Testing the P-map hypothesis: Coda devoicing*

Shigeto Kawahara and Kelly Garvey

Abstract

Many languages disfavor coda voiced stops, but the number of ways in which languages resolve coda voiced stops is limited: i.e. languages alter voiced stops by devoicing but not by any other phonological means. For example, underlying /ab/ can become [ap], but not *[am], *[aba] or *[a]. To explain this observation, Steriade (2001/2008) claims that (i) speakers maximize the perceptual similarity between inputs and outputs, assuming that (ii) devoicing universally yields an outcome that is perceptually most similar to the original form. This paper reports a series of similarity judgment experiments to test the premise (the clause (ii)). The results are mixed: several orthography-based judgment studies of English speakers demonstrate that devoicing yields an outcome that is most similar to the target forms, compared to nasalization, deletion, or epenthesis. However, the auditory follow-up experiments reveal a more complex picture. The experiments overall provide support for the P-map hypothesis, but they also suggest a nuanced picture of what the underlying knowledge of similarity must be.

1 Introduction

1.1 The aim of the experiments

Many languages disfavor coda voiced stops, but the number of ways in which languages resolve a constraint against coda voiced stops is limited. A host of historically unrelated languages alter voiced stops in codas by devoicing; see Myers (to appear) for a recent survey. However, languages do not resolve a constraint against coda voiced stops by any other phonological means (Lombardi, 2001; Steriade, 2001/2008). For example, underlying /ab/ can become [ap], but not *[am], *[aba]

*This project is partly funded by a Research Council Grant from Rutgers University. We would also like to express our thanks to research assistants at the Rutgers phonetics lab—Lara Greenberg, Sophia Kao, and Shanna Lichtman—for their help in developing this research program. Many thanks to the audiences at NELS 40, SUNY and 2010 Rutgers Aresty Undergraduate Conference, as well as Osamu Fujimura, Kazu Kurisu, Toshio Matsuura, Jeremy Perkins, and Kyoto Yamaguchi for their comments on this project. Any remaining errors are ours

or *[a].[1] Relatedly, in loanword adaptation, when a recipient language which lacks voiced stops adapts voiced stops, they are rendered as voiceless stops rather than nasals (Adler, 2006; Kenstowicz, 2003). The absence of some strategies is an example of a "too-many-solutions" problem; the number of the ways in which languages resolve some marked structures is limited (Steriade, 2001/2008). In order to explain why languages only resort to devoicing to resolve a restriction against coda voiced stops, Steriade (2001/2008) claims that (i) speakers maximize the perceptual similarity between inputs and outputs, assuming that (ii) devoicing yields an outcome that is perceptually most similar to the original form. The mechanism underlying this proposal is called the P-map, which is general knowledge of similarity, or in words of Steriade, "a set of statements about relative perceptibility of different contrasts, across the different contexts where they might occur" (p. 151).

The first part of Steriade's (2001/2008) proposal has been extended to explain other phonological phenomena (broadly construed, including loanword adaptation and verbal art patterns) (e.g. Adler 2006; Côté 2004; Jun 2004; Kaplan 2010; Kawahara 2006, 2007; Kawahara & Shinohara 2009; Kenstowicz 2003; Kwon 2005; McCarthy 2010; Steriade 2001a, 2003; Wilson 2006; Zuraw 2007).[2] However, the premise (the clause (ii)) of this proposal, which should be part of the P-map or speakers' similarity knowledge, has never been systematically tested. This paper reports experiments that fill that gap.

One caveat is in order, before proceeding: the aim of this paper is to test the premise of the P-map hypothesis rather than proving its validity as a general theory of the too-many-solutions problem. In order for the P-map hypothesis to be correct, its premise needs to be supported; however, the P-map hypothesis can be falsified even if our results turn out to support its premise.[3] See McCarthy (2008) for a review of other proposed solutions to the too-many-solutions problem.

Another note on similarity judgment patterns—or speech perception systems that underlie

[1] In Viamu Picardo, such apparent nasalization does occur, but after nasalized vowels (Jose & Auger, 2004). This case can be considered as assimilation in nasality. We will thus set this case aside, and focus on coda positions preceded by oral vowels.

[2] Implemented within Optimality Theory (Prince & Smolensky, 1993/2004), this thesis means that "correspondence constraints are ranked as a function of the relative distinctiveness of the contrasts they refer to" (Steriade 2001/2008: 164).

[3] One prominent challenge to the P-map hypothesis is how to account for cases of heterogeneity of processes (McCarthy, 2002), in which a marked structure is resolved by multiple phonological means. A clear example is nasal-voiceless stop clusters, which are resolved in multiple ways (Pater, 1999). One idea proposed by Steriade (2001b, 2003) is that what P-map dictates is a default state, but other factors—such as diachronic sound changes—can create situations which do not follow the P-map. This idea—that phonetic naturalness in phonology is default but not an absolute requirement—was anticipated by some previous work (Dresher, 1981; Postal, 1968), and has been pursued by some recent work (Hayes et al., 2009; Kawahara, 2008; Wilson, 2006). This proposal in particular predicts that novel phonological patterns—for example, new phonological patterns in new loanwords or in an artificial language learning environment—follow patterns predicted by P-map (Kawahara, 2008; Wilson, 2006). Steriade (2001b) claims that even in languages whose native phonology resolves nasal-voiceless stop clusters in ways other than post-nasal voicing, innovative patterns (like loanword adaptation and innovative dialects) show post-nasal voicing, suggesting that a change in [voice] is the default option.

them—is in order. It is not the case that speech perception systems—and the P-map that builds on them—are entirely universal, because language-specific phonetic/phonological knowledge affects speech perception (Dupoux et al., 1999; Massaro & Cohen, 1983; Moreton, 2002; Pitt, 1998). However, it does not mean that there cannot be universal speech perception patterns, which can sometimes be observed, for example, in the perception patterns of non-speech sounds (Holt et al., 2004; Kingston, 2005; Kingston et al., 2009; Mann, 1986). Therefore, we believe that it is too soon to give up the P-map hypothesis, despite that our linguistic knowledge affects speech perception. See section 8 for more discussion.

1.2 A brief review of the previous studies

We now turn to a brief review of the previous psycholinguistics literature on which Steriade (2001/2008) bases her claim to show that devoicing yields a form that is perceptually most similar to the unaltered form. One kind of source comes from multi-dimensional scaling analyses of similarity judgments, which show that manner of articulation (e.g. sibilancy and sonorancy) generally provides more important auditory dimensions than voicing (Walden & Montgomery, 1975; Peters, 1963). The same tendency is observed in a similarity judgment task (Bailey & Hahn 2005: 360) and, to some degree, a word association task (Greenberg & Jenkins, 1964). See Steriade (2001/2008) for a summary of other studies.[4]

Another line of support comes from the studies of verbal art in several languages. There has been a long-standing observation that speakers can create verbal art—rhymes, alliteration, and puns—by combining two "similar" but not identical sounds (e.g. *hatred* vs. *make it*: Zwicky 1976). Studies of verbal art have shown that speakers are more willing to combine two consonants that differ in voicing than those that differ in other phonological dimensions. The more frequent appearance of voicing mismatches has been documented in Japanese rap lyrics (Kawahara, 2007) and Japanese pun patterns (Kawahara & Shinohara, 2009; Shinohara, 2004). A voicing contrast mismatch has also been found to be common in English imperfect puns (Lagerquist 1980: 186; Zwicky and Zwicky 1986: 500), Pinsky's slant rhymes found in his translation of the Inferno of Dante (Hanson, 2003),[5] Romanian half rhymes, especially post-nasally and word-finally (Steriade 2003: section 5), and limited so in English rock rhymes (Zwicky 1976: 685-686).

[4]A voicing contrast is known to be robust under white noise (Miller & Nicely, 1955). This robustness may be due to the fact that a voicing contrast is conveyed by various durational cues (Kingston & Diehl, 1994; Lisker, 1986; Port & Dalby, 1982), and that white noise does not cover durational cues well. This difference between voicing and other manner features is in fact diminished under signal-dependent noise, to the degree that the difference becomes non-significant (Benkí 2003: 140). See also Shepard (1972: 106-107) for a discussion of this potentially ambivalent behavior of a voicing contrast.

