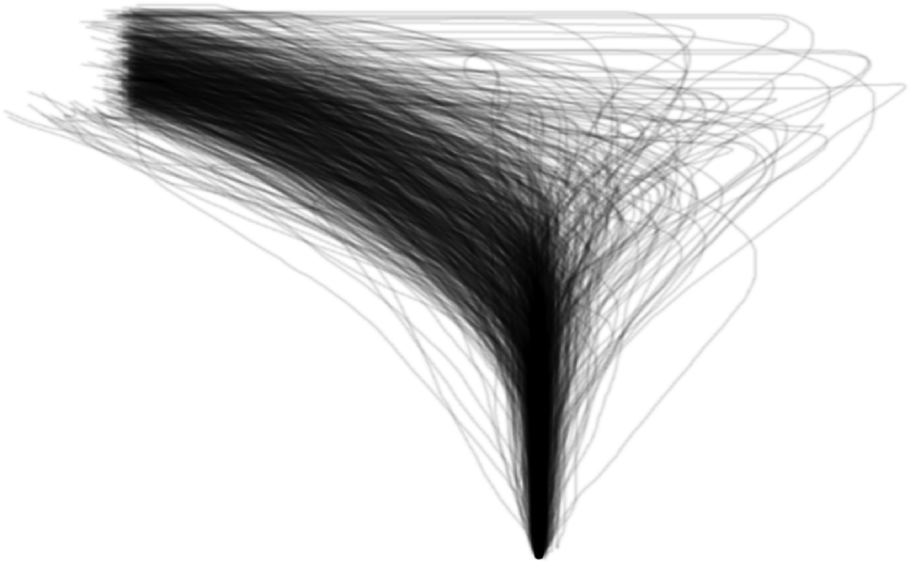


Figure A.4.: Experiment 2: Trajectory heatmap, with all mouse trajectories aligned on common start coordinates and flipped horizontally

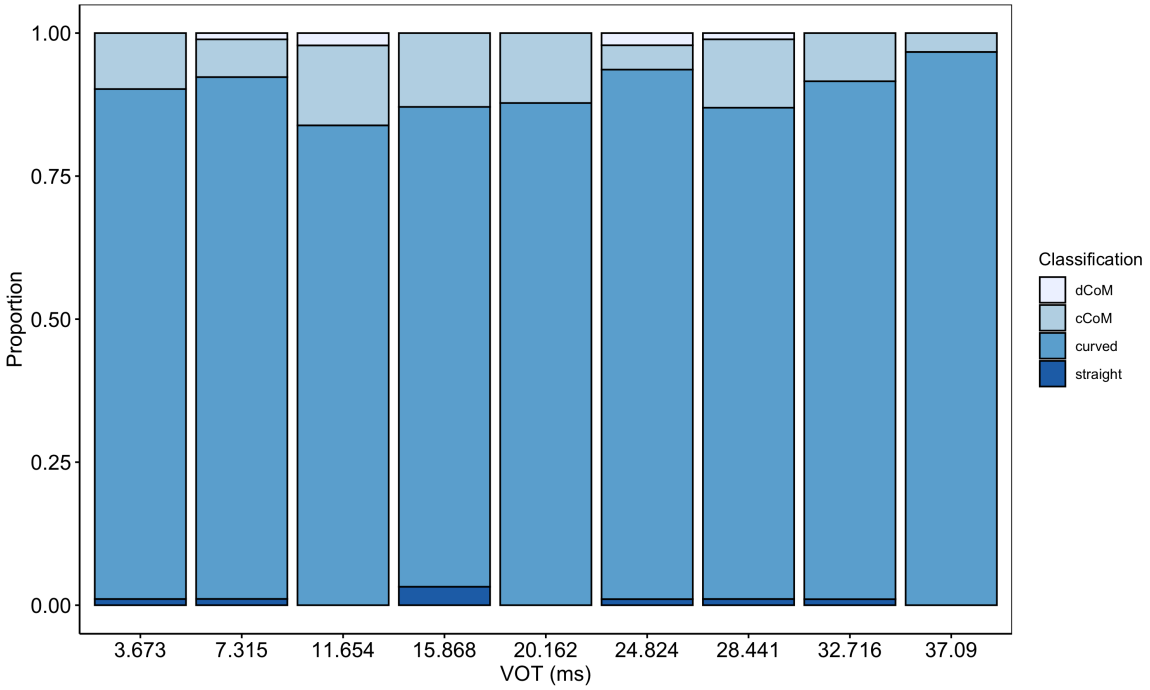


Figure A.5.: Experiment 2: Proportions of the four classified trajectory types on the y-axis, with stimulus VOT on the x-axis

