Detailed phonetic memory for multi-word and part-word sequences

TRAVIS WADE and BERND MÖBIUS

Universität Stuttgart

Abstract

Humans recognize previously heard spoken words better when repetitions of the words involve identical productions than productions by a different speaker. Such findings have been taken as evidence that perceived instances of words or sub-lexical units are stored in a detailed form in memory, and that collections of these memory traces comprise or are linked to mental lexical representations. This study tested a different possibility, that detailed acoustic memory occurs during spoken language processing but does not necessarily correspond to words or other traditionally defined units. Two experiments examined lexical access and recognition memory for continuous speech sequences, extracted from a spoken language corpus, as a function of sequence length and onset phase (with respect to word onset), and speaker. Qualitatively different patterns between word identification and memory performance based on these three variables provide little evidence for a role of the word level of representation in memory for the sequences, and suggest that memory-based processing may more independent of this level than has been assumed.

1. Introduction

When asked to remember a list of spoken words, humans recognize stimuli as previously heard more accurately when the words are repeated in the same voice and at the same speaking rate than when they differ in these or other dimensions (Palmeri et al. 1993, Goldinger 1996, Bradlow et al. 1999). This seems to indicate that memory for word tokens includes not only an abstract reference to encountered lexical items but also at least some of the acoustic detail that is traditionally assumed to be stripped away during the recognition process. The finding has often been taken further, as evidence for episodic lexical models (Goldinger 1997, 1998), which assert that knowledge of word categories is effectively *comprised* of detailed memories of multiple occurrences of the categories, and it remains

JLP 1 (2010), 283–294
DOI 10.1515/LABPHON.2010.014

1836-6346/10/01–283

central to the behavioral empirical evidence for exemplar and usage-based approaches to linguistic knowledge in general (see, e.g., Dahan and Magnuson 2006).

Of course, despite common assumption that *words* are objects of perception during spoken language processing (cf, e.g., Goldinger and Azuma 2003), lexical items are seldom heard as isolated productions, but rather as part of continuous utterances which provide a context that variously helps to specify their spectral and temporal properties, probability of occurrence, and meaning, and from which they are in general not straightforwardly extracted acoustically. If it is assumed that word tokens are stored in exemplar fashion, questions arise as to whether and how these effects of context, and the temporal nature of speech in general, might be represented. In fact, there is evidence from phonological analysis (Bybee 2002a), a corpus study of pronunciation (Binnenpoorte et al. 2005), and word-monitoring experiments (Sosa and MacFarlane 2002) that at least frequent multi-word expressions are accessed from memory as units in the same way that words are assumed to be. On the other hand, there is some indication that any detailed storage of the type measured in recognition memory experiments involves *sub-lexical* representations (Jesse et al. 2007) instead of words.

More generally, inference over multiple, detailed storage of categories is often invoked to account for patterns observed above (Bod 1998) and below (Johnson 1997, 2006, Pierrehumbert 2001) the word level, with a gradually increasing emphasis on how the apparent processes at these different levels (or time scales) might work together in comprising linguistic understanding (e.g., Hay and Bresnan 2006, Goldinger and Azuma 2003). In this context, it seems appropriate to revisit previously observed speaker effects in memory for isolated word productions in the context of *connected* speech, with the aim of evaluating the role of the word in driving these effects. This study measured recognition memory for continuous speech sequences extracted from a multi-talker spoken language corpus, as a function of the length (number of words) and onset phase (with respect to word onset) of the sequences, and the speaker. It was designed partly as a first exploratory study of memory for continuous speech as it relates to these variables, but was also intended to provide information as to the role of lexical access in memory for spoken language. In one experiment, listeners transcribed words from sequences extracted at random from the corpus, to estimate the probability of lexical access as a function of sequence length, phase, and speaker. In a second experiment, recognition memory for these same sequences was measured and compared with word recognition accuracy data. It was predicted that, if storage and/or retrieval for the sequences critically involves the word level of representation, fine-grained effects of sequence length, phase, and speaker on word recognition performance should be reflected in memory for the same sequences. On the other hand, indication that these variables affected recognition and memory more independently would suggest that detailed phonetic memory for the sequences is not driven by lexical access.

2. Experiment 1

Experiment 1 was designed to measure listeners' ability to recognize words from sequences extracted from continuous productions.

