American Psychological Association

Journal of Experimental Psychology: Learning, Memory, and Cognition

2020, Vol. 46, No. 1, 189–199
http://dx.doi.org/10.1037/xlm0000704

Specialized Memory Systems for Learning Spoken Words

James M. McQueen
Radboud University and Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

Frank Eisner
Radboud University

Merel A. Burgering and Jean Vroomen
Tilburg University

Learning new words entails, inter alia, encoding of novel sound patterns and transferring those patterns from short-term to long-term memory. We report a series of 5 experiments that investigated whether the memory systems engaged in word learning are specialized for speech and whether utilization of these systems results in a benefit for word learning. Sine-wave synthesis (SWS) was applied to spoken nonwords, and listeners were or were not informed (through instruction and familiarization) that the SWS stimuli were derived from actual utterances. This allowed us to manipulate whether listeners would process sound sequences as speech or as nonspeech. In a sound–picture association learning task, listeners who processed the SWS stimuli as speech consistently learned faster and remembered more associations than listeners who processed the same stimuli as nonspeech. The advantage of listening in "speech mode" was stable over the course of 7 days. These results provide causal evidence that access to a specialized, phonological short-term memory system is important for word learning. More generally, this study supports the notion that subsystems of auditory short-term memory are specialized for processing different types of acoustic information.

Keywords: word learning, phonological loop, STM, sine-wave synthesis, speech

Supplemental materials: http://dx.doi.org/10.1037/xlm0000704.supp

Learning new words requires, as one key component, the ability to decode and store novel sound patterns. In this series of experiments, we ask whether specialized, phonological memory processes contribute to the learning of new words or, alternatively, whether it is supported only by general auditory perception and memory systems.

Although it is quite well understood how verbal information is maintained in short-term memory (STM), we know little about the mechanisms and constraints involved in learning and remembering other types of auditory material. In existing models of auditory STM, such as Baddeley and Hitch's working memory model (Baddeley, 2012; Baddeley & Hitch, 1974), auditory information is stored in a component called the *phonological loop* and maintained using articulatory rehearsal. Despite its name, however, the phonological loop in the multicomponent model of working memory does not differentiate between different types of sounds, such as speech, nonverbal vocalizations, music, or environmental sounds (Baddeley, Gathercole, & Papagno, 1998). Even nonverbal auditory features, such as pitch, are assumed to be stored in the loop and maintained by articulatory rehearsal (Williamson, Baddeley, & Hitch, 2010). Although there has been some disagreement about this assumption (Baddeley, 2012; Shah & Miyake, 1999), not much is known about how nonspeech sounds are processed in STM. Recent research on nonverbal auditory materials that cannot be articulated, such as musical timbre, has shown that such sounds behave differently from speech in a verbal interference memory task and has raised the possibility that there are, in fact, subsystems of auditory STM that specialize in different types of auditory objects (Soemer & Saito, 2015).

Word learning requires not only STM but also long-term memory processes, as newly learned words are integrated with existing lexical representations in the mental lexicon. In a recent account of novel-word learning by Davis, Di Betta, Macdonald, and Gaskell

This article was published Online First March 18, 2019.

James M. McQueen, Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University, and Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands; Frank Eisner, Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University; Merel A. Burgering and Jean Vroomen, Department of Cognitive Neuropsychology, Faculty of Social and Behavioral Science, Tilburg University.

Merel A. Burgering is now at the Department of Molecular Cell Biology, Vrije Universiteit Amsterdam.

The research was funded by the Gravitation programme "Language in Interaction" from the Dutch Science Foundation (NWO). We thank Natalie van Dis and Ferry Venselaar for their assistance, Judith Kroll for her permission to use the nonobject drawings, and Florian Hintz for computing the Levenshtein distances in Experiment 4. Part of this work was presented at the Annual Meeting of the Psychonomic Society, Chicago, Illinois, November 2015.

Correspondence concerning this article should be addressed to James M. McQueen, Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University, Nijmegen, Montessorilaan 3, 6525 HR, Nijmegen, the Netherlands. E-mail: j.mcqueen@donders.ru.nl

(2009), which is based on the more general complementary-systems memory model (McClelland, McNaughton, & O'Reilly, 1995), a novel spoken word is linked to episodic memory involving hippocampal systems, and from there is transferred to long-term memory representations in the mental lexicon in a medial temporal-lobe network (Bakker, Takashima, van Hell, Janzen, & McQueen, 2014, 2015; Davis et al., 2009; Davis & Gaskell, 2009; Gaskell & Dumay, 2003; Lindsay & Gaskell, 2013). This model does not make strong claims about whether the initial STM encoding involves specialized phonological processes or only general auditory processes.

If long-term memory encoding of novel words is dependent on STM, the quality of short-term encoding should have long-term consequences. Indeed, there is correlational evidence for a link between phonological STM, as measured by nonword repetition tasks, and long-term learning of novel phonological word forms (Baddeley et al., 1998). The present study aimed to establish whether there is a causal link between STM encoding of novel word forms and long-term retention. We tested whether learning and memory for sound patterns is more effective when acoustic information can be represented phonologically. More specifically, we asked whether learning associations between acoustic stimuli and pictures of nonsense objects is more effective when the acoustic stimuli can be represented as consisting of phonological units and sequences of those units, that is, as the vowels and consonants of human speech and the ways those sequences are combined in spoken language. If so, this would show that word learning depends, in part, on a specialized, phonological STM system.

Sine-wave synthesis (SWS) was used as a means to manipulate whether an acoustic signal is likely to be represented phonologically or not. SWS creates a "replica" of the original speech signal by tracking the first three formants and replacing them with time-varying tones (Remez, Rubin, Pisoni, & Carrell, 1981), while discarding all other information in the signal. To naïve listeners, the three overlaid tones resemble something like computer-generated bleeping and are not normally perceived spontaneously as degraded speech. In contrast, it has been shown that when listeners are informed that the signal is in fact recognizable speech (i.e., they are encouraged to listen in "speech mode"), they can process the signal phonologically and, in many cases, can understand what was being said (e.g., Baart, Vroomen, Shaw, & Bortfeld, 2014; Vroomen & Baart, 2009).

It is important to emphasize that being in speech mode goes beyond being able to detect that the signal originated from a human vocal tract. Consider the case of listening to an unknown (but undegraded) foreign language. In this case, listeners can certainly identify that they are listening to human speech, but depending on the phonological similarity of the new language to languages that the listeners do speak, they may have a limited ability to build and maintain phonological representations of the content of the foreign speech and the ways in which that content is sequenced. That is, they may be limited in their ability to engage speech mode.