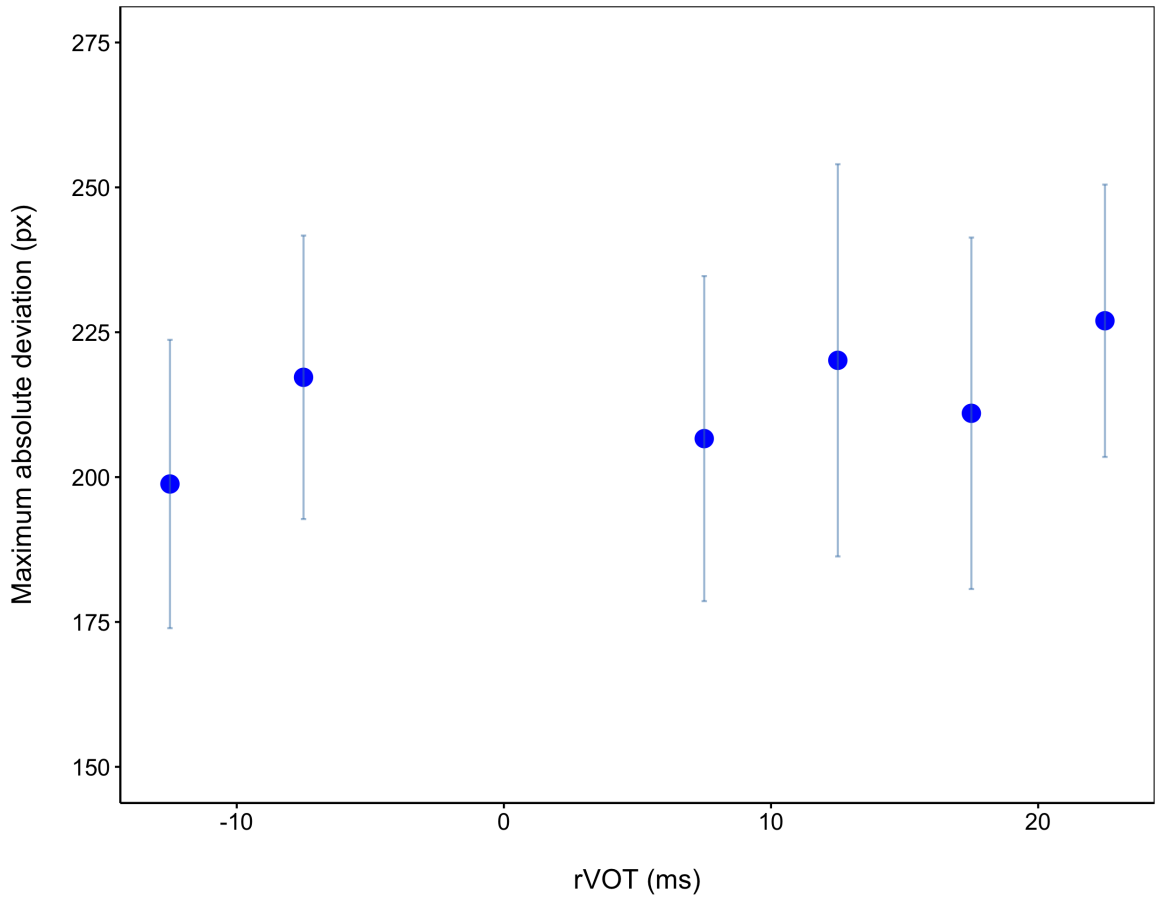


Figure A.6.: Experiment 1: The mean across-subject MAD values (in px) along the y-axis, with relative VOT (in ms) on the x-axis.