[5]Hanson (2003: 391) also notes that "the practice of allowing differences in voice in rhymes seems to have a longer history in English than is commonly acknowledged", and cites other examples of rhyme patterns that contain rhyme pairs that differ in voicing. However, she also notes later (p. 322) that a voicing mismatch is limited to coronal pairs.

1.3 The current experiments

Therefore, there are good reasons to believe that the P-map's premise is on the right track. However, so far no systematic studies have directly tested it. One exception is Myers (to appear), which has shown that English utterance-final stops are semi-devoiced, and hence can be confusable with their voiceless counterparts. This study however only tests the confusability of coda voiced stops with the voiceless couterparts; i.e. it does not test the confusability of coda voiced stops with, for example, nasalized forms.

This paper thus aims to test the premise of the P-map hypothesis in the most direct way. To this end, this paper reports six similarity judgment experiments that test the premise. Our findings are mixed. Four orthography-based tests dominantly support the premise; the audio-based experiments however show that similarity judgment patterns depend on detailed information of phonetic implementation. The experiments overall provide support for the P-map hypothesis, but they also suggest a nuanced picture of what the underlying knowledge of similarity must be.

Before proceeding, some comments are in order as to why we started with orthography-based studies. One reason is the efficiency; we can obtain a large amount of participants rather easily, especially in the form of online-based studies that we employ in the current project. Second, we can avoid the effect of misperception—in an auditory presentation, some stimuli might fail to be distinguished by the listeners, so that the results could be a mixture of discriminability and judged similarity. Third, by not presenting sounds, we may tap speakers' abstract knowledge of similarity that goes beyond particular phonetic implementation patterns. The first four experiments are thus orthography-based. Two follow-up audio-based experiments are reported as Experiment V and VI.

2 Experiment I: A multiple-choice similarity judgment task

2.1 Introduction

The first experiment was a multiple-choice similarity judgment experiment. In this experiment, English speakers were first presented with a form that contains coda voiced stops and then various forms which each undergo devoicing, nasalization, deletion and epenthesis. They were then asked to choose the form that sounds most similar to the original form.

2.2 Method

The target stimuli contained coda voiced stops, and for each target, the participants were presented with four options that each represent the output of devoicing, nasalization, deletion and epenthesis.[6] For example, for [ab], the four options were [ap], [am], [a], and [apa]. (The epenthetic vowel is most likely pronounced with a reduced vowel by native speakers of English; we address the issue of epenthetic quality in Experiments IV-VI.) The target stimuli were [ab], [ad], [ag], [itab], [itad], and [itag]. Disyllabic stimuli were added because speakers may disfavor deletion in monosyllabic stimuli because of the minimal word requirement in English (Hammond, 1999; McCarthy & Prince, 1986): i.e. [a] may be disfavored because it sounds "too short". Since all target items involve coda voiced stops, 6 fillers were added: [am], [an], [na], [ma], [da], and [ga].

The experiment was administered online through Sakai[7] using English orthography. The experiment site first showed the instructions to the experiment as well as the consent form to a human subject experiment. For each target word, four choices were presented, and the participants were asked which of the option sounds most similar to the target word. The order of these choices and the presentation of the 12 stimuli were randomized by Sakai. The nasalization of [g] was represented by [ng] with a note that these letters represent the last sound in "sang". Although the experiment was based on English orthography, the participants were asked to read the stimuli before they answer the questions and base their judgments on their auditory quality. 32 native speakers of English completed Experiment I.

For a statistical analysis, the posited null hypothesis was that each speaker was responding randomly and therefore would choose devoicing 1.5 (=$6/4$) times out of 6 target items. A non-parametric within-subject Wilcoxon test was used to assess this null hypothesis. All statistic analyses in this paper were performed using R (R Development Core Team, 1993-2010), which also generated illustrative graphs.

2.3 Results and discussion

Table 1 illustrates the percentages of forms that were judged to be the most similar to the target items with coda voiced stops. Figure 1 provides pie-graph representations of Table1.

For all targets, the speakers most often chose the devoiced outcome as most similar to the original forms. Recall that the null hypothesis was that each speaker was responding randomly and hence would choose devoicing 1.5 (=$6/4$) times out of 6 items. A non-parametric Wilcoxon test shows that the preferences toward devoicing did not arise by chance ($p < .001$). This result

[6]There are three other phonological strategies that could possibly satisfy a constraint against coda voiced stops (Steriade, 2001/2008): segmental reversals, featural transfer, and lenition to glides. The current experiments do not consider the first two alternatives because they can be ruled out based on independent phonological grounds. First, segmental reversal (/tab/ → [bat]) is independently ruled out because long-distance metathesis is known not to occur in natural languages (Carpenter, 2002; Hume, 2001; McCarthy, 2006). Featural transfer (/tab/ → [dap]) is not possible when there are no voiceless onset consonants on which [+voice] can dock. Moreover, we do not know of any clear case in which a consonantal feature flops across a vowel onto a distant consonant. Finally, lenition to glides (/tab/ → [taw]) is not considered in this experiment, because corresponding glides for [d] and [g] are not clear in English.

[7]`https://sakai.rutgers.edu/portal`

Table 1: The percentages of forms that were judged to be most similar to the target items with coda voiced stops.

stems	ab	ad	ag	itab	itad	itag	average
devoicing	75%	71.9%	68.8%	87.5%	78.1%	68.8%	75%
nasalization	12.5%	12.5%	15.6%	9.4%	3.1%	12.5%	10.9%
deletion	3.1%	6.3%	3.1%	0%	9.4%	15.6%	6.3%
epenthesis	9.4%	9.4%	12.5%	3.1%	9.4%	3.1%	7.8%

shows that English speakers do think that devoiced forms are most similar to the original, unaltered forms, and that this preference does not arise by chance.

3 Experiment II: A binary-choice similarity judgment task

3.1 Introduction

To further verify the results of Experiment I and to compare the similarity differences caused by each phonological process, a follow-up similarity judgment task was conducted with binary comparisons. The design involved all binary comparisons of four phonological processes (devoicing, nasalization, deletion, and epenthesis).

3.2 Method

The target stimuli included [ab], [ad], [itab], and [ikad].[8] The stimuli with coda dorsal stops were excluded because the binary comparison design in Experiment II involves more comparisons (4*3/2 = 6 comparisons between four different phonological processes) than Experiment I, and it was expected (at the time of this experiment) that there would be a psychological limit on how many questions speakers can focus on in an online test (Hayes et al., 2009). Hayes et al. state in footnote 19 (p. 840) that "our experience in a pilot study was that most participants would not complete more than about fifteen items [in an online linguistics test]."[9] The stimuli with coda dorsal stops were excluded because English does not offer an orthography to represent nasalized [g] (i.e. [ŋ]), and hence it may not be ideal. (We did not consider it to be a crucial problem given

[8]In this experiment, as well as in Experiment III, V, VI, we used [ikad] instead of [itad], because speakers may disfavor [itat] because of the two [t]s in adjacent syllables (Coetzee 2008 and references cited therein). This consideration was not made when we ran Experiment I and Experiment IV (Experiment IV was a direct follow-up of Experiment I). The inter-experimental consistency here is not intentional.

[9]Having run many other online-based linguistic experiments, we now believe that this assumption was too conservative. Those speakers who are willing to participate in experiments can deal with many more questions than 15. See Reips (2002) for general discussion on pros and cons of online-based experimentation.