We hypothesized that Dutch listeners in speech mode are able to construct phonological representations of novel, sine-wave-synthesized Dutch nonwords in STM, whereas naïve listeners have to rely on general auditory processing alone. Once phonological representations have been formed, they can be maintained in STM through (subvocal) rehearsal (Baddeley, 2012; Baddeley et al., 1998; Baddeley & Hitch, 1974). If access to a phonological STM representation is causally linked with novel word learning, then informed listeners should be able to learn and remember new words more easily than naïve listeners. Based on this rationale, an association learning task was administered in a between-subjects design. Two groups of listeners, a naïve group and an informed (speech mode) group, were asked to learn 12 pairs of auditory nonwords and visual nonobjects. The difference in treatments of the groups was in the initial phase of the experiment, when participants were familiarized with the 12 stimuli they had to learn. The two groups received slightly different instructions, such that the informed group was instructed to learn associations between pictures and distorted spoken nonwords, whereas the naïve group was instructed to learn associations between pictures and computer-generated sounds. Effects were measured both during the encoding phase and in a later recognition memory test.

According to our hypothesis, participants in the informed group should be in speech mode, and hence should be able to represent the SWS nonwords phonologically (i.e., as a sequence of vowels and consonants), and hence should be able to access phonological STM systems during learning. They should therefore be able to memorize the sound–picture associations more quickly and show better long-term retention. Participants in the naïve group, in contrast, are expected to process the same sounds as nonspeech using general auditory systems alone and therefore should be less efficient during learning and should remember fewer of the learned associations at test.

Experiment 1

Method

Participants. Thirty-five participants were recruited through the online research participation system at Radboud University and received course credit or monetary compensation. None reported hearing problems and all had normal or corrected-to-normal vision. They were all native speakers of Dutch. They were randomly assigned to one of the two groups, in a protocol aiming for usable data from 16 participants per group. Two participants in the naïve group were replaced because they did not complete the training within 20 min, and one participant in the informed group was replaced for failing to perceive the sine-wave sequences as speech. The final set thus comprised 32 participants (16 in each group), aged 18 to 27 years ($M = 22.3$, $SD = 2.4$), of which 24 were female. The study was approved by the ethics committee of the social sciences faculty at Radboud University. All participants gave written, informed consent prior to taking part.

Materials. The nonobjects were vectorized versions of images developed by Kroll and Potter (1984) and digitized by Brooks and Bieber (1988). These images depict line drawings of shapes that superficially have an object-like appearance but have little resemblance or association with real objects. A subset of 12, with the Numbers 1, 2, 4, 5, 9, 25, 27, 28, 31, 32, 37, and 38 in the original article (Kroll & Potter, 1984), was selected. The digitized images were processed using the trace tool in Adobe Illustrator (CS5) in order to reduce pixilation and then saved in a higher resolution bitmap format. An example nonobject stimulus is shown in Figure 1.

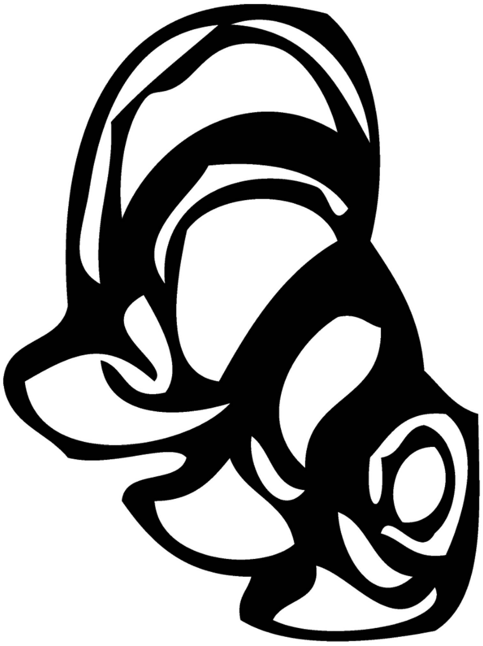

Figure 1. An example nonobject.

Speech materials consisted of six bisyllabic and six trisyllabic nonwords that were phonologically legal in Dutch (*boridal*, *gerikel*, *gruma*, *hemer*, *kaplavij*, *luinter*, *nimsel*, *plaker*, *trimonee*, *vertropel*, *vuimel*, and *zaandium*). Recordings were made of a female native Dutch speaker producing these nonwords in isolation and with a slightly falling intonation contour, using a Shure SM57 microphone connected to a Mac Pro running Audacity (2.0.6). Recordings were saved with a 44.1-kHz sampling rate at 16-bit quantization and processed further with Praat (Boersma & Weenink, 2014). Twelve nonword utterances were cut out at zero crossings and scaled to the same peak amplitude. Sine-wave replicas of these sounds were made by combining three time-varying sinusoids, tracking the first three formants (Remez et al., 1981) using a Praat script (Darwin, 2003). Figure 2 provides examples of a stimulus before and after conversion to SWS. Audio files of this stimulus (*trimonee*) before and after conversion are available in the online supplemental materials.

Design and procedure.

Overview. The experiment consisted of four phases: three training phases and a test phase (see also Table 1). The second and third training phases and the test phase were identical for the two groups; they differed only the first training phase, with respect to the instructions they received about the SWS stimuli and how they were familiarized with those stimuli.

Training. During the first phase, participants received on-screen written instructions and were familiarized with the sounds. The naïve group was told that they had to learn associations between nonexisting objects and computer-generated sounds. They then heard each of the 12 sine-wave replicas three times in a row. In contrast, the informed group was instructed to learn associations between nonexisting objects and made-up Dutch words that had been distorted. This group heard each of the sine-wave sequences twice, with the respective unmanipulated version of the nonword in between (i.e., SWS, clear, SWS).

During the second training phase, each sine-wave sequence was presented simultaneously with its associated picture, which remained centered on the screen for 1.5 s. The interval between two presentations of a sound–picture pair was 500 ms. The pairings were always the same and were always presented in the same order.

The third phase was the actual training and consisted of blocks of 12 trials. A trial started with the simultaneous playback of a sine-wave sequence and the display of a 2 × 2 array of nonobjects, one of which was associated with the sound. Participants were instructed to select the object that belonged to the sound using a computer mouse. Following their selection, they immediately received feedback as to whether the choice was correct or incorrect, and then the sound and its associated picture were presented again after the distractors disappeared from the screen. Within a block, each pair occurred once in a random order, and each picture occurred once as a target and three times as a distractor. Screen position of targets and distractors was pseudorandomized such that each had an equal probability of occurring in a particular position in the 2 × 2 grid. The number of correct responses within a block was tracked, and the training phase ended when a criterion of at least 80% correct responses ($\geq$10 out of 12) was reached or when the training had gone on for more than 20 min. Participants who exceeded the 20-min limit were excluded from the analysis and replaced; for all others, the number of blocks needed to reach criterion was used as the dependent measure of speed of learning.

