Published in final edited form as:
*J Exp Psychol Hum Percept Perform*. 2014 August ; 40(4): 1479–1490. doi:10.1037/a0036656.

# Visual Speech Acts Differently Than Lexical Context in Supporting Speech Perception

**Arthur G. Samuel**[1,2,3] and **Jerrold Lieblich**[3]

[1]IKERBASQUE, Basque Foundation for Science, Bilbao 48011 Spain

[2]Basque Center on Cognition Brain and Language, Donostia-San Sebastian 20009 Spain

[3]Department of Psychology, Stony Brook University

## Abstract

The speech signal is often badly articulated, and heard under difficult listening conditions. To deal with these problems, listeners make use of various types of context. In the current study, we examine a type of context that in previous work has been shown to affect how listeners report what they hear: visual speech (i.e., the visible movements of the speaker's articulators). Despite the clear utility of this type of context under certain conditions, prior studies have shown that visually-driven phonetic percepts (via the "McGurk" effect) are not "real" enough to affect perception of later-occurring speech; such percepts have not produced selective adaptation effects. This failure contrasts with successful adaptation by sounds that are generated by lexical context – the word that a sound occurs within. We demonstrate here that this dissociation is robust, leading to the conclusion that visual and lexical contexts operate differently. We suggest that the dissociation reflects the dual nature of speech as both a perceptual object and a linguistic object. Visual speech seems to contribute directly to the computations of the perceptual object, but not the linguistic one, while lexical context is used in both types of computations.

### Keywords

Visual speech; Lexical processing; Perceptual versus linguistic processing

Consider the following three scenarios: (1) You move to a new country where people speak a language you do not know. As you listen to people speak, you are struck by how fast their speech is, and how it is one continuous blur, with few apparent breaks. (2) You have now lived in the country for a few months, and you have been trying to learn the language. As you listen to people speak, the pace seems a little slower, and when a word that you have learned occurs, it seems to pop out of the blur that is most of the rest of the speech. (3) You have lived in the country for several more months, and worked diligently on learning the local language. You can now understand perhaps half of what people say. You find that when you watch people speak, rather than just listen to them, you can understand significantly more of what they are saying.

Corresponding Author: Arthur Samuel, Basque Center on Cognition Brain and Language, Donostia-San Sebastian, 20009 Spain, a.samuel@bcbl.eu.

These three scenarios should ring true with anyone who has learned a second language. In fact, they reflect principles of spoken language processing that can be shown in normal adult native language use under appropriate experimental conditions. The first scenario illustrates the nature of the speech signal, a very compressed interleaving of information from nominally sequential speech segments ("the speech code", Liberman et al., 1967), with no natural separation of words. The apparent difference in speaking rate for a language we know, and the seemingly successive words, are illusions afforded us by the exquisite processing mechanisms that we bring to bear by virtue of proficiency in the language. Similarly, the perceptual pop-out of known words from the drone of unrecognized foreign speech is a consequence of *lexical access*, the activation of the relevant lexical representation. The visual information that comes from watching a speaker ("lip reading") seems to provide a comparable gain in recognition performance.

Superficially, lexical context and visual context appear to be similar. Both provide additional information that can be used to overcome the poor quality and high variability of the acoustic-phonetic signal. For example, if the first few sounds of a word have been tentatively identified as /swIf/ ("swif"), lexical information can predict that the next sound should be /t/, even if the acoustic support for that sound is poor or nonexistent. Similarly, if the acoustics suggest that either "type" or "pipe" was said, visible lip movements can determine which of these it is.

There are many experimental demonstrations of these two contextual sources affecting speech perception. For example, lexical influences on perception are reflected in the "Ganong effect" (Ganong, 1980; Pitt & Samuel, 1993) and in phonemic restoration (Samuel, 1981; Warren, 1970). Ganong created stimuli with an ambiguous segment (e.g., a sound midway between /d/ and /t/) that could occur in two different contexts, e.g., followed by /æʃ/ ("ash"), or by /æsk/ ("ask"). Ambiguous sounds were heard as different phonemes as a function of the lexical constraint: An ambiguous alveolar stop consonant was more often heard as /d/ before /æʃ/, but as /t/ before /æsk/, because this led to hearing a real word (e.g., "dash" or "task") rather than a nonword (e.g., "dask" or "tash"). In the seminal work on phonemic restoration, Warren replaced part of a spoken word with a coughing sound, and found that listeners consistently reported the speech as intact. Samuel showed that restoration was influenced by the lexicon, with stronger restoration of missing speech segments in real words than in matched pseudowords.

Visual influences on perception are also well documented. For example, Sumby and Pollack (1954) presented listeners with speech under various levels of noise masking and found that word recognition was significantly improved when listeners could see the speaker talking. There was little impact when the speech was relatively clear, but under difficult conditions the visual cues were quite powerful, improving word recognition by as much as 40%. Perhaps the best known visual speech phenomenon is the "McGurk effect". McGurk and MacDonald (1976) dubbed mismatching syllables onto videos (e.g., acoustic /ba/ dubbed onto a visual /ga/) and showed that in many cases subjects heard either the visually-presented stimulus, or some compromise between the acoustic and visual inputs (e.g., /da/, given visual /ga/ and acoustic /ba/). McGurk and MacDonald's study has garnered over 3500 citations, and this number rises daily, reflecting a widespread interest in the impact of

visual speech on perception. In addition to many behavioral studies, there is also a growing literature examining the neural underpinnings of the effect (e.g., Besle, Fort, Delpuech, & Giard, 2004; Klucharev, Möttönen, & Sams, 2003; van Wassenhove, Grant, & Poeppel, 2005).

However, despite these developments, and despite the clear utility of visual cues under some circumstances, there are a number of recent findings that suggest that visual context does not have the same status as lexical context. We report the results of two studies we have conducted that pursue this difference. These studies, coupled with existing studies of the effect of lexical context, confirm that the two types of context operate differently. In the General Discussion, we take this evidence, together with recent converging findings, to develop the view that there are two fundamentally different kinds of processing that occur when speech is heard. The two processing routes reflect the fact that spoken language is both a perceptual object, and a linguistic object. Adding to the complexity in this stimulus domain, speech inherently includes multiple types of information (including both visual and lexical context), at multiple possible levels of analysis.

Although is it clear that several types of information affect the processing of spoken language, Norris, McQueen, and Cutler (2000) have pointed out that contextual effects may reflect decision-level influences, rather than true perceptual effects. This distinction can be illustrated by Samuel's (1981) study of phonemic restoration. Listeners were presented with two stimulus types: speech in which a segment had been replaced by noise, or speech in which noise was added to a segment. For each stimulus, the listener judged whether the speech was intact or not. Because listeners produced both types of possible errors (reporting truly intact stimuli as missing a segment, and the reverse), it was possible to compute signal detection measures of perceptual discriminability (d') and bias (Beta) toward reporting a stimulus as intact. The logic of the study was that if listeners perceptually restore the missing speech in a stimulus that is not intact, the resulting percept should be similar to what is heard when noise is simply added to an intact word. Such a perceptual effect would make the two stimulus types difficult to discriminate, lowering d'. In fact, listeners showed worse discriminability (lower d' scores) for segments in real words than in pseudowords, consistent with a true perceptual effect that is driven by lexical activation. The same study also examined whether sentential context led to stronger phonemic restoration – would a missing speech segment be restored more if it was in a word that was predicted by the sentential context than if it was in an unpredicted word? Miller and Isard (1963) had demonstrated that sentence context allows listeners to identify words under difficult conditions, but the locus of this advantage was not determined. In the phonemic restoration test, predictive sentence context biased listeners to report words as being intact, but it did not produce the discriminability (d') difference that was found for real words versus pseudowords (cf. Connine, 1990; Connine & Clifton, 1987). The signal detection analyses suggest that the lexical effect was perceptual (cf. Mirman, McClelland, & Holt, 2005), whereas the sentential effect was taking place at a decision level (cf. van Alphen & McQueen, 2001). In the current study, the central goal is to examine when and how visual context affects speech processing, and to compare this to the effect of lexical context. As we have noted, there are many demonstrations of the power of visual context to affect speech

report. Our focus here is on determining whether the impact of visual speech is found under conditions that exclude decision-level effects.

As Norris, McQueen, and Cutler (2000) noted, most studies of contextual influences in spoken word recognition depend on tasks that are potentially subject to decision level interpretations; it is actually quite difficult to construct testing conditions that are not open to such an interpretation. One way to address this concern is to look for evidence of such contextual effects under conditions in which the listener does not make decisions about the speech segments that are potentially affected by the context. With this approach, one instead looks for *consequential* effects that should occur if the listener had indeed perceived the speech in accord with the context. A methodology that naturally lends itself to this type of consequential test is the selective adaptation paradigm. In an adaptation experiment, listeners first identify members of a set of syllables that comprise a continuum (e.g., with /da/ at one end, and /ta/ at the other). After producing such a baseline measure of how they hear these syllables, the listeners go through an adaptation phase. In this phase, a sound (the “adaptor”) is played repeatedly, with occasional breaks during which listeners again identify syllables from the test series. As Eimas and Corbit (1973) originally showed, adaptation produces a contrastive effect, changing how people identify the test syllables. For example, if /da/ is the adaptor, fewer test syllables will be identified as /da/ after adaptation than on the baseline; if /ta/ is the adaptor, there will be fewer reports of /ta/. In the current study, we employ the consequential adaptation paradigm to assess the role of visual information in audiovisual speech perception.