2.1. *Method*

2.1.1. *Subjects*

Subjects were 19 adult native speakers of German with no reported speech or hearing deficit. Subjects were paid for their participation.

2.1.2. *Stimuli*

Stimuli were continuous sequences extracted from the Kiel corpus of read speech, a multitalker database of Standard German (Kohler 1996). Sequences varied in length from 0.5–3 words, in 0.5 word steps, and in onset phase with respect to word onset, beginning either at the beginning of a word or halfway (in time) between word onset and offsets. Sequences were selected at random from the corpus, without considering word or sequence frequency or probability or syntactic constituent boundaries (although, since the corpus consists almost entirely of sentence-level productions, sequences that spanned sentence boundaries were extremely rare). Sixteen sequences at each length and phase condition were selected, for a total of 192 unique sequences, selected separately for each listener to maximize coverage. Sequences were normalized to the same total RMS amplitude, and 25-ms linear on/off ramps were applied.

2.1.3. *Procedure*

The experiment was divided into two subtests, separated by a short break. In each subtest, the listener heard each of the 192 sequences over AKG K-501 headphones and was prompted to type the words that he or she recognized from the sequence. The order in which sequences appeared was randomized separately for each subtest. In the second subtest, half of the sequence productions at each length/phase condition were identical to those appearing in the first subtest, and half were productions of the same sequences (taken from the same sentence contexts) by a different speaker in the corpus. The experiment took about an hour; listeners were encouraged to take their time and to type carefully.

2.2. *Predictions and analysis*

Independent of the various effects of speakers and production patterns, syntactic structure, and word frequency, neighborhood density, confusability, etc., at least two factors were considered likely to influence overall recognition accuracy of

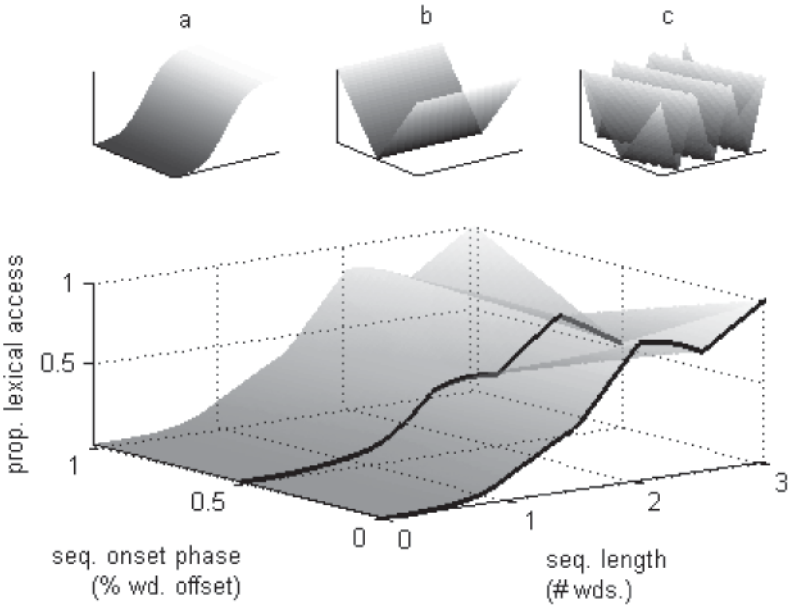

Figure 1. Top *(axes correspond to bottom): predictions of average sequence length (a), onset (b) and offset (c) phase effects on lexical access. Performance varies psychometrically with length, increases linearly with proximity of sequence boundaries to word boundaries.* Bottom*: trivial combination of these effects a* × (*b* + *c*) *demonstrating the fine-grained predictions listed in section 2*

words from connected speech contexts. First, due to the influence of both local coarticulatory and longer-range effects of acoustic context (Ladefoged and Broadbent 1957, Lindblom and Studdert-Kennedy 1967), lexical access is likely to be better in longer sequences, probably varying in a psychometric manner (i.e. equally bad for sequences shorter than some critical range and at ceiling level above it). Second, access is likely to be better if sequence onsets and/or offsets correspond to word onsets and offsets (e.g. Mattys et al. 2005).

