Barbara Höhle | Tom Fritzsche | Katharina Meß | Mareike Philipp | Adamantios Ionannis Gafos

Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning

Suggested citation referring to the original publication:
Developmental Science 23 (2020) 5, Art. e12950 pp. 1 - 16
DOI https://doi.org/10.1111/desc.12950
ISSN 1363-755X, 1467-7687

Journal article | Version of record

Secondary publication archived on the Publication Server of the University of Potsdam:
Zweitveröffentlichungen der Universität Potsdam : Humanwissenschaftliche Reihe 868
ISSN: 1866-8364
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-516674
DOI: https://doi.org/10.25932/publishup-51667

Received: 2 April 2019 | Revised: 22 December 2019 | Accepted: 27 January 2020

DOI: 10.1111/desc.12950

PAPER

Developmental Science WILEY

Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning

Barbara Höhle | Tom Fritzsche | Katharina Meß | Mareike Philipp | Adamantios Gafos

Department of Linguistics, Cognitive Sciences, University of Potsdam, Potsdam, Germany

Correspondence
Barbara Höhle, Department of Linguistics, Cognitive Sciences, University of Potsdam, Karl-Liebknecht-Str. 24-25, Potsdam D-14476, Germany.
Email: hoehle@uni-potsdam.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 317633480 – SFB 1287

Abstract

Seminal work by Werker and colleagues (Stager & Werker [1997] *Nature*, 388, 381–382) has found that 14-month-old infants do not show evidence for learning minimal pairs in the habituation-switch paradigm. However, when multiple speakers produce the minimal pair in acoustically variable ways, infants' performance improves in comparison to a single speaker condition (Rost & McMurray [2009] *Developmental Science*, 12, 339–349). The current study further extends these results and assesses how different kinds of input variability affect 14-month-olds' minimal pair learning in the habituation-switch paradigm testing German learning infants. The first two experiments investigated word learning when the labels were spoken by a single speaker versus when the labels were spoken by multiple speakers. In the third experiment we studied whether non-acoustic variability, implemented by visual variability of the objects presented together with the labels, would also affect minimal pair learning. We found enhanced learning in the multiple speakers compared to the single speaker condition, confirming previous findings with English-learning infants. In contrast, visual variability of the presented objects did not support learning. These findings both confirm and better delimit the beneficial role of speech-specific variability in minimal pair learning. Finally, we review different proposals on the mechanisms via which variability confers benefits to learning and outline what may be likely principles that underlie this benefit. We highlight among these the multiplicity of acoustic cues signalling phonemic contrasts and the presence of relations among these cues. It is in these relations where we trace part of the source for the apparent paradoxical benefit of variability in learning.

KEYWORDS

acoustic variability, habituation-switch paradigm, infant word learning, minimal pairs, phonological development, visual variability

1 | INTRODUCTION

Even very young infants have an amazing capacity to learn words. Already at an age of 6 months, they show a basic understanding of highly familiar words (Bergelson & Swingley, 2013), start to produce their first words around their first birthday and their second year of life is characterized by a rapid increase in vocabulary size (e.g. Fenson et al., 1994).

In a seminal study, Stager and Werker (1997) reported findings showing that 14-month-old English-learning infants have the

capacity to learn word-object associations after a short exposure in an experimental setting that is now known as the habituation-switch paradigm. In this experimental paradigm, infants are habituated to two label-object pairs until their looking times to the visually presented objects decrease below a pre-specified level. In the immediately following testing phase, the infants are presented with the habituated label-object pairs (so-called 'same' trials) and with trials in which the learnt association is disturbed by combining one of the objects with the label that had been habituated with the other object (so-called 'switch' trials). Infants' longer looking to the presented object during the switch compared to the same trials (switch effect) is considered as an indication that infants have formed label-object associations during the habituation phase, a first step of word learning. In their original study, Stager and Werker (1997) revealed a switch effect for 14-month-old infants when the labels used within one experiment were phonologically dissimilar (e.g. [lɪf] vs. [ni:m]) but not when the labels formed a minimal pair (e.g. [bɪ] vs. [dɪ]). This result has been replicated in several follow-up studies using various sound contrasts (Archer, Ference, & Curtin, 2014; Pater, Stager, & Werker, 2004; Rost & McMurray, 2009; Werker, Fennell, Corcoran, & Stager, 2002). As infants discriminated the sound contrast when no referent object was presented (Stager & Werker, 1997), the puzzling question about the source of this failure in learning minimal pairs arose.

Subsequent studies have shown that infants' performance in learning minimal pairs can be modulated by various factors. Successful learning has been found when the task contained trials in which familiar words were presented (Fennell & Waxman, 2010; Yoshida, Fennell, Swingley, & Werker, 2009) or when the infants had been familiarized with the novel objects prior to the word exposure (Fennell, 2012). Better performance has also been found when the labels were presented within naming phrases that are typical in speech to children (e.g. *There is the x* or *Look at the x*; Fennell & Waxman, 2010).

Other studies have tested whether specific phonological or phonetic properties of the presented labels may enhance infants' responsiveness to the critical sound contrast. Minimal pair word learning was successful when the contrast was implemented by liquids (Archer & Curtin, 2018) or by vowels contrasting in height but not in backness (Curtin, Fennell, & Escudero, 2009). Altvater-Mackensen and Fikkert (2010) report asymmetries in the direction of change for a stop-initial and fricative-initial minimal pair: a switch effect was found when the test item involved replacing a fricative by a stop but not when a stop was replaced by a fricative.

Another factor that modulates performance is the degree of acoustic variability of repeated instances of the words. In minimal pair word learning settings using the habituation-switch paradigm, 14-month-old infants showed better performance when the acoustic stimuli were drawn from recordings of multiple speakers compared to a single speaker (Quam, Knight, & Gerken, 2017; Rost & McMurray, 2009, 2010) and also when a high number of different exemplars from a single speaker who was instructed to produce them with varying pitch and duration was used as stimuli (Galle, Apfelbaum, & McMurray, 2015). Bortfeld and Morgan (2010) found that even 7.5-month-old infants' ability to detect words in fluent speech was enhanced when these words alternated between emphatic and non-emphatic productions. Emphatic stress (and its concomitant acoustic repercussions in duration, intensity and pitch) is not contrastive and yet variability in the presence and degree of emphatic stress during familiarization was shown to improve segmentation.

Research Highlights

- Fourteen-month-old German infants learn minimal pairs in the habituation-switch paradigm when labels were spoken by multiple speakers but from one speaker.
- Visual variability of the presented objects did not improve minimal pair word learning (with a single speaker).
- Identification of the relevant phonetic dimensions for phonemic contrasts is unlikely to proceed on the basis of finding invariant individual cues.
- Relations among (individually varying) cues offer a plausible basis for the identification of phonemic contrasts.

Such findings highlight the key problem we take up in this paper. Variability is considered in many contexts to be synonymous with noise, a curse or unwanted property of the input which obscures crucial signal dimensions. However, in the results just reviewed, variability is actually beneficial. These results present, in a nutshell, a case of what students of cognitive science consider to be a hallmark of natural versus artificial cognition, namely, that humans do well in complicated and noisy cases but performance decays or breaks down in artificially simplified cases. In the present instance of this problem, as demonstrated by Rost and McMurray (2009), 14-month-olds infants succeed in learning the minimal pair at the more complex acoustic training scenario (multiple speakers, each with their speaker-specific values for the cues conveying the contrast in the minimal pair) and fail at the simpler scenario (single speaker). Understanding what basic principles subserve this ability remains a major open problem both in the domain of language development and other domains of cognition where category formation is crucial and where benefits of variability on category formation and generalization have been shown (Posner & Keele, 1968; Quinn & Bhatt, 2010).

The present paper aims to both verify and further better delimit effects of input variability in minimal pair learning. We verify previous findings (and extend these to German infants) by demonstrating success in the habituation-switch paradigm under a multiple speakers condition but not under a single speaker condition. We then aim to better delimit the role of variability by asking whether effects of variability on minimal pair word learning are restricted to speech-specific variability or whether variability in a different modality (i.e. visual) would also affect performance. In what follows, we focus on experimental results and theoretical proposals on the role of variability in word learning within the habituation-switch paradigm, turning to the motivation of our studies next.

1.1 | Variability in the habituation-switch paradigm

Rost and McMurray (2009) investigated English 14-month-olds' learning of minimal pairs that instantiated a voiced-voiceless contrast (/puk/ vs. /buk/) in a single speaker and a multiple speakers condition using exactly the same procedure and the same visual stimuli across the two conditions. Acoustic stimuli of natural recordings were used: a single exemplar of each label recorded from a single female native speaker of English in the single speaker condition and 54 different exemplars of each label recorded from 18 different native speakers in the multiple speakers condition. Evidence for learning the minimal pairs was only found in the multiple speakers condition. Perhaps the most important implication of these results is that failure in the habituation-switch task (in the single speaker setting) cannot be ascribed to infants' performance limitations in terms of ability to store or process acoustic details (Rost & McMurray, 2009, p. 347). Hence, the disparity first noted in Stager and Werker (1997) between the acuity of perceptual discrimination and the failure to notice the mismatching word-object pairing in the habituation-switch paradigm cannot have its basis on limitations in registering signal details. This is because the multiple speakers setting presents a more complex acoustic environment to the infants than the single speaker setting and yet it is in the former scenario where success at the habituation-switch task is demonstrated. Beyond this important conclusion, why variability confers benefits to learning cannot be ascertained from these results (as the authors are careful to point out).