We have noted that Samuel’s (1981) signal detection study of phonemic restoration indicated that lexical context produces a true perceptual effect – lexically-determined missing phonemes are perceived by the listeners, rather than being the result of a decision bias toward reporting words as being intact. A second study (Samuel, 1997) provides a converging test of this conclusion, using the consequential adaptation paradigm that we will be using in the current study. As noted, the key to this test is that listeners do not make any responses to the restored phonemes, eliminating the possible role of decision processes; effects are instead assessed by a consequence of whether the listeners perceived restored phonemes. Listeners first identified members of a /bI/-/dI/ (“bih” – “dih”) continuum. Adaptation was then conducted, with the adaptors being words in which either /d/ (e.g., “armadillo”) or /b/ (e.g., “exhibition”) had been replaced by white noise. Even though these adaptors did not have any acoustic basis for /d/ or /b/, identification of the members of the /bI/-/dI/ continuum was significantly different after adaptation with restored /b/’s than with restored /d/’s. This shift indicates that listeners had perceived the missing speech sounds, and that these restored sounds functioned as perceived sounds do – they affect the later perception of test syllables. Because listeners made no judgments about the adapting words themselves, the results are not subject to an interpretation based on decision-level bias. Moreover, there were no shifts in a control condition in which the same adaptors were used, except that silence was left (rather than white noise) where a segment had been removed. Listeners do not restore missing phonemes under these conditions, and no shifts occurred when these stimuli were used as adaptors.

Samuel (2001) conducted a series of experiments that used the same logic, but the lexical manipulation was based on the Ganong effect, rather than phonemic restoration. In these experiments, the adaptors were words that either end with /ʃ/ (e.g., “abolish”) or with /s/ (e.g., “arthritis”). For all of the adaptors, the final fricative was replaced by a sound midway between /s/ and /ʃ/. Recall that Ganong (1980) had shown that lexical context causes an ambiguous sound to be heard differentially – the mixture is heard as /ʃ/ when preceded by “aboli_”, but as /s/ when “arthriti_” is the context. If these fricative sounds are actually perceived, rather than being generated at a decision level, then hearing them repeatedly should affect the later identification of sounds in a test continuum. These adaptors did indeed produce adaptation shifts on /Is/ - /Iʃ/ (“iss”-”ish”) test syllables.

Two control conditions solidify the conclusions that can be drawn from these experiments. In one, no final fricatives (normal or ambiguous) were included, and no shifts occurred, indicating that the shifts for the experimental condition were not due to any remaining cues in the word stems (e.g., “aboli_”, or “arthriti_”). The second control condition provided an even stronger test of this because the original words were rerecorded and intentionally mispronounced so that the wrong fricative was originally present (for example, “arthritish”). The mispronounced fricatives were removed and replaced with the ambiguous segment, and these adaptors produced a replication of the results from the experimental condition. Samuel and Frost (in preparation) have recently replicated the major findings of this study, and shown that highly proficient non-native English speakers also demonstrate the lexically-driven adaptation effects, while less proficient non-native speakers do not. Thus, the results for both the restoration-based sounds, and the Ganong-based sounds, indicate that lexical representations can generate true percepts of their component phonemes. These consequential tests converge with the findings from the signal-detection methodology of phonemic restoration, supporting the view that lexical context acts to support the perception of phonemic segments.

In the current study, we test whether visual context can also support the perception of phonemic segments in a way that is sufficient to generate adaptation shifts, meeting the criterion of indirect measurement. In fact, there are prior studies in the literature that have taken this approach to test the perceptual status of audiovisually-determined percepts. Roberts and Summerfield (1981) had subjects identify /bɛ/-dɛ/ (“beh” – “deh”) test syllables, before and after adaptation. The critical adaptor was presented audiovisually, comprised of a visual /gɛ/ paired with an auditory /bɛ/. Recall that McGurk and MacDonald (1976) had shown that the combination of an auditory /b/ and a visual /g/ typically causes listeners to hear a /d/. However, unlike the lexical cases, this procedure did not produce adaptation based on the contextually-determined percept (/dɛ/). Instead, the shifts were identical to those found with /bɛ/ (the auditory component of the audiovisual adaptor). Saldaña and Rosenblum (1994) conducted a follow-up to this study, using improved stimuli and procedures, and replicated the results: the audiovisual adaptor acted just like the purely auditory one. Van Linden (2007) tested a similar condition in which a clear auditory /b/ was paired with a clear visual /d/, and also found effects quite similar to those found for the auditory stimulus alone; these effects were larger with larger numbers of presentations of the adaptor. She contrasted this audiovisual case to one in which the auditory component was

phonetically ambiguous (between /b/ and /d/), and replicated the findings of Bertelson, Vroomen, and de Gelder (2003): This pairing produces effects opposite to those found for adaptation.

The results for visual context thus conflict with those for lexical context. Lexical context produces percepts that can sustain adaptation, but audiovisual context does not, even though both types of context create persuasive subjective experiences. In the current study, we present two sets of experiments (Study 1 and Study 2) that pursue these conflicting results. One goal is to clarify when context produces fully functional phonemic codes and when it does not, in order to delineate the relationship between bottom-up and top-down processing in spoken word recognition. A second goal is to provide evidence that bears on the question of how and when different forms of contextual information (lexical; visual) are incorporated with the acoustic-phonetic signal.

The experiments in the current study use the consequential methodology described above, in order to isolate purely perceptual effects from any decision-level factors. Study 1 is very similar to earlier studies (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994) using this approach, but adds a lexical component that could potentially strengthen the audiovisual effect. Study 2 is based on a different audiovisual phenomenon, one examined by Green and Norrix (2001). Those authors found that identification of members of an auditory /ili/ - /iri/ (“eelee” – “eeree”) continuum could be shifted by the presentation of (silent) visual speech in which /b/ was presented just before the /l/ or /r/. In both Studies, we test whether a phonemic percept that is generated by visual speech can produce subsequent adaptation shifts of the type found with lexically-generated phonetic segments (Samuel, 1997, 2001). In the General Discussion, we consider the implications of the results for theories of spoken language perception.

## Study 1: Can McGurk-generated phonetic segments produce adaptation?

The two studies that found no adaptation by an audiovisually-determined percept both used nonlexical simple consonant-vowel stimuli (Roberts & Summerfield, 1981; Saldaña and Rosenblum, 1994). Brancazio (2004) and Barutchu et al. (2008) have shown that the McGurk percept is strengthened when the audiovisual combination yields a real word. Given this, it might be possible to boost the audiovisual context effect by giving it lexical support, and thereby produce an adaptation effect with audiovisual adaptors if they correspond to real words. To maximize comparability to the lexical literature, we modeled our stimuli closely on those used in Samuel’s (1997) lexical adaptation study. But, rather than producing a /d/ adaptor by inserting white noise in place of the /d/ in words like “armadillo”, we used the McGurk effect to produce the /d/ by pairing a visual nonword (e.g.,“armagillo”) with an auditory nonword (e.g.,“armabillo”). The question addressed in Study 1 is whether an audiovisual percept can produce adaptation if the audiovisual percept has lexical support.

### Experiment 1

#### METHOD

**Stimuli:** The test series was the same set of syllables used by Samuel (1997), an eight step /bI/ - /dI/ continuum synthesized on the Klatt synthesizer, in its cascade (serial) mode.

All syllables were 220 ms long, including 155 ms of steady state vowel. The primary cue that varied across the continuum was the path of the second (F2) and third (F3) formants. At the /b/ end, these started at relatively low frequencies (F2: 1422 Hz; F3: 2264 Hz), while for the /d/ endpoint these values were higher (F2: 1800 Hz; F3: 2600 Hz). Additional details can be found in Samuel (1997).

To construct the adaptors, digital videorecordings were made of the first author pronouncing each of the five critical words ("academic", "armadillo", "confidential", "psychedelic", and "recondition"), in three different versions: normal (e.g., "armadillo"), /b/-version (e.g., "armabillo"), and /g/-version (e.g., "armagillo"). The video was framed as a headshot, and several recordings of each item were made. The videotapes were digitally transferred to an Apple iBook G4, and FinalCut Express software was used to select and recombine the video and auditory tracks as needed.

**Adaptation Conditions:** Three adaptation conditions were tested. The critical experimental condition was the **McGurk** case, in which visual versions of each word produced with a /g/ (e.g., "armagillo") were paired with auditory /b/ versions (e.g., "armabillo"). In the **Auditory** condition, the same adaptors were used, but the video monitor was turned off. The **Real-/d/** adaptors were audiovisual versions of the normally pronounced words (e.g., "armadillo").

**Procedure:** Each participant did a baseline identification test followed by an adaptation test. The baseline test included 22 randomizations of the 8 test syllables. The first two passes were practice and were not scored. After each syllable was presented, the participant responded by pushing one of two labeled buttons (B vs D). The adaptation test included 20 randomizations of the test syllables, presented for the same judgment. However, before each randomization, a 40-second adaptation sequence was presented. Each adaptation sequence consisted of five randomizations of the five adaptor stimuli (= 25 tokens), presented at a rate of approximately one item every 1.6 seconds (this included approximately a half second of video fade out and fade in, to minimize discontinuities).

Participants were instructed to merely attend to the adaptors, but to respond to the test syllables. In the McGurk condition, a white dot was superimposed near the speaker's mouth on 2, 3, or 4 of the 25 adaptors in any given adaptation phase. Whenever a dot appeared, participants were instructed to push a response button, ensuring attention to the mouth area (Bertelson, Vroomen, & de Gelder, 2003).

Participants were tested individually in a sound shielded chamber. The videos were shown on a 20.1 inch ViewSonic VP201b LCD monitor. The audio was presented at a comfortable listening level over Harman/Kardon HK195 speakers. The speakers were placed next to the monitor, with one on each side.

Each of the three conditions was run with two separate groups of 16 participants, under slightly different conditions. For the McGurk condition, in one case the adaptors were made with a 6 dB increase in the amplitude of the /b/ portion of the waveform; this produced a somewhat more persuasive McGurk percept in pilot testing. The second McGurk group was

tested with adaptors in which there was no such manipulation of the amplitude. The two versions of the Auditory condition had the same distinction, since they were conducted with the same adaptors with the video turned off. For the Real-/d/ condition, in one version participants were explicitly instructed not to speak during the adaptation phase; this was done as a check on the other version, in which no such instructions were given, and there was the theoretical possibility that any effects could be due to participants "talking along" with what they were seeing.