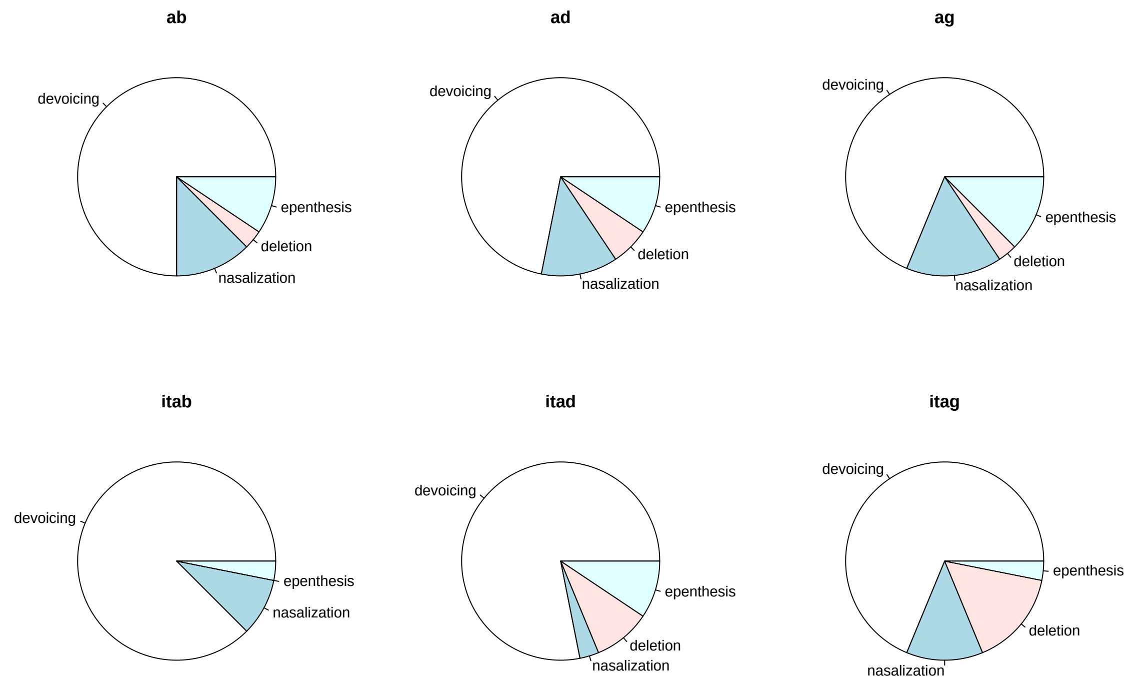

Figure 1: The percentages of forms that were judged to be most similar to the target items with coda voiced stops.

the results in Experiment I, but since we needed to exclude something, we excluded this less-than ideal case.) As a result, Experiment II contained 24 stimuli (6 comparisons * 4 stems).

The procedure was almost identical to Experiment I. Given binary choices, the participants were asked which of the options sounds most similar to the target word. For example, one question asked which one of [ap] and [aba] is more similar to [ab]. The order of these two choices was randomized by Sakai. The overall experiment was organized into two smaller blocks. The first block contained all 12 monosyllabic stimuli (6 comparisons * 2 target stems) followed by a break sign where the participants were encouraged to take a break. After the break, the second block contained all 12 disyllabic stimuli. The order of the stimuli within a block was randomized by Sakai. 35 native speakers of English participated in Experiment II.

Within-subject non-parametric Wilcoxon tests were used to assess any skews in the speakers' responses. The alpha level was adjusted according to the number of multiple comparisons.

3.3 Results and discussion

Table 2 illustrates the percentages of devoiced forms chosen as more similar to the target forms than the outcomes of other phonological processes. Figure 2 and Figure 3 visually illustrate the comparisons separately for monosyllabic and disyllabic stimuli.

Table 2: The percentages of devoicing chosen as more similar than other phonological processes.

	against nasalization	against deletion	against epenthesis
ab	69.4%	88.9%	61.1%
ad	86.1%	83.3%	58.3%
itab	75%	91.7%	72.2%
ikad	80.6%	91.7%	66.7%
average	77.8%	88.9%	64.6%

The participants chose devoicing as yielding more similar outcomes than any other phonological operations in all the comparisons in all the stems ($p < .001$). Among the three competitors, epenthesis seems to be the strongest competitor against devoicing; i.e. epenthesis may yield the second most similar forms to the unaltered form with a coda stop.

We also note that generally, devoicing was considered to be similar to the original forms more often in disyllabic forms than in monosyllabic forms. We may conjecture that word-final voicing is weaker in longer words than in shorter words because of the lowering of subglottal pressure, and hence word-final voiced stops have a weaker voicing contrast in long words than in short words. This hypothesis is speculative, though, and needs to be tested in a systematic acoustic study.

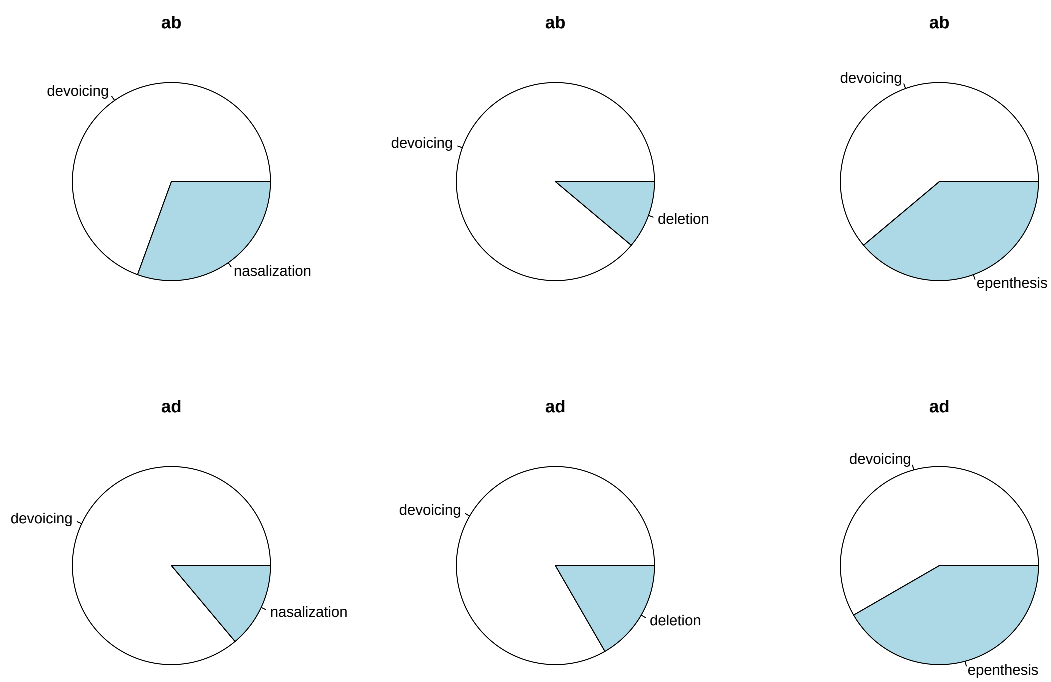

Figure 2: Comparisons involving devoicing: monosyllabic stimuli.