In a subsequent paper, Fennell and Waxman (2010) pointed to a potential explanation of the multiple speakers effect: the social convergence established by different speakers uttering the same word during the presence of the same object may have clarified the referential character of the task, thus leading to better performance in word learning. In forming a word-object association, the learner must somehow infer that the acoustic event accompanying the visual stimulus is the name for that stimulus. What makes acoustic events have the status of words (as opposed to any other sounds generated by a human) is their use in inter-speaker interactions. The consistent replication of the acoustic event-object relation across several speakers served, in Fennell and Waxman's (2010) interpretation, to establish the referential status of the presented acoustic events.

However, in a follow-up study, Galle et al. (2015) demonstrated a switch effect with a single speaker who was instructed to produce the stimuli in a prosodically highly variable fashion, indicating that variability can boost minimal pair learning even in a setting where the Fennell and Waxman (2010) interpretation cannot be applied (as there was one speaker only).[1] Yet it is conceivable that in the Galle et al. (2015) task, as a consequence of the instructions to (the single speaker to) produce the labels in prosodically highly variable infant-directed ways, the resulting utterances resembled variegated contexts where words known to the infants were learned in past communicative exchanges and thus helped clarify the referentiality of the novel labels. It does not appear straightforward at present to know the extent to which any given experimental design clarifies the referential role of the to-be-learned labels. Fennell and Waxman (2010) do not explicitly specify which experimental settings clarify referentiality, but rather demonstrate two markedly different designs they consider to do so and where the outcome is success in the habituation-switch task.

A further study by Rost and McMurray (2010) and a subsequent modelling paper by Apfelbaum and McMurray (2011) aimed to identify experimentally and make explicit via a computational model likely mechanisms via which variability leads to better learning. Rost and McMurray (2010) contrasted a condition where speaker identity was held constant but voice onset time (VOT) was varied by drawing exemplars from a bimodal VOT distribution (with a mode for /b/ at some low VOT value and another mode for /p/ at some high VOT value) with a condition where speaker identity varied but VOT was held constant across the 18 speakers (that is, within each of the /b/ and /p/ categories, all speakers had practically the same VOT value). This latter condition is certainly unlike what occurs naturally, due to the well-known speaker specificity of VOT values (Allen, Miller, & DeSteno, 2003), but this design was motivated by an attempt to tease apart the source of the learning benefit, with the expectation that infants 'should succeed at the switch task when exemplars contain lots of variability, but minimal within-category variability in contrastive cues' (Rost & McMurray, 2010, p. 621). Fourteen-month-olds succeeded in this latter condition but not in the single speaker condition with VOT variation. This result then indicated that it is variability in the irrelevant dimensions (such as speaker voice and F_0 among other parameters) that is crucial to learning. Such variability in irrelevant dimensions, the authors argued, prevents these dimensions from being associated with the objects, thus promoting the selection of the crucial signal dimensions such as VOT which encode the phonemic contrast for the minimal pair.

Apfelbaum and McMurray (2011) offered a modelling study that aimed at replicating (among others) these two experimental conditions. Using parameter values that qualitatively reflect properties of the stimuli used in prior experiments, they trained a connectionist model which includes two visual units, representing the two objects in the habituation-switch task, and three auditory cues, F_0, VOT and an indexical cue which serves as a stand in for whatever parameters characterize speaker voice. The network begins in a state where each auditory cue is connected to both visual units with a weight of zero (no associations formed). Each training trial updates the strengths of the associations between the auditory cues and the visual units. Thus, when /buk/ is uttered with some value of F_0 the association weight between that F_0 value and the visual unit it was presented with is increased (and so on for the values of the other auditory cues). As trials accumulate, the association weights change in a way that reflects properties of the co-occurrence between specific cue values and visual units. In the multispeaker scenario, at each trial, a different value of F_0 and a different value of the indexical cue increment their association weights with the visual unit presented at that trial. As a result, in the multispeaker scenario, no single F_0 value and no single indexical cue value will come to be robustly associated with any of the two visual units, because different speakers with

speaker-specific F_0 and indexical cue values utter both /buk/ and /puk/. In contrast, for the VOT cue, different (modes of) values of VOT end up building robust associations to the two objects because the unit corresponding to /b/ in the VOT cue is always (as assumed in the simulations) activated along with the /buk/ visual unit and the unit corresponding to /p/ in the VOT cue is always activated with the /puk/ visual unit. In the sea of variability of the multispeaker training, then, the relevant auditory cues to the phonemic contrast (here, VOT) float up as the crucial dimensions over which that contrast is specified. At test, when an object is presented with the 'wrong' acoustic stimulus, say, the object A which during habituation is presented along with acoustic-/buk/ is now presented with acoustic-/puk/, the VOT cue of /puk/ strongly activates the other object leading to rejection of that acoustic stimulus-object pairing. This is not so in the single talker training scenario. In this case, the speaker's voice or F_0 does not change across different object presentations and thus these cues will not be less associated to any of the two objects than the VOT cue. As a consequence, at test, when an object is presented with the 'wrong' acoustic stimulus (switch trial), the VOT cue does activate the other object but because all other cues are associated with both objects (in contrast to the multispeaker scenario) they contribute activation to both objects, thus making rejecting the auditory stimulus-object pairing harder than in the multiple speakers scenario. Apfelbaum and McMurray (2011) infer from this 'that *greater variability* on *irrelevant* dimensions is what is important' (Apfelbaum & McMurray, 2011, p. 1,108) and that 'relative variability among cues may play a crucial role' (Apfelbaum & McMurray, 2011, p. 1,109).

Although the experimental results from Rost and McMurray (2010) and attendant modelling in Apfelbaum and McMurray (2011) agree, note that one cannot infer from this that relevant cues must vary minimally also in the general case. The relevant cues did vary in the Rost and McMurray (2009) experimental results obtained with unmodified natural stimuli and in the real world scenario. Thus, the inference that relevant cues must vary minimally or must vary less than irrelevant cues is based on a special case (Experiment 3 in Rost & McMurray, 2010), a training environment where speaker identity varies but the VOT within each category was manipulated to be the same across the 18 speakers. This special case departs from the Rost and McMurray (2009) training environment which had natural stimuli and where VOTs within each category did vary. It also departs from the training environment in the Galle et al. (2015) study which demonstrated a switch effect in a task where a single speaker was instructed to produce the words in prosodically highly variable ways.

It is important to highlight here a distinction between a major thesis and a specific mechanism across the works reviewed above. The specific mechanism is that learners home in on the relevant signal dimensions by tracking the extent of variability in individual cues. This specific mechanism is distinct from the major thesis and accompanying findings that the works of Rost and McMurray (2009, 2010) and Galle et al. (2015) clearly bring out, namely, that variability in what appear to be irrelevant dimensions (either across speakers or within a speaker) seems to benefit learning. We return to the specifics of the likely mechanisms via which variability benefits learning in the General Discussion section of the paper.

1.2 | Current studies

Past work indicates that variable input leads to better learning. However, what properties of the input facilitate or hinder word learning is not yet well understood. Furthermore, the mechanisms via which variability may facilitate learning are also not entirely clear.

As outlined in the previous section, one interpretation of the evidence reviewed so far is that it is variability in the acoustic dimensions of the speech signal per se that fosters better performance at the habituation-switch task, because it aids in selecting or differentially weighing task-relevant signal dimensions (Apfelbaum & McMurray, 2011; Rost & McMurray, 2010; see also Bortfeld & Morgan, 2010 for infants' word segmentation).

Another interpretation is that variability by itself is helpful regardless of the dimensions over which it is expressed. Varying input might be more attractive, helping learners to stay focused on the task. Much experimental work from visual perception indicates that successful performance in perception tasks requires attention (Denison, Adler, Carrasco, & Ma, 2018; Mack & Rock, 1998; Simons & Chabris, 1999). An instructive example of how attention may affect performance in infants' word processing comes from the study of Bortfeld and Morgan (2010). In a typical word segmentation study, the effect size of successful segmentation was largest when the words were presented during familiarization and during test with alternations of emphatic and non-emphatic stress. In two other experiments of this series, infants were familiarized with words that were realized either with emphatic or with non-emphatic stress, exclusively. Interestingly, after familiarization with the emphatically stressed words, infants showed overall longer orientation times during testing. As the authors suggest, emphatic stress during familiarization may have better maintained infants' attention during the test phase. Given that emphatic stress is associated with higher ranges in pitch and duration across the words used in the experiment, it is conceivable that this acoustic variation contributes to higher attention levels. Transferred to the minimal pair word learning scenario, these findings may imply that a higher variability in the stimuli themselves may help to maintain infants' attention during the experimental procedure and thereby contributes to an enhanced performance.

