The time-course of talker-specificity effects for newly-learned pseudowords: Evidence for a hybrid model of lexical representation

Helen Brown, M. Gareth Gaskell

Department of Psychology, University of York, UK

h.brown@psych.york.ac.uk, g.gaskell@psych.york.ac.uk

Abstract

Whilst research shows that talker information affects recognition of recently studied words, it remains unclear whether this information is stored in long-term memory. Three experiments explored whether talker-specificity effects (TSEs) for pseudowords changed over time and were affected by within- and between-talker variability during study. Results showed TSEs immediately after study in all experiments, consistent with episodic models, but TSEs remained a week later only for pseudowords studied in a single voice. Furthermore, source memory data suggested that talker information becomes less accessible over time, supporting hybrid models that incorporate aspects of both episodic and abstract lexical representation.

Index Terms: speech perception; first language acquisition.

1. Introduction

Over the years there has been substantial debate as to whether lexical representations of words have an episodic form, containing all perceptual and contextual details specific to each single occurrence of a word, or whether representations are more abstract, with words being reduced to a sequence of idealized phonemes through normalization processes that strip away all perceptually and contextually specific details.

One advantage of episodic models is that they allow talker information to be retained during spoken word recognition since any given speech input provides information not only about the linguistic message, but also about the speaker and speaking environment [1]. There are now several studies showing that talker-specific information is retained in memory immediately after words are encountered, consistent with the predictions of episodic models [e.g. 2]. However, given the absence of TSEs in other experiments [e.g. 3] there has been a gradual convergence towards hybrid models of lexical representation that involve aspects of both episodic and abstract representation [4]. One such model is the complementary learning systems framework [5] in which newly-encountered information is rapidly encoded into highly-detailed, pattern-separated representations within the hippocampal network [6], but over time memory becomes gradually more reliant on distributed representations in neocortical regions that may be more abstract in nature.

Goldinger [7] examined the time-course of TSEs for existing words by exposing participants to 150 words, each spoken by a single talker from a set of two, six, or ten talkers. TSEs were tested using an old/new categorization task in which both studied and unstudied items were presented, and an identification-in-noise task. Participants were tested either immediately after studying the existing words, one day later, or one week later. Whilst TSEs were present at all time-points in the identification-in-noise task, TSEs decreased across the course of a week in the old/new categorization task, suggesting that information about the study talker became gradually less influential in the recognition of previously studied words as time progressed. These findings clearly indicate that immediately after study recognition of studied words relies on highly-detailed episodic traces. The question is whether recognition of studied words after a delay relies on representations in long-term memory that are abstract in nature, consistent with hybrid models of lexical representation, or whether the lexicon consists of multiple exemplars of each individual word, consistent with episodic models. It is difficult to differentiate between these two possibilities when studying existing words that already have pre-existing representations in long-term lexical memory because even episodic models would predict a decrease in TSEs as recognition memory becomes more reliant on long-term memory. According to episodic models all previous episodic traces of an existing word will be activated during spoken word recognition, rendering recent talker information less influential as a result of participants having encountered the existing words many times prior to the experimental session, spoken by many different speakers in a variety of contexts.

A different way to address the nature of lexical representation in long-term memory is to examine TSEs over time for pseudowords, where the number of exposures to each item, and the number of talkers per item can be controlled more closely. According to hybrid models TSEs should decrease over time for recently-encountered pseudowords as they develop abstract representations within long-term lexical memory. This should be the case regardless of the amount of talker variability encountered during initial encoding of the pseudowords. In comparison, episodic models predict that TSEs for pseudowords should only decrease over time when each item is spoken by multiple talkers during study. Here we present three experiments that systematically manipulated the degree of talker-variability during the study of pseudowords, allowing us to differentiate between the predictions of episodic and hybrid models outlined above. Experiment 1 examined the case where only one token from a single talker was presented during study whilst Experiments 2 and 3 explored the effects of increasing variability during study, examining within- and between-talker variability respectively. Given the similarity in the design of all three experiments the method and results sections will describe all experiments together, allowing cross-experiment comparisons to be made more easily.