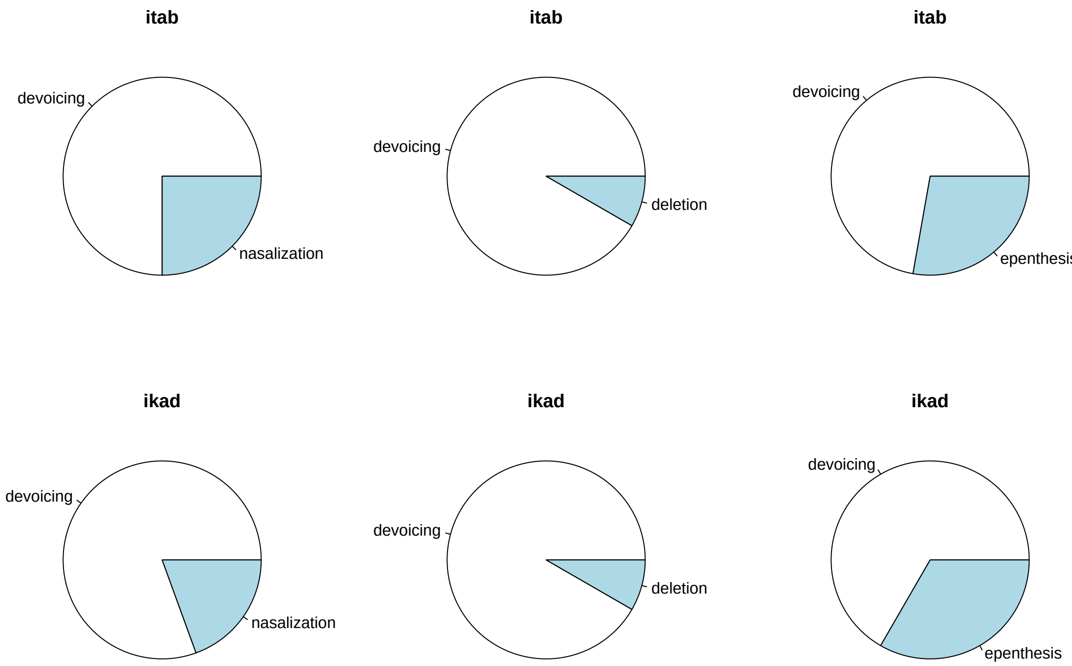

Figure 3: Comparisons involving devoicing: disyllabic stimuli.

Table 3 illustrates comparisons that do not involve devoicing, and Figure 4 and Figure 5 visually illustrate them. Epenthesis generally yielded a form that is more similar to the original form than nasalization and deletion did (except in one condition, [itab], where nasalization was chosen to be more similar than epenthesis). Nasalized forms were always judged to be more similar to the original form than forms with deletion.

Table 3: Processes that were chosen as more similar than the other, and their percentages of times that they were chosen to be more similar.

	ep vs. nas	ep vs. del	nas vs. del
ab	ep (66.7%)	ep (65.7%)	nas (69.4%)
ad	ep (72.2%)	ep (69.4%)	nas (52.8%)
itab	nas (55.6%)	ep (55.6%)	nas (80.0%)
ikad	ep (63.9%)	ep (66.7%)	nas (61.1%)

The total order of judged similarity, therefore, is: devoicing $>$ epenthesis $>$ nasalization $>$ deletion (where "$>$" means "is judged to be more similar to the unaltered form"). However, when we adjusted the alpha level for multiple comparisons ($\alpha = .05/3 = .016$), the last two comparisons—(i) epenthesis against nasalization and deletion and (ii) nasalization against deletion—did not turn out to be significant. Recall that the comparison of devoicing with the other three phonological repairs was significant ($p < .001$), highlighting the special status of devoicing.

To summarize, Experiment II shows that speakers treat devoicing differently from the other three phonological processes, again supporting the premise of Steriade (2001/2008) that devoicing yields a form that is most similar to the original, unaltered form. Comparisons among other phonological processes did not reveal significant differences, although epenthesis showed some hint toward yielding the second most similar form to the original form.

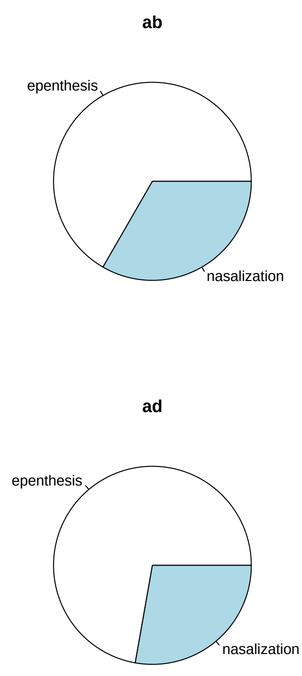

Figure 4: Comparisons involving non-devoicing processes: monosyllabic stimuli.

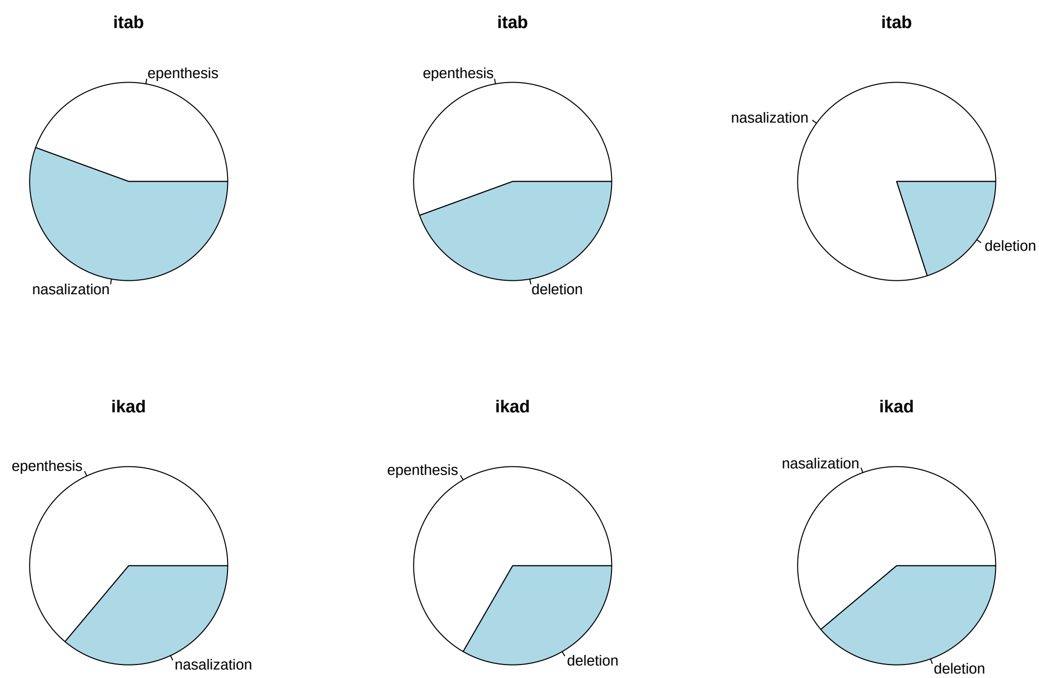

Figure 5: Comparisons involving non-devoicing processes: disyllabic stimuli.

4 Experiment III: A similarity rating task

4.1 Introduction

To confirm that the results obtained in Experiment I and II are not an artifact of task design (i.e. forced choice methods), a rating task was run. In this task, speakers were asked to judge the perceptual differences between two stimuli on a given scale.

4.2 Method

The target stimuli were [ab], [ad], [itab], [ikad]. These four stems were each compared to the outcome of four phonological operations (e.g. [ab]-[ap], [ab]-[am], [ab]-[a], and [ab]-[aba]). The total number of stimuli was thus 4*4=16. The stimuli with coda [g] were not included because 16 test items would have been sufficient for an online test (Hayes et al., 2009) (though see footnote 9). The scale was a 5-point scale: (A) almost identical, (B) very similar, (C) similar, (D) not so similar, and (E) completely different (the scale was provided alphabetically because it was the only option in Sakai). The responses were later converted numerically as follows: "almost identical"=5, "very similar"=4, "similar"=3, "not so similar"=2 and "completely different" =1.

The entire experiment was organized into two smaller blocks, preceded by a practice session with 3 items. The design included a practice session so as to allow participants to establish their subjective scale of similarity before proceeding to the main session. The first block contained monosyllabic stimuli with 8 items (2 stems * 4 comparisons) followed by a break sign. After the break, the second block contained all disyllabic stimuli. The order of the stimuli within a block but not the order of the options was randomized by Sakai. 27 native speakers of English participated in Experiment III.

A within-subject Wilcoxon contrast test compared the judged similarity scores between the devoiced form and the average of the other three forms.

4.3 Results and discussion

Table 4 and Figure 6 illustrate the averages of similarity ratings of the four forms in each condition.

