Auditory Selective Adaptation Moment by Moment, at Multiple Timescales

Arthur G. Samuel [a,b,c] and Nicolas Dumay [a,d]

[a] *Basque Center on Cognition, Brain and Language, Donostia, Spain*

[b] *IKERBASQUE, Basque Foundation for Science*

[c] *Stony Brook University, Dept. of Psychology, Stony Brook, NY, the United States of America*

[d] *University of Exeter, Dept. of Psychology, Exeter, United Kindgom*

WORD COUNT: 13,654

Corresponding author:

Arthur G. Samuel

Department of Psychology

Stony Brook University

Stony Brook, NY 11794-2500

arthur.samuel@stonybrook.edu

ABSTRACT

Over the course of a lifetime, adults develop perceptual categories for the vowels and consonants in their native language, based on the distribution of those sounds in their environment. However, in any given listening situation, the short-term distribution of sounds can cause changes in this long-term categorization. For example, if the same sound (the "adaptor") is heard many times in a short period of time, listeners *adapt* and become less prone to hearing that sound. Although hundreds of speech selective adaptation experiments have been published, there is almost no information about how long this adaptation lasts. Using stimuli chosen to produce very large initial adaptation, we test adaptation effects with essentially no delay, and with delays of 25 min, 90 min, and 5.5 hr; these tests probe the duration of adaptation both in the (single) ear to which the adaptor was presented, and in the opposite ear. Reliable adaptation remains 5.5 hours after exposure in the same-ear condition, whereas it is undetectable at 90 min in the opposite ear. Surprisingly, the amount of residual adaptation is largely unaffected by whether the listener is exposed to speech between adaptation and test, unless the speech shares critical acoustic properties with the adapting sounds. Analyses of the shifts on three time scales (seconds, minutes, and hours) provide information about the multiple levels of analysis that the speech signal undergoes.

Public Significance Statement: Spoken language is the primary mode of communication for most humans. The current work helps to specify the ways in which the speech signal is decoded, helping us to understand how humans accomplish this challenging task.

Keywords: speech perception; auditory selective adaptation; levels of speech analysis; time course; recovery time; testing effect

An optimal perceptual system must have two seemingly conflicting properties. It should be stable – it should produce the same analysis despite irrelevant variation of the input. However, it should also be flexible – if conditions change, it should be able to modify its analysis accordingly. There is an abundance of evidence, across multiple perceptual domains and levels of complexity, that humans achieve both of these goals. Stability seems to be grounded in perceptual representations that are built from a person's long-term interactions with the world; flexibility seems to come from being sensitive to recent experiences that provide evidence that the long-term statistics need to be modified, at least temporarily.

As discussed in a recent review by Snyder et al. (2015), perceptual flexibility is demonstrated by contrastive adjustments, in which experiencing a clear, unambiguous stimulus of some type induces a change in how a somewhat similar stimulus is perceived. There are many well-known contrastive effects in the visual domain, implicating dimensions such as motion, tilt, brightness, and color. In the "color afterimage" effect, for instance, after staring at a color patch (e.g., red), a person will see its complementary color (e.g., green) when then looking at a white wall. Perceptual stability, in contrast, is achieved by means of attractive, rather than contrastive, adjustments. For example, the way in which a viewer interprets a perceptually bi-stable stimulus (like a Necker cube, which can be seen with its face pointing either up or down) tends to re-impose itself if a similar bi-stable stimulus (i.e., another Necker cube) is presented immediately after.

In the present study, we are concerned with how listeners map the acoustic stimuli they hear onto the sounds of their language. Despite this change of modality, the two types of perceptual adjustment that we have described so far apply in the speech domain, just as they do in visual perception. A listener's perceptual categories are grounded in years of experience with the language, but they are not immutable. In fact, given the substantial variability that characterizes speech, perceptual flexibility is essential. In any given listening situation, the short-term distribution of sounds may diverge from long-term averages, and modifying these long-term averages adaptively can yield improved perceptual performance. Just as Snyder et al. (2015) identified two complementary types of perceptual adjustment, there are two well-documented phenomena that demonstrate modifications of speech categories as a function of particular short-term deviations from the long-term distributions: phonemic recalibration (also sometimes called “perceptual learning”, or “retuning”) and selective adaptation (or "habituation"). Recalibration involves an attractive process in Snyder et al.'s terminology, whereas adaptation involves a contrastive process.

Evidence for phonemic recalibration comes from two seminal papers (Bertelson et al., 2003; Norris et al., 2003) published in the same year, but developed independently in two different labs. In Bertelson et al. (2003) participants were initially exposed to speech tokens chosen to be perceptually ambiguous, midway between /aba/ and /ada/. Critically, the ambiguous speech was dubbed, on some trials, onto a video of a talker saying /aba/, and on other trials, onto a video of a talker saying /ada/. Perceptual recalibration was found when the participants identified auditory-only tokens immediately following the audiovisual exposure: ambiguous test sounds were perceived more as /aba/ when the preceding speech had been dubbed onto a visual /aba/ than onto a visual /ada/. Bertelson et al.'s interpretation was that during audiovisual exposure, unambiguous visual phonemic information disambiguates the speech signal, and this causes the phonemic category boundary to be adjusted. In other words, the long-term representations are modified in accord with current listening conditions.

In Norris et al. (2003), exposure involved listening to (auditory-only) words ending with a consonant modified to be ambiguous between /s/ and /f/ while performing a lexical decision task. Like the visual speech in Bertelson et al., the lexical context helped to disambiguate the ambiguous consonants. All ambiguous tokens for a given participant occurred in words ending with a particular consonant, either /f/ (e.g., "photogra?") or /s/ (e.g., “compa?”), with other words ensuring that participants heard unambiguous tokens of the opposing category (/s/ or /f/). As in Bertelson et al., a post-exposure test showed that participants had recalibrated the boundary between the two categories.

In both the audiovisual and the lexically-driven versions of recalibration, long-term phonemic category boundaries get modified through the presentation of ambiguous stimuli that come with a teaching signal, some kind of information that allows the system to interpret the ambiguity. In selective adaptation, by comparison, the critical sounds in the exposure phase are not ambiguous – they are good exemplars of a phonemic category. The short-term adjustment of phonemic boundaries is driven simply by many exposures to the good exemplar. In a pioneering paper, Eimas and Corbit (1973) repeatedly exposed listeners to one of the two unambiguous endpoint sounds of a continuum between /ba/ and /pa/ (e.g., "/ba/"). In a subsequent labelling test in which all tokens of the continuum were played in random order, the authors found fewer reports favoring the adaptor category (i.e., "/b/") compared to pre-adaptation levels. Note that these *contrastive* shifts are in the opposite direction of the *attractive* shifts involved in perceptual recalibration.

Though clearly not all (e.g., Favreau, 1976; McCullough, 1965), many adaptation effects, most of which have been studied with visual stimuli, implicate fatigue (i.e., saturation) in the neural response of early sensory areas. For example, classic single-cell recordings (e.g., Barlow & Hill, 1963) showed that stimulation of cells sensitive to one motion direction makes them less able to compete with fresher cells sensitive to the opposite direction of motion. These effects are typically short-lived and decline over the space of a few seconds. In comparison, attractive effects appear to engage wider networks, involved in higher order computations and integrative processes (e.g., Kaiser et al., 2013; Killian-Hutten et al., 2011; Schwiedrzik et al., 2014).

Given this dissociation in terms of breadth of neural engagement, Snyder et al. (2015) put forward a Bayesian model which deals with both contrastive and attractive effects by means of two separate mechanisms. In this model attractive effects involve a *long-term* change in the probability distribution of likely percepts given past experiences (i.e., a change in "priors)", whereas contrastive effects involve a *transient* reduction in the evidence currently available for a given percept, favoring the alternative interpretation of the stimulus. The resulting percept is thus a product of the combination of these two independent probability functions.

Snyder et al.'s (2015) approach contrasts with the single-mechanism Bayesian account that Kleinschmidt and Jaeger (2015) offered to explain attractive and contrastive effects in speech perception. In their model, perceptual recalibration and selective adaptation are explicitly seen as two consequences of a *single* adjustment process. With phonemic perception depending on the predictive nature of available cues and the likelihood of the category, their "belief-updating" model assumes that recalibration and adaptation both are a result of exposure modifying the distribution of the target category (e.g., Clayards et al., 2008; Maye et al., 2002). Whereas exposure to an ambiguous exemplar in the presence of a disambiguating cue (as in recalibration) enlarges the breadth (i.e., the variability) of a category, exposure to the prototype (as in selective adaptation) acts in the opposite fashion by sharpening the category around its mean.

In the Kleinschmidt and Jaeger (2015) model, phoneme perception is driven by the distributional properties of sound categories and the extent to which these overlap with one another. Exposure has no effect other than that of shaping and reshaping the parameters of the category. In the more general model proposed by Snyder et al. (2015), in contrast, exposure has the potential to have both short-term and long-term effects depending on whether the stimulus acts as an adaptor or instead supports a change in the underlying distribution. Thus, only according to the two-mechanism approach should we expect phonemic recalibration and adaptation to be affected differently by the passage of time.

It is already clear that even the two variants of phonemic recalibration do not decay at the same rate: whereas audiovisual recalibration is gone after a few tens of seconds (Vroomen et al., 2004), lexically-driven recalibration persists a matter of hours, if not a few days (Eisner & McQueen, 2006; Zhang & Samuel, 2014). Surprisingly, while the question of persistence seems central to elucidating the mechanisms behind phonemic adaptation, Sharf and Ohde (1981) appears to be the only paper in the large literature on speech adaptation to have assessed the recovery time (i.e., how long it takes categorization to return its pre-adaptation boundary). Using the /pa/ endpoint of a /ba/-/pa/ continuum as the adaptor, Sharf and Ohde measured how much of the initial adaptation shift was left after 1, 4, 7 and 28 min; this was done within-participant, but counterbalancing the interval condition. Relative to the shift obtained during the adaptation phase (5.4%), they found adaptation effects reduced by half after 1 min (2.6%) and by about two thirds after 4 and 7 min (1.9% and 1.6%, respectively). While these effects were reliable and of reasonable size (Cohen's d (derived from their t-values reported on p. 83) = .78, .43, and .81, respectively), the adaptation remaining at the longest interval was negligible (0.5%), suggesting that 28 min was sufficient for essentially complete recovery.

Based on this single case, it would appear that phonemic adaptation dissipates in a matter of minutes. The rapid recovery that they found is similar to the time course reported by Schweinberger et al. (2008) on voice adaptation effects, which also dissipated within minutes. However, we know from the speech adaptation literature that the size of the shift can vary considerably depending on the nature of the continuum tested as well as on the parameters of the procedure. Sharf and Ohde (1981) used a continuum of stop consonants with an adaptor presented binaurally, and the observed recovery time could well be bound to the magnitude of the shift that they achieved by their procedure. Therefore, the first aim of our study was to assess the recovery time after adaptation using adaptors and test stimuli that have been shown to produce extremely large initial adaptation shifts.

In contrast to Sharf and Ohde (1981), we use monaural presentation of the adaptor and test stimuli, because previous studies using this procedure have provided evidence for more than one level of speech analysis. By adopting this approach, we can determine whether each level of analysis has its own time course of recovery. The studies that compared ipsilateral (same ear) to contralateral (different ear) adaptation have implicated two processing levels: a "peripheral" level (i.e., before information from the ears is combined, which is quite early – subcortical – in audition) and a "central" level (i.e., based on information from the two ears combined). A "peripheral" level is implied if adaptation is found only when test syllables are presented in the same ear as the adaptor. Conversely, a “central” level is implied if contralateral adaptation is as effective as ipsilateral adaptation. Assuming that these two levels are effectively distinct, adaptation in the same ear as the test stimuli should reveal the workings of central and peripheral processes combined, whereas identification of test stimuli in the other ear (than the one used for the adaptor) should reveal the workings of central processes only.

Study	Continuum	Adaptor	% Ear Transfer	*N*	Design
Eimas et al. (1973)	da-ta	Ta	~100%	5	
Ades (1974)	bae-dae	Bae	~50%	4	
Ades (1974)	bae-dae	Dae	~50%	4	
Sawusch (1977)	bae-dae	Bae	~50%	12	btwn
Sawusch (1977)	bae-dae	Dae	~50%	12	btwn
Ganong (1978)	bi-di	Ti	~33%	13	
Jamieson et al. (1986)	ba-pa	Ba	~25%	19	
Jamieson et al. (1986)	ba-pa	Pa	~80%	19	
Jamieson et al. (1986)	da-ta	Da	~30%	9	
Jamieson et al. (1986)	da-ta	Ta	~90%	9	
Samuel (1988)	ba-wa (V*)	wa (V*)	~40%	12	
Samuel (1988)	ba-wa (V*)	ba (V*)	~33%	12	
Samuel (1988)	ba-wa (W*)	wa (W*)	~40%	12	
Samuel (1988)	ba-wa (W*)	ba (W*)	~95%	12	

Table 1. Result summary of studies that compared ipsilateral to contralateral phonetic adaptation. “% Ear Transfer” expresses the size of the contralateral shift as a percentage of the ipsilateral shift. The mean Ear Transfer is 54.8%; the median Ear Transfer is 50%. V* = voiced stimuli; W* = whispered stimuli. btwn = between-subject design, with a within-subject design for cases not marked as between. Numbers are approximations based on what was available in the original paper (e.g., tables, text, graphs).