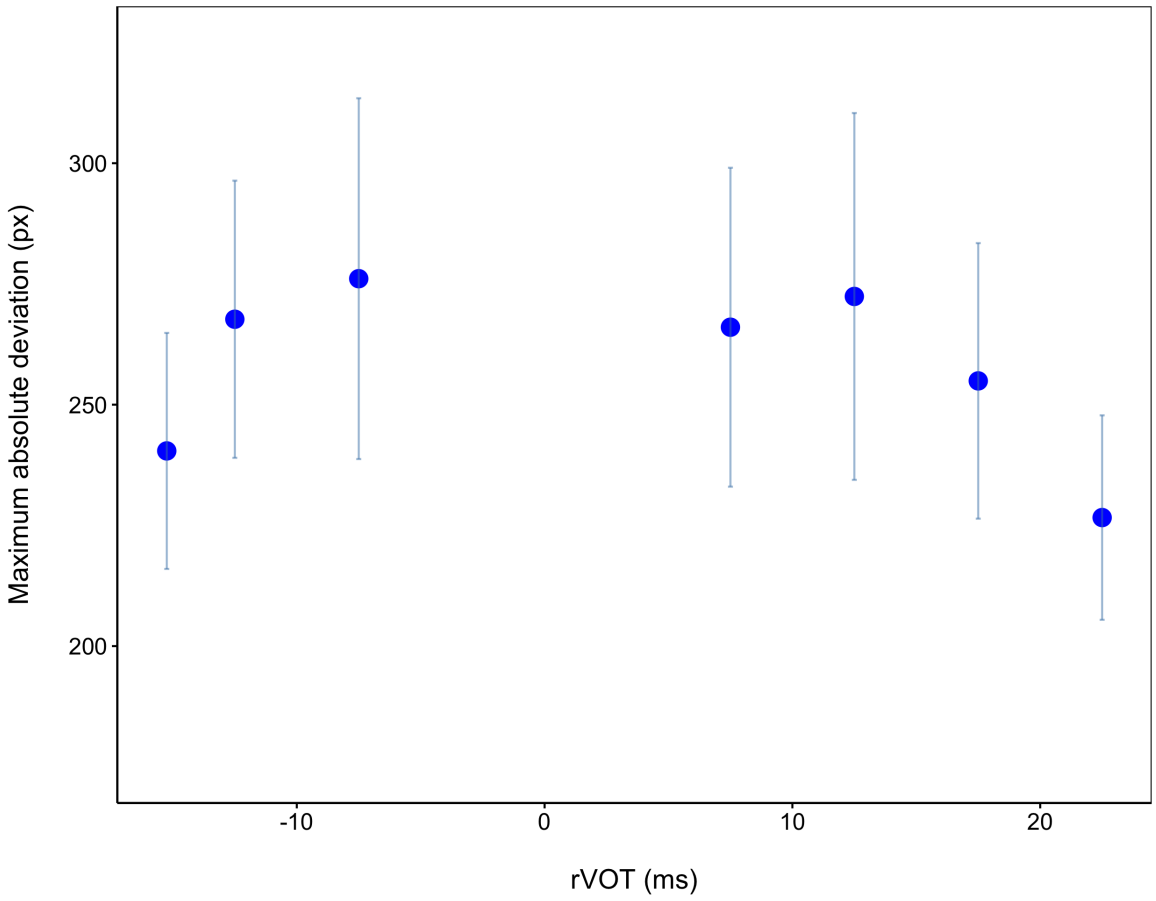


Figure A.7.: Experiment 2: The mean across-subject MAD values (in px) along the y-axis, with relative VOT (in ms) on the x-axis.

Appendix B.

Supplementary tables

Step no.	VOT	Step size	Stimulus
1	3.673 ms	0 ms	265.0 ms
2	7.315 ms	3.642 ms	268.6 ms
3	11.654 ms	4.339 ms	272.9 ms
4	15.868 ms	4.214 ms	277.2 ms
5	20.162 ms	4.294 ms	281.4 ms
6	24.824 ms	4.662 ms	286.1 ms
7	28.441 ms	3.617 ms	289.7 ms
8	32.716 ms	4.275 ms	294.0 ms
9	37.090 ms	4.374 ms	297.9 ms

Table B.1.: VOT duration, continuum step size, and stimulus duration for all nine steps.

