Cognitive Psychology 88 (2016) 88–114

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration

Arthur G. Samuel *

Basque Center on Cognition, Brain and Language, Donostia, Spain
IKERBASQUE, Basque Foundation for Science, Spain
Stony Brook University, Dept. of Psychology, Stony Brook, NY, United States

ARTICLE INFO

Article history:
Accepted 30 June 2016
Available online 16 July 2016

Keywords:
Perceptual recalibration
Attention in recalibration
Lexical processing time

ABSTRACT

In listening to speech, people have been shown to apply several types of adjustment to their phonemic categories that take into account variations in the prevailing linguistic environment. These adjustments include selective adaptation, lexically driven recalibration, and audiovisually determined recalibration. Prior studies have used dual task procedures to test whether these adjustments are automatic or if they require attention, and all of these tests have supported automaticity. The current study instead uses a method of targeted distraction to demonstrate that lexical recalibration does in fact require attention. Building on this finding, the targeted distraction method is used to measure the period of time during which the lexical percept remains malleable. The results support a processing window of approximately one second, consistent with the results of a small number of prior studies that bear on this question. The results also demonstrate that recalibration is closely linked to the completion of lexical access.

* Address: Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor, 20009 Donostia (San Sebastián), Spain.
E-mail address: a.samuel@bcbl.eu

http://dx.doi.org/10.1016/j.cogpsych.2016.06.007
0010-0285/

1. Introduction

Perception is a product of mental processes that operate on sensory information. As such, there is an indirect relationship between the actual stimulus and the percept – perception is not veridical, and as mental processes change, so does the percept. The disconnect between an objective description of what the sensory systems provide and what is perceived can be seen across different sensory systems, and across different types of input. For example, because of the physical layout of the receptors on the retina, in each eye there is a region that lacks receptors – the "blind spot". If perception were veridical, then an observer looking at a pattern with only one eye open should see a disruption in the pattern at the spatial location corresponding with the blind spot. This does not occur. Instead, the perceiver fills in the missing part of the visual field (Ramachandran, 1992), using information at the edges of the blind spot to create surprisingly rich patterns (Spillman, Otte, Hamburger, & Magnussen, 2006).

In perceiving spoken language there is an analogous filling in process. Warren (1970) removed part of a word and replaced it with an extraneous noise. Listeners did not perceive the gap in the spoken word, just as observers do not perceive a break in the visual pattern at the blind spot. Rather, they heard the speech as being intact, with the extraneous sound heard as a separate perceptual object.

In fact, speech perception abounds with examples of how constructive the perception process is. Probably the best-known phenomenon in the speech domain is categorical perception (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). In this case, when an experimenter creates a continuum of stimuli (e.g., from /ba/ to /da/) by changing one syllable into the next by a series of linear physical changes (e.g., increasing the starting frequencies of the formants by a fixed amount), the change in perception across the continuum is entirely nonlinear, with essentially imperceptible changes within each category, and a sudden perceptual shift at the category boundary. Thus, the mental processes that operate on the linearly changing sensory signal yield percepts that do not follow the linear changes. More sensitive measures (e.g., reaction time, Pisoni & Tash, 1974, and eyetracking, McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008) or testing procedures (e.g., Samuel, 1977) reveal that more continuous information is present within the processing system, but the subjective experience is largely categorical.

Two other cases of non-veridical perception of speech are more closely related to the issues that will be probed in the current study. When we teach students about speech perception, we often provide two examples of how percepts deviate from the actual signal: (1) Listeners hear words as being separated by short breaks, whereas the reality is that there are no such breaks in the signal – looking at a waveform in general provides no obvious clue about where words begin and end. (2) Speech from a language we don't know seems very fast (and without the breaks between words), but slows down to "normal" when we become proficient in this second language. Both of these extremely intuitive phenomena illustrate the way that perception is constructed from the sensory signal, guided by highly sophisticated knowledge. The first case demonstrates that perceptual processes have developed that take variable lengths of the speech signal and partition the speech into a series of perceptually discrete objects – words. The apparent slowing down of second-language speech with increasing expertise highlights the impressive adaptability of the speech processor – available information is used to optimize performance.

In Part I of the current study, the focus will be on this adaptability of perceptual processing of speech. There are multiple literatures that examine different aspects of how listeners adjust the perceptual processes that they bring to bear on the sensory input, over various time scales. The apparent slowing down of a second language takes place over months or years of exposure to the language. Other adjustments may take place over days or weeks of training, such as learning how to understand speech with a heavy non-native accent (e.g., Bradlow & Bent, 2008). There have been many training studies in which listeners initially have great trouble understanding degraded speech (e.g., through poor synthesis Fenn, Nusbaum, & Margoliash, 2003, or vocoding Davis, Johnsrude, Hervais-Ademan, Taylor, & McGettigan, 2005), but after several hours of training performance is significantly better.

Two complementary forms of adjustment that have been extensively studied operate on a much shorter time-scale, a matter of minutes. The first, selective adaptation of speech, was initially reported by Eimas and Corbit (1973), followed by many others (for a review, see Samuel, 1986). In selective

adaptation, after a listener has heard a particular sound (the "adaptor") many times over the course of several minutes, perception of similar speech sounds is shifted in a contrastive way: Given the same sensory input, the listener is now less likely to perceive the input as being a member of the phonetic category of the adaptor.

A second adjustment process that operates on a short time-scale is perceptual recalibration. This adjustment is complementary to selective adaptation because the effect is one of assimilation, rather than contrast (but see Kleinschmidt & Jaeger, 2015, for an attempt to treat both phenomena within a single model). In recalibration, the auditory input is phonetically ambiguous (e.g., a sound midway between /b/ and /d/), and some type of context provides disambiguation. Bertelson, Vroomen, and de Gelder (2003) used audiovisual presentation, with the visual (lipreading) information providing the disambiguation of the auditory signal. Norris, McQueen, and Cutler (2003) used purely auditory presentation, with lexical context providing the disambiguation of a phonetically ambiguous sound (e.g., a sound that was heard as /f/ as the end of the word "giraffe", but as /s/ as the end of a word like "horse"). In both cases, in addition to an immediate contextual effect (e.g., subjects hearing the ambiguous /sf/ mixture as /f/ in "giraffe"), perception minutes (Kraljic & Samuel, 2005) or even hours (Eisner & McQueen, 2006) later was also shifted in an assimilative direction – ambiguous tokens were now more likely to be heard in accord with the earlier contextually-driven manner. Note that this effect is opposite in direction to what occurs in selective adaptation.

Over the years, there have been a number of studies that investigated whether selective adaptation and/or perceptual recalibration operate automatically, without requiring attention. For adaptation, all of these studies (Baart & Vroomen, 2010; Mullennix, 1986; Samuel & Kat, 1998; Sussman, 1993) have yielded an affirmative answer. Similarly, both of the previous tests of automaticity of recalibration (Baart & Vroomen, 2010; Zhang & Samuel, 2014) have concluded that these adjustments are inherent in speech processing and do not consume attentional resources. Experiment 1 revisits this issue for recalibration, using a new and more sensitive procedure than in previous work.

Recall that one of the standard examples of non-veridical speech perception is the apparent breaks between words, when none are present in the signal. It was suggested above that this phenomenon stems from the speech processor's mapping stretches of the waveform onto discrete perceptual objects – words. This process of chopping the input into discrete perceptual structures must operate over some kind of window. That is, given the continuous nature of speech, the formation of any given object must be based on a certain amount of information spread over time. In Part II of the current study, the method that is used in Experiment 1 to investigate attentional effects on perceptual recalibration is extended to study the "perceptual window" for recognizing spoken words. To oversimplify a bit, the method provides a measure of the window in which providing the system with more processing time leads to more perceptual recalibration. This can be taken as the time during which perception of a word is still malleable, before the final conscious percept is produced.

Thus, Part I uses a new technique to test whether attention is needed to recalibrate phonetic categories on the basis of contextual disambiguation of phonetic information, and Part II uses this technique to measure the duration of a word's perceptual window. Finally, in Part III, Experiment 3 provides a refinement of the interpretation of the attentional results, Experiment 4 clarifies how best to think about the duration of the perceptual window, and Experiment 5 shows that the strength of recalibration patterns with the success of lexical access. Collectively, the five experiments yield very clear answers to two basic questions about the cognitive processes needed to perceive spoken language: (1) Is attention involved in revising perceptual categories? and (2) how long does the perception of a word remain malleable and open to revision by information gathered by these processes?

2. Part I: Is attention needed to recalibrate phonetic categories?

For the past half century the central question in research on speech perception and spoken word recognition has been how listeners can overcome the very substantial variation that is present in the signal. As noted above, one key to success seems to be the constant adjustment of the perceptual process, over multiple time scales. In the current study, we focus on an adjustment process that seems to operate over the course of minutes, and to last for at least 12 h (Eisner & McQueen, 2006), but prob-

ably less than a week (Zhang & Samuel, 2014): recalibration of phonetic category boundaries driven by the lexical interpretation of ambiguous phonetic segments (Norris et al., 2003; see Samuel and Kraljic (2009), for a review). In Experiment 1, we pursue an issue about recalibration that has been examined in several previous studies, but to date remains an open question: Does recalibration occur automatically, or does it instead depend on the allocation of attention?

2.1. Experiment 1

There have been three mostly independent literatures that examine shifts in the perceptual boundaries between phonetic categories: Selective adaptation (seminal study: Eimas & Corbit, 1973), audiovisual recalibration (seminal study: Bertelson et al., 2003), and lexically-driven recalibration (often and originally referred to as perceptual learning for speech) (seminal study: Norris et al., 2003). These three phenomena share the property that they manifest in shifting category boundaries, typically measured either by changes in identifying members of a test series (usually a set of 5–10 syllables that change from one category to another across the continuum), or by measuring the identification of items selected to be at the boundary between two categories. The three cases also involve loosely equivalent triggering conditions in the sense that each normally includes an initial exposure period on the order of minutes, rather than seconds, hours, or days/weeks, with a subsequent test phase.

There are, however, substantial differences among the three cases: The two recalibration effects are assimilative while adaptation is contrastive; audiovisual recalibration has a very short duration (less than a minute: Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004), adaptation lasts minutes or hours, and lexically driven recalibration lasts at least 12 h (Eisner & McQueen, 2006) but less than a week (Zhang & Samuel, 2014); audiovisual recalibration peaks with about eight exposure tokens whereas adaptation keeps building across hundreds of exposure tokens (Vroomen, van Linden, de Gelder, & Bertelson, 2007).

One commonality among the three is that the possible role of attention in driving each effect has been tested in prior research, and in each case the existing evidence supports the conclusion that these adjustment processes operate automatically (however, see Scharenborg, Weber, & Janse, 2015, for some evidence that lexical recalibration may be related to attention-switching ability). For example, Samuel and Kat (1998); see also Mullennix, 1986, and Sussman, 1993) tested the role of attention by giving subjects tasks to do during the time when the adaptor was being played repeatedly. One group had to solve a series of visual arithmetic problems presented rapidly. Another group was shown a rapid series of visually presented word triplets (e,g., "such" – "touch" – "hutch", or "bear" – "chair – "near") during the adaptation phase, and had to make rhyme judgments (YES for the first example, and NO for the second). The size of the adaptation shift in these divided attention conditions was compared to a baseline case with no second task, and there was no decrease in the size of shift at all, even for the phonemic rhyme task.