Test. The last phase of the experiment took place 2 days after training. The test procedure was identical to Phase 3 of the training, except that no feedback and no repetitions were provided and that the experiment ended after three blocks. The proportion of correct responses across the three blocks was the dependent measure of memory retention.

Both parts of the experiment were run in MATLAB (R2013b) with the Psychophysics Toolbox (Brainard, 1997) on a Dell Precision 3610 desktop computer. Sounds were played over dynamic open-back headphones (Beyerdynamic DT990Pro) at a comfortable level. Pictures were approximately 6 × 6 cm and displayed in black on a white background on a 24-in. LCD monitor (BenQ XL2420Z). Participants were seated approximately 50 cm from the screen. They could move on to the next trial by pressing a button on the keyboard.

Results

Data from the training and test phases were analyzed separately using the Statistics Toolbox in MATLAB (R2014b). Effect sizes were calculated using the Effect Size Toolbox for MATLAB (v. 1.3, Hentschke & Stüttgen, 2011); we report eta squared (η^2) for analyses of variance and Hedges' g for t tests. The training and test results are summarized in Figure 3. A two-tailed t test for independent samples showed that the naïve group needed significantly more training blocks than the informed group to reach criterion (respective mean number of blocks = 5.25 and 2.88), $t(1, 30) = 2.61$, $p = .014$, $g = 0.9$. The naïve group also gave fewer correct responses than the informed group in the test phase (respective mean proportion correct = 0.705 and 0.802), $t(1, 30) = 2.57$, $p = .015$, $g = 0.89$.

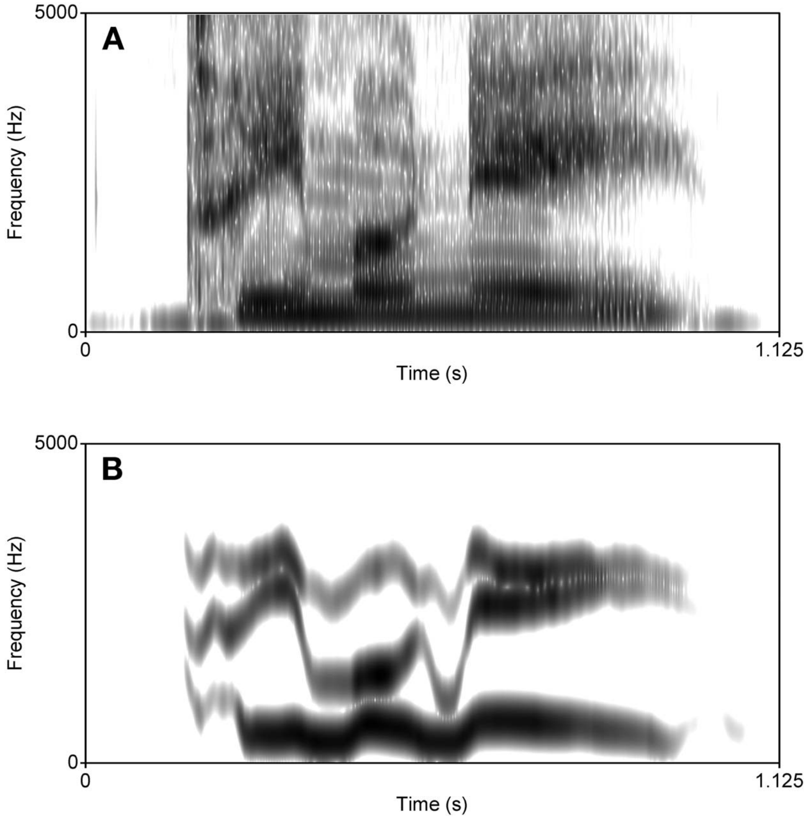

Figure 2. Spectrograms of the nonword *trimonee* [trimɔne:] in (A) clear and (B) sine-wave synthesized form.

Experiment 2

Memories for newly learned words change over time as a result of (sleep-enhanced) memory consolidation processes (Bakker et al., 2014, 2015; Davis et al., 2009; Davis & Gaskell, 2009; Gaskell & Dumay, 2003). The 2-day delay between training and the final test in Experiment 1 was therefore used to ensure that effects at final test reflected long-term memory retention, after consolidation, rather than immediate memory for the new picture–sound associations. Consolidation processes, however, continue for at least a week (Bakker et al., 2014; Gaskell & Dumay, 2003). An important question, therefore, is whether the benefits of phonological representation of the SWS nonwords continues beyond 2 days. In order to assess whether the memory advantage in speech mode is stable over time, Experiment 2 tested recognition memory after a delay of 1 week.

Method

Participants. Recruitment, consent, and replacement procedures were identical to Experiment 1. None of the 37 participants had taken part in Experiment 1. In the naïve group, one participant was replaced for failing to reach criterion within 20 min of training, and three were replaced because they reported hearing speech spontaneously. One participant in the informed group was replaced for failing to perceive the sine-wave sequences as speech. The final set therefore included 32 participants (16 per group, as planned) aged 18 to 28 years ($M = 22.3$, $SD = 2.7$), of which 20 were female.

Materials, design, and procedure. Experiment 2 was identical to Experiment 1 except that the test phase took place 1 week after training instead of after 2 days (see Table 1).

Results

As in Experiment 1, the naïve group needed more training blocks than the informed group to reach criterion (respective means = 4.81 and 3.18), $t(1, 30) = 2.11$, $p = .044$, $g = 0.73$. Again, the naïve group gave fewer correct responses than the informed group in the test phase (respective means = 0.665 and 0.783), $t(1, 30) = 2.52$, $p = .017$, $g = 0.87$. The combined results of Experiments 1 and 2 were analyzed in two 2 × 2 ANOVA, with factors Group (naïve, informed) and Experiment (2 days, 7 days), separately for the training and test phases. Both showed a main effect of group (training, $F[1, 60] = 11.24$, $p = .001$, $\eta^2 = 0.157$; test, $F[1, 60] = 12.79$, $p = .001$, $\eta^2 = 0.173$), but no significant main effect of experiment and no interaction between group and experiment (all η^2s $< .013$; see Figure 3). The speech-mode advantage observed in Experiment 1 was thus replicated in both the learning and test phases, and there was no significant decay in recognition after 1 week.