It is not a priori clear which of these two interpretations is correct. In a habituation-switch task, auditory stimuli of spoken novel words are presented along with pictures of objects, the intended referents of these stimuli. Hence, at least two sensory modalities are involved, auditory and visual. In the general case, the memory trace of a word consists at least in what that word sounds like (involving representations and processes in the auditory domain) and what its referent looks like (involving representations and processes in the visual domain), along with other aspects which can be considered outside the purview of the habituation-switch task such as how the object referred to by the word may be used

which would in turn implicate action-oriented, along with tactile and kinaesthetic elements from the motor and parietal areas (Allport, 1985; Emmorey, McCullough, Mehta, & Grabowski, 2014; Skipper, Devlin, & Lametti, 2017). It is thus conceivable that, just as variability in the acoustic form of the stimuli serves to enhance learning, variability in the visual form (the other sensory modality which constitutes an integral part of the word-object link) may also contribute benefits to learning. Prior evidence on how visual variability may affect performance in the habituation-switch task and at the age group of our studies is non-existent. However, evidence for supportive effects of visual variability on children's word learning has been reported in a referent selection task with 2-year-olds. Twomey, Ma, and Westermann (2018) showed that children trained in novel word-object pairings better retained these pairings when the objects were presented in a variable colour background compared to an invariable colour background. A contrasting view is that it is speech signal variability per se that helps highlight the relevant properties of the acoustic signal (Apfelbaum & McMurray, 2011; Rost & McMurray, 2009, 2010). If so, then supportive effects in word learning are expected for speech-specific variability but not for visual variability.

Our study, thus, set out to compare for the first time the effects of speech-specific versus visual variability on minimal pair learning in the habituation-switch paradigm. In our first two experiments, we tested 14-month-olds' ability for minimal pair learning in a single (first experiment) and a multiple speakers condition (second experiment). In a third experiment, the objects whose names were uttered by a single speaker (as in the first experiment) were presented with different types of visual transformations but with no acoustic variability. This allowed us to assess whether variability in a different sensory modality implicated in the habituation-switch paradigm confers comparable benefits in performance.

2 | MINIMAL PAIR WORD LEARNING IN A SINGLE SPEAKER SETUP

Our first experiment can be considered a conceptual replication of Rost and McMurray's (2009) Experiment 1, but extended here to German learning infants. The acoustic stimuli were just two exemplars (each repeated many times), one for each word in the minimal pair /buk/, /puk/, spoken by single speaker.

2.1 | Methods

2.1.1 | Participants

Twenty-two 14-month-old children (between 13.0 and 14.8 months) growing up in monolingual German speaking families participated in Experiment 1. The infants were recruited from the participant pool of the BabyLAB Potsdam. For all children, parents reported that they were born full-term and typically developing. Data from five children had to be excluded because of failure to reach the habituation criterion (1) or parental interference, excessive movements or distractions (4). Thus, data from 17 children (8 girls; mean age: 13.6 months, range: 13.0–14.8 months) were included in the analysis. Informed consent was obtained from the children's caregivers before the experiment was run. The study was approved by the ethics committee of the University of Potsdam.

2.1.2 | Stimuli

The acoustic stimuli used in the experiment were the same phonemic sequences that Rost and McMurray (2009, 2010) used in their studies with English-learning children (/puk/ and /buk/) but spoken with a long vowel by a German native speaker, that is, [bu:k] and [p^{h}u:k]. The two words form a phonotactically legal minimal pair in German but do not correspond to any existing German word. Like in English, the crucial dimension on which the members of the minimal pair differ is the voicing of the initial consonant. The words were recorded from a female native speaker of German in different contexts. For the use in the experiment, a single token of each word produced in a focused context (*Look…X*; *Schau mal…puk*) was chosen. Acoustic measurements (Table 1) revealed comparable values for duration and pitch for the two words. The values of VOT for the initial consonant in [bu:k] and [p^{h}u:k] were 17 and 111 ms

TABLE 1 Acoustic parameters of the labels used in Experiment 1 and 2

Word	Property	Single speaker (Exp. 1)	Multiple speakers (Exp. 2)		
			Mean	Range	*SD*
/bu:k/	Duration (ms)	996	964	623 to 1936	253.4
	Mean F_0 (Hz)	288	246	120 to 383	64.2
	Max. F_0 (Hz)	476	458	147 to 618	120.0
	VOT (ms)	17	3.1	−157 to 35	39.0
/pu:k/	Duration (ms)	942	956	537 to 1764	232.2
	Mean F_0 (Hz)	332	273	127 to 395	64.1
	Max. F_0 (Hz)	514	485	188 to 614	97.7
	VOT (ms)	111	89.6	42 to 140	21.6

Abbreviation: VOT, voice onset time.

respectively. For each word, a sequence containing seven repetitions of that same word with an inter-stimulus interval of 2 s and an overall length of 14 s was created and stored as an audio file.

As visual stimuli, three objects with different shapes and colours (Figure 1) were selected from the NOUN database (image no. 2061, 2002, 2015; Horst & Hout, 2016). Adult ratings (Horst & Hout, 2016) indicate that these objects are perceived as rather novel (green object: 47%, red: 78%, blue: 63%). The green and the red object served as referents in the word-learning task. They are among the 16 most highly dissimilar objects in the NOUN database based on the adult ratings (Horst & Hout, 2016). The blue object served as the stimulus for the novel condition of the test phase.

2.1.3 | Procedure

The procedure corresponded to the classical habituation-switch paradigm as introduced by Stager and Werker (1997) and was implemented using Habit 2 (version 2.1.25, Oakes, Sperka, & Cantrell, 2015). Infants were seated on their caregiver's lap in front of a monitor. Caregivers were instructed to close their eyes, sit still, and avoid any interaction with their child during the experiment. The experimenter sat in a curtained-off part of the test room. During the habituation phase of the experiment, in each trial one of the two object referents was presented on the screen together with the speech file containing the word for this object (see Figure 2 for an illustration of the procedure). The maximal duration of a trial was 14 s. During the trial, infants' duration of looking to the visual presentation was coded online by the experimenter by pressing a button on the keyboard. When the infant looked away for more than two consecutive seconds the presentation was stopped and the next trial started. The habituation criterion was reached when the average looking durations of four consecutive trials dropped by 50 percent compared to the mean of the first four trials. If an infant did not reach this criterion within 30 trials, the habituation was stopped. Object-word parings were blocked into two trials such that the same pairing was repeated at most once in two consecutive trials.

The test phase started immediately after the infant had reached the habituation criterion. During the test phase of the experiment, only one of the two objects presented during habituation was used in two test trials (counterbalanced across infants). In one of these test trials, the object was presented with the same label as in the habituation phase (same trial). In the other test trial, the object was presented with the label that had been presented with the other object during the habituation (switch trial). In a third trial of the test phase, one of the two labels was presented with the novel object that had not been presented before (novel trial). The speech files presented during the test phase were the same as in the abituation phase and looking times were measured in the same way as in that phase. The visual stimuli were presented as static pictures during habituation and testing. Depending on the individual looking durations and the number of trials that the infant needed to reach the habituation criterion, the experiment had a duration between five and eight minutes. Object-label pairing and the order of stimulus presentation during habituation and test was counterbalanced across the participants.

2.2 | Results and discussion

On average, infants reached the habituation criterion after the presentation of 15.7 trials (*SD* = 7.3) with a duration of exposure to the stimuli of 134 s (*SD* = 63). The looking times for the same trials were compared to the looking times of the switch and the novel trials (see Figure 3) by fitting a linear mixed model with the package lme4 (version 1.1.21, Bates, Maechler, Bolker, & Walker, 2015) in R (version 3.6.1, R Core Team, 2019). The contrast between the conditions was coded as treatment contrast, participants were specified as random component and *p*-values were obtained by the package lmerTest (version 3.1.0, Kuznetsova, Brockhoff, & Christensen, 2017). The statistical analysis revealed no significant differences: looking times in switch trials were not significantly different from looking times in same trials (estimate: –710 ms, $t = 0.674$, $p = .505$), and looking times in novel trials were not longer than in same trials (estimate: 471 ms, $t = 0.448$, $p = .658$).

These results do not provide any evidence that 14-month-old German-learning infants learn minimal pairs in these experimental conditions, that is, in the habituation-switch task when the labels are presented without any acoustic variation. Recall that in our experiment, only one exemplar from one speaker was used for each label in the habituation as well as in the test phase. Thus, our findings are in line with findings from previous studies using only one or a small set of different exemplars of the labels from a single speaker (Rost & McMurray, 2009; Stager & Werker, 1997) and extend them to children learning a different language than English. In the next experiment, we tested whether 14-month-old German infants benefit

FIGURE 1 The three objects used in the study: The green (left) and the red (centre) object were labelled by the novel words in the habituation and tested with correct or incorrect labels in the test phase. The blue (right) object always served as the novel referent in the test phase