Comparable tests have been conducted for the two recalibration effects, with comparable null effects. For audiovisual recalibration, before each set of recalibration-inducing audiovisual stimuli, Baart and Vroomen (2010) gave subjects information to maintain in memory, to be probed at the end of the recalibration stimulation. One group of subjects had to keep moving dot patterns in memory, and a different group of subjects had to maintain a set of three, five, or seven letters in mind while exposed to the audiovisual input. Neither secondary task reduced the recalibration effect. Baart and Vroomen also used these distractor tasks in a version of the selective adaptation paradigm, and found no decrease in that effect either. Zhang and Samuel (2014) looked for attentional effects on lexically driven recalibration using both the memory load and the simultaneous task approaches. One group had to keep a set of consonants in mind during presentation of sentences that contained the critical ambiguous segments. Another group was given a visual search task of letter arrays during the exposure sentences. Neither task reduced the recalibration effect.

Given the apparently substantial evidence that these perceptual adjustments are automatic in the sense that they do not seem to be reduced by attentionally-demanding simultaneous tasks, a reasonable question would be: Why revisit this issue? The answer to this question is that the prior studies usually include an important disclaimer. For example, with regard to lexically driven recalibration, Zhang and Samuel (2014) noted that the null effect of each load task "indicates that the retuning is

a relatively robust process that does not require full attention. This is not to say that perceptual learning can always operate normally regardless of other demands on the system (p. 212)". For the audiovisual case, Baart and Vroomen (2010) said "Admittedly, the critical part of the exposure phase that induces recalibration – the part in which a participant hears an ambiguous segment while seeing another phonetic segment – is very short, and there is no guarantee that participants were – at that specific time – actually engaged in repeating the memory items. Unfortunately, we cannot offer an obvious solution for this because it is a very general problem in dual-task paradigms where there is always uncertainty about strategic effects in performing the primary and secondary task (p. 580)".

What both caveats were getting at is indeed a very general problem with the dual-task approach to assessing attentional effects: Subjects are clever creatures, and may find a way to complete both tasks using some kind of time-sharing. To the extent that they can do so, their performance on the task of interest will not show costs of the distractor task. Hence, the caution shown in both of these papers – the null effect of a second task could either be showing that the primary task is indeed automatic, or it might just be the result of clever time-sharing strategies.

Thus, in order to provide a more rigorous test of the automaticity of the recalibration process, the test must be structured in a way that makes time sharing difficult or impossible. If that can be accomplished, then a null effect of the second task would indeed be suggestive of true automaticity. Experiment 1 introduces a new methodology that is designed to provide this rigorous test of automaticity. The key new feature of the method is the use of a distractor task that is tightly timed to the moment at which the recalibration computations should be occurring. This approach can be thought of as "targeted distraction", and as will become clear, such targeted competition offers a very general and useful tool.

2.1.1. Targeted distraction

Experiments 1–4 in the current study follow the same general procedures as in prior work on lexically driven recalibration, with an initial exposure phase followed by a test phase. The procedures used by Kraljic and Samuel (2005); these were very similar to those in the seminal work by Norris et al., 2003) provide the best reference point for the work here because the critical stimuli in the current study were taken from the Kraljic and Samuel study. In that study, the exposure task was an auditory lexical decision task (100 words, 100 pseudowords) that included 20 critical items for each subject. For half of the listeners these 20 items were words in which an "s" was replaced by an ambiguous mixture of "s" and "sh" (e.g., "dino[s/sh]aur"). For the other half of the subjects, the 20 critical items were words with an "sh" replaced by the ambiguous mixture (e.g., "offi[s/sh]al"). After the exposure phase, subjects identified items from a continuum that ranged between "asee" and "ashee". Robust perceptual recalibration was observed, with significantly more report of "ashee" by the subjects who had been exposed to critical items like "offi[s/sh]al" than by those who heard items like "dino[s/sh]aur".

In Experiment 1, the same critical words and the same test series are used. The important change is in the task done during the exposure phase. Two versions of Experiment 1 were run. In both versions, during the exposure task listeners did 100 trials, with each trial including the dichotic presentation of two items. One item was always a word said in a female voice. The other item was either a word (50%) or a pseudoword (50%), said in a male voice. Fig. 1 shows the waveforms for the female voice (top) and the male voice (bottom) from one trial. Critically, as the figure shows, the female word always began 200 ms before the male item. The assignment of the male and female items to the right or the left ear on each trial was randomly determined, so listeners could not pre-focus attention on one speaker or the other. In Experiment 1a, the subject's task was to make a lexical decision on the *male* item. However, the critical (ambiguous) segments were always in the *female* voice. Thus, in Experiment 1a, the task requirements were designed to force attention away from the critical information needed to induce perceptual recalibration. This is *targeted distraction* because the timing of the distractor is designed to impact processing during the moment when any computations underlying recalibration would be taking place. In particular, by interrupting processing of the female word just 200 ms after it begins, the procedure should disrupt lexical access of the female word. Given that recalibration depends on successful lexical access (Norris et al., 2003), if the targeted disruption impairs lexical processing, it should also reduce or eliminate recalibration.

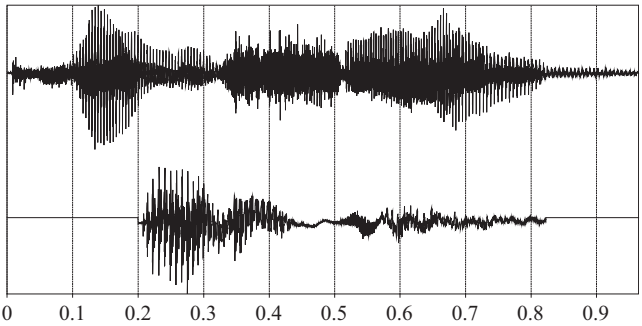

Fig. 1. Waveforms for a dichotic trial, with the female item (top) beginning 200 ms before the male item (bottom). Time (in s) is shown along the *x*-axis.

Experiment 1b is a control experiment that addresses a plausible concern that could be raised about the procedure in Experiment 1a: Because the male item is presented almost simultaneously with the critical female word, an absence of perceptual recalibration could occur as a result of simple masking – if the listener cannot hear the female word clearly, it would not be surprising to see no recalibration from it. Presenting the stimuli to opposite ears, in two very different voices, was designed to eliminate energetic masking and to minimize informational masking, but some masking could still occur. Thus, in Experiment 1b, the physical conditions were identical to those in Experiment 1a, equating any potential masking effects. However, the subjects in Experiment 1b were given a different task than those in Experiment 1a. On each trial, they were to count the number of syllables in the *female* word, and to push one of three buttons corresponding to the three possibilities (all words were two, three, or four syllables long). This task was chosen to direct the listener's attention to the stimuli that contained the ambiguous segments, allowing the subjects to complete their lexical processing of the female words. If the results of Experiment 1b are similar to those found in our previous work with these stimuli (i.e., robust perceptual recalibration), then the new procedure is not imposing any major masking issues. If so, then we can interpret the results of Experiment 1a with confidence: If similar recalibration occurs, then attention is not necessary. If recalibration disappears, then we can conclude that recalibration is not automatic – resources are required to complete the lexical processing that generates recalibration.

2.2. Method

2.2.1. Participants

Our lab has conducted over a dozen previous experiments that examined various properties of lexically driven perceptual recalibration. For relatively basic effects in this paradigm we typically have tested about 50 subjects per experiment. For cases examining a more subtle, second order effect, we have generally tested about 65 subjects per experiment. McQueen and his colleagues have also maintained an active research program in this domain (e.g., Eisner & McQueen, 2005, 2006; Jesse & McQueen, 2011; Norris et al., 2003). Across these studies, the number of subjects per experiment has ranged from about 24 to 48, typically 30–35. In the current study, as some of the conditions are testing for the abolition of the recalibration via targeted distraction, we tested larger samples so that any null shift would not be attributable to a lack of power. The exact number of subjects in any given experiment varies a bit, but all experiments had many more subjects than in previous work, with a median of 96 subjects per experiment. In Experiment 1a, 94 subjects were tested. In Experiment 1b, 77 (different) subjects participated. Given that well over a dozen previous studies (including ones using this same stimulus set) have shown that 30-50 subjects provide enough power to observe an effect when it exists, the substantially larger samples here ensure that we will be able to detect an effect if one occurs. Subjects were undergraduates at Stony Brook University and received credit toward a research requirement for one of their classes. All subjects were native English speakers with no known hearing problems.

2.2.2. Stimuli

As noted above, stimuli were digital copies of those used by Kraljic and Samuel (2005). In that study, perceptual recalibration was tested using stimuli from both a female and from a male speaker. The current study used 60 of the filler female words, the 20 female critical "s" words, and the 20 female critical "sh" words. In addition, 50 of the male filler words and 50 of the male filler nonwords were used. The only occurrences of "s" or "sh" were in the critical female words. The critical ambiguous sounds were always syllable-initial, and occurred relatively late in the words to allow strong lexical access (see Jesse & McQueen, 2011, for evidence that the latter is important). Table 1 shows the duration of the 40 critical female words, the onset/offset times of the critical "s" or "sh" segments, and the location of the critical "s" or "sh" relative to each word's uniqueness point (the segment at which only the word itself, or a close morphological relative, would be a possible completion). As the uniqueness point values show, all words became lexically unique within one segment of the critical fricative, with the vast majority becoming unique at or before that segment. Thus, when the

Table 1
Stimulus information (times in ms; uniqueness point is in segments, relative to the critical fricative) for the critical female words.

Word	Word Duration	Fricative Onset	Fricative Offset	Fricative Duration	Uniqueness Point
Arkansas	870	369	501	132	−2
Coliseum	964	324	517	192	0
Compensate	855	462	596	134	−1
Democracy	978	530	711	182	−3
Dinosaur	841	303	472	169	1
Embassy	708	266	447	182	−1
Episode	849	231	404	173	1
Eraser	840	370	555	185	1
Hallucinate	1051	307	476	169	−1
Legacy	820	325	523	198	−1
Literacy	812	346	526	180	0
Medicine	818	286	474	188	0
Obscene	881	203	406	203	1
Parasite	1038	356	534	178	1
Peninsula	903	342	501	160	−2
Personal	792	152	312	160	0
Pregnancy	927	479	627	148	−4
Reconcile	1017	419	557	138	0
Rehearsal	713	325	485	160	−3
Tennessee	884	287	465	178	−1
Ambition	918	399	542	144	0
Beneficial	883	433	597	165	−2
Brochure	853	181	387	206	0
Commercial	789	362	525	163	−2
Crucial	794	239	418	179	0
Efficient	956	337	481	144	−1
Flourishing	1001	452	647	196	−1
Glacier	771	266	464	197	0
Graduation	1020	541	705	164	−1
Impatient	1010	438	567	129	0
Initial	765	268	430	163	0
Machinery	900	110	291	182	0
Negotiate	1129	361	529	168	−1
Official	819	346	501	155	−1
Parachute	941	361	540	179	0
Pediatrician	1045	588	755	167	−6
Publisher	889	376	544	169	0
Reassure	922	228	449	221	0
Refreshing	901	392	548	156	−1
Vacation	967	475	637	162	−1
Mean	896	346	516	170	−0.8

ambiguous segment occurs, listeners have sufficient information to access and select the lexical carrier for that segment. See Kraljic and Samuel (2005) for details of stimulus construction.

