CATEGORICAL VS. EPISODIC MEMORY FOR PITCH ACCENTS IN ENGLISH

Amelia E. Kimball[1], Jennifer Cole[1], Gary Dell[1], Stefanie Shattuck-Hufnagel[2]

University of Illinois at Urbana Champaign[1], Massachusetts Institute of Technology[2]

akimbal2@illinois.edu, jscole@illinois.edu, gdell@illinois.edu,sshuf@mit.edu

ABSTRACT

Phonological accounts of speech perception postulate that listeners map variable instances of speech to categorical features and remember only those categories. Other research maintains that listeners perceive and remember subcategorical phonetic detail. Our study probes memory to investigate the reality of categorical encoding for prosody—when listeners hear a pitch accent, what do they remember? Two types of prosodic variation are tested: phonological variation (presence vs. absence of a pitch accent), and variation in phonetic cues to pitch accent (F0 peak, word duration). We report results from six experiments that test memory for pitch accent vs. cues. Our results suggest that listeners encode both categorical distinctions and phonetic detail in memory, but categorical distinctions are more reliably retrieved than cues in later tests of episodic memory. They also show that listeners may vary in the degree to which they remember prosodic detail.

Keywords: Prosody, memory, episodic memory, categorical perception, pitch accent.

1. INTRODUCTION

All speech sounds are variable in their acoustic manifestations. Speakers' accents, local phonological context, and the broader linguistic context can create many acoustically distinct instances of a speech sound that phonological theory would classify as "the same sound." The lack of invariance between the phonological unit and its acoustic realization creates questions for language scientists: How is the phonological category defined in relation to the many acoustically distinct instances of that sound unit? How do speakers perceive, record and recognize different instances of a phonological sound unit as "the same", i.e. belonging to the same category? Given a finite memory capacity and a lifetime of exposure to many varied sounds, what information about a perceived sound do listeners retain in memory, after initial processing and recognition of the sound?

There are two theoretical positions that address these questions. One, stemming from the tradition of formal phonology and generative linguistics, characterizes language in terms of a set of abstract representations. Rules (or constraints) operate over those representations, and together form a grammar that models variable and error-ful input and output. Under this theory, remembering a word involves encoding and storing its abstract representation. The second theoretical position invokes episodic (or 'exemplar') models for speech perception [5], and hold that all acoustic information can be represented in memory, including details that are not linguistically meaningful.

These two theories make separate predictions for what is stored in memory when a listener perceives speech. The formal phonological approach suggests that each phonetic instantiation of a sound unit is mapped to an abstract category. This analysis predicts that if you ask a listener to recall two acoustically distinct sounds that are mapped to the same phonological category, they will report the two sounds as being the same because they encoded only the category of each sound in their memory, rather than the phonetic realization. The second approach predicts that within-category acoustic differences will be remembered because memory for speech is episodic, and all perceived acoustic information is recorded in memory to some extent.

There is evidence to support both of these views. A classic example used to support phonological abstraction is the categorical perception of phonemes (e.g. [6]). Listeners hear two sounds that may or may not vary in voice onset time and must identify whether the sounds are the same or not. It turns out that they often fail to detect the differences in the sounds if they both belong to the same category. The tasks used to measure these distinctions are usually called perception tasks, but they are in a sense short-lag memory tasks, in that comparing two things involves holding the first in memory, however briefly. Because listeners do not report a difference when hearing two sounds that belong to the same category, it is argued that only the category is encoded in memory and retrievable after the fact.

Another example of this in the realm of prosody is "stress deafness." Speakers of languages with fixed lexical stress report that unfamiliar words pronounced with different stress patterns are the same [4] (even though in at least one study EEG readings have shown that the listeners do in fact hear the differences [3]). These results suggest that listeners do not remember acoustic distinctions that are not meaningful in their language, even when they can perceive these distinctions.

Earlier findings on categorical perception and memory for linguistic features are challenged by two findings from recent research. Firstly, it has been shown that speech perception is not strictly categorical. Recent studies show that the categorical responses measured in perception experiments reflect continuous encoding of speech units. For example, in an eyetracking study using the visual world paradigm McMurray and colleagues [10] found that although speakers responded categorically when identifying sounds, eye fixations to pictures of targets (e.g. beach) and competitors (e.g. peach) suggested that subcategorical variation is driving the eyes. These results were mirrored in later ERP results [12] which showed that listeners were sensitive to within-category differences in voice onset time, leading Toscano et al. to conclude that "at perceptual levels, acoustic information is encoded continuously, independent of phonological information."