**Post-test Questionnaire:** After completing the baseline and adaptation tests, each participant completed a questionnaire to assess what the person's subjective experience had been of the adaptors. Each of the five adapting words was printed in all three forms (e.g., "armabillo", "armadillo", "armagillo"), and participants were asked to circle whether a given form was heard: "All of the time", "Most of the time", "A few times", or "Never". The responses were converted to numbers (e.g., "Never" = 0, "All" = 3), and a sum of the five /d/ versions was computed to index how often the stimuli were perceived as having a /d/. If listeners had heard a given item as its true /d/ form all of the time (e.g., in the Real /d/ condition), this would yield a total score of 15 (5 stimuli × a rating of 3).

**Participants:** 96 native English speakers with self-reported normal hearing participated (32 per condition).

**RESULTS AND DISCUSSION—**The average report of "D", for each stimulus, in each condition, was computed for all participants, on the baseline and post-adaptation tests. For each of the three conditions, an analysis of variance was conducted on the differences between baseline and post-adaptation, using average D report for the middle four tokens of the test series (Pitt & Samuel, 1993; Samuel, 1986). Preliminary analyses tested whether there were any differences in the adaptation effects across the two versions of each condition (i.e., with or without a 6dB increase; with or without explicit directions to listen silently). In all three cases, there was no hint of any difference between the two versions (for all three, $F<1$); therefore, in the main analyses the two versions were collapsed, providing 32 participants per condition.

The left panel of Figure 1 shows the results of adapting with normal words (audiovisual Real-/d/), a reliable 6.8% reduction in D report, $F(1,31) = 12.82$, $p<005$. This condition provides the upper boundary for any potential McGurk adaptation. The middle panel of the figure presents the corresponding results for the Auditory (/b/) condition. Because adaptation is a contrastive effect, adapting with /b/ should decrease report of B. There was a nonsignificant 2.1% increase in D (decrease in B) report ($F(1,31) < 1$). This is exactly the pattern reported by Samuel (1997) with comparable stimuli.

Recall that in both previous studies of McGurk-driven adaptation, the McGurk adaptor produced results similar to the auditory part of the McGurk stimulus, rather than the dominant percept. The central question is whether McGurk adaptors that have lexical support will behave more like real /d/, or will instead still behave like their auditory (/b/) component. The right panel of Figure 1 shows the results, which are unambiguously similar to those shown in the middle panel for the Auditory /b/ condition. There was a small (0.9%)

nonsignificant increase in D report ($F(1,31) < 1$), just as there was in the Auditory only condition.

A two-factor ANOVA confirmed that the McGurk condition was unlike the Real-/d/ condition. One factor was Baseline versus post-adaptation labeling, and the second factor was condition (McGurk versus Real-/d/). The significant interaction ($F(1,62) = 8.33$, $p<.005$) shows that different shifts occurred. In contrast, a comparable ANOVA with McGurk versus Auditory (/b/) showed no hint of any such interaction, $F(1,62) < 1$. Thus, despite the lexical support, McGurk adaptation was determined by the auditory component, not the conscious percept.

Of course, if the McGurk stimuli failed to produce a clear /d/ percept, then the results could be attributed to simple adaptation by the /b/-containing auditory component. The post-test questionnaire results provide an assessment of what subjects consciously perceived. The average ratings for both the Real-/d/ adaptors (13.9) and the McGurk adaptors (12.9) were near the maximum possible “D-like” score of 15, though the one-point difference just did reach significance, $F(1,62)=4.05$, $p=.05$.

The auditory (/b/) adaptors (7.9) yielded scores much lower than those for the McGurk adaptors ($F(1,62)=35.57$, $p<.001$). The rating of the McGurk percept was very similar to that of a Real /d/ and unlike that of the Auditory /b/, yet its adaptation effect was quite different from that for the Real /d/, and indistinguishable from the effect of the Auditory /b/ adaptor. The current results, despite the addition of lexical support, are completely consistent with the previous studies showing that audiovisual percepts do not support adaptation (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994).

## Study 2: Can “unopposed” visually-generated phonemic segments produce adaptation?

We now have results from multiple studies that unambiguously demonstrate that the phonetic percepts generated by McGurk stimuli cannot produce a consequential adaptation effect. These results contrast sharply with those from structurally very similar tests (Samuel 1997, 2001) in which lexical context, rather than visual context, was used to generate phonetic percepts. The obvious conclusion is that visual and lexical types of context affect phonetic processing in different ways, despite their very similar phenomenological effects.

Before accepting such a conclusion, however, it seems prudent to consider any procedural differences between the studies testing lexical and visual context effects in the adaptation paradigm. Although in almost all respects the two cases are well matched, there is one important difference. By the very nature of the McGurk effect, the visual context tests have relied on stimuli in which the subject receives contradictory information: Although /d/ is perceived, the perceptual system is simultaneously receiving clear auditory evidence for /b/. In the lexical context experiments, there is no such clear contradictory evidence potentially competing with the synthesized percept. In the phonemic restoration case (Samuel, 1997) white noise replaced the critical phoneme, and white noise does not constitute a clear competing phonetic segment. In the Ganong-based case (Samuel, 2001) the critical phonetic

information was carefully tuned to be ambiguous, midway between two phonemes. As such, it again does not offer a strong phonetic alternative to the lexically-generated percept. Thus, it is possible that the McGurk-driven percepts have failed to produce adaptation because of strong phonetic competition, a factor not present in the two lexically-driven cases.

To provide a test of visual context that does not have this potential disadvantage, we would need a visually-driven phonetic shift in which there is no clear competing phonetic information. In fact, Green and Norrix (2001) have reported just such an effect. They synthesized an /ili/ - /iri/ test continuum, and either presented the items in purely auditory form, or accompanied by a silent video in which the speaker had articulated /ibi/ (with the /b/ timed to occur just before the /l/ or /r/). In previous experiments the authors had found that the phonetic boundary between /l/ and /r/ shifted if an auditory /b/ was included – the /l/ - /r/ boundary is different in /ili/ - /iri/ than in /ibli/ - /ibri/. The audiovisual test revealed a comparable shift, as though the subjects had integrated the visual /b/ with the auditory tokens.

For our purposes, the utility of this effect is that it potentially allows us to shift the perception of an ambiguous auditory token (at the phonetic boundary for /ili/ - /iri/) by presenting silent visual speech without the conflicting phonetic information that is inherent in the McGurk effect; the critical visual information pertains to a preceding stop consonant (/b/), not to the liquids being probed. In fact, the Green and Norrix stimuli offer a test that is isomorphic to the Ganong-driven adaptation case used by Samuel (2001): Just as Samuel had adaptors in which (lexical) context affected an auditory segment designed to be midway between two sounds (/s/ and /ʃ/), we can use (visual) context to affect an auditory segment designed to be midway between two sounds (/l/ and /r/). If the previous failures to find adaptation for visually-generated phonemes were due to the presence of conflicting phonetic information, then this alternative method for generating such phonemes should now produce adaptation effects like those found for lexically-driven phonemes. If the adaptation effects still do not appear, then a more fundamental difference in the role of lexical and visual context is implicated.

## Experiment 2a

Experiment 2a is a preliminary test to establish the visually-driven phonetic shift in our laboratory using a modified version of the Green and Norrix (2001) procedure. The most important modification is the use of two different visual contexts, rather than the single /ibi/ visual context used in the original study. The visual appearance of /g/ reflects its velar place of articulation, while the visual appearance of /b/ is clearly labial. Thus, we expected that visual /igi/ should produce a shift in the opposite direction from that caused by a visual /ibi/ -- seeing the visual cues for /g/ should increase /l/ report, while seeing the visual cues for /b/ should decrease it. Because we wish to produce as large a visually-driven phonetic shift of the /l/-/r/ boundary as possible, we therefore contrasted a silent visual /ibi/ context with a silent visual /igi/ context.

Experiment 2a provides a measurement of whether these contexts produce different enough identification functions for the /ili/ - /iri/ continuum members to support the desired adaptation-based test. In particular, we wish to determine if the visually-driven shifts in

identifying the /ili - /iri/ stimuli are as large as the lexically-driven shifts that produced significant adaptation effects in Samuel's (2001) study. If they are, then we can move forward and conduct an audiovisual adaptation experiment in which an identical auditory stimulus (an ambiguous member of the /ili/ - /iri/ continuum) is paired with a silent visual /ibi/ and with a silent /igi/. The former should be perceived as /ibri/, and the latter as /igli/, allowing us to test whether listeners report fewer /l/ percepts when adapted with audiovisually-determined /igli/ than with audiovisually-determined /ibri/.

## METHOD

**Stimuli:** Eight tokens from the 10-step /ili/ - /iri/ test series used by Green and Norrix (2001) were used here. The tokens were generated on the Klatt (1980) synthesizer with a fundamental frequency of 126 Hz. The initial /i/ had four formants kept at constant frequencies of 250, 2090, 2900, and 3300 Hz; the final /i/ had the same fixed formant values. To make the /l/ - /r/ portions of the stimuli, F2 and F3 varied. For all steps, F2 was initially set to 1200 Hz for 80 msec, with a following 95 msec transition up to its steady state frequency of 2090. The /l/ - /r/ distinction was carried by the F3 formant transition. A low initial frequency (most extreme: 1300 Hz) cued /ri/, while a high initial frequency (most extreme: 3100 Hz) signaled /li/. Step size was 200 Hz for the third formant onset frequency. These transitions were flat for 80 msec before shifting toward the steady state value (2900) Hz over the following 95 msec. All tokens were 750 msec long. See Green and Norrix (2001) for more details.