The participants judged the devoiced forms (the leftmost bars) to be more similar to the original forms than other forms. A Wilcoxon contrast test compared the judged similarity scores between the devoiced form and the average of the other three forms, and revealed a statistically significant difference ($p < .001$). The result from the magnitude estimation study again supports the P-map's assumption (Steriade, 2001/2008) that devoicing is judged to yield a form that is most similar to the form with a coda voiced stop.

Table 4: Similarity ratings in each comparison.

	ab	ad	itab	ikad	average
devoicing	3.46	2.85	3.46	3.46	3.31
nasalization	1.77	2.00	2.5	2.5	2.19
deletion	1.56	1.62	2.42	2.5	2.02
epenthesis	2.23	2.04	2.16	2.19	2.15

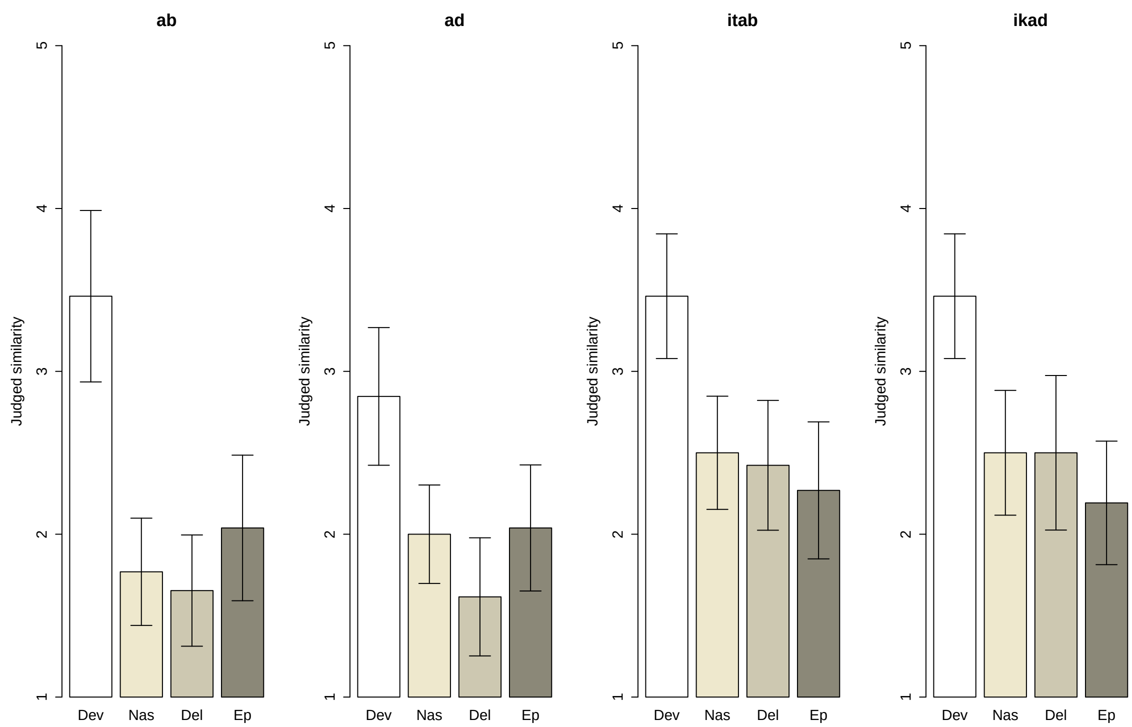

Figure 6: The average similarity ratings of the four types of forms with respect to the original forms with coda voiced stops. The error bars represent 95% confidence intervals based on variability across 27 speakers.

5 Experiment IV: High vowel epenthesis

5.1 Introduction

The concern that Experiment IV addressed is that in the three preceding experiments, the epenthetic candidate was created with an epenthetic [a]. Although the orthographic [a] in the final syllable of the epenthetic form (e.g. [aba]) is likely to have been read with a reduced vowel, there was also some concern. If the word-final epenthetic vowel was read as a full vowel, then [a] is in general longer in duration than other vowels (Lehiste, 1970; Lindblom, 1968; Peterson & Lehiste, 1960; Umeda, 1975). To address this concern, this experiment included the epenthetic target in which the epenthetic vowel is [i]. [i] is shorter (Peterson & Lehiste, 1960) and less intense (Fairbanks et al., 1950) than [a] in English[10]. This vowel is moreover arguably used for epenthesis in English (Yip, 1987) and other languages (Howe & Pulleyblank, 2004).

5.2 Method

The details of the experimental design were identical to Experiment I, except for a few aspects. As with Experiment I, there were 6 target stems ([ab], [ad], [ag], [itab], [itad], [itag]), together with 6 filler items. The task was a three-way forced choice similarity test, with the options being the devoiced candidate, a candidate with an epenthetic [a], and a candidate with an epenthetic [i]. The participants were gathered primarily through "Psychological Research on the Net" maintained by Dr. John H. Krantz, which hosts many online psychology experiments.[11] 35 native English speakers completed the survey.

5.3 Results and discussion

Table 5 and Figure 7 illustrate the percentages of each form chosen as being most similar to the original forms.

Speakers considered the devoiced forms as the most similar to the target forms ($p < .001$). Speakers did not consider forms with [i]-epenthesis more similar than the forms with [a]-epenthesis. As anticipated above, English orthographic representation *aba* is likely to have been read with a reduced second vowel (i.e. schwa), so that the second vowel may have been actually shorter in duration than [i]. To summarize, Experiment IV shows that devoicing is considered to be more similar to the original form than epenthesis, even when the epenthetic form contains a short high vowel [i].

[10] Though see Parker 2002 for a result that does not necessarily support the weaker intensity of high vowels with respect to low vowels.

[11] `http://psych.hanover.edu/research/exponnet.html`

Table 5: The percentages of forms that were judged to be most similar to the target items with coda voiced stops.

	ab	ad	ag	itab	itad	itag	average
devoicing	80.0%	77.1%	68.6%	74.3%	74.3%	77.1%	75.2%
[i]-epenthesis	2.9%	2.9%	5.7%	2.9%	5.7%	8.6%	4.8%
[a]-epenthesis	17.1%	20.0%	25.7%	22.9%	20.0%	14.3%	20%

6 Experiment V: Auditory Experiment I

6.1 Introduction

The previous four experiments support the premise of the P-map hypothesis. The four experiments used orthography-based tests, and recall that that was without reason (see section 1.3). However, since the P-map hypothesis is about perceptual similarity, a follow-up sound-based similarity judgment experiment was run.

6.2 Method

The target stimuli in this experiment included [ab], [ad], [ag], [itab], [ikad], [itag]. As with the previous experiments, for each form we prepared devoiced forms, nasalized forms, forms with deletion, forms with epenthetic [i] (e.g. for [ab], the four options were [ap], [am], [a], [abi]). [i] was used for the epenthetic vowel because it is short and non-intense, and also because it can be used as an epenthetic vowel in English (see section 5.1). Two female native speakers of English (both from New Jersey) were recorded in a sound-attenuated booth (one speaker is the second author). Each syllable was written on a separate index card, and the order was randomized. Their speech was recorded through an AT4040 Cardioid Capacitor microphone with a pop filter and amplified through an amplifier (DIGITAL MPA by ART). The recorded speech was digitized at a 44k sampling rate. The stimuli were placed in a frame sentence: “Please say the word X twice.” Speakers always put stress on the stem vowel [a] to avoid stress differences among different forms.

After the recording, the target stimuli were extracted at zero crossings using Praat (Boersma & Weenink, 1999-2010). Since the speakers did not assign a uniform pitch contour to all target syllables, the stimuli were re-synthesized with a flat pitch contour at 250Hz using Praat’s PSOLA function. The peak amplitude was adjusted to 0.7. Two tokens from each speaker were used for the listening experiment.