Secondly, work by Goldinger [5] and Pufal & Samuel [11] shows that even detailed information that is not linguistically meaningful is stored in memory. A word recognition memory task shows that at time lags of up to one week listeners are better at identifying whether they have heard a word before or making a judgement about a word if the word is spoken with the same voice and with the same background noise as when they first heard it. This is true even for background noise that is meaningless or non-linguistic, such as a dog barking. These findings are taken as evidence that listeners create episodic memories of speech that include acoustic detail of within-category variation as well as information that is not linguistically meaningful.

The evidence we have presented thus far appears to be contradictory: in the cases of categorical perception and stress deafness listeners appear not to remember phonetic information that is not meaningful in their language. However, recognition memory tasks and priming tasks show that listeners do remember phonetic details that are not meaningful in their language, because they are faster to recognize or make a judgement about words that are presented exactly as they first heard them. Our experiment is designed to directly address the question of whether listeners encode subcategorical detail for speech. We focus on the perception of intonational pitch accent and ask whether listeners remember subcategorical detail in the acoustic parameters that encode pitch accents. We examine pitch accents for three reasons: First, pitch accents in English differ from other phonological features in that they do not mark lexical contrasts, which are clearly categorical, but function instead to mark information status distinctions related to focus and accessibility, which are potentially gradient. Second, although it has been investigated [9], there is not yet strong evidence that pitch accents are categorically perceived. Third, while many studies of American English find acoustic correlates of pitch accent in measures of intensity, duration and/or f0, [1] there is less evidence to indicate which of these acoustic properties listeners pay attention to in perceiving and interpreting pitch accents.

We test memory for two types of variation related to pitch accent: phonological variation (presence vs. absence of a pitch accent), and variation in the phonetic cues to pitch accent (F0 peak values, duration of accented word). If listeners encode and remember subcategorical acoustic detail related to pitch accent, they will be sensitive to variation both in the accent status and phonetic cues of stimulus utterances. If listeners instead create and store an abstract representation of accent status that does not include specific phonetic cues, they will only be sensitive to differences in accent status. We report results from a set of six experiments that test these predictions.

2. METHOD

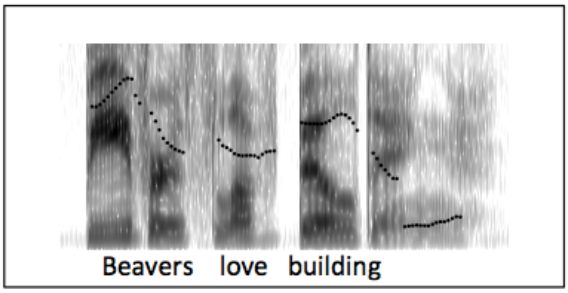
The study consists of two sets of three experiments. All experiments involved listeners hearing a speech sample, and then hearing a second test sample, in some cases after delay or interference. They must judge whether or not the test sample is the same recording as the study sample. When the sample is not the same, it can differ in accent categorically, or with respect to a subcategorical change in a cue to accent.

2.1. Stimuli

Stimuli were words excised from natural productions of sentences of American English designed to have mostly voiced segments (e.g. "Beavers love building"; see Fig. 1). Twelve nouns were used from six sentences (e.g., "beavers" and "building"). Each sentence was recorded with four accent patterns (first noun H* and second noun H*; first noun H* and second noun unaccented; first noun unaccented and second noun H*; neither noun

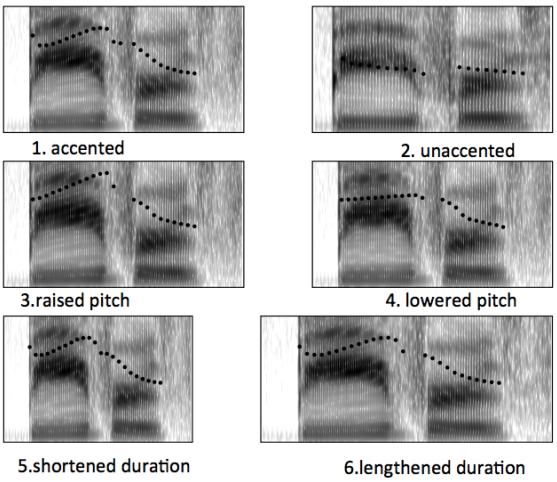
accented) by a trained linguist who was not part of the research team and was not aware of the research goals. Target nouns in the sentences were spliced out in their accented or unaccented form.