Subject No.	Boundary (ms)	% /ba/	% /pa/
1	18.05	38.89	61.11
2	15.06	41.11	58.89
3	15.87	37.78	62.22
4	15.05	36.67	63.33
5	16.04	36.67	63.33
6	18.47	42.22	57.78
7	15.78	35.56	64.44
8	16.15	44.44	55.56
9	15.59	38.89	61.11
10	15.87	45.56	54.44

Table B.2.: Experiment 1: Estimated category boundaries by subject, and individual proportions of /ba/- or /pa/-responses

Subject No.	Boundary (ms)	% /ba/	% /pa/
1	17.12	38.89	61.11
2	18.92	37.78	62.22
3	15.59	42.22	57.78
4	15.47	44.44	55.56
5	15.9	35.56	64.44
6	15.88	38.89	61.11
7	14.15	33.33	66.67
8	18.02	34.44	65.56
9	13.68	46.67	53.33
10	14.19	34.44	65.56

Table B.3.: Experiment 2: Estimated category boundaries by subject, and individual proportions of /ba/- or /pa/-responses

	RT Slow		RT Default	
VOT (ms)	Mean (ms)	SD (ms)	Mean (ms)	SD (ms)
3.673	1321.72	315.71	984.86	280.29
7.315	1332.20	314.15	964.95	279.53
11.654	1347.15	314.29	1014.78	261.51
15.868	1422.61	384.69	1032.59	274.62
20.162	1372.79	313.88	988.56	240.06
24.824	1340.50	260.84	981.36	249.87
28.441	1277.44	243.92	985.53	251.66
32.716	1300.13	303.25	988.38	230.96
37.09	1300.95	255.00	987.48	271.62

Table B.4.: Mean reaction times per stimulus VOT and experiment

	MAD Slow		MAD Default	
VOT (ms)	Mean (px)	SD (px)	Mean (px)	SD (px)
3.673	204.35	48.22	263.67	36.94
7.315	221.02	41.51	265.60	27.24
11.654	241.69	72.62	290.28	72.24
15.868	236.01	77.20	294.63	69.36
20.162	212.77	42.84	261.61	48.63
24.824	205.09	62.23	257.75	52.76
28.441	220.91	79.61	273.78	75.80
32.716	207.52	62.13	255.87	53.21
37.09	220.01	53.09	229.68	39.05

Table B.5.: Mean Maximum Absolute Deviations per stimulus VOT and experiment

Setting	VOT (ms)	straight	curved	cCoM	dCoM
Slow	3.673	9.0 %	90.0 %	1.0 %	0.0 %
Slow	7.315	8.2 %	85.7 %	6.1 %	0.0 %
Slow	11.654	8.1 %	81.8 %	10.1 %	0.0 %
Slow	15.868	8.2 %	84.5 %	6.2 %	1.0 %
Slow	20.162	5.2 %	88.7 %	6.2 %	0.0 %
Slow	24.824	10.1 %	84.8 %	5.1 %	0.0 %
Slow	28.441	10.0 %	82.0 %	7.0 %	1.0 %
Slow	32.716	8.2 %	87.8 %	4.1 %	0.0 %
Slow	37.09	9.3 %	83.5 %	6.2 %	1.0 %
Default	3.673	1.0 %	88.9 %	10.1 %	0.0 %
Default	7.315	1.0 %	89.8 %	8.2 %	1.0 %
Default	11.654	0.0 %	85.0 %	13.0 %	2.0 %
Default	15.868	2.1 %	84.4 %	11.5 %	2.1 %
Default	20.162	0.0 %	88.9 %	11.1 %	0.0 %
Default	24.824	2.0 %	91.9 %	4.0 %	2.0 %
Default	28.441	2.0 %	85.9 %	11.1 %	1.0 %
Default	32.716	1.0 %	90.0 %	9.0 %	0.0 %
Default	37.09	0.0 %	96.9 %	3.1 %	0.0 %

Table B.6.: Proportions of classified trajectory types per stimulus VOT and experiment. Respective percentages may not sum to exactly 100% due to rounding

Bibliography

Allen, J. Sean & Joanne L. Miller (2001). "Contextual Influences on the Internal Structure of Phonetic Categories: A Distinction between Lexical Status and Speaking Rate". In: *Perception & Psychophysics* 63.5, pp. 798–810. DOI: 10/bmhbq6.