Two stimulus lists were constructed. Both lists had the same 50 male words and 50 male pseudowords, as well as the same 60 female filler words. The lists differed in the makeup of the critical female words. One list included the 20 "s" words with the original segments replaced by the ambiguous mixtures of "s" and "sh", together with the normal versions of the 20 "sh" words. The other list included the converse – 20 normal "s" words and 20 "sh" words with ambiguous segments replacing the "sh" segments. Each female word was randomly paired with a male word or pseudoword.

The stimuli for the test phase were the same as those used in Kraljic and Samuel (2005): A set of six items that ranged between "asee" and "ashee". These were constructed by using differently weighted mixtures of an original "asee" token and "ashee" token produced by the same female speaker who produced the words used in the exposure phase.

2.2.3. Apparatus and procedure

In both Experiments 1a and 1b, subjects were randomly assigned to one of the two exposure stimulus lists. Subjects were tested in groups of up to three at a time in sound attenuated chambers. Stimuli were presented dichotically over high quality headphones (Sony MDR-V900). The assignment of channels was determined randomly on each trial, but the onset of the female word was always 200 ms before the onset of the male word or nonword. In Experiment 1a, subjects used two buttons on a button board to indicate whether the male voice had produced a word or a pseudoword. In Experiment 1b, the subjects used three buttons on the board to indicate whether the female word had two, three, or four syllables. Subjects were encouraged to respond both accurately and quickly. Each trial began one second after the response(s) had been received on the previous trial, with a timeout maximum of 2500 ms. The exposure phase lasted approximately 10 min. Following the exposure task, all subjects did the same phonetic identification task. They heard 14 randomizations of the 6-step "asee" – "ashee" continuum, and used two buttons on the response board to identify each item as either "asee" or "ashee".

2.3. Results and discussion

Following our usual procedures for recalibration experiments (e.g., Kraljic, Brennan, & Samuel, 2008; Kraljic & Samuel, 2005, 2006, 2007, 2011; Kraljic, Samuel, & Brennan, 2008), in this and in all of the following experiments the labeling performance of each subject was initially screened to assure that the participant was able to systematically identify the members of the "asee" – "ashee" continuum. This screening eliminates any participants who made no effort to do the task, or who for some reason could not do so. Two exclusion criteria were used. The main criterion for exclusion was a failure to have at least a 35% difference in "sh" report between the most extreme "s" and the most extreme "sh" on the six-step continuum (as will be seen in the data below, the norm is a difference of 80–90%). Subjects who do not show at least this much ability to discriminate the endpoint members of the test series are effectively not doing the task that they have been asked to do. A second exclusion criterion was the presence of a very non-monotonic change in identification across the continuum; if such a "bump" in the function affected at least two of the six members of the continuum, the subject's data were excluded. Collectively, across all of the experiments in the current study, the first criterion eliminated 11% of the subjects, and the second eliminated less than 2%. The screening eliminated nine of the 94 participants in Experiment 1a, and ten of the 77 participants in Experiment 1b. The exclusion procedure creates a very slight bias against finding recalibration effects (when they are present) because it tends to eliminate a few subjects per experiment who essentially always respond "sh" after being in the ambiguous "sh" condition (and it is not possible to know if such heavy "sh" responding is due to very strong recalibration or to a subject simply pushing one button all of the time).

Data analysis also followed the same procedures used in previous studies of this type. For each participant, the average percentage of "sh" identification for the middle four items of the 6-step continuum was computed. This is the region that is usually most sensitive to any recalibration effects, and using an average of the middle items also leads to scores that are usually far from the percentage extremes (i.e., it avoids floor or ceiling compression effects). By using data from four tokens, each sub-

ject provides a large enough sample (56 responses) to provide a good estimate of the true value. To test for recalibration, a single-factor between-subject analysis of variance compared identification of the test series by those who had been exposed to ambiguous segments in "sh" words to those whose exposure to those segments was in "s" words. Recalibration is inferred if the test series is heard as being more "sh"-like by the first group than by the second group.

Fig. 2 presents the average identification of the test continuum for each exposure group. The left panel shows the results for subjects who had made a lexical decision on the male voice during the exposure phase, and the right panel shows the corresponding results for the listeners who counted the syllables in the female words. The results are exceptionally clear: The targeted distraction imposed in Experiment 1a completely abolished the recalibration effect. The loss of recalibration cannot not due to the male voice masking the female voice, as shown by the quite robust recalibration found in Experiment 1b. If the absence of recalibration in Experiment 1a had been caused by the male voice masking the female voice, preventing subjects from having a sufficient signal to recognize the female words, the same absence of recalibration would necessarily have been found in Experiment 1b because the acoustics in the two experiments were identical.

Summarizing these differences more quantitatively, with targeted distraction the average "sh" report in Experiment 1a was 65.8% for participants who had heard ambiguous sounds in "s" words, and 66.0% for those who heard the ambiguous segments in "sh" words, a shift of 0.2%, $F(1,81) = 0.002$, n.s. In contrast, under identical physical listening conditions, in Experiment 1b the "s" exposure group produced 57.4% report of "sh", versus 71.0% for the "sh" exposure group, a shift of 13.6%, $F(1,65) = 11.126$, $p < .001$. This shift is similar to, but slightly smaller than, the effect (17.5%) Kraljic and Samuel (2005) observed with the same stimuli when there were only female stimuli, with subjects making a lexical decision on those stimuli during the exposure phase.

The results of Experiment 1 show that the caution expressed by Baart and Vroomen (2010) and by Zhang and Samuel (2014) was warranted. Those studies, along with several others that found no reduction in identification shifts under dual task or memory load conditions, had suggested that recalibration is an automatic process that does not require attentional resources. The targeted distraction procedure in the current study shows that this is not the case – when subjects were required to process the male voice in order to make a lexical decision, no recalibration occurred.

The success of the targeted distraction procedure opens up a number of interesting possible avenues of investigation. In Part II, the technique is used to track the time course of the lexical processing that generates recalibration.

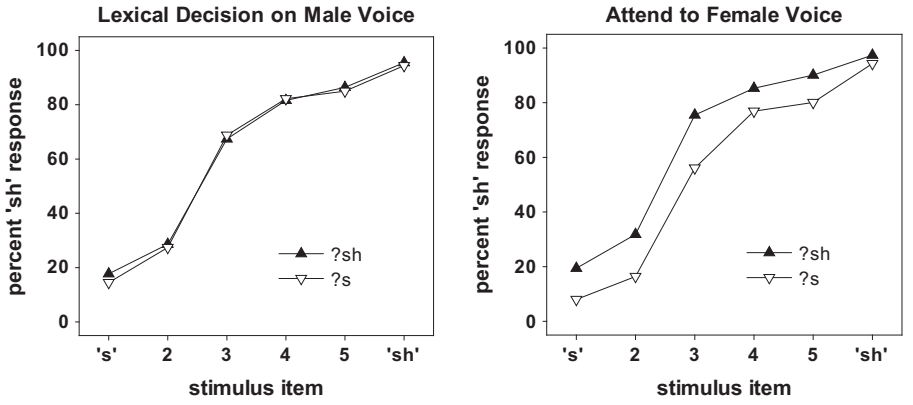

Fig. 2. Average identification of the test series items as "ashee" as a function of whether the listener had previously heard ambiguous "sh" items or ambiguous "s" items. The left panel shows the results when subjects made a lexical decision on a distractor (male word/nonword). The right panel shows the results when subjects counted syllables of the female words.

3. Part II: How long does a word's perception remain malleable?

As discussed in the Introduction, perception is quite non-veridical. One of the examples discussed before is the perception of speech as a series of discrete words, when in fact there are no systematic breaks between the words. The separate words are constructed by perceptual processes, giving the perceiver a cleaned up version of the highly overlapping and messy speech stream. There has been a fair amount of research on how the system imposes boundaries between words – the segmentation problem (e.g., Mattys, White, & Melhorn, 2005). However, there has been much less attention paid to a consequence of the need to construct each word: The developing percept must remain malleable for some period of time while the word is under construction. Moreover, because in many cases the information needed to know that a word has ended does not arrive until some time in the following word, this malleable period often must extend well past the physical end of the word in the signal. And, all of this must be done in a way that hides what is going on from consciousness, as what we hear is a lovely series of discrete words, even though many of these words could not actually be known when they end. This is just an example of our perception of speech being indirect.

Consider, for example, someone hearing the sentence "The police were told to deter more protesters from entering the park". Although the listener presumably hears this sentence perfectly well, with a sense of individual words, it is not possible to be sure that "deter" was in fact a word until some time into "more" because the sentence might have actually been "The police were told to determine how many protesters were in the park". That is, the initial /m/ in "more" could have been part of "determine", rather than the onset of the next word. Thus, the perception of "deter" must have remained malleable, in essence suspended, until enough of "more" was available to exclude "determine" as the word. Because embedding of words is rampant in English and many other languages (e.g., McQueen, Cutler, Briscoe, and Norris (1995), have calculated that over 80% of English words include at least one embedded word), such a delayed closure of the word recognition process must also be the norm, rather than an exception. Yet, as noted, the phenomenology gives no hint of this delay. Instead, it seems that the word construction process takes the time it needs to converge on the correct word and only after this has been accomplished a conscious output becomes available. Note that this means that our perception must be lagging behind the input while information is being incorporated into the malleable percept. If so, then a quite interesting question is: How long does a developing word percept remain malleable, open to further relevant information, before the final percept is "sealed"? Put another way, how long is the window during which words remain under construction? Experiment 2 uses the targeted distraction technique developed in Experiment 1 to address this question.

3.1. Experiment 2

As the "deter more"/"determine" example illustrates, accurate perception depends on collecting information over time. Klatt (1980) articulated this as his "principle of delayed commitment": In general, more accurate perception will be achieved by delaying a commitment to the percept, allowing additional information to be taken in, and more processing of the information to occur. But as he pointed out, at some point the commitment must be made: The percept must become available to the perceiver. Once that percept has become available, later information or computations cannot change it; if they did, we would constantly be "re-perceiving" all of the speech that we hear, which would presumably make speech communication impossible.

In Experiment 2, the targeted distraction technique is used to measure how long the perceptual system delays commitment by systematically increasing the processing time before the distractor is imposed. At some duration, enough time will be available to allow some degree of success in lexical access. Assuming that recalibration occurs when there is some conflict (i.e., the phonetic ambiguity) with the accessed lexical representation, the recalibration effect should start to be seen; at some longer duration the effect should be similar to the effect found with full attention. This longer time is a measure of the processing window's duration – additional time beyond this point does not change the percept, presumably because the system has made a commitment to the lexical percept by that time. As shown in Table 1, the critical stimuli were constructed so that when the listener gets the

ambiguous fricative, there is sufficient information to achieve lexical selection, allowing the recalibration calibrations to occur. We know from a great deal of previous work, using priming measures, eyetracking measures, and other techniques that initial lexical activation of word candidates begins very early, within the first 150–200 ms of a word's onset. The issue being examined here is, given such activation of lexical candidates, when does the system end lexical competition and settle on the word that will become available to the listener? Until that point, the final word choice has not been made and emerging information (either in the signal, or as a result of any ongoing computations) could affect the winning candidate.