Experiment 3

This experiment examined more closely how being in speech mode can be induced. A potential concern about the previous

Table 1
Overview of Design Across Experiments

Phase	Experiment 1		Experiment 2		Experiment 3		Experiment 4	Experiment 5	
	Naïve	Informed	Naïve	Informed	Naïve	Informed	Informed	Naïve	Informed
Training: Preexposure	—	—	—	—	—	—	SWS Clear SWS	—	SWS Clear SWS
Training: SWS familiarization	SWS SWS SWS	SWS Clear SWS	SWS SWS SWS	SWS Clear SWS	SWS SWS SWS	SWS + T SWS + T SWS + T	SWS SWS SWS	SWS SWS SWS	SWS SWS SWS
Training: Associations	SWS-Picture pairs	SWS-Picture pairs	SWS-Picture pairs	SWS-Picture pairs	SWS-Picture pairs	SWS-Picture pairs	SWS-Picture pairs	SWS-Picture pairs	SWS-Picture pairs
Training: Learning to criterion	>80%	>80%	>80%	>80%	>80%	>80%	>80%	>80%	>80%
Test	After 2 days	After 2 days	After 7 days	After 7 days	After 2 days	After 2 days	After 2 days	After 2 days	After 2 days

Note. There was no naïve condition in Experiment 4; data in the informed condition in Experiment 4 were compared statistically with the naïve condition in Experiment 1. SWS = sine-wave speech; T = with orthographic transcription.

experiments is that the informed group knew not only that the sound sequences were manipulated speech but also that they had heard a clear, undistorted version of each nonword during the familiarization phase. To be sure that the effect we observed in Experiments 1 and 2 is driven by how listeners process the sine-wave speech, rather than by linking the sine-wave sequences to undistorted templates of each nonword held in memory, the familiarization procedure was changed. In Experiment 3, the two groups heard exactly the same as each other (only three times the sine-wave version of each nonword), but the speech mode group was again told that the sounds were spoken nonwords, and critically, this time they saw an orthographic transcription during familiarization. There was thus no clear acoustic template that could be linked to the distorted sounds. The main training and test sessions were identical to those in Experiments 1 and 2; the test was given after 2 days (as in Experiment 1).

Method

Participants. Seventy participants were recruited from the Tilburg University psychology subject pool via an online enrollment system. They were aged 18 to 28 (M = 19.6, SD = 3.12); 61 were female. When online enrollment ended, there were 35 participants each in the informed and naïve groups.

Materials, design, and procedure. Experiment 3 was identical to Experiment 1 (see Table 1), except that participants no longer heard an undistorted version of the training items during familiarization. Instead, they saw an orthographic transcription, which remained on the screen while three repetitions of each sine-wave speech token were played.

Results

The informed group was again significantly faster than the naïve group to reach criterion during training (Figure 4; respective means = 4.17 and 6.97), $t(1, 68) = 4.13$, $p = .0001$, $g = 0.98$, and identified more pairs correctly in the test phase (respective means = 0.77 and 0.70), $t(1, 68) = 2.52$, $p = .014$, $g = 0.60$. The speech-mode advantage observed in Experiments 1 and 2 was thus replicated using orthographic transcriptions instead of undistorted audio. To test whether the size of the advantage is modulated by the way in which speech mode was induced (auditory vs. orthographic), two ANOVAs with the factors Experiment (1 and 2 vs. 3) and Group (naïve, informed) were run on the training and test phase data. Critically, there was no significant interaction between group and experiment, neither in the training nor in the test data (all Fs $< .8$). Experiment 3 thus confirmed that speech mode can be induced as reliably using written forms of the nonwords as using natural spoken forms.

Experiment 4

Experiments 1, 2, and 3 showed significantly faster learning in the informed group during the training phase and significantly more remembered associations in the test phase. In Experiment 3, the unmanipulated auditory nonwords were no longer presented during familiarization but instead were presented together with an orthographic transcription. It is possible, however, that by reading the transcriptions during famil-

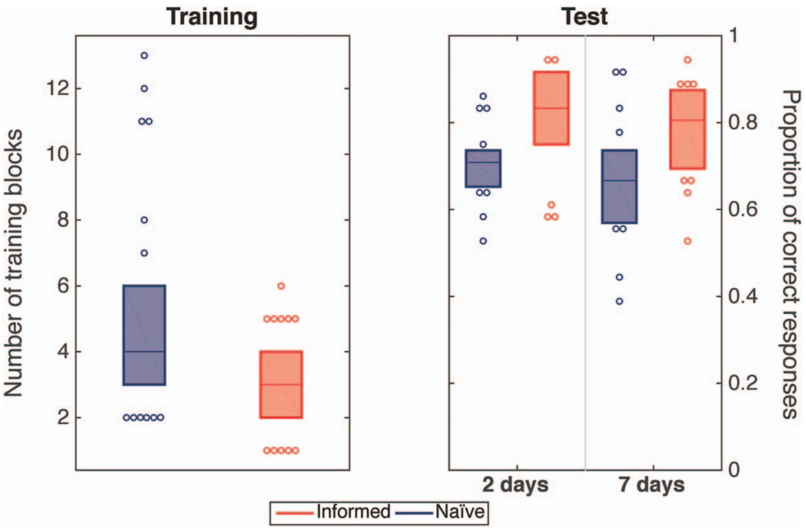

Figure 3. Boxplots showing the combined data from the training phase in Experiments 1 and 2 on the left, and performance in the test phase after 2 (Experiment1) and 7 days (Experiment 2) on the right. Horizontal lines represent the median, boxes show the interquartile range, and circles show data points outside the interquartile range. See the online article for the color version of this figure.

iarization, participants may have subvocalized during reading and/or rehearsal, thus constructing an undistorted acoustic template from the written input. In order to establish whether the opportunity for subvocal rehearsal of the training items may have driven the advantage in the informed group, we conducted a further experiment in which participants no longer were told what the training items were but instead were familiarized with sine-wave speech versions of different items in a preexposure phase. This meant that the participants had no exposure to the base words from which the critical SWS stimuli had been made, neither in spoken nor written form. Only this informed condition was tested; the data were compared with those from the naïve condition in Experiment 1.