Allopenna, Paul D., James S. Magnuson, & Michael K. Tanenhaus (1998). "Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models". In: *Journal of Memory and Language* 38.4, pp. 419–439. DOI: 10/d3bqsc.

Andruski, Jean E., Sheila E. Blumstein, & Martha Burton (1994). "The Effect of Subphonetic Differences on Lexical Access". In: *Cognition* 52.3, pp. 163–187. DOI: 10/fvpv29.

Boersma, Paul & David Weenink (2016). *Praat: Doing Phonetics by Computer*. Version 6.0.22. URL: http://www.praat.org.

Calcagnì, Antonio, Luigi Lombardi, & Marco D'Alessandro (2019). "A State Space Approach to Dynamic Modeling of Mouse-Tracking Data". In: *arXiv preprint*. arXiv: 1907.08387. URL: http://arxiv.org/abs/1907.08387 (visited on 07/27/2019).

Cho, Taehong & Peter Ladefoged (1999). "Variation and Universals in VOT: Evidence from 18 Languages". In: *Journal of Phonetics* 27.2, pp. 1–23. DOI: 10/bfrsns.

Cisek, Paul & John F Kalaska (2005). "Neural Correlates of Reaching Decisions in Dorsal Premotor Cortex: Specification of Multiple Direction Choices and Final Selection of Action". In: *Neuron* 45.5, pp. 801–814. DOI: 10/bx442c.

Farmer, Thomas A., Sarah E. Anderson, & Michael J. Spivey (2007). "Gradiency and Visual Context in Syntactic Garden-Paths". In: *Journal of Memory and Language*. Language-Vision Interaction 57.4, pp. 570–595. DOI: 10/b825f7.

Fischer, Martin H. & Matthias Hartmann (2014). "Pushing Forward in Embodied Cognition: May We Mouse the Mathematical Mind?" In: *Frontiers in Psychology* 5, p. 525. DOI: 10.3389/fpsyg.2014.01315.

Freeman, Jonathan B. (2018). "Doing Psychological Science by Hand". In: *Current Directions in Psychological Science* 27.5, pp. 315–323. DOI: 10/gfhdmn.

Freeman, Jonathan B. & Nalini Ambady (2010). "MouseTracker: Software for Studying Real-Time Mental Processing Using a Computer Mouse-Tracking Method". In: *Behavior Research Methods* 42.1, pp. 226–241. DOI: 10/crr63c.

Freeman, Jonathan B., Rick Dale, & Thomas A Farmer (2011). "Hand in Motion Reveals Mind in Motion". In: *Frontiers in Psychology* 2, pp. 1–6. DOI: 10/b57b72.

Friedman, Jerome, Trevor Hastie, & Robert Tibshirani (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. OCLC: 428882834. New York, NY: Springer-Verlag New York. URL: doi.org/10.1007/978-0-387-84858-7 (visited on 07/29/2019).

Hehman, Eric, Ryan M. Stolier, & Jonathan B. Freeman (2015). "Advanced Mouse-Tracking Analytic Techniques for Enhancing Psychological Science". In: *Group Processes & Intergroup Relations* 18.3, pp. 384–401. DOI: 10/f7bqtv.

Holt, Lori L. & Andrew Lotto (2010). "Speech Perception as Categorization". In: *Attention, Perception & Psychophysics* 72.5, pp. 1218–1227. DOI: 10/bctgj6.

Jessen, Michael (1999). *Phonetics and Phonology of Tense and Lax Obstruents in German*. John Benjamins. DOI: 10.1075/sfsl.44.

Kieslich, Pascal J. & Felix Henninger (2017). "Mousetrap: An Integrated, Open-Source Mouse-Tracking Package". In: *Behavior Research Methods* 49.5, pp. 1652–1667. DOI: 10/gb4dhw.