How long might this window of malleability remain open? There are some hints in the literature that the period is approximately one second. Connine, Blasko, and Hall (1991) presented listeners with sentences that included an ambiguous segment that created an ambiguous word. For example, because of an ambiguous initial stop consonant, a word might be either "dent" or "tent". When the preceding context was about a car, "dent" tended to be perceived, whereas "tent" was reported more often in the context of camping. The critical result for the current purpose is that the disambiguating context could also follow the critical word, but this only worked if the disambiguation arrived within about one second of the ambiguous word. Similarly, Samuel (1979) measured phonemic restoration effects in sentences and found that context coming within about one second after a word with a sound to restore behaved like context that preceded the critical word; later context was ineffective. Swinney's (1982) work on semantic ambiguity also provides an estimate of malleability of about one second. He found that a semantically ambiguous spoken word (e.g., "bug") primed both of its meanings (e.g., "spy", or "ant") if the to-be-primed target was presented within that time frame, but that with longer delays only the contextually consistent meaning showed priming. These results are not definitive, but they are suggestive of a malleable window lasting approximately one second.

With these findings as a guide, Experiment 2 uses the same targeted distraction technique that successfully blocked recalibration in Experiment 1a, but across five experiments more and more time is given to the system before the interruption is imposed. Specifically, rather than the 200 ms SOA used in the first experiment, the SOAs are: 400 ms (Experiment 2a), 600 ms (Experiment 2b), 800 ms (Experiment 2c), 1000 ms (Experiment 2d) and 1200 ms (Experiment 2e). The goals are to determine (a) the minimum time that is needed to see any evidence of recalibration, and (b) the point at which the recalibration effect starts to approximate the amount of recalibration found without targeted disruption.

3.2. Method

3.2.1. Participants

As noted above, because the expectation was that with short SOAs the targeted distraction would prevent recalibration (given the results of Experiment 1a), null effects were anticipated in those conditions. To be able to distinguish between such a meaningful null effect and one due to a lack of power, much larger sample sizes were employed than in previous work in this domain. Recall that prior work generally used between 30 and 50 participants per experiment, and that in Experiment 1b very robust effects were found using the current version of the paradigm with 77 subjects. The following numbers of participants took part in each experiment: 94 (Experiment 2a), 94 (Experiment 2b), 119 (Experiment 2c), 117 (Experiment 2d), and 96 (Experiment 2e). The subjects were from the same population and met the same criteria as in Experiment 1, only participated in one of the current experiments, and had not participated in Experiment 1.

3.2.2. Stimuli

The two stimulus lists used in Experiment 1a were used in all of the current experiments. The only change was the increase in the SOA across experiments.

3.2.3. Apparatus and procedure

The apparatus and procedure were identical to those in Experiment 1a.

3.3. Results and discussion

The data were treated as in Experiment 1. Using the same exclusion criteria, the numbers of subjects eliminated were 14 (Experiment 2a), 12 (Experiment 2b), 22 (Experiment 2c), 15 (Experiment 2d), and 13 (Experiment 2e). The data for the remaining subjects were analyzed as before.

Before we consider the recalibration effects as a function of the processing time given to the subject before the targeted disruption, it is worth making sure that performance on the distracting task – lexical decision on the male voice – was as it should be. Table 2 shows the accuracy and reaction time results for the words and pseudowords for the five SOA conditions of Experiment 2, as well as those for the comparable condition in Experiment 1a (SOA 200). The results are in fact as one would expect. Accuracy was very high for all six SOA cases, averaging over 96% correct. As is typical of lexical decision, subjects were consistently better at recognizing words than they were at rejecting pseudowords, an effect that (given the very high accuracy) was primarily visible in the reaction time data. The word advantage was statistically robust for each SOA in the reaction time data, and was reliable in the accuracy data for all but the longest SOA (where it was marginally significant). Finally, as the SOA was increased, reaction times decreased. This would be expected for two reasons. First, there is less and less overlap between the male and female tokens as the SOA gets longer, reducing any masking of the male token and thus speeding responses. Second, the onset of the female token essentially provides a warning signal to the subject, allowing the subject to predict the onset time and ear of the male token. As more and more time is given after that warning signal (i.e., as the SOA increases), subjects can be more ready for the male token, reducing response times. Note that if subjects use the "warning signal" to shift their attention to the male token (their assigned task), and if attention is needed for the recalibration process, then this would *reduce* the recalibration shift with increasing SOA. As will be seen shortly, the actual recalibration results are exactly the opposite of this, and as such cannot be attributed to some task-specific artifact.

Focusing now on the recalibration data, recall that in Experiment 1a, the targeted distraction with an SOA of 200 ms completely eliminated any hint of recalibration, with a 0.2% shift. When processing was interrupted after 400 ms in Experiment 2a, the group exposed to "s" ambiguities produced 59.2% "sh" report, while the group exposed to "sh" ambiguities produced 61.8% "sh" report. This 2.6% shift was still indistinguishable from zero, $F(1,78) = 0.411$, n.s. In Experiment 2b, with a 600 ms SOA, the shift doubled to 5.4%, based on 58.2% "sh" report for the "s" group, and 63.6% "s" report for the "sh" group, but the recalibration was still not reliable, $F(1,80) = 1.734$, $p = .192$. The size of the shift stayed in the same range for the 800 and 1000 ms SOA conditions, but the reliability clearly increased. In Experiment 2c, the "s" group (59.8% "sh") and the "sh" group (65.8% "sh") produced a marginally significant 6.0% recalibration effect, $F(1,97) = 3.515$, $p = .064$. Similarly, in Experiment 2d, the 5.7% difference between the "s" group (58.4% "sh") and the "sh" group (64.1% "sh") was marginally significant, $F(1,99) = 3.398$, $p = .068$. It was only when the SOA was extended to 1200 ms in Experiment 2e that the recalibration was both large and fully reliable. The "s" group (61.3% "sh") clearly differed from the "sh" group (72.4% "sh"), with the 11.1% difference being statistically robust, $F(1,81) = 10.561$, $p = .002$.

Table 2
Lexical decision task accuracy and reaction times.

SOA	Word Accuracy (%)	Pseudoword Accuracy (%)	Word RTs (ms)	Pseudoword RTs (ms)
200	97.4	95.5	1035	1173
400	96.8	93.7	979	1143
600	98.2	93.7	974	1131
800	97.4	96.4	965	1051
1000	97.8	95.7	943	1039
1200	97.8	95.4	884	948
Mean	97.6	95.1	963	1081

It is useful to look at the observed effects for the six SOA conditions in terms of how the effect size changes as more and more processing time is available before the targeted disruption is imposed. Omega-squared (ω^2) is a measure of effect size that allows one to characterize effects as being "small" (ω^2 of .01), "medium" (ω^2 of .06), or "large" (ω^2 of .15) (Keppel & Wickens, 2004). The pattern of shifts in terms of effect size is quite systematic. For the 200 ms SOA (Experiment 1a) and 400 ms SOA conditions, the observed *F*-ratios of less than 1 yield effective ω^2 values of zero – there is no effect when lexical processing is disrupted so early. For SOAs of 600, 800, and 1000 ms, there is a small effect, with ω^2 values of .01, .03, and .02, respectively. When the SOA was extended to 1200 ms, the effect size of .10 reached the medium-large range. For comparison, in Experiment 1b, when listeners were directing their attention to the female stimuli (to count the number of syllables), the effect size was .13, approaching a large effect.

Using this effect size breakdown, Fig. 3 shows the identification functions for the six SOA values broken down by whether there was no effect (SOAs 200 and 400), a small effect (SOAs 600, 800, and 1000), or a medium-large effect (SOA 1200). Even with an extremely large sample size (163 subjects), the "no effect" case was just that, $F(1,161) = 0.560$, n.s.; the effect size remains zero. In contrast, the combined results for the three cases that produced small effect sizes become quite reliable, $F(1,282) = 8.418$, $p = .004$; the effect size remains small, with ω^2 of .026, but the huge sample allows the reliability of this small effect to be seen clearly.

The targeted distraction technique thus functioned exactly as desired by halting processing after successively longer windows during which the processes underlying lexically driven recalibration could operate. This method allows us to see how much time is needed to complete the lexical access process, assuming that such lexical support is essential for successful recalibration (Norris et al., 2003; also see Experiment 5 below). Given that substantial recalibration did not emerge before the 1200 ms SOA condition (though significant effects did start to appear after a half second), it appears that lexical processing continues for over a second after word onset.

In fact, even the effect at 1200 ms SOA (a shift of 11.1%) was smaller than the "attend to the female voice" condition in Experiment 1b (13.6%) and the "full attention" case from Kraljic and Samuel (2005) (17.5%). This suggests that we would need to bring the SOA out even further, perhaps to 1500 or 1600 ms, before the observed shift would match these attended cases. Such a timeframe is actually longer than the window hinted at in prior studies (Connie et al., 1991; Samuel, 1979; Swinney, 1982), which suggest that perception remains malleable for about one second. One possible explanation for this apparent difference will be examined in Experiment 4. This issue notwithstanding, Experiment 2 has clearly confirmed the conclusion of Experiment 1 that recalibration is not automatic – its success varies systematically with the amount of processing time provided before the targeted disruption. Some recalibration can be produced when between a half second and one second of uninterrupted processing is available, but full calibration appears to take longer.

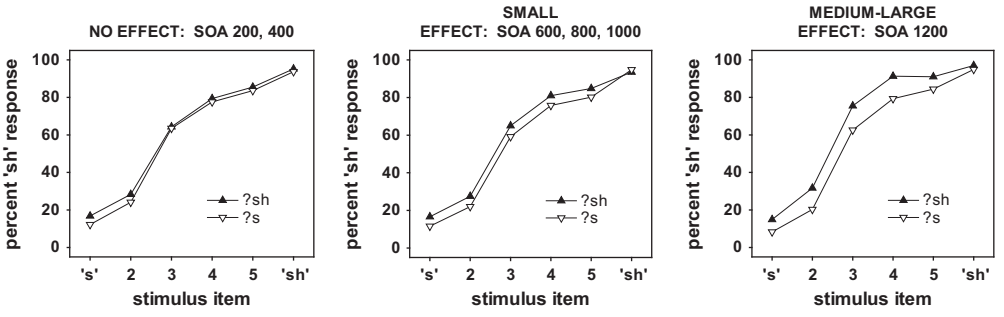

Fig. 3. Average identification of the test series items as "ashee" as a function of whether the listener had previously heard ambiguous "sh" items or ambiguous "s" items. The left panel shows the average results for conditions with an effect size of zero (SOA 200 and 400). The middle panel shows the average results for conditions that yielded a small effect size (SOA 600, 800, and 1000). The right panel shows the results for the condition that produced a medium-large effect size (SOA 1200).