2. Method

1.1. Participants

Forty-eight participants completed Experiment 1, and thirty-two completed Experiments 2 and 3. All were undergraduate students at the University of York, and were native British English speakers reporting no hearing, speech, or language impairments at the time of testing. Informed consent was obtained from all participants prior to the first experimental session. Participants took part in only one of the three experiments.

1.2. Stimuli

Forty-eight pseudowords and their corresponding foil items were selected from stimuli used in a previous longitudinal study of word learning [8]. Targets and foils differed only at the final consonant cluster (e.g., *dolpheg-dolphess*). The items were split into two lists of 24 items, matched on initial phoneme, number of syllables, and as closely as possible on number of phonemes, and CELEX frequency [9].

For Experiment 1 (*no variability*) two native British English speakers (one male and one female) recorded all pseudowords and their foils, with the clearest exemplar of each pseudoword being used in the phoneme monitoring task. The same speakers recorded 18 unique tokens of each pseudoword for the phoneme monitoring task in Experiment 2 (*within-talker variability*), varying both speech rate and intonation. An additional token of each pseudoword was recorded by each speaker for the old/new categorization task, with average speech rate and intonation. Finally, for Experiment 3 (*between-talker variability)* nine of the tokens from each speaker used in Experiment 2 were selected, and two additional speakers, one male and one female, recorded nine tokens of each pseudoword with varied speech rate and intonation. Old/new categorization tokens were the same as those used in Experiment 2. All stimuli were recorded in a sound attenuated booth using a Tascam DR-100 and Sennheiser ME40 microphone. The stimuli were digitized at 44.1Hz sampling rate with 16-bit analogue-to-digital conversion. Peak amplitude was normalized using Adobe Audition.

1.3. Design and Procedure

During the *study phase* of the experiment participants were exposed to one list of 24 pseudowords, counterbalanced across participants, in a phoneme monitoring task in which participants listened for the presence or absence of specified phonemes. In Experiments 1 and 2, 12 of the novel words were spoken consistently by the male talker, and 12 consistently by the female talker, with the talker of each novel word during study counterbalanced across participants. In Experiment 1 participants heard a single token of each word repeated 18 times, whereas Experiment 2 included 18 different tokens of each novel word, all spoken by the same talker. In Experiment 3 each item was heard in two voices at study, one male and one female, with 9 tokens of each pseudoword spoken by each of the talkers. The male talker from Experiments 1 and 2 was paired with a new female speaker, and the female talker from Experiments 1 and 2 was paired with a new male speaker.

To examine TSEs for the pseudowords half of the items changed talker between study and test in all three experiments, whilst half of the items remained in the same voice. In Experiment 3 the two new talkers that had been added to the study phase of the experiment in order to generate between-talker variability were not heard during test. Thus, only two talkers were used during the test phase of all three experiments, allowing the same materials and counterbalancing to be used for all experiments.

During the *test phase* of all experiments, participants first completed an old/new categorization task in which the 24 pseudowords from their studied list, as well as the 24 corresponding foil items were heard. Participants were instructed to classify studied items as "old", and unstudied/foil items as "new". Participants were then asked to rate their confidence in this old/new categorization using a scale from 1-7, where "1" corresponded to "definitely new" and "7" corresponded to "definitely old". This decision was included in order to examine the roles of recollection and familiarity in memory for the pseudowords by plotting the confidence ratings as receiver operating characteristic (ROC) curves [10]. Finally, in order to determine whether participants could explicitly access talker information, participants were asked to indicate whether the studied items had been heard in a male or a female voice during the study task. Each item was presented only once, at the beginning of each trial, with the three questions occurring in the same order on every trial. Response times were measured from word onset for the old/new categorization task, and from the onset of onscreen prompts for the confidence rating and male/female categorization tasks. Note that the male/female task was not included in Experiment 3 since all pseudowords were encountered in both a male voice and a female voice during study.