The experiment was run on Sakai, which allows us to embed sounds; the sound files were converted to mp3 files for embedding. For each question, we first presented a target stem with a coda voiced stop, followed by four options, all in auditory format (they appeared as play buttons

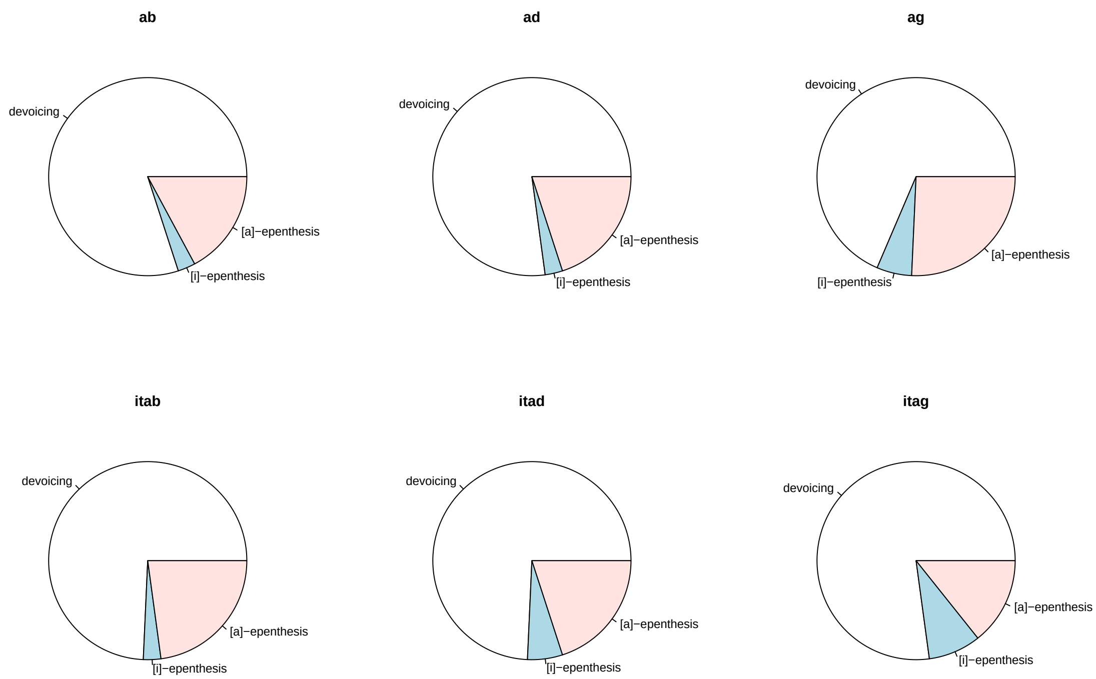

Figure 7: The results of the three-way similarity judgment task comparing devoiced forms, forms with [i]-epenthesis, and forms with [a]-epenthesis.

without orthography). To avoid misperception of stimuli, participants were allowed to listen to the stimuli as many times as they liked; in fact, they were encouraged to listen to all the sounds at least twice to avoid being confused which play button corresponds to which sound. All participants used high quality headphones (Sennheiser HD 280 Pro), and the experiment was run in a sound-attenuated room at the Rutgers phonetics laboratory. Listeners started with three practice questions before proceeding to the main session.

The participants were recruited from the psychology subject pool at Rutgers University and students of introductory linguistics classes. They primarily participated in the experiments for course requirements or extra credit. 30 native speakers of English completed this experiment. A within-subject Wilcoxon test was used to access the null hypothesis that listeners were responding randomly.

6.3 Results and discussion

Table 6 represents the percentages of forms that were judged to be most similar to the target items with coda voiced stops; Figure 8 represents the result visually.

The English listeners considered the devoiced forms as the most similar to the target forms,

Table 6: The percentages of forms that were judged to be most similar to the target items with coda voiced stops.

	ab	ad	ag	itab	ikad	itag	average
devoicing	56.1%	45.0%	47.5%	48.3%	56.7%	46.2%	50.0%
nasalization	21.1%	20.0%	18.6%	21.7%	9.2%	21.0%	18.6%
deletion	14.9%	20.8%	18.6%	22.5%	25.8%	23.5%	20.0%
epenthesis	7.9%	14.2%	15.3%	7.5%	8.3%	9.2%	10.4%

about 50% of the time. This skew in response was significantly different from chance ($p < .01$), although the skew is not as strong as in the previous experiments. A sound-based similarity judgment experiment yet again supports the hypothesis that devoicing yields the most similar outcome, although less straightforwardly so than the previous experiments.

One reason for the less extreme skew toward devoicing was because there were some item-specific effects. For example, one speaker (but not the other) flapped word-final [d] in one form: because the following word in the frame sentence was “twice”, she degeminated the final [d] with the following [t] and flapped it. The listeners thought nasalization was the most similar to this token, because both a flap and a nasal are both sonorants. The same speaker also produced one nasalization candidate which contained [ŋ] with a small nasal burst. This token was considered to be very similar to [g]. These results show that similarity judgments can be affected by phonetic details of particular tokens, and although devoicing turns out to be chosen as most similar most often, this preference can depend on how particular sounds are phonetically implemented.

The result then implies that, to the extent that the P-map hypothesis is on the right track as a theory of the too-many-solutions problem, its basis involves some abstraction: it is not the case that devoiced forms always yield forms that are most similar to the original forms—but devoicing does so most often. The similarity knowledge underlying phonology thus must be based on abstraction over experiences listening to multiple tokens.

7 Experiment VI: Auditory Experiment II

7.1 Introduction

In Experiment V, devoicing was chosen as most similar, but not as often as the previous four experiments. The result shows that similarity judgement patterns can be influenced by particular phonetic implementation patterns. Building on this result, the final experiment had two aims: (i) to further test the effect of particular phonetic implementations of phonological forms, and (ii) to use schwa for an epenthetic candidate.

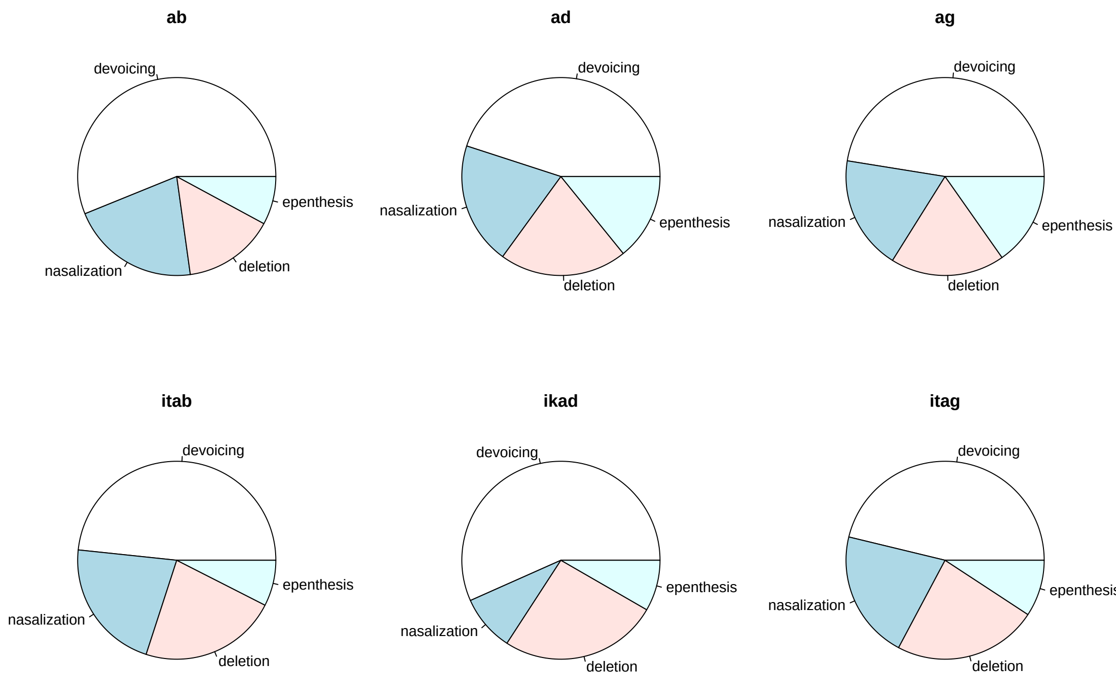

Figure 8: The percentages of each form chosen as most similar to target forms in the sound-based similarity judgment task.