Table 1 displays the results of six studies that differed in the specifics of their stimuli and protocol, but which included the same- versus different-ear manipulation. None of this earlier work manipulated the delay between adaptation and test—all tests were conducted immediately after adaptation. The % Ear-Transfer represents the size of the contralateral shift as a function of the shift obtained in same-ear testing. For example, Sawusch (1977) found that both a /bae/ adaptor and a /dae/ adaptor produced shifts that were about half as big using contralateral compared to ipsilateral adaptation. In fact, although there is clearly some variability across studies, the average contralateral shift (in terms of both mean (55%) and median (50%)) is about half as large as the ipsilateral shift. This pattern has been taken to support the existence of two separate levels of speech analysis, with one level operating on “peripheral” and another level operating on “central” representations.

This interpretation is bolstered by results from studies that used adaptors that shared certain abstract properties with the test continua, but that were not acoustically as similar to the test items as the typical endpoint adaptor. For example, in Sawusch (1977) the "high" adaptors were based on the endpoint tokens of the /bae/-/dae/ test continuum, but the adaptor’s formants were shifted up by 1.5 critical bands. These adaptors thus have similar frequency patterns as the continuum endpoints, but they do not sound like them nor overlap with them acoustically. As Table 2 shows, the "Efficacy" of these more abstractly-related adaptors for test items presented in the same ear is on average a bit less than half of the effect obtained with the original endpoints. These “abstract” adaptors are tapping into a level of processing that is beyond an initial analysis that is more tightly tied to the signal.

Study	Continuum	Adaptor	Efficacy	% Ear Transfer	*N*	Design
Sawusch (1977)	bae-dae	"high" bae	~50%	~100%	12	btwn
Sawusch (1977)	bae-dae	"high" dae	~30%	~100%	12	btwn
Samuel (1988)	ba-wa (V*)	wa (W*)	~100%	~70%	12	
Samuel (1988)	ba-wa (V*)	ba (W*)	~30%	~65%	12	
Samuel (1988)	ba-wa (W*)	wa (V*)	~65%	~60%	12	
Samuel (1988)	ba-wa (W*)	ba (V*)	~45%	~100%	12	

Samuel & Kat (1996)	ba-da	pa, ta	~30%	~100%	12
Samuel & Kat (1996)	ba-da	F2F3b,d	~30%	~100%	12

Table 2. Result summary of studies that used abstract adaptors. In Sawusch (1977), abstract adaptors were the original endpoints with their formants frequency shifted up by 1.5 critical bands. In Samuel (1988), they were whispered variants of voiced test items, or voiced variants of whispered test items. In Samuel and Kat (1996), they were either the /pa/ and /ta/ voiceless counterparts of the /ba/-/da/ endpoints, or versions of the /ba/-/da/ endpoints generated with only the second and third formants (hence, the F2F3 notation)—the lack of F1 produces sounds with very little phonetic quality (i.e., they do not sound very speechlike). “Efficacy” expresses the size of abstract adaptation effects as a percentage of the effect obtained with the continuum’s original endpoints. The mean efficacy rate is 48%; the median efficacy rate is 38%. “% Ear Transfer” expresses the size of the contralateral shift as a percentage of the ipsilateral shift. The mean Ear Transfer is 87%; the median Ear Transfer is 100%. V* = voiced stimuli; W* = whispered stimuli. btwn = between-subject design, with a within-subject design for cases not marked as between. Numbers are approximations based on what was available in the original paper (e.g., tables, text, graphs).

A natural way of relating these findings to those shown in Table 1 is to assume that abstract adaptors engage mostly central, as opposed to peripheral, processes. If this interpretation is correct, abstract adaptors should be as effective contralaterally as they are ipsilaterally. As indicated by the rates of ear transfer in Table 2, this is essentially the case: unlike the original endpoint adaptors, abstract adaptors produce categorization shifts of virtually the same magnitude in contralateral tests as in ipsilateral tests (with a mean transfer rate of almost 90%). In view of the results reported in Tables 1 and 2, the prevailing view is that adaptation effects implicate two distinct levels of speech analysis: the first of these levels extracts acoustic features, while the second level operates on more abstract representations.

Given the above, and in contrast with Sharf and Ohde (1981), the present research therefore distinguished between testing in the ear that received the adaptor versus the ear that did not. This allowed us to map out, on three different time scales (i.e., hours, minutes seconds), the time course and recovery time for phonetic adaptation, separately for central and peripheral levels. It might be expected that peripheral analyses that are closely linked to the acoustic properties of the signal should operate over a relatively short timescale, whereas central analyses should involve more abstract representations that endure longer—this is the pattern generally found for the visual aftereffects mentioned above.

We based our experiments on a quite different contrast than the /ba/-/pa/ contrast used by Sharfe and Ohde (1981). Prior adaptation experiments with a /ba/-/wa/ contrast (Samuel, 1988) showed (a) extremely large adaptation shifts and (b) a strong difference between ipsilateral and contralateral adaptation. These two properties are exactly those that are needed for the current study, and thus we employed those stimuli here.

While tracking the recovery from this likely to be very strong phonemic adaptation, we also assessed the impact of repeated exposure to the full continuum of syllables during the post-adaptation test. In domains in which the input is typically not as transient as speech, adaptation shifts grow weaker the longer the duration of the test-stimulus (e.g., Leopold et al., 2005). Similarly, attractive effects in speech (i.e., recalibration) also seem to be rapidly undone by multiple rounds of test-trials (Liu & Jaeger, 2018). We therefore gauged adaptation after every pass and every trial within a pass in order to capture this "unlearning" process.

The experiments and analyses are presented in four parts. In Part 1, we report an experiment that is comparable to those presented in Table 1. This experiment establishes that the stimuli and procedures produce results that are consistent with the literature on ipsilateral versus contralateral adaptation effects. In Part 2, we report four experiments that represent the crossing of three factors: (1) the ipsilateral versus contralateral nature of adaptation, (2) the delay between adaptation and test -- either 25-min or 90-min, and (3) the amount of speech exposure during the adaptation-test interval (i.e., substantial versus little speech exposure; see Earle & Myers, 2015, for a demonstration of the impact of such exposure on learning non-native phonetic contrasts). Part 3 extends the adaptation-test delay out to 5.5 hours, keeping the same-ear vs. different-ear manipulation. The four experiments in Part 2 are grouped together because their delay durations allow us to control whether the listeners hear speech between the adaptation phase and the test phase. This cannot be done in Part 1 because there is essentially no time between these phases; in Part 3, this time period is too long, as a practical matter, to eliminate all speech input. Finally, in Part 4, we use the results from all six experiments to look at ipsilateral and contralateral adaptation over three time scales: hours (i.e., the interval between a pretest and a posttest), minutes (i.e., from one pass to the next, within the test or adaptation phase), and seconds (i.e., within a pass through the test-syllables of the continuum).

PART 1: Immediate Adaptation Effects

The first experiment is intended to establish the basic Same Ear versus Different Ear difference, and to provide a baseline that can be used to measure the reduction in the size of the adaptation effect over time. Our goal is to produce a large initial adaptation shift and a substantial difference between ipsilateral and contralateral adaptation. We use Samuel's (1988) /ba/-/wa/ voiced continuum and /wa/ adaptor because in that study, the /wa/ adaptor generated a very large adaptation shift and the ear manipulation yielded very robust results representative of those reported in the literature (i.e., contralateral effects were approximately half as large as ipsilateral effects; see our Table 1).

To specify what we mean by a large shift and a large ipsilateral/contralateral difference, Cohen’s d for the /wa/ adaptor was 3.80 for the post-adaptation shift and 2.42 for the ear effect in Samuel (1988). Note that the adaptation remaining after 1 min in Sharf and Ohde (1981) (Cohen’s d of 0.85) would be considered as "large" according to Cohen's classification (i.e., > 0.8), but it certainly was much less robust than our expected base effect. Thus, with very large initial adaptation shifts and a clear difference between ipsilateral and contralateral adaptation to start with, we will be able to track changes in the size of these effects as the test phase is moved farther and farther away from the adaptation phase. As noted above, it is likely that the time for which an effect remains measurable depends on how large it is to begin with. By using an adaptor that produces very large immediate effects, we can therefore obtain an estimate of adaptation duration that is likely to be near the upper bound, which was our goal.

For a similar reason, we use a larger sample than is the norm in the adaptation literature. As Tables 1 and 2 show, such studies typically have about a dozen participants; this relatively small number is usually sufficient because adaptation is a very reliable and robust phenomenon. However, because our study involved pushing the test away from the adaptation phase, we anticipated that effects would get smaller and smaller. As such, they would no longer be as robust as in a typical post-adaptation immediate test. Therefore, in each condition we ran 32 participants (with an expectation of about 10% attrition), a sample thus two-to three-times larger than the norm. As with adaptor strength, it is conceivable that an even larger sample might extend the measurable period slightly, but our large sample size should provide a good estimate of the upper bound.

All 192 participants in the current study were volunteers from the Psychology Department Subject Pool at Stony Brook University. The Subject Pool is 65% female, and after excluding students in the Pool younger than age 18 (who were excluded from our experiments), approximately 98% of the students in the Pool are between the ages of 18 and 25. We did not systematically keep track of the gender of the participants, but were able to retrospectively determine this information for over half of the 192 participants. For this set, 69% were female, with this value in the individual experiments varying between 57% and 75%. The research was approved by the Institutional Review Board at Stony Brook University.

EXPERIMENT 1

Method

Participants

A total of 32 participants took part. As in all experiments reported in this paper, they were American English native speakers with no self-identified hearing problems and they received course credits for their participation.

Stimuli

Test syllables: The test items were a set of synthetic speech syllables that instantiated a /ba/-/wa/ continuum. The stimuli came from an 8-step test series used by Samuel (1988). To better center the continuum, the most extreme /ba/ stimulus was not used, leaving a 7-step continuum. All stimuli were 300 msec long; they differed in the duration (and therefore slope) of the consonantal formant transitions. The most extreme /ba/ token used here had 25 msec transitions, while the most extreme /wa/ token had 55 msec transitions; the transition durations of the intermediate steps changed in 5-msec increments. The items had been generated using the cascade branch of the Klatt synthesizer, using a fundamental frequency range (122 HZ for the first 100 msec, dropping linearly to 90 Hz at offset) typical of an adult male voice. See Samuel (1988) for more details of the synthesis.

Adaptor: The endpoint /wa/ token of the test series was used as the adaptor.

Apparatus and Procedure

Participants were tested in sound-shielded chambers, with 1 or 2 participants tested at a time. They listened to the speech over SONY MDR-V900 stereo headphones, and responded by using two labeled buttons on a response pad in front of them. To balance ear of stimulus presentation, half of the participants wore the headphones with the standard orientation (i.e., with the left channel presented to the left ear), and half wore them with the orientation reversed (i.e., with the left channel presented to the right ear).

The test session included an initial identification task, followed by an adaptation task, followed by a final identification task. During the initial identification task, listeners heard 22 randomizations of the seven members of the /ba/-/wa/ continuum; the first two randomizations were practice and were not scored. The odd-numbered passes were presented to one channel, and the even-numbered passes were presented to the other channel, yielding 10 observations for each member of the continuum heard in the left ear, and 10 observations for each member of the continuum in the right ear. Participants identified each syllable by pushing one of two labeled buttons (“B” on the left key, “W” on the right key). Trials began 500 msec after the previous responses, with a maximum of 3000 msec allowed before moving on to the next trial.

During the adaptation task, participants made the same judgments, on the same syllables, with the same timing. There were 20 passes (no practice passes), again providing

10 observations per test item per ear. However, each randomization was preceded by an adaptation phase in which the endpoint /wa/ was played 30 times, with an interstimulus interval of 500 msec. After the final adaptor was played, an additional 500 msec interval was included before the presentation began of the randomization of the test continuum items. The adaptor was always presented to the same single channel; with the headphones reversed for half of the subjects, the adaptation was presented to the left ear for half of the participants, and to the right ear for the other half. Previous adaptation studies (e.g.. Samuel, 1986) have shown no effect of left- versus right-ear of adaptor presentation. Because the test item randomizations switched ears (again, the odd-numbered passes went to one ear, and the even-numbered passes went to the other), this procedure provides identification results for stimuli presented to the same ear as the adaptor, and to the opposite ear.

The final task of the session was the second identification task. This task was identical to the initial identification task. Including reminding participants of how the identification task worked, and entering the necessary information into the program running the task, it typically took about 1-2 min to begin the final test after completion of the adaptation task. We will refer to this condition as the "Immediate Test" condition, in comparison to the much longer delays used in subsequent experiments. Each identification task took about 5 min, and the adaptation task took about 14 min, so the entire session lasted about 25 min.

Results and Discussion

As in previous studies (e.g., Samuel 1989, 2016, 2020), participants who were unwilling or unable to do the task were identified on the basis of their labeling of the test syllables. Each participant produced four identification functions – responses presented to stimuli in the left ear and in the right ear, for the initial identification test and for the final identification test. If for any of these four functions the percentage of “W” report for the most /w/-like token was not at least 50% greater than the percentage for the most /b/-like item, the listener was classified as not having done the required task. Data for three of the 32 participants were eliminated on this basis, leaving usable data for 29 participants. Figure 1 shows the results for the remaining 29 participants.