Two short video clips were recorded of the first author saying /ibi/ and /igi/. The videos were head shots, and the speech rate was designed to match the /ili/ - /iri/ tokens. Final Cut software was used to dub each of the eight speech tokens onto each of the two silent video clips, with the visual stop closure timed to occur just before the onset of the /l/ or /r/ formant transitions. Thus, depending on which visual token is paired with which member of the continuum, audiovisual integration will yield /ibri/, /ibli/, /igri/, or /igli/. The visual clips were 1550 msec long.

**Procedure:** The resulting 16 audiovisual stimuli were presented to participants following the procedures used for the baseline identification test in Experiment 1. Twelve randomizations of the 16 stimuli were presented. After each audiovisual stimulus was presented, the participant pushed one of two labeled buttons (L versus R) on a response panel to indicate whether it contained /l/ or /r/. Participants were instructed to watch the screen, and nothing was said about the /b/ or /g/.

**Participants:** 13 individuals from the same population as in Experiment 1 participated in Experiment 2a.

**RESULTS AND DISCUSSION—**Average report of "L" for each member of the /ili/ - /iri/ continuum, in each visual context, was computed for each participant. The 13 participants clearly fell into two distinct groups. Four of the participants showed no effect of the visual context – their identification functions were essentially the same for the two visual contexts. The other nine participants showed extremely large effects of the visual context. To put this

distribution in perspective, the success of the widely-cited McGurk effect varies enormously across studies. Nath and Beauchamp (2012) report success rates in the literature ranging from 26% to 98%. For example, Sams et al. (1998) found a 32% incidence of audiovisually-determined fusions for visual /ka/ with auditory /pa/, whereas in the seminal McGurk and MacDonald (1976) paper this pairing produced an 81% fusion rate (in both of these papers, there were also significant numbers of people who simply reported the visual stimulus). Thus, observing an audiovisual effect for 69% of the participants in Experiment 2a is quite consistent with the McGurk literature.

Recall that the goal of Experiment 2a was to determine if the silent visual context could shift identification of the /ili/ - /iri/ stimuli as much as the lexical context had done in Samuel's (2001) demonstration of significant adaptation by lexically-driven phonetic perception. For the subset of participants in Experiment 2a who were affected by the visual context, the effect was in fact quite large. An analysis of variance on identification of the middle four members of the continuum confirmed that report of "L" was significantly higher in the context of a visual /g/ than in the context of a visual /b/, $F(1,8) = 28.14$, $p<.001$. The right panel of Figure 2 shows the spread of the identification functions for these listeners as a function of the visual context (/ibi/ versus /igi/). For comparison purposes, the left panel of the figure shows the lexical shift produced by stimuli comparable to those used by Samuel (2001). As the figure makes clear, the visual context produced a shift that was at least as large as the lexical shift. Thus, it appears that the Green and Norrix (2001) effect provides an ideally-matched context effect that can be used to compare the effect of visual context and lexical context.

## Experiment 2b

Experiment 2b uses the visual context effect that we have now shown to be effective in order to determine if the phonetic percept driven by the visual context will produce adaptation effects comparable to those found for lexical context (Samuel, 1997, 2001). Critically, by using the Green and Norrix (2001) procedure, there is no clear conflicting phonetic information, unlike the McGurk-based test in Experiment 1 and in previous studies (Roberts & Summerfield, 1981; Saldaña and Rosenblum, 1994).

### METHOD

**Stimuli:** The same audiovisual stimuli used in Experiment 2a were used in Experiment 2b. In addition, the audio-only versions of the /ili/ - /iri/ test series were used.

**Procedure:** Each participant took part in two sessions held on different days. The first session included two tasks. The first task was the same audiovisual identification task used in Experiment 2a – participants heard all eight members of the /ili/ - /iri/ continuum dubbed onto the two visual contexts, and responded by pushing the "L" or the "R" button for each item. When the participant had completed this task, the two identification functions (one for the visual /ibi/ context, and one for the visual /igi/ context) were inspected by the experimenter. Based on this inspection, the experimenter selected an individually-determined ambiguous token from among the four middle items of the 8-item /ili/ - /iri/

continuum. The selection was designed to pick an item that was primarily identified as “L” in the /igi/ context and as “R” in the /ibi/ context.

When this choice had been made, the experimenter initiated the adaptation test. The auditory component of each adaptor was the token that had been individually selected for the participant. It was accompanied by either the visual /ibi/, or by the visual /igi/. Half of the participants saw /ibi/ during the first session, and half saw /igi/; during the second session, each participant received the “other” visual context. Critically, the auditory token was always identical across the two sessions for a given participant.

The adaptation phase of each pass included 20 presentations of the appropriate audiovisual adaptor, followed by 12 audio-only /ili/ - /iri/ tokens. Each 1550 msec audiovisual adaptor was followed by 500 msec of black screen, resulting in adaptation phases of 41 seconds. The 12 tokens on each pass included one randomization of the eight /ili/ - /iri/ tokens, plus four of those tokens played a second time; on the following pass, the listeners heard another randomization plus the remaining four tokens. Thus, across pairs of consecutive passes, listeners received three randomizations of the eight-member /ili/ - /iri/ test series. Participants were instructed to watch and listen during the adaptation periods, and to only respond to the audio-only stimuli (by pushing the “L” or the “R” button on the response panel). As in Experiment 1, on 2, 3, or 4 of the audiovisual adaptors in each pass we superimposed a small white dot near the mouth, and subjects were instructed to push a different button on the response panel each time a white dot appeared. There were eight passes in an adaptation session, providing 12 observations for each test token.

**Participants:** 30 individuals from the same population tested in the previous experiments participated in Experiment 2b.

**RESULTS AND DISCUSSION—**Seven of the participants could not reliably distinguish /ili/ from /iri/ on one or both of the baseline identification tests, leaving 23 participants. Of these, 16 participants showed strong visual context effects. Our analyses of adaptation effects are based on the 16 participants who identified the test items reliably, and who also were strongly affected by the visual context. This selection ensures that the adaptors were heard as we wished them to be – as /l/ with visual /igi/, and as /r/ with visual /ibi/. Figure 3 shows the audiovisual identification data for these 16 participants. As in Experiment 2a, the visual context effect for these participants was very large, $F(1,15) = 20.21$, $p < .001$.

Given these robust effects of the visual context, we can now examine the central question: Will the phonetic percepts that are determined by audiovisual speech integration produce adaptation shifts in the same way that lexically-driven ones have been shown to do? Figure 4 provides the relevant data, and the results are unambiguous: There was absolutely no hint of any differential adaptation, despite the extremely large visual context effect for these participants, $F(1,15) = 0.04$, n.s. These results converge with the previous studies of McGurk-based adaptation attempts – in all cases, the audiovisual percepts fail to generate any consequential adaptation effects.

Because this critical result relies on a null effect, it seems prudent to be sure that such a null effect does not reflect some artifact or lack of power in the design. For example, the test assumes that syllable-initial /l/ or /r/ can be adapted by the /l/ or /r/ in a consonant cluster (/br/, /bl/, /gr/, or /gl/), and there is evidence that under some circumstances adaptation effects do not transfer across syllable position (e.g., Samuel, 1989; Samuel, Kat, & Tartter, 1984). Thus, we conducted a control experiment using procedures comparable to those of Experiment 2b, but with actual auditory /ibri/ and /igli/ as the adaptors, rather that adaptors created by audiovisual integration. The adaptors were natural speech in which the speaker roughly imitated the qualities of the synthetic /ili/ - /iri/ tokens, producing speech with approximately the same fundamental frequency, relatively flat pitch contour, and similar duration. The adaptors clearly were not tightly matched to the test items acoustically, but these differences would only reduce any adaptation effects, so that if we find clear effects that is not an issue. A new group of 34 listeners participated in two sessions (order of the sessions was counterbalanced), one of which had the naturally-produced /ibri/ as the adaptor, and one of which had the /igli/. Eight of the 34 did not label the /ili/ - /iri/ cleanly, leaving adaptation data for 26 listeners. Figure 5 shows the adaptation results, and as is very clear in the figure, adaptation with auditory /ibri/ produced very different identification functions than adaptation with auditory /igli/. There was an 10.9% shift across the two conditions that was extremely reliable, $F(1,25) = 18.92$, $p < .001$. Thus, the lack of adaptation found in Experiment 2b for the corresponding audiovisual adaptors does not trace to an artifact or to a lack of power – it is a function of their audiovisual nature.

## GENERAL DISCUSSION

We began by noting that the complexity and variability of spoken words might lead listeners to use contextual information to aid word recognition. Phonemic restoration, the Ganong effect, and the McGurk effect all seem to be examples of such contextual influences. However, prior adaptation studies have found evidence for only lexical context producing the consequential effects that unambiguously are associated with the perceptual level (Samuel, 1997, 2001); similar studies have found contrasting negative results for audiovisual processing (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994; van Linden, 2007). The current study was undertaken to try to reconcile these conflicting results.

Study 1 produced two critical findings. First, we replicated the non-effect of McGurk percepts as adaptors. As in the prior work, such adaptors behaved just like their auditory components. Second, this non-effect of the McGurk adaptors obtained even though the McGurk percept resulted in a lexical item. There was a striking dissociation between people's conscious percepts and the observed adaptation effects. Study 2 produced the same dissociation, with clear percepts of /l/ or /r/ failing to produce adaptation shifts that would normally be found for acoustically-driven percepts of these sounds. Critically, unlike the McGurk-based tests, the stimuli in this case did not present the listener with competing acoustic inputs. In fact, the structure of this test was essentially identical to the structure of the Ganong-based lexical tests that have produced successful adaptation shifts (Samuel, 2001; Samuel & Frost, in preparation).