We also asked the participants to transcribe the SWS stimuli in an additional posttest phase. We then counted the total number of stimuli that were transcribed as speech and computed the phonological distance between the participants' transcriptions and transcriptions of the original nonwords. This made it possible to ask whether these informed participants were indeed able to represent the stimuli phonologically and, more specifically, whether there was a correlation between their transcriptions and their memory performance. We predicted that participants who were better able to identify the stimuli phonologically (in terms of the total number of phonological transcriptions and/or the correctness of those transcriptions) would reach the training criterion earlier and/or remember more associations at final test. Because the Experiment

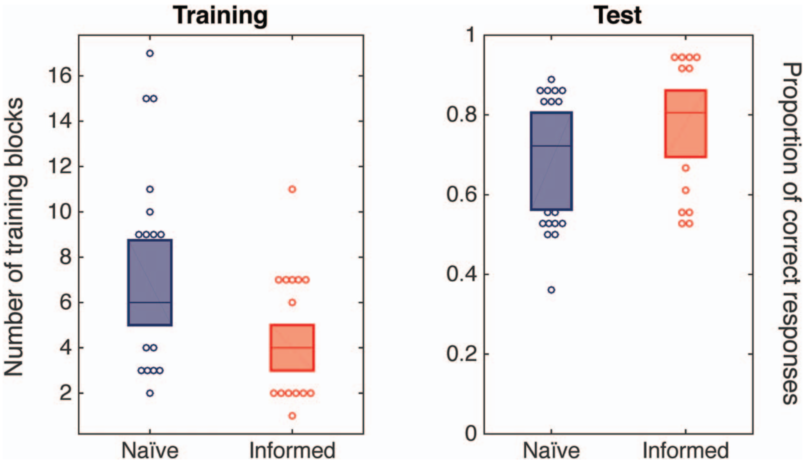

Figure 4. Boxplots showing results from the training phase (left) and test phase (right) in Experiment 3. Horizontal lines represent the median, boxes show the interquartile range, and circles show data points outside the interquartile range. See the online article for the color version of this figure.

4 participants neither heard any untransformed versions of the nonwords (as in the informed condition in Experiments 1 and 2) nor saw transcriptions of the original nonwords (as in the informed condition in Experiment 3), their posttest transcriptions indicate whether they could build phonological representations solely on the basis of experience with the SWS stimuli when in speech mode.

Method

Participants. Recruitment and consent procedures were identical to Experiments 1 and 2. Sixteen participants aged 18 to 26 years (*M* = 20.8, *SD* = 2.3; 14 were female) completed the experiment in the informed condition. None had taken part in the previous experiments.

Materials, design, and procedure. Experiment 4 was identical to the informed condition in Experiment 1, except for the following changes (see Table 1): First, the training phase began with a preexposure that was intended to familiarize listeners with the sine-wave manipulation. During preexposure, listeners heard 12 nonwords that were not part of the training set (six bisyllabic and six trisyllabic phonologically legal sequences in Dutch, which had been recorded in the same session as the training items: *daarster*, *duiderde*, *molmissie*, *neling*, *omidan*, *pongel*, *potering*, *rifteling*, *soerket*, *soperij*, *tiekel*, and *uifer*). Listeners heard each word twice as sine-wave replicas and the respective unmanipulated version once in between. Second, in the subsequent familiarization with the training items, no undistorted stimuli were played and, instead, participants heard only the distorted versions three times each.

Third, after the test session there was a brief posttest. Participants heard each of the 12 SWS items, in turn, and were asked to transcribe what they heard using the computer keyboard. These transcriptions were scored in two ways: We counted the number of stimuli that were transcribed as speech and we computed the Levenshtein distance (LD) between each transcription and transcriptions of the original nonwords. The LD (more precisely here, the Damerau-Levenshtein distance) was, for each of the 12 nonwords, the number of edits (additions, substitutions, deletions or transpositions) required to change the participants' transcriptions into the standard transcription. A smaller LD thus indicates that two transcriptions were more similar (LD = 0 means they are identical).

Results

Training and test performance were compared with that of the naïve group from Experiment 1 (see Figure 5). The informed group required fewer training blocks to reach criterion than the naïve group (respective means = 2.56 and 5.25), *t*(1, 30) = 2.91, *p* = .007, *g* = 1.00, and identified more pairs correctly in the test phase (respective means = 0.78 and 0.70), *t*(1, 30) = 2.12, *p* = .043, *g* = 0.73. These results replicate the previous experiments in both training and test by showing faster learning and better retention in the informed group. Experiment 4 suggests that the advantage for the informed group is not based on knowing the identity of the nonwords in advance but on how these listeners processed the distorted sounds.

The posttest responses are shown in Table 2. These transcriptions suggest that participants were largely able to extract phonological information from the distorted nonwords. On average, participants identified 10.6 of 12 items as speech (*SD* = 2.3). As can be seen in Table 2, in many cases the responses were correct or very close to the original nonword (mean LD = 4.6 edits, *SD* = 1.3). No significant correlations were observed between the number of stimuli participants transcribed as speech and either the number of training blocks they required to reach criterion, *r*(14) = −0.40, *p* = .13, or proportion of pairs identified correctly

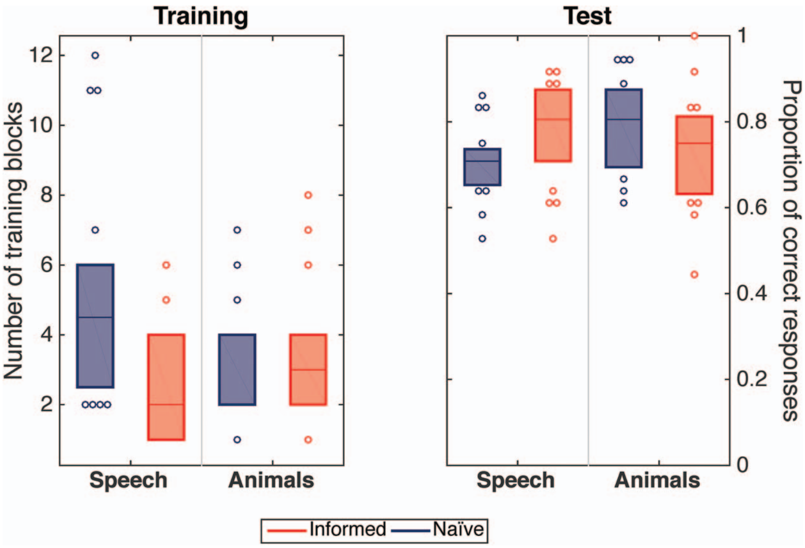

Figure 5. Boxplots contrasting performance with speech and animal vocalizations. The speech data are taken from Experiments 1 and 4 for the naïve and informed conditions, respectively. The animal vocalization data are taken from Experiment 5. Horizontal lines represent the median, boxes show the interquartile range, and circles show data points outside the interquartile range. See the online article for the color version of this figure.