Again following the procedures used in previous work (e.g., Samuel 1989, 2016, 2020), adaptation effects for each participant were assessed by computing scores that were based on identification of stimuli near the middle of the test continuum, where adaptation effects are typically most robust (note, for example, that in Figure 1, effects become smaller closer to the endpoints due to identification levels approaching floor or ceiling, as expected). Specifically, for both the same-ear and different-ear conditions, for each participant, the average report of /wa/ for items 3, 4, and 5 was computed for the initial identification test, for the adaptation test, and for the final identification test. These averages were used in the statistical tests. This measurement procedure has proven to be very robust in previous work. For the interested reader, the Appendix describes this method and others that are typically used to measure identification changes in experiments of this sort.

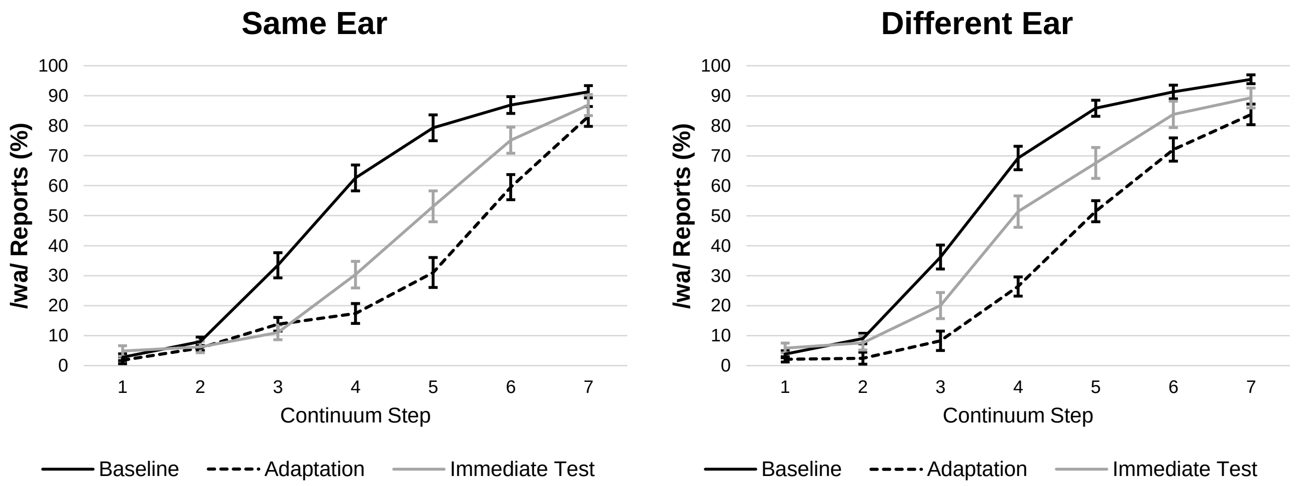

***Figure 1**: Identification of the members of the /ba/-/wa/ test series when the final identification test immediately followed the adaptation task. The left panel shows the results when test syllables were presented to the same ear as the adaptor, and the right panel shows the results when test syllables were presented to the opposite ear. "Baseline" shows identification before adaptation, "Immediate Test" shows identification on the final test (i.e., 1-2 mins after the adaptation phase), and "Adaptation" shows identification within the adaptation phase. Error bars show SEs.*

As Figure 1 shows, the results of Experiment 1 meet our goals for it: The endpoint /wa/ adaptor produced very large adaptation effects, and these effects were noticeably bigger when the test syllables were presented ipsilaterally to the adaptor, rather than contralaterally. Given our goal of tracking the strength of adaptation over time, the most important results are based on the difference in percentage of /wa/ reports between the pre- and post-adaptation identification tests; this is the measure that we will track over increasing delays. Based on the middle three items of the test series, for the same-ear condition, the initial 58.4% /wa/ report dropped to 31.5% on the final identification test, a difference of 26.9%, t(28) = 8.11, p < .001, Cohen's d = 1.51. When the test items were instead contralateral to the adaptor, the initial 63.8% /wa/ report dropped to 46.3%, a difference of 17.4%, t(28) = 6.08, p < .001, Cohen's d = 1.13. While both effects were quite robust (as desired), the contralateral shift was 65% as large as the ipsilateral case, a statistically significant difference in magnitude, one-tailed t-test: t(28) = 3.17, p < .002, Cohen's d = .60. This is well within the range of effects reported in Table 1.

The results of Experiment 1 establish that the stimuli and testing conditions are suitable for addressing the primary questions of the current study: (1) how long do adaptation effects endure?, (2) are the patterns of reduction over time the same for effects based on peripheral (same-ear) processes as for effects based on central (opposite-ear) processes?, and (3) does exposure to speech during the post-adaptation period play a major role in returning the categorization pattern back to its pre-adaptation state?

PART 2: Adaptation Effects after 25 vs. 90 min

In order to address these three core questions, Part 2 reports the results of four experiments. In Experiments 2a and 2b, we imposed a 25-min delay between the completion of the adaptation test and the presentation of the final identification test (recall that in Experiment 1 this delay was about 1-2 min, the time needed to set up and start the final post-adaptation test). In Experiments 2c and 2d, this post-adaptation delay was extended to 90 min. Within each pair of experiments, we varied the amount of speech exposure during this interval. Whereas in Experiments 2a and 2c the activities that participants were given to do during the interval entailed listening to speech, in Experiments 2b and 2d the tasks were chosen to be ones that did not involve hearing speech (see Kraljic & Samuel, 2005 for a similar test of recalibration). The only speech sounds that participants heard in these experiments were the very brief instructions for starting the last part (basically, two or three short sentences, as the participants were already familiar with the task).

Based on the only precedent that we are aware of (Sharf & Ohde, 1981), one might expect to find very small effects in Experiments 2a and 2b (with their 25-min delay) and no effects in Experiments 2c and 2d (after 90 min). As noted in the Introduction, adaptation in the Sharf and Ohde (1981) study was virtually gone at the longest (i.e., 28-min) delay that they tested. However, the literature also makes it clear that the size of the initial adaptation shift is very much dependent on the details of the stimuli, and to some extent, on the details of the test procedure. Our stimuli were chosen to yield very strong adaptation and to be very sensitive to the ear manipulation (which was not included in the Sharf & Ohde (1981) paper). Therefore, it is an open question whether the time course found for a voicing continuum (/ba/-/pa/) with a voiceless adaptor (/pa/) will resemble the time course for a stop-continuant continuum (/ba/-/wa/) with a continuant adaptor (/wa/).

EXPERIMENT 2

Method

Participants

A total of 128 participants took part; none had participated in Experiment 1.

Stimuli, Apparatus and Procedure

Stimuli, apparatus and procedure for the pre- and post-adaptation identification tests and the adaptation phase were the same as in Experiment 1.

Experiment 2a Delay Procedures: In Experiment 2a there was a 25-min delay between the end of the adaptation task and the beginning of the final identification task. When the adaptation task finished, the participants were told to come out of the sound-shielded chamber and to sit in front of a computer screen at a nearby table. They were given headphones in which only one ear was fed sound; the ear alignment was set up to match the ear in which the participant had heard the adapting sound. They watched a 23-min episode of “Pinky and the Brain” (a cartoon about two talking mice, one of whom has a rather deep voice, and one of whom has a higher-pitched voice); all participants watched the same episode. When the episode finished, they returned to the sound-shielded chamber to complete the final identification task.

Experiment 2b Delay Procedures: In Experiment 2b there was also a 25-min delay, and participants also left the sound-shielded chamber to sit in front of the computer screen at a nearby table. However, rather than watch a cartoon with sound, they were offered a set of computer games that did not have any sound (e.g., solitaire, checkers, …). They played the game(s) that they preferred for 23 min and then returned to the sound shielded chamber to compete the final identification task.

Experiment 2c Delay Procedures: In Experiment 2c, there was a 90-min delay between the end of the adaptation task and the beginning of the final identification task. In order to keep participants engaged, we gave them several different activities during this rather long interval, rather than asking them to do one thing for an hour and a half.

When they finished the adaptation task, they came out of the sound shielded chamber and sat in front of the computer screen at a nearby table. The first activity they were given was to watch the 23 min episode of “Pinky and the Brain” that was used in the 25-min condition, with the sound track presented over headphones to the same ear that had received the adapting sound. When this episode finished, they returned to the sound shielded chamber. They did a 20-min task, wearing headphones that again only presented sound to the same ear. The task was to listen to a series of trials in which a real word (e.g., "computer") or a slight deviation from a real word (e.g., "comtuter") was presented in a female voice, followed by a matching or slightly mismatching token presented in a male voice that had undergone distortion to lower the intelligibility. The judgment was whether the two items were phonemically the same or different. When this task was completed, the participants left the sound shielded chamber again to watch a second episode of "Pinky and the Brain" with sound presented to the same ear; all participants watched the same second episode. After that 23-min activity, they returned to the sound shielded chamber for the final filler activity. This task involved listening to a male voice presenting short phrases (e.g., "African parrot") in the same ear as before. Participants used three buttons on their response pad to categorize each phrase as "animal", "vegetable", or "mineral". This task took about 15 min, and was followed by the final identification test. With the transition times between different tasks, the time between finishing the adaptation task and starting the final identification task was approximately 90 min.

Experiment 2d Delay Procedures: As in Experiment 2c, we imposed a delay of 90 min between the adaptation task and the final identification test. However, in this case, this interval minimized exposure to speech. Again, we used multiple filler tasks to keep the participants engaged. For the first filler task, participants remained in the sound shielded chamber. They heard a series of melodies, played on a piano. The melodies were ones chosen to be relatively familiar to most people, but they had been subjected to varying degrees of distortion (involving flipping segments of the melody on the time axis). Each melody was played twice in the same ear that the adaptor had been in. Participants used the four buttons on their response pads to classify each melody as ranging from not at all recognizable to very recognizable. This task took about 20 min.

After completing the first filler task, participants left the sound shielded chamber and sat at a nearby table in front of a computer screen. They had the same choice of games (e.g., solitaire, checkers, …) offered to the participants in Experiment 2b. They played the game(s) of their choice for 25 min, and then returned to the sound shielded chamber.

After putting on their headphones, they did two same-different tasks with stimuli presented to the same ear as before. On each trial, two sounds were presented successively. The second sound was either identical to the first, or was very slightly changed. During the first task, the sounds were tonal patterns, and during the second task the sounds were degraded versions of something that sounded more or less like “uh”. Together, the two same-different tasks took about 10 min.

After completing the same-different tasks, subjects again left the sound shielded chamber and returned to the table where they played the computer-based game(s) of their choice for 25 min. Finally, they returned to the sound-shielded chamber to do the final identification test. With the transition times between different tasks, the time between finishing the adaptation task and starting the final identification task was approximately 90 min.

Results and Discussion

The same criteria were used to identify participants who did not do the task as instructed. The data set from one participant was eliminated from Experiment 2a (leaving 31), no participants’ data were eliminated from Experiment 2b, two participants’ data sets were eliminated from Experiment 2c (leaving 30), and four participants' data sets were eliminated from Experiment 2d (leaving 28).

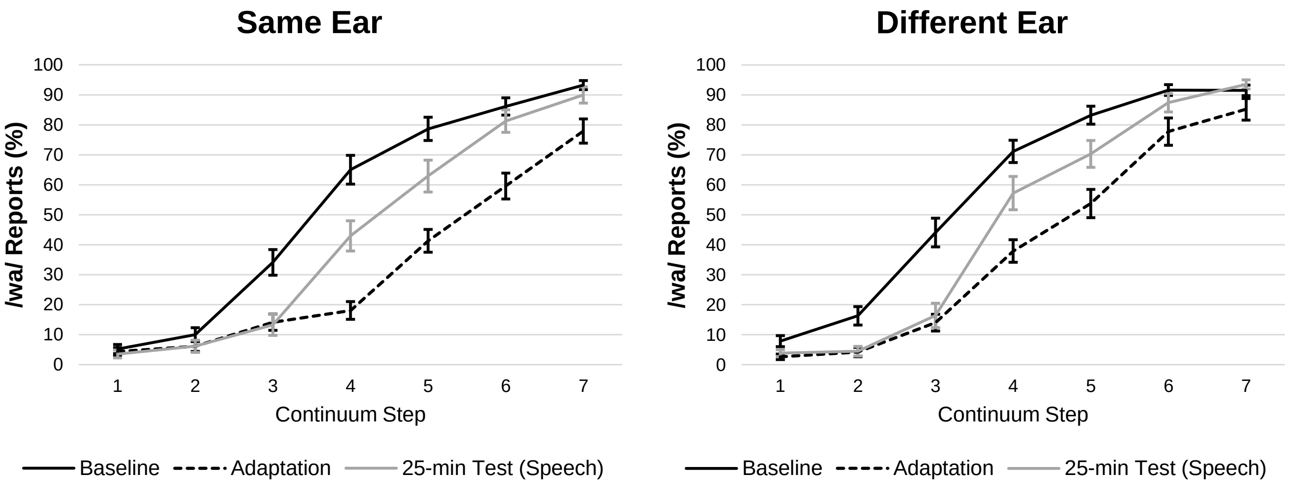

***Figure 2**: Identification of the members of the /ba/-/wa/ test series for the 25-min delay case with speech heard during the delay (Experiment 2a). The left panel shows the results when test syllables were presented to the same ear as the adaptor, and the right panel shows the results when test syllables were presented to the opposite ear. "Baseline" shows identification before adaptation, "25-min Test" shows identification on the final test, and "Adaptation" shows identification within the adaptation phase. Error bars show SEs.*

Figure 2 presents the results for Experiment 2a, in which the final identification test followed a delay of 25 min during which the participants heard speech in the same ear that had received the adapting sounds. Figure 3 presents the results for Experiment 2b, which had the same delay but with very little speech heard during the 25 min. These figures make it clear that adaptation effects can last much longer than would have been expected based on Sharf and Ohde's (1981) results. In the presence of speech (Experiment 2a) and in its absence (Experiment 2b), robust and sizeable adaptation is still present after a 25-min delay.