What does the pattern of successful versus unsuccessful adaptation tell us about how context is combined with the speech signal? We begin by considering the conditions that have produced such consequential effects, and then consider those that have not. The successful cases have all involved lexical context, including both sounds perceived via phonemic restoration (Samuel, 1997) and those perceived through the Ganong effect (Samuel, 2001; Samuel & Frost, in preparation). For these tests of lexical context, it is worth noting that all of the successes have come from situations in which the signal was ambiguous, and the listeners had strong lexical representations that could be used to disambiguate the signal. In the restoration test, the consequential effect was obtained when segments (/b/ or /d/) were replaced by noise in lexical contexts that unambiguously provide a known interpretation; given “arma*illo”, with noise replacing the /d/, the lexicon specifies the /d/. However, such a specification is not sufficient – there is no restoration, and no consequential adaptation, if the signal is unambiguous: When the critical /b/ or /d/ segments were replaced by silence rather than white noise, listeners heard the gap, did not restore, and did not produce adaptation shifts. Top-down influences will not overrule clear bottom-up input. If they did, listeners would be prone to hallucinations, which is obviously quite undesirable. Rather, lexical context operates when the input signal is less than ideal, a very common occurrence in the real world. This system is effective because speakers almost always produce real words, rather than nonwords, making a lexical bias on perception successful most of the time.

The successful adaptation effects using Ganong stimuli also meet the criterion of having acoustically-ambiguous critical segments because in Ganong studies, the critical phoneme is carefully designed to be acoustically ambiguous. As in the restoration case, when silence was used (e.g., “arthriti_”) instead of an ambiguous segment, no adaptation occurred (Samuel, 2001). Recent results show that along with signal ambiguity, the listener must have a strong lexical representation available to drive perception of the ambiguous segment. Samuel and Frost (in preparation), using the English materials that Samuel (2001) had used, found that highly proficient non-native English speakers (native Hebrew speakers living in Israel) produced adaptation shifts that were quite similar to those found by Samuel (2001) for native English speakers, but less-proficient non-native English speakers (native Arabic speakers in Israel) did not.

Collectively, the successful adaptation cases indicate that lexical context can drive phonetic perception when (1) there is an ambiguous phonetic signal, and (2) the listener has strong lexical information available to disambiguate the signal. What about visual speech context? The prior adaptation tests of visual speech both violated the first of these conditions: In Roberts and Summerfield (1981), Saldaňa and Rosenblum (1994), and in van Linden (2007), the use of McGurk stimuli entailed the presentation of a clear acoustic-phonetic segment that was inconsistent with the audiovisual percept (that is the core property of McGurk stimuli). In our first Study, using McGurk stimuli bolstered by lexical support, the same violation was present. All of these tests thus do not meet the hypothesized constraint that top-down influences will not overrule clear bottom-up input. Consistent with this violation of the premise, in all of these studies, the shifts were dominated by the auditory information, rather than by the conscious percept.

There is a substantial body of audiovisual research by Vroomen and his colleagues (e.g., Bertelson, Vroomen, & de Gelder, 2003; Vroomen, van Linden, de Gelder, & Bertelson, 2007) that does meet the criterion of having ambiguous acoustic-phonetic input. In these studies, such ambiguous acoustic stimuli are paired with unambiguous visual information, showing a mouth producing, for example, a clear /b/ or a clear /d/. These studies have produced perceptual recalibration effects that go in the opposite direction to those found for adaptation. There are generally somewhat different procedures used in these studies than in a standard adaptation study, though Vroomen et al. (2007) approximated the adaptation procedures and still failed to observe adaptation for visually-determined sounds.

Additional visual adaptation tests in the literature also contrast with the successful lexically-driven effects. In both Roberts and Summerfield (1981) and in Saldaňa and Rosenblum (1994), visual-only adaptation conditions were ineffective: If subjects watched silent video presentations, showing the lip movements consistent with one or the other endpoint of the test continua, there was no effect on their identification of the audio-only test syllables. This is not because visual presentation is unable to produce adaptation per se – classic color aftereffects demonstrate such contrastive effects. More to the point here, if subjects watch a video of lips producing the gestures for /m/, versus the gestures for /u/, their identification of silent video clips on an /m/-/u/ stimulus continuum is contrastively shifted (Jones et al., 2010). It is the audiovisually determined percept that has consistently failed to produce adaptation effects on the identification of spoken test items.

Although the failures with McGurk adaptors are consistent with the notion that context will only drive phonetic perception of ambiguous segments, Vroomen's studies and the results of Study 2 show that ambiguity is not sufficient. The adaptors in Study 2 were individually selected for each listener in the same way that individual adaptors were selected for lexical tests using the Ganong effect (Samuel, 2001; Samuel & Frost, in preparation): The acoustic-phonetic input was selected to be ambiguous. Nonetheless, the audiovisual adaptors were entirely unable to drive perception, even though the participants were carefully chosen to be ones who produced extremely large visually-driven changes in identification of the /ili/ -/iri/ stimuli. Formally, the lexically-based experiments and the visually-based ones are identical, yet only the lexical context has proven to be capable of driving phonetic perception in a way that produces the consequential effects that we have associated with perception.

We thus have a very curious situation. As anyone who has experienced the McGurk effect or the Ganong effect will confirm, both visual (McGurk) and lexical (Ganong) context produce compelling phenomenological experiences. Despite this, the results of these two types of experiences produce divergent outcomes with respect to the criterion of consequential impact. In the Introduction, we alluded to the fact that the speech signal is both a perceptual object, and a linguistic object. A possible resolution of the empirical dissociation of lexical versus visual context might be that visual context plays a direct role for speech as a perceptual object, but not as a linguistic object, while lexical context directly impacts both. One way to think about this distinction is to note that from a linguistic perspective, phonetic segments are naturally associated with the lexical representations that contain them; in contrast, the visual pattern that is associated with a word does not have the part-whole relationship of segments and words.

This is clearly quite speculative at this point, but the literature provides some evidence that is consistent with this speculation. Our suggestion is that lexical processing is an inherent property of spoken language recognition, with critical operations carried out in posterior cortical regions (e.g., superior temporal sulcus/gyrus), while visual speech processing is a supportive property, with critical operations housed in more anterior regions (e.g., left inferior frontal gyrus). A reasonably large literature using fMRI supports this localization difference, but in general studies using fMRI do not have the temporal resolution needed to interpret a potential early perceptual role for visual speech. Evidence for early involvement of anterior regions during speech perception requires measures with the necessary temporal resolution, such as EEG and MEG.

There is, in fact, a small literature using these techniques that is relevant to our suggestion. Two studies of audiovisual speech perception looked for differences in the N1 ERP component, a very early response associated with auditory processing. Besle, Fort, Delpuech, and Giard (2004) presented four different syllables to their subjects in auditory, visual, or audiovisual form. The task was to respond to one of the four syllables, thus requiring identification. There was a decrease in the amplitude of the N1 component in the audiovisual case compared to the combination of the auditory and visual cases, which the authors took as evidence for visual influences on phonetic perception. These effects were seen on a component 120–190 msec after stimulus presentation, which certainly qualifies as an early effect. A similar result was obtained by van Wassenhove, Grant, and Poeppel (2005). They also presented syllables in auditory, visual, or audiovisual form, and they also observed a reduction in the amplitude of N1 (and of the slightly later P2).

For our purposes, the major issue with respect to these findings is whether they reflect phonetic integration effects across modalities, or if they instead are due to an alerting effect. Because visual speech usually provides evidence to the perceiver before the acoustic signal (the articulators can be seen to move about 200 msec before the sound is emitted), perceivers who can see the speaker are alerted to the imminent arrival of the acoustic signal. Both sets of authors argue against an alerting interpretation of the N1 effect, but in neither case is the argument fully compelling. For example, Besle et al. assert that an alerting effect would be expected to increase N1 amplitude rather than decrease it, but there is not much evidence in the literature for this under comparable circumstances. Van Wassenhove et al. make the argument that the amplitude decrease was independent of the identity of the triggering consonant while a latency effect was sensitive to consonant identity. This is somewhat more plausible, but not overly persuasive.

An elegant pair of studies by Stekelenburg and Vroomen (2007; Vroomen and Stekelenburg, 2010) provides strong evidence that the N1 attenuation is due to a type of alerting effect, one that has nothing to do with phonetic encoding. In their first series of experiments, the authors looked for audiovisual N1 attenuation with both speech stimuli and nonspeech stimuli – two hands clapping, or tapping a spoon against a cup. Note that in all three cases, the visual information stream provides evidence before the auditory stream – the lips move before sound begins, the hands move towards each other before their impact produces a clap, and the spoon moves toward the cup before the tapping sound occurs. Stekelenburg and Vroomen found comparable N1 attenuation in all three cases compared to the sum of the

visual-only and auditory-only cases, and clearly there is no phonetic integration involved in the nonspeech situations. In another experiment, they compared the N1 attenuation for a speech case in which the phonetic information was consistent across the two modalities to a case in which it was not (auditory /bi/ paired with visual /fu/). The N1 attenuation did not differ for these two situations, as it should have if the N1 attenuation is related to phonetic integration of the auditory and visual speech.

In follow-up work, Vroomen and Stekelenburg (2010) used a display in which a rectangle appeared to be “squeezed” for 240 msec. The visual squeezing was accompanied by a tone. In one condition the squeezing appeared to be caused by two circles that approached the rectangle from each side, with the tone/squeezing onsetting at the moment of impact. Vroomen and Stekelenburg demonstrated that the critical N1 attenuation was found if and only if the visual information preceded the auditory event in a predictable way; visual information that was available too soon, too late, or unpredictably did not lead to the attenuated N1. Collectively, the set of experiments provides strong evidence that the attenuated N1 does not reflect audiovisual phonetic integration. Rather, when a perceiver gets a reliable visual cue about a soon-to-occur auditory event, the processing load when the auditory signal arrives is reduced, leading to a smaller (and often more rapid) N1. The visual information is clearly playing a useful perceptual role, but not one that is tied to phonetic perception, consistent with our suggestion that visual speech contributes to perceptual processing but not to linguistic encoding. Presumably direct involvement in linguistic encoding is necessary to generate phonetic percepts that are capable of sustaining selective adaptation on a phonetic continuum.