4. Part III: What is the nature of the disruption, and when does it begin?

The first two experiments have shown that recalibration requires attention because it depends on lexical processing that requires attention, and that lexical processing can remain active for a relatively long period of time. Part III includes three experiments that are designed to clarify (a) the type of attentional resources that are being tapped, (b) the onset timing of the recalibration process, and (c) the link between recalibration and lexical access.

4.1. Experiment 3

The task used to impose targeted distraction in the first two experiments was auditory lexical decision. This task was chosen as a natural way to impose competition with the lexical processing needed to generate recalibration. The use of this task leads to the question of whether such specific competition plays a necessary role in the targeted distraction.

Experiment 3 addresses this question by replicating the test conditions of Experiment 1a, but with a very different distracter task. As in Experiment 1a, there is a 200 ms SOA between the onset of the female word (carrying the critical segmental ambiguity) and the onset in the opposite ear of the distracter that the subject must respond to. However, in Experiment 3, the distracters are not words. In fact, they are not even speech. Rather, on each trial the distracter is an environmental sound, such as a doorbell or a dog barking. The subject's task on each trial is to judge whether the sound comes from something living (e.g., the dog barking) or nonliving (e.g., the doorbell). If the halting of lexical processing by targeted distraction in Experiments 1 and 2 was due to the distracter task's need for lexical processing, then the nonlexical task in Experiment 3 should leave recalibration intact. If instead the disruption traces to a more general need for attentional resources, then we should see a similar absence of recalibration when subjects make judgments about the nonspeech sounds in the current experiment.

4.2. Method

4.2.1. Participants

100 subjects from the same population, meeting the same criteria as in Experiments 1 and 2, participated in Experiment 3. They had not participated in the previous experiments.

4.2.2. Stimuli

There were two stimulus lists, as in the previous experiments, with one list including ambiguous segments in "s" words, and the other including the ambiguous segments in "sh" words. However, rather than randomly pairing words and pseudowords from a male speaker with each of the female words, environmental sounds were randomly paired with them. The environmental sounds were edited versions of sounds used in prior studies (Gregg & Samuel, 2008, 2009; Pufahl & Samuel, 2014). The male words had lengths that generally ranged between 500 and 900 ms, with a mean of a little under 700 ms. The environmental sounds were edited to be of similar lengths (mean of 804 ms). Each environmental sound began 200 ms after the onset of a female word.

4.2.3. Apparatus and procedure

The apparatus and procedure were identical to those in Experiments 1 and 2. During the exposure phase, subjects were told that they would hear a word and a sound on each trial, and that their task was to decide whether the sound on a given trial came from a living thing or from a non-living thing. They responded by pushing labeled buttons ("living", "non-living") on the button boards. The test phase, with the "asee" – "ashee" test items, was identical to the previous experiments.

4.3. Results and discussion

The data were treated as in Experiments 1and 2. Using the same exclusion criteria, the data from 14 subjects were eliminated. The central question is whether recalibration will occur when the distracter task is nonlexical. As Fig. 4 shows, there is a very weak trend toward recalibration. Subjects exposed to the ambiguous segments in "s" words reported 62.0% of the "asee" – "ashee" test items as "sh", while those exposed to the segments in "sh" words reported 63.9% "sh", a shift of 1.9%. This trend did not approach significance, $F(1,84) = 0.326$, n.s. Like the SOA 200 and SOA 400 conditions with lexical decision providing the targeted distraction, the F-ratio under 1.0 leads to an effective ω^2 of zero.

While null effects should always be considered cautiously, it is important to keep in mind that this test included about twice as many subjects as have typically been used in experiments of this sort. Thus, a lack of power is very unlikely to account for the absence of recalibration. In principle, another possible cause of a null effect could be masking of the female words by the environmental sounds. The dichotic presentation makes this quite unlikely. Further evidence against a masking account comes from an acoustic analysis of the spectral overlap of the sounds and the words, compared to the same analysis of the male items and the female words in the previous experiments. Cooke (2006) developed "glimpse" analyses to assess how much of a speech signal is available, given masking by another sound (the "glimpses" are what remain of the speech, after the masking). The remains are expressed as percentages of the original signal. Glimpse analyses of the speech stimuli in Experiment 1a show that the remaining signal for the female words averages 46.2% of the original. The comparable analysis with the environmental sounds yields 47.4% remaining. The essentially identical remainders, together with the dichotic presentation (the glimpse analyses return what would remain if the signals were actually mixed), provide no support at all for a masking account of the impact of the environmental sounds on recalibration. Thus, at the very least, it seems safe to conclude that the nonlexical targeted distraction was largely effective in preventing recalibration from occurring. This indicates that recalibration requires attentional support, and that the needed resources are not specific to lexical processing.

4.4. Experiment 4

Recall that the results of Experiment 2 suggested that even the SOA of 1200 ms was not yet giving the system enough uninterrupted time to complete the lexical processing needed to produce recali-

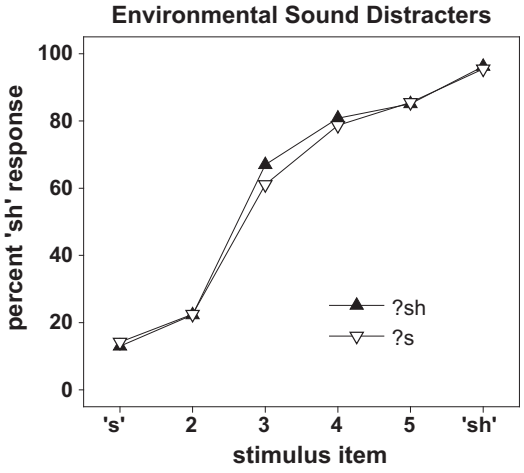

Fig. 4. Average identification of the test series items as "ashee" as a function of whether the listener had previously heard ambiguous "sh" items or ambiguous "s" items. The identification test followed an exposure task in which listeners heard environmental sounds, rather than words.

bration effects comparable to the full attention cases – a rough extrapolation suggested that perhaps with a 1500–1600 ms SOA, more or less, the full effect might be seen. This estimate of the processing window for a word is longer than the suggestions in prior work of about a one second window.

However, in the current study, it is not obvious where one should "start the clock" on the recalibration processing. Using the SOA between the female word carrying the critical ambiguity and the point of interruption seems reasonable, but so are at least two other possibilities. One alternative would be to assume that recalibration can begin as soon as the necessary information has been received. Because the ambiguous segments were designed to occur at or after a word's lexical uniqueness point (see Table 1), the necessary information is available when the ambiguous segment has been heard. Another alternative is that recalibration processes can begin when the carrier word has been fully presented, including the ambiguous segment. These alternatives are of course very highly correlated, but differ by about 400 ms because the ends of the critical segments were on average about 500 ms into words that were on average about 900 ms long (see Table 1 and Fig. A.1 in Appendix A). If what matters is either the end of the word, or the end of the critical segment, the actual window would be much shorter than the estimate based on SOAs.

Because of the way that the data patterned in Experiment 2, there is a way to test whether the SOA is what matters in defining the window, or if instead the time from the end of the carrier word (or its tight correlate, the location of the critical segment) is what matters. The critical result that enables the test is the large jump in the amount of recalibration between the SOA 1000 condition (a shift of 5.7%, and a small effect size, $\omega^2 = .02$) and the SOA 1200 condition (a shift of 11.1%, and a medium-large effect size, $\omega^2 = .10$). The test in Experiment 4 leverages this pattern by creating trials that maintain the SOA of the SOA 1000 case, but that simultaneously offer the listener the same amount of post-carrier-word time before interruption that was available in the SOA 1200 case. Observing a small shift with these stimuli would suggest that SOA is the relevant factor, whereas finding a large shift would favor the post-carrier-word or post-critical-segment interpretation.

The way that this can be done is by compressing the duration of all of the carrier words by 200 ms, and then using them in the SOA 1000 condition. Fig. 5 provides an example of how this works. The top half of the figure shows a pair of waveforms from the SOA 1000 condition that was tested in Experiment 2d – the male token's onset is at 1.0 s, relative to the female token beginning at 0.0 s. Note that there is about 150 ms of silence after the female word ends before the male distracter begins. The bottom half of Fig. 5 shows the same female and male items, but the (leading) female item is now 200 ms shorter than before (about 650 ms, rather than 850 ms). Because the female and the male tokens begin at exactly the same times that they begin in the top panel, the SOA is still 1000. Critically, however, in the compressed condition the female word now ends at the same time relative to the onset of the male token as it did in the SOA 1200 case of Experiment 2e. One way to think about this is that the SOA 1200 case could be constructed by taking the female token in the top panel and sliding it to the left by 200 ms (so that it would now start at −0.2 s, and would now end 200 ms sooner than what is shown for this case).

The "short 1000" condition thus preserves the SOA 1000 onset-to-onset timing but offers the system the same extra time after the female token that the SOA 1200 condition offered relative to the SOA 1000 condition. If the total time available for processing the word is what matters, then matching SOA is key, and the results for the "short 1000" case should be similar to what was found in Experiment 2d, a shift of about 5.7%. If instead what matters is processing time after the carrier word is recognized (using its offset as an approximation), then the "short 1000" condition should produce a much larger shift, more like the 11.1% effect found in Experiment 2e.

4.5. Method

4.5.1. Participants

96 subjects from the same population, meeting the same criteria as in Experiments 1–3, participated in Experiment 4. They had not participated in the previous experiments.

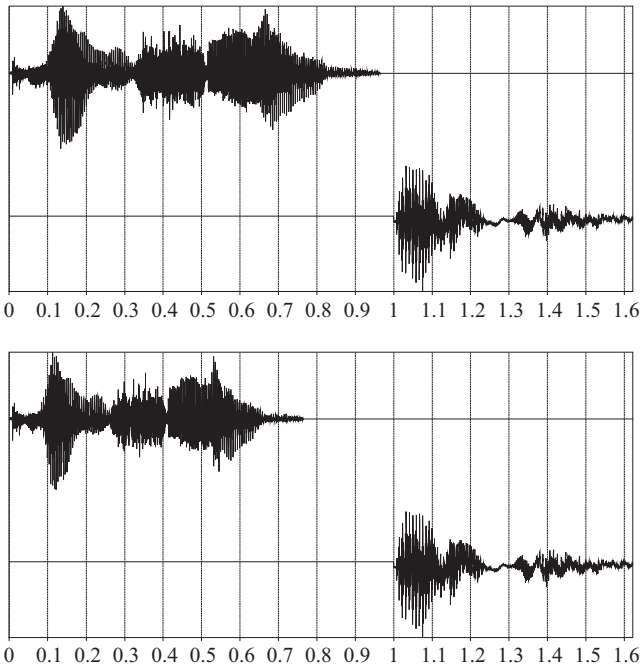

Fig. 5. Waveforms for two dichotic trials, with the female item (top waveform in each pair) beginning 1000 ms before the male item (bottom waveform in each pair). Time (in seconds) is shown along the *x*-axis of each panel. Note that although the SOA is 1000 ms in both cases (the bottom waveform starts at the same time relative to the top waveform in each pair), the amount of silence is 200 ms longer in the bottom pair than in the top pair, due to the compression of the female token in the lower panel.