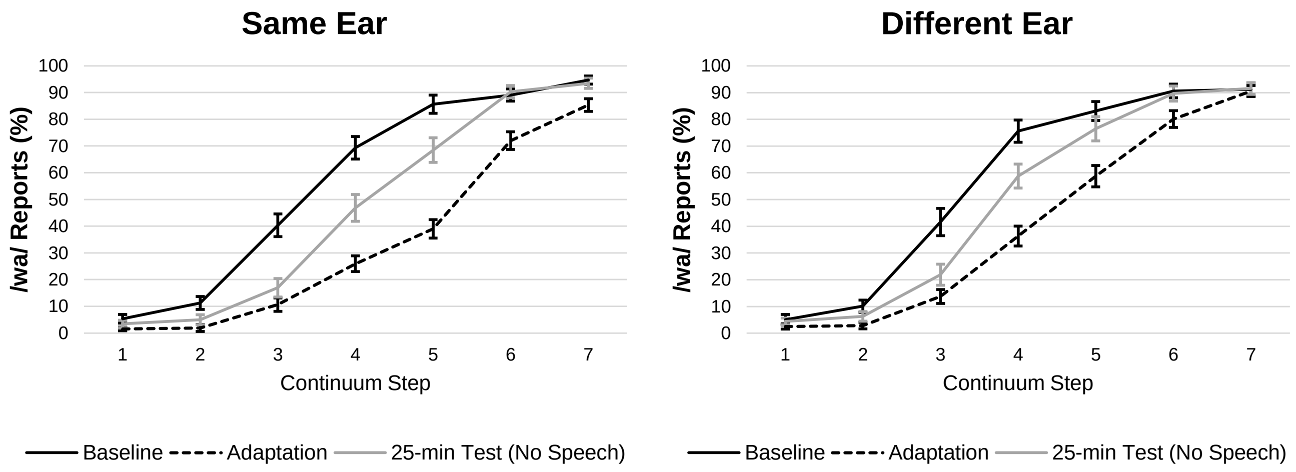

***Figure 3**: Identification of the members of the /ba/-/wa/ test series for the 25-min delay case with little to no speech heard during the delay (Experiment 2b). The left panel shows the results when test syllables were presented to the same ear as the adaptor, and the right panel shows the results when test syllables were presented to the opposite ear. "Baseline" shows identification before adaptation, "25-min Test" shows identification on the final test, and "Adaptation" shows identification within the adaptation phase. Error bars show SEs.*

For the speech-present case, using our measure based on the middle items of the test series, for the same-ear condition, the initial 59.2% /wa/ report dropped to 39.7% on the final identification test, a difference of 19.6%, t(30) = 6.48, p < .001, Cohen's d = 1.16. When the test items were instead contralateral to the adaptor, the initial 66.2% /wa/ report dropped to 48.0%, a difference of 18.1%, t(30) = 5.97, p < .001, Cohen's d = 1.07. The contralateral shift is 93% the size of the ipsilateral shift, and not statistically different, one-tailed t-test: t(30) = .36, p > .36, Cohen's d = .06. For the speech-absent 25-min delay, for the ipsilateral case, the baseline of 65.1% /wa/ report dropped to 44.1%, a difference of 21.1%, t(31) = 5.39, p < .001, Cohen's d = .95 For the contralateral case, the baseline of 66.8% /wa/ report was reduced to 52.4%, a difference of 14.4%, t(31) = 3.86, p < .005, Cohen's d = .68. In this case, the contralateral shift is 68% the size of the ipsilateral shift, a difference close to significance, one-tailed t-test: t(31) = 1.59, p = .06, Cohen's d = .28, and very similar to what was found in the Immediate test (i.e., Experiment 1).

Collectively, the results from Experiment 2a and 2b demonstrate that there is surprisingly little reduction in the categorization shift over the course of 25 min: Averaging across Experiments 2a and 2b, for same-ear testing the effects after a delay were 75% of those in the Immediate test, and for different-ear testing that percentage was 93%. These values are quite different than the very small residual effect reported by Sharf and Ohde (1981) for a voiceless adaptor on a voicing continuum.

Figures 4 and 5 present the results for the 90-min delay cases, and it is immediately apparent that the additional hour's delay produced very different results than those seen after the 25-min delay. In particular, while the shifts after 25 min were quite similar to those in the Immediate test, the shifts after 90 min were much smaller. For the speech-present case, for the same-ear condition, the initial 61.7% /wa/ report dropped to 55.2% on the final identification test, a difference of 6.5%, t(29) = 1.95, p = .06, Cohen's d = .36. When the test items were instead contralateral to the adaptor, the initial 65.4% /wa/ report dropped to 64.3%, a difference of 1.1%, t(29) = 0.39, p = .70, Cohen's d = .07.

The 1.1% contralateral effect, far from significance and negligible in standard effect size, suggests that a delay of 90 min might exceed the duration of the adaptation effect in the contralateral ear. The 6.5% ipsilateral effect is at the boundary of significance, leaving it unclear whether there is still some residual adaptation. The difference between the ipsilateral and the contralateral effect again approaches significance, one-tailed t-test: t(29) = 1.59, p = .06, Cohen's d = .29. However, in view of the small-to-medium size of the ipsilateral effect, we will suspend judgment on this question until we see the results of the longer delay, examined in Experiment 3.

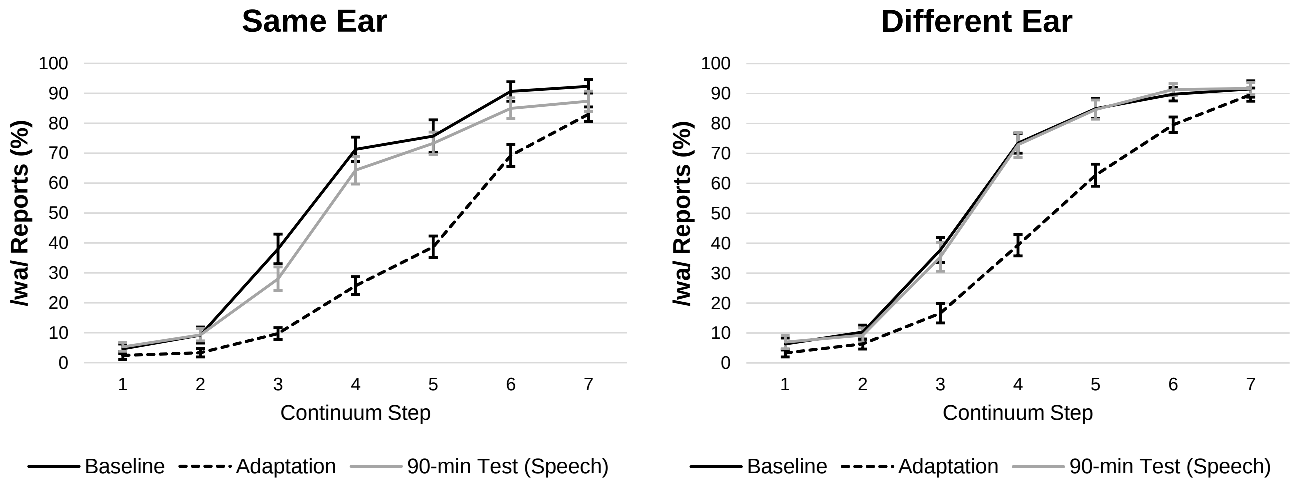

***Figure 4**: Identification of the members of the /ba/-/wa/ test series for the 90-min delay case with speech heard during the delay (Experiment 2c). The left panel shows the results when test syllables were presented to the same ear as the adaptor, and the right panel shows the results when test syllables were presented to the opposite ear. "Baseline" shows identification before adaptation, "90-min Test" shows identification on the final test, and "Adaptation" shows identification within the adaptation phase. Error bars show SEs.*

The effects for the speech-absent 90-min delay were numerically larger. For the ipsilateral case, the baseline of 61.9% /wa/ report dropped to 50.2%, a difference of 11.7%, t(27) = 3.82, p < .001, Cohen's d = .72. For the contralateral case, the baseline of 65.0% /wa/ report was reduced to 57.7%, a difference of 7.2%, t(27) = 2.14, p < .05, Cohen's d = .40. In this case, the contralateral shift is 62% the size of the ipsilateral shift. Though not statistically reliable, one-tailed t-test: t(28) = 1.07, p > .14, Cohen's d = .20, this difference in shift size reproduces again what was found at the Immediate test (i.e., Experiment 1).

We noted that there was little reduction in the categorization shift over the course of 25 min. In contrast, a delay of 90 min severely reduced the size of the shift. Averaging across Experiments 2c and 2d, for same-ear testing the effects after a delay were 34% of those in the Immediate test, and for different-ear testing that percentage was 24%. Note that contrary to the expectation that peripheral effects might fade while more central ones endure, the loss is, if anything, greater for the contralateral condition, a condition that was assumed to reflect central representations.

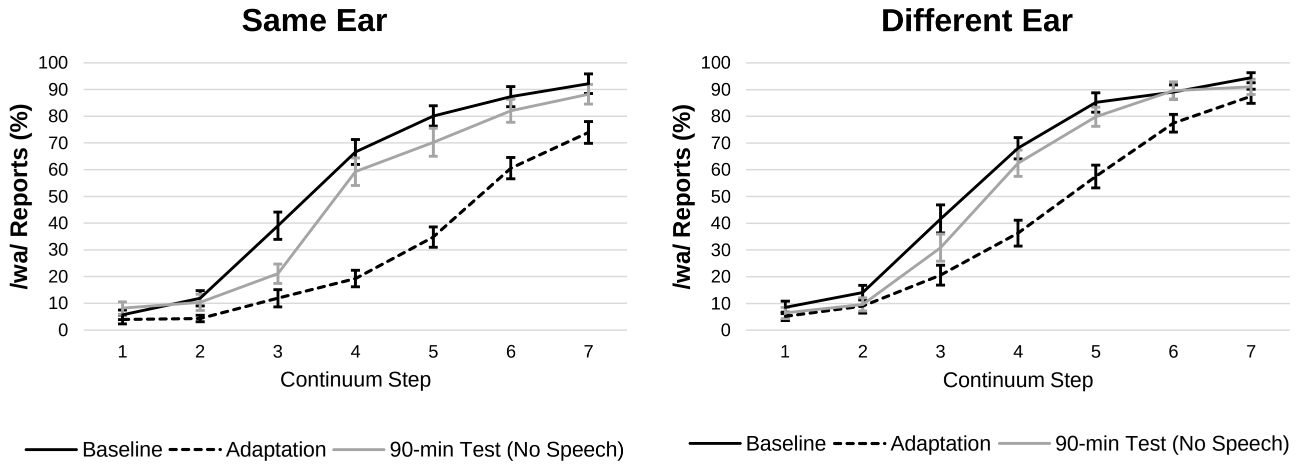

***Figure 5**: Identification of the members of the /ba/-/wa/ test series for the 90-min delay case with little to no speech heard during the delay (Experiment 2d). The left panel shows the results when test syllables were presented to the same ear as the adaptor, and the right panel shows the results when test syllables were presented to the opposite ear. “Baseline” shows identification before adaptation, “90-min Test” shows identification on the final test, and “Adaptation” shows identification within the adaptation phase. Error bars show SEs.*

Recall that one of the central questions of the current study was whether exposure to speech is an important factor in the eventual disappearance of adaptation shifts. For both audiovisual recalibration (e.g., Vroomen et al., 2004) and for lexically-driven recalibration (Liu & Jaeger, 2018), there is evidence that hearing speech tokens reduces the recalibration effect (but see Kraljic & Samuel, 2005, for somewhat different findings). To examine this question here, for both the 25-min and 90-min delays we ran paired experiments – one experiment in each pair exposed participants to speech post-adaptation, and the other experiment minimized speech exposure post-adaptation. Although there seems to be a small trend toward an effect of this manipulation, particularly with the longer delay, the manipulation did not have a robust effect.