If selective adaptation were the only domain in which visual speech failed to produce language-based effects, we would be more hesitant to suggest the dissociation that we have put forward. However, there are additional results in the literature to support this dissociation. For example, there are several studies of “compensation for coarticulation” in which lexical support for the percept can affect the result (Elman & McClelland, 1988; Magnuson, McMurray, Tanenhaus, & Aslin, 2003; Samuel & Pitt, 2003), but lip-read information does not (Vroomen & de Gelder, 2001). A recent study provides intriguing converging evidence for functionally separate processing of the linguistic and perceptual aspects of spoken language, with visual speech directly contributing to the latter but not the former. Ostrand, Blumstein, and Morgan (2011) conducted a semantic priming study in which audiovisual primes preceded auditory test items; participants made lexical decision judgments for the auditory test items. For example, a prime could have auditory “bamp” combined with video of a speaker articulating “damp”, which produces a percept of “damp” through visual capture (similar to the McGurk effect). Alternatively, the prime could have an auditory word (e.g., “beef”) combined with a visual nonword (e.g., “deef”), producing a percept of “deef”. The central question was whether semantic priming (e.g., for auditory test items like “wet” or “pork”) is generated by the audiovisual percept. Ostrand et al. found that if the auditory component of the audiovisual prime was a word (e.g., “beef”) then semantic priming was found even if the perception of the audiovisual prime was a nonword (e.g., “deef”). In other words, even when people perceived “deef”, the (unperceived) auditory signal “beef” was able to engage the lexical system and produce priming; no such priming occurred when a prime percept (e.g., “deef”) was based on an audiovisual combination that

was purely a nonword (e.g., both audio and video “deef”). As Ostrand et al. note, this suggests that “the auditory signal determines the word actually activated in the lexicon while the combined audio and visual information determines the item the comprehender believes she has received” (p. 1380). This is exactly the duality that we are positing: One process drives linguistic coding while another drives the percept, with the visual speech only involved in the latter.

Clearly, much more research is needed before one can be confident that visual speech directly contributes to speech as a perceptual object but not to speech as a linguistic object. Nonetheless, there is now evidence from two quite different testing situations – selective adaptation and semantic priming – that is consistent with this dissociation. There is also now a relatively robust data set that demonstrates the difference between lexical context and visual speech context: Five separate studies (Roberts & Summerfield,1981; Saldaňa & Rosenblum, 1994; van Linden, 2007; Study 1 and Study 2 here) have looked for adaptation effects driven by visual speech, and all five have produced very clear negative results. In contrast, three separate studies (Samuel, 1997; Samuel, 2001; Samuel & Frost, in preparation) have shown reliable lexically-driven adaptation effects. These three studies demonstrate that lexical context can be used by a listener to generate a functional phonetic percept from an ambiguous acoustic-phonetic signal.

A system that allows lexical top-down effects to enhance phonemic perception, but that prohibits top-down effects to overrule clear bottom-up input, is in many ways an optimal system. Because humans often must communicate under less than ideal conditions, the signal is often incomplete and/or ambiguous. Allowing lexical information to refine the phonemic encoding when the signal is not definitive produces a much higher hit rate for speech than would be achieved otherwise. Of course, if the input were made up of nonwords, rather than real words, such a system would work very badly. But, our systems have evolved to deal with the normal case, which is a stream of real words. By imposing our hypothesized constraint on top-down effects – clear bottom-up signals cannot be overruled – the system behaves optimally in the sense that it does not hallucinate. It would be a very poor design to allow expectations to dominate clear sensory information, but it would also be a poor design to ignore existing knowledge about the phonetic structure of lexical entries. It appears that our systems have evolved to avoid both of these poor options, providing us with the ability to recognize the often-noisy words we encounter in the real world.

## Acknowledgments

We thank Donna Kat for her invaluable help and Elizabeth Cohen for her assistance in data collection. Julia Irwin provided great help with the videorecording, and Doug Whalen kindly made facilities at Haskins Labs available (NIH grant HD-01994). We reviewers for their constructive suggestions. Support was provided by NIMH Grant R0151663 and NSF Grant 0325188.

## References

Barutchu A, Crewther S, Kiely P, Murphy M, Crewther DP. When /b/ill with /g/ill becomes /d/ill: evidence for a lexical effect in audiovisual speech perception. European Journal of Cognitive Psychology. 2008; 20(1):1–11.

Bertelson P, Vroomen J, de Gelder B. Visual recalibration of auditory speech identification: A McGurk after effect. Psychological Science. 2003; 14(6):592–597. [PubMed: 14629691]

Besle J, Fort A, Delpuech C, Giard M-H. Bimodal speech: early suppressive visual effects in human auditory cortex. European Journal of Neuroscience. 2004; 20:2225–2234. [PubMed: 15450102]

Brancazio L. Lexical influences in audiovisual speech perception. Journal of Experimental Psychology: Human Perception and Performance. 2004; 30:445–463. [PubMed: 15161378]

Connine CM. Altmann GTM. Effects of sentence context and lexical knowledge in speech processing. Cognitive models of speech processing. 1990:281–294.

Connine CM, Clifton C Jr. Interactive use of lexical information in speech perception. Journal of Experimental Psychology: Human Perception and Performance. 1987; 13:291–299. [PubMed: 2953858]

Eimas PD, Corbit JD. Selective adaptation of linguistic feature detectors. Cognitive Psychology. 1973; 4:99–109.

Elman JL, McClelland JL. Cognitive penetration of the mechanisms of the mechanisms of perception: Compensation for coarticulation of lexically-restored phonemes. Journal of Memory and Language. 1988; 27:143–165.

Ganong WF. Phonetic categorization in auditory perception. Journal of Experimental Psychology: Human Perception and Performance. 1980; 6:110–125. [PubMed: 6444985]

Goldinger SD. Words and voices: Episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996; 22(5): 1166–1183.

Green KP, Norrix LW. Perception of /r/ and /l/ in a stop cluster: Evidence of cross-modal context effects. Journal of Experimental Psychology: Human Perception and Performance. 2001; 27:166–177. [PubMed: 11248931]

Jones BC, Feinberg DR, Besterlmeyer PEG, DeBruine LM, Little AC. Adaptation to different mouth shapes influences visual perception of ambiguous lip speech. Psychonomic Bulletin & Review. 2010; 17(4):522–528. [PubMed: 20702872]

Klucharev V, Möttönen R, Sams M. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Cognitive Brain Research. 2003; 18(1):65–75. [PubMed: 14659498]

Liberman AL, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review. 1967; 74:431–461. [PubMed: 4170865]

Magnuson JS, McMurray B, Tanenhaus MK, Aslin RN. Lexical effects on compensation for coarticulation: the ghost of Christmash past. Cognitive Science. 2003; 27:285–298.

Mann VA, Repp BH. Influence of preceding fricative on stop consonant perception. Journal of the Acoustical Society of America. 1981; 69:548–558. [PubMed: 7462477]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264:746–748. [PubMed: 1012311]

Miller GA, Isard S. Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Verbal Behavior. 1963; 2:217–228.

Mirman D, McClelland JL, Holt LL. Computational and behavioral investigations of lexically induced delays in phoneme recognition. Journal of Memory and Language. 2005; 52:424–443.

Nath AR, Beauchamp MS. A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. Neuroimage. 2012; 59:781–787. [PubMed: 21787869]

Norris D, McQueen JM, Cutler A. Merging information in speech recognition: Feedback is never necessary. Behavioral & Brain Sciences. 2000; 23:299–370. [PubMed: 11301575]

Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Perception & Psychophysics. 1998; 60(3):355–376. [PubMed: 9599989]

Ostrand, R.; Blumstein, SE.; Morgan, JL. When hearing lips and seeing voices becomes perceiving speech: Auditory-visual integration in lexical access. In: Carlson, L.; Hölscher, C.; Shipley, T., editors. Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society; 2011. p. 1376-1381.

Pitt MA, Samuel AG. An empirical and meta-analytic evaluation of the phoneme identification task. Journal of Experimental Psychology: Human Perception and Performance. 1993; 19:1–27.

Roberts M, Summerfield Q. Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. Perception & Psychophysics. 1981; 30:309–314. [PubMed: 7322807]

Saldaňa AG, Rosenblum LD. Selective adaptation in speech perception using a compelling audiovisual adaptor. Journal of the Acoustical Society of America. 1994; 95:3658–3661. [PubMed: 8046153]

Sams M, Manninen P, Surakka V, Helin P, Katto R. McGurk effect in Finnish syllabes, isolated words, and words in sentences: effects of word meaning and sentence context. Speech Communication. 1998; 26:75–87.

Samuel AG. Phonemic restoration: Insights from a new methodology. Journal of Experimental Psychology: General. 1981; 110:474–494. [PubMed: 6459403]

Samuel AG. Red herring detectors and speech perception: In defense of selective adaptation. Cognitive Psychology. 1986; 18:452–499. [PubMed: 3769426]

Samuel AG. Does lexical information influence the perceptual restoration of phonemes? Journal of Experimental Psychology: General. 1996; 125:28–51.

Samuel AG. Insights from a failure of selective adaptation: Syllable-initial and syllable-final consonants are different. Perception & Psychophysics. 1989; 45:485–493. [PubMed: 2740189]

Samuel AG. Lexical activation produced potent phonemic percepts. Cognitive Psychology. 1997; 32:97–127. [PubMed: 9095679]

Samuel AG. Knowing a word affects the fundamental perception of the sounds within it. Psychological Science. 2001; 12:348–351. [PubMed: 11476105]

Samuel AG, Frost R. Do lexical representations support phonetic perception in non-native listening as they do in native listening?. (in preparation).