4.5.2. Stimuli

There were two stimulus lists, as in the previous experiments, with one list including ambiguous segments in "s" words, and the other including ambiguous segments in "sh" words. The lists were identical to those used in Experiment 2. However, all 100 female words were compressed by 200 ms, using the speech compression option in GoldWave that preserves the pitch of the speech. Because the algorithm is well designed, and because the compression was not severe, the resulting stimuli sounded quite natural.

4.5.3. Apparatus and procedure

The apparatus and procedure were identical to those in the previous experiments.

4.6. Results and discussion

The data were treated as in Experiments 1–3. Using the same exclusion criteria, the data from 10 subjects were eliminated. The middle panel of Fig. 6 shows the results. For comparison, the left panel shows the results for the SOA 1000 condition of Experiment 2d, and the right panel shows the corresponding results for the SOA 1200 condition of Experiment 2e.

The question posed in Experiment 4 is whether the "1000 short" condition created by compressing the female carrier words produces effects more like the small effect of the SOA 1000 case, or the medium-large effect of the SOA 1200 case. The results are unambiguous: The compressed stimuli produced a large shift (14.8%), based on 56.0% "sh" report for the subjects in the "s" condition, versus 70.8% "sh" report for the subjects in the "sh" condition. This difference was extremely robust, $F(1,84) = 13.707$, $p < .001$. Looking at the three cases in terms of effect size, the new condition's effect

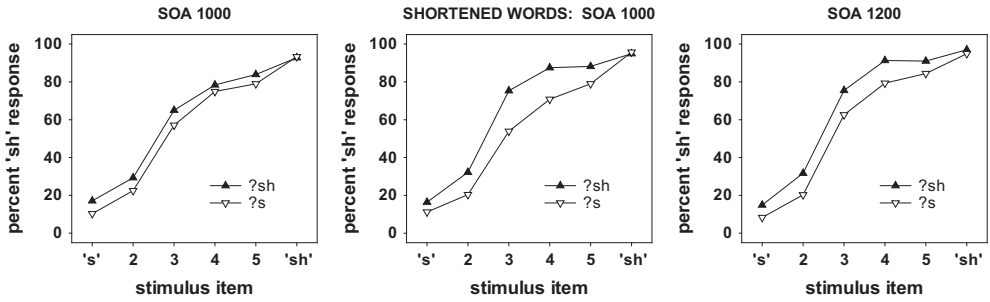

Fig. 6. Average identification of the test series items as "ashee" as a function of whether the listener had previously heard ambiguous "sh" items or ambiguous "s" items. The center panel shows the results of Experiment 4. For comparison, the left panel shows the results for stimuli matched by SOA (Experiment 2d), and the right panel shows the results for stimuli matched by time from offset of the female word until onset of the male token (Experiment 2e).

size (ω^2 = .13) is clearly much closer to the ω^2 of .10 for the SOA 1200 case than to the ω^2 of .02 for the SOA 1000 case.

The large recalibration effect found for the compressed carrier word stimuli indicates that the computations involved in recalibration do not start when the word begins (as an effect tied to SOA would imply), consistent with the findings of Jesse and McQueen (2011). Rather, the time being used for these computations begins at the end of the word, or after the critical segment (the current data do not distinguish these two possibilities, but they are highly correlated and only differ by a few hundred ms). This window makes sense, as it suggests that once the system is using lexical information to resolve the phonetic ambiguity, recalibration is a natural consequence. That resolution cannot begin until the ambiguity has been encountered, and sufficient lexical information has been received to guide the resolution and begin the recalibration process.

4.7. Experiment 5

In Experiments 1–4 various manipulations were used to affect the time available to the listener to recognize the female tokens that contained the critical ambiguous fricative segments. Collectively, the results clearly show that if processing is interrupted relatively early, recalibration is reduced or eliminated. This pattern was used to argue that there is a window of time during which lexical processing operates, with the final lexical output remaining malleable during this window. The size of the recalibration effect was taken to be a consequence of the degree to which lexical access finished.

An alternative possibility is that the interruption of processing is not of lexical access, but is instead an interruption of the recalibration process more specifically. That is, perhaps what was being measured in Experiment 2 was the time needed to complete the computations for recalibration itself, rather than the time for the lexical access assumed to be necessary to drive the recalibration process. Experiment 5 uses the same exposure procedure as in the previous experiments, but includes a following test that is intended to measure lexical processing more directly. The idea is to see whether this measure of lexical access follows the same pattern as what we saw for recalibration. If so, that would suggest that it is the degree of lexical access that was responsible for the recalibration results.

The task that follows the exposure test here is a medium term repetition priming test. Many studies have shown enhanced processing of a word that had recently been heard: If a listener hears "dog", and then 10 or 20 minutes later "dog" is presented for a lexical decision, reaction time will be faster than if "dog" had not been recently experienced (e.g., Bowers, 2000; McLennan, Luce, & Charles-Luce, 2003; Sumner & Samuel, 2007). This priming task thus provides a measure of recent lexical access. In Experiment 5, it is used to measure how well lexical access succeeds in the exposure task that has been used in the previous experiments. For example, we observed almost no recalibration when listeners were interrupted 400 ms after the female word began, whereas we saw very strong recalibration when there was a 1200 ms SOA. In Experiment 5, we can test whether medium term repetition

priming is correspondingly small or absent for words that are interrupted with a 400 ms SOA, but present for words that are interrupted with a 1200 ms SOA. If the recalibration effects that we have observed are in fact a consequence of interrupting the lexical support that recalibration requires, then the medium term repetition priming effects should follow the same pattern as the recalibration effects. Three specific hypotheses are tested: (1) For a short SOA condition (400 ms) that produced no recalibration, no medium term priming should be found. (2) For a long SOA condition (1200 ms) that produced robust recalibration, robust medium term priming should be found. (3) For an intermediate SOA (800 ms) that produced marginal recalibration, marginal medium term priming should be found.

4.8. Method

4.8.1. Participants

36 subjects from the same population, meeting the same criteria as in Experiments 1–4, participated in Experiment 5. They had not participated in the previous experiments. Because the priming paradigm is more sensitive than the recalibration, and because the priming effect can be measured within-subject, a much smaller sample size can be used than in the recalibration experiments.

4.8.2. Stimuli

The stimuli for the initial exposure task were based on those used in Experiment 2, except that no ambiguous fricatives were presented – the original forms of the critical words were used, with their original (unambiguous) fricatives. The 100 trials were divided into four sets of 25 trials, with each set of 25 including 5 "s" words, 5 "sh" words, and 15 of the filler words; the distribution of word length in each set of 25 was similar. For a given subject, one set of 25 trials was presented with the 400 ms SOA, one set was presented with the 800 ms SOA, one set was presented with the 1200 ms SOA, and the fourth set was not presented. Using a Latin Square, the four stimulus sets were rotated through these four conditions so that each set was presented equally often with a given SOA. These four presentation sets were then distributed, using another Latin Square, across three presentation orders so that the short, medium, and long SOA conditions were presented equally often as the first, second, or third block. These two Latin Squares yielded 12 possible presentation lists; three participants were assigned to each list.

A nonspeech filler task followed the initial exposure task. The stimuli were eight short excerpts of familiar melodies played on a piano. Each excerpt was about 8–15 s long. The melodies had been subjected to signal processing (local time reversal) to make them more difficult to recognize. Each melody was played twice, after which subjects used a four-point scale to indicate how recognizable the melody was.

The final task was an auditory lexical decision task. The stimuli in the lexical decision task were the 100 female words (75 of which had been presented in the exposure task) and 100 nonwords produced by the same speaker (like the words, taken from Kraljic and Samuel (2005)).

4.8.3. Apparatus and procedure

The apparatus was identical to that in the previous experiments. The exposure task was broken into three blocks of about two minutes each, with SOA fixed within a block. There was a short break (approximately 1–2 min) between blocks. The task was the same one used in Experiment 2 – subjects were told to make a lexical decision for the male words and pseudowords that were presented in the ear opposite to the female words; as before, no judgments were made on the female words. The melody task followed the exposure task after a couple of minutes needed to give instructions for the task. The task itself took 4–5 min. With the instructions, the filler task, and the instruction time for the final task, there was about a 10 min separation of the end of the exposure task from the beginning of the final lexical decision task. Subjects were told that on the final task they were to judge whether each (female) token was a real English word or not. The final task took 6–7 min.

4.9. Results and discussion

The critical predictions for Experiment 5 pertain to the final lexical decision task: If the pattern of recalibration effects that was found in Experiment 2 reflects interruption of lexical processing then we should see a corresponding pattern of long term priming as a function of the timing of the targeted disruption. As noted above, this means that there should be little or no evidence of lexical activation for the 400 ms SOA exposure case, whereas there should be evidence of strong lexical activation for the 1200 ms SOA case. The 800 ms SOA condition was included as an intermediate case, as we observed marginal recalibration shifts for this SOA.

Overall, 33 of the 36 subjects performed very accurately on the final lexical decision task, averaging 95.3% correct on real words and 94.8% correct on nonwords. Three subjects were quite inaccurate on nonwords (accuracies of 71%, 74%, and 75%, respectively) and were replaced.

Medium term priming was assessed by comparing the average reaction time for previously experienced words to those that a given subject had not heard during the exposure task. Our first prediction was that words that had been heard in the long (1200 ms) SOA condition would produce robust repetition priming. This prediction was confirmed: Subjects were 21 ms faster to recognize words that had been heard in the 1200 ms SOA condition (936 ms) than words that they had not been exposed to (957 ms), $F(1,35) = 5.514$, $p < .03$. The second prediction was that little or no priming should be seen for words that had been interrupted quickly – those in the 400 ms SOA condition. This prediction was also confirmed, with virtually identical response times for the experienced (956 ms) and the unheard ones (957 ms), $F(1,35) = 0.040$, n.s. Finally, for the 800 ms SOA that had produced marginal recalibration effects, we found a 17 ms repetition priming trend, with an average response time of 940 ms, $F(1,35) = 2.457$, $p = .126$.

Fig. 7 provides a visual comparison of the phonetic recalibration and the lexical recalibration effects for the three SOA conditions. As the figure suggests, the patterns are indeed similar. For the 400 ms SOA case, both recalibration and priming had *F*-ratios under 1.0, with both thus having an ω^2 of zero. For the 800 ms SOA, the recalibration ω^2 was .03, and the lexical priming ω^2 was .02; in both cases, a small effect. Finally, for the 1200 ms SOA, the recalibration effect size was .10, versus .06 for the priming measure, a medium-large versus a medium effect size. Given the different sensitivities of the two tasks, and the different units of measurement, there is no simple way to compare the results across tasks (e.g., it would not make sense to compute any measure of a linear relationship). However, it does make sense to ask whether exposure conditions that afford lexical access also produce recalibration, and they do; similarly, one can ask if conditions that prevent lexical access also show no recalibration. Again, the results are consistent. This was the logic that generated the predictions for Experiment 5, and the results are what would be expected if the blocking of recalibration is a consequence of a blocking of lexical access.