Kieslich, Pascal J. et al. (2019a). "Design Factors in Mouse-Tracking: What Makes a Difference?" In: *Behavior Research Methods*. DOI: 10/gf4qzh.

Kieslich, Pascal J. et al. (2019b). "Mouse-Tracking: A Practical Guide to Implementation and Analysis". In: *A Handbook of Process Tracing Methods*. 2nd edition. New York, NY, US: Routledge. URL: https://osf.io/zuvqa (visited on 07/08/2019).

Kowler, Eileen et al. (1995). "The Role of Attention in the Programming of Saccades". In: *Vision Research* 35.13, pp. 1897–1916. DOI: 10/bdh9gq.

Liberman, Alvin M. et al. (1957). "The Discrimination of Speech Sounds within and across Phoneme Boundaries." In: *Journal of Experimental Psychology* 54.5, pp. 358–368. DOI: 10/bfcz77.

Liberman, Alvin M. et al. (1961). "The Discrimination of Relative Onset-Time of the Components of Certain Speech and Nonspeech Patterns". In: *Journal of Experimental Psychology* 61.5, pp. 379–388. DOI: 10/dg82fh.

Lindblom, Björn (1996). "Role of Articulation in Speech Perception: Clues from Production". In: *The Journal of the Acoustical Society of America* 99.3, pp. 1683–1692. DOI: 10/dx97qz.

Lisker, Leigh (1986). ""Voicing" in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees". In: *Language and Speech* 29.1, pp. 3–11. DOI: 10/gf4343.

Lisker, Leigh & Arthur S. Abramson (1964). "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements". In: *Word* 20.3, pp. 384–422. DOI: 10/gf26qs.

Magnuson, James S. (2005). "Moving Hand Reveals Dynamics of Thought". In: *Proceedings of the National Academy of Sciences* 102.29, pp. 9995–9996. DOI: 10/dxzm7q.

Maldonado, Mora, Ewan Dunbar, & Emmanuel Chemla (2018). "Manipulated Decision Tasks to Decode Behavioral Measures: The Case of Mouse-Tracking".

Massaro, Dominic W. & Michael M. Cohen (1983). "Categorical or Continuous Speech Perception: A New Test". In: *Speech Communication* 2.1, pp. 15–35. DOI: 10/fr92qc.

Mathôt, Sebastiaan, Daniel Schreij, & Jan Theeuwes (2011). "OpenSesame: An Open-Source, Graphical Experiment Builder for the Social Sciences". In: *Behavior Research Methods* 44.2, pp. 314–324. DOI: 10/ft2dgc.

Matin, Ethel, K. C. Shao, & Kenneth R. Boff (1993). "Saccadic Overhead: Information-Processing Time with and without Saccades". In: *Perception & Psychophysics* 53.4, pp. 372–380. DOI: 10/bcvht5.

McMurray, Bob et al. (2008). "Gradient Sensitivity to Within-Category Variation in Words and Syllables." In: *Journal of Experimental Psychology: Human Perception and Performance* 34.6, pp. 1609–1631. DOI: 10/dhbdrz.

Miller, Joanne L. et al. (1983). "A Possible Auditory Basis for Internal Structure of Phonetic Categories". In: *The Journal of the Acoustical Society of America* 73.6, pp. 2124–2133. DOI: 10/dp87c3.

O'Reilly, Christian & Réjean Plamondon (2011). "Can Computer Mice Be Used as Low-Cost Devices for the Acquisition of Planar Human Movement Velocity Signals?" In: *Behavior Research Methods* 43.1, pp. 229–238. DOI: 10/cghzqj.

Pfister, Roland et al. (2013). "Good Things Peak in Pairs: A Note on the Bimodality Coefficient". In: *Frontiers in Psychology* 4. DOI: 10/gfkmrc.

Pisoni, David B & Jeffrey Tash (1974). "Reaction Times to Comparisons within and across Phonetic Categories". In: *Perception & Psychophysics* 15.2, pp. 285–290. DOI: 10/fmvx2p.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org.

Repp, Bruno H. (1984). "Categorical Perception: Issues, Methods, Findings". In: *Speech and Language*. Ed. by Norman J. Lass. Vol. 10. Elsevier, pp. 243–335. DOI: 10.1016/B978-0-12-608610-2.50012-1.