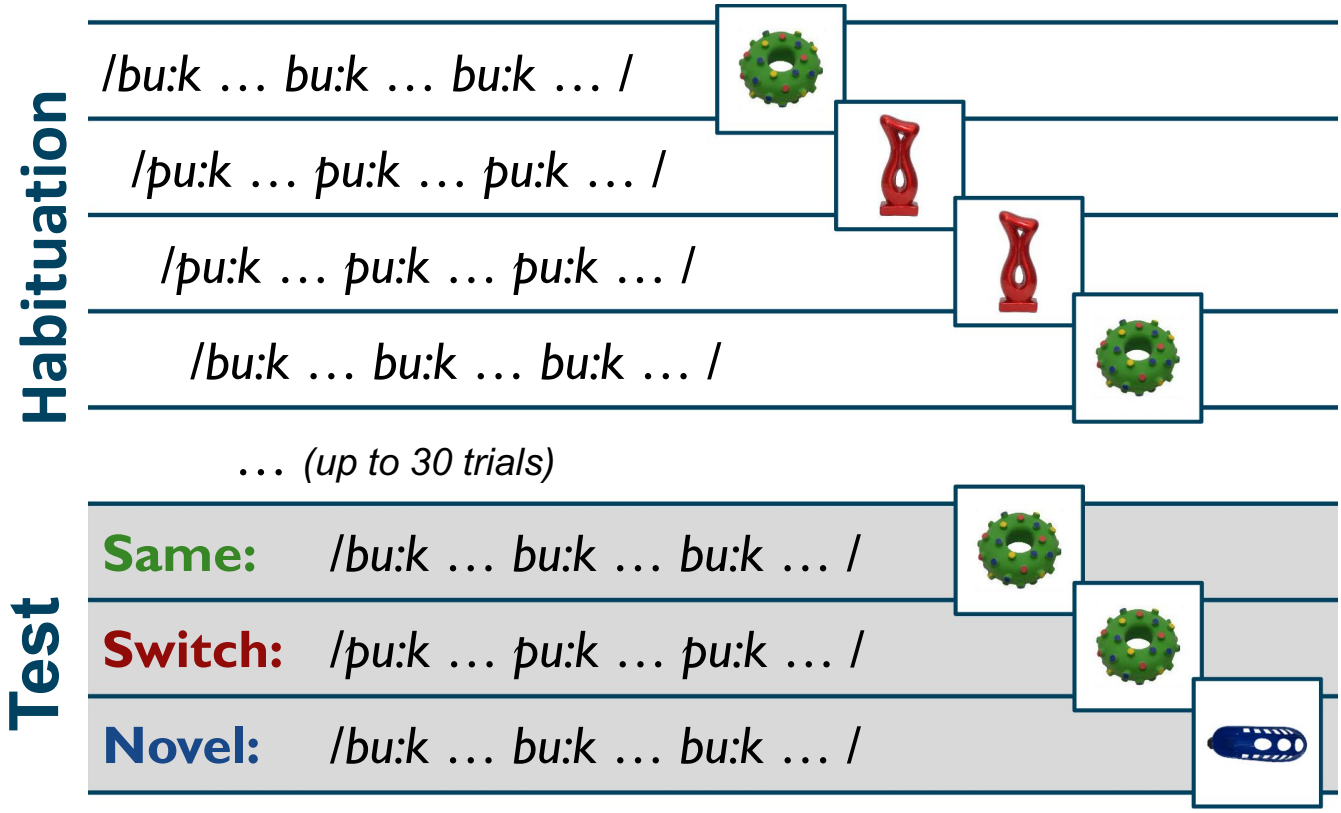

FIGURE 2 Experimental procedure

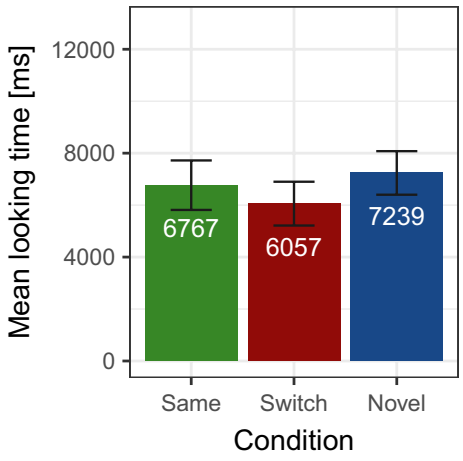

FIGURE 3 Mean looking times in the three test trials in Experiment 1

from the presentation of multiple exemplars of the labels recorded by multiple speakers.

Previous studies using this paradigm typically found significantly longer looking times in novel trials than in same trials (e.g. Rost & McMurray, 2009; Werker, Cohen, Lloyd, Casasola, & Stager, 1998). The reason for including the novel trials is to provide evidence that infants are still attentive during the test phase by showing that they are responsive to the occurrence of an event not encountered before (Werker et al., 1998). Unexpectedly, looking times for novel trials were not significantly longer than for same trials in our experiment. One interpretation for the failure of finding such an effect in our experiment may be that that the lack of variability in the acoustic stimuli substantially attenuated infants' attention. However, our experimental procedure matched as closely as possible the one in the study by Rost and McMurray (2009) who also used only one exemplar of each word and static pictures of the objects. Potentially there were differences in acoustic properties of the stimuli or in the saliency of the visual objects across the two studies that caused longer looking times for the novel trial in the Rost and McMurray (2009) study but not in ours.

3 | MINIMAL PAIR WORD LEARNING IN A MULITPLE SPEAKER SET UP

3.1 | Methods

3.1.1 | Participants

Twenty-nine 14-month-old children (between 13.0 and 15 months) growing up in monolingual German speaking families participated in Experiment 2. Again, the infants were recruited from the participant pool of the BabyLAB Potsdam, none of the infants tested in Experiment 2 had also participated in Experiment 1. For all children, caregivers reported that they were born full-term and typically developing. Data from twelve children had to be excluded because of

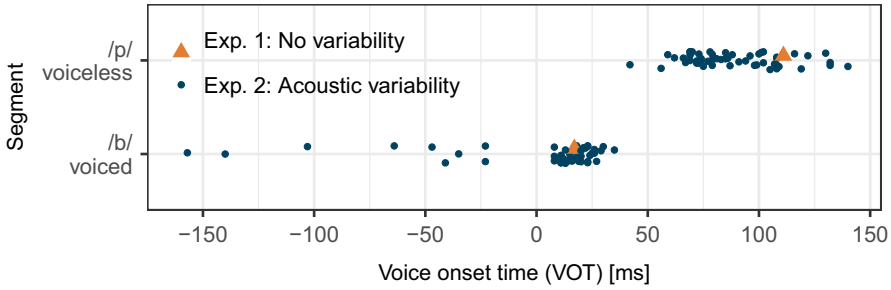

FIGURE 4 Voice onset time (VOT) values for each exemplar in Experiment 2 (multiple speakers, blue dots) and VOT values for the two exemplars (of voiced /b/ and voiceless /p/) of Experiment 1 (single speaker, orange triangles). As expected for a Germanic voicing system, VOT values for /b/ mostly fall within a range of short VOT values (10–25 ms) along with a few negative values indicating occasional voicing during the closure of stop (Beckman, Jessen, & Ringen, 2013; Jessen, 1998) and VOT values for /p/ are much higher with a correspondingly substantial degree of inter-speaker variability (Kuberski et al., 2016; Tobin et al., 2018)

failure to reach the habituation criterion (5), parental interference, excessive movements or distractions (2), no looks to the screen in one of the test trials (3), or experimenter error (2). Thus, results from 17 children (8 girls; mean age: 13.9 months, range: 13.1–15.0 months) were included in the analysis. Informed consent was obtained from the children's caregivers before the experiment was run. The study was approved by the ethics committee of the University of Potsdam.

3.1.2 | Stimuli

The same words as in Experiment 1 were used but recorded from 18 different speakers (12 female and 6 male). Each speaker produced three exemplars of each word in three different contexts (focused as in Exp. 1; in isolation; in a question: *Is this a…X?*). Acoustic measurements (Table 1; Figure 4) indicate that the labels from the multiple speakers presented a distribution for which the stimulus from the single speaker used for Experiment 1 was a representative exemplar. Speech files were created in the same way as described for Experiment 1.

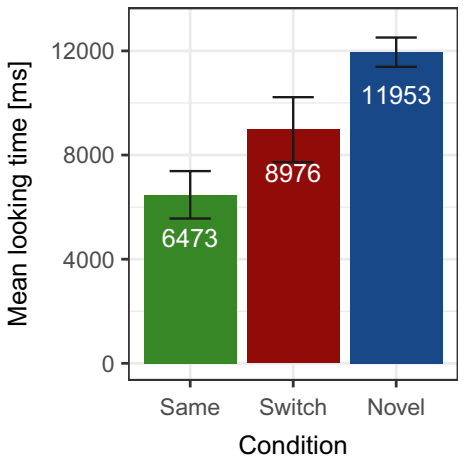

FIGURE 5 Mean looking times in the three test trials in Experiment 2

3.1.3 | Procedure

The procedure was identical to that of Experiment 1.

3.2 | Results and discussion

On average, infants reached the habituation criterion within 15.1 trials (*SD* = 5.8), with an average habituation duration of 146 s (*SD* = 63) which is statistically not different from the habituation duration in Experiment 1, $t(32) = -0.534$, $p = .597$. Again, the looking times in the three conditions (Figure 5) were compared by fitting a linear mixed model specified as in Exp. 1. This analysis revealed that the looking times in switch trials were significantly longer than in same trials (estimate: 2,503 ms, $t = 2.176$, $p = .037$). Moreover, the looking times in novel trials were also significantly longer than in same trials (estimate: 5,480 ms, $t = 4.765$, $p < .001$).