We ran two ANOVAs to assess the effect of this factor, one for testing in the ear ipsilateral to the adaptor, and one for testing in the ear contralateral to the adaptor. Each ANOVA included two factors—the speech versus no-speech manipulation, and the length of the delay (25 versus 90 min). In both ANOVAs, the effect of Delay was quite robust, reflecting the big drop in adaptation effects after the longer delay (Ipsilateral: $F(1,117) = 11.042$, $p = .001$, $\eta_p^2 = .086$; Contralateral: $F(1,117) = 13.805$, $p < .001$; $\eta_p^2 = .106$). In both cases, there was no effect of the Speech manipulation (Ipsilateral: $F(1,117) = 1.010$, $p = .32$, $\eta_p^2 = .009$; Contralateral: $F(1,117) = 0.137$, $p = .71$, $\eta_p^2 = .001$), or of the interaction between Delay and Speech (Ipsilateral: $F(1,117) = 0.330$, $p = .57$, $\eta_p^2 = .003$; Contralateral: $F(1,117) = 2.307$, $p = .13$, $\eta_p^2 = .019$).

The results of Experiment 2 provide very useful information with respect to our core questions. Adaptation was only modestly diminished after a 25-min delay, and was still present after a 90-min delay, at least for testing done in the same ear as for the adapting sound. The speech versus no-speech comparisons suggest that the return to the pre-adaptation categorization pattern is not driven by exposure to speech, at least not to speech that comes from a different source than the adapting sounds (see Part 4).

PART 3: Adaptation Effects after a Delay of 5.5 Hours

Experiment 2 showed that although the effects after 90 min were significantly smaller than those observed initially and after a 25-min delay, there were still some significant residual shifts. Given this, in Experiment 3, we imposed a much longer post-adaptation delay before the final identification test, to determine if adaptation is still detectable after a delay of 5.5 hours.

EXPERIMENT 3

Method

Participants

A total of 32 participants took part; none had participated in Experiment 1 or 2.

Stimuli, Apparatus and Procedure

The stimuli, apparatus and procedure for the pre- and post-adaptation identification tests and the adaptation phase were the same as Experiments 1 and 2. During the initial session, they did the initial identification task, followed by the adaptation task. They then left the lab, and returned 5.5 hours later. When they returned, they did the final identification task.

Results and Discussion

Using the same criteria used to identify participants who did not do the task as instructed, data from two participants were eliminated from Experiment 3. Figure 6 presents the results for the remaining 30 participants.

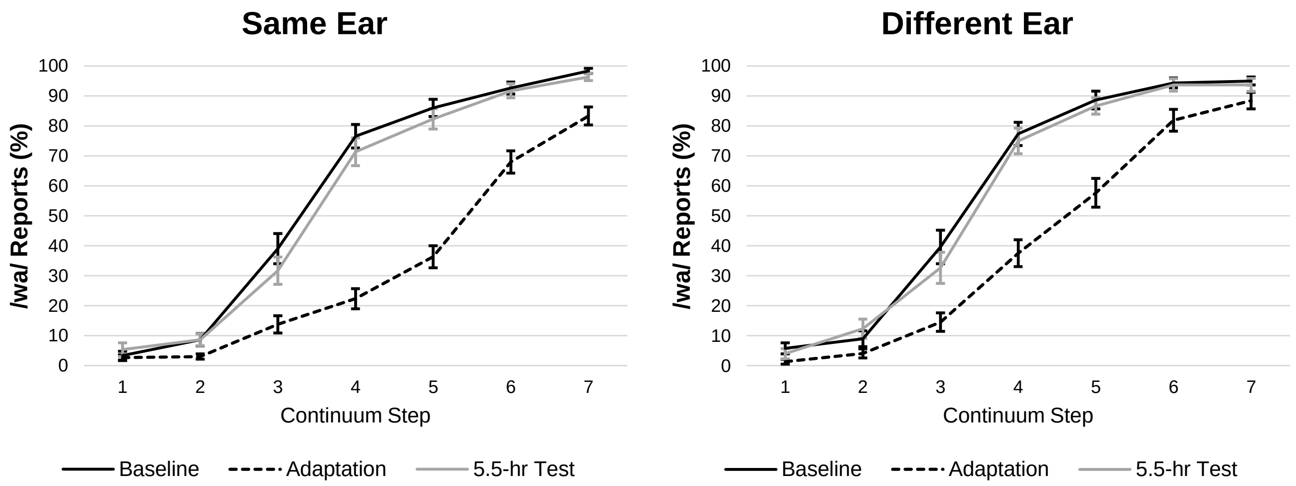

Figure 6*: Identification of the members of the /ba/-/wa/ test series for the 5.5-hr delay case (Experiment 3). The left panel shows the results when test syllables were presented to the same ear as the adaptor, and the right panel shows the results when test syllables were presented to the opposite ear. “Baseline” shows identification before adaptation, “5.5-hr Test” shows identification on the final test, and “Adaptation” shows identification within the adaptation phase. Error bars show SEs.*

During the 5.5 hours that the participants were gone from the lab, we assume that they were exposed to a fair amount of speech (we did not ask them to report what they did). From this perspective, it is interesting that the results in Experiment 3 are quite similar to those for the 90-min delay that included exposure to speech. Recall that in that condition, there was a 6.5% marginally significant shift for the ipsilateral case, and that the effect was nearly zero (1.1%) for the contralateral case. After a 5.5-hour delay, for the ipsilateral case, the baseline of 67.2% /wa/ report dropped to 61.8%, a difference of 5.4%, t(29) = 2.31, p < .03, Cohen's d = .42. Given the marginal statistical effect after 90 min, we had suspended judgment of that 6.5% shift. In light of a similar (5.4%) effect that is in fact both significant and reasonably powerful after 5.5 hours, it seems appropriate to assume that there was indeed still adaptation present after 90 min for the ipsilateral test—this seems much more likely than the effect dissipating within 90 min, and then rebounding at 5.5 hours. For the contralateral case, the baseline of 68.5% /wa/ report was reduced to 64.8%, a difference of 3.8%, t(27) = 1.17, p = .25, Cohen's d = .22. Thus, at both a 90-min and a 5.5-hour delay, any residual adaptation in the contralateral ear is too weak to be measured reliably. Due to the small magnitude of the effects by this point, statistically there was no difference between the contralateral and the ipsilateral shift, t(29) = .44, p > .33, Cohen's d = .08.

PART 4: Adaptation Effects on Three Time Scales

In Part 4, we pull together the results from all six experiments in order to describe the time course of adaptation, for both the ipsilateral and contralateral cases, at three time scales. Just as Cutting (1976) identified qualitatively different levels of processing by examining how different types of dichotic fusion patterned across parametric variation, we examine similarities and differences in how adaptation effects pattern across the combination of laterality and time scales. We first summarize the results shown in Figures 1-6 in order to illustrate how adaptation is reduced over the course of hours. We then focus on the first 50 min of adaptation, in this case beginning with the onset of the adaptation phase, rather than its conclusion. Finally, we drill down to the level of seconds, looking at how the adaptation effect changes within the course of a single block (i.e., one randomization of the seven-item test series that follows a series of adapting tokens), a time scale of approximately 20 seconds.

Adaptation on a Scale of Hours

Figure 7 summarizes the results of the six experiments. The left panel shows the adaptation shift sizes (defined as the difference in /wa/ report between the initial and final identification tests) for items presented in the same ear as the adapting sounds; the right panel shows the corresponding effects for the contralateral case. For the 25 and 90-min delays, the plotted data are averaged across the Speech and No-Speech conditions; recall that there was no significant effect of this manipulation.

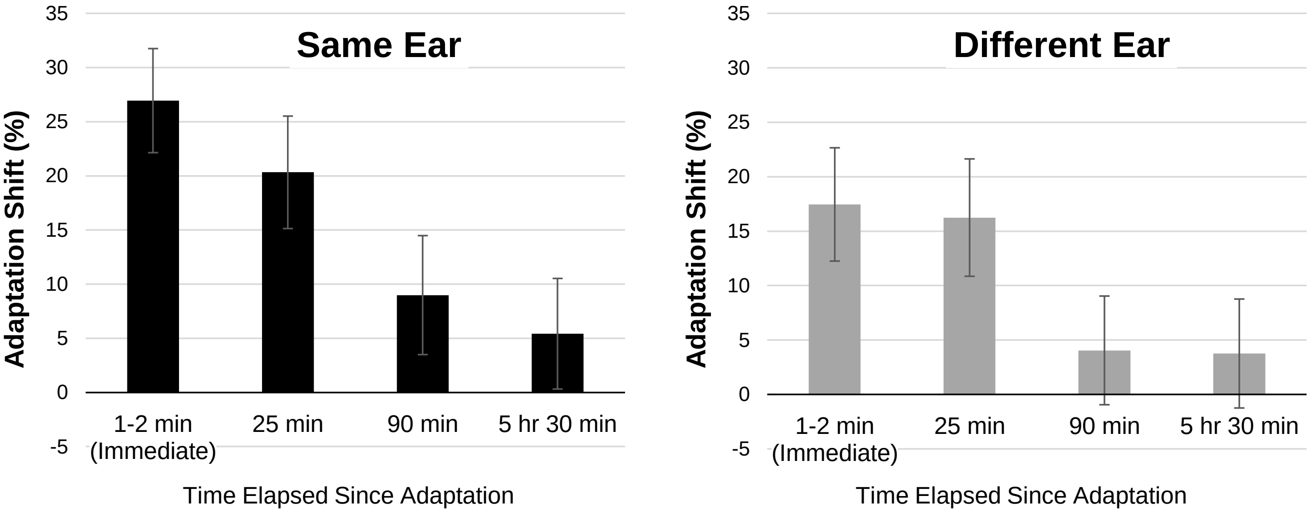

***Figure 7**: Adaptation shifts after four different delays between adaptation and the final identification test. The left panel shows the results when test syllables were presented to the same ear as the adaptor, and the right panel shows the results when test syllables were presented to the opposite ear. For the 25-min and 90-min cases, the plotted data are the averages for the Speech and No-Speech conditions. Error bars show standard error of the effect mean.*

As the analyses in the individual experiments confirmed, same-ear adaptation effects gradually decreased, but they remained stable enough to still reach significance after a 5.5-hour delay. In contrast, when test tokens were presented to the contralateral ear, the adaptation remained stable for 25 min, but at longer delays the small residual effects were not significantly different from zero. These graphs illustrate the quite different duration of adaptation found in the current study than that found by Sharf and Ohde (1981) – effects here were quite robust after 25 min, whereas theirs were negligible by that point. This difference cannot be due to the number of adaptors heard because Sharf and Ohde presented their adapting sound 95 times in each pass, versus the 30 per pass here. Their test was based on a voicing contrast for stop consonants, whereas we used a manner contrast with an adaptor chosen from earlier research to produce both large adaptation effects and a large difference between ipsilateral and contralateral adaptation. This difference underscores the need for a broad empirical base if we are to understand short-term changes in categorization through processes such as adaptation and recalibration. These effects clearly vary for different kinds of speech sounds. Such differences have been a hallmark of speech research from its early inception (e.g., the large differences in the degree of categorical perception across stop consonants, fricatives, and vowels).

Adaptation on a Scale of Minutes

We turn now to a shorter time scale, one that includes the adaptation phase itself (approximately 20 min) and the first 30 min after adaptation ends. Figure 8 shows the relevant results, broken down as usual by whether the identification is of stimuli presented to the same ear as the adaptor, or contralateral to it. Recall that during the adaptation task, participants heard a block of 30 presentations of the endpoint /wa/ in one ear, and then identified one randomization of the /ba/-/wa/ test items in one ear. This sequence was repeated 20 times, with the ear of adaptor presentation fixed while the ear of test items alternated (i.e., blocks 1, 3, 5,…19 to the contralateral ear, and blocks 2, 4, 6,… 20 to the ipsilateral ear). Each of the 20 blocks took approximately 1 min. The left side of Figure 8 tracks the development of the change in phoneme categorization across the 10 blocks of ipsilateral testing and the 10 blocks of contralateral testing.