These influences are schematized in Figure 1. The estimate of the effects' combined predictions in the bottom panel of the figure qualitatively demonstrates at least three predictions for the set of stimuli included in the present experiments: (1) accuracy should increase overall with sequence length, (2) zero-phase sequences (those with onsets corresponding to word onsets) should be better overall, and (3) a length × phase interaction should result from advantages of onset and offset agreement between sequences and spoken words. Two separate measures of lexical access were considered: *accuracy* (whether listeners' responses matched words listed in the corpus) and *consistency* (whether words in listeners' responses to a sequence matched across the two presentations of the sequence in the test).

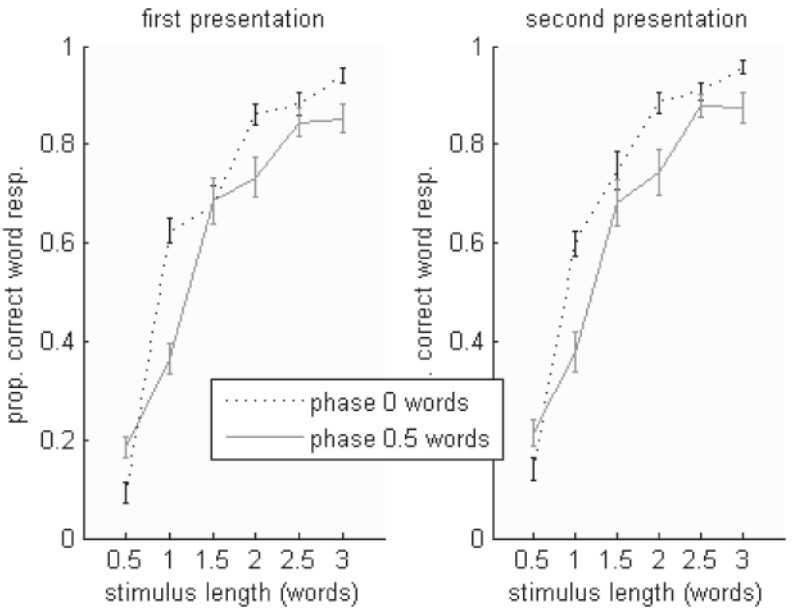

Figure 2. *Mean (and std. error) correct recognition responses to sequences in Experiment 1.*

2.3. *Results and discussion*

In evaluating listeners' responses, punctuation and capitalization were ignored, and predictable alternate letter sequences (for example, vowel + 'e' vs. umlauted vowel) were treated as equivalent; otherwise, misspellings or typing errors were treated as incorrect (when comparing a response to the word sequence listed in the corpus) or different (when comparing a response to another response by the same listener). In both experiments, "lexical access" was defined as at least one word being recognized (accurately or consistently) from a sequence.

Figure 2 shows recognition accuracy as a function of sequence length and phase. Listeners were slightly more accurate overall in the second subtest, resulting from practice effects and/or familiarity with the stimulus materials. The three predictions shown in Figure 1 are also all clearly visible. Accuracy increases from near chance level for the shortest segments to ceiling performance at approximately 2.5 words. Zero phase sequences were responded to more accurately overall ($t(18) = 4.54$, $p < 0.001$), and a Length × Phase interaction ($F(5,18) = 14.1$; $p < 0.001$) indicates that performance was better when word and sequence boundaries coincided. This last effect is more clearly demonstrated in the left panel of Figure 3, which shows about a 10% advantage for both types of coincidence. (It should be noted that data for word onsets and offsets overlap substantially due to the limited set of length and phase settings used; however, sequence length and (at least for