Participants completed the test phase of the experiment at three separate time points, immediately after study (Day 1), one day later (Day 2), and one week later (Day 8), in order to explore whether TSEs for pseudowords changed over time. Participants were tested individually in a sound-attenuated room. Tasks were run using DMDX experimental software [11], with stimuli presented binaurally at a comfortable listening level.

3. Results

1.4. Study Phase

Mean error rate in the phoneme monitoring task was 5.3% (*SD* = 2.8), 5.1% (*SD* = 2.3%), and 5.4% (*SD* = 2.1%) in Experiments 1, 2, and 3 respectively, indicating that participants paid close attention to the pseudowords during the study-phase of the experiment.

1.5. Test Phase

All data were analyzed using repeated-measures ANOVAs with variability (none, within-talker, between-talker), test-talker (same *vs.* different) and day (1, 2, 8) included as within-subjects factors, and list (1 *vs.* 2) included as a between-subjects factor in order to reduce the estimate of random variation [12]. For all tasks participants and items with error scores more than 2.5*SD* above the grand mean were removed prior to analysis. Means and standard errors for same- and different-talker items in each test session are reported in Table 1. F- and p-values are reported only for significant or marginally significant main effects and interactions.

1.5.1. Old/New Categorization

Mean accuracy rate was 83.1% (*SD* = 6.0%), 78.5% (*SD* = 5.9%) and 82.3% (*SD* = 6.7%) for Experiments 1, 2, and 3 respectively. The data were analysed using signal detection theory [13], with the number of hits and misses for same- and different-talker items being calculated separately for each session.

Analysis of the d' data revealed a non-significant main effect of variability. There were however significant main effects of test-talker, $F(1,103) = 48.09$, $p < .001$, and day, $F(2, 206) = 4.46$, $p < .05$, and the interaction between test-talker and day approached significance, $F(2,206) = 2.68$, $p = .071$. Most importantly, there was a significant interaction between variability and test-talker, $F(2,103) = 7.66$, $p < .001$. Pairwise comparisons between the three experiments revealed that the interaction was significant when comparing no-variability to both within-talker variability, $F(1,75) = 8.80$, $p < .01$, and between-talker variability, $F(1,74) = 10.99$, $p = .001$, but not when comparing within- and between-talker variability,

suggesting that introducing any type of variability significantly reduced the size of the TSEs during recognition of the pseudowords (Figure 1). Variability did not interact with day, suggesting that the effects of variability on TSEs were stable across the course of a week.

Separate analysis of each experiment revealed significant main effects of test-talker in Experiments 1, $F(1,46) = 66.63$, $p < .001$, and 2, $F(1,29) = 10.62$, $p < .01$, but only a marginally significant main effect of test-talker in Experiment 3, $F(1,28) = 3.10$, $p = .089$, suggesting that TSEs were significant only when each item was heard in a single voice during study. However, further analysis of Experiment 3 revealed a significant main effect of test-talker on Day 1, $F(1,29) = 8.09$, $p < .01$, but not on Days 2 and 8, consistent with data from existing words where TSEs were present immediately after study, but not following a delay [7]. For day, the main effect was significant only in Experiments 2, $F(2,58) = 3.74$, $p < .05$, and 3, $F(2,56) = 3.76$, $p < .05$, suggesting that the introduction of variability resulted in performance decreasing over the course of a week in the old/new categorisation task. Finally, whilst the interaction between test-talker and day was not significant in any of the experiments individually, the marginal interaction in the combined analysis suggests that TSEs in the old/new categorisation task decreased slightly over the course of a week.