7.2 Method

The stimulus structure is the same as Experiment V, except that the epenthetic candidate had a schwa. Three female native speakers of English (all from New Jersey), including the two speakers who produced the stimuli for the last experiment, were recorded in a sound-attenuated booth, digitized at 44K. The stimuli were placed in a frame sentence: "Please say the word X three times." To avoid flapping and reduction of word-final consonants, speakers were encouraged to release all the word-final consonants. They placed stress on the stem vowel [a]. The speakers repeated each token 10 times.

The target stimuli were extracted at zero crossings using Praat (Boersma & Weenink, 1999-2010); the stimuli were re-synthesized with a flat pitch contour at 250Hz. The peak amplitude was adjusted to 0.7. Out of 10 repetitions, those that had phonetic distortions (e.g. clipping, heavy creakiness, unintended vowel qualities, misplaced stress, nasal burst) were excluded. As a result, for all the stimuli, four tokens from each speaker were used for the listening experiment. Our pilot result shows that given the stimuli with (heavily) released word-final consonants, listeners chose epenthesis as a most similar form to the original, unaltered form (73%), presumably because vocalic releases in the original tokens resemble schwa very much (Silverman, to appear) (see Figure

9, Left). Therefore, we spliced off all heavy releases from voiced stops (Figure 9, Right); also, we spliced off heavy releases from voiceless stops and attached normally released stop bursts.

The procedure was identical to the previous experiment. The experiment started with a practice block with three items. The main session was organized into three blocks, each block presenting tokens from one speaker. We blocked the experiments by speaker so that the listeners would not be distracted by individual speech style differences. All sound tokens were repeated twice. Within each block, the same number of fillers were included. As a result each block contained 24 target items (6 stems * 4 tokens * 2 repetitions) and 24 fillers. The participants were students of introductory psychology or linguistics classes. 18 native speakers of English completed this experiment.

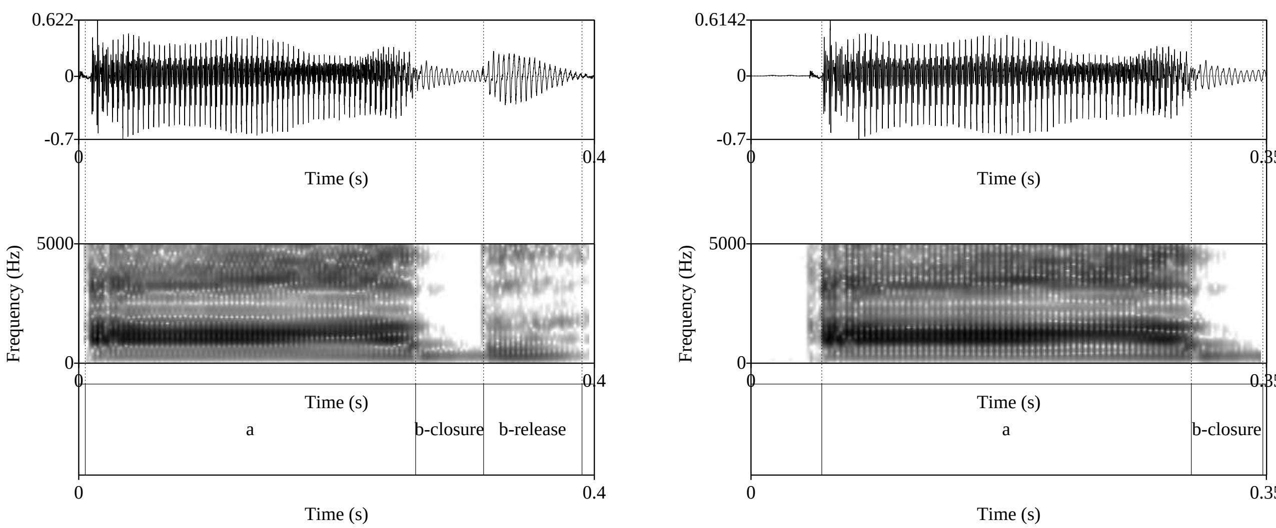

Figure 9: A released token of [b] (unedited, left) and its spliced token (right). The time scales are different.

7.3 Results and discussion

Table 7 and Figure 10 represent the percentages of forms that were judged to be most similar to the target items. Despite the fact that we spliced off audible releases, the listeners still chose the epenthetic candidate as most similar to the original form most often ($p < .05$). Devoicing on the other hand was not very often chosen as similar to the unaltered form.

The result of the last two auditory experiments shows that patterns of similarity judgments can crucially depend on details of phonetic implementation. In particular, audible release in coda voiced stops, even after they are spliced off, can be judged to be similar to epenthetic schwa.

Table 7: The percentages of forms that were judged to be most similar to the target items with coda voiced stops.

	ab	ad	ag	itab	ikad	itag	average
devoicing	27.8%	17.4%	8.88%	22.3%	23.1%	19.9%	19.9%
nasalization	8.8%	14.6%	12.1%	13%	13.9%	8.8%	11.9%
deletion	22.7%	20.2%	24.8%	25.1%	23.6%	9.2%	23.3%
epenthesis	40.7%	47.9%	54.2%	39.5%	39.4%	48.1%	45%

The challenge to the P-map hypothesis is then, if a language releases final voiced consonants, why don't speakers of that language repair coda voiced stop with an epenthetic schwa? Recall that the P-map hypothesis aims to explain the *universal* lack of phonological strategies other than devoicing to resolve a constraint against coda voiced stops. If there is *some* phonetic implementation of coda voiced stops which can be judged to be similar to the epenthesized form, then why can't that form provide a basis to induce phonological epenthesis?[12]

8 Summary and conclusion

8.1 Summary

The first four orthography-based experiments show that English speakers find the devoiced outcome as most similar to the original forms compared to the outcomes of other phonological forms. This finding supports the premise of Steriade (2001/2008). The phonological strategy that speakers chose as yielding the most similar form to the original form is actually the phonological strategy observed in natural languages. On the other hand, the final auditory experiment with an epenthetic schwa shows that schwa-epenthesis was judged to be the most similar form, when the target consonants are heavily released (even after the releases themselves are removed). This result raises a challenge to the P-map hypothesis; to the extent that forms with a schwa epenthesis are most similar to the original forms, then the P-map hypothesis predicts that languages can resort to schwa epenthesis, if the knowledge of similarity is based on released coda voiced stops.

[12] Jeremy Perkins (p.c) suggested that the P-map hypothesis be restricted to comparisons at the level of features (and not the level of segments). This amendment does solve this problem, but it at the same time significantly narrows down the original scope of the P-map hypothesis. This revision to the original P-map hypothesis, though consistent with the results of Experiment VI, does not explain any longer why languages do not use epenthesis to rescue coda voiced stops.