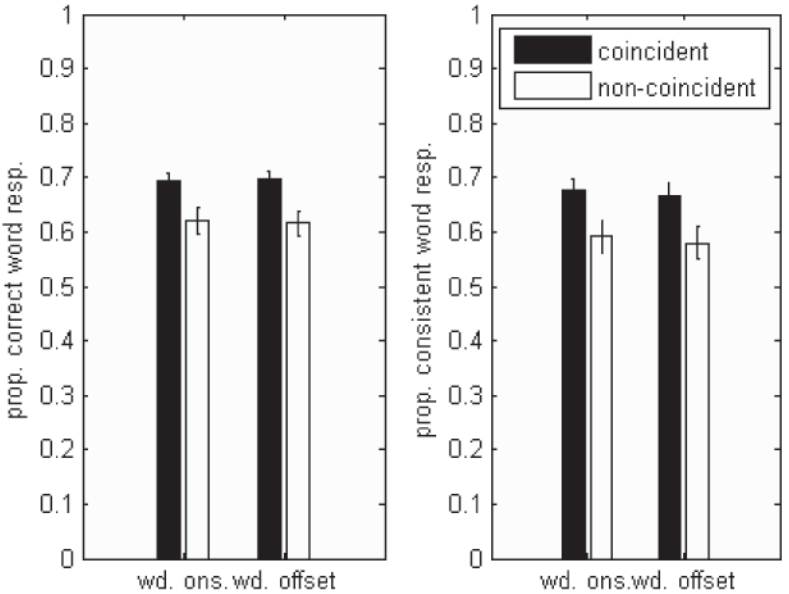

Figure 3. *Correctness and consistency of responses depending on onset and offset agreement between sequences and words.*

word onsets) phase were the same on average across conditions, so it seems reasonable to attribute the differences to word-sequence boundary agreement.)

Figure 4 shows response consistency results across subtests, as a function of sequence length and phase and whether the second presentation of a sequence involved the same speaker (and production) as the first. The three predictions based on sequence length and phase are also clearly present (word/sequence coincidence effects are also summarized in the right panel of Figure 3). Additionally, it can be seen that consistency was slightly greater when the two presentations of a sequence involved identical waveforms than when they involved different speakers' productions. Perhaps interestingly, the magnitude of this same-speaker advantage decreased linearly over the range of sequence lengths considered ($r = -0.96$; $p = 0.0025$), from about 10% for the shortest sequences to (not surprisingly) near zero where overall performance was at ceiling level.

3. Experiment 2

3.1. *Methods*

Subjects were 16 adult native German speakers not reporting hearing or language impairment. Stimuli and procedures were identical to those of Experiment 1, ex-

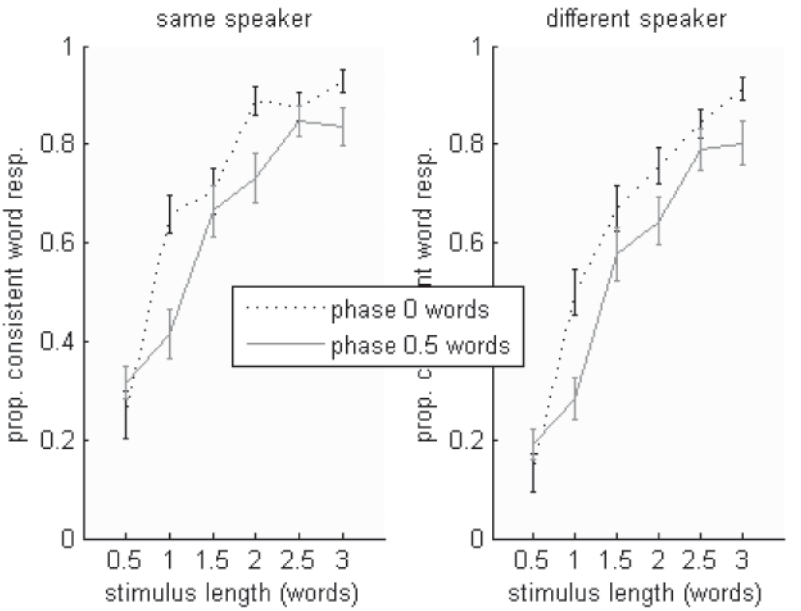

Figure 4. *Consistent responses across Experiment 1 subtests.*

cept that the recognition task in the second subtest was replaced by a recognition memory task. That is, after hearing a sequence listeners were prompted to respond whether or not they had heard the same sequence (words were not explicitly referenced in the instructions) in the first subtest.

