Current Psychology (2024) 43:17500–17524
https://doi.org/10.1007/s12144-024-05614-4

Consciously choosing and shaping what to comprehend: a mixed-methods approach to first-person aspects of mental agency in ambiguous speech perception

Johannes Wagemann[1] · **Annika Walter**[2]

Accepted: 31 December 2023 / Published online: 8 February 2024

Abstract
Speech perception plays a key role in many fields of human development and social life but is often impaired by ambiguities on various levels of processing. While these phenomena have been extensively researched in the cognitive (neuro-) sciences according to empirical paradigms that adhere to the third-person perspective of externally measurable behavior, their first-personal and agentive dimensions remain mostly elusive. However, particularly the latter should not be neglected as they can in principle not completely be mapped on quantitative data but are crucial for people in lifeworld situations. We explored this point in the contexts of cognitive penetrability and mental action and conducted a mixed-methods study with qualitative reports on speech perceptual reversal ($N=63$) as part of a series of related studies on other modalities. Exposed to respective stimuli, one half of the participants was instructed to voluntarily change their verbal percept, while the other half were told to hold a deliberately chosen word. Qualitative data analysis revealed four typical forms of mental activity, various strategies, and accompanying forms of intention and metacognitive feelings. On the one hand, this activity structure replicates that found in already published studies on vision and non-linguistic audition and thus lends itself to refinement of Posner and Petersen's (*Annual Reviews in Neuroscience, 13*, 25–42, 1990) classic model of attentional shift. On the other hand, statistical testing of the quantified data strengthened our hypotheses about mental activities across conditions and modalities, thus also arguing for a cultivable agentive attention awareness in speech perception that even penetrates early stages of speech processing.

Keywords Ambiguous speech perception · Perceptual reversal, first-person method · Mental strategies · Metacognitive feelings · Cognitive penetration of perception, mental action

Introduction

Comprehending spoken language plays a crucial role in many forms of human interaction by providing access to a shared understanding of manifold aspects of social life and cooperative work. However, as research has shown, understanding spoken utterances is far more complex than simply mapping heard sounds onto corresponding meanings (Harley, 2014; Poeppel, 2012). At each of the processing stages of speech perception, ranging from physical stimuli to complete and successful comprehension, ambiguities may occur that must be resolved. While much of neuro- or biolinguistic research suggests, that this is an exclusive matter of brain processes (Bickerton, 2014; Friederici, 2017; Hickok, 2012), it can be asked critically which aspects of speech perception might evade neuroscientific explanation (Johnson, 2009; Mondal, 2022) and to which extent phenomenal consciousness and mental agency might have access to them (Kee, 2020). However, answering these questions, which are both of theoretical interest (e.g., for mind-body correlations) and have practical implications (e.g., for language learning and self-development), seems to depend on the processing stages at which ambiguities lurk and how they are managed. Therefore, we first give an overview of the descriptive levels at

✉ Johannes Wagemann
johannes.wagemann@alanus.edu

[1] Institute for Waldorf Education, Inclusion and Interculturalism, Alanus University of Arts and Social Sciences, Campus Mannheim, Am Exerzierplatz 21, 68167 Mannheim, Germany

[2] Institute for Education and Social Innovation, Alanus University of Arts and Social Sciences, Thomas-Mann-Straße 36, 53111 Bonn, Germany

which determinants of speech perception can be considered, before connecting them regarding their cognitive penetrability. Against this background, we then introduce an empirical first-person approach to perceptual reversals and develop hypotheses to address (potentially) conscious mental activities as one aspect of the outlined questions.

Starting at the distal end of speech comprehension, the acoustic signal a listener receives during an act of verbal communication might be polluted by other (verbal and/or non-verbal) sounds. A disturbing acoustic environment (Le Prell & Clavier, 2017) or inherently ambiguous speech stimuli might challenge or even threaten the success of communicative acts, although in acoustically ambiguous situations, physical cues in the stimulus itself can aid disambiguation for the listener (Gow & Gordon, 1995; Lehiste, 1972; Montagne & Zhou, 2016) and support successful comprehension. Apart from the stimulus itself, physiological issues in the ear or auditory system, or impairments in auditory processing pathways and connected neural networks, or a combination of internal and external factors (Liu et al., 2018), can hinder comprehension and classification of any auditory information.

While these determinants of speech perception can undeniably be explained at the physical or physiological level, it still relies on meaning-bearing, higher cognitive functions, such as experience, memory retrieval, and predictive abilities (Frank & Willems, 2017), whose mapping onto the neural substrate is much more difficult to accomplish (Brehm & Goldrick, 2016; Buzsaki, 2019; Poeppel, 2012). The particular complexity of speech perception can be illustrated as follows: Even though neonates already seem to prefer spoken language over equally complex non-linguistic sounds (Vouloumanos & Werker, 2007) and discriminate between different languages (Mehler et al., 1988; Moon et al., 1993), infants do take much longer to acquire more specific perceptual skills in their native language than in visual perception (Johnson, 2010; Kuhl et al., 2008). Once acquired, however, expectation and prediction operate very quickly at a pre-reflective and automated level of processing, which may suggest that they can be reduced to neural activity. A striking example is sine-wave speech, which is derived from natural speech by simulating its frequency and amplitude patterns with a few sine tones and is perceived by untrained individuals as incoherent whistling or science fiction sounds when first heard (Davis & Johnsrude, 2007; Remez et al., 1981). However, after being exposed to the natural utterance or informed about its verbal content, individuals are able to completely comprehend its degraded sine-wave version. This clearly demonstrates a top-down effect on perceptual grouping based on phenomenally conscious knowledge; but the mechanism of this change is beyond listeners' insight and control, as they cannot switch back to their first, incoherent experience. Voluntary reversals in ambiguous speech perception, however, may provide a greater extent of processual insight and agentive control, as will be shown.

In any case, sine-wave and other forms of distorted speech pointedly illustrate that even the normal, undistorted stream of verbal utterances does not contain clearly separable linguistic elements (Redford & Baese-Berk, 2023; Roberts & Summers, 2010), which reflects the mapping problem and has been highlighted in the context of Chomsky's poverty of the stimulus argument (Laurence & Margolis, 2001). Hence, the underdetermination of successful speech comprehension by its acoustic input requires listeners to organize the latter into linguistic substructures, such as phones, phonemes, morphemes, words, and phrases. For example, the distinction of "gray day" and "grade A" depends on whether the phone [d] (i.e., the d-sound) is assigned to either the preceding or the following phoneme /eɪ/ (written in English as "a" or "ay") constituting the different words with different meanings (for further examples, see Lehiste, 1960). In well-pronounced utterances, listeners can exploit phonetic correlates such as pre-boundary lengthening and pitch accent (Beckman & Pierrehumbert, 1986; Wightman et al., 1992), but in ambiguous speech signals these phonological percepts are vague or not available at all, which opens a scope for word boundary interpretation (Lee et al., 2020). This process, also called lexical segmentation, is potentially one of the most significant elements in successful speech comprehension as it constitutes a bridge between incoming speech stimuli and linguistic structure formation (Klatt, 1989), which is particularly highlighted in educational contexts (Field, 2008; Goh & Wallace, 2018).

The question to which extent people can consciously access and control the subtleties of their linguistic processing can first be placed in the debate on cognitive penetrability of perception, where the role and relevance of top-down versus bottom-up neural processing are discussed. On the one hand, opponents of cognitive penetrability conceive perceptual subsystems, especially close to the sensory level, to be informationally encapsulated and thus shielded against influences from higher processing levels (Clarke, 2021; Firestone & Scholl, 2016; Fodor, 1983). On the other hand, proponents of penetrability interpret ventral (as opposed to dorsal) neural streams as top-down influences and propose corresponding interactionist models of perception (Gregory, 1966; Hommel et al., 2001; Rock 1983). Specifically for speech perception, the *TRACE* model was introduced by McClelland and Elman (1986; McClelland et al., 2006), which provides for bidirectional interaction between three linguistic levels (words, phonemes, phoneme features). Although, in this sense, there is much experimental work favoring top-down aspects even in early stages of auditory processing (Getz & Toscano, 2019; Heald and Nussbaum, 2014; Patel et al., 2022), Norris and McQueen's *Merge B* model argues for a probability-based explanation of

top-down feedback connections from the lexical to the prelexical level (2008; Norris et al., 2016). This in turn calls into question the influence of lexical knowledge on speech perception, so the outcome of this debate seems undecided.

As a mostly underexposed aspect of cognitive penetrability, it should be noted that both mechanisms, not only bottom-up but also top-down, are usually classified as operating below the level of phenomenal consciousness and therefore would not allow for any conscious control either. This is where first-person experimental designs and the examination of mental agency may provide a new perspective, since the study of speech perception in ambiguous situations is mostly limited to third-person paradigms such as button press (e.g., Barraza et al., 2016), the tracking of mouse trajectories (e.g., Lee et al., 2020) or reaction times (e.g., Maciuszek, 2018). Yet, even from a neuro-centric perspective, the study of perceptual reversal is closely connected to the conscious experience of subjects and has been linked to neural components indicating an involvement of higher-order processing networks. In both visual (Pitts & Britz, 2011) and auditory perceptual reversal (Davidson & Pitts, 2014), EEG measurements imply not only the participation of networks associated with sensory processing, but also hint at higher-level areas connected to the content of conscious thought. Therefore, while changes in perception due to passively primed object knowledge or semantic effects have been extensively investigated (see Firestone and Scholl's (2016) reference guide: http://perception.yale.edu/TopDownPapers), consciously intended and voluntarily executed perceptual reversals provide an experimental route focusing on the agentive dimension of cognitive penetration.

The mental action debate, however, has so far only dealt with higher cognitive functions such as deciding (Peacocke, 2007), judging (Owens, 2009), remembering (Arango-Muñoz & Bermúdez, 2018), and reasoning (Valaris, 2023; for overview see Fiebich & Michael, 2015). Since sensory perception is traditionally viewed to be even more beyond conscious control than higher cognitive functions, efforts have been limited to clarifying the agentive status of the latter. There are only a few exceptions to this, such as exploring the affordance character of perception for mental (and bodily) actions (McClelland, 2019) and the agentive awareness occurring during attentional shifts in visual or auditory perception (Watzl, 2017). However, as we have shown in previous studies, the definitional criteria for mental actions, such as conscious intention (O'Shaughnessy, 2000), trying (Proust, 2001), and evaluative control by metacognitive feelings (Proust, 2010/2015), can be transferred to perception (Wagemann & Raggatz, 2021, Wagemann, 2023). Although in this respect, as elsewhere (Brent & Titus, 2023; Wu, 2013), the role of consciousness for mental action is of increasing interest, the move toward systematic first-person empiricism has not yet been made more broadly.

At this point, the indicated threads of cognitive penetrability, mental agency, and perceptual reversal shall be connected and substantiated by preceding work to provide methodological and conceptual cornerstones for the current study. The mixed-methods approach of task-based introspective inquiry (TBII) was originally developed by the first author to more explicitly and deeply incorporate individuals' first-person perspectives into research on perceptual reversals (Wagemann, 2020, 2023; Wagemann et al., 2018). More generally, this approach draws on cognitive or other tasks the execution of which is documented by participants' qualitative self-reports, followed by in-depth content analysis and coding of the data, and statistical analyses based on different levels of "late" quantification. In contrast to common mixed-methods designs collecting qualitative and quantitative data in different stages and with different instruments (Creswell, 2009), "late" quantification means that only qualitative data is recorded and first analyzed, safeguarded by intercoder reliability tests, before it is quantified and subjected to statistical hypothesis testing. Here, quantification can directly exploit quantitative aspects of the self-reports (e.g., word frequencies) or build on nominal or metrical variables derived from qualitative coding, thus avoiding incommensurability problems (Small, 2011).

This first-person experimental procedure (to be explained below in detail) has already been applied to visual and (non-linguistic) auditory perceptual reversal tasks under different conditions and yielded results at a cross-modal level which directly address the issue of cognitive penetrability in terms of phenomenally conscious mental activity. While participants in both the visual (Wagemann et al., 2018) and auditory (Wagemann, 2023) experiments were instructed to switch between different percepts being faced with ambiguous stimuli, the task of holding a certain percept with stimuli continuously changing between ambiguous and unambiguous versions has so far only been tested for the visual case (Wagemann 2020). As the core finding of these studies, for both modalities and conditions a common structure of mental activities could be confirmed, as inspired by Witzenmann's (2022) structure phenomenology: Conscious mental activity can be distinguished in perceptual reversals in terms of (1) Turning Away (from the stimulus), (2) Producing (anticipatory mental content), (3) Turning Towards (the stimulus while searching), and (4) Perceiving (the changed percept with full certainty). In view of cognitive penetration of perception and corresponding mental agency, these mental micro-activities normally proceed subconsciously, but under experimental conditions can be raised to the level of conscious observation and, to some extent, control. This framework of conscious and agentive attention regulation could also be found (in modified forms) in visual counting of moving objects (Wagemann & Raggatz, 2021), nonverbal social interaction (Wagemann & Weger, 2021; Wagemann

et al., 2022), and directed thought (Wagemann, 2022). Comparison of modalities and experimental conditions yielded different frequency patterns for the mental activity forms, which, however, did not yet allow any clear conclusions to be drawn. In addition to the four micro-activities, performance-related or metacognitive emotions were reported, for example, in a visual hold task (Wagemann, 2020), and different forms of conscious intention were found under auditory change condition (Wagemann, 2023).