Our results corroborate previous findings from English-learning infants: 14-month-olds seem to benefit from hearing the phonetic realizations of the contrast that separates the two labels encountered during the experiment in a high number of exemplars produced by multiple speakers.[2] However, it is still an open question why infants' performance is improved by receiving input from multiple speakers. Fennell and Waxman (2010) have hypothesized that hearing multiple speakers producing the same word during the presence of the same object may enhance the referential status of the acoustic stimulus and therefore foster the forming of object-label associations. However, Galle and colleagues (2015) questioned this explanation based on their finding that presenting multiple exemplars of the labels with high acoustic variability produced by the same speaker also boosts 14-month-olds' capacity to learn minimal pairs in the habituation-switch paradigm. One interpretation of these results is that variability by itself in irrelevant dimensions ('noise') is helping the learner. Varying input might be more attractive to the learner, helping infants' attention to stay focused on the task. Under this interpretation, one would expect differences in habituation duration between tasks containing variable input versus invariable input. These differences could go either way, resulting in longer habituation durations with

variable stimuli because attention is better maintained or in shorter habituation durations because due to more focused attention learning is undistracted and can proceed faster. However, the habituation durations of Experiment 1 and 2 were quite comparable, thus providing no indication that input variability in and of itself creates differences in attention. Rost and McMurray (2009, 2010) also found no differences in habituation durations when comparing their experiments that involved speaker variability (and in which infants showed minimal pair learning) and experiments that did not involve speaker variability (and in which infants did not show minimal pair learning). We thus suspect, as others have done before us (Galle et al., 2015; Rost & McMurray, 2009, 2010), that properties of the variability itself are what fosters better performance at the habituation-switch task in the multispeaker scenario. This interpretation raises one basic but yet unaddressed question concerning different kinds of variability. If acoustic variability per se is what helps highlight the relevant properties of the acoustic signal, supportive effects should only occur within the same modality. Under this hypothesis the beneficial effect of variability in word learning is only expected for acoustic variability (as found in Experiment 2) but not for visual variability. This is what we tested in Experiment 3.

4 | MINIMAL PAIR WORD LEARNING WITH VISUAL VARIABILITY

4.1 | Method

4.1.1 | Participants

Twenty-six monolingual 14-month-old infants (between 13.1 and 14.7 months) growing up in monolingual German speaking families were tested in this experiment. None of them had participated in Experiment 1 or 2. The infants were recruited from the participant pool of the BabyLAB Potsdam. For all children, parents reported that they were born full-term and typically developing. Data from nine children had to be excluded because of failure to reach the habituation criterion (4), no looks to the screen in one of the test trials (4), or experimenter error (1). Thus, data from 17 children (9 girls; mean age: 13.7 months, range: 13.1–14.7 months) remained in the analysis. Informed consent was obtained from the children's caregivers before the experiment was run. The study was approved by the ethics committee of the University of Potsdam.

4.1.2 | Stimuli

The speech material for this study corresponded exactly to the speech stimuli used in Experiment 1. The same objects as in Experiments 1 and 2 were used but visual variability was induced by applying four different transformations to the images (Figure 6). First, we used two different sizes, the original size (100%) and a smaller version (85% of the original size). Second, three levels of orientations were created: not rotated (0°), rotated by 20° to the left and rotated by 20° to the right. Third, we created three levels of dimensional scaling: unscaled, scaling 1 (the original object width was increased to 110% while reducing height to 85%) and scaling 2 (width increase to 120%, height to 70%). Fourth, three different background colours were used (matched in luminance): light red, light yellow and grey. All possible combinations of these four transformations (2 × 3 × 3 × 3) yielded 54 different versions for each object. This number of different visual exemplars mirrored the number of 54 different acoustic exemplars that were presented in the multispeaker Experiment 2. The visual changes to the presented object occurred within a trial every two seconds synchronous to the onset

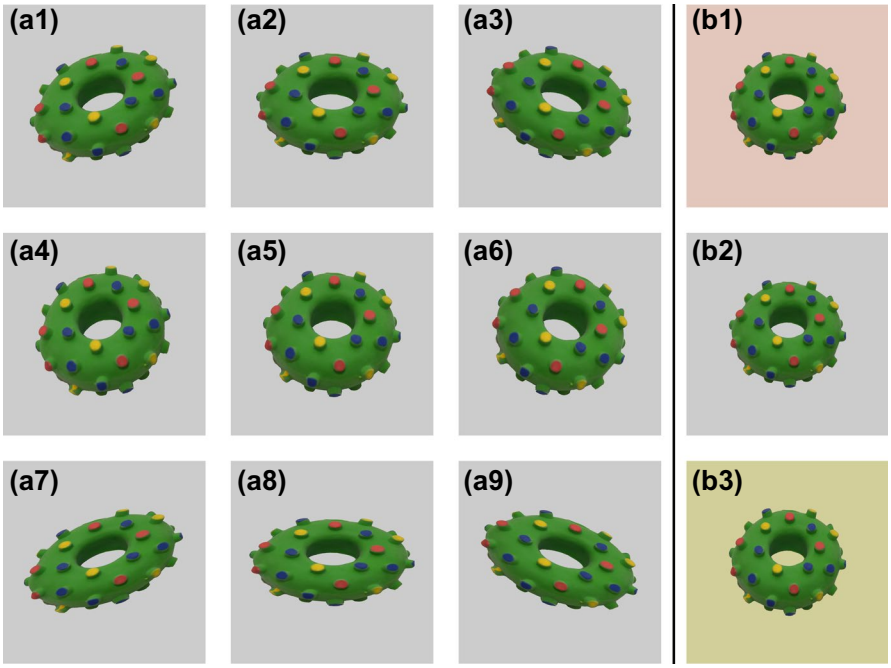

FIGURE 6 Examples of the visual stimuli in Experiment 3. A5 shows the green object unchanged as used in Experiment 1 and 2. On the horizontal axis, A4 shows the rotation to the left, A6 to the right. Vertically, A2 shows scaling 1 and A8 scaling 2. The panels in the corners (A1, A3, A7, A9) show the combinations of rotation and distortion. Column B on the right displays the green object in smaller size on the different background colours used

of the auditory stimulus for the word such that each word repetition was accompanied by a different variant of the object.

4.1.3 | Procedure

The procedure was identical to that of Experiment 1 and 2.

4.2 | Results and discussion

On average, infants were presented with 15.7 habituation trials (*SD* = 6.0) with an overall duration of 144 s (*SD* = 72) until they reached the habituation criterion. Habituation duration was statistically not different from Experiment 2 ($t(32) = 0.093$, $p = .926$. Statistical comparisons (with a linear mixed model as in Experiment 1 and 2) of the looking times in the three test trials (Figure 7) revealed the following picture. Looking times between the same and switch trial were statistically not different (estimate: 339 ms, $t = -0.266$, $p = .792$). However, the looking times in novel trials were longer than in same trials (estimate: 3,505 ms, $t = 2.753$, $p = .001$).

The results of Experiment 3 do not provide any evidence that the visual variability implemented via presenting the unfamiliar objects in different sizes, orientations, dimensional scales or on different background colours increased infants' ability to selectively associate the labels to the objects. Thus, in comparison to Experiment 1 (single speaker, no acoustic variability) and to Experiment 2 (multispeaker setting with acoustic variability), the added visual variability in Experiment 3 did not benefit learning appreciably. This indicates that supportive effects of variability in the habituation-switch paradigm are not equally present across different kinds of variability. We place these findings in the context of prior results on the role of variability in speech and other domains in the following section.

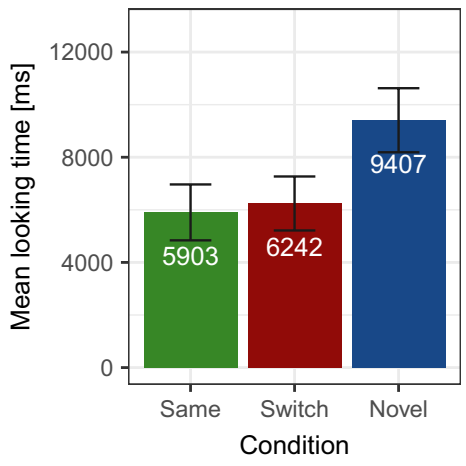

FIGURE 7 Mean looking times in the three test trials in Experiment 3

5 | GENERAL DISCUSSION

Our study compared effects of phonetic variability (in auditory labels) versus effects of visual variability (in how the unfamiliar objects for the auditory labels were presented) on minimal pair word learning using the habituation-switch-paradigm. More specifically, in our first two experiments, we tested 14-month olds' ability for minimal pair word learning in a single speaker (first experiment) versus a multiple speakers condition (second experiment). We have found, as others have found before, that additional acoustic variability created by having the labels produced by more than one speaker resulted in 14-month olds succeeding to learn the minimal pair. In a third experiment, the objects whose labels were uttered by a single speaker—as in the first experiment—were presented with different types of visual transformations (visual variability). We asked whether visual variability in that setting would confer the same benefit on minimal pair learning as the phonetic variability in the multiple speakers condition. Results indicated that in contrast to variability contributed by multiple speakers, visual variability does not enhance minimal pair learning in this experimental paradigm. In short, not all kinds of variability are beneficial to learning. In the following, we turn to place these results in the context of other results in the habituation-switch paradigm as well other paradigms which explore the effect of various sources of variability on learning.