3.2. *Predictions*

As mentioned above, this experiment was designed in part to test between two (not exhaustive and not completely mutually exclusive) accounts of the apparently detailed memory for perceived speech. According to one account, words are recognized and explicitly stored as discrete units. Alternatively, we might simply remember the acoustic sequences that we hear, independent of (or absent) any analysis of these words into traditionally described units such as words. These two possibilities suggest the following partly overlapping sets of predictions concerning memory accuracy for connected speech sequences: (1) in a word-based model, memory should follow lexical access. Specifically, the three predictions relating sequence length and phase to word recognition should all be seen in memory for the same sequences. In addition, memory should be limited by lexical access; where no words are recognized, memory should be at chance level. Finally, the same-speaker advantage observed in Experiment 1, including its inverse correlation with sequence length, should be maintained, since in a word-based exemplar

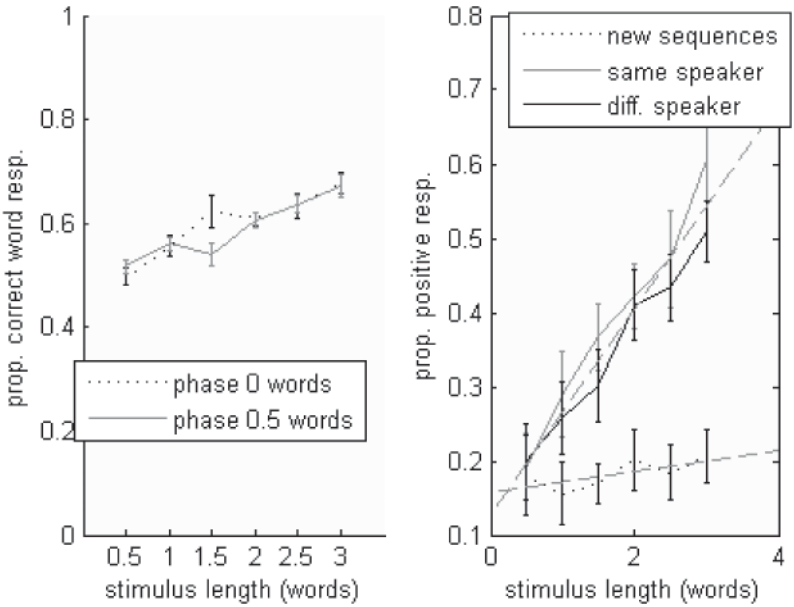

Figure 5. *Mean (and std. error) correct recognition responses (left), positive responses (right) in Experiment 2. Dashed lines represent linear regression for positive responses to old and new items.*

model lexical access and recognition memory are assumed to be essentially the same process. (2) According to a more flexible continuous sequence model, memory performance should also increase with sequence length, since more acoustic material should lead to richer, more unique/distinct representations. However, other than this overall increase, the fine-grained effects of sequence length and word phase shown in Figure 1 (and Figures 2 and 4) should not appear. Memory should not be strictly limited by lexical access, although, again, it should be poor for very short sequences. Finally, no specific link is predicted between the same-speaker advantage seen in Figure 3 and any same-speaker memory advantage, since sequence memory and lexical access are different processes.

3.3. *Results and discussion*

The left panel of Figure 5 shows recognition accuracy as a function of sequence length and phase. Overall, accuracy seems to increase linearly over the range of sequence lengths considered, and there is no apparent effect or interaction involving onset phase. The lack of fine-grained length and phase effects relating to lexical access can be seen more clearly in Figure 6, which compares the effects of coincidence between word and sequence onsets and offsets on recognition accuracy during the first subtest (left) and on memory performance (right). As in Experiment 1,

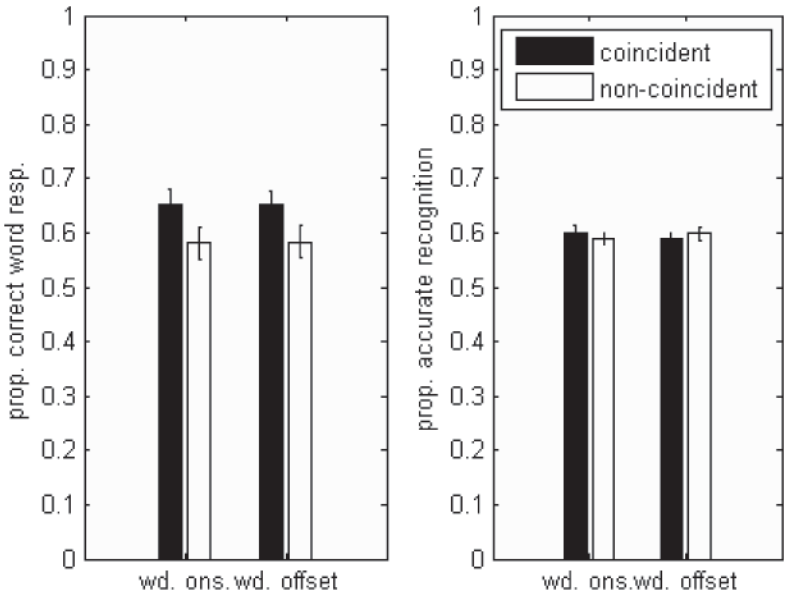

Figure 6. *Recognition (left) and memory (right) correctness of responses depending on onset and offset agreement between sequences and words in Experiment 2.*