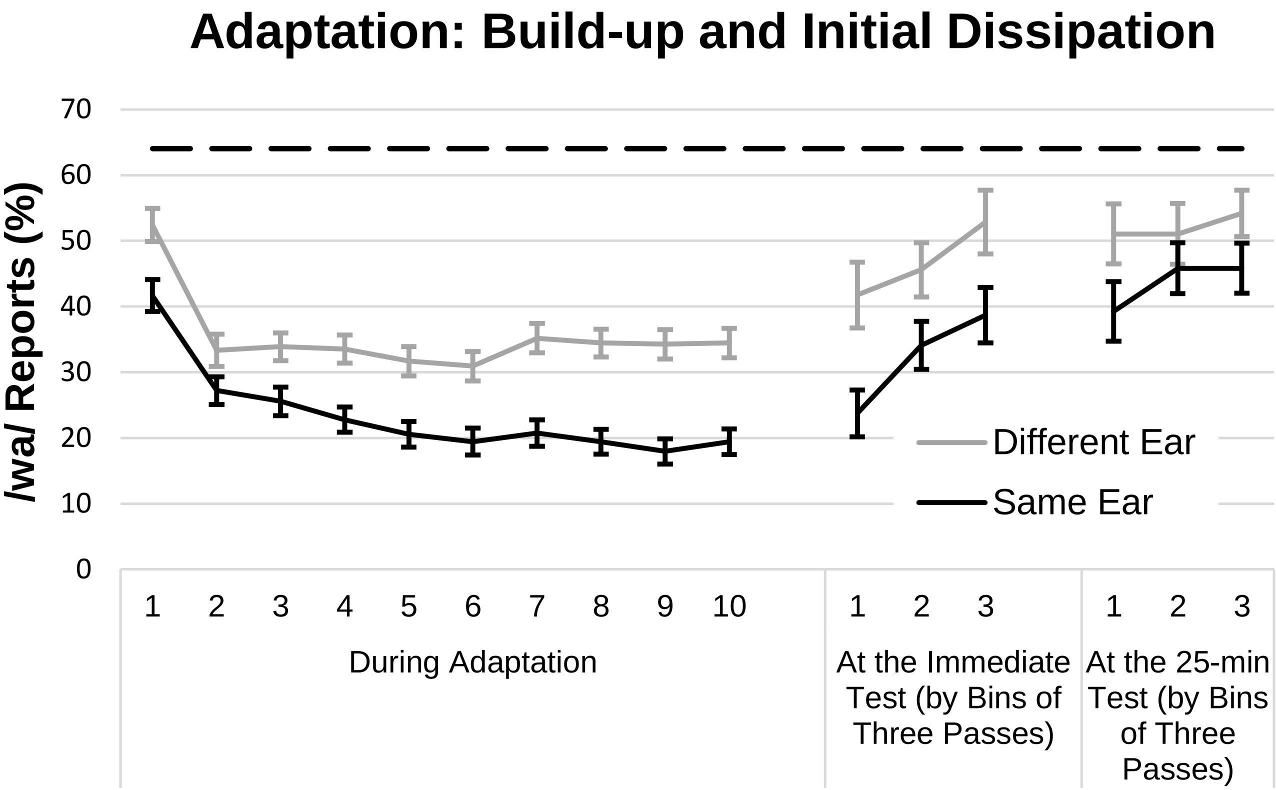

***Figure 8**: Identification averages across approximately 50 min from the beginning of adaptation, broken down by whether the test syllables were presented in the Same Ear as the adaptors or in the Different Ear. The left side of the figure shows identification for each of the 10 adaptation-identification cycles (based on all 180 subjects). The second set of points come from the condition in which the identification test was run immediately after the adaptation phase finished (based on 29 subjects). The final set of points come from the condition in which the identification test was run 25 min after the adaptation phase finished (with little to no speech during that interval, based on 32 subjects). For the identification-task-based points, the first two (practice) passes were skipped, and data were analyzed by bins of three passes. Error bars show SEs. The dashed line at the top of the figure shows the 64% average identification for the initial (pre-adaptation) identification task (averaged across all conditions). This is the level that the system would be at if any adaptation effect has fully dissipated.*

As is clear from the figure, most of the adjustment took place during the first two blocks, particularly for the different-ear case – the function is essentially flat from cycles 2 to 10 in that condition. For the same-ear case, although the strongest adaptation also occurred during the first two blocks, it appears that some additional shifting took place out to perhaps the eighth or ninth block. If we assume that same-ear adaptation is similar to the adaptation one would see if the signal goes to both ears, the results here are generally similar to those reported by Vroomen et al. (2007). Those authors compared adaptation to audiovisually-driven recalibration on an /aba/-/ada/ test series. As we noted previously, they reported that recalibration reached its asymptotic level after only eight exposures, but adaptation increased with increasing exposures out to the maximum of 256 exposures that they tested. Since we had 30 adaptor presentations per block, 256 would correspond to between 8 or 9 blocks here. To the extent that one can compare their curves to Figure 8, the results look similar.

We conducted two ANOVAs on the data shown on the left side of Figure 8. In the first ANOVA, one factor was ipsilateral versus contralateral test Ear, and the second factor was Adaptor Block (1 – 10). Both main effects were robust (Ear: $F(1, 179) = 69.364$, $p < .001$, $\eta_p^2 = .028$; Block: $F(9, 1611) = 18.493$, $p < .001$, $\eta_p^2 = .094$); for the interaction, $F(9, 1611) = 1.842$, $p < .06$, $\eta_p^2 = .010$. Although this analysis is reasonable, it is potentially being overly influenced by the outlying size of the shifts during the very first block. Therefore, we ran the same ANOVA, but excluded the first block. Consistent with what is visible in the figure, the effect of Ear remained robust, $F(1, 179) = 66.929$, $p < .001$, $\eta_p^2 = .27$, but the effect of Block did not, $F(8, 1432) = 1.407$, $p = .19$, $\eta_p^2 = .008$. Perhaps most interesting, the interaction was reliable, $F(8, 1432) = 2.104$, $p < .04$, $\eta_p^2 = .012$ This pattern indicates that the shift in the contralateral ear was essentially constant across Blocks 2-10, while the shift continued to grow in the ipsilateral ear. If we view these results in terms of a peripheral and a central level of processing, the pattern suggests that the central level saturates relatively quickly, while the peripheral level continues to be modified with additional presentations of the adaptor.

The right half of Figure 8 shows the remaining adaptation effects during the Immediate test and during the 25-min delay (No Speech) final identification tests, in each case broken down into roughly the first, second, and last third of each test. For the Immediate test, there is a clear and substantial return toward the baseline (shown with the dashed line); for the later test, there is a trend of that sort, but clearly with a flatter slope. It is interesting that the ending points from the Immediate test (Bin 3) are at about the same level as the initial points from the 25-min test (Bin 1). This pattern suggests that it is the test itself that is driving the categorization back toward baseline. This would account directly for the slope seen within the Immediate test, and for the lack of a bigger return toward baseline seen at the beginning of the 25-min test. In fact, Liu and Jaeger (2018) observed exactly this pattern in their lexically-driven recalibration experiments. They tracked the size of the recalibration shift through the course of the final identification test, and found that the effect was substantial at the beginning of that test, but became much smaller through the course of the test. On the other hand, this pattern was not evident in the adaptation study by Vroomen et al. (2004). In their study, after 50 presentations of an adaptor (audiovisual /aba/ or /ada/), they presented 60 test items from their /aba/-/ada/ stimuli. There was very little change in the size of the adaptation shift through the course of the 60 items (about 2.5 min). On balance, the data do suggest that hearing a balanced set of stimuli (which is what is heard with test items) in the same voice as the exposure items will reduce the adaptation effect. The pattern seems to be very similar for the ipsilateral and contralateral conditions.

Adaptation on a Scale of Seconds

Our final temporal window of analysis is much shorter – the time that it took to present one randomization of the 7-item test series, about 15-20 seconds. Figure 9 shows these results for the ipsilateral and contralateral conditions, using data from all 180 participants and all 20 passes.

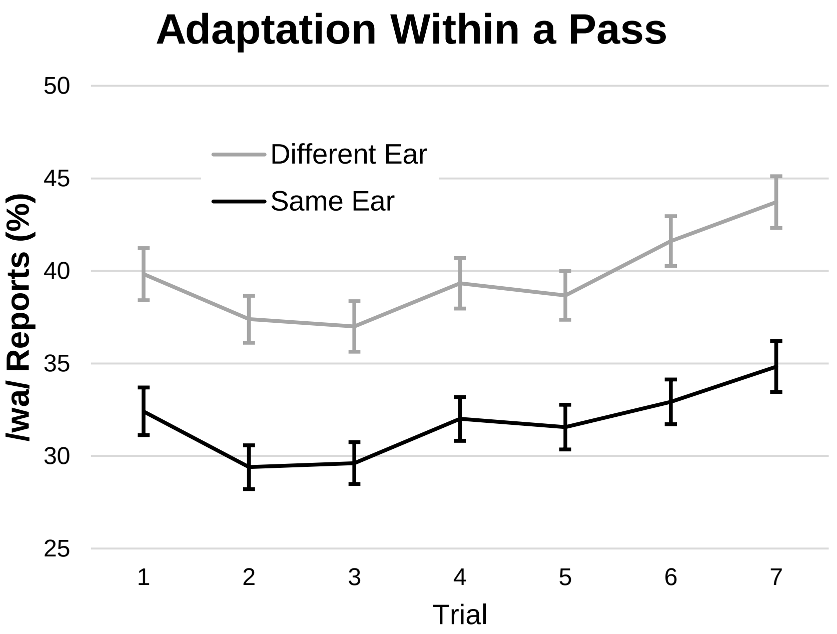

***Figure 9**: Identification percentages for the seven stimuli within one adaptation pass, ordered by whether the test item was the first one after adaptation finished, the next one, …, the last one. Note that unlike the rest of the data presentation, these results are for all seven members of the /ba/-/wa/ test series, rather than just for the middle three items. The averages are based on all 20 passes for each listener. As shown in Figure 8, average pre-adaptation /wa/ identification was 64%. Error bars show SEs.*

As usual, there is a robust difference as a function of whether testing was done in the same ear as the adapting sound. For test positions 2-7, there is a clear trend back toward baseline, though in absolute terms this is rather small (about a 5% recovery). The results for the first test item heard after the adaptation sequence ended are a bit odd—the first point is a few percent higher than the second point. Based on the overall shape of the functions, we would expect the first point to be the lowest one. Moreover, the literature on “contrast effects” (e.g., Diehl, 1981) would predict that the first point would have an extra push downward.

Given these two facts, our interpretation is that this first point is contaminated by occasional confusion by participants in which their attention lapses during the repetitive adaptation sequence. If a person occasionally responds randomly to the first item after the repetition because of such a lack of attention, that would essentially be mixing 50% on those trials together with the true value. Such noise would be sufficient to move the true value up by a few percentage points, yielding the observed results.

To assess the results shown in Figure 9, we ran an ANOVA with two factors – same versus different Ear of the test items, and Position within the testing sequence (first item presented, second item presented, … seventh item presented). Both main effects were robust (Ear: $F(1, 179) = 69.282$, $p < .001$, $\eta_p^2 = .28$; Position: $F(6, 1074) = 7.004$, $p < .001$, $\eta_p^2 = .038$). As Figure 9 suggests, the functions for the two Ear conditions were parallel to each other, producing a non-significant interaction, $F(6, 1074) = 0.214$, $p > .97$. $\eta_p^2 = .001$. A comparable ANOVA that omitted the suspect first position produced essentially identical results.

We are only aware of one prior adaptation study that included an analysis of the short-term dissipation of adaptation effects (other than the Vroomen et al., 2004, study mentioned previously), and the testing conditions were somewhat different. Jamieson and Cheeseman (1986) used the endpoints of a /da/-/ta/ test series as adaptors, and included the ipsilateral versus contralateral contrast. After each sequence of adaptors, they presented four randomizations of their 7-item /da/-/ta/ test series, and they looked at the adaptation shift for each of the four randomizations (with each such randomization taking approximately 15 seconds). For contralateral tests, the adaptation effect was essentially the same across the four randomizations, both for /da/ and for /ta/ as adaptors. They found a similar result for /ta/

in the ipsilateral condition. However, when /da/ was the adaptor, there was a very large ipsilateral adaptation effect during the first randomization of the seven test items, followed by results similar to the other cases after that. Given the differences in stimuli and procedure, and the somewhat odd single data point, it is not clear what to make of their results.

Our own results, shown in Figure 9, show a very clear but fairly small reduction in adaptation during the 15-20 seconds after the adapting sequence ends. The source of this change could either be the simple passage of time, or it could be the effect of hearing non-endpoint test stimuli (as was discussed in the context of Figure 8). The latter seems more likely, though the data in hand do not allow us to definitively reject the alternative.

GENERAL DISCUSSION

In the current study we set out to answer three questions that bear on the processes listeners use to adjust their long-term phonemic categorization criteria when the prevailing speech environment differs from the long-term distribution of speech sounds. We will discuss each of these in turn.

(1) For how long can adaptation effects be detected? The first question was stimulated by the almost complete absence of data on how long adaptation lasts: When the short-term statistics are inconsistent with the prior long-term distribution, for what period of time does the system modify its categorization of the affected sounds? The only prior study we are aware of suggested that the modification gradually gets smaller and is largely gone within a half hour (Sharf & Ohde, 1981). Our results provide a very different answer: The modified categorization can persist for at least 5.5 hours. As we noted above, finding different time courses for different types of speech sounds (voiceless adaptors on a voicing continuum

versus continuants on a stop-continuant continuum) is not surprising, but is informative. We would expect to find time courses shorter than what we have observed for adaptors that are not as effective as our carefully-chosen /w/, and it is possible that somewhat longer time courses might be found with an even more effective adaptor.

Recall that adaptation is one of two well-established ways in which phonemic categorization changes as a function of the short-term distribution of speech sounds in the environment. Adaptation results in a reduction of report of sounds that are similar to a clear exemplar that is greatly over-represented in the current environment. Recalibration, in contrast, occurs when the environment includes several ambiguous tokens that are accompanied by disambiguating contextual information; the adjustment leads to an increase, rather than a decrease, in report of sounds related to the sound that violates previous norms. Prior recalibration studies have shown that the disambiguation can either be provided by visual speech (e.g., Bertelson et al., 2003) or by lexical context (e.g., Norris et al., 2003). The contrasting patterns echo those found in sensory systems more generally, with contrastive changes often following clear exemplars, and attractive changes often affecting more ambiguous stimuli (Snyder et al., 2015).