Where does the advantage in learning when multiple speakers utter the to-be-learnt minimal pairs derive from? As established in the Introduction, there are at least two views on why minimal pair presentation by multiple speakers confers an advantage to learning over presentation by a single speaker. In the first of these views, variability induced in the signal by including multiple speakers may be beneficial not because of the variability in the signal per se, but because participants experience multiple communicative interactions with different speakers which in turn may enhance the referential role of the presented acoustic stimuli (Fennell & Waxman, 2010). Now, crucially, we did not present the faces of the (multiple) speakers along with their utterances but only auditory signals of spoken utterances from multiple speakers were presented. Thus, any information, including the information leading to the projection of a multiplicity of communicative partners by our participants, must have been extracted from the speech signal itself. Note that we are not arguing that referentiality of the stimuli is irrelevant. We are merely pointing out that, in our study, the mechanisms via which referentiality was inferred must have relied on the sole source of information offered to the participants, namely, the speech signal itself. In sum, it follows that the differences we observe in the results (between the single speaker and multiple speakers settings) must derive from properties inherent to the speech input. This assumes that infants can recognize that the stimuli were produced by several speakers. Although even newborns can discriminate their mother's voice from another female voice (DeCasper & Fifer, 1980) and 7-month olds notice a switch from one voice to another when producing

sentences (Fecher & Johnson, 2018; Johnson, Westrek, Nazzi, & Cutler, 2011), it cannot be taken for granted that 14-month olds have discriminated the voices of our different speakers given that the speakers only produced repetitions of two highly similar words. Talker recognition based on voice properties is still not developed to an adult-like level at preschool age (Creel & Jimenez, 2012) and even adults are not always 100 percent correct in deciding whether they listen to one or multiple speakers (Galle et al., 2015; Mann, Diamond, & Carey, 1979).

Let us now consider another view on how phonetic variability can benefit minimal pair word learning. In considering this view, it is useful to take some clues from Fennell and Waxman's (2010) hypothesis on how presentation of the to-be-learnt words by multiple speakers clarifies referentiality: 'After all, when a range of different speakers consistently applies the very same word to a novel object, this social convergence signals that that word is the name of that object' (p. 1,381). We use the original wording as it aptly demonstrates the point that we wish to make next. It is not at all obvious that infants can recognize different acoustic exemplars of the same phonemic sequence as instances of the same word. For example, it has been demonstrated that word recognition across different speakers, emotional affects, pitch levels or stress conditions is initially limited in young infants (Houston & Jusczyk, 2000; Singh, Morgan, & White, 2004; Singh, Nestor, & Bortfeld, 2008; Singh, White, & Morgan, 2008) and continues to pose a challenge well into the second year of life (Fecher, Paquette-Smith, & Johnson, 2019; Mulak & Best, 2013).

What would be required of the infant to infer that a range of speakers applies the very same word? The general answer for both speech and other areas of perception is: identifying variations in one system that correspond to and therefore 're-present' relevant differences in another system. For speech perception, the two systems (within which variations must stand in correspondence) are auditory signals, on the one hand, and mental representations in the mind-brain of the listener on the other hand. Consider the example of a /buk/-/puk/ pair. The child is exposed to some number of auditory stimuli while visually presented with object A (say, the object whose auditory label is /buk/) and some number of auditory stimuli while visually presented with object B (say, the object whose auditory label is /puk/). In word learning, the child must map differences between the various auditory instances of /buk/ and /puk/ to differences in the mental representation of the labels /buk/ and /puk/. The former auditory differences are multidimensional, that is, expressed not in terms of a single variable but in terms of sets of spectrotemporal primitives (amplitude, burst spectrum, VOT, among others) which are neurophysiologically encoded and which we assume, as is standard in linguistic parlance, have their functional equivalents in (presumed) abstract symbolic representations, in this case, [+Voice] for /b/ and [−Voice] for /p/. Overall, then, differences in one system (auditory signals) are mapped to differences in the other system (here, [+Voice] vs. [−Voice]). This establishing of a correspondence (across the two information encoding systems, auditory and linguistic representation) is the problem the infant is faced with in the minimal pair task.

We can gain a first appreciation of the nature of this correspondence problem by trying to simplify it. Let us assume for now that this correspondence problem can be solved by making use of what is usually taken to be the most relevant cue for the voiced-voiceless contrast in /buk/-/puk/ for both English and German, namely, the VOT of the initial plosive (we will drop this limitation later). In effect, VOT differences between /b/-/p/ in the auditory signal must be mapped to differences in the encoding of the /b/-/p/ contrast in the mental representations of the infant. However, VOTs for the initial plosive vary substantially across speakers (for English, see Allen et al., 2003; for German, see Tobin, Hullebus, & Gafos, 2018). VOT distributions are highly speaker dependent, both in terms of their means and their variances. For example, in a sample of forty-two speakers of German, mean VOT for the voiceless plosives /ka/ and /ta/ were found to vary from 44 to 104 ms (Kuberski, Tobin, & Gafos, 2016; Tobin et al., 2018). Moreover, VOT within a single speaker varies with syllable duration (the longer the syllable, the longer the VOT). Both in normal speech as well as in experiments with infants, the to-be-learnt words are typically uttered in various prosodic modulations which surely affect the duration of the syllables wherein these plosives are placed. It then follows that if, as we have shown, infants succeed in the multiple speakers scenario, then they must necessarily also track information about how individual speakers express the difference between /b/-/p/. The ability to infer that a set of different speakers 'applies the very same word' in referring to a visually presented object highlights the issue at hand: despite meeting a phonetically more[3] complex environment in the multiple speakers than in the single speaker setting, infants succeed in the former but not in the latter setting; this is the major finding of Rost and McMurray (2009), the one extended to German learning infants in one of our experiments here, and the one supporting the major thesis that variability in what appear to be irrelevant dimensions in the acoustic signal is beneficial to learning (Galle et al., 2015; Rost & McMurray, 2009, 2010).

Before we address possible mechanisms that may be at play in establishing this correspondence let us highlight one key phonetic property distinguishing the stimuli in the multiple speakers versus the single speaker condition. Both in our study and in that of Rost and McMurray (2009), the training sets consist of multiple (varying) exemplars for each word (/buk/, /puk/) in the multiple speakers condition but only one exemplar for each word (repeated several times) in the single speaker condition. The presence of variation in acoustic dimensions (see Figure 4) and, as we emphasize in what follows, the co-variation among these dimensions in the multiple exemplars provides a basis for setting up expectations for category membership in the former case but not the latter case. In the single speaker condition, no such (co-)variation is to be found. Each object was paired with a single auditory stimulus. These two stimuli had highly distinct phonetic values (e.g. VOTs were 17 ms for [bu:k] and 111 ms for [p^{h}u:k]), but the stimuli did not vary within each category. Note that, in principle, a learner could construct category expectations from just two different, repeated stimuli: many (identical) presentations of an auditory stimulus for /buk/ and many (identical) presentations

of a different auditory stimulus corresponding to /puk/. Our results as well as those from Rost and McMurray (2009) indicate that mere repetition of two clearly separable but identical auditory labels does not confer advantages to learning.

One proposal for a learning mechanism that may be involved in the word learning scenario focuses on the extent of variability in individual cues (Galle et al., 2015; Rost & McMurray, 2010). The idea is that learners home in on the right cues for the crucial phonetic contrast by tracking the extent of variability in different candidate cues. Cues that do not vary or vary less are promoted as the basis for discrimination and cues that vary more or substantially are demoted. Such a criterion, the argument goes, helps in deciding whether any given cue is relevant or irrelevant for word learning, with 'less variable cues being more relevant for word learning than variable cues' (Galle et al., 2015, p. 68; see also Rost & McMurray, 2010). The proposal that a key criterion for whether a cue serves as the basis of a lexically relevant contrast is the extent of individual cue variability, however, seems to meet challenges when one considers the richness of cues in speech perception and the phonetics of the multiple speakers scenario over and beyond the habituation-switch paradigm.

In speech perception, multiple cues interact in complex ways to convey phonological contrasts (Repp, 1982). There are two important aspects to this point. One concerns the multiplicity of cues. Another concerns relations between cues. Cases where a phonemic contrast (like voicing) is expressed either exclusively or mostly by a single cue are rare. However, even in such cases, the idea that it is the extent of variability which allows a listener to home in to the right cues meets issues in the face of the flexibility of phonetic expression in speech production. This can be illustrated with the perception of short versus long consonants. Consider deciding whether a heard word is 'topic' versus 'top pick'. The duration of silence during the closure of the medial consonantal interval (either [p] in 'topic' or [pp] in 'top pick') provides the primary acoustic cue. In a forced-choice task, the 50% decision boundary between the single [p] and the double [pp] is not invariant but depends on the rate of the utterance these words are part of. The faster the rate, the lower the boundary value (that is, a duration of silence which at normal or slow rates is judged to represent [p] will be judged with higher probability to represent [pp] at faster rates). Listeners are sensitive to this rate-dependent change in the signal. A given duration of silence is judged differently depending on the rate of the utterance (Summerfield, 1981). Similar results hold even when closure duration serves as a basis for lexical contrast as for example in /s/ versus /ss/ in Japanese, a language which unlike English has distinctive consonant length (see Miller, 1981 for a review). In sum, the considerable flexibility with which contrasts are expressed in production is met by a corresponding graceful adaptation in perception to this flexibility in production.