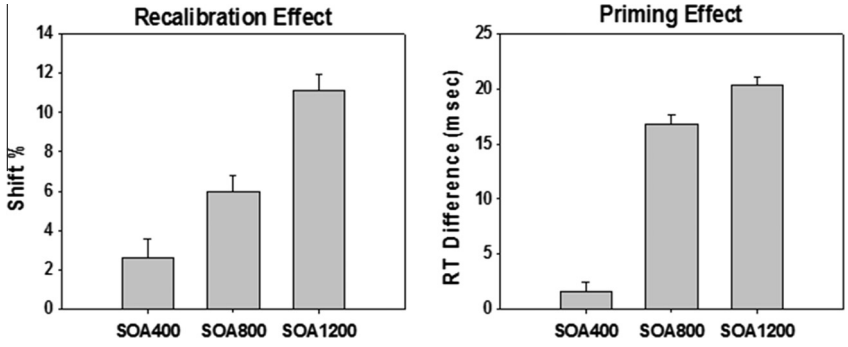

Fig. 7. A comparison of the size of the observed recalibration shifts and the lexical priming effects, for the short (400 ms SOA), medium (800 ms SOA), and long (1200 ms SOA) conditions. The error bars represent standard errors of the means.

5. General discussion

As noted in the Introduction, for over a half century, research on speech perception and spoken word recognition has focused on the high variability of the signal, and how the perceptual system might cope with this variability. One potentially important coping mechanism is perceptual adjustment to the variability. It has been known for a very long time that our perceptual systems, in many domains, scale their responses to take into account what the prevailing input has been. For example, after an observer looks at a red display for a while, a white field will appear green – an adaptation that reduces the influence of the prevailing wavelengths. This type of adjustment is critical in maintaining perceptual constancy across different environments, such as walking into a house lit with traditional incandescent bulbs. Although this drastically changes the wavelengths reflecting off of the surfaces (e.g., the person's clothes) because the incandescent lighting is heavily weighted toward longer wavelengths, color constancy is largely maintained by down-weighting those longer wavelengths.

A comparable adaptation effect for speech was first reported by Eimas and Corbit (1973), who showed that repeated exposure to one speech sound reduced the listener's likelihood of perceiving sounds like that one. As more was learned about this speech adaptation phenomenon, several investigators tested whether this type of adjustment occurred automatically, or if instead it required attention. These tests all used some type of concurrent task to see whether the adaptation effect was reduced when such a concurrent load was tying up attentional resources, and in all cases the adaptation shifts were unaffected by the manipulations (Mullennix, 1986; Samuel and Kat, 1998; Sussman, 1993). Three decades after selective adaptation of speech was first studied, two additional adjustment mechanisms were discovered. Norris et al. (2003) found that listeners retune their phoneme category boundaries when lexical context directs the disambiguation of a phonetic segment. Bertelson et al. (2003) observed the same kind of recalibration, in this case with the disambiguation being provided by unambiguous lip movements accompanying the ambiguous speech. As in the case of adaptation, researchers tested whether these adjustments were automatic or if they required attention, using the same dual-task methods. And, the same null effects on the shifts were found, both for the lexical case (Zhang & Samuel, 2014) and for the audiovisual case (Baart & Vroomen, 2010).

The current study was undertaken as a result of the limitation of the dual-task method acknowledged by the authors of these studies: It is impossible to know if a subject is defeating the experimenter's manipulation by engaging in some kind of time-sharing across the two tasks. The approach here was to use a technique that is essentially a descendent of the postmasks used in research on visual word recognition. Such postmasks were designed to halt whatever visual processing might have begun, up to the moment when the mask was presented. A quite interesting property of masking in that domain was that its effectiveness depended on properties of the mask itself. For example, different results were found for masks that consisted of simple dot patterns, or just bright light, compared to "patterned postmasks", with the latter made up of line segments and angles. As McClelland and Rumelhart (1981) argued, the patterned postmasks could engage many different letter patterns, effectively wiping out activation of any particular letter, but these masks would not engage word patterns. This accounted for better performance at recognizing a letter when it was part of a word than when it was presented in isolation.

In the current study, the "postmasks" were the stimuli that accompanied the female words and their critical ambiguous segments. Experiment 1 established that this procedure works very well in the domain of spoken word recognition by demonstrating completely different results under conditions that were physically identical but attentionally distinct. In particular, because recalibration was intact when listeners were given a task that caused them to attend to the critical female words, we can be confident that the loss of recalibration was not due to masking. The complete abolition of recalibration when listeners attended to the male token that began 200 ms after the female word provides powerful evidence for the role of attention in recalibration. Indeed, these results are completely counter to what was found with the dual-task procedures used in previous research, making a time-sharing explanation in those studies most likely.

It is useful to treat the two conditions of Experiment 1 as two poles, with the abolition of recalibration caused by targeted distraction at one end, and the intact recalibration for attended stimuli at the

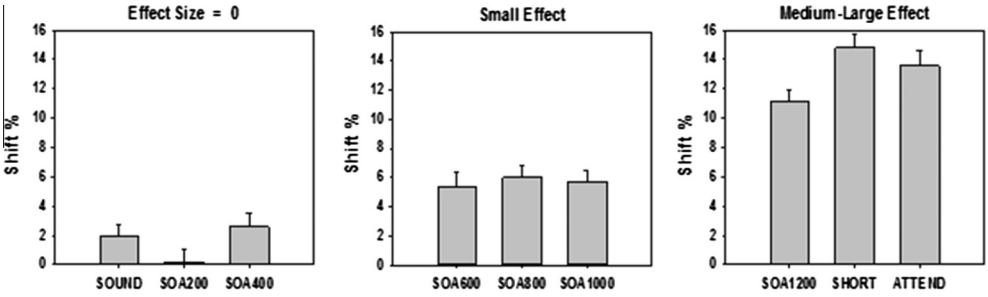

Fig. 8. Summary of the recalibration shifts, broken into three categories. Left: $\omega^2 = 0$; middle: $\omega^2 = .01–.03$; right: $\omega^2 = .10–13$. The error bars represent standard errors of the means.

other. Fig. 8 summarizes the results of the nine sub-experiments using these two cases as anchor points. Conceptually, the nine cases break into three outcomes: cases with an effect size of zero, cases with small effect sizes (.01–.03), and cases with medium-large effect sizes (.10–.13). The dominant factor in whether a condition falls into one of these three sets is obviously the timing that was manipulated across experiments: The two shortest SOA conditions belong to the no-effect group, and the longest SOA condition is in medium-large effect set, with the three intermediate SOA cases making up the small-effect group. This pattern demonstrates that the targeted distraction manipulation did exactly what it was intended to do, serving a role analogous to the postmasks in visual word recognition research.

Given this partitioning of the experiments, it is instructive to see how the three "non-standard" tests grouped, as each of these bears on an important theoretical question. The "attend to the female voice" experiment is in fact the pole representing the set of medium-large recalibration effects. As noted above, the theoretical implication of this experiment grouping here is that attentional allocation is critical in observing recalibration: Two conditions that were physically identical produced fundamentally different outcomes as a function of the attentional instructions.

The fact that the environmental sound experiment groups with the no-effect cases addresses the issue of the type of attentional resources that are required for recalibration to proceed. Recall that in this experiment the targeted distraction did not involve speech sounds – the listeners did not need to engage lexical processing, as they did in the other experiments. Nevertheless, recalibration was preempted. This result indicates that the type of processing done during word recognition that yields recalibration is not being done in some isolated, independent part of the perceptual system.

The third "non-standard" experiment was the "short 1000" manipulation in which the female words were compressed by 200 ms, maintaining the same onset to onset time as in the standard SOA 1000 case, but giving listeners the same offset to onset time as in the SOA 1200 condition. This experiment took advantage of the big jump in recalibration between the SOA 1000 and SOA 1200 experiments. The short 1000 condition clearly belongs to the set of experiments that produced medium-large effects, a result that provides important information about the window during which lexical processing is taking place. Given some reasonable assumptions, the duration can be estimated as being about one second.

This estimate comes from first extrapolating that a full recalibration effect would need a bit more time than the SOA 1200 case provided (perhaps around 1500–1600 ms), given that previous experiments using these stimuli without targeted distraction found somewhat larger effects (Kraljic & Samuel, 2005). Then, the results of the short-1000 experiment indicate that the critical time window should begin either at word offset (rather than the onset implied by using SOA), or more plausibly with the highly correlated point at which a listener has enough information to recalibrate – when there is sufficient information to access the correct word and to have recognized the ambiguous segment. The stimulus information in Table 1 (also see the figures in Appendix A) shows that on average this point should be reached about 500 ms after word onset. If we subtract out this 500 ms from the 1500 to 1600 ms SOA that we have extrapolated for full recalibration, we get a processing window of 1000–1100 ms, or about one second. Note that this window reflects the point at which additional

uninterrupted processing no longer increases the amount of recalibration. This does not mean that lexical access requires 1000 ms – we know from many previous studies that lexical activation begins much sooner than this. The one second period is an estimate of the point at which the system normally commits to a lexical candidate. The duration reflects the balance that must be struck between delaying commitment to optimize performance (Klatt, 1980) and the need to keep up with incoming speech.

It should be clear that saying that there is a processing window of about one second is not a suggestion that every word will take exactly one second to process. Even in the absence of such an overly ambitious claim, the results do provide a reasonable degree of specificity. As an analogy, in George Miller's classic paper (Miller, 1956), the title referred to "the magical number seven, plus or minus two". Miller was quite explicit in saying that there was in fact a particular range of short term memory capacity: It did not hold one or two items, nor did it hold 19 or 20. At the same time, his estimate of seven allowed about a 25% deviation. The suggestion here of a lexical processing time window of about one second should be taken in the same spirit: Lexical processing does not conclude within 200 ms, nor does it continue for four or five seconds. If we adopt the 25% deviation range that Miller suggested, the claim of "about one second" means that under the conditions being looked at, lexical processing finishes around 750–1250 ms after the necessary lexical information is available. A 250 ms range around the average is in fact consistent with the variation in the stimuli used in the current study: For all 40 of the critical words (the female "s" and "sh" words), the time between the onset (or offset) of the critical fricative and the end of the word is within 240 ms of the average time difference between the fricative's location and word offset; for over 90% of the items, the variation is less than 160 ms from the mean (see Appendix A for the distribution of individual item values). The consistency of this timing means that the estimate of the average processing time (about a second) is not undercut by large variations in the point at which listeners will have the necessary information for lexical access and will have recognized the ambiguous segment.