Samuel AG, Kat D, Tartter VC. Which syllable does an intervocalic stop belong to? A selective adaptation study. Journal of the Acoustical Society of America. 1984; 76:1652–1663. [PubMed: 6520303]

Samuel AG, Pitt MA. Lexical activation (and other factors) can mediate compensation for coarticulation. Journal of Memory and Language. 2003; 48:416–434.

Schacter DL, Church BA. Auditory priming: Implicit and explicit memory for words and voices. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1992; 18(5):915–930.

Stekelenburg JJ, Vroomen J. Neural correlates of multisensory integration of ecologically valid audiovisual events. Journal of Cognitive Neuroscience. 2007; 19(12):1964–1973. [PubMed: 17892381]

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America. 1954; 26:212–215.

van Alphen PM, McQueen JM. The time-limited influence of sentential context on function word identification. Journal of Experimental Psychology: Human Perception and Performance. 2001; 27(5):1057–1071. [PubMed: 11642695]

van Linden S. Recalibration by auditory phoneme perception by lipread and lexical information. Doctoral Thesis, Tilburg University, Tilburg, the Netherlands. 2007 ISBN978-90-5335-122-2.

van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(4):1181–1186. [PubMed: 15647358]

Warren RM. Perceptual restoration of missing speech sounds. Science. 1970; 167:392–393. [PubMed: 5409744]

Vroomen J, de Gelder B. Lipreading and the compensation for articulation mechanism. Language and Cognitive Processes. 2001; 16(5):661–672.

Vroomen J, Stekelenburg JJ. Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. Journal of Cognitive Neuroscience. 2010; 22:1583–1596. [PubMed: 19583474]

Vroomen J, van Linden S, de Gelder B, Bertelson P. Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting buildup courses. Neuropsychologia. 2007; 45:572–577. [PubMed: 16530233]

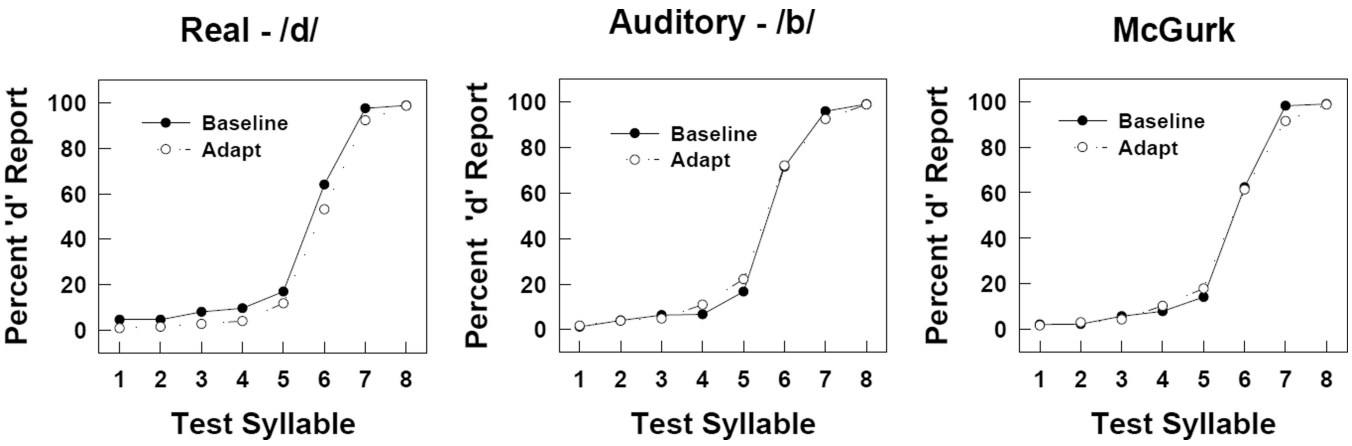


**Figure 1.**

Identification of the /bI/-/dI/ test syllables, before and after adaptation. The x-axis represents the eight test syllables, with stimulus 1 being most like /b/, and 8 being most like /d/. Left Panel: Real-/d/ condition; Middle Panel: Auditory-/b/ condition; Right Panel: McGurk condition.

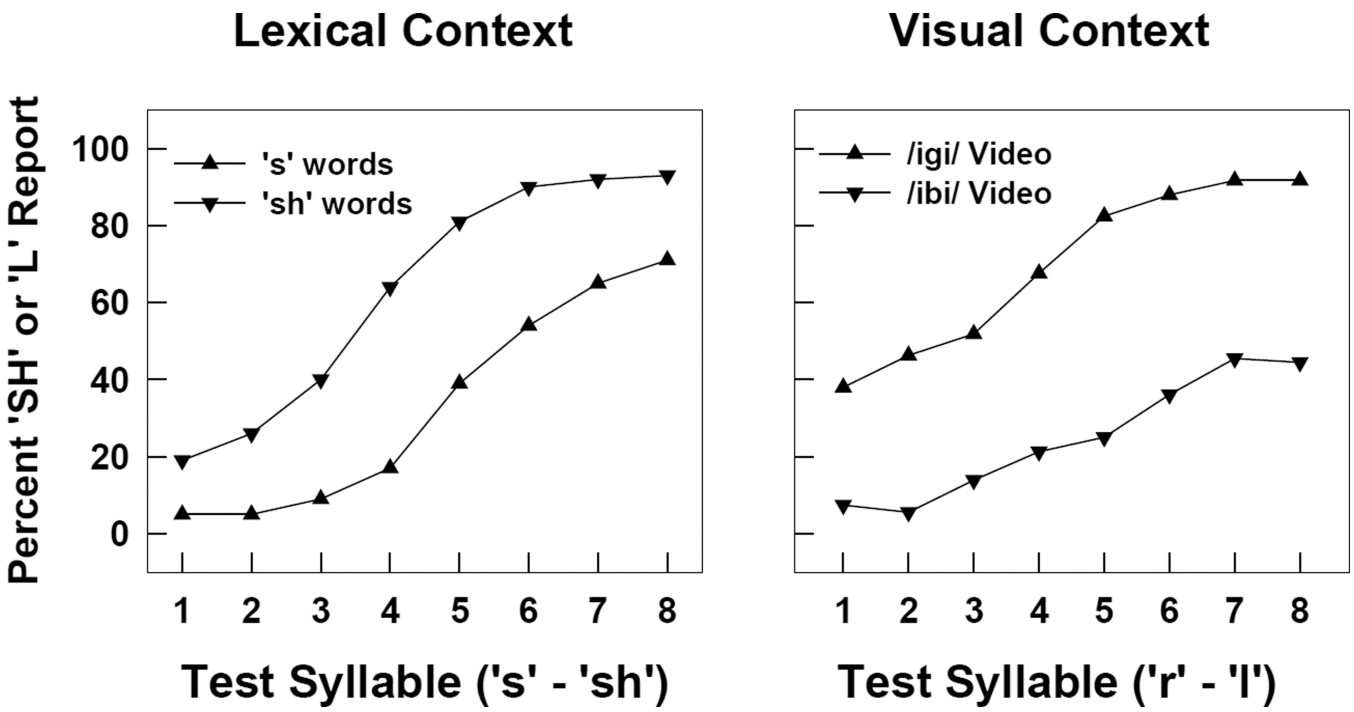


**Figure 2.**

Left Panel: Identification of test items, adapted from Pitt and Samuel (2006). The x-axis represents the eight test items, with stimulus 1 being most like /Is/ (“iss”) and stimulus 8 being most like /I∫/ (“ish”). Right panel: Identification of the /ili/ - /iri/ test items in Experiment 2a. The x-axis represents the eight test items, with stimulus 1 being most like /r/, and 8 being most like /l/.