Figure 1: Spectrogram of the sentence "beavers love buildings" with both nouns pitch accented.



Each accented word was resynthesized to create a large, perceptually salient phonetic difference that stayed within the accent category of the original production (i.e., the manipulated word was within the distribution of either unaccented or accented tokens for that acoustic measure). To manipulate pitch, we stylized the pitch contour using Praat [2] and manually adjusted the pitch peak up 25 Hz or down 25 Hz. We also used PSOLA resynthesis in Praat to decrease or increase the duration of the entire word by 10%. These phonetic differences were found in pilot studies to be detected at the same rate as the accented/unaccented difference for our materials, in an AX task. Thus, the differences were chosen to equate the perceptual salience of the differences for the three conditions.

Figure 2: Spectrograms of word "beavers." 1 and 2 are different recordings, 3-6 are resynthesized versions of 1.



2.2. Participants

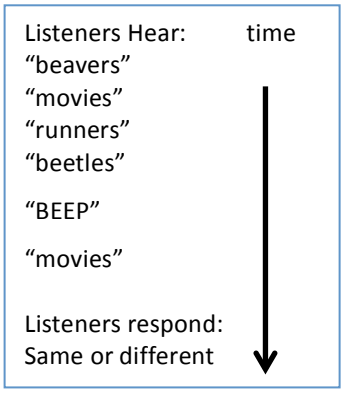
193 total subjects participated in six separate experiments. All subjects were self-reported native English speakers located in the United States. Their ages ranged from 19-59 (mean=31,s.d. 8.4). Results reported here do not include subjects who did not finish the task (8) or self-reported bilinguals (5), leaving 30 subjects in each of the six experiments.

2.3. Procedure

All experiments were conducted online using Amazon Mechanical Turk. In the first three experiments participants heard two words with one second of silence between. They were immediately asked to click on a button to indicate if they were "the exact same recording or different recordings." (an AX task) The pairs of words were either the same recording (1/2 of the trials), or they differed in one of three ways. In experiment 1, words varied in accent status, meaning that participants would hear a naturally produced accented recording of a word and then a naturally produced unaccented recording of the same word (or vice versa). In experiment 2, participants heard a shortened version of a recording and then a lengthened version of the same recording (or vice versa). In experiment 3 participants heard a version of a recording with lowered pitch peak and then a version with raised pitch peak (or vice versa). In experiments 2 and 3 all stimuli were resynthesized, so subjects were never asked to distinguish between a naturally produced token and a resynthesized token.

Experiments 4, 5 and 6 use the same stimuli but add a delay and interference, to make recognition more difficult. Listeners heard four different words (exposure), then a tone, and then another presentation of a word from the exposure phase (test). They were asked to report whether the test word was exactly the same recording as the exposure version. This task is more difficult than the AX tasks in Experiments 1-3 due to the added interference from the following words, the time delay between encoding and retrieval, and the increased working memory load because the subject must hold all words in memory until they hear what the test word is.

Figure 3: procedure for experiments 4,5,6.



3. RESULTS

Results of the AX task show that listeners are well above chance at discriminating all three contrasts, (exp 1=77%, exp 2=85%, exp 3= 75%) meaning they correctly marked pairs of stimuli that were the same as "Same" (hits) and pairs of stimuli that differed as "Different" (correct rejections) above and beyond the rate that would be expected if they were guessing.[1] Critically, listeners did not differ significantly in their ability to hear accent status differences compared with duration or pitch changes. This suggests that the three differences we presented are equivalently salient and easy to differentiate in the AX task with a very short time lag between.

When a longer time delay and interference are added, listeners are still accurate at remembering accent differences: listeners did not differ significantly in their recognition accuracy for accent between the immediate response task (AX) and the delayed response task (77% AX vs. 83% with delay). For the phonetic differences, listeners were still significantly above chance in the delay condition, but were significantly worse at recognizing phonetic differences after delay and interference than in the AX task. This was true both for pitch differences (75% AX vs. 54% with delay) and duration differences (85% AX vs. 67% with delay).

Table 1: Mean percentage correct by experiment. Light grey values do not differ significantly.

	Accent	Duration	Pitch
AX (Exp 1,2,3)	77%	85%	75%
Delay (Exp 4,5,6)	83%	67%	54%

Overall, these results show that the accent difference, which was equally as salient as the pitch and duration differences, is still remembered after a time lag and in the presence of interference, while duration and pitch differences are detectable at a rate above chance, but are much less accurately remembered.