Richman, Joshua S. & J. Randall Moorman (2000). "Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy". In: *American Journal of Physiology-Heart and Circulatory Physiology* 278.6, H2039–H2049. DOI: 10/gfrbk4.

Roettger, Timo B. (2019). "Researcher Degrees of Freedom in Phonetic Research". In: *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10.1, p. 1. DOI: 10/gf5gz4.

Roettger, Timo B. & Michael Franke (2019). "Evidential Strength of Intonational Cues and Rational Adaptation to (Un-)Reliable Intonation". In: *Cognitive Science* 43.7, e12745. DOI: 10/gf5ksc.

Roettger, Timo B. & Mathias Stoeber (2017). "Manual Response Dynamics Reflect Rapid Integration of Intonational Information during Reference Resolution". In: Proceedings of CogSci 2017.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. URL: http://www.rstudio.com/.

Samuel, Arthur G. (1982). "Phonetic Prototypes". In: *Perception & Psychophysics* 31.4, pp. 307–314. DOI: 10/cxv7f6.

Scherbaum, Stefan & Pascal J. Kieslich (2018). "Stuck at the Starting Line: How the Starting Procedure Influences Mouse-Tracking Data". In: *Behavior Research Methods* 50.5, pp. 2097–2110. DOI: 10/gfdtg6.

Schouten, Bert, Ellen Gerrits, & Arjan van Hessen (2003). "The End of Categorical Perception as We Know It". In: *Speech Communication*. The Nature of Speech Perception 41.1, pp. 71–80. DOI: 10/bvf6kj.

Song, Joo-Hyun & Ken Nakayama (2009). "Hidden Cognitive States Revealed in Choice Reaching Tasks". In: *Trends in Cognitive Sciences* 13.8, pp. 360–366. DOI: 10/fm27vs.

Spivey, Michael J. (2007). *The Continuity of Mind*. Oxford Psychology Series 44. OCLC: 837300199. Oxford: Oxford Univ. Press. 428 pp.

Spivey, Michael J., Marc Grosjean, & Guenther Knoblich (2005). "Continuous Attraction toward Phonological Competitors". In: *Proceedings of the National Academy of Sciences* 102.29, pp. 10393–10398. DOI: 10/df8phh.

Stillman, Paul E., Xi Shen, & Melissa J. Ferguson (2018). "How Mouse-Tracking Can Advance Social Cognitive Theory". In: *Trends in Cognitive Sciences* 22.6, pp. 531–543. DOI: 10/gdmmzx.

Tanenhaus, Michael K. et al. (2000). "Eye Movements and Lexical Access in Spoken-Language Comprehension: Evaluating a Linking Hypothesis between Fixations and Linguistic Processing". In: *Journal of Psycholinguistic Research* 29.6, pp. 557–580. DOI: 10/dpvptc.

Teruya, Hideko & Vsevolod Kapatsinski (2019). "Deciding to Look: Revisiting the Linking Hypothesis for Spoken Word Recognition in the Visual World". In: *Language, Cognition and Neuroscience* 34.7, pp. 861–880. DOI: 10/gf4qzm.

Toscano, Joseph C & Bob McMurray (2016). "Voicing in English Revisited: Measurement of Acoustic Features Signaling Word-Medial Voicing in Trochees". In: *The Journal of the Acoustical Society of America* 132.3, pp. 1–1. DOI: 10/gf4qzn.

Wulff, Dirk U., J. M. B. Haslbeck, & M. Schulte-Mecklenbeck (2018). "Measuring the (Dis-) Continuous Mind: What Movement Trajectories Reveal about Cognition". In: *Manuscript in preparation*.

Wulff, Dirk U. et al. (2019). "Mouse-Tracking: Detecting Types in Movement Trajectories". In: *A Handbook of Process Tracing Methods*. 2nd edition. New York, NY, US: Routledge. URL: https://osf.io/6edca (visited on 07/08/2019).

Zgonnikov, A. et al. (2017). "Decision Landscapes: Visualizing Mouse-Tracking Data". In: *Royal Society Open Science* 4.11. DOI: 10/gf6rqn.