Table 2
Responses from the Experiment 4 Posttest

Participant number	Boridal	Gerikel	Gruma	Hemer	Kaplavij	Luinter	Nimsel	Plaker	Trimonee	Vertropel	Vuimel	Zaandium
S4–10		krekel	chlora	neem maar	schilderij	zuigen	mimetol		neem maar mee		vluigel	champignon
S4–11	Oliedom	Krekel	Schommel	Neemaar	Katlafijn	Luideur	Nimmuzzel	Klaker	Tribonee	Serdrogol	Zwembroek	Sabijun
S4–12							biebetu	latu			abu	
S4–13	Olie . . .	. . . winkel		schemer	pak . . .	lui . . . \	. . . wissel	laken	neem mooi mee		zwembroek	sabium
S4–14	omidol	sleutel	owa	emer	daszofijn!	luister	nimesol	later	neem me mee	centrofon	zuivel	zaandium
S4–15			ole	immuun	kaklohey	mijntje	minescuul	kwaaku	simone		zuivel	fabien
S4–16	Oliedol	cevukum	ole	eemuu	kakluizijn	luintuul	nimzul	la-ul	symonie	suntro-un	zuimloel	saambium
S4–17	vogel	begrepen	mensen		afgevinkt	besproeid	onbemand	vreselijk	ingekort	onbeschreven	vervaagd	ontdaan
S4–1	euro			Evert	Kukeleku	landje	mimisoe	lachen!	Blijf bij mij		euro	champion
S4–2	sowieso	geregel	hoger	eeuwer	opderijm	later	nimbusul	later	trebonee	centrofoon	zuivel	zaambium
S4–4	oridol		saowee	emur	ozofijn	luideur	mimsol	luier	emonnee	centro-oon	zuivel	sabien
S4–5	poridoor	gewikkel	lomo	emer	ozofijn	luister	mimezol	lager	simone	verdwoku	zuimel	zadium
S4–6	oliedom	zuwikoe	cho-u	ieeua	kokjezuim	luinter	nimwuzoe	klaken	tsimonei	certwouzoen	zuivel	zaambium
S4–7	ferico	ferico	faclafei	leri	faclafei	zaandium	nimmisseu	lager	rimonee		zeundum	zaandium
S4–8	Doridom	Surikum	Homme	Emur	Katlevein	Luinder	Nimbusel	Laker	Simone	Surtropfung	Suinbloem	Zaandium
S4–9	oliedom	suzidom	govert	evert	katlazijn	luister	meemoezel	laaster	siemomee	seedroogstam	suifel	sabiun

in the test phase, $r(14) = -0.27$, $p = .32$, presumably because the number of transcriptions was close to ceiling. Although there was also no significant correlation between LD and proportion correct at test, $r(14) = -0.05$, $p = .84$, there was a positive correlation between LD and number of training blocks, $r(14) = 0.52$, $p = .04$. As predicted, participants who could more accurately transcribe the SWS stimuli were those who required less training to reach criterion.

Experiment 5

The previous experiments suggest that the consistent advantage in learning novel associations in the informed group is driven by being able to construct phonological representations of the auditory input. If this is the case, the advantage should be specific to speech and not apply to other types of auditory stimuli, for which a phonological representation is not possible. Experiment 5 addressed this issue by replacing the SWS versions of spoken nonwords with SWS versions of animal vocalizations.

Method

Participants. Recruitment, consent, and replacement procedures were the same as in Experiments 1, 2, and 4. Thirty-six participants were tested, but three were excluded (all in the informed condition) because they were either taking medication that could affect cognitive function ($n = 2$) or suffering from claustrophobia ($n = 1$). The final data set thus consisted of 33 participants (22 female; 16 in the naïve condition), aged 18–34 ($M = 22.4$, $SD = 4.0$).

Materials, design, and procedure. This experiment replicated the design of Experiment 4 except that participants learned to pair pictures with animal vocalizations instead of speech, and a naïve group was tested alongside an informed group (see Table 1). The stimulus materials for the training in the main experiment were 12 clips of vocalizations from a range of different animals: cat, crocodile, dog, elephant, frog, horse, rhinoceros, sea lion, and four different bird sounds (goose, hawk, owl, and whippoorwill). A further 12 vocalizations were used for preexposure. All 24 vocalizations were obtained from sound repositories and then sine-wave synthesized. Audio files for the horse vocalization, before and after transformation, are provided in the online supplemental materials. As in Experiment 4, there was a preexposure phase for the informed group only, in order to familiarize participants with the SWS manipulation without playing undistorted audio versions of the actual training items. During this phase, participants heard each of the 12 sounds three times, in distorted–clear–distorted order, as in Experiment 4.

Both groups then had the same familiarization with the training items, that is, they heard each sound three times in SWS form. Familiarization in the naïve condition was thus analogous to that in Experiment 1. The other phases of training and test were identical to those in Experiments 1 and 4 except that (as in only the latter experiment) there was also a posttest phase in which participants were asked to identify the SWS stimuli. Participants in both groups heard each of the 12 items in turn and were asked to indicate (through typed keyboard responses) what they thought the original sound was. These responses were analyzed in two ways: We computed, first, the proportion of stimuli each participant identified as being an animal vocalization of any type (as opposed to other kinds of sound), and second, the proportion of stimuli identified correctly (i.e., as a particular vocalization). In the latter case, responses to the four birds were counted as correct if they were labeled either as the correct bird, or as another bird, or as "bird" or "birds."

Results

Performance in both the training and test phases of Experiment 5 was, in general, very good but certainly not perfect (see Figure 5). The informed group required a similar number of training blocks as the naïve group (respective means = 3.47 and 3.63), $t(1, 31) = 0.25$, $p = .805$, $g = 0.085$, and remembered a similar proportion of pairs in the test phase (respective means = 0.79 and 0.73), $t(1, 31) = 1.26$, $p = .216$, $g = 0.430$. In order to compare the overall performance with animal sounds to speech, these data were also analyzed with an ANOVA that included the equivalent groups from Experiment 1 (naïve) and Experiment 4 (informed). Separate 2 × 2 ANOVAs were conducted for the training and test phases, with the factors Group (naïve, informed) and Experiment (speech, animal sounds). In the training data, there was a significant main effect of group, $F(1, 61) = 6.60$, $p = .013$, $\eta^2 = 0.090$, and a significant interaction with experiment, $F(1, 61) = 5.25$, $p = .026$, $\eta^2 = 0.072$, but no main effect of experiment, $F(1, 61) = 0.42$, $p = .519$, $\eta^2 = 0.006$. The test data also showed a significant interaction between group and experiment, $F(1, 61) = 5.48$, $p = .023$, $\eta^2 = 0.082$, but no significant effects of group, $F(1, 61) = 0.19$, $p = .661$, $\eta^2 = 0.003$, or experiment, $F(1, 61) = 0.32$, $p = .576$, $\eta^2 = 0.005$. There was thus no difference in overall task difficulty between the two experiments, and the absence of an advantage for the informed group in Experiment 5 cannot be attributed to the task being too difficult. Critically, these results show that there was no advantage of being in the informed condition when the sounds to be learned were animal vocalizations instead of speech.