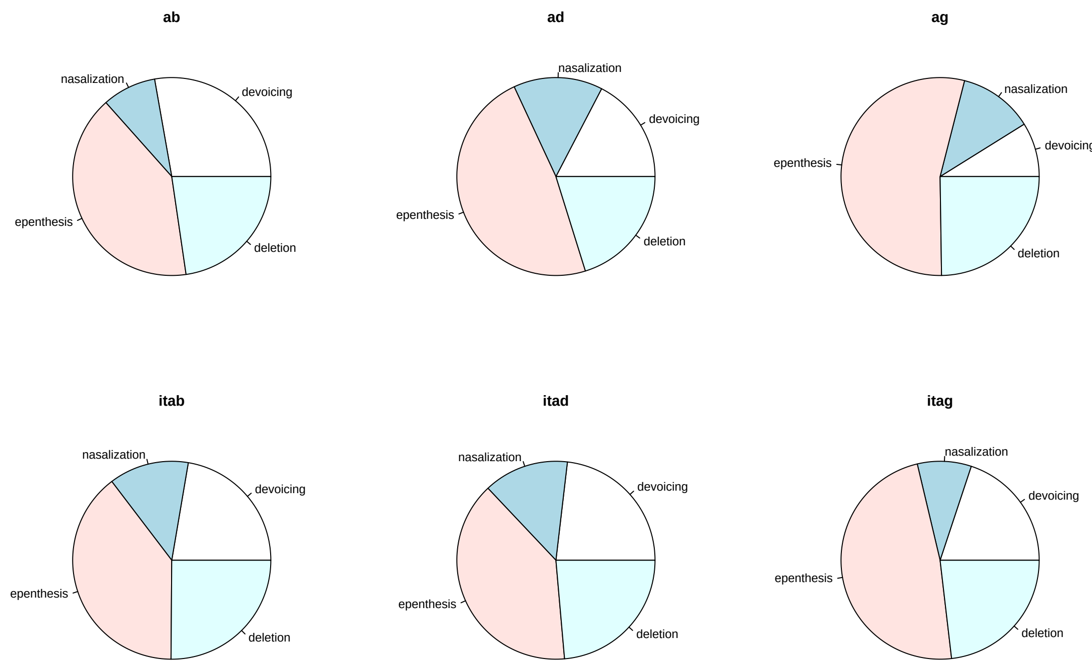

Figure 10: The percentages of forms that were chosen to be most similar in the second sound-based similarity judgment task.

8.2 Discussion and topics for future experimentation

The result of Experiment VI in and of itself does not immediately falsify the premise of the P-map hypothesis because speakers may not canonically release word-final voiced stops. Myers (to appear) has shown that English utterance-final coda stops are semi-devoiced, and such sounds are readily confused as their voiceless counterparts. In our current experiments, the first four experiments presented the stimuli in isolation, whereas the last two experiments presented stimuli which were originally uttered sentence-internally. In Experiment VI, the final stops were fully released. It could be that coda devoicing has its roots in the utterance-final semi-devoicing. Indeed this utterance-final semi-devoicing could be the cause of the special status of devoicing we observed in Experiment I-IV: since such voiced stops are already half-way to becoming voiceless as they are pronounced in isolation in the listeners' head, devoicing is judged to be the form that is most similar to the original form. On the other hand, the stimuli in Experiment V and VI were recorded in sentence-internal positions and were not semi-devoiced. Therefore, devoicing was not judged to be so similar to the original unaltered forms. To address this hypothesis, one possible follow-up experiment would be to use tokens that contain utterance-final coda stops for auditory similarity judgment experiments.

Another project that is left for future research is whether the similarity judgment patterns observed in the current experiments—especially those found in Experiment I-V)—hold for speakers of other languages. To the extent that the P-map hypothesis attempts to explain the universal lack of non-devoicing strategies, the same similarity judgment patterns should hold in other languages. In particular, it would be interesting to investigate languages that lack coda consonants altogether. Such a language provides a testing ground to investigate similarity patterns in coda positions without an influence of their language background. On this note, Broselow et al. (1998) show that Mandarin speakers spontaneously show coda devoicing when acquiring a second language, and it would be interesting to investigate whether this coda devoicing has its root in their P-map knowledge.

Finally, similarity measures can and should be investigated through other experimental paradigms, such as identification experiments under noise and discrimination experiments. While all of these experiments are necessary to further test the foundation of the P-map hypothesis, they are left as topics for future investigation, both due to time and space limitation.

To summarize, our orthography-based experiments generally support the premise of the P-map hypothesis; English speakers do find devoiced forms to be most similar to the original, unaltered form with coda voiced stops. Two auditory experiments show that similarity judgment patterns crucially depend on phonetic implementation patterns. Therefore, a question remains how language learners acquire the similarity knowledge underlying phonological patterns. Future investigation should address this issue.

References

Adler, Allison (2006) Faithfulness and perception in loanword adaptation: A case study from Hawaiian. *Lingua* **116**: 1021–1045.

Bailey, Todd & Ulrike Hahn (2005) Phoneme similarity and confusability. *Journal of Memory and Language* **52**: 339–362.

Benkí, José (2003) Analysis of English nonsense syllable recognition in noise. *Phonetica* **60**: 129–157.

Boersma, Paul & David Weenink (1999-2010) Praat: Doing phonetics by computer. A software.

Broselow, Ellen, Su-I. Chen, & Chilin Wang (1998) The emergence of the unmarked in second language phonology. *Studies in Second Language Acquisition* **20**: 261–280.

Carpenter, Angela (2002) Noncontiguous metathesis and adjacency. In *University of Massachusetts Occasional Papers in Linguistics 26: Papers in Optimality Theory II*, Angela Carpenter, Andries Coetzee, & Paul de Lacy, eds., Amherst: GLSA, 1–26.

Coetzee, Andries (2008) Grammaticality and ungrammaticality in phonology. *Language* **84**(2): 218–257.

Côté, Marie-Hélène (2004) Syntagmatic distinctness in consonant deletion. *Phonology* **21**: 1–41.

Dresher, Elan (1981) On the learnability of abstract phonology. In *The Logical Problem of Language Acquisition*, C. Baker & John J. McCarthy, eds., Cambridge, MA: MIT Press, 188–209.

Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, & Jacques Mehler (1999) Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* **25**: 1568–1578.

Fairbanks, Grant, Arthur House, & Eugene Stevens (1950) An experimental study of vowel intensities. *Journal of the Acoustical Society of America* **22**: 457–459.

Greenberg, Joseph & James Jenkins (1964) Studies in the psychological correlates of the sound system of American English. *Word* **20**: 157–177.

Hammond, Michael (1999) *The phonology of English: A prosodic Optimality-Theoretic approach*. New York: Oxford University Press.

Hanson, Kristin (2003) Formal variation in the rhymes of Robert Pinsky's the Inferno of Dante. *Language and Literature* **12**: 309–337.

Hayes, Bruce, Kie Zuraw, Péter Siptár, & Zsuzsa Londe (2009) Natural and unnatural constraints in Hungarian vowel harmony. *Language* **85**(4): 822–863.

Holt, Lori L., Andrew J. Lotto, & Randy Diehl (2004) Auditory discontinuities interact with categorization: Implications for speech perception. *Journal of the Acoustical Society of America* **116**: 1763 – 1773.

Howe, Darin & Douglas Pulleyblank (2004) Harmonic scales as faithfulness. *Canadian Journal of Linguistics* **49**: 1–49.

Hume, Elizabeth (2001) Metathesis: formal and functional considerations. In *Surface syllable structure and segment sequencing*, Elizabeth Hume, Norval Smith, & Jeroen van de Weijer, eds., Leiden: HIL, 1–25.

Jose, Brian & Julie Auger (2004) (Final) nasalization as an alternative to (final) devoicing: The case of Vimeu Picard. In *Indiana Working Papers Online*, Jose Brian & Ken De Jong, eds., Bloomington: Inidiana University Linguistics Club.

Jun, Jongho (2004) Place assimilation. In *Phonetically based Phonology*, B. Hayes, R. Kirchner, & D. Steriade, eds., Cambridge: Cambridge University Press., 58–86.

Kaplan, Abby (2010) *Phonology shaped by phonetics: The case of intervocalic lenition*. Doctoral dissertation, University of California at Santa Cruz.

Kawahara, Shigeto (2006) A faithfulness ranking projected from a perceptibility scale: The case of voicing in Japanese. *Language* **82**(3): 536–574.

Kawahara, Shigeto (2007) Half-rhymes in Japanese rap lyrics and knowledge of similarity. *Journal* *of East Asian Linguistics* **16**(2): 113–144.

Kawahara, Shigeto (2008) Phonetic naturalness and unnaturalness in Japanese loanword phonology. *Journal of East Asian Linguistics* **18**(4): 317–330.

Kawahara, Shigeto & Kazuko Shinohara (2009) The role of psychoacoustic similarity in Japanese puns: A corpus study. *Journal of Linguistics* **45**(1): 111–138.

Kenstowicz, Michael (2