Table 1: *Mean scores (and standard error values) for same- and different-talker items in each test session*

	Exp	Day 1		Day 2		Day 8	
		Same	*Diff*	*Same*	*Diff*	*Same*	*Diff*
Old/ new d'	1	2.55 (.07)	1.73 (.07)	2.30 (.08)	1.78 (.06)	2.41 (.07)	1.75 (.07)
	2	2.36 (.09)	2.05 (.10)	2.31 (.09)	2.06 (.09)	2.19 (.11)	1.90 (.10)
	3	2.32 (.12)	1.95 (.11)	2.08 (.09)	2.02 (.10)	1.96 (.13)	1.80 (.13)
AUC values	1	.97 (.01)	.89 (.01)	.95 (.01)	.88 (.01)	.95 (.01)	.88 (.01)
	2	.94 (.01)	.92 (.01)	.94 (.01)	.91 (.01)	.92 (.01)	.89 (.01)
	3	.93 (.03)	.92 (.03)	.93 (.02)	.91 (.03)	.90 (.04)	.90 (.04)
Male/ female error (%)	1	16.3 (2.4)	42.9 (3.5)	18.8 (2.1)	47.9 (2.6)	22.2 (2.4)	51.4 (3.0)
	2	15.5 (2.4)	42.8 (3.4)	16.6 (2.8)	48.7 (3.0)	23.7 (3.2)	46.0 (3.8)

1.5.2. Confidence Ratings

Data from the confidence ratings were plotted as ROC curves, with the false positive rate plotted against the true positive rate as a function of confidence [10]. Since recollection typically leads to higher confidence ratings, the greater the contribution of recollection to recognition memory, the closer the curve is to the upper left hand corner of the figure. Area under the curve (AUC) provides a quantitative measure of how close the curve is to the top-left hand corner, with larger AUCs indicating a larger contribution of recollection to recognition memory, reflecting greater reliance on highly-detailed episodic traces. AUC, calculated for same- and different-talker conditions for each session, was used as the dependent variable in all analyses.

For the AUC data the main effect of variability was not significant. There were however significant main effects of both test-talker, $F(1,101) = 41.03$, $p < .001$, and day, $F(2,202) = 9.68$, $p < .001$. Notably, the only significant interaction was between test-talker and variability, $F(2,101) = 8.97$, $p < .001$, with this interaction remaining significant only when comparing no-variability to either within-talker variability, $F(1,74) = 13.03$, $p = .001$, or between-talker variability, $F(1,74) = 13.42$, $p < .001$, suggesting, as in the d' data, that introducing any kind of talker variability during study significantly decreased the size of the TSEs observed in the ROC curves.

Separate analyses for each experiment revealed that the main effect of test-talker was significant only in Experiments 1, $F(1,45) = 82.53$, $p < .001$, and 2, $F(1,29) = 5.34$, $p < .05$, with participants being more confident about responses to same-talker items, but only when each item was heard in a single voice at study. The main effect of day was significant or marginally significant in all experiments (Exp 1 - $F(2,90) = 2.89$, $p = .06$; Exp 2 - $F(2,58) = 3.74$, $p < .05$; Exp 3 - $F(2,54) = 4.32$, $p < .05$) indicating that recollection processes became less involved when making old/new categorisation responses as the week progressed regardless of the amount of variability during study. The lack of interaction between test-talker and day in each experiment (as well as in the combined analysis) suggests that the decrease in reliance on recollection processes was equivalent for same- and different-talker items.

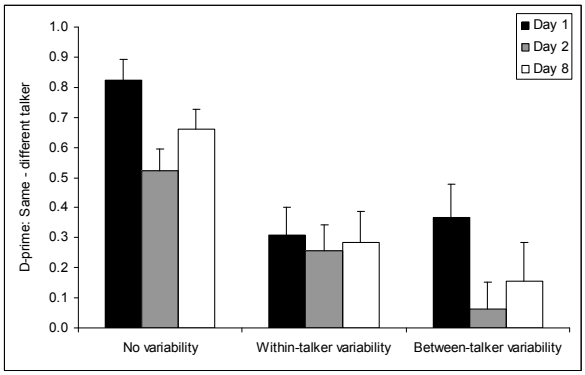

Figure 1: Differences between d' scores for same- and different-talker items in old/new categorization. Error bars indicate standard error of the mean.