Consider now the second aspect of the phonetic basis for phonological contrasts, that is, the one concerning relations between cues. Even though VOT is one (and in the word initial context after a pause, perhaps the most) important cue for signalling the distinction between voiced and voiceless plosives in both English and German, other cues expressing spectral information (e.g. the frequency at the onset of the first formant, F1) have been shown to play a role (Summerfield & Haggard, 1977) and moreover to become more crucial than the 'main' cue in certain contexts, for example when the stimuli are presented in noise (Jiang, Zhang, & McGilligan, 2006). In perception, the temporal dimension of VOT and the spectral F1 onset frequency component may be traded for one another: the lower the frequency of F1 at the onset of voicing, the longer the VOT required to produce a voiceless percept. When VOT is set to values that are ambiguous (between the voiced and voiceless category), then category identification can be demonstrated to rely on the other spectral cue (Summerfield & Haggard, 1977). The point we wish to bring out here is that for successful perception neither VOT nor F1 are required to be fixed or less variable. Rather, cues enter in relations with one another such that each cue can vary individually but not independently from the other.

It thus seems more likely that the mechanisms of identifying the basis of a phonological contrast track relations among cues, rather than the degree of variability at the individual cue level. Tracking (the functional equivalents of) cue relations embraces the speaker dependent nature of phonetic expression (because the individual cues are allowed to vary) while elevating consistency at the level of higher order relational properties among cues.[4] Our proposal concerning the beneficial effect of variability then is that what appears to be 'noise', when viewed at the level of individual acoustic parameters, is in fact crucial to the detection of these relational properties (regardless of whether this noise derives from indexical properties of the speaker or from properties that are considered linguistic but are still highly speaker-dependent). Variability is essential to the identification of relational properties between cues because to find a relation between two or more parameters, these parameters must be allowed to vary individually (and the greater the range of their individual variabilities, the more robust the evidence for the presence of a relation among them). Rost and McMurray (2010) and Apfelbaum and McMurray (2011) have emphasized how variability helps prune out irrelevant parameters. We agree and emphasize that also variability in relevant parameters is beneficial because it helps highlight relational properties among cues. Even though so far only very few studies have been devoted to cue relations in infants, the evidence available points to the relevance of such relations in discrimination tasks. Thus, early research suggests that 4-month-olds' discrimination of syllables depends on the coherence of the relational properties among spectral and temporal signal dimensions which work together in cueing the contrast between these syllables (Eimas, 1985; Miller & Eimas, 1983). There are also hints for developmental changes in the precise form of relations among cues (Morrongiello, Robson, Best, & Clifton, 1984). Further research should explore the effects of cue relations for word learning, especially for minimal pair word learning in infants. Nevertheless, the theoretical argument we make here is supported by the experimental evidence available and it is an argument which traces the apparent paradoxical benefit of variability in learning to (at least in part) relations among cues. Rather than requiring relative

invariance at the individual cue level (an assumption too stringent to meet the flexibility and multidimensionality of phonetic expression), the identification of relations among cues in fact requires the presence of variability in individual signal dimensions.

Let us now turn to visual variability. How does the lack of a beneficial role of visual variability compare to past results on visual variability and word learning? Our study is the first to ask whether visual variability implemented via object transformations improves minimal pair learning in the classic paradigm used for assessing minimal pair word learning, that is, the habituation-switch paradigm. Specifically, the results of Experiment 3 indicate that visual variability implemented via the scaling of several object dimensions did not result in infants' success in the habituation-switch task.

As previewed earlier, Twomey et al. (2018) have presented results indicating that visual variability may be beneficial to establishing robust word-object associations. In their cross-situational word learning task with 22-month-old children, a novel object was repeatedly presented together with its auditory label in the context of various other objects. In one of the conditions (the constant colour condition) the objects were always presented on a white background while in the other condition (the variable colour condition) the background colour changed across the learning trials. In the retention phase of the experiment, when probing the learnt object-label association after a 5 min break, objects were always presented on a grey background (note that this background colour was different from that used in the training phase of the constant colour condition). In their results, Twomey et al. (2018) find that only the children who were trained in the variable colour condition had formed sustained associations between the novel word and its referent object.

The results from our work and those from Twomey et al. (2018) are not necessarily contradictory. One potential resolution of the apparent discrepancy between ours and their results is that the tasks in the two studies are markedly different. Twomey et al. (2018) varied the background but not the properties of the objects, included familiar objects/words into their materials, the labels were embedded in phrases, multiple objects were presented during the test phase, and perhaps most importantly the task did not address the learning of minimal pairs but rather learning of phonologically dissimilar words. Another potential resolution, which is not mutually exclusive with the first, may be related to specifics of the design of the crucial test phase in that task. In the Twomey et al. (2018) procedure, after three warm-up trials and 15 referent selection trials, both the constant-colour-trained and variable-colour-trained participants received a single warm-up trial with a change in colour (followed by the six retention trials). The variable-colour-trained participants were by that time well-trained with variability of exactly the same nature (change in colours). In contrast, the constant-colour participants saw this change for the first time. Conceivably, then, it is not the beneficial role of training with visual variability before the break, but rather the distracting effect of presenting the constant-cohort children with something they have not seen before that accounts for the differences in outcomes between the constant-colour-trained and variable-colour-trained participants.

6 | CONCLUSION

In sum, our results both confirm and better delimit the role of variability in minimal pair word learning in the habituation-switch paradigm. Our results confirm previous findings and extend these to German-learning infants by demonstrating that presentation of the to-be-learnt words by multiple speakers confers advantages to learning over presentation by a single speaker. Our results also better delimit the source of the variability benefit by demonstrating that such benefit is specific to the speech sensory modality in our studies. Speech-specific variability seems to be inductively privileged over non-speech (visual) variability for minimal pair word learning in the habituation-switch paradigm. Finally, we review contrasting proposals on mechanisms via which variability confers benefits to learning and outline what may be likely reasons that underlie this benefit. We highlight among these the multiplicity of acoustic cues signalling phonemic contrasts and the presence of relations among these cues. It is in these relations where we trace part of the source for the apparent paradoxical benefit of variability in learning. Whereas previous work has emphasized the role of variability in weeding out the irrelevant cues, we emphasize that variability in relevant cues is also crucial. It is due to the presence of such variability that the relational properties among the relevant signal parameters can be discerned. Rather than requiring (relative) invariance at the individual cue level (an assumption too stringent to meet the flexibility of phonetic expression), the identification of relations among cues requires the presence of variability in individual signal dimensions.

ACKNOWLEDGEMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Projektnummer 317633480 – SFB 1287, Project C03.

CONFLICT OF INTEREST

None of the authors has to declare a conflict of interest.

DATA AVAILABILITY STATEMENT

Data will be made available on a public data repository when accepted for publication.

ORCID

Barbara Höhle https://orcid.org/0000-0002-9240-6117
Tom Fritzsche https://orcid.org/0000-0002-7917-514X

ENDNOTES

[1] Note that this does not imply that the Fennell and Waxman (2010) hypothesis is false—it only implies that multiple speakers is not a necessary condition for learning. Furthermore, there is a caveat to this result. Strictly speaking the inference that 'the present findings refute the notion that multiple talkers are necessary for successful performance in this task.' (Galle et al., 2015, p. 75) is valid only if one can safely exclude that participants interpreted their stimuli as coming from multiple speakers. In profiling their stimuli with adult listeners '79% of test trials were classified as single talker.' That is, a fair

proportion of the test trials (1/5) were judged (by adult listeners) to come from multiple speakers. We have no information on how the infant participants interpreted the stimuli.

[2] As two of our reviewers point out, variability in Experiment 2 is due to both speaker differences and multiple exemplars of the same words spoken by each speaker. In referring to Experiment 2 throughout the paper, our use of the term 'multiple speakers' (or 'multispeaker' condition elsewhere) should not be taken to imply that the benefits seen in Experiment 2 accrue exclusively from only one of these sources (namely, the different speakers). This experiment was meant to replicate the Rost and McMurray (2009) finding and extend it to German and hence, by design, we are not aiming to tease these two sources of variability apart.

[3] The 'more' here refers to a metric of complexity from the experimenter's perspective. The fact that the more complicated case is the one where the infants perform better serves as a humble reminder that the notion of what is more or less complex from the operational point of view may not directly translate to any notion of complexity relevant to the infant.

[4] That such relational properties may serve as the functional units of perception does not imply that perception via these properties requires the registration of the individual first-order cues that are so related (Kingston & Diehl, 1995; Repp, 1982). An example from vision may help clarify this point further. In visual discrimination, velocity is perceived without mediation of the more primitive parameters of distance and time (Algom & Cohen-Raz, 1984; Lappin et al., 1975; Lappin et al., 2011); e.g., 'Definitions and evaluations of derivatives are not necessarily derived from the specific values involved in a function or change' (Lappin et al.; 2011, p. 2,378). To return to speech, the operational and functional sense of the notion of cue should not be conflated (Bailey & Summerfield, 1980; McNeill & Repp, 1973). Manipulation by the experimenter of a single parameter leading to different percepts illustrates the operational sense of the notion of cue. The use of cues in this sense continues to be an appropriate means for constructing stimuli for perceptual experiments since the very early stages of systematic research in the field (Delattre, Liberman & Cooper, 1955 *et seq.*). Demonstration of shifts in sound categorisation as a consequence of a single cue manipulation does not necessarily imply a functional role for that cue as an isolated primitive in perception.