The outlined activity structure can be contextualized by other work, such as the attentional shift paradigm. According to Posner and Petersen (1990), the process of attentional shift has been divided into three distinct sub-phases, namely:

I. Disengagement from current focus of attention
II. (Re-) orientation towards new (intended) focus
III. Engagement with the new focus of attention

In our framework, Posner & Petersen's paradigm was not only validated, but the second or third phase of their model could even be further refined by mental activities of Producing, Turning Towards, and Perceiving. For example, (Re-) orientation (II) could be subdivided into the former two activities, or Engagement with new focus (III) could be subdivided into the latter two activities. In any case, our approach provides a more fine-grained dynamic and a phenomenal and agentive access to attentional shift. Another reference can be found in EEG studies on visual perceptual reversal revealing a temporal dynamic of two or three neural components (event-related potentials) which are interpreted as contributing to destabilization of a preceding percept and restabilization of a new percept (Kornmeier et al., 2019). Also, in relation to this work, our framework offers a sophisticated and complementary approach by which perhaps even previously undetected neural components can be predicted. That the outlined activity structure has been discovered only by including first-person experience and mental agency at a qualitative level underlines the relevance of this methodological extension and suggests utilizing it also in the current study.

The indicated methodological and conceptual gaps can be transferred directly to the field of speech perception, as is evident from the preceding considerations. Consequently, we want to investigate to what extent the results from our visual and auditory studies can be replicated and adapted for the case of ambiguous speech perception. In terms of replication, we deploy the proven methodology (TBII) and experimental design of a perceptual reversal task with change and hold conditions, while we use a slightly increased sample size to be on the safe side with statistical analyses and focus more strongly on metacognitive emotions and conscious intentions to provide a suitable basis for assessing the agentive status of mental activities. More precisely, the following research questions are raised and then condensed into two quantitative hypotheses with qualitative complements.

A first, methodological question is whether suitable tasks (change/hold) can be designed with a demand characteristic comparable to the visual and auditory studies and whether their execution can stimulate participants' awareness for mental micro-activities. This question was worked on in the preparatory phase of this study and answered positively based on trial runs with students. Second, given that the outlined activities could be reliably coded in the data, it would be of interest whether and how their frequency patterns relate to those of the other modalities (vision, audition). If significant differences could be found across modalities and justified theoretically, then the relevance of mental activities for perceptual reversal would be strengthened for vision, (non-linguistic) audition, and speech perception. Third, the question whether code frequencies of activities depend on the experimental conditions (change/hold) needs to be pursued and outcomes to be explained, which can be combined with sensory modalities. Fourth, the common question of cognitive penetrability and mental agency whether and how deep phenomenal consciousness reaches into the linguistic and probably even auditory processing stages and can influence them via intentional commands is central from a more philosophical perspective.

For the quantitative hypotheses, questions (2) and (3) about code frequencies of mental activities are connected. Here, we specify the already mentioned deviation of visual and auditory frequency patterns in that Producing seems to be higher for audition than for vision (change and hold), and, conversely, Turning Toward appears to be lower for audition than for vision (change and hold, Wagemann, 2023, see Fig. 5). Theoretically, this can be explained by the more inward orientation of audition compared to vision (O'Callaghan, 2009) leading, on the one hand, to an increased awareness of stimulus-averted activities (e.g., Producing) in relation to stimulus-oriented activities (e.g., Turning Toward) for audition. On the other hand, such a shift in introspective awareness could be justified with limited cognitive resources according to the Global Workspace/Working Memory model (Baars, 1988). Against this background, speech perception can be considered to be even more inwardly oriented than non-linguistic audition, since it builds on the abstract and highly differentiated rules of linguistic levels and their interrelations and thus involves higher order (top-down) cognitive processing, as indicated above.

Therefore, we expect higher frequencies of Producing compared to vision and audition (Hypothesis 1). As a qualitative complement, we hypothesize that a variety of more sophisticated forms of interchangeable mental strategies will be included in Producing, as such strategies have already

been observed in the visual and auditory cases, but not as pronounced in terms of variety.

In view of experimental conditions, a relatively high frequency of Turning Away was salient for holding an intended visual percept while being faced with an ambiguously changing stimulus (Wagemann, 2020, see Fig. 5). As we assume that, for speech perception, the hold condition is also associated with a stronger confrontation of participants with disturbing or distracting aspects of the stimulus, we expect higher frequencies of Turning Away here, as in the change condition (Hypothesis 2). As a qualitative complement for both conditions and referring to the above question (4), we expect supportive data in terms of intentions and metacognitive feelings as criteria for mental action, but do not make specific hypotheses here.

Findings supportive for these hypotheses would contribute to the indicated research gaps as follows. In view of Hypothesis 1, the combination of a quantitatively pronounced and qualitatively differentiated status of Producing would strengthen cognitive penetration and mental agency, in particular for higher, stimulus-remoter processing stages of speech perception and, at the same time, confirm the cross-modal relevance of the mental activity structure. As for the quantitative aspect of Hypothesis 2, the same can be claimed, with the difference that we are concerned here with the stimulus-nearer processing stage of Turning Away and its susceptibility to intramodal experimental conditions. Supportive findings in terms of the qualitative part of Hypothesis 2 would additionally contribute to a phenomenally conscious approach to mental agency in speech perception. In general, at the methodological level, qualitatively rich and statistically significant results in line with our hypotheses would show that the chosen procedure can shed light on key aspects of speech perception that would otherwise not be accessible.

Experimental procedure

Stimuli and tasks

For the two conditions, change and hold, different computer-generated speech stimuli were designed to fine-tune the demand characteristic and level of the tasks. To minimize unintended low-level differences between conditions, the same voice was used for both stimuli. First, for the change condition, one segmentally ambiguous two-word sequence was chosen, which can be heard as either "Ice cream" or "I scream" (as in the famous song titled "I scream, you scream, we all scream for ice cream"). These approximately homophonous sequences are equal regarding their phonetic spelling /aɪ-s-kriːm/, whereas the disambiguated speech percept depends on the assignment of the s-sound to the preceding or following phonemes (Lee et al., 2020; Lehiste, 1960). To create a stimulus with maximum ambiguity in this regard, from the available synthetic voices in the used text-to-speech app, one was chosen that was as expressionless as possible, and the s-sound was placed in an intermediate position between the preceding and following phonemes through trial and error. Since there were no "right" or "wrong" percepts to process from the stimulus in articulative or semantic regard, there was no need to enrich it with further sub-phonetic or prosodic features or to embed it in a carrier sentence, as is common in linguistically more specialized studies. The duration of the acoustic information in the stimulus was 1.0 s, followed by 3.4 s of silence. The stimulus was presented in a loop and recommended to be heard with earphones.

Similarly, for the hold condition, a one-word stimulus with the phonetic spelling /flaɪ/ ("Fly") was used, which was presented in a continuous loop with 0.2 s silence after the 0.3 s' long word. Due to the fast succession of the end of one perceptual chunk and the beginning of the next one, the stimulus can be not only be heard as "Fly" but also as "Life", known as the *verbal transformation effect* (Barraza et al., 2016; Warren & Gregory, 1958). Here, it is the assignment of the f-sound to the preceding or following phonemes which determinates the disambiguated percept. Interestingly, this effect can easily be created without technical aids by repeatedly pronouncing the same word ("fly" or "life") in rapid succession. To observe this effect introspectively, it is even sufficient to speak silently just by moving the tongue, which already anticipates one aspect of data analysis.

For both conditions, the trial was designed to last one consecutive week and required subjects to perform the task daily for 5 min at their own responsibility. The one-week period allowed for both familiarization and initial training, on the one hand, and repetition of the task with learning effects, on the other hand, as has been proven in previous studies. With less time, participants would have difficulties to gain access to the (normally) unaccustomed and untrained first-person mode of observation of mental processes; more time would increase the risk that participants lose interest and commitment and possibly tend to repeat their own notes or even begin to add confabulations to their protocols. Thus, this design represents a reasonable compromise between different constraints. The procedure consisted of the following steps and aspects: After participants received the stimulus as a mp3 file via email, they had to first familiarize for 2 to 3 days with the stimulus and practice to safely perceive the different variants without any further intention. Following this initial phase, participants were instructed to perform the task, which included behavioral and observational components. In terms of behavior, in the change condition, participants were asked to repeatedly switch between the different percepts at will, whereas in the hold condition, they were instructed to voluntarily hold one perceptual variant over

as long a period as possible without switching to the other. As for observation, in both conditions they were asked to describe what they experienced while performing the task, what they did (mentally) to accomplish the behavioral part of the task, and to report how they succeeded. Furthermore, the instructions included a brief explanation of multistable perception with ambiguous (speech) stimuli and the recommendation to adjust the volume carefully. Participants were required to submit their protocols via email immediately after completion of the one-week trial. They were also instructed not to communicate with each other about the experiment during the trial and until the submission deadline.

Of course, since the study was not conducted in the laboratory, we were not able to directly control whether participants completed the task in a satisfactory manner. However, the qualitative reports allowed to assess whether participants understood the instructions and how they performed the task in terms of individual commitment and external conditions. Regarding individual commitment, there were certain differences, but this did not mean that individual protocols had to be excluded from analysis. External conditions were captured in qualitative coding but gave just as little reason to doubt a satisfactory task execution (see below Qualitative analysis section). As regards validity, frequencies of first-person pronouns in the protocols were also measured to assess whether data are based on introspective observation, which could be confirmed (see below Protocol length and first-person pronouns section).

Participants

The experiment was conducted between September 2021 and June 2022 at Alanus University (Campus Mannheim). Participants were recruited from undergraduate students in a variety of majors and levels and received partial course credit in phenomenology or anthropology courses through participation. In sum, sixty-three persons (51 females, 12 males) between 19 and 30 years ($M = 23.3$) participated in the study. Subjects were randomly assigned to conditions; 32 were assigned to the change condition and 31 to the hold condition. Neither before nor during data collection was the content of the study discussed with the participants, and they were not informed of hypothetical explanations for phenomena that might occur.

From a qualitative perspective, the total sample size seems more than sufficient, considering that $N = 20$ to 30 is generally recommended for qualitative in-depth studies (Dworkin, 2012; Fugard & Potts, 2015) and even smaller samples are accepted for thematic saturation (Guest et al., 2020). However, in view of the reference studies (Visual change: $N = 25$; visual hold: $N = 22$; Auditory change: $N = 26$) related to the individual experimental conditions the sample sizes are more in line with each other. The fact that they are a bit higher in the current study is due to the quantitative perspective of possibly also being able to statistically deal with more subtle phenomena. For the initial exploratory study with the first 16 participants already revealed a remarkable variety of mental strategies that can be assigned to the activity form of Producing. This encouraged us to investigate this aspect in more detail, not only qualitatively but also quantitatively. Depending on the different constellations, the expected statistical power for chi-square tests ranges between 0.67 and 0.72, based on a medium effect size $w = 0.3$ (Cohen, 1988), $\beta/\alpha = 4$ (e.g., $\beta = 0.2$ and $\alpha = 0.05$), and total sample sizes from 53 (speech hold vs. visual hold) to 63 (e.g., speech change vs. speech hold) (Faul et al., 2007). For independent samples t-tests, expected statistical power ranges between 0.66 and 0.70 under the same conditions. While these values are below the commonly recommended power of 0.8, we consider this to be a reasonable trade-off within our mixed-methods design.

Data acquisition and analysis

Consistent with the reference studies, data were collected via open-ended written self-reports submitted by participants via mail. Since we already justified the use of this method for perceptual-change studies at length in comparison to both standard methods and other first-person accounts (e.g., Wagemann, 2020; Wagemann et al., 2022), only some aspects will be briefly mentioned again here. First and foremost, since written self-reports are not influenced by the content of interview questions or questionnaire items, they provide a relatively open access to participants' first-person experience, which is not biased by experimenter expectations or predefined constructs. In the case of content-empty interview questions aimed only at re-evoking the experience in question (e.g., Vermersch, 1999; Petitmengin, 2006), written self-reports still have the advantage of excluding subliminal forms of nonverbal communication or other social dynamics. Second, this form of data collection fits well with participants' independent, time-flexible performance of the task because it does not require the deployment of additional staff or laboratory equipment. Third, the further processing of the data already available in text form is resource-friendly regarding the subsequent time-consuming qualitative analysis steps, not least also in view of the current sample size.

As indicated in the introduction, data analysis was conducted according to the following mixed-methods procedure: First, text data were qualitatively analyzed and coded, second, the qualitative results were quantified in terms of code frequencies which then were subjected to statistical analyses. In a sense, this procedure lies between what Creswell (2009) called concurrent and sequential approaches: Since data are collected only once (instead of successively collecting

different types of data) and have both qualitative and quantitative aspects, it is a concurrent approach. However, since the data are first analyzed from a qualitative perspective, the results of which are the starting point for the quantitative analysis, this is a sequential approach. Both analytical steps shall be explained in the following.

Qualitative analysis

Qualitative analysis followed the steps of conventional (bottom-up/inductive) and directed (top-down/deductive) content analysis (Hsieh & Shannon, 2005; Mayring, 2000). Using the first method, at Level 1 twenty-one categories and subcategories emerged from a data-driven analysis of multiple aspects of first-person experience (Table 1). At Level 2, the four main forms of mental micro-activities were adopted as categories from the visual and auditory reference studies, corresponding to a top-down approach, and then differentiated according to the task-specific data in the present study using the bottom-up principle, which resulted in eleven (sub-)categories (Table 2). At Level 3, three forms of intention or trying were adopted from the reference studies without further adaptation (Table 3). Level 2 and 3 categories are explained in more detail in the next paragraph. Thus, in this hierarchical coding procedure, Levels 2 and 3 directly address the research question of mental agency in speech perception, while Level 1 mostly serves to embed it in broader contexts and allow for a complete coding of the data. One exception of this are metacognitive feelings (Cat. 6), which partly also refer to mental micro-activities at Level 2. In quantitative terms, 98% of the total text data (based on characters) was coded, with the remainder consisting of numbering characters, dates, blanks, and unclear or fragmentary statements that could not be assigned. Coding units ranged from partial to whole sentences, resulting in a total of 2200 coded segments. For Level 1, a code coverage of 66% of the text (1384 segments) was achieved, while Level 2 resulted in 28% (613 segments) and Level 3 in 8% (202 segments). The fact that the percentages add up to 102% indicates a slight overlap in the codes. A structured overview of the category system and the coding levels is given in Fig. 1.