Framing the processing window this way raises an intriguing question: If the word must be recognized along with the ambiguity before recalibration begins, then what is going on in this window? It is not word recognition in a simple (or perhaps simplistic) sense if recognition must precede recalibration – only by knowing the word does the listener know that the ambiguous segment should have been "s", or should have been "sh". This conception treats lexical access as too black-and-white. Recall the basic reason that some window of processing is needed in the first place, which is that the word's identity is often not knowable with certainty until well after the physical ending of the word itself (the "deter more" versus "determine" problem). This reality calls for a model of lexical access that is gradual, rather than all-or-none. This type of gradual emergence of the correct candidate is what underlies the activation metaphor, with various lexical candidates each enjoying increased activation as a function of the support they receive from the input. With this conception, a word need not be fully and finally recognized before it can affect both phonetic encoding and recalibration. It need only be active enough to produce these effects. From this perspective, both the phonetic encoding and word recognition processes occur over time during a processing window of about one second, with mutual resonance leading to a convergence on both the correct phonetic sequence and the correct word. The results of the current experiments indicate that recalibration is a natural consequence of this resonant period. As Experiment 5 demonstrated, the size of the recalibration effect follows the same time course as the degree of successful lexical access, supporting the view that recalibration is a consequence of resolving phonetic ambiguity during lexical access. As Norris et al. (2003) suggested, the lexical information can provide a teaching signal to the phonetic categorization process, teaching that process which phonetic category the ambiguous region belongs to.

Collectively, the findings in Parts 1, 2, and 3 support three major conclusions: First, lexically driven recalibration does not happen automatically, if by automatic we mean without the need for attention. The process is automatic in the sense that it appears to occur as a natural consequence of lexical access, and such lexical access does occur in the absence of any particular task (e.g., in all but one of the recalibration experiments, subjects were directed to attend to something other than the female words, yet they processed them enough to generate recalibration). Second, the attentional resources involved in recalibration are not speech specific, as requiring subjects to attend to environmental sounds blocked recalibration in much the same way that imposing a lexical access task did. Third, lexical processing extends past a word's physical ending, with this window of processing having a dura-

tion on the order of one second, consistent with the few prior results in the literature that bear on this question (Connine et al., 1991; Samuel, 1979; Swinney, 1982).

In providing these answers the current work has, not surprisingly, raised new questions. One concerns the breadth of the conclusion that is called for: We now know that lexically driven recalibration is not automatic, but we don't know whether we can extend this new knowledge to the two other phonetic adjustment processes that have been studied – selective adaptation (e.g., Eimas & Corbit, 1973) and audiovisually determined recalibration (e.g., Bertelson et al., 2003). Recall that prior work on all three adjustment processes had suggested that all three are automatic. While the current work shows that this was not correct for lexically driven recalibration, it does not necessarily follow that attention is needed for the other two. Selective adaptation in particular might well be automatic, but this now becomes a renewed empirical question.

A second question that arises concerns the pattern of recalibration effects found across the nine experiments. Because this is such a large set of experiments using the same test stimuli, and because of the range of effect sizes shown in Fig. 7, it is possible to look at how much of the shift came from recalibration after hearing the ambiguity in "s" words versus how big the shifts were after hearing the ambiguity in "sh" words. There have been a few studies on lexical recalibration that have reported asymmetric shifts, though usually the data collection conditions do not permit one to see whether effects are symmetric or asymmetric. Zhang and Samuel (2014), using a contrast between /s/ and /f/, found that robust recalibration occurred for subjects in the "s" condition, but no shifts were found for those hearing ambiguous segments in "f" words. They reanalyzed data reported by Eisner and McQueen (2006), and found the same asymmetry, with "s" being effective and "f" being ineffective. In the current study, there is in fact an asymmetry, but in this case the "s" exposure was not very effective, while "sh" was. This can be seen by comparing the identification values for each side, broken down by whether the effect size was zero, small, or medium-large. On average, the no-effect cases yielded 62.3% "sh" report for the "s" exposure groups, versus 63.9% for the "sh" exposure groups. There was a relatively small change for the "s" groups for the small-effect cases (58.8%, a 3.5% change) or for the medium-large cases (58.2%, a 4.1% change relative to the no-effect cases). The change in the effect was almost twice as large for the "sh" exposure cases, going from the no-effect (63.9%) to the small-effect (64.5%) to medium-large effect (71.4%) cases. To date, no one has offered a clear explanation for these asymmetries, yet they are widespread in the results from all three adjustment phenomena. Clearly, it would be desirable to know why the effects vary in this way.

The current results place very interesting boundaries on the starting and ending times for the window during which lexical processing seems to be converging on a final percept. In future work, it would be very interesting to flesh out exactly what is occurring for the small-effect situation – the pattern found for SOAs greater than a half-second but not more than a second. Presumably this small-but-present recalibration pattern is found because on a trial-by-trial basis, sometimes recalibration occurs, and sometimes it does not. It is possible that this mixture might be due to across-item differences in when a word becomes lexically unique, and when the ambiguous segment occurs. Even though the 40 critical words were rather tightly distributed around their average timing difference between the critical segment and the end of the word, it is plausible that some of them were more likely to afford listeners the information needed to begin recalibration computations before others. It is possible to imagine a set of stimuli selected to systematically vary in the moment when recalibration could theoretically start, and the prediction that follows from the current work is that the size of the observed recalibration would be a function of when these computations are running, relative to the moment when the targeted distraction stops the process. The complications imposed by the actual words in the lexicon would make this a very difficult experiment to conduct, but in principle (and possibly in the right language) it could be done. Such a precisely controlled test would offer the possibility of more finely specifying the window of lexical activation that culminates in the percept that listeners become aware of. This would bring us even closer to being able to reconcile the complexities and intermingling of words in the signal, with the listener's phenomenology of a neatly lined up set of sequential words.

Acknowledgment

Support was provided by Ministerio de Ciencia E Innovacion Grant #PSI2014-53277 and by Ayuda Centro de Excelencia Severo Ochoa SEV-2015-0490. I thank Mark Pitt for providing the uniqueness point calculations, and Martin Cooke for calculating the glimpse analyses.

Appendix A

Fig. A.1, showing the distribution of items in terms of (a) the time between the onset of the critical phoneme and the end of the word, and (b) the time between the offset of the critical phoneme and the end of the word. In both cases, the items are tightly clustered around the mean (550 ms from onset, 379 ms from offset).

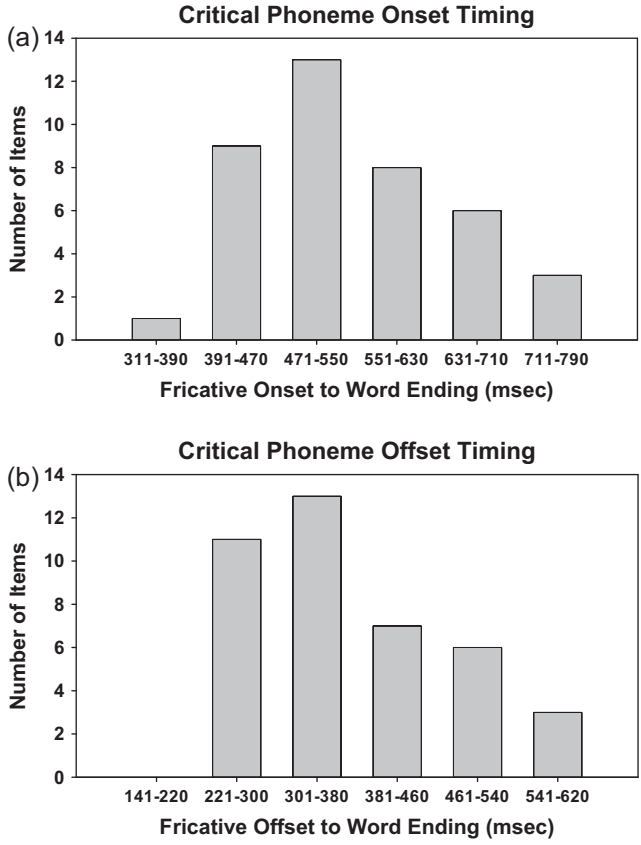

Fig. A.1.

References

Baart, M., & Vroomen, J. (2010). Phonetic recalibration does not depend on working memory. *Experimental Brain Research, 203*, 575–582.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk after effect. *Psychological Science, 14*, 592–597.

Bowers, J. S. (2000). In defense of abstractionist theories of word identification and repetition priming. *Psychonomic Bulletin & Review, 7*, 83–99.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*, 707–729.

Connine, C. M., Blasko, D. M., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language, 30*, 234–250.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America, 119*, 1562–1573.

Davis, M. H., Johnsrude, I. S., Hervais-Ademan, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*, 222–241.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*, 99–109.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*, 224–238.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of Acoustical Society of America, 119*, 1950–1953.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*, 614–616.

Gregg, M. K., & Samuel, A. G. (2008). Change deafness and the organizational properties of sounds. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 974–991.

Gregg, M. K., & Samuel, A. G. (2009). Semantics versus acoustics: Which is more important in auditory representations? *Attention, Perception & Psychophysics, 71*, 607–619.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin and Review, 18*, 943–950.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: An researcher's handbook* (4th ed.) : . Pearson.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Kleinschmidt, D., & Jaeger, T. F. (2015). Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological Review, 122*, 148–203.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition, 107*, 54–81.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*, 141–178.

Kraljic, T., & Samuel, A. G. (2006). How general is perceptual learning for speech? *Psychonomic Bulletin and Review, 13*, 262–268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*, 1–15.

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition, 121*, 459–465.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science, 19*, 332–338.

Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception, part 1: An account of basic findings. *Psychological Review, 88*, 375–405.

McLennan, C., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 529–553.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1609–1631.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes, 10*, 309–331.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Mullennix, J. W. (1986). *Attentional limitations in the perception of speech* Unpublished doctoral dissertation. Buffalo, NY: State University of New York at Buffalo.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204–238.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15*, 285–290.

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology, 70*, 1–30.

Ramachandran, V. S. (1992). Filling in the blind spot. *Nature, 356*, 115.

Samuel, A. G. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics, 22*, 321–330.

Samuel, A. G. (1979). *Speech is specialized, not special* Unpublished doctoral dissertation. San Diego, La Jolla, CA: University of California.

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology, 18*, 452–499.

Samuel, A. G., & Kat, D. (1998). Adaptation is automatic. *Perception & Psychophysics, 60*, 503–510.
Samuel, A. G., & Kraljic, T. (2009). Perceptual learning in speech perception. *Attention, Perception & Psychophysics, 71*, 1207–1218.
Scharenborg, O., Weber, A., & Janse, E. (2015). The role of attentional abilities in lexically guided perceptual learning by older listeners. *Attention, Perception & Psychophysics, 77*(2), 493–507.
Spillman, L., Otte, T., Hamburger, K., & Magnussen, S. (2006). Perceptual filling-in from the edge of the blind spot. *Vision Research, 46*, 4252–4257.
Sumner, M., & Samuel, A. G. (2007). Lexical inhibition and sublexical facilitation are surprisingly long lasting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 769–790.
Sussman, J. E. (1993). Focused attention during selective adaptation along a place of articulation continuum. *Journal of the Acoustical Society of America, 93*, 488–498.
Swinney, D. A. (1982). The structure and time-course of information interaction during speech comprehension: Lexical segmentation, access, and interpretation. In J. Mehler, E. C. T. Walker, & M. Garrett (Eds.), *Perspectives on mental representation*. Hillsdale, NJ: Erlbaum.
Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia, 45*, 572–577.
Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: Dissipation. *Speech Communication, 44*, 55–61.
Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392–393.
Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 200–217.