1.5.3. Male/Female Categorisation

Data from this task provided a measure of whether talker information for newly-encountered pseudowords could be explicitly recalled. Only data from studied items were included in the analysis since foil words were encountered for the first time in the categorization task on Day 1. Since participants were instructed to focus on accuracy when making male/female categorizations response time measures were not analyzed. Mean accuracy was 66.4% (*SD* = 12.4%) in Experiment 1, and 69.1% (*SD* = 9.6%) in Experiment 2. Although low, these accuracy scores were significantly above chance in both experiments (Exp 1 – $t(49) = -9.29$, $p < .001$; Exp 2 – $t(31) = 11.34$, $p < .001$).

Cross-experiment analyses revealed no significant main effect or interactions involving variability (none *vs.* within-talker), suggesting that recall of talker identity associated with newly-encountered pseudowords was not affected by the presence or absence of within-talker variability during the study-phase of the experiment. In both experiments there was a significant main effect of test-phase talker (Exp 1 – $F_1(1,46) = 75.03$, $p < .001$, $F_2(1,46) = 181.04$, $p < .001$; Exp 2 – $F_1(1,30) = 34.27$, $p < .001$, $F_2(1,44) = 139.03$, $p < .001$), with more errors being made to different-talker items. There was also a main effect of day in Experiment 1, $F_1(2,92) = 11.53$, $p < .001$, $F_2(2,92) = 9.36$, $p < .001$, and a marginal main effect of day in Experiment 2, $F_1(2,60) = 2.90$, $p = .063$, $F_2(2,88) = 5.23$, $p < .01$, revealing that ability to recall information about the study talker decreased over the course of a week regardless of whether there was within-talker variability in the input during study or not. There was however no significant interaction between test-phase talker and day, suggesting that memory declined equally for same- and different-talker items over the course of a week.

4. Discussion

The aim of this paper was to investigate how newly-encountered pseudowords are represented in long-term memory, particularly whether there is a change in how much detail is retained in lexical representations over time.

The presence of TSEs immediately after study in the d' data from all three experiments and AUC data from Experiments 1 and 2 support studies using existing words [e.g., 2, 7] suggesting that recognition of recently studied lexical items relies initially on highly-detailed episodic representations. In addition, the continued presence of TSEs in both the d' and AUC data from Experiments 1 and 2 suggest that episodic representations continue to be involved in recognition of newly-learned words for a considerable period of time after the items are initially encountered, contradicting findings from Goldinger's [7] study showing that TSEs for existing words were not observed one week after the items were initially studied. Interestingly, the absence of TSEs in Experiment 3 on Days 2 and 8 suggests that introducing between-talker variability during study may result in multiple episodic traces of a single lexical item in long-term memory, with talker information becoming less useful as all of those traces are activated during spoken word recognition, consistent with episodic models of lexical representation. Nevertheless, it could be argued that the absence of TSEs in Experiment 3 arose due to the fact that participants heard only nine tokens of each item per talker, whereas they heard 18 tokens per talker in Experiments 1 and 2, potentially resulting in weaker memory for talker information in Experiment 3. The significant difference between TSEs for Experiments 1 and 2 in the d' and AUC data argues against this, instead suggesting that variability during study affects retention of speaker information even when the number of tokens per talker is matched. This finding suggests that effects typically described as TSEs may arise, at least in part, due to matches in variables such as speech rate and intonation, as well as being due to repetition of the word by the same talker.

Despite the evidence described above which seems to favour an episodic model of long-term lexical memory, our experiments also provide evidence that recognition of newly-encountered pseudowords becomes gradually more reliant on abstract representations as time progresses. Firstly, overall decreases in confidence ratings in all three experiments suggest that participants rely less on recollection processes and more on familiarity processes as the week continued. Interestingly, evidence suggests that whilst recollection relies primarily on activation within the hippocampus, familiarity processes are more dependent on activity within surrounding medial temporal lobe regions [14], consistent with a complementary learning systems account [5] of our findings. Secondly, data from the male/female categorization task in Experiments 1 and 2 indicates that talker information becomes less explicitly accessible in later test sessions. Notably, these effects occur despite the fact that, in Experiment 1 at least, accuracy in the old/new categorisation task does not decrease significantly across the course of a week suggesting that this is not merely a result of forgetting, but rather an effect of abstraction.