More detailed information on Level 2 and Level 3 categories adopted from previous studies (esp. Wagemann, 2023) and adjusted according to the task is given below. Firstly, in this sense, the core of the four mental micro-activities at Level 2 can be defined as follows:

1) Turning Away refers to all formulations of activities that indicate aversive gestures such as pushing back, fading out, disengaging from the unwanted variant or corresponding aspects of the stimulus. This includes expressing what the person wants to get away from to get to something other. However, the focus here is *not* on a positive decision *for* a particular percept, but on a decision and activity *against* something. Physical aids to disengage from distracting stimuli, such as closing the eyes, are not a part of this category to concentrate on the contribution of purely mental activity.
2) Producing includes first the decision for and explicit awareness of the word to which one wants to shift to or stick to. However, this goal is not prevalent here regarding its perceptual dimension but only in terms of the conceptual aspects which can support or constitute it. This means bringing forth and shaping mental content which appears to be helpful in the task context and can be assigned to individual mental strategies. Strategies related to external sensory perceptions, such as reading written words or other body-related strategies, are excluded here.
3) Turning Toward focuses attention on auditory processes and the perceptual stimuli mediated by them. In contrast to Turning Away, attentional activity is motivated and directed by specific content provided by Producing and searches for anchor points in the stimulus that might confirm the intended word variation. However, actually finding and confirming the intended variant at the stimulus does not belong to this category. Rather, this activity transitions from hearing an unintended variant or ambiguous stimulus to the intended variant without already perceiving it.
4) Perceiving enables the person to clearly confirm success in view of their perceptual intention. Success does not necessarily have to be perfect (in terms of perceptual quality) or complete (in terms of a certain duration of the trial), but the intended word variant is at least partially heard and confirmed as such.

To demarcate perceptual intention and trying as criteria for mental action at Level 3 from Level 2 activities, indicators must be determined that are common for the three forms of intention and those by which they can be clearly distinguished. To begin with the former, words, phrases, or contexts are searched for, which indicate that an agent wants to achieve or succeeds in achieving something by certain means. Typical examples are "I do … *in order to* achieve …", "I *try* to … *by* …", or "If I do … then … happens". In these cases, it can be assumed that an activity of the agent does not occur unintentionally or automatically (and is observed and reported like any arbitrary mental event or state) but arises as a direct consequence of a conscious intention or attempt (Proust, 2010). Without being able to cite here all linguistic forms of expression coded in this context, this defines the common feature of the three forms of intention. The distinction of different forms of intention initially builds on corresponding definitions in the philosophy of (mental) action delineating distal intentions (D-intentions)

Table 1 First coding level. Categories with subcategories, short descriptions, and exemplary excerpts from the data

LEVEL 1 – Multiple aspects	Short description	Examples
1. Stimulus, percepts, and coupled mental phenomena	Appearance and assessment of the stimulus and percepts, coupled mental states and dynamics (passive or preparatory to the task)	
1.1 Stimulus	Character of artificial speech stimulus and corresponding sensations	"Monotony of the voice and constant repetition" (WP1_16_H) / "[…] annoyed me the lack of accentuation" (WP3_04_C)
1.2 Mental State/Attitude	Accompanying mental states and attitudes to the task	"This allowed me to achieve a kind of meditative state while listening" (WP1_16_H) / "The next day, I approached the exercise with a more open mind" (WP3_20_H)
1.3 Thoughtless Listening	The stimulus is listened to without intentional thoughts	"My thoughts are empty." (WP1_03_C) / "I just tried to listen without thinking about it" (HP_12_H)
1.4 Passive Imagination	Spontaneously occurring thoughts or imaginations while listening	"Both statements create images in my head" (WP3_03_C) / "The longer I watch FLY, the stronger an image of a flying bird becomes" (HP_03_H)
1.5 Verbal Percept	Unintended hearing of a particular word or search for other word variations	"[…] until I heard 'ice cream' for the first time" (HP_01_C) / "Then I started trying to hear other things, e.g., 'slide,' 'wife,' 'drive,' 'slice' (WP1_15_H)
1.6 Preference	One of the two task variants s heard more often	"[…] because I heard 'I scream' more often" (WP1_06_C) / "[…] but 'ice cream' occurred more often." (WP3_25_C)
1.7 Perceptual Character	Character of the perceived word or perceptual change	"It seems slightly distorted, not pronounced correctly" (WP1_10_H) / "There was always a kind of 'stumbling' of the words after a few repetitions and 'fly' became 'life'" (HP_12_H)
1.8 Mixed Percept	A mixture of both word variants is perceived	"I repeatedly perceive an amalgamation of the words" (WP1_12_H) / "Chaos breaks out, then I hear everything mixed" (HP_20_C)
1.9 Alternation	An alternation of the word variants is perceived	"[…] new experience, because now I could listen alternately once 'I Scream' and once 'Ice Cream' (WP1_02_C)
1.10 Automaticity	Involuntary or uncontrollably arising perception or perceptual change	"I do not have to do anything, it changes automatically" (WP3_13_C) / "I had the feeling that I no longer had any control at all" (HP_10_H)
2. Body sensation/reaction	Perceived posture, condition, accompanying sensations and reactions	"I feel more awake and more present in my body" (WP_12_H) / "I perceive how I move my mouth along while listening" (HP_19_H)
3. External behavior/body support	Targeted use of body-related means, external objects or tools	"I bought ice cream today and ate it at about 3 pm" (HP_09_C) / "I wrote the words on a paper and read what I wanted to hear while playing the sequence" (HP_17_C)
4. External conditions/procedure	Where, under what conditions and how exactly was the task performed	"I take my mobile with the audio track into the forest, hoping to find more peace there." (HP_16_C) / "Then I play the audio file two to three times" (HP_02_C)
5. Affective emotions	Reactive-affective emotions relating to the task as a whole or its contents	"Depending on which word I understand, it triggers certain emotions in me" (WP3_03_C) / "It bugs me that I have to write sentences" (HP_09_C)
6. Metacognitive feelings	Task performance is experienced as easier or harder, other feelings about one's own cognition (subject-side of process evaluation)	"The change was not a challenge" (WP3_25_C) / "There is a fascination that goes along with the observation of my perception" (WP1_16_H) / "When I held on to "life," something inside me bristled" (HP_12_H)
7. Task evaluation (pos./neg.)	Success or failure of the task, perceived quality, possibly depending on external or internal conditions (object-side of process evaluation)	"It was more successful that way" (WP3_08_H) / "The perception of 'Fly' keeps interrupting me" (WP1_12_H)

Table 1 (continued)

LEVEL 1 – Multiple aspects	Short description	Examples
8. Concentration/mental effort	Concentration, focus, attention span, mental effort	"For today, I had resolved to concentrate more on the exercise" (HP_09_C) / "If I am only inattentive for a small moment […]" (WP3_14_H) / "If I try very hard mentally […]" (WP3_03_C) /
9. Distraction/mind wandering	Exo-/endogenous Distractions, mind wandering, digressions	"If I drift off […]" (WP1_08_C) / "Various thoughts, which have nothing to do with Ice cream or I scream, keep wandering through my head." (HP_16_C)
10. Learning	Advancement/regression in skill development, use of various methods	"I also notice how I surprisingly become more creative and discover more and more new mechanisms/images" (WP1_16_H) / "[…] I first faded out the images while listening and then also left them out as preparation (WP1_04_C)
11. Reflective thought	Associations, speculations, and conclusions about the experiment	"I think about the meaning of this exercise" (HP_09_C) / "[…] maybe because I focused it for so long on that one day" (WP3_08_H)

as future-directed or goal-oriented and proximal intentions (P-intentions) as more process-related (Buckareff, 2005; Mele, 1992). D-intentions are present before and at the beginning of a (e.g., perceptual) task as well as during the attempt to achieve the goal and therefore remain unchanged until the goal is reached. In the context of speech perception, D-intentions aim at acoustically perceiving a certain word und thus are connected to this mental activity (see Level 2). In contrast, P-intentions refer to specific options and strategies that can be discovered and used in the task context, and thus can and often do change over the course of task performance. Because they refer to strategic mental content to be actively brought about and deployed, they are connected to Producing (Level 2) and, for speech perception, comprise the whole range of semantic and articulative strategies as outlined in Table 2. As a third form, executive intentions (E-intentions) have been introduced to explain intentional access to those activities that establish the transitions between the conceptual (Producing) and the perceptual (Perceiving) side of the process (Wagemann, 2023). Therefore, E-intentions refer to the complementary activities of Turning Aaway and Turning Toward which are necessary to perform a full perceptual change.

Metacognitive feelings (MCFs) as a further criterion for mental action are analyzed as interlevel relations between Category 6 (Level 1) and mental micro-activities (Level 2). In general, MCFs express how difficult or easy a cognitive performance is perceived by an agent and to what extent they are satisfied with its outcome. In this sense, MCFs refer evaluatively to already completed cognitive (sub-)processes but can also precede them as a motivating factor (Proust, 2015). In our context, similar to intentions, we can distinguish whether metacognitive feelings refer to the whole process of perceptual reversal or to individual activities involved in it. To get the most accurate assessment of the agentive nature of mental micro-activities, we focus here on MCFs occurring in the same segments (as partial sentences) or in immediately adjacent segments before or after (additionally checked by context). It should be noted, however, that the assignment of MCFs to individual mental activities is ambiguous when they cluster in the same or adjacent segments. Since activity-related MCFs represent only a subset of all reported MCFs, we did not provide a separate analytic level for them as we did for intentions.

Intercoder reliability was tested only for Level 2 due to the close definitional relationships between mental activities and corresponding intentions (Level 3) and metacognitive feelings (Level 1), as explained above. One hundred fifty coded segments (about one-quarter of all codings at Level 2) were randomly selected, blinded, and then independently reassigned to the eleven categories by two coders who were not involved in the development of the Level 2 categories. One of them was the second author of this study,

Table 2 Second coding level. Categories with subcategories, short descriptions, and exemplary excerpts from the data

LEVEL 2 – Micro-activities	Short description	Examples
1. Turning away from the stimulus	Averting from unwanted parts of the stimulus, a specific percept, or other disturbances	"It's about ignoring/pushing away […] what you don't want to hear" (WP1_03_C) / "Blanking out of everything else, of all other possibilities, as if one were turning a deaf ear to everything else" (HP_11_H)
2. Producing	Deciding to change to or hold a certain variant; mentally productive attitude and performance with different strategies	
2.1 Producing (general)	Decision, awareness, or thoughts of what is to be perceived are mentally produced; concentrating or focusing on one variant	"The change between the two variants is without exception only brought about by a thought" (HP_08_C) / I actually create the connection between the sounds (WP3_02_C)
2.2 Imagining (general)	Imagining one variant in the form of figurative ideas	"[…] pictured them in a certain way" (WP1_01_C) / "Figurative idea to get to the desired word" (WP1_05_C)
2.3 Situation/symbol	Imagination of typical or individually significant situations/symbols	"I imagine a big fly" (WP1_12_H) / "A little screaming girl" (HP_04_C) / "reconnected completely with the image of the licking tongue over the colorful ice cream scoops" (HP_08_C)
2.4 Printed/written word	The printed or written word is imagined, combined with spelling the individual letters	"The word is handwritten in white, on a black background" (WP1_10_H) / "the lettering LIFE" (HP_03_H) / "I mentally imagine how one would spell which statement or word (HP_05_C)
2.5 Inner Speech	The word is spoken with inner, subvocal speech	"Then I say with an inner voice one of the two words" (WP1_03_C) / "[…] if I speak it silently" (WP1_14_H)
2.6 Phoneme placement	Consciously focusing on, emphasizing, or shifting back and forth certain sounds or related letters	"In 'Ice Cream', I pull the S forward, while in 'I scream', I pull the S backward (HP_02_C) / "Depending on which initial letter I focus on, F or L" (HP_07_H)
2.7 Productive emotions	Supportive emotions are evoked, e.g., in connection with imaginations, perceptual qualities, or by adjusting the inner attitude	"I try to think of negative, or upsetting stories in advance" (HP_16_C) / "Thus, 'ice cream' was pronounced high and soft internally" (WP3_16_C) /"The change from 'ice cream' to 'I scream' requires an inward turning or a separation from the outside world" (WP3_12_C)
3. Turning toward the stimulus	Aligning with the stimulus and searching the intended word in it	"[…] and concentrated on hearing out the word 'I Scream" (HP_04_C) / "I had to listen in into it first" (WP1_09_C)
4. Perceiving	Partial or complete certainty regarding the intended percept	
4.1 Preliminary/partial	Partial success, success in performance that deviates from the task (e.g., change in hold condition), spontaneous success (i.e., without prior activities)	"Initially, I still sporadically hear the other word through" (WP1_10_H) / "I could jump back and forth between the […] words without any problem" (WP1_13_H) /
4.2 Success	(Almost) full success (following at least one of the other three activity forms)	"Listening to it, I now understood 'Ice Cream'" (WP1_02_C) / "[…] and then I could hear 'Life'" (WP1_11_H)

Table 3 Third coding level. Categories with descriptions and exemplary excerpts from the data.