REFERENCES

Algom, D., & Cohen-Raz, L. (1984). Visual velocity input–output functions: The integration of distance and duration onto subjective velocity. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(4), 486–501. https://doi.org/10.1037/0096-1523.10.4.486

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552. https://doi.org/10.1121/1.1528172

Allport, D. A. (1985). Distributed memory, modular subsystems and dysphasia. In S. K. Newman & R. Epstein (Eds.), *Current perspectives in dysphasia* (pp. 207–244). Edinburgh, Scotland: Churchill Livingstone.

Altvater-Mackensen, N., & Fikkert, P. (2010). The acquisition of the stop-fricative contrast in perception and production. *Lingua*, *120*(8), 1898–1909. https://doi.org/10.1016/j.lingua.2010.02.010

Apfelbaum, K., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138. https://doi.org/10.1111/j.1551-6709.2011.01181.x

Archer, S. L., & Curtin, S. (2018). Fourteen-month-olds' sensitivity to acoustic salience in minimal pair word learning. *Journal of Child Language*, *45*(5), 1198–1211. https://doi.org/10.1017/S0305000917000617

Archer, S., Ference, J., & Curtin, S. (2014). Now you hear it: Fourteen-month-olds succeed at learning minimal pairs in stressed syllables. *Journal of Cognition and Development*, *15*(1), 110–122. https://doi.org/10.1080/15248372.2012.728544

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(3), 536–563. https://doi.org/10.1037/0096-1523.6.3.536

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics*, *49*(2), 259–284. https://doi.org/10.1017/S0022226712000424

Bergelson, E., & Swingley, D. (2013). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bortfeld, H., & Morgan, J. L. (2010). Is early word-form processing stress-full? How natural variability supports recognition. *Cognitive Psychology*, *60*(4), 241–266. https://doi.org/10.1016/j.cogpsych.2010.01.002

Creel, S. C., & Jimenez, S. R. (2012). Differences in talker recognition by preschoolers and adults. *Journal of Experimental Child Psychology*, *113*(4), 487–509. https://doi.org/10.1016/j.jecp.2012.07.007

Curtin, S., Fennell, C., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word–object associative learning. *Developmental Science*, *12*, 725–731. https://doi.org/10.1111/j.1467-7687.2009.00814.x

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, *208*(4448), 1174–1176. https://doi.org/10.1126/science.7375928

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, *27*(4), 769–773. https://doi.org/10.1121/1.1908024

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(43), 11090–11095. https://doi.org/10.1073/pnas.1717720115

Eimas, P. D. (1985). The equivalence of cues in the perception of speech by infants. *Infant Behavior and Development*, *8*(2), 125–138. https://doi.org/10.1016/S0163-6383(85)80001-1

Emmorey, K., McCullough, S., Mehta, S., & Grabowski, T. J. (2014). How sensory-motor systems impact the neural organization for language: Direct contrasts between spoken and signed language. *Frontiers in Psychology*, *5*, 484. https://doi.org/10.3389/fpsyg.2014.00484

Fecher, N., & Johnson, E. K. (2018). Effects of language experience and task demands on talker recognition by children and adults. *Journal of the Acoustical Society of America*, *143*(4), 2409–2418. https://doi.org/10.1121/1.5032199

Fecher, N., Paquette-Smith, M., & Johnson, E. K. (2019). Resolving the (apparent) talker recognition paradox in developmental speech perception. *Infancy*, *24*(4), 570–588. https://doi.org/10.1111/infa.12290

Fennell, C. T. (2012). Habituation procedures. In E. Hoff (Ed.), *Research methods in child language: A practical guide* (pp. 3–16). Chichester, UK: Wiley-Blackwell.

Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development*, *81*(5), 1376–1383. https://doi.org/10.1111/j.1467-8624.2010.01479.x

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), 1–173. https://doi.org/10.2307/1166093

Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The role of single talker acoustic variation in early word learning. *Language Learning and Development*, *11*(1), 66–79. https://doi.org/10.1080/15475441.2014.895249

Horst, J. S., & Hout, M. C. (2016). The novel object and unusual name (NOUN) database: A collection of novel images for use in experimental

research. *Behavior Research Methods*, *48*(4), 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582. https://doi.org/10.1037/0096-1523.26.5.1570

Jessen, M. (1998). *Phonetics and phonology of tense and lax obstruents in German*. Amsterdam, the Netherlands: John Benjamins. https://doi.org/10.1075/sfsl.44

Jiang, J. J., Zhang, Y., & McGilligan, C. (2006). Chaos in voice, from modeling to measurement. *Journal of Voice*, *20*(1), 2–17. https://doi.org/10.1016/j.jvoice.2005.01.001

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011. https://doi.org/10.1111/j.1467-7687.2011.01052.x

Kingston, J., & Diehl, R. L. (1995). Intermediate properties in the perception of distinctive feature values. In A. Arvaniti & B. Connell (Eds.), *Papers in Laboratory Phonology IV* (pp. 7–27). Cambridge, UK: Cambridge UP.

Kuberski, S. R., Tobin, S. J., & Gafos, A. I. (2016). A landmark-based approach to automatic voice onset time estimation in stop-vowel sequences. *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, *2016*, 60–65. https://doi.org/10.1109/GlobalSIP.2016.7905803

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lappin, J. S., Bell, H. H., Harm, O. J., & Kottas, B. (1975). On the relation between time and space in the visual discrimination of velocity. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(4), 383–394. https://doi.org/10.1037/0096-1523.1.4.383

Lappin, J. S., Norman, J. F., & Phillips, F. (2011). Fechner, information, and shape perception. *Attention, Perception, & Psychophysics*, *73*(8), 2353–2378. https://doi.org/10.3758/s13414-011-0197-4

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.

Mann, V. A., Diamond, R., & Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology*, *27*(1), 153–165. https://doi.org/10.1016/0022-0965(79)90067-5

McNeill, D., & Repp, B. (1973). Internal processes in speech perception. *The Journal of the Acoustical Society of America*, *53*(5), 1320–1326. https://doi.org/10.1121/1.1913473

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39–74). Hillsdale, NJ: Erlbaum.

Miller, J. L., & Eimas, P. D. (1983). Studies on the categorization of speech by infants. *Cognition*, *13*(2), 135–165. https://doi.org/10.1016/0010-0277(83)90020-3

Morrongiello, B. A., Robson, R. C., Best, C. T., & Clifton, R. K. (1984). Trading relations in the perception of speech by 5-year-old children. *Journal of Experimental Child Psychology*, *37*(2), 231–250. https://doi.org/10.1016/0022-0965(84)90002-X

Mulak, K. E., & Best, C. T. (2013). Development of word recognition across speakers and accents. In L. Cogate & G. Hollich (Eds.), *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence* (pp. 242–269). Hershey, PA: Information Science Reference. https://doi.org/10.4018/978-1-4666-2973-8.ch011

Oakes, L. M., Sperka, D. J., & Cantrell, L. (2015). *Habit 2*. Unpublished software. Davis, CA: Center for Mind and Brain, University of California, Davis. Retrieved from http://habit.ucdavis.edu/

Pater, J., Stager, C., & Werker, J. F. (2004). The perceptual acquisition of phonological contrasts. *Language*, *80*(2), 384–402. https://doi.org/10.1353/lan.2004.0141

Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3, Pt.1), 353–363. https://doi.org/10.1037/h0025953

Quam, C., Knight, S., & Gerken, L. (2017). The distribution of talker variability impacts infants' word learning. *Laboratory Phonology*, *8*(1), 1–27. https://doi.org/10.5334/labphon.25

Quinn, P. C., & Bhatt, R. S. (2010). Learning perceptual organization in infancy: The effect of simultaneous versus sequential variability experience. *Perception*, *39*(6), 795–806. https://doi.org/10.1068/P6639

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110. https://doi.org/10.1037/0033-2909.92.1.81

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074. https://doi.org/10.1068/p281059

Singh, L., Morgan, J. L., & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, *51*(2), 173–189. https://doi.org/10.1016/j.jml.2004.04.004

Singh, L., Nestor, S. S., & Bortfeld, H. (2008). Overcoming the effects of variation in infant speech segmentation: Influences of word familiarity. *Infancy*, *13*(1), 57–74. https://doi.org/10.1080/15250000701779386

Singh, L., White, K., & Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, *4*(2), 157–178. https://doi.org/10.1080/15475440801922131

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language*, *164*, 77–105. https://doi.org/10.1016/j.bandl.2016.10.004

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*, 381–382. https://doi.org/10.1038/41102

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095. https://doi.org/10.1037/0096-1523.7.5.1074

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *The Journal of the Acoustical Society of America*, *62*(2), 435–448. https://doi.org/10.1121/1.381544

Tobin, S., Hullebus, M., & Gafos, A. (2018). Immediate phonetic convergence in a cue-distractor paradigm. *The Journal of the Acoustical Society of America Express Letters*, *144*, 528–534. https://doi.org/10.1121/1.5082984

Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive Science*, *42*(S2), 413–438. https://doi.org/10.1111/cogs.12539

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word–object associations by 14-month-old infants. *Developmental Psychology*, *34*(6), 1289–1309. https://doi.org/10.1037/0012-1649.34.6.1289

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, *3*(1), 1–30. https://doi.org/10.1207/S15327078IN0301_1

Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, *12*(3), 412–418. https://doi.org/10.1111/j.1467-7687.2008.00789.x

How to cite this article: Höhle B, Fritzsche T, Meß K, Philipp M, Gafos A. Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning. *Dev Sci*. 2020;23:e12950. https://doi.org/10.1111/desc.12950