there is a clear effect (again about 10%) of coincidence on lexical access, but this effect is conspicuously lacking in memory. A 2 (task/subtest) × 2 (coincidence status) mixed-model ANOVA revealed a significant interaction ($F(1,15) = 4.7$; $p = 0.035$), indicating that fine-grained effects of "lexical plausibility" did not affect memory in the same way that they did the initial parsing of the sequences.

The absolute range of accuracy values seen in Figure 5 is also interesting. The right panel of the figure breaks these data down further into previously heard (by the same or a new speaker) and previously unheard sequences. Positive responses for previously heard sequences are more frequent than false positive responses to new stimuli for all sequence lengths, and (incidentally, since the study was not powered to make comparisons within length conditions) individual pairwise comparisons are significant at all but the shortest (0.5 word) sequences. In fact, comparison of linear regression functions between new and old sequences would suggest that accuracy would remain above chance for sequences greater than about 0.25 words in length. Considering the lexical access estimates shown in Figures 2 and 4, it is very unlikely that any words would be recognized from such short sequences. This is also inconsistent with a strictly word-based model, since it suggests that sequence memory is not limited by lexical access.

A more direct way of estimating whether word recognition limits memory is to compare memory for sequences that were correctly recognized and those that were not. Indeed, repeated sequences were robustly more likely to be recognized as

previously heard if at least one word was identified correctly during the first subtest. However, this was probably due to word identification and memory being similarly limited by auditory factors for the most difficult stimuli; these items tended to be the shortest in absolute (time) length, both overall and when considered within word length conditions. In order to determine whether, on average, sequences with completely incorrect word identification responses were remembered at better than chance accuracy, a one-sample t-test compared positive responses ("previously heard") to these sequences in the second subtest with the false positive response rate (to previously unheard sequences, weighted in length and phase conditions to match those of the incorrectly identified old sequences). A significant difference was observed ($t(15) = 2.38$; $p = 0.031$), indicating that listeners remembered even those sequences from which they did not recognize any words.

Finally, the right panel of Figure 5 shows that sequences presented by the same speaker were remembered better than productions from a new speaker ($t(15) = 2.61$, $p = 0.02$ overall). Contrary to the recognition results of Experiment 1, the size of this effect tended to increase with sequence length ($r = 0.7$; $t = 0.12$). In fact, the size of the same-speaker advantage in memory tended to correlate inversely ($r = -0.74$; $t = 0.092$) over sequence lengths with that of the same effect in word recognition. Thus, again consistent with the more flexible connected speech sequence account, the same-speaker effect in memory seems to operate independently of that for word recognition.

4. General discussion and conclusions

In summary, two experiments measured word identification and recognition memory for connected speech sequences extracted at random from a spoken language corpus. As predicted, identification was aided by increasing sequence length and by agreement of sequence onsets and offsets with word boundaries, and was more consistent across repeated presentations of the same production of a sequence than across productions by different talkers. Comparison of these data with recognition memory results offered little evidence that memory is related to parsing at a lexical level. Memory was generally better for longer sequences but lacked any of the fine-grained effects of word-sequence boundary coincidence seen in identification. Memory did not seem to be limited by word identification accuracy in general, and same-speaker advantages in the two tasks correlated inversely across sequence lengths.