LEVEL 3 – Intention/Trying	Short description	Examples
1. Distal Intention	Intending the effect or result *that* one succeeds in the task / Trying to principally bring about a specific change or hold performance with one word Related to Perceiving	"[…] became aware of my intention again before listening" (WP1_16_H) "[…] to decide which word I want to hear" (HP_12_H) "Then I try to understand 'Icecream' and after that 'I scream' (WP1_03_C)
2. Proximal Intention	Intending different options as to *how* one could manage the task / Trying to have and follow a specific strategy to achieve the goal Related to Producing	"In the course of this I wanted to see if I can mentally shift this break" (WP3_25_C) "I now produce […] inner images in an attempt to stay with one variant" (HP_19_H) "I have changed my approach to focus only on the first letter F" (WP3_10_H)
3. Executive Intention	Intending / Trying to perform the averting or engaging mental activities required for task performance Related to Turning Away or Turning Toward	"I have to concentrate very consciously so that the word does not switch over in my head" (HP_10_H) "[…] and I had to keep telling myself to listen carefully" (WP1_16_H)

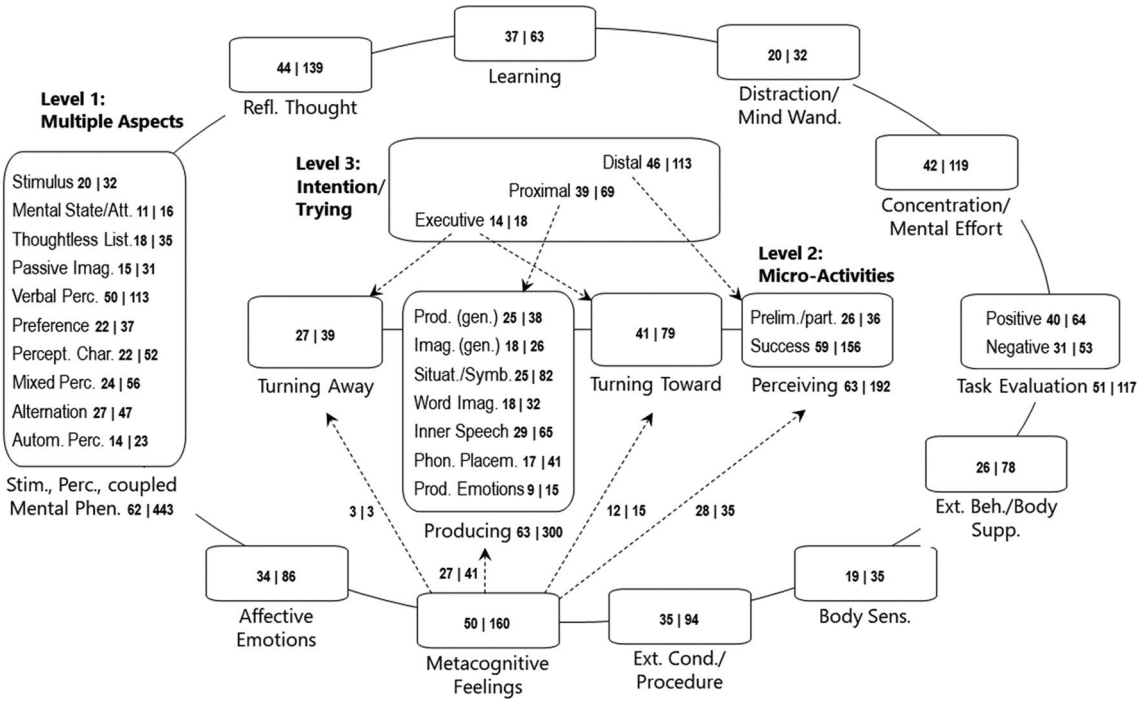

Fig. 1 Structured category system and coding levels. Numbers indicate in how many data sets (participants) data were encoded and how many segments were coded according to a certain category. Solid lines display intra-level connections, while dashed arrow lines show the relations between Levels 2 and 3

the other was not involved in the study at all. This resulted in Cohen's kappa values of $\kappa_1 = 0.67$ and $\kappa_2 = 0.74$, which on average already represents substantial (Landis & Koch, 1977) or moderate agreement (McHugh, 2012). To improve coding consistency, a feedback session was held with each coder to discuss and, where possible, clarify the discrepancies in the ratings (Campbell et al., 2013; O'Connor & Joffe, 2020). In almost all cases, inconsistencies turned out to rely on misunderstandings concerning code definitions and demarcations or missing context of isolated segments and could be resolved resulting in $\kappa_1 = \kappa_2 = 0.99$ (perfect agreement). To provide transparency here (Cheung & Tai, 2021), the most important points concerned the sharpening of the mental strategies referring to "words" and the activity of Turning Toward the stimulus. Firstly, on the one hand, it was stated that the unspecific "thinking about" or "focusing on" task-relevant "words" belongs to the more general category 2.1 of Producing, whereas category 2.4

requires the explicit formulation of a figurative imagining of the written or printed word. On the other hand, to demarcate this from 2.6 (phoneme placement), it was argued that the former refers to the whole word and letters play a role only insofar as words are spelled out typographically, whereas in the latter the focus is on individual sounds and associated letters or pauses (in the sense of an inner speaking or listening). Secondly, divergent assignments around Turning Toward (Cat. 3) highlighted the proximity and intermediate position of this category with respect to Producing (2.1) and Preliminary/partial perceiving (4.1). For Turning Toward, on the one hand, explicit attention to the auditory sense or something acoustically happening was emphasized here, whereas for Producing, attention is on the purely mental, self-initiated process. On the other hand, regarding 4.1, Turning Toward does not involve hearing the intended word already with full clarity and certainty, even if this intention may be formulated as a goal. Therefore, for example, the expression "to pick out" has a preliminary or transient role in the context of Turning Toward, while it has a final role in Perceiving, which can be verified in each case by the course of the sentence. Finally, the few changes that resulted from these clarifications were incorporated into the final coding of the data, which served as the basis for the next step of statistical analysis.

Quantitative analysis

Prior to quantitative analyses, which further process the results of qualitative coding, we conducted some tests directly related to quantitative aspects of the text data. As elementary parameters of open-ended introspective text data, we determined the protocol length in words and the proportion of first-person pronouns and compared them between conditions and modalities (Chung & Pennebaker, 2007; Seih et al., 2011). Statistical tests used for this purpose were t-tests for independent samples.

For statistical analyses based on qualitative coding three different variants were deployed. For the first two, the quantities of coded segments per category and data set (participant protocol) were binarized so that only the information on whether a category was present in a protocol or not was examined. This way, firstly nominal variables were generated from the codes and investigated by chi-square tests complemented by an exact test for frequencies below five (Boschloo, 1970). In the second variant metrical variables were derived from the number of coded categories per data set and analysis level and again explored by t-tests for independent samples. For the third variant a ratio scale variable (the activity cluster index) was derived from relations of coding quantities and the topological feature of proximity of code occurrence in the protocols, as will be explained in more detail below. For this variant, a one-way ANOVA was used to test for dependence.

Results

Protocol length and first-person pronouns

To begin with some purely quantitative results independent of qualitative analysis, the numbers of written words and proportions of first-person pronouns in the data sets were compared for experimental conditions and sensory modalities (corresponding to the previous studies). Across experimental conditions, protocol length appeared to be nearly constant and did not change significantly between Speech Change ($M = 398.7$, $SD = 176.6$) and Speech Hold ($M = 394.2$, $SD = 197.1$), $p = 0.929$. However, while protocol length for Speech Change was lower than for Auditory Change ($M = 459.2$, $SD = 217.9$), although not significantly, $p = 0.270$, it was significantly higher for speech than for vision in both conditions, in detail for Visual Change ($M = 196.0$, $SD = 112.8$), $t(56) = 4.84$, $p < 0.001$, $d = 1.8$, and for Visual Hold ($M = 266.3$, $SD = 138.9$), $t(53) = 2.76$, $p = 0.008$, $d = 0.9$. The difference between proportions of first-person pronouns (I, my, me) in Speech Change ($M = 9.2\%$, $SD = 2.2\%$) and Speech Hold ($M = 9.8\%$, $SD = 1.6\%$) was not significant, $p = 0.223$. Compared with the visual case, first-person pronouns were higher for Speech Change than for Visual Change ($M = 8.7\%$, $SD = 2.8\%$), although not significantly, $p = 0.470$, but marginally significantly higher for Speech Hold than for Visual Hold ($M = 8.8\%$, $SD = 2.3\%$), $t(53) = 1.72$, $p = 0.091$, $d = 0.4$. The average of 4.99% for various forms of written language can be cited here as a significantly lower comparative value (Pennebaker et al., 2015). Since the frequency of first-person (singular) pronouns used by participants in the protocols provides general information about their attentional focus during the task (Rude et al., 2004), this measure can be used in conjunction with protocol length to assess the required mode of self-focused introspective observation and the amount of information gained by it. In view of the relatively high occurrence of first-person pronouns in both speech conditions lying clearly above averages for different genres of writing (Pennebaker et al., 2015) and the relatively high protocol length (compared to the visual case and only slightly below auditory change), we can draw two initial conclusions: First, these results strengthen the methodological validity of the study, and second, they suggest greater proximity between non-linguistic auditory and speech perceptual reversal as opposed to visual reversal.

Level 1: Multiple aspects of first-person experience

As mentioned earlier, we cannot fully explore the multiple aspects of first-person experience at Level 1 in the context of this study but limit ourselves to those that are most important in qualitative terms, also with relation to Levels 2 and 3, and, in quantitative regard, are most salient or vary most markedly across conditions. As a first qualitative aspect that will be relevant to the issue of cognitive penetrability, participants reported highly differentiated experiences on the relationship between the auditory stimulus and clearly perceived words (Category 1). Particularly at the beginning of the trials, but also later, many participants noted that they were able to listen *intentionally without (content-related) intent* (Cat. 1.3), observing a phenomenality of the (proximal) stimulus that appeared gradually deprived of meaning. Here, we can distinguish four levels of deprivation or decomposition, starting with monotony, neutrality, or slight distortion of perceived words, e.g., "… it seems slightly distorted, not pronounced correctly" (Cat. 1.7, WP1_10_H), continuing with ambiguously mixed percepts (Cat. 1.8, see Table 1), further increasing with loss of meaning, e.g., "… feeling that the sounds dissolve more and more and the spoken loses more and more meaning" (Cat. 1.7, WP3_02_C), and culminating in the fully decomposed stimulus, e.g., "… I do not recognize which statement it is ultimately about" (Cat. 1.1, HP_02_C), "… the words themselves lost all meaning and were only a common sound" (Cat. 1.3, WP3_25_C).

Besides this, several phenomena are captured by Level 1 codes describing passive or reactive aspects of experience, such as hearing certain word variants without explicit perceptual intention (e.g., Categories 1.5 and 1.10) or having passive imaginations (Cat. 1.4) or affective emotions (Cat. 5) accompanying certain perceptions. While, complementary to this, aspects of *mental* agency in perception are assigned to Levels 2 and 3, active or agentive aspects can be found at Level 1 in terms of bodily or external behavior such as body-related strategies (Cat. 3) or external conditions of task performance (Cat. 4). Another interesting connection to Levels 2 and 3 can be seen in learning processes (Cat. 10), in which participants take up aspects they initially experienced passively (e.g., affective emotions, Cat. 5) and then use them intentionally and systematically in their task performance (e.g., productive emotions, Cat. 2.7), which will be discussed in more detail below.

When it comes to quantitative analyses, as shown in Fig. 2, the three most frequent categories remaining relatively constant across conditions are briefly mentioned. As would be expected in a speech perception experiment, unintentional hearing of a particular word variant occurs quite frequently in the protocols (Cat. 1.5). Also, very often metacognitive feelings (Cat. 6) and concentration/mental effort (Cat. 8) can be found. While the verbal percept represents what results on the object side, concentration/mental effort is what participants invest from their (subject) side in the perceptual process, and metacognitive feelings are what

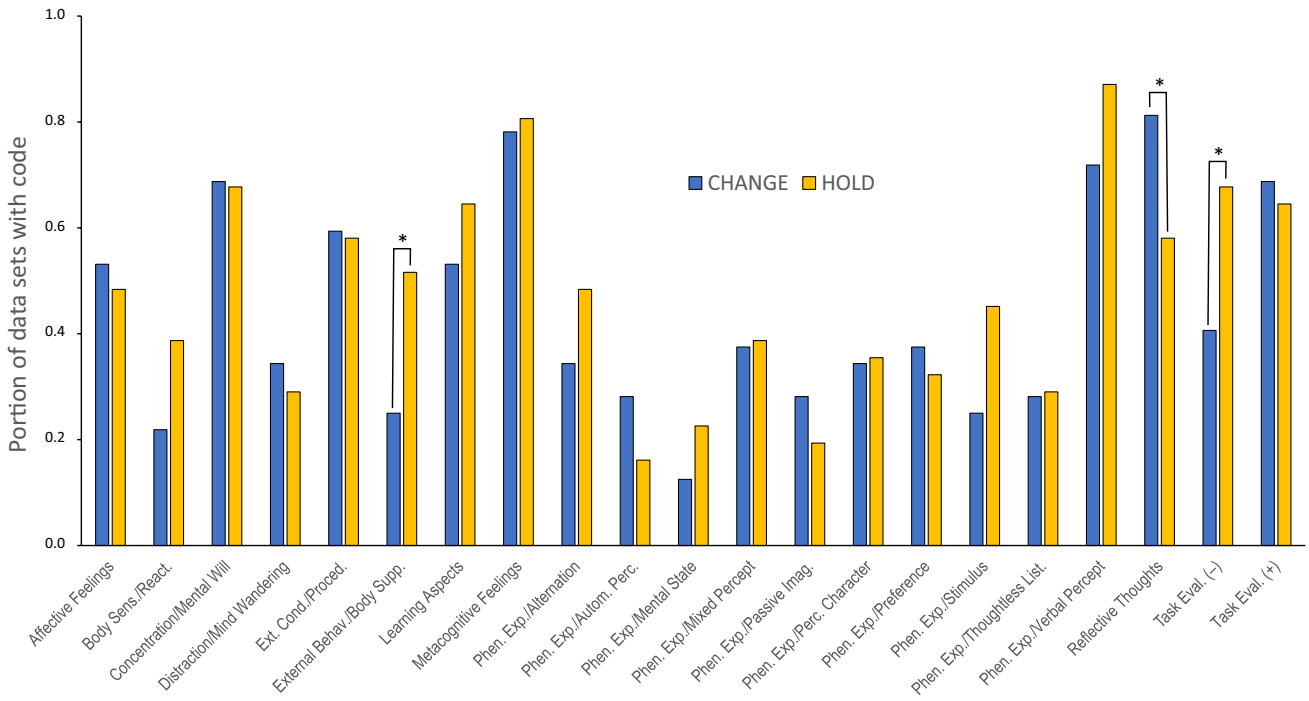

Fig. 2 Level 1: Multiple aspects of first-person experience. *$p < .046$, all others not significant, $p > .093$

they experience retrospectively evaluating the process and specific strategies, again from the subject side. Insofar as concentration/mental effort can be seen as a still undifferentiated expression of Level 2 micro-activities, and (at least the common forms of) metacognitive feelings reactively refer to completed processes, this again shows how Level 1 categories complementarily embed and contextualize Level 2 and 3 categories. The interlevel relations between metacognitive feelings and mental micro-activities will be presented below.