Our finding that adaptation effects last on the order of hours can be compared to what is known about the durability of the two versions of recalibration. The literature suggests that visually-driven recalibration is a very local effect. Not only does it reach asymptote with only eight exposure tokens (Vroomen et al., 2007), it also disappears very quickly. The effect drops during the first few test items, and is gone after about a half-dozen such exposures, a matter of tens of seconds (Vroomen et al., 2004). Lexically-driven recalibration lasts much longer. Kraljic and Samuel (2005) showed that the effect was fully intact after a 25-min delay.

Eisner and McQueen (2006) extended the timeline considerably and found that significant recalibration was present after a 12-hour delay, and that the size of this effect was not affected by whether the 12 hours included sleep or not. Zhang and Samuel's (2014) results suggest an upper bound on the duration, with no effects remaining after a one-week delay. We (Zheng & Samuel, in preparation) have recently examined the duration question by testing two different contrasts at three different delays. The results are consistent with those of Eisner and McQueen (2006) and Zhang and Samuel (2014), with robust recalibration present after a 24-hour delay, but not after one week. Collectively, the duration data show quite different patterns for the three types of adjustment: Visually-driven recalibration only affects speech categorization for at most a matter of minutes, adaptation changes the system for several hours, and lexically-guided recalibration operates for days. It is not immediately clear how the Kleinschmidt and Jaeger (2015) "belief-updating" model can account for these differences in recovery time, as it treats recalibration and selective adaptation as products of the same adjustment function. In contrast, dual accounts that allow exposure to trigger adjustments at more than one level (Snyder et al., 2015) would have the required flexibility.

(2) Is Hearing Speech Critical for Undoing Adaptation? The second core question of the current study was whether the return to a listener's long-term categorization pattern is simply a matter of time, or is instead a matter of continuing experience with the statistical distribution of sounds in the environment. To address this question, for the 25 and 90-min delay conditions we compared the adaptation effects found when these delays were filled with speech (and thus provided new distributions) or not (and thus provided no new statistical information to the system). Perhaps surprisingly, there was no reliable effect of this manipulation – the remaining adaptation effects were not significantly different for the two

cases. On its face, this result would seem to suggest that time is more important than additional information about the distribution of sounds.

However, the results of our finer-grained temporal analyses lead us to question that conclusion. In particular, the right side of Figure 8 shows two relevant results. First, within the final identification test in the Immediate condition (Experiment 1), there is a very marked return toward baseline through the course of the several minutes of that test. Second, when the final identification test of the 25-min delay condition begins, identification has not moved any closer to baseline, even though an additional 20 min have gone by. Taken together, these results suggest that it is exposure to the test items, more than time, that is gradually eliminating the adaptation shift. This interpretation is bolstered by the results of the even finer-grained temporal analysis shown in Figure 9, where there is a clear movement toward baseline over the course of the test items within each adaptation block. As noted above, Liu and Jaeger (2018) reported exactly this effect after lexically-driven recalibration. Note that this effect could contaminate measurement of adaptation effects across multiple sessions if a study used a within-participant design. This concern motivated the between-participant approach here. The presence of a testing effect *within a session* suggests that exposure to speech tokens that span the testing range plays a significant role in the recovery process. Thus, in phonetics as in other domains where adaptation has been found, increased sampling of the alternatives along the dimension of interest speeds up the return to a pre-adaptation state (e.g., Leopold et al., 2005).

How can we resolve the apparent contradiction between the non-significant effect of the speech versus no-speech manipulation, and the effect of hearing speech that is seen in the finer-grained temporal analyses? One option would be to assume that speech exposure matters over the short-term, but not on a longer time scale. This is possible, but seems unlikely. Instead, we suggest that exposure to speech is important for remodifying the categorization process, but that this is primarily based on within-talker statistics. Note that what listeners were hearing within a final identification test was speech from the same (synthetic) talker, with tokens spanning the whole continuum range, whereas the voices heard during the longer intervals were not. There are thus two properties that differed: The effective case included sounds from the same source/talker, and it included the full range of sounds. Our design does not allow us to choose between these two factors, but there are findings in the literature that support the role of source/talker.

If listeners change their categorization of speech sounds on a per-talker basis, then hearing other voices should not have a substantial effect. In fact, in their study of lexically-driven recalibration, Kraljic and Samuel (2005) demonstrated that hearing “normal” tokens of sounds that had generated recalibration had no effect if these subsequent normal tokens were from a different talker, but that hearing such normal tokens did reduce the recalibration if they were from the same talker (cf. Eisner & McQueen, 2005). This remodification apparently requires quite a bit of new evidence, compared to the original recalibration, as a hundred good tokens only weakened, but did not eliminate, the effect generated by twenty ambiguous ones (see Kraljic et al., 2008 for further evidence of the primacy of earlier-arriving information). Although within-talker adjustments seem to be important, this does not preclude more general effects under some circumstances. For example, adaptation in one voice (e.g., a female voice) can cause shifts on test items in another voice (e.g., a male voice; see Bowers et al., 2016, for an example of this). It would be very interesting to know whether initial modifications to the long-term statistics can be induced on a non-talker-specific basis, but that the return back toward “normal” is more dependent on talker-specific input.

(3) Do Adaptation Effects Dissipate Differently for Ipsilateral than for Contralateral Cases? The third major question of the current study related to two qualitatively different levels of speech processing that have been posited on the basis of previous studies that included a comparison of ipsilateral to contralateral adaptation effects. These studies produced results that implicated one level of processing (“peripheral”) that takes input from one ear and does analyses that are relatively closely tied to the acoustic properties of the input. The second level (“central”) receives input from both ears, and does analyses that abstract across acoustic and/or phonetic properties (e.g., this level tolerates frequency shifts that preserve the general directions of frequency change (Sawusch, 1977), and tolerates qualitative changes such as whether sounds are normally voiced or are instead whispered (Samuel, 1988)).

The focus of our investigation of these two putative levels was whether ipsilateral tests (assumed to tap both the peripheral and central levels) and contralateral tests (assumed to tap only the central level) would produce comparable patterns of adaptation over time. If there are indeed two qualitatively different types of processing, then we might expect to see differences in their time courses. At the longest time scale, we did observe different patterns. As Figure 7 shows, over the course of 5.5 hours, ipsilateral testing yields a gradual reduction of the adaptation effect, with a significant residual effect at the longest interval tested. In contrast, not only was the immediate effect smaller in contralateral testing (consistent with the results summarized in Table 1), it was only reliable out to 25 min; neither the 90-min nor the 5.5-hour conditions yielded significant remaining adaptation. The larger and longer-lasting effects for same-ear tests are consistent with adaptation occurring at both levels, and with the peripheral processes returning to baseline more slowly. Interestingly, for visual motion aftereffects, Favreau (1976) also reported larger and longer-lasting effects for same-eye tests versus between-eye tests.

When we examined the onset of adaptation, shown on the left side of Figure 8, there were both similarities and differences for the ipsilateral/contralateral cases. For both, the bulk of the adjustment process took place within the first minute or two of adaptation. However, at that point the contralateral case leveled off, with no growth in adaptation after that. In contrast, the size of ipsilateral effects continued to grow for most of the adaptation phase. If the contralateral test specifically taps a central processing stage, that stage saturates quite quickly, while the peripheral level of analysis continues to be modified by additional sequences of the adapting sound.

There are two aspects of the time course for which the ipsilateral and contralateral conditions produced strikingly parallel results. First, when we look at the return toward baseline shown in the right side of Figure 8, the two functions are rising at the same rate. Second, when we look at the return toward baseline within an adaptation block (Figure 9), we again see impressively similar slopes for the two conditions. Recall that the ipsilateral test is assumed to reflect the operation of both peripheral and central processes, while the contralateral test only is sensitive to central processes. Given this, the most parsimonious account of both findings of parallelism is that we are seeing how the central processes recover from adaptation as they are exposed to test stimuli with a balanced distribution of sounds (i.e., the functions are parallel because they reflect the recovery of the same component that is operating in both tests). Note that if the central processes do return toward their long-term values more quickly (as this account would suggest), that would lead to the pattern we found over the longest time course (Figure 7), in which the centrally-driven contralateral effects only remained reliable through the first 25 min (while the peripherally + centrally driven effects for ipsilateral testing were still reliable at 5.5 hours). A priori, one might have expected that more central representations (i.e., those based on more abstract properties of the signal) would last longer than ones more directly tied to the acoustics, but the results at all three temporal scales suggest otherwise.

Overall, our combination of the ipsilateral/contralateral manipulation with analyses at three different time scales provides additional empirical support for the theoretical distinction between monaurally-driven analyses and binaurally-driven analyses. In light of arguments made recently by one of us (Samuel, 2020), it is worthwhile to briefly outline the possible structure of these multiple levels of analysis. A key argument in that paper was that, in general, it is not a good research strategy to take a representational unit in linguistic theory and conduct experiments to “prove” that the unit plays a critical role in speech perception. Such a strategy has a 50-year history of failure. It is not that linguistic units cannot play a role in perception; rather, it is that many different units are possible, and there has not proven to be any special role in speech perception for linguistically-grounded ones. The adaptation literature provides a nice illustration of an alternative approach, as does a set of non-adaptation experiments reported by Cutting (1976) (see below).

The results summarized in Tables 1-3 provide broad-based empirical support for positing qualitatively different levels of analysis, with monaurally-driven analyses being more sensitive to the acoustic details than analyses done on binaurally-available input. Sawusch’s (1977) results offer a particularly elegant demonstration of this distinction: By shifting the adaptors' formant patterns up in the frequency domain so that the energy was being fed into different critical bands (a basic defining property of the auditory system) than the test syllables, Sawusch halved the efficacy of the sounds as adaptors compared to the non-shifted adaptors. This illustrates the importance of matching acoustics at the peripheral level. The fact that these shifted adaptors were just as effective under contralateral presentation as under ipsilateral shows that the binaurally-driven processes are not tied tightly to the acoustics, and instead are affected by more abstract patterns.

In order to flesh out the properties of the posited central-level representations, Samuel and Kat (1996) adopted an additional methodological tool that Samuel (1986) had introduced. Samuel (1986) had used reaction time analyses to complement the more typical identification percentages because the reaction times can be used to look for adaptation effects within a phonetic category, rather than just focusing on a shift in the boundary between two sides of a contrast (cf. Miller, 1975).

The key comparison in the Samuel and Kat (1996) study, using both identification percentages and reaction times, was between their "F2F3" adaptors and their voiceless adaptors (/pa/ and /ta/). Their test series was /ba/-/da/, and the F2F3 adaptors were versions of the endpoint /ba/ and /da/ in which there was no first formant. Stimuli with only the second and third formants have been used as "nonspeech" stimuli in speech research because they do not have a clear phonetic quality. In contrast, the /pa/ and /ta/ adaptors have clear phonetic qualities and differ from each other in the same way that /ba/ and /da/ differ – the first member of each pair is labial, and the second is alveolar. Thus, we can classify the F2F3 stimuli as sharing complex acoustic properties with the test series, but not phonetic quality; the voiceless syllables share phonetic qualities with the test series but are acoustically rather different because of the lack of voicing. The critical results were that (1) both types of adaptors produced substantial adaptation effects using the standard percentage identification measure; (2) only the F2F3 adaptors also produced the corresponding reaction time changes; and (3) the same results were obtained under contralateral testing.

The fact that both adaptors types worked contralaterally as well as ipsilaterally implicates a central locus. The fact that one type of adaptor generates only boundary shifts, whereas the other type also affects tokens throughout the category, indicates that the two types are engaging different types of (central) processes. Given the relationships between the two adaptor types and the test series, we could say that the F2F3 adaptors engaged central representations that are sensitive to complex acoustic patterns, and that the voiceless adaptors engaged central representation that are sensitive to phonetic information. Together with the monaurally-driven analyses, this yields a model with a peripheral level that responds to detailed acoustic properties, a central level that responds to abstract/complex acoustic properties, and a central level that responds to phonetic properties. Note that rather than starting with linguistically-defined phonetic units, this research program identified units of this sort, along with non-linguistic units, based on the pattern of results that emerged from a broad set of experiments.

The same is true of the levels of analysis identified by Cutting (1976) through his examination of six different ways that sounds can fuse when different input is given to each ear. His approach was to identify qualitatively different combination situations, and to characterize how each situation was affected by variations in onset time, amplitude, or frequency differences between the two ears. Using this approach, he identified three different classes of central processing. The simplest of these three is the sound localization process, in which the same input goes to both ears, but one ear receives the input slightly later (and perhaps with slight attenuation). This leads to the perception of a single stimulus, with its location determined by the interaural delay. The delay must be extremely brief for fusion to occur, and any frequency difference must be very small, allowing this level to be separated from the other levels.

Three other types of dichotic fusion share similar sensitivities to variation of onset time, amplitude, or frequency differences, leading Cutting (1976) to consider them as evidence for a second central level. Psychoacoustic fusion (e.g., /ba/ in one ear, plus /ga/ in the other, yielding a percept of /da/), spectral fusion (the first formant to one ear, and the second to the other ear, yielding an intact syllable), and spectral/temporal fusion (the second formant transition to one ear, the rest of a syllable to the other ear, yielding a percept of the full syllable in one ear and a percept of the formant transition (chirp) in the other) all show similar tolerance for variations in the three acoustic properties. Thus, Cutting treats all of these as forms of integrating acoustic features. This level seems to be comparable to the level implicated in the adaptation studies that we called “abstract/complex acoustic” processing.