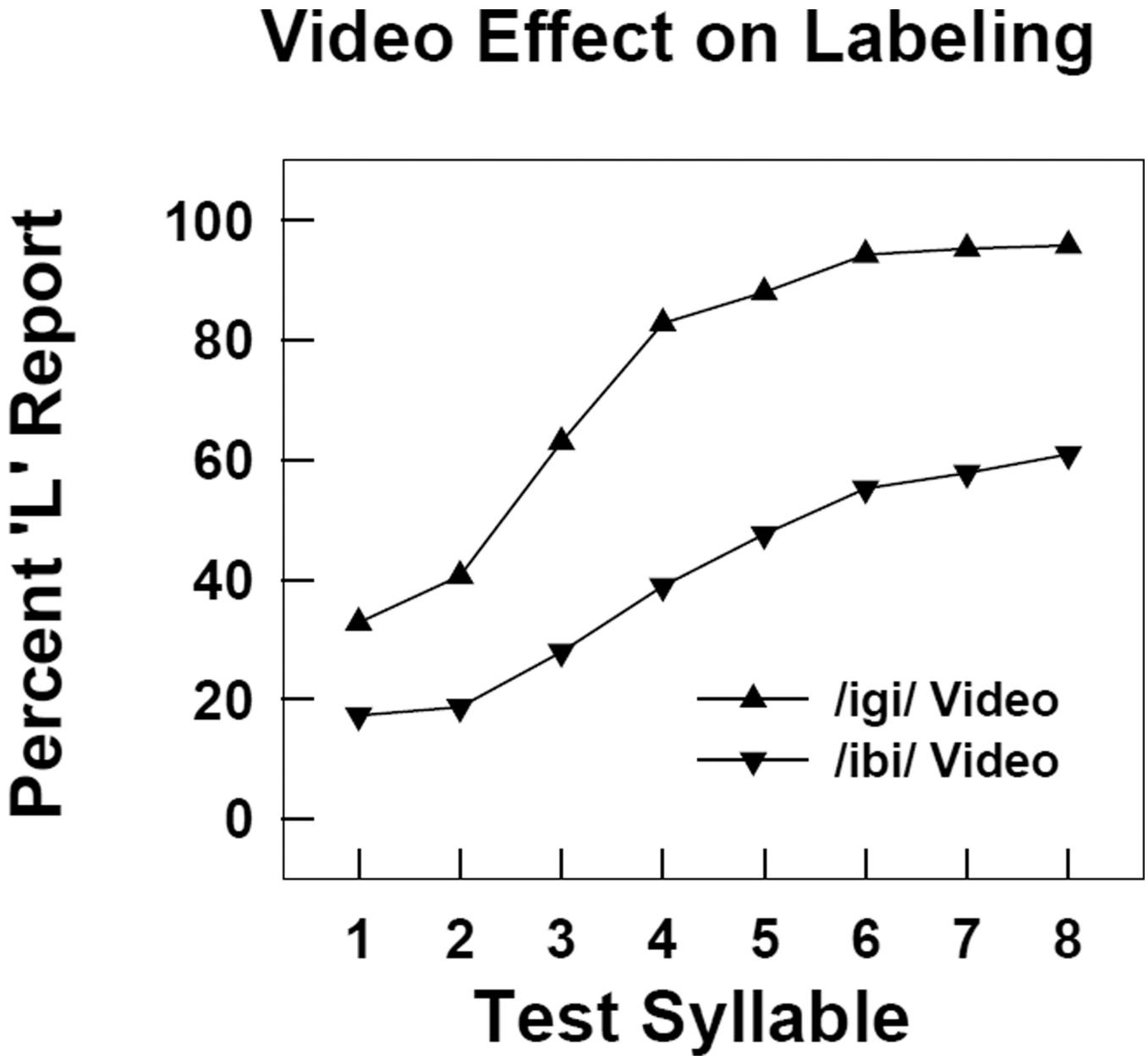


**Figure 3.**

Identification of the /ili/ - /iri/ test items in Experiment 2b during the baseline identification task. The x-axis represents the eight test items, with stimulus 1 being most like /r/, and 8 being most like /l/.

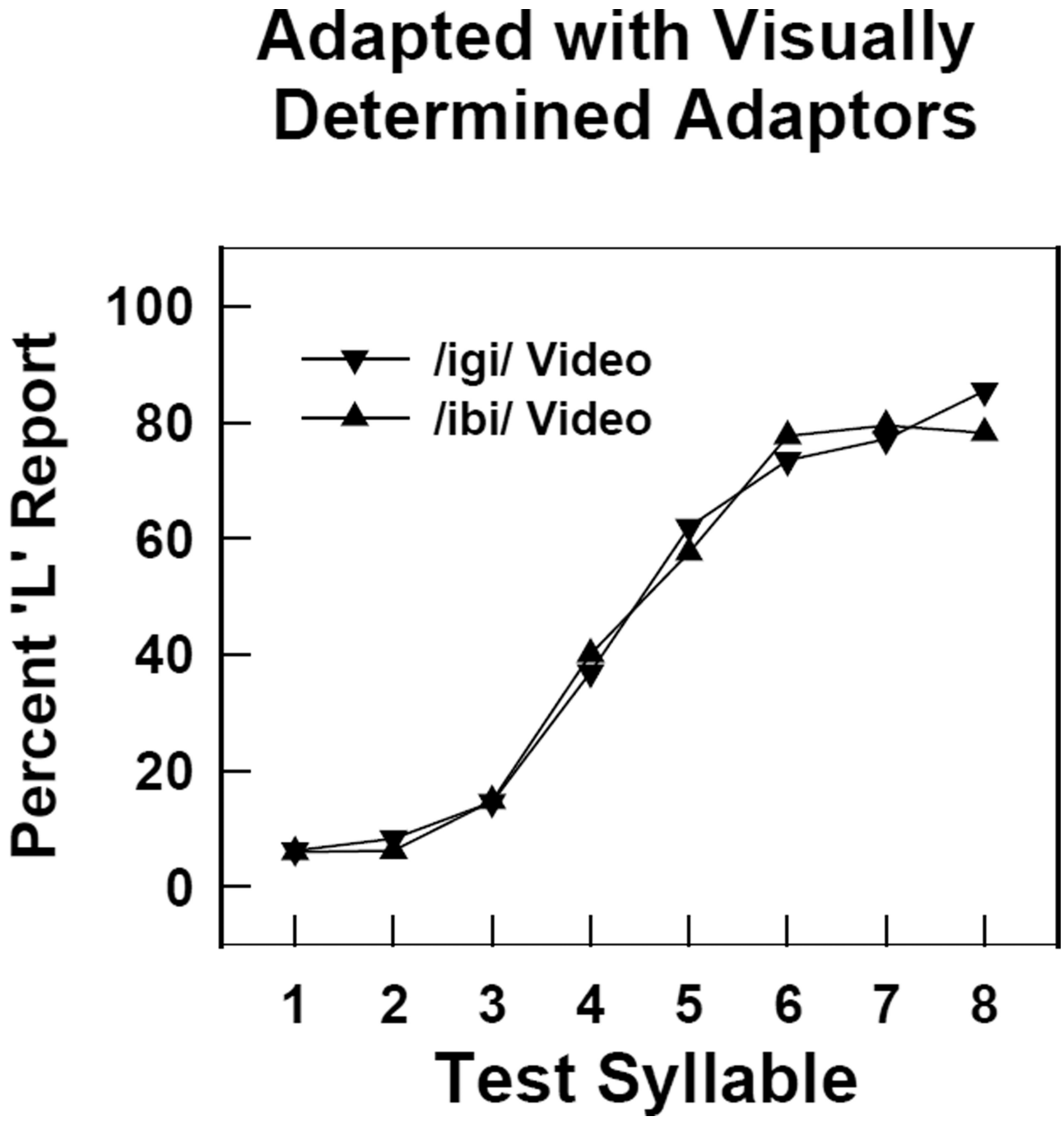


**Figure 4.**
Identification of the /ili/ - /iri/ test items in Experiment 2b during the adaptation task. The x-axis represents the eight test items, with stimulus 1 being most like /r/, and 8 being most like /l/.

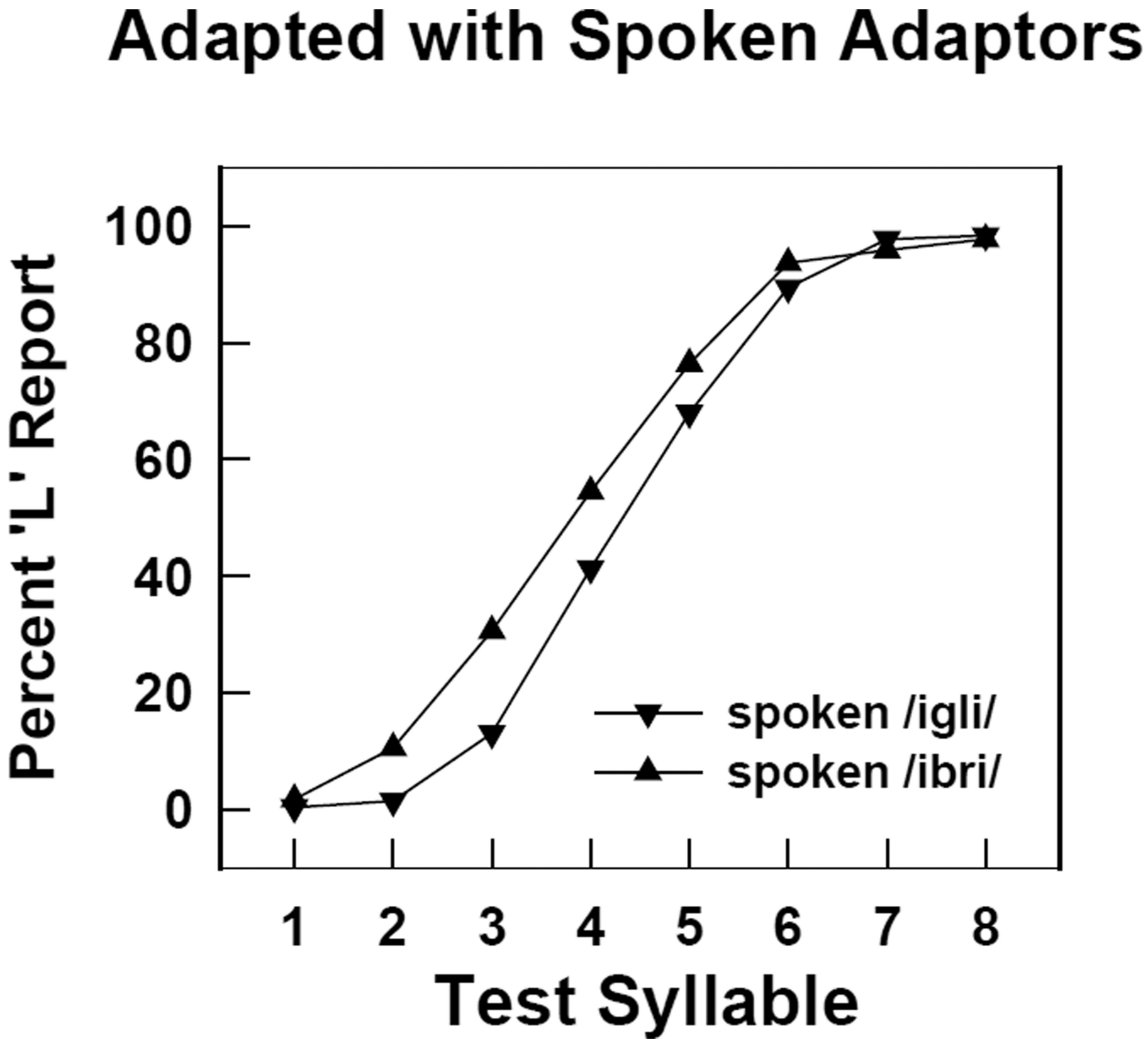


**Figure 5.**

Identification of the /ili/ - /iri/ test items during the adaptation task in the control experiment in which the adaptors were naturally-produced /ibri/ and /igli/. The x-axis represents the eight test items, with stimulus 1 being most like /r/, and 8 being most like /l/.