Three categories were identified that showed significant differences between the conditions. External behavior/body support (Cat. 3) seemed to be more deployed for hold ($M = 51.6\%$) than for change ($M = 25.0\%$), $\chi^2(1, N = 63) = 4.7$, $p = 0.030$, $w = 0.27$. Negative task evaluation (Cat. 7) also appeared to be higher for hold ($M = 67.7\%$) than for change ($M = 40.6\%$), $\chi^2(1, N = 63) = 4.7$, $p = 0.031$, $w = 0.27$. Finally, reflective thought (Cat. 11) was reported more frequently for change ($M = 81.3\%$) than for hold ($M = 58.1\%$), $\chi^2(1, N = 63) = 4.0$, $p = 0.045$, $w = 0.25$. The relevance of these exploratory investigations for the hypotheses will be discussed below.

Finally, the number of coded categories per data set was slightly higher for hold than for change but did not differ significantly ($M_{Change} = 9.2$, $SD_{Change} = 3.5$, $M_{Hold} = 10.0$, $SD_{Hold} = 3.2$), $p = 0.342$.

Level 2: Mental micro-activities

As with Level 1, we begin with some qualitative features that emerged during the bottom-up coding of the data, because while the basic structure of the four mental micro-activities was adopted top-down from the previous studies, their inner differentiation was still undetermined. What stands out and can be considered an important result of this study is the high differentiation of mental strategies in Producing (Cat. 2, qualitative part of Hypothesis 1), which has implications both for speech perception theory and mental agency. The main category of Producing is further divided into seven subcategories with three hierarchical levels of generality (Table 2). At medium level, quasi-visual (2.2 Imagining), quasi-articulative/auditory (2.5 Inner Speech), and active-emotional (2.7. Productive Emotions) strategies can be distinguished. This is much more than in the non-linguistic auditory study, where only two types of Producing were differentiated in terms of imaginative strategies and quasi-auditory anticipation of specific sounds, and without further differentiation (Wagemann, 2023). For the speech experiment, in contrast, there is an increase in semantic and thus also emotional aspects as well as aspects concerning the articulative structure of phonemes. At the most specific level, categories also show relations to each other, for example in imagining printed or written words (2.4), which in combination with letter spelling has a connection to inner speech or phoneme placement, or in productive emotions which are often (not always) induced by specific imaginations.

A second qualitative finding was that the distribution of micro-activities in the protocols sometimes showed an immediate succession of different forms. This could indicate an increased and differentiated agentive awareness among participants as opposed to scattered reporting of mental activities. To analyze this phenomenon quantitatively, an *activity cluster index* (ACI) was calculated as a ratio scale variable per data set by dividing the number of immediately adjacent codings (at Level 2) by the total number of codings minus one (to map the full range between 0 and 1). ACI scores varied between $M_{Visual\_Hold} = 0.40$ and $M_{Visual\_Change} = 0.57$ but not significantly across conditions (change, hold) and modalities (vision, audition, speech), which was tested by a one-way ANOVA, $F(4, 132) = 1.59$, $p = 0.181$. To give an impression of this phenomenon, some examples of activity clusters occurring in single sentences are shown in Fig. 3.

With regard to quantitative aspects, binarized relative frequencies of the four micro-activities and their subcomponents were compared with each other and between conditions and modalities. Firstly, Turning Away was significantly higher for hold ($M = 54.8\%$) than for change ($M = 28.1\%$), $\chi^2(1, N = 63) = 4.6$, $p = 0.031$, $w = 0.27$, whereas this was reversed for Turning Toward in that change ($M = 75.0\%$) was significantly higher than hold ($M = 45.2\%$), $\chi^2(1, N = 63) = 5.9$, $p = 0.016$, $w = 0.30$ (Fig. 4, Hypothesis 2). Secondly, comparing activity frequencies for different modalities and conditions, Producing for speech (change and hold, $M = 100\%$) was significantly higher than for auditory change ($M = 84.6\%$) and visual Producing frequencies (Hypothesis 1). This had to be demonstrated by an exact test according to Boschloo (1970) due to frequencies below five; the odds ratio was used to assess the effect size, which was corrected according to Haldane (1940) and Anscombe (1956) in case of zero values, $p = 0.029$, $OR = 13.00$ (Fig. 5). Thirdly, while frequencies of Producing in the speech experiment were constant across conditions, its subcomponents (subcategories) differed in several ways. While Imagining (General) and its subcategories Situation/Symbol and Emotions were reported significantly more often for change than for hold, Inner Speech was used significantly more often in the hold condition (see Fig. 6 and Table 4).

Finally, the number of coded categories per data set was slightly higher for change than for hold but did not differ significantly ($M_{Change} = 4.8$, $SD_{Change} = 1.7$, $M_{Hold} = 4.4$, $SD_{Hold} = 1.6$), $p = 0.283$.

Level 3: Intention and trying

As assumed by the qualitative part of hypothesis 2, differentiated forms of intentions and metacognitive feelings

Fig. 3 Level 2. Single sentence coding examples: Clustering of mental activities. Producing and perceiving are not differentiated into subcategories

a) After I had then practiced understanding this variant and was able to hear the speech stimulus more clearly by imagining myself shouting, I was also able to switch between the two variants more easily. (WP1_04_C)

b) As I try to hear both out, I imagine the desired and my brain blocks out that which prevents me from understanding the desired word. (WP1_05_C)

c) The moment I thought of "life," I also heard this. (WP1_14_H)

d) [...] I subtly (without neglecting active listening) try to make a sense or a symbol for it (picture bird, lettering LIFE). Likewise, I try to stay with the initial letter of the word so that I don't get distracted by the other word. (HP_03_H)

e) During the transition from one variant to the other, I first heard both words alternating, blending into each other, before it settled back on one of the two. (HP_12_H)

f) I noticed while listening how I was saying this word to myself in my head in order to actually recognize it from what I heard and to ascertain myself. (HP_15_H)

g) I used the short pauses between the sequences to listen away, and only by mentally speaking along (even during the pauses) was I able to achieve a change in listening. (WP3_04_C)

h) As soon as I focused more on the space after the "I", the bigger the pause seemed to me, that "I" moved more to the foreground and I understood two single words, i.e. "I scream". (WP3_12_C)

Turning Away Producing (with subcodes) Turning Toward Perceiving (with subcodes)

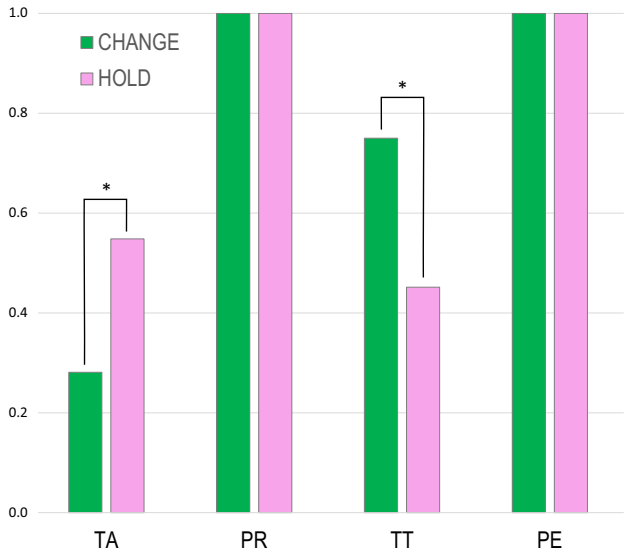

Fig. 4 Level 2: mental micro-activities (change vs. hold). TA: turning away (*$p = .031$); PR: producing; TT: turning toward (*$p = .016$); PE: perceiving

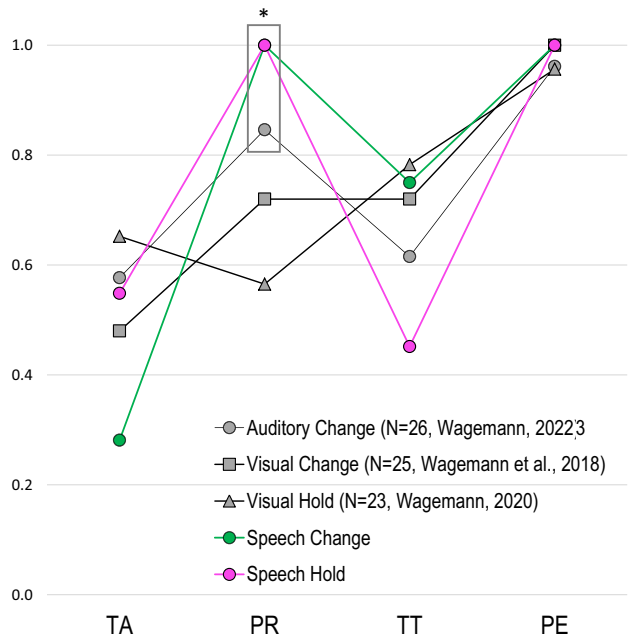

Fig. 5 Level 2: mental micro-activities (modalities and conditions). TA: turning away; PR: producing (*$p = .029$); TT: turning toward; PE: perceiving

were found in the data, which even allowed for quantitative analysis. Binarized frequencies of the three forms of intention showed several significant differences between modalities and conditions (Fig. 7). Firstly, executive intentions were significantly higher for auditory change ($M = 50.0\%$) than for speech change ($M = 21.9\%$), $\chi^2(1, N = 58) = 5.0$, $p = 0.025$, $w = 0.29$. Secondly, distal intentions were significantly more pronounced for speech hold ($M = 90.3\%$) than for speech change ($M = 59.4\%$), $\chi^2(1, N = 63) = 8.0$, $p = 0.005$, $w = 0.36$. Two further differences were observed for averaged speech conditions: Executive intentions ($M = 22.2\%$) were reported significantly less often than distal intentions ($M = 74.6\%$), $\chi^2(1, N = 63) = 17.3$, $p < 0.001$, $w = 0.52$, and proximal intentions ($M = 61.9\%$), $\chi^2(1, N = 63) = 10.2$, $p < 0.005$, $w = 0.40$.

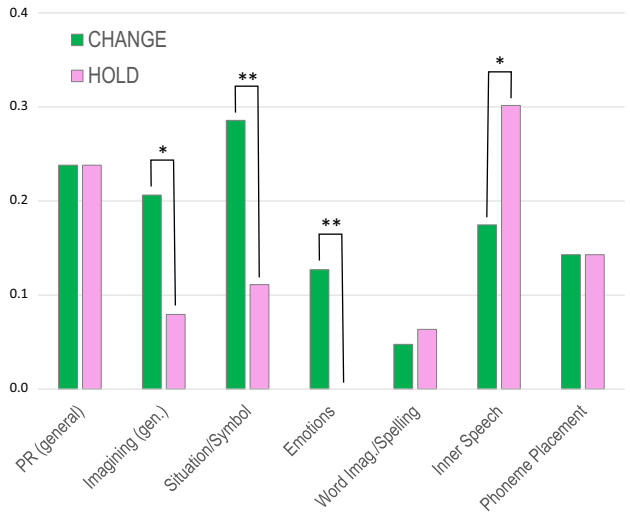

Fig. 6 Level 2: mental micro-activities (subcodes of producing). $*p < .033$, $**p < .0064$ (others not significant)

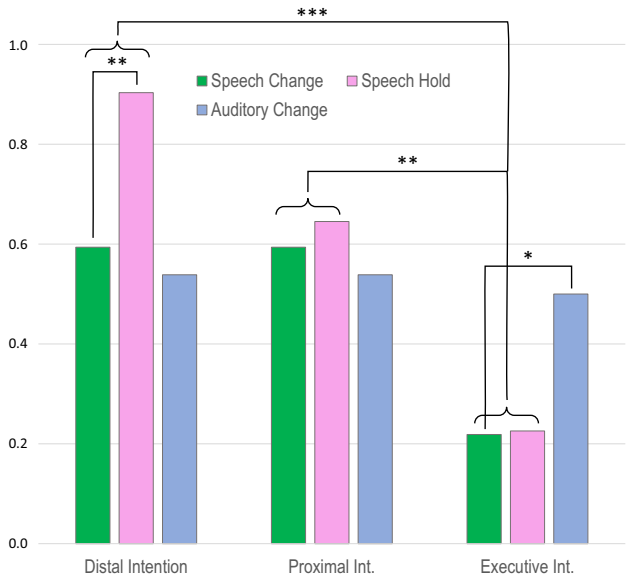

Fig. 7 Level 3: forms of intention across conditions and modalities. $*p = .025$, $**p < .005$, $***p < .001$ (others not significant)

Level 1–2: Metacognitive feelings and mental micro-activities

A subset of 63 data segments (distributed over 36 participants) out of a total of 160 coded under MCFs directly relates to the four forms of mental micro-activity as explained above (2.3.1). Due to clustering of different activities in identical or adjacent segments, individual MCFs are often assigned to more than one activity form, which means lower discriminatory power (and explains the higher sum of segments in Fig. 1). Qualitatively, beyond prevalent statements about ease or difficulty of task performance, negative MCFs (e.g., irritation, discomfort, frustration, unfamiliarity) were reported in more detail than positive MCFs (e.g., fascination, relaxation). In quantitative terms, MCFs occurred less frequently overall than intentions (203 segments in 62 participants) and less frequently than MCFs in the auditory change study (Figs. 8). More precisely, MCFs associated with Turning Away were significantly higher in auditory change ($M = 0.308$) than in speech change ($M = 0.031$), $p = 0.006$, $OR = 13.78$ (with exact test, see above), just as with Perceiving which was significantly higher in auditory change ($M = 0.885$) than in speech change ($M = 0.469$), $\chi^2(1, N = 58) = 11.0$, $p = 0.0001$, $w = 0.44$. Differences between speech change and hold partly seem to correspond with those of micro-activities (for Turning Away and Turning Toward, see Fig. 4) but were not significant.