These results are consistent with a model of perception and memory in which detailed episodic storage occurs for acoustic sequences that are of variable size and composition, potentially corresponding to multiple words or phrases and not necessarily coinciding with words as discrete units. We have suggested previously (Wade 2007, Wade and Möbius 2007, Wade et al. 2010) that memory for speech

generally involves sequences that are longer (perhaps corresponding to entire utterances) than potential units of analysis (such as segments or words), so that units at any level are stored as part of a larger temporal context, and in fact are never considered apart from this context. During perception, newly encountered patterns are compared – along with their surrounding contexts – with sequences of similar length in memory, so that correlations between different types of spectral information in adjacent regions (like phonetic context effects related to coarticulation), between different types of spectral information at the same location (like speaker or gender effects), and between spectral information and temporal organization (like speaking rate effects) are all implicitly "normalized for" in the recognition process. Similarly, during production in such a model, exemplars may be selected with probability proportional to the similarity of their original context with the relevant neighboring sounds in the current production, resulting in context-appropriate selection, potentially including acoustic patterns at lower (e.g. syllable, segment) levels specific to frequent sequences such as words, phrases or collocations (Bybee 2002a,b, 2006, Binnenpoorte et al. 2005).

Obviously, the experiments described here do not comprehensively evaluate either the role of the word in driving processing or memory for spoken language or the influences of variables such as sequence length, phase, and speaker on parsing, storage, or retrieval. At the very least, it will be necessary to consider a larger range of stimulus lengths that is more finely graded in both length and phase. It will be necessary to determine whether the psychometric and linear trends in Figures 2, 4 and 5 continue for shorter or longer sequences, and how consistent the coincidence advantages observed are with those predicted in Figure 1 more generally. Consideration of other variables such as the probability and acoustic/phonetic content of sequences and their status related to (assumed) constituents at other time scales or linguistic levels will be informative as well.

Correspondence e-mail address: Travis.wade@positscience.com

References

Binnenpoorte, D., Cucchiarini, C., Boves, L., & Strik, H. (2005). Multi-word expressions in spoken language: An exploratory study on pronunciation variation. *Computer Speech and Language*, 19:433–449.

Bod, R. (1998). *Beyond Grammar: An Exeperience-Based Theory of Language*. CSLI Publications/ Cambridge University Press.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics*, 61:206–219.

Bybee, J. (2002a). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition*, 24:215–221.

Bybee, J. (2002b). Word frequency and context of use in the lexical diffusion of phonetically-conditioned sound change. *Language Variation and Change*, 14:261–290.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82:529–551.

Dahan, D. & Magnuson, J. S. (2006). Spoken word recognition. In Traxler, M. J. & Gernsbacher, M. A., eds., *Handbook of Psycholinguistics (Second Edition)*, 137–157. Benjamins, Amsterdam.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *J. Exp. Psych.*, 22:1166–1183.

Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In Johnson, K. & Mullenix, K., eds., *Talker Variability in Speech Processing*, 33–66. Academic Press, San Diego.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psych. Rev.*, 105:251–279.

Goldinger, S. D. & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, 31:305–320.

Hay, J. & Bresnan, J. (2006). Spoken syntax: The phonetics of giving a hand in New Zealand English. *The Linguistic Review*, 23:321–349.

Jesse, A., McQueen, J. M., & Page, M. (2007). The locus of talker-specific effects in spoken-word recognition. In *Proceedings of the 16th International Conference of Phonetic Sciences*, 1921–1924, Saarbrücken.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson, K. & Mullenix, K., eds., *Talker Variability in Speech Processing*, 145–165. Academic Press, San Diego.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *J. Phonetics*, 34:485–499.

Kohler, K. J. (1996). Labelled data bank of spoken standard german; the kiel corpus of read/spontaneous speech. In *Proceedings of ICSLP 1996*, 1938–1941, Philadelphia.

Ladefoged, P. & Broadbent, D. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.*, 29:98–104.

Lindblom, B. E. F. & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am.*, 42:830–843.

Mattys, S. L., White, L., & Melhorn, J. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General*, 134:477–500.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19:309–328.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J. & Hopper, P., eds., *Frequency and the Emergence of Linguistic Structure*, 137–157. Benjamins, Amsterdam.

Sosa, A. V. & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language*, 83:227–236.

Wade, T. (2007). Implicit rate and speaker normalization in a context-rich phonetic exemplar model. In *Proceedings of the 16th International Conference of Phonetic Sciences*, pages 765–768, Saarbrücken.

Wade, T., Dogil, G., Schütze, H., Walsh, M., & Möbius, B. (2010). Syllable frequency effects in a context-sensitive segment production model. *Journal of Phonetics*, 38:227–239.

Wade, T. & Möbius, B. (2007). Speaking rate effects in a landmark-based phonetic exemplar model. In *Proceedings of Interspeech 2007*, 402–405, Antwerp.