Detailed Phonetic Memory for Multi-Word and Part-Word Sequences

Travis Wade & Bernd Möbius

Institute for Natural Language Processing, University of Stuttgart, Germany;
{travis.wade, bernd.moebius}@ims.uni-stuttgart.de

Humans recognize previously heard spoken words better when the words are repeated in the same voice and at the same speaking rate than when they differ in these or other dimensions (e.g., Goldinger, 1996). Such findings have usually been taken as evidence that perceived instances of either words or sublexical units are stored in a detailed form in memory, and that collections of these episodic memory traces comprise or are somehow linked to mental lexical representations. A different possibility is that, while detailed information about the speech signal is remembered, this information is not explicitly segmented or organized into discrete, traditionally assumed units such as words. Instead, episodic memories may involve sequences of connected speech that vary in length depending on the nature of the listening situation, perhaps comprising entire phrases or utterances. Such a possibility is more in line with the temporal and context-dependent nature of speech perception and production than traditionally defined exemplar models, since it predicts that more local information is always stored and (potentially) considered as part of a larger context. Indeed, there is evidence from corpus studies of pronunciation (Binnenpoorte, Cucchiarini, Boves, & Strik, 2005), phonological analysis (Bybee, 2002 for a review), and word-monitoring experiments (Sosa and MacFarlane, 2002) that at least very frequent multi-word expressions are in some sense accessed as “units” in the same way that words are usually assumed to be. However, it is unknown whether variable-length stretches of connected speech in general exhibit properties of exemplar-based storage or perception, since previous studies on memory have considered only isolated word productions.

In this study, word identification and recognition memory accuracy were measured for sequences of words extracted from a large multi-speaker corpus of German sentence productions (Kohler, 1996). Sequences varied in their length (0.5-3 words, in 0.5-word steps), onset phase with respect to word onset (beginning either at a word onset or halfway between a word onset and offset), and speaker (target sequences were repeated with the same or a different voice). Sequences were extracted at random from the corpus, without considering word or sequence frequency or probability or syntactic constituent boundaries. In Experiment 1, listeners (n=19) heard 384 sequences (selected separately from the corpus for each listener to maximize coverage) in two blocks and identified the words they perceived. Blocks contained the same 192 word (or part-word) sequences extracted from the same sentence contexts; half of the words in the second block were identical repetitions of first-block stimuli and half were produced by a different speaker. Experiment 2 also involved two blocks. The first block was identical to that of Experiment 1; in the second block, listeners (n=16) categorized sequences as new or previously encountered. Half of repeated sequences (one fourth of the trials in the second block) were identical repetitions, and half involved a new speaker.

Word identification results (Experiment 1) were consistent with previous findings concerning the effects of local acoustic context on phonetic perception, and the influence of lexical information on speech segmentation (e.g., Mattys, White, and Melhorn, 2005). By several measures of accuracy and consistency, listeners were better at identifying words in longer sequences, approaching ceiling performance for sequences longer than about 2.5 words. Sequence onset phase was also important, with robustly better identification where stimulus onsets and/or offsets coincided with word onsets/offsets. Responses were more consistent across repetitions of a sequence by the same speaker than repetitions by different speakers, especially at shorter sequence lengths.

Recognition memory results (Experiment 2) differed from these patterns in several critical ways. Recognition accuracy increased linearly with sequence length over the range of stimuli considered, even after identification accuracy reached ceiling performance. There was no effect of sequence onset phase on memory, and there was a significant difference in the size of the benefit of stimulus and word onset/offset coincidence between identification and memory tasks (no benefit in the memory task), indicating that lexical

cues which aided segmentation of sequences did not affect the encoding of these same sequences in memory.

Both overall and within length and phase conditions, repeated sequences were more likely to be recognized as previously heard if at least one word was identified correctly during the first block than if no words were correctly identified. However, better-than-chance memory performance was seen even for sequences where no words were correctly identified. Moreover, if completely misidentified (no words correct) sequences were discarded, the likelihood of correct recognition of sequences of a given length was not related to the number of words that were originally correctly identified. This indicates that both segmentation/identification and memory were limited by auditory/perceptual factors for the most difficult stimuli; these stimuli tended to be the shortest in absolute (time) length, both overall and when considered within word length conditions. However, there was no direct evidence that words were remembered as discrete units, or that memory was directly mediated by segmentation.

Recognition memory was better for sequences repeated with the same voice than for sequences repeated with a different voice. This was true whether sequences that were completely misidentified in the first subtest were discarded or included in the analysis. The size of this same speaker recognition memory benefit tended (n.s.) to increase with sequence length, even at lengths where word identification was at or near ceiling performance. In fact, over the range of sequence lengths considered, the same speaker benefit in memory correlated inversely with the same speaker benefit in identification consistency.

In summary, a same-speaker recognition memory advantage was seen for sequences of connected speech ranging from 0.5 to 3 words in length. Memory was generally better for longer sequences, but neither overall performance nor the size of the same-speaker benefit seemed to be influenced by parsing at a lexical level, as measured by word identification. 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. It seems further possible that processes involving various levels of a linguistic hierarchy (both phonetic and lexical, for example) may reference, or result implicitly from the structure of, the same set of multi-word episodes at multiple time scales.

References

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

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

Goldinger, S.D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1166-1183.

Kohler, K.J. (1996). Labelled data bank of spoken standard German; the Kiel Corpus of Read/Spontaneous Speech. In *Proceedings of ICLSP 1996, Philadelphia*.

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.

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.