Discussion

Summary and hypothesis-related evaluation of results

In the following, the major results of this study are summarized and implications for the hypotheses raised above are given, except for the qualitative part of Hypothesis 2 which is discussed in the next section. Initially, as a basis for more detailed considerations, the four-phase structure of mental micro-activities in perceptual reversals could be

Table 4 Level 2: **Producing across conditions**. An exact text according to Boschloo (1970) was used when frequencies occurred below five and supplemented by the odds ratio for effect size corrected according to Haldane (1940) and Anscombe (1956). Only significant results are shown

	Producing			
	Imagining (Gen.)	Situation/Symbol	Emotions	Inner Speech
Change	0.206	0.286	0.127	0.175
Hold	0.079	0.111	0.000	0.302
p	.031	.006	.001	.032
$\chi^2(1, N = 63)$	4.6	7.5	Exact test (Boschloo)	4.6
Effect Size	$w = 0.27$	$w = 0.34$	$OR = 25.47$	$w = 0.27$

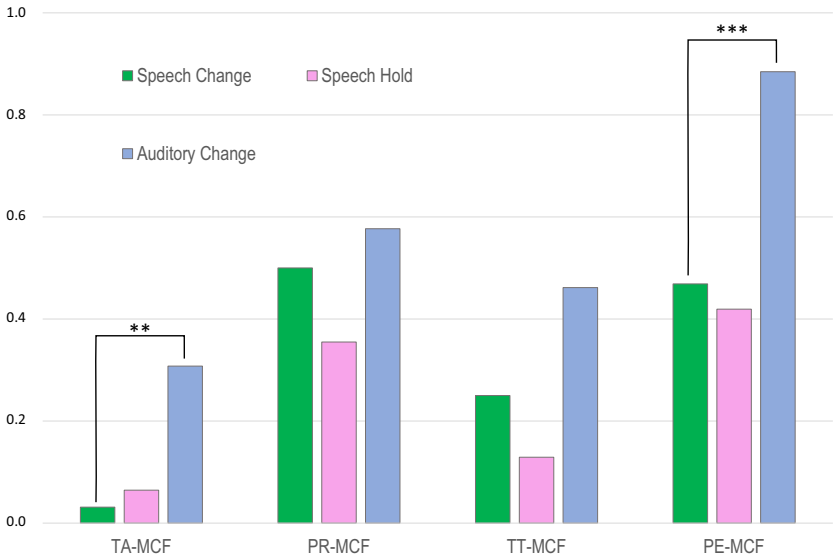

Fig. 8 Level 1–2: metacognitive feelings across conditions and modalities. ** $p = .006$, *** $p = .0001$ (others not significant)

reliably replicated and is thus extended from vision and non-linguistic audition to speech perception. This reinforces both the cross-modal nature of this activity structure and its potential for refining the classic three-phase attentional shift paradigm of Posner and Petersen (1990). Strictly speaking, we can even assume a five-phase dynamic if preliminary or partial Perceiving (category 4.1) is considered as a separate stage. Besides the classification of general forms of mental activity, the differentiation of Producing into seven aspects or three typical strategies (semantic – emotional – articulatory) represents a crucial qualitative outcome, which will be discussed below. Furthermore, the quantitative dependence of mental activities and their sub-aspects on experimental conditions and sensory modalities also supports their place in a perceptual change scenario integrating the first- and third-person perspective. As confirmation of Hypothesis 1, Producing takes a prominent position in that it was reported in speech perception by all participants (in both conditions), and significantly more often than in non-linguistic audition and vision (Fig. 5). Interpreting the higher frequency of a mental activity form as resulting from a more conscious exercise by participants, this finding appears consistent with the qualitatively more differentiated substructure of Producing (again, compared to audition and vision), as the qualitative part of Hypothesis 1. However, it would not have to follow from this that an increase of the frequencies (however achieved) for the other activity forms would lead to their further qualitative differentiation, since Turning Away (TA) and Turning Toward (TT) have a merely executive character related to the respective strategy. Rather, this could be a specific feature of Producing for the case of speech perception.

Also, Hypothesis 2 seems to be strengthened in that TA is significantly more often observed in the hold condition than in the change condition (Fig. 4). Moreover, the frequency of TT reacts inversely as it is significantly lower for hold than for change. Initially, this can be explained by differences in stimulus presentation, as participants in the hold condition were continuously exposed to auditory signals, whereas in the change condition they had some seconds of silence between stimulus presentations. Therefore, in the first case, the activity of suppressing unwanted parts of the stimulus might have been more challenged, whereas in the second case, the activity of anticipating the next stimulus presentation might have been more prominent. In the context of the Global Workspace/Working Memory model (Baars, 1988), this can be interpreted as selective attention, which due to a "constant-capacity storage mechanism" is limited to three to five chunks of information (Cowan et al., 2004, p. 634). However, since here it is not only about sensory attentional targets competing with each other, but also about discriminating self-performed mental activity forms, this situation could be understood as a sensory-mental dual-task. Therefore, in terms of a "hierarchical shifting of attention" between different levels of goals (Cowan, 2001, p. 93; see also Watzl, 2017), the observed effect could be explained by "dual-task costs to memory accuracy that favor a shared resource structure of working memory" (Doherty et al., 2019, p. 1549).

From here, we can also relate to a heretofore unsuspected effect for the substructure of Producing, as semantically driven imagination of suitable situations or symbols was higher for change, while the articulative strategy of inner or subvocal speech was higher for hold (Fig. 6). So, although Producing was mentioned equally often (by all) participants as the comprising activity form or phase in both conditions, the respective favored strategies can be broken down according to the experimental conditions. In this context, the relationship shown for TA and TT seems to be reversed insofar

as the imaginative-semantic strategy preferred for change acts rather distanced from the auditory stimulus, whereas for hold the subvocal articulation directly refers to stimulus-related aspects. In this respect, the choice of the mentally productive strategy could be seen as a compensation for the one-sidedness observed with regard to the executive activities (TA, TT). This is supported by the third strategy type of productive emotions occurring exclusively for change, as it is mostly related to imaginative or affective content and thus also more remote from the auditory stimulus than inner speech.

Perceptual penetrability and mental agency

Having thus addressed the results on speech perceptual reversal that are attentional in the more general sense, which, as indicated, can also be placed in models assuming an essentially unconscious or predominantly automatic conception of cognitive processes, we now come to the questions raised above about the connection between mental action and cognitive penetrability. Before evaluating the above considerations and our findings on intentions and metacognitive feelings in this context (qualitative part of Hypothesis 2), the phenomenological descriptions about gradually decomposed speech perception captured at Level 1 already provide an illuminating aspect. For how would perceptual reversals appear from a first-person perspective if perception were indeed cognitively impenetrable due to informational encapsulation of neural modules? According to this scenario, conscious experience would always be confronted with apodictic results of auditory or linguistic processing stages such as phonemes, syllables, entire words, and so on. Obviously, however, subjects experience not only unambiguously determined (intermediate) outcomes of modular processing stages, but also ambiguous and, above all, meaningless transitional forms between them (see examples given above). Even though this is rather the opposite of cognitive penetration, i.e., a gradual cognitive decomposition, we see this as a first questioning of a rigid encapsulation of linguistic and especially early auditory processing stages.

Furthermore, what argues for cognitive penetrability in the strict sense of conscious access and control are mental micro-activities that, contrary to phenomena of decomposition, explain the gradual construction or recomposition of word percepts. This extends the debate from cognitive or perceptual content to the volitional dynamics of processing as it appears in first-person experience, rather than limiting it to neural computation. In our framework, two stimulus-averted (TA, PR) and stimulus-oriented (TT, PE) phases or, respectively, two stimulus-nearer (TA, PE) and stimulus-remoter (PR, TT) phases can be distinguished. Conversely, the micro-activities can be classified according to their relationships with the conceptual structures to be produced and actualized for guiding the mental strategy. In sum, this is consistent, for example, with Steiner's (1861–1925) and Witzenmann's (1905–1988) structure-phenomenological approach to cognition in which conscious (e.g., perceptual) experience emerges from a dynamic intertwining of universal concepts and decomposed stimuli enabled by participatory mental activity (Steiner, 1988; Witzenmann, 2022). More recently, O'Callaghan has outlined a similar (and cross-modal) conception of perceptual objects as coherent compositions of sensory individuals, although he does not take mental activity into account (O'Callaghan, 2008). Some other studies do consider mental activity or mental acts in the context of cognitive penetrability but not in a more sophisticated way, let alone in empirical first-person research (e.g., Gross, 2017; Stins & Beek, 2012). If, on the other hand, mental activity is regarded as the driving force of cognitive penetration, which in turn can be subjected itself to cognitive penetration by introspective observation, an examination of its agentive status is required.

To this end, we first mention intentions directly related to mental activities, which were reported in at least one of their three forms by almost all participants. That distal (D-) intentions were significantly higher for speech hold than for speech change (Fig. 7), can be explained again by the continuous exposition of participants with challenging stimulus material, which is also reflected by the significantly higher negative task evaluation for hold (Level 1, Cat. 7, Fig. 2). That executive (E-) intentions for speech averaged across conditions were significantly weaker than the other two forms can again be explained by competing selective attention with respect to Producing. Conclusions about the agentive status of explicitly intended activities can be drawn by an analogy: In the context of criminal cases, suspects are examined not only based on circumstantial evidence, but also questioned about their motives. In the case of D-intentions, participants admittedly cannot be fully charged for subsequent mental processes, insofar as they simply follow the instructions and cooperatively try to meet the demands of the task (Gross, 2017; Orne, 1962). However, by expressing P- or E-intentions, they show that the mental acts associated with them, which are uninstructed but obviously necessary for successful task performance, are their own responsibility and are initiated by conscious attentional commands. Comparing P- and E-intentions, however, the former qualify their target activity (PR) more strongly as mental action than the latter, because here different strategies can be individually chosen and combined, whereas TA and TT represent rather "mechanical" basics of mental action with fewer possibilities for variation. Accordingly, we can state increasing evidence for mental actions via D- and E- up to P-intentions.

Concerning metacognitive feelings (MCFs) as a second criterion for mental action, two different kinds of *reference objects* can be distinguished: When performance of one

or more activities is experienced as more vs. less difficult (Arango-Muñoz & Michaelian, 2014), MCFs indicate points where resistance occurs in the process that can be dealt with worse or better. Therefore, on the one hand, they refer to what challenges or hinders the mental agent to implement their intentions and which can be found in the "persistence" of the stimulus adhering to unintended meaning (change condition) and its "unreliability" in not accepting the intended meaning (hold condition). From this ambivalence between conceptual determinateness and indeterminacy of the stimulus, together with the above findings about a decomposed or meaning-deprived phenomenality, inferences for the McDowell-Dreyfus debate on non-conceptual content of perception could be drawn (Schear, 2013; Witzenmann, 2022), which is out of the scope of this paper. At least we can point out that there seem to be both non-conceptual and conceptually imbued manifestations of perception, which is reasonable against the background of our dynamic approach. On the other hand, MCFs refer to what activity the mental agent performs and to how this resonates with the difficulties described, which is illustrated by opposite expressions like "frustration" and "fascination". As shown, the reference objects of MCFs in this case are the individual forms of mental activity, and even if these cannot always be unambiguously associated with particular MCFs in the data, they ultimately refer to the (potentially) conscious agent herself by whom they are performed. Overall, we think that both task- and self-related aspects of MCFs further strengthen the idea of a developable "agentive attention awareness" (Watzl, 2017, p. 232) comprising not only cognitive and volitional but also emotional dimensions.

This idea, however, could be relativized by subpersonal approaches to metacognition (de Sousa, 2009; Fields & Glazebrook, 2020; Proust, 2013), because the MCFs discussed so far are delivered to subjects in a receptive or reactive way and thus might stem from principally unconscious sources. However, just as perceptions which in everyday life usually appear as apodictic and ready given, emotions or feelings can possibly also be traced back to consciously executable micro-activities. Specifically, for MCFs we propose an extension from the standard case of receptive-reactive manifestations to productive and phenomenal-performative forms. First, productive emotions can be mentioned as a mental strategy that was autonomously developed by participants observing that certain reactively occurring feelings were associated with the semantic context of the target percepts. Based on this experience, they proceeded to intentionally induce precisely these feelings, whether through the detour of semantically appropriate imaginings or by directly assuming certain attitudes, in order to better achieve their perceptual goal. Since the feelings generated in this way do not refer primarily to specific external reference objects (these can vary greatly for one and the same intended feeling), but rather to a semantic self-stimulation of attention regulation, they can be justified as *productive metacognitive feelings*. Therefore, while previous research demonstrated effects of self-generated or self-induced feelings on neural processing (Damasio et al., 2000), sport performance (Rathschlag & Memmert, 2015), and emotion regulation (Zysberg & Raz, 2019), we suggest extending them to agentive attention awareness.