Finally, Cutting grouped two other types of fusion together because of their similar sensitivity to the three acoustic parameters. Phonetic feature fusion occurs when, e.g., /ba/ goes to one ear and /ta/ goes to the other, yielding a percept of either /da/ or /pa/. These percepts could be produced by taking the voicing feature from one input and fusing it with the place of articulation from the other input. In phonological fusion, e.g. /da/ to one ear, and /ra/ to the other, yielding /dra/, the whole set of features of the /r/ get merged with the /da/ syllable, but this obeys the phonotactic rules of the language that only allow the /r/ to come after the /d/. Cutting classified these final two types of fusion as ones that involve the disruption and recombination of phonetic features.

Thus, setting aside the less interesting case of localization, Cutting's analysis of a large literature on diverse forms of dichotic fusion produced two emergent central levels. One operates on complex acoustic patterns, and the other operates on phonetic codes. As with the adaptation literature, the research did not start with a linguistic unit and try to verify it. Instead, looking across a wide and rich body of data, two qualitatively different levels of analysis emerged. These two levels from one large literature appear to map rather closely to the two central levels that emerge from the large adaptation literature. This is exactly the kind of convergence that one hopes for in a scientific enterprise.

The results of the six experiments in the current study complement those in the existing adaptation literature. They provide additional support for the distinction between a peripheral level of speech analysis, and a central level. In this case, the distinction is grounded in the different time courses we observed for recovery from adaptation. Recent work on phonemic recalibration, an assimilative process, has shown a very short recovery period for audiovisually-based recalibration (a few minutes, at most), and a quite long period for lexically-driven recalibration (longer than a day, shorter than a week). The current study provides evidence that speech adaptation (a contrastive effect) falls in between these two ranges, lasting for several hours when the initial effect is driven by a very strong adaptor. Models of recalibration and adaptation (e.g., Kleinschmidt & Jaeger, 2015) will need to demonstrate that they can accommodate the temporal patterns now available in the literature.

Author Note

This work was supported by grants from the Economic and Social Research Council (UK) (grant ES/R006288/1) and from the Spanish Ministery for Science and Innovation (grant PSI2017-82563-P), by the Basque Government, through the BERC 2018-2021 program, and by the Spanish State Research Agency, through the BCBL Severo Ochoa excellence accreditation (SEV-2015-0490). The data from this project are available on the OSF repository: https://osf.io/2fjkb/?view_only=265878a996174dfab3c6d74bb563da0f

REFERENCES

Ades, A. (1974). How phonetic is selective adaptation? Experiments on syllable position and vowel environment. *Perception & Psychophysics*, *16*, 61-66.

Barlow, H.B., & Hill, R.M. (1963). Evidence for a physiological explanation for the waterfall phenomenon and figure aftereffects. *Nature*, *200*, 1345-1347.

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., Kazanina, N., & Andermane, N. (2016). Spoken word identification involves accessing position invariant phoneme representations. *Journal of Memory and Language*, *87*, 71-83.

Clayards, M.A., Tanenhaus, M.K., Aslin, R.N., & Jacobs, R. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108 (3), 804-9.

Cutting, J.E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, *83*, 114-140.

Diehl, R. (1981). Feature detectors for speech: A critical reappraisal. *Psychological Bulletin*, *89*, 1-18.

Earle, F.S., & Myers, E.B. (2015). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41 (6),* 1680-1695.

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 the Acoustical Society of America*, *119*, 1950-1953.

Favreau, O.E. (1976). Motion aftereffects: Evidence for parallel processing in motion perception. *Vision Research*, *16*, 181-186.

Ganong W.F. (1978). The selective adaptation effects of burst-cued stops. *Perception, & Psychophysics*, *24*, 71-83.

Jamieson, D.G., & Cheeseman, M.F. (1986). Locus of selective adaptation in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *12(3)*, 286-294.

Kaiser, D., Walther, C., Schweinberger, S.R., & Kovacs, G. (2013). Dissociating the neural bases of repetition-priming and adaptation in the human brain for faces. *Journal of Neurophysiology, 110,* 2727-2738.

Kilian-Hutten, N., Vroomen, J., & Formisano, E. (2011). Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. *Neuroimage, 57,* 1601-1607.

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., & Samuel, A.G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141-178.

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.

Liu, L., & Jaeger, T.F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55-70.

Leopold, D.A., Rhodes, G., Muller, K.M., & Jeffery, L. (2005). The dynamics of visual adaptation to faces. *Proceedings of the Royal Society: B., 272,* 897-904.

Maye, J., Werker, J.F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-11.

McCollough, C. (1965). Color adaptation of edge-detectors in the human visual system. *Science*, *149*, 1115-1116.

Miller, J. (1975). Properties of feature detectors for speech: Evidence from the effects of selective adaptation on dichotic listening. *Perception & Psychophysics*, *18*, 389-397.

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

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

Samuel A. (1988). Central and peripheral representation of whispered and voiced speech. *Journal of Experimental Psychology: Human Perception and Performance*, *14(3)*, 379-388.

Samuel, A.G. (1989). Insights from a failure of selective adaptation: Syllable-initial and syllable-final consonants are different. *Perception & Psychophysics*, *45*, 485-493.

Samuel, A.G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive Psychology*, *88*, 88-114.

Samuel, A.G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 1-12.

Samuel, A.G., and Kat, D. (1996). Early levels of analysis of speech. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 676-694.

Sawusch, J. (1977). Peripheral and central processes in selective adaptation of place of articulation in stop consonants *Journal of the Acoustical Society of America*, *62*, 738-750.

Schweinberger, S.R., Casper, C., Hauthal, N., Kaufmann, J.M., Kawahara, H., Kloth, N., Robertson, D. M.C., Simpson, A.P., & Zäske, R. (2008). Auditory Adaptation in Voice Perception, *Current Biology, 18,* 684-688.

Schwiedrzik, C.M., Ruff, C.C., Lazar, A., Leitner, F.C., Singer, W., & Melloni, L. (2014). Untangling perceptual memory: hysteresis and adaptation map into separate cortical networks. *Cerebral Cortex, 24,* 1152-1164.

Sharf, D.J., & Ohde, R.N. (1981). Recovery from adaptation to stimuli varying in voice onset time. *Journal of Phonetics*, *9*, 79-87.

Snyder, J.S., Schwiedrzik, C.M., Vitela, A.D., & Melloni, L. (2015). How previous experience shapes perception in different sensory modalities. *Frontiers in Human Neuroscience*, *9*, 594-603.

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.

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.

Zheng, Y., & Samuel, A.G. (in preparation). The time course of lexically-driven speech recalibration.

Appendix

Approaches to Measuring a Shift in Identification

In selective adaptation studies, the standard procedure is to have listeners identify members of a test series under different conditions: (1) after adaptation with an adaptor intended to affect one side (e.g., one endpoint of the continuum); (2) after adaptation with an adaptor intended to affect the other side (e.g., the other endpoint); or (3) in the absence of adaptation (the "baseline"). Studies vary in whether all three of these are done, versus just two of them. For example, in the current study, we used only the /w/ side of the /b/-/w/ contrast, and compared it to a no-adaption baseline

When the two or three conditions have been run, experimenters need to determine whether a pair of conditions yielded differences in the identification of the test items. Adaptation is actually just one of several paradigms that require this step. For example, in the Ganong (1980) paradigm, members of a continuum are often tested to see whether lexical context causes a shift (e.g., are the items in a /g/-/k/ continuum identified differently in "gift"-"kift" than in "giss"-"kiss"?). Similarly, in both the lexical recalibration paradigm (Norris et al., 2003) and the audiovisual recalibration paradigm (Bertelson et al., 2003), the goal is also to determine if members of a test continuum are labeled differently in one condition than the other.

There have been hundreds of studies across the various paradigms with this property. Given this, it should not be surprising that more than one approach has been used to measure a shift. In the current study, we use an approach that the first author has developed in over two dozen papers: Present the full continuum to listeners for identification, but only use a few items from near the middle of the continuum in the shift analysis. We will briefly review the approaches that have been taken in the literature, and explain why we believe the approach used here is the best option. To be clear, this is not the only good option, and in most experiments the conclusions will be the same across different ways of measuring the shift.

Most of the measurement approaches can be sorted on the basis of three factors: (1) which members of the test series are presented to listeners, (2) which members of the test series are used in the shift analysis, and (3) what method is used to do the statistical test. We will briefly consider each of these.

Which members of the test series are presented to listeners? Looking across the hundreds of studies, the most common choice is probably to test listeners on the full set of stimuli. In this approach, the experimenters typically generate between about six to ten stimuli, evenly spaced, such that one endpoint item is identified as one category at or near 100% of the time, and the other endpoint item is identified as the other category at or near 100% of the time. A limitation of this approach is that often, several of the test items produce ceiling performance and thus provide little useful information. For this reason, some experimenters reduce the number of test stimuli by only using items closer to the midpoint. There are multiple versions of this approach. Some researchers just choose four or five items near the middle of the continuum (e.g., Norris et al., 2003), perhaps with an expected identification range between about 30% and 70%. Sometimes researchers choose three or four items in this midrange, but also include the continuum endpoints, skipping items in between. The most extreme approach (e.g., Bowers et al., 2016) is to pick a single item intended to be the most ambiguous member of the test series. A better version of this approach is to do a pretest for each listener to identify the most ambiguous item by-subject, and to use that item plus one item “below” it and one item “above” it in the measurement (e.g., Vroomen et al., 2004).

The decisions for this design factor involve tradeoffs in terms of testing time, individual differences in listeners’ category boundaries, and the “naturalness” of the listeners’ experience. Generally speaking, if listeners only hear tokens that really are not reasonable exemplars of either category (i.e., only items in the middle), the task is relatively unnatural. If a reasonable number of items are clearly instances of one category, and a reasonable number of items are clearly instances of the other, then judgments of the middle items are more likely to reflect “normal” processing. Given the substantial individual differences that typically exist in where people’s category boundaries are, when only middle items are used, some listeners will either produce ceiling/floor identification, or may force their judgments into categories that they really are not hearing.

For these reasons, we believe that presenting the full set of stimuli is preferable. These tests are generally short enough that the extra testing time is just a few minutes. Using the full range also provides an important way to identify participants who cannot or will not do the task: If the endpoint items are not consistently identified differently, it makes no sense to include a listener’s results in the analyses. This “sanity check” for each listener is not available if only ambiguous items are tested.

<u>Which members of the test series are used in the shift analysis?</u> When only the items near the middle of the continuum are presented to listeners, there really is no decision to be made — identification of those items will be the measure. However, when the full set of items is presented, the experimenter must decide whether to base the analyses on all of the items,

or on a subset. This decision may interact with the analysis method (see below). In general, it is preferable to focus the analysis on the items that provide the most relevant information. On these grounds, it is best to use a predetermined subset that excludes steps that are likely to be near ceiling or floor levels. In general, we use the middle four items when the test continuum has six, eight, or ten steps, and we use the middle three items when the test continuum has seven or nine steps. Because shifts are typically (though not always) most robust for ambiguous stimuli, focusing on these stimuli provides a sensitive measure. Thus, combining this decision with the first one, our approach gives listeners a relatively natural task (by including clear tokens of both categories), while basing the analyses on items most likely to reflect the theoretically relevant differences.

What method is used to do the statistical test? A number of different approaches can be found in the literature. In the earliest work in selective adaptation, researchers often fit a curve (an ogive) to the identification function in each condition, and compared the curve parameters. Over time, researchers moved toward simpler measures, often just averaging the identification scores and doing a t-test. Currently, the most common approaches to compare the two conditions are t-tests, analyses of variance, or linear mixed effects models. In the latter two, researchers typically include step in the continuum as a factor, and when this approach is used, the analysis is based on all of the tested steps of the continuum.

We generally prefer to do a simple t-test, using the average identification across the middle items (e.g., the middle three steps of the seven-step continuum used in the current study). This measure has proven to be extremely robust across the dozens of experiments we have run over the years. The score has two important advantages. First, each score is based on enough observations to be very stable, reducing noise in the measurement. For example, in the current study we had ten observations per token, meaning that each score was based on 30 observations (10 observations per token x 3 middle tokens). Second, because the items come from the middle range of the continuum, the vast majority of the scores fall near the middle of the percentage range, mostly between 25% and 75%. This is an optimal range for t-tests (or ANOVA), as there are no non-linearity issues (these issues do arise with extreme percentages, those above about 93% or under about 7%). On occasion, we have compared the ANOVA results for these scores to ones based on arcsine-transformed scores, and they have always been virtually identical because the arcsine transform only matters for extreme values.

As we noted at the outset, the conclusions one gets from different design/analysis decisions will typically be very similar, as long as those decisions are not outliers. The method used in the current study has proven to be extremely reliable across dozens of previous studies. It has five important advantages: (1) By presenting listeners with the full range of tokens, the task is relative natural — participants make decisions that are not just on "weird" boundary items. (2) We can objectively identify participants who did not do what they were instructed to do, by looking at identification of items that have well-defined correct answers — the endpoints. (3) The analyses are based on the tokens that will typically show the most robust differences, using a range that can accommodate the individual differences that exist across listeners in their category boundaries. (4) The scores that are analyzed are based on a substantial number of observations, reducing measurement noise. (5) The resulting scores strongly cluster in the range in which t-tests and ANOVA are robust.