Performance in both groups was comparable with that in the informed conditions in Experiments 1 to 4. This suggests that, at least numerically, it was slightly easier for participants to learn the sound–picture associations when the sounds were derived from animal vocalizations than when they were derived from human speech. As just noted, however, there was no significant difference in difficulty between experiments. Furthermore, the absolute level of performance in Experiment 5 is not the critical result; rather, it is the relative levels of performance in the two conditions. The lack of difference between the informed and naïve groups with stimuli based on animal vocalizations contrasts with differences between the groups in all the experiments with stimuli based on human speech. This contrast suggests that there is facilitation of learning in the informed group only when participants can use speech mode.

Analyses of the posttest data indicated that participants in the informed group were better able to identify the SWS stimuli as animal vocalizations than those in the naïve group. One participant was excluded from these analyses because they failed to give any responses on the posttest. The advantage for the informed group was observed both on the proportion of stimuli identified as any type of animal vocalization (informed, $M = 0.85$, $SD = 0.18$; naïve, $M = 0.44$, $SD = 0.22$), $t(30) = 5.73$, $p < .001$, and on the proportion of stimuli identified correctly (informed, $M = 0.27$, $SD = 0.14$; naïve, $M = 0.18$, $SD = 0.10$), $t(30) = 2.24$, $p = .033$. These analyses indicate not only that the between-groups manipulation was effective in encouraging the informed group to treat the SWS stimuli as animal vocalizations but also that the lack of difference between the two groups in learning was not because of either a ceiling or a floor effect (i.e., participants in both groups did not treat stimuli either always or never as vocalizations).

General Discussion

This series of experiments investigated whether there is a causal link from phonological STM to novel word learning. Experiment 1 showed that participants who had been informed that a set of auditory stimuli were distorted speech were faster to learn sound–picture associations and retained more of the associations after a 2-day delay than participants who were unaware that the sounds were in fact speech. STM and long-term memory encoding appears to be enhanced when listeners are able to process sounds phonologically. Experiment 2 replicated the speech-mode advantage in the training phase and showed that it persists in the recognition task after 1 week. Experiments 3 and 4 showed further that the better performance in the informed group did not depend on exposure to undistorted auditory versions of the SWS nonwords, and Experiment 4 indicated, in addition, that there was a correlation between the speed with which informed participants learned the sound–picture associations and the degree to which they correctly identified the phonological content of the original (undistorted) nonwords. Experiment 5 demonstrated that there was no benefit for the informed group when the SWS stimuli were made from nonhuman vocalizations.

The speech-mode advantage observed in Experiments 1 to 4 suggests that phonological processing, over and above general auditory processing, facilitates both encoding and retention of the novel sound–picture associations. One way of thinking about this result is in terms of schema theory (Bartlett, 1932). Existing sources of phonological knowledge about the vowels, consonants, and phonotactics of spoken language can be considered as schemata that support how the SWS stimuli are encoded in memory and retrieved from it. Although schema theory is usually applied to long-term memory, we suggest that the schema-like benefit in the present case arises primarily through the engagement of a STM system that is specifically phonological (e.g., the phonological loop; Baddeley & Hitch, 1974). A direct test of the idea that the present effects reflect engagement of phonological schemata would be to manipulate the phonological regularity of the nonwords. The group difference should decrease as the phonological properties of the nonwords (their vowels and consonants; their phonotactics) more strongly mismatch the phonological properties of the participants' native language. Such investigations could, in turn, start to specify which aspects of phonological knowledge modulate the operation of the STM system.

Once the SWS stimuli have been encoded phonologically, they can be maintained through rehearsal processes. It remains to be determined how crucial rehearsal is for the difference we observed between the informed and uninformed groups. It would therefore be informative to test, in future studies, whether this difference would be weakened under conditions of articulatory suppression.

Such a study would help to specify which components(s) of the STM system underlie the speech-mode advantage.

Our account is consistent with the results of studies on the neural response to SWS stimuli. Activity in posterior regions of the superior temporal sulcus and gyrus is enhanced after SWS stimuli (through training and instruction) start to be treated as speech (Dehaene-Lambertz et al., 2005; Möttönen et al., 2006). These regions are part of the dorsal stream (Hickok & Poeppel, 2007), which links perceptual processing to speech motor processing; this pathway would be required if the content of SWS speech is indeed maintained through articulatory rehearsal. A recent electrocorticography (ECoG) study (Khoshkhoo, Leonard, Mesgarani, & Chang, 2018) supports this view by providing evidence that there was activity in inferior frontal cortex (i.e., the endpoint of the dorsal stream) only when SWS stimuli were comprehended (i.e., only when the ECoG patients were in speech mode). It would be interesting to test, in an experiment based on the current design, whether the dorsal stream is more strongly engaged in the informed group than in the uninformed group.

Our findings are consistent with previous claims that phonological STM is critical for learning novel words (Baddeley, 2003; Baddeley et al., 1998; Gathercole, 2006) but extends this work by showing that there is not only a correlation but also a causal connection. Because informed and naïve participants had exactly the same physical exposure during learning, and the experimental manipulation took place before learning began, the groups differed only in how they processed the stimulus materials during learning. If informed participants were able to use specialized phonological memory systems, they could map the distorted nonwords onto existing phonological structures, such as phonetic categories and phonotactic patterns, and hence engage rehearsal processes to maintain these representations in STM. Naïve participants, in contrast, would not be able to utilize existing phonological categories and had to encode the sine-wave stimuli through general auditory processes alone. The specialized phonological systems thus appear to play a causal role in word learning. This interpretation is consistent with previous studies showing that categorical information being available along with sensory input assists in the STM encoding of visual (Olsson & Poom, 2005) and auditory (Li, Cowan, & Saults, 2013) objects.

The ability to use a specialized, phonological STM system would facilitate integration of new words with existing lexical entries and the transfer to long-term memory—an integration process that is still used frequently in adults as they learn new vocabulary. The fact that there seemed to be a knock-on effect from more efficient short-term encoding during the learning phase to more efficient retention in the long-term memory test is functionally consistent with the two-step complementary-systems account of word learning (Bakker et al., 2014; Davis et al., 2009; Davis & Gaskell, 2009; McClelland et al., 1995). However, our results suggest that the initial encoding of the associations into episodic memory already takes advantage of the specialized phonological STM system. Clearly, however, word learning cannot be achieved through this system alone. Most obviously, additional mechanisms of associative memory are required to link the sounds (however they are represented) to the pictures of the nonsense objects.

The posttest results of Experiment 4 offer further support for the claim that the benefits observed in the informed group reflect the use of a STM system that is phonological in nature. Participants who were faster to learn the sound–picture associations were better able to transcribe the SWS stimuli phonologically (i.e., their transcriptions were closer to the form of the original nonwords from which the SWS stimuli had been derived). Associative learning appears to be easier when the phonological processes of speech mode can be more strongly engaged.

Experiment 5 found no benefit for the informed condition when the auditory materials were animal vocalizations instead of speech. Although animal vocalizations are likely to have a categorical organization in memory, this is unlikely to be as rich and detailed as the phonetic and phonological systems for speech and does not have a structure equivalent to the mental lexicon into which the novel stimuli could be integrated for long-term storage. The results of Experiment 5 are thus further evidence that the speech-mode advantage is specific to speech because only speech can be processed and represented phonologically. Being informed about what type of sound the distorted stimuli were originally and, hence, being better at identifying the vocalizations was thus not beneficial on its own for associative learning.

The observed advantage for the informed groups does not depend on having an undistorted acoustic template available for subvocal rehearsal, as both Experiments 3 and 4 show that the advantage can be driven only by an expectation about what may be in the signal, without having prior knowledge about the actual content. More generally, these findings (including the posttest identification data in Experiment 4) imply that whether a sound enters phonological STM or not can be determined by prior knowledge about only the type of content of the SWS stimuli (i.e., that they are derived from spoken nonwords) rather than about their actual content.

Overall, these results suggest that speech has a special status in phonological STM. Together with some previous studies (Golubock & Janata, 2013; Soemer & Saito, 2015), however, our findings raise questions about an assumption in the phonological loop model, namely, that sounds that are not phonological and cannot be verbalized are in fact stored in a phonological code and rehearsed by an articulation-based mechanism (Williamson et al., 2010). Our findings suggest that that the multicomponent model of working memory might need to be extended to account for different types of memory representation for different types of auditory stimuli.

Conclusions

Being able to process novel words through a specialized, phonological STM system leads to faster learning and better long-term retention, lasting at least 1 week. Our findings are new evidence for the existence of such a specialized component of STM as well as for a causal link between phonological STM and word learning. Auditory STM appears to consist of distinct components for different types of auditory objects.

References

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition, 130,* 31–43. http://dx.doi.org/10.1016/j.cognition.2013.09.006

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36,* 189–208. http://dx.doi.org/10.1016/S0021-9924(03)00019-4

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63,* 1–29. http://dx.doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158–173. http://dx.doi.org/10.1037/0033-295X.105.1.158

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York, NY: Academic Press.

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Competition from unseen or unheard novel words: Lexical consolidation across modalities. *Journal of Memory and Language, 73,* 116–130. http://dx.doi.org/10.1016/j.jml.2014.03.002

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2015). Changes in theta and beta oscillations as signatures of novel word consolidation. *Journal of Cognitive Neuroscience, 27,* 1286–1297. http://dx.doi.org/10.1162/jocn_a_00801

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.

Boersma, P., & Weenink, D. (2014). Praat (Version 5.4) [Computer software]. Retrieved from http://www.fon.hum.uva.nl/praat/

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Brooks, J. O., & Bieber, L. L. (1988). Digitized nonobjects for use with the Apple Macintosh computer. *Behavior Research Methods, 20,* 433–434.

Darwin, C. J. (2003). SWS Praat script. Retrieved from http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/

Davis, M. H., Di Betta, A. M., Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience, 21,* 803–820. http://dx.doi.org/10.1162/jocn.2009.21059

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 364,* 3773–3800. http://dx.doi.org/10.1098/rstb.2009.0111

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage, 24,* 21–33. http://dx.doi.org/10.1016/j.neuroimage.2004.09.039

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89,* 105–132. http://dx.doi.org/10.1016/S0010-0277(03)00070-2

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 27,* 513–543. http://dx.doi.org/10.1017/S0142716406060383

Golubock, J. L., & Janata, P. (2013). Keeping timbre in mind: Working memory for complex sounds that can't be verbalized. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 399–412. http://dx.doi.org/10.1037/a0029720

Hentschke, H., & Stüttgen, M. C. (2011). Computation of measures of effect size for neuroscience data sets. *European Journal of Neuroscience, 34,* 1887–1894. http://dx.doi.org/10.1111/j.1460-9568.2011.07902.x

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8,* 393–402. http://dx.doi.org/10.1038/nrn2113

Khoshkhoo, S., Leonard, M. K., Mesgarani, N., & Chang, E. F. (2018). Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. *Brain and Language, 187,* 83–91. http://dx.doi.org/10.1016/j.bandl.2018.01.007

Kroll, J. F., & Potter, M. C. (1984). Recognizing words, pictures, and concepts: A comparison of lexical, object, and reality decisions. *Journal of Verbal Learning & Verbal Behavior, 23,* 39–66. http://dx.doi.org/10.1016/S0022-5371(84)90499-7

Li, D., Cowan, N., & Saults, J. S. (2013). Estimating working memory capacity for lists of nonverbal sounds. *Attention, Perception, & Psychophysics, 75,* 145–160. http://dx.doi.org/10.3758/s13414-012-0383-z

Lindsay, S., & Gaskell, M. G. (2013). Lexical integration of novel words without sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 608–622. http://dx.doi.org/10.1037/a0029243

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457. http://dx.doi.org/10.1037/0033-295X.102.3.419

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., & Sams, M. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *NeuroImage, 30,* 563–569. http://dx.doi.org/10.1016/j.neuroimage.2005.10.002

Olsson, H., & Poom, L. (2005). Visual memory needs categories. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 8776–8780. http://dx.doi.org/10.1073/pnas.0500810102

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947–949. http://dx.doi.org/10.1126/science.7233191

Shah, P., & Miyake, A. (1999). Working models of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 1–27). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139174909.004

Soemer, A., & Saito, S. (2015). Maintenance of auditory-nonverbal information in working memory. *Psychonomic Bulletin & Review, 22,* 1777–1783. http://dx.doi.org/10.3758/s13423-015-0854-z

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition, 110,* 254–259. http://dx.doi.org/10.1016/j.cognition.2008.10.015

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory & Cognition, 38,* 163–175. http://dx.doi.org/10.3758/MC.38.2.163

Received October 31, 2018
Revision received February 1, 2019
Accepted February 5, 2019 ■