As a second extension, the feeling of what is it like to perform different mental micro-activities themselves can be regarded as a MCF in the context of the cognitive phenomenology (CP) debate (Bayne & Montague, 2011). To account for this, similarly to receptively registered emotions, the restriction of CP to cognitive states must be overcome to cognitive processes and their stages that are also introspectively accessible to the extent shown here. Then, in view of phenomenal contrasts between activity forms as confirmed by reliable coding (see above), the activities possess a performative phenomenality accordingly felt by the reporting subjects. Since, again, this phenomenality does not refer to the partially sensory mediated or associated reference objects of the activities (e.g., stimulus, mental representations), but to them themselves, it can even be evaluated as proprietary or irreducible. Ultimately, both kinds of extended MCFs, performative phenomenality and productive emotions, seem to fulfill the requirements of the CP thesis that "this phenomenology must be caused or determined by the cognitive attitude itself" (Arango-Muñoz, 2019, p. 5).

Conclusion

It goes without saying that the philosophical debates addressed cannot be treated with due depth within the framework of an empirical study. Nevertheless, this study proceeded from an experimental investigation of speech perception with two conditions (change vs. hold) in a first-person mixed methods design to the identification of a basic structure of mental micro-activities providing findings with significant theoretical and practical relevance. At a general level, the theoretical contribution of this study is that the addressed philosophical debates, largely unrelated, can be linked through a systematic analysis and interpretation of the first-person data. On the one hand, it has been shown that the issue of *cognitive penetrability* ultimately leads to the question of what– gradually decomposed stimuli – is to be penetrated (or not) with what – conceptual structures – as discussed in the *McDowell-Dreyfus debate*. On the other hand, the question of how this penetration is achieved can be answered by structured mental micro-activities, which can be classified as *mental actions* according to first-person criteria such as intention and metacognitive feelings. Moreover, touching on the cognitive phenomenology debate,

productive metacognitive feelings and the differentiation of four micro-activities establish a performative phenomenality that does not seem to be reducible to sensory input, reactive emotions, or other state-like mental contents, as they independently describe the conscious process quality of changing or holding a certain percept. In sum, this can be seen as strengthening agentive self-awareness in cognition in the context of participatory reality formation (Froese, 2022; Steiner, 1988; Witzenmann, 2022).

In terms of speech perception, the major finding of this study with theoretical implications consists in the unexpected high degree of conscious access particularly to stimulus-near stages of the perceptual process. In the lexical segmentation task, participants were able to use both productive mental strategies and executive activities to advance to the formation of phonemes from phones or even more incoherent or ambiguous stimulus fragments and to intentionally influence it according to the experimental conditions. The frequency patterns of the cross-modal activity structure behaved characteristically in terms of reported activity forms and strategies closer and farther away from the stimulus, which is consistent with limited attentional resources in the context of a sensory-mental dual task and demonstrates the connectivity of our findings to the Global Workspace Theory (Baars, 1988). Through the lens of Construal Level Theory (Trope & Liberman, 2010), the interplay of different forms of intention in perceptual change provides a dynamic integration of high-level and low-level presentations of the same object, i.e., the speech stimulus presented. Distal intentions operate at a higher, more abstract level of construal encompassing the goal of perceptual change but not the strategic-executive way in which it is achieved. The latter, however, is incorporated in proximal and executive intentions aiming at specific micro-activities which thus highlight self-control at a lower construal level and extend the conventional view that self-control only increases with construal level (Fujita et al, 2006). This can be explained by our finding that perceptual change cannot be achieved by distal intentions or high-level construal alone (Hansen, 2019), but is significantly supported by both concrete aspects of strategic content (proximal intentions) and steps of processing (executive intentions). Optimal self-control is therefore probably not established through a one-sided prioritization of a certain construal level, but rather through their dynamic and balanced interaction. Referring back to cognitive penetrability of perception this can also be understood in reverse as perceptual or attentional penetrability of cognitive processes. Thus possibly even constitutive aspects of speech perception which are usually thought to be inaccessible to consciousness are shown to possess a first-person phenomenal and agentive side.

While this, of course, does not disprove the relevance of neural processing, it shifts the ratio between conscious and unconscious aspects of linguistic cognition toward the former, which has both implications for practical applications and future research. To begin with the former, mental action and strategy use play a crucial role in second language (L2) learning (e.g., Burns & Richards, 2018) but have so far been difficult to explain (Macaro (2006), or only with reference to subpersonal, connectionist or working memory theories (Moonen et al., 2006; Driessen et al., 2008). Here, our consciousness-immanent approach to mental agency in speech perception can be considered not only for novel theory building (see above) but also in L2 educational settings, where the role of self-regulated listening and the impact of metacognitive skills on it are increasingly recognized (Chamot et al., 1999; Field, 1998; Teng et al., 2021; Yokomoto et al., 2021). In this context, Goh (2008, p. 191) explicitly recommends teachers "to show learners the mental activities that they engage in to construct their understanding of listening texts", and Vandergrift (2003, p. 487) describes his approach to listening instruction as "orchestrating strategies in a continuous metacognitive cycle". This does not only apply to the level of phrases and words but also to lexical segmentation (Vandergrift, 2004) which needs practice in perception skills (Goh, 2002; Hulstijn, 2001) and especially attention to pause-bounded linguistic units (Harley, 2000). Here, we point to our findings underlining the option to become conscious of mental micro-activities in meaning anticipation and pre-listening strategies (Goh, 2002; Ur, 1984; see change condition) and ambiguity tolerance (Chu et al., 2015; Varasteh et al., 2016; see hold condition) and deliberately make use of them. In particular, the broad range of productive (semantic, emotional, articulatory) strategies integrates top-down and bottom-up approaches to listening instruction, instead of polarizing them (Goh, 2002; Hulstijn, 2001), and offers an experiential, self-efficacious, and playful handling of language and access to otherwise abstract linguistic concepts.

Beyond the specific focus on language learning, our findings may also have practical significance for broader education contexts, as the basic structure of micro-activities could be strengthened not only for speech perception but also for vision and non-linguistic audition and, in modified forms, even for thought processes and social interaction (see introduction). Regarding the critical role of self-development for motivation and learning (McCombs, 1990; Dutta & Dubey, 2008), we propose to integrate these dimensions of agentive self-awareness and participatory reality formation into cross-curricular aspects of higher education such as self-regulated learning (Zimmerman, 2002), student agency (Inouye et al., 2022), or mindfulness training (Reavley, 2018). For example, the current study itself took place in a higher education context of teacher training, where students performed not only as participants but also practiced exemplary data analysis and learned experientially and theoretically about

the significance of self-awareness and self-control in mental agency for professional development. In that way, referring to our above considerations about Construal Level Theory, deeper and more sustainable forms of learning can be developed in which students not only have to work through abstract content but also intrinsically connect with it by involving themselves into concrete cognitive processes – and thus become more concrete themselves as cognitive agents.

Before giving an outlook on future research, the generalizability and limitations of the current study should be outlined. Although most psychological studies are based on students as participants, this obviously does not satisfy the needs of generalizability (Hanel & Vione, 2016). Nonetheless, while it is hardly possible to transfer the results of this study to other populations, it is reasonable to assume that, at least for this population, the basic structure of micro-activities demonstrated not only for vision and audition, but now also for speech perception, can be generalized to other sensory modalities and perhaps even other cognitive processes, as our own studies on thought processes and social cognition suggest. Another limitation refers to the first-person methodology deployed in this study, as it uses only one kind of qualitative verbal data. Although it is exactly this approach that leads to the significant results of this study, their scope and validity could certainly be further enhanced by triangulation with other types of data (e.g., external behavior, neurophysiological measures). Therefore, in terms of future research, replications with varied tasks and populations as well as methodological extensions are needed to further develop the approach of this study and to improve the basis for generalizability. Specifically, a neurophenomenological extension of our mixed-methods approach to (speech) perception would lend itself to further inquiry, not only because cognitive neuroscience has become a gold-standard of research but also because our findings imply a detailed research agenda. More precisely, specific tasks could be developed focusing only on one mental micro-activity or strategy at a time to identify corresponding neural correlates which then could be traced in complete perceptual reversal settings. By triangulating first- and third-person data in this way, it may be possible to gain further insight into the temporal dynamics of mental and neural phenomena (e.g., in relation to the phases of attentional shift), paving the way for a more fine-grained exploration of the nature of their connection.

Acknowledgements We would like to thank Jonas Raggatz for supporting the study with intercoder reliability testing.

Authors' contributions JW: Study design and implementation, qualitative and quantitative analyses, and structure and main body of the article. AW: Data acquisition, intercoder reliability testing, literature review and parts of the introduction, methodological discussion, and article revision.

Funding This study was conducted without external funding.

Data availability The speech stimuli and raw datasets (German) generated and analyzed during this study are available under https://osf.io/dnxfc/?view_only=b18a5a2f4f764ba4b53cf668fce64c81

Declarations

Ethical statement This study has been approved by the Institutional Board of Alanus University (Campus Mannheim) University. Participation in this project involved no risks that went beyond the risks of normal life. Participants received information about the study and its purpose and participated voluntarily.

Consent for publication Participants consented to the publication of their anonymized verbal reports in research contexts.

Competing interests The authors declare that they have no competing interests.

References

Anscombe, F. (1956). On estimating binomial response relations. *Biometrika, 43*, 461–464.

Arango-Muñoz, S. (2019). Cognitive phenomenology and metacognitive feelings. *Mind and Language, 34*(2), 247–262.

Arango-Muñoz, S., & Bermúdez, J. P. (2018). Remembering as a mental action. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New directions in the philosophy of memory* (pp. 75–96). Routledge.

Arango-Muñoz, S., & Michaelian, K. (2014). Epistemic feelings, epistemic emotions: Review and introduction to the focus section. *Philosophical Inquiries, 2*(1), 97–122.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Barraza, P., Jaume-Guazzini, F., & Rodríguez, E. (2016). Pre-stimulus EEG oscillations correlate with perceptual alternation of speech forms. *Neuroscience Letters, 622*, 24–29. https://doi.org/10.1016/j.neulet.2016.04.038

Bayne, T., & Montague, M. (2011). *Cognitive phenomenology*. Oxford University Press.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonation structure in Japanese and English. *Phonology, 3*(01), 255–309.

Bickerton, D. (2014). Some problems for biolinguistics. *Biolinguistics, 8*, 73–96.

Boschloo, R. D. (1970). Raised conditional level of significance for the 2x2-table when testing the equality of two probabilities. *Statistica Neerlandica, 24*(1), 1–9.

Brehm, L., & Goldrick, M. (2016). Empirical and conceptual challenges for neurocognitive theories of language production. *Language, Cognition and Neuroscience, 31*(4), 504–507.

Brent, M., & Titus, L. M. (Eds.). (2023). *Mental action and the conscious mind*. Routledge.
Buckareff, A. A. (2005). How (not) to think about mental action. *Philosophical Explorations, 8*(1), 83–89.
Burns, A., & Richards, J. C. (Eds.). (2018). *The Cambridge guide to learning English as a second language*. Cambridge University Press.
Buzsaki, G. (2019). *The brain from inside out*. Oxford University Press.
Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294–320.
Chamot, A. U., Barnhardt, S., El-Dinary, P. B., & Robbins, J. (1999). *The learning strategies handbook*. Longman.
Cheung, K. K. C., & Tai, K. W. H. (2021). The use of intercoder reliability in qualitative interview data analysis in science education. *Research in Science & Technological Education.* https://doi.org/10.1080/02635143.2021.1993179
Chu, W. H., Lin, D. Y., Chen, T. Y., Tsai, P. S., & Wang, C. H. (2015). The relationships between ambiguity tolerance, learning strategies, and learning Chinese as a second language. *System, 49*, 1–16. https://doi.org/10.1016/j.system.2014.10.015
Chung, C. K., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social Communication* (pp. 343–359). Psychology Press.
Clarke, S. (2021). Cognitive penetration and informational encapsulation: Have we been failing the module? *Philosophical Papers, 178*, 2599–2620.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–185.
Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science, 15*(9), 634–640. https://doi.org/10.1111/j.0956-7976.2004.00732.x
Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Sage Publications.
Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L. B., Parvizi, J., & Hichwa, R. D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience, 3*(10), 1049–1056.
Davidson, G. D., & Pitts, M. A. (2014). Auditory event-related potentials associated with perceptual reversals of bistable pitch motion. *Frontiers in Human Neuroscience, 8*, Article 572. https://doi.org/10.3389/fnhum.2014.00572
Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research, 229*(1–2), 132–147. https://doi.org/10.1016/j.heares.2007.01.014
De Sousa, R. (2009). Epistemic feelings. *Mind and Matter, 7*(2), 139–161.
Doherty, J. M., Belletier, C., Rhodes, S., Jaroslawska, A., Barrouillet, P., Camos, V., Cowan, N., Naveh-Benjamin, M., & Logie, R. H. (2019). Dual-task costs in working memory: An adversarial collaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(9), 1529–1551. https://doi.org/10.1037/xlm0000668
Driessen, D., Westhoff, G., Haenen, J., & Brekelmans, M. (2008). A qualitative analysis of language learning tasks: The design of a tool. *Journal of Curriculum Studies, 40*(6), 803–820.
Dutta, J., & Dubey, P. K. (2008). Teacher-student relationships and interactions on self-development and motivation. In I. D. George, J. G. Valan Arasu, P. Agrawal, & M. Gupta (Eds.), *Quality education: prospects and challenges* (pp. 166–179). APH Publishing.
Dworkin, S. L. (2012). Sample size policy for qualitative studies using in-depth interviews. *Archives of Sexual Behavior, 41*, 1319–1320.
Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/bf03193146
Field, J. (1998). Skills and strategies: Towards a new methodology for listening. *ELT Journal, 52*, 110–118.
Fiebich, A., & Michael, J. (2015). Mental actions and mental agency. *Review of Philosophy and Psychology, 6*(4), 683–693.
Field, J. (2008). *Listening in the language classroom*. Cambridge: Cambridge University Press.
Fields, C., & Glazebrook, J. F. (2020). Do process-1 simulations generate the epistemic feelings that drive process-2 decision making? *Cognitive Processing, 21*(4), 533–553.
Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. The Behavioral and Brain Sciences *39*, e229. https://doi.org/10.1017/S0140525X15000965
Fodor, J. A. (1983). *The modularity of mind*. The MIT Press.
Frank, S., & Willems, R. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience, 32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109
Friederici, A. D. (2017). *Language in our brain: The origins of a uniquely human capacity*. MIT Press.
Froese, T. (2022). Scientific observation is socio-materially augmented perception: Toward a participatory realism. *Philosophies, 7*(2), 37. https://doi.org/10.3390/philosophies7020037
Fugard, A. J. B., & Potts, H. W. W. (2015). Supporting thinking on sample sizes for thematic analyses: A quantitative tool. *International Journal of Social Research Methodology: Theory & Practice, 18*(6), 669–684.
Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and selfcontrol. *Journal of Personality and Social Psychology, 90*(3), 351–367. https://doi.org/10.1037/0022-3514.90.3.351
Getz, L. M., & Toscano, J. C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological Science, 30*(6), 830–841.
Goh, C. (2002). Exploring listening comprehension tactics and their interaction patterns. *System, 30*, 185–206.
Goh, C. C. M. (2008). Metacognitive instruction for second language listening development: Theory, Practice and Research Implications. *RELC Journal, 39*, 188–213.
Goh, C. C. M., & Wallace, M. (2018). Lexical segmentation in listening. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1379–1385). John Wiley & Sons.
Gow, D. W., Jr., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21*(2), 344–359.
Gregory, R. (1966). *Eye and brain: The psychology of seeing*. McGraw-Hill Book Company.
Gross, S. (2017). Cognitive penetration and attention. *Frontiers in Psychology, 8*, Article 221. https://doi.org/10.3389/fpsyg.2017.00221
Guest, G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PloS One, 15*(5), e0232076.