Group effects hold when analyzed with a mixed effect logit model with random slopes and intercepts to account for individual variability. However, examining individual performance in the AX task shows that listeners' memory for prosodic features is variable. The standard deviation of scores in the AX pitch task was significantly higher than the standard deviation of scores in the AX duration task ($F(29,29) = 2.7492$, $p<.01$) or AX accent task ($F(29,29) = 3.1501$, $p<.01$), meaning performance varied more from listener to listener in the pitch task than in the duration or accent task. This holds despite the fact that these same listeners were excellent at discriminating a pure tone difference of the same magnitude pitch in a post-test (mean=91% correct, s.d. =.133%).

4. DISCUSSION

Results of experiments 1, 2, and 3 show that pitch accent status and the phonetic cues that express pitch accent are perceived, encoded and available for immediate access. In contrast, experiments 4, 5, and 6 provide evidence that after a delay and interference some information about pitch and duration is accessible, but phonological accent status is much more accessible in memory. This evidence is consistent with the hypothesis that listeners encode detailed instances of pitch accents, but that phonetic detail related to pitch accent quickly becomes less accessible in memory compared to the categorical accent distinction.

In the introduction we argued that there were two schools of thought on memory for speech, the abstractionist and episodic views, and that they make separate predictions. However, results of our study meet both predictions. We believe that these two predictions are compatible if a distinction is made between encoding and retrieval, that is, if all acoustic information is encoded, but not all information is retained or accessible later at retrieval.

We also found that listeners varied in their accuracy rates across the different tasks, and particularly in the pitch manipulation. This is surprising given that the pitch difference we tested is well above the *just noticeable difference* (JND) in this range [12], and in light of wide agreement that f0 patterns are one way that speakers signal pitch accent in American English [1]. Our study provides evidence that listeners as a group have poor memory for within-category pitch differences, but does not provide answers as to why within our group of subjects some listeners are better able to report pitch differences in both pure tones and speech. We speculate that musical and linguistic experience could have an effect on accuracy of pitch memory, and we believe that this merits further study.

Taken together, our results suggest that (1) listeners encode both categorical distinctions and phonetic detail in memory, but categorical distinctions are more accessible at retrieval in an explicit judgment task and (2) listeners may vary in the degree to which they remember or can access prosodic detail.

5. REFERENCES

[1] Breen, M., Fedorenko, E., Wagner, M., Gibson, E., 2010. Acoustic correlates of information structure. *Language and Cognitive Processes* 25, pp.1044-1098

[2] Boersma, P, Weenink, D. 2015. Praat: doing phonetics by computer [Computer program]. Version 5.4.08, retrieved 24 March 2015 from http://www.praat.org/

[3] Dohmahs, U., Knaus, J., Orzechowska, P.,Weise R. 2012. Stress "deafness" in a language with fixed word stress: an ERP study on Polish. *Frontiers in Psychology* 3, 439

[4] Dupoux, E., Sebastian-Galles, N., Navarrete, E., Peperkamp, S. 2008. Persistent stress 'deafness': The case of French learners of Spanish. *Cognition* 106, (2).

[5] 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 (5) pp.1166-1183.

[6] Goldinger, S.D. 2007. A complementary-systems approach to abstract and episodic speech perception. *Proc. 16th ICPhS Saarbrucken*

[7] Goto, H. 1971. Auditory perception by normal Japanese adults of the sounds `l' and `r' *Neuropsychologia* 9 pp. 317–323

[8] Klatt, D., Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *Journal of the Acoustical Society of America* 53, (8)

[9] Ladd, D.R., Morton, R., 1997 The perception of intonational emphasis :continuous or categorical? *Journal of Phonetics* 25, pp.313-342

[10] McMurray, B., Tanenhaus, M., Aslin, D., 2002. Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86 (2).

[11] Pufal, A., Samuel, A. G. (2014). "How lexical is the lexicon? Evidence for integrated auditory memory representations" *Cognitive Psychology* 70, 1-30.

[12] Toscano, J.C., McMurray, B., Dennhardt, J., and Luck, S.J.2010 . Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science* 21 (10) 1532-1540

[1] Statistical significance holds when results are converted to *d'*. Values were as follows: AX task: accent= 0.79, duration=1.09, pitch =0.78. Delay task: accent=1.14, duration=0.51, pitch=0.15