Thus, it seems that the data are most consistent with a hybrid model of lexical representation in which both episodic and abstract representations are important. Nevertheless, the fact that TSEs do not decrease significantly across the course of a week in Experiments 1 and 2 in either d' or AUC analyses suggests that the change in reliance between different types of representation is a very gradual process, consistent with data from Takashima et al. [15] who showed activation of hippocampal regions during recognition of novel images even three months after initially studying the items.

Together these data highlight the importance of examining the time-course of TSEs for items that do not have pre-established lexical representations in order differentiate between the predictions of hybrid and episodic models. Moreover, it is clearly important to use a number of different measures of TSEs in order to fully understand the nature of long-term lexical representation and the qualitative changes that occur during the formation of new lexical entries. Our data suggest that initially memory relies on highly-detailed episodic traces, but that over time there is some evidence of abstraction, suggesting that representations in long-term lexical memory are likely to be abstract in nature.

5. Acknowledgements

This research was funded by an Economic and Social Research Council grant (RES-063-27-0061).

6. References

[1] Lachs, L., McMichael, K. and Pisoni, D.B., "Speech perception and implicit memory: Evidence for detailed episodic encoding in phonetic events", in J.S. Bowers and C.J. Marsolek (Eds.), *Rethinking Implicit Memory*, 215-235, Oxford, 2003.

[2] Bradlow, A.R., Nygaard, L.C. and Pisoni, D.B., "Effects of talker, rate, and amplitude variation on recognition memory for spoken words", *Percept Psychophys*, 61(2):206-219, 1999.

[3] Luce, P.A., McLennan, C.T. and Charles-Luce, J., "Abstractness and specificity in spoken-word recognition: Indexical and allophonic variability in long-term repetition priming", in J.S. Bowers and C.J. Marsolek (Eds.), *Rethinking Implicit Memory*, 215-235, Oxford, 2003.

[4] Goldinger, S.D., "A complementary-systems approach to abstract and episodic speech perception", *ICPhS* XVI, 49-54, 2007.

[5] McClelland, J., McNaughton, B. and O'Reilly, R., "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory", *Psych. Review*, 102:419-437, 1995.

[6] Bakker, A., Kirwan, C.B., Miller, M. and Stark, C.E.L., "Pattern separation in the human hippocampal CA3 and dentate gyrus", *Science*, 319: 1640-1642, 2008.

[7] Goldinger, S.D., "Words and voices: Episodic traces in spoken word identification and recognition memory", *JEP:LMC*, 22(5):1166-1183, 1996.

[8] Tamminen, J. and Gaskell, M.G., "Newly learned spoken words show long-term lexical competition effects", *QJEP*, 61(3):361-371, 2008.

[9] Baayen, R.H., Piepenbrock, R. and van Rijn, H., "The CELEX lexical database [CD-ROM]", Philadelphia: Linguistic Data Consortium, 1993.

[10] Yonelinas, A.P., "The nature of recollection and familiarity: A review of 30 years of research", *JML*, 46(3):441-517, 2002.

[11] Forster, J.C. and Forster, K.I., "A Windows display program with millisecond accuracy", *Behavioural Research Methods, Instruments & Computers*, 35:116-124, 2003.

[12] Pollatsek, A. and Well, A.D., "On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis", *JEP:LMC*, 21(3): 785-794, 1995.

[13] Green, D.M. and Swets, J.A., "*Signal detection theory and psychophysics*", New York: Wiley, 1966.

[14] Yonelinas, A.P., Otten, L.J., Shaw, K.N. and Rugg, M.S., "Separating the brain regions involved in recollection and familiarity in recognition memory", *J Neurosci*, 25(11):3002-3008, 2005.

[15] Takashima A., Petersson K.M., Rutters F., Tendolkar I., Jensen O., Zwarts M.J., McNaughton B.L. and Fernandez G., "Declarative memory consolidation in humans: a prospective functional magnetic resonance imaging study", *PNAS USA*, 103:756 –761, 2006.