Haldane, J. (1940). The mean and variance of the moments of chi-squared, when used as a test of homogeneity, when expectations are small. *Biometrika, 29*, 133–143.

Hanel, P. H., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PloS one, 11*(12), e0168354.

Hansen, J. (2019). Construal level and cross-sensory influences: High-level construal increases the effect of color on drink perception. *Journal of Experimental Psychology: General, 148*(5), 890–904. https://doi.org/10.1037/xge0000548

Harley, B. (2000). Listening strategies in ESL: Do age and LI make a difference? *TESOL Quarterly, 34*, 769–776.

Harley, T. A. (2014). *The psychology of language: From data to theory* (4th ed.). Psychology Press.

Heald, S. L., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*, 35. https://doi.org/10.3389/fnsys.2014.00035

Hickok, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders, 45*(6), 393–402.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding: A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849–931.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge University Press.

Inouye, K., Lee, S., & Oldac, Y. I. (2022). A systematic review of student agency in international higher education. *Higher Education*, 1–21. Advance online publication. https://doi.org/10.1007/s10734-022-00952-3

Intaitė, M., Koivisto, M., Rukšėnas, O., & Revonsuo, A. (2010). Reversal negativity and bistable stimuli: Attention, awareness, or something else? *Brain and Cognition, 74*, 24–34. https://doi.org/10.1016/j.bandc.2010.06.002

Johnson, G. (2009). Mechanisms and functional brain areas. *Mind and Machines, 19*, 255–271.

Johnson, S. P. (2010). How infants learn about the visual world. *Cognitive Science, 34*(7), 1158–1184.

Kee, H. (2020). Horizons of the word: Words and tools in perception and action. *Phenomenology and the Cognitive Sciences, 19*(5), 905–932.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 169–226). The MIT Press.

Kornmeier, J., Friedel, E., Hecker, L., Schmidt, S., & Wittmann, M. (2019). What happens in the brain of meditators when perception changes but not the stimulus? *PLoS ONE, 14*(10), e0223843. https://doi.org/10.1371/journal.pone.0223843

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 363*(1493), 979–1000.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science, 52*(2), 217–276.

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica, 5*, 5–54.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America, 51*(6, Pt. 2), 2018–2024.

Le Prell, C. G., & Clavier, O. H. (2017). Effects of noise on speech recognition: Challenges for communication by service members. *Hearing Research, 349*, 76–89. https://doi.org/10.1016/j.heares.2016.10.004

Lee, Y., Kaiser, E., & Goldstein, L. (2020). I scream for ice cream: Resolving lexical ambiguity with sub-phonemic information. *Language and Speech, 63*(3), 526–549.

Liu, Y. W., Cheng, X., Chen, B., Peng, K., Ishiyama, A., & Fu, Q. J. (2018). Effect of tinnitus and duration of deafness on sound localization and speech recognition in noise in patients with single-sided deafness. *Trends in Hearing, 22*, 2331216518813802. https://doi.org/10.1177/2331216518813802

Macaro, E. (2006). Strategies for language learning and for language use: Revising the theoretical framework. *Modern Language Journal, 90*, 320–337.

Maciuszek, J. (2018). Lexical access in the processing of word boundary ambiguity. *Social Psychological Bulletin, 13*(4), Article e28690. https://doi.org/10.32872/spb.v13i4.28690

Mayring, P. (2000). Qualitative content analysis. Forum: Qualitative. *Social Research, 1*(2). Retrieved October, 05, 2022, from https://www.qualitative-research.net/index.php/fqs/article/view/1089/2385

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10*(8), 363–369. https://doi.org/10.1016/j.tics.2006.06.007

McClelland, T. (2019). Representing our options: The perception of affordances for bodily and mental Action. *Journal of Consciousness Studies, 26*(3-4), 155–180.

McCombs, B. L., & Marzano, R. J. (1990). Putting the self in self-regulated learning: the self as agent in integrating will and skill. *Educational Psychologist, 25*, 51–69.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*(2), 143–178.

Mele, A. R. (1992). *Springs of action*. Oxford University Press.

Mondal, P. (2022). A critical perspective on the (neuro)biological foundations of language and linguistic cognition. *Integrative Psychological & Behavioral Science*. https://doi.org/10.1007/s12124-022-09741-0

Montagne, C., & Zhou, Y. (2016). Visual capture of a stereo sound: Interactions between cue reliability, sound localization variability, and cross-modal bias. *The Journal of the Acoustical Society of America, 140*(1), 471. https://doi.org/10.1121/1.4955314

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior & Development, 16*(4), 495–500.

Moonen, M., de Graaff, R., & Westhoff, G. (2006). Focused tasks, mental actions and second language learning. Cognitive and connectionist accounts of task effectiveness. International Journal of. *Applied Linguistics, 152*(1), 35–53.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*, 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience, 31*(1), 4–18.

O'Callaghan, C. (2008). Object perception: vision and audition. *Philosophy Compass, 3*(4), 803–829.

O'Callaghan, C. (2009). Audition. In S. Robins, J. Symons, & P. Calvo (Eds.), *Routledge Companion to the Philosophy of Psychology* (pp. 579–591). Routledge.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1–13. https://doi.org/10.1177/1609406919899220

O'Shaughnessy, B. (2000). *Consciousness and the world*. Oxford University Press.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–783.

Owens, D. (2009). Freedom and practical judgement. In L. O'Brien & M. Soteriou (Eds.), *Mental actions* (pp. 121–237). Oxford University Press.

Patel, P., van der Heijden, K., Bickel, S., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2022). Interaction of bottom-up and top-down neural mechanisms in spatial multi-talker speech perception. *Current Biology, 32*(18), 3971–3986.e4. https://doi.org/10.1016/j.cub.2022.07.047

Peacocke, C. (2007). Mental action and self-awareness (I). In B. McLaughlin & J. D. Cohen (Eds.), *Contemporary debates in philosophy of mind* (pp. 358–376). Blackwell.

Pennebaker, J. W., Boyd, R., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. University of Texas at Austin. https://doi.org/10.15781/T29G6Z

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive Sciences, 5*, 229–269.

Pitts, M. A., & Britz, J. (2011). Insights from intermittent binocular rivalry and EEG. *Frontiers in Human Neuroscience, 5*, 107. https://doi.org/10.3389/fnhum.2011.00107

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology, 29*(1-2), 34–55.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Reviews in Neuroscience, 13*, 25–42.

Proust, J. (2001). A plea for mental acts. *Synthese, 129*, 105–128.

Proust, J. (2010). Mental acts. In T. O'Connor & C. Sandis (Eds.), *A companion to the philosophy of action* (pp. 209–217). Wiley-Blackwell.

Proust, J. (2013). *The philosophy of metacognition*. Oxford University Press.

Proust, J. (2015). The representational structure of feelings. In T. Metzinger & J. M. Windt (Eds.), *OpenMind*. Frankfurt am Main: Johannes-Gutenberg Universität. https://doi.org/10.25358/openscience-77

Rathschlag, M., & Memmert, D. (2015). Self-generated emotions and their influence on sprint performance: An investigation of happiness and anxiety. *Journal of Applied Sport Psychology, 27*(2), 186–199.

Reavley, N. J. (2018). Mindfulness training in higher education students. *The Lancet Public Health, 3*(2), e55–e56. https://doi.org/10.1016/S2468-2667(17)30241-4

Redford, M. A., & Baese-Berk, M. (2023). Acoustic theories of speech perception. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.742

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*(4497), 947–950. https://doi.org/10.1126/science.7233191

Roberts, B., Summers, R. J., & Bailey, P. J. (2010). The perceptual organization of sine-wave speech under competitive conditions. *The Journal of the Acoustical Society of America, 128*(2), 804–817.

Rock, I. (1983). *The logic of perception*. MIT Press.

Rude, S. S., Gortner, E.-M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion, 18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Schear, J. K. (Ed.). (2013). *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.

Seih, Y. T., Chung, C. K., & Pennebaker, J. W. (2011). Experimental manipulations of perspective taking and perspective switching in expressive writing. *Cognition and Emotion, 25*(5), 926–938.

Small, M. L. (2011). How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual Review of Sociology, 37*, 57–86.

Steiner, R. (1988). *The science of knowing: Outline of an epistemology implicit in the Goethean world view* (1st ed., 1886). Mercury Press.

Stins, J. F., & Beek, P. J. (2012). A critical evaluation of the cognitive penetrability of posture. *Experimental Aging Research, 38*(2), 208–219.

Teng, M. F., Wang, C., & Wu, J. G. (2021). Metacognitive strategies, language learning motivation, self-efficacy belief, and English achievement during remote learning: a structural equation modelling approach. *RELC Journal*. https://doi.org/10.1177/00336882211040268

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463. https://doi.org/10.1037/a0018963

Ur, P. (1984). *Teaching listening comprehension*. Cambridge: Cambridge University Press.

Vandergrift, L. (2003). Orchestrating strategy use: Toward a model of the skilled second language listener. *Language Learning, 53*, 463–496.

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics, 24*, 3–25.

Valaris, M. (2023). Reasoning and mental action. In M. Brent & L. M. Titus (Eds.), *Mental action and the conscious mind* (pp. 142–163). New York: Routledge.

Varasteh, H., Ghanizadeh, A., & Akbari, O. (2016). The role of task value, effort-regulation, and ambiguity tolerance in predicting EFL learners' test anxiety, learning strategies, and language achievement. *Psychological Studies, 61*, 2–12. https://doi.org/10.1007/s12646-015-0351-5

Vermersch, P. (1999). Introspection as practice. *Journal of Consciousness Studies, 6*(2–3), 17–42.

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science, 10*(2), 159–164.

Wagemann, J. (2020). Mental action and emotion—What happens in the mind when the stimulus changes but not the perceptual intention. *New Ideas in Psychology, 56*, Article 100747.

Wagemann, J. (2022). Exploring the structure of mental action in directed thought. *Philosophical Psychology, 35*(2), 145–176.

Wagemann, J. (2023). Voluntary auditory change: First-person access to agentive aspects of attention regulation. *Current Psychology, 42*, 15169–15185.

Wagemann, J., Edelhäuser, F., & Weger, U. (2018). Outer and inner dimensions of brain and consciousness—Refining and integrating the phenomenal layers. *Advances in Cognitive Psychology, 14*(4), 167–185.

Wagemann, J., & Raggatz, J. (2021). First-person dimensions of mental agency in visual counting of moving objects. *Cognitive Processing, 22*(3), 453–473.

Wagemann, J., & Weger, U. (2021). Perceiving the other self. An experimental first-person account to non-verbal social interaction. *American Journal of Psychology, 134*(4), 441–461.

Wagemann, J., Tewes, C., & Raggatz, J. (2022). Wearing face masks impairs dyadic micro-activities in nonverbal social encounter. A mixed-methods first-person study on the sense of I and Thou. *Frontiers in Psychology, 13*, Art. 983652. https://doi.org/10.3389/fpsyg.2022.983652

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology, 71*, 612–613.

Watzl, S. (2017). *Structuring mind*. Oxford University Press.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*(3), 1707–1717.

Witzenmann, H. (2022). Structure phenomenology. In T. Vine, & J. Wagemann (Eds.), *Preconscious formation in the epistemic disclosure of reality*. Bloomsbury Academic. (Original work published 1983).

Wu, W. (2013). Mental action and the threat of automaticity. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the will* (pp. 244–261). Oxford University Press.

Yokomoto, K., Tsunemoto, A., & Suzukida, Y. (2021). Effects of awareness-raising activities on Japanese university students' listening comprehension of World English pronunciation. *Lingua (Sophia University Bulletin), 32*, 97–113.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Zysberg, L., & Raz, S. (2019). Emotional intelligence and emotion regulation in self-induced emotional states: Physiological evidence. *Personality and Individual Differences, 139*(1), 202–207.

The full reference list can be accessed via.
https://osf.io/dnxfc/?view_only=b18a5a2f4f764ba4b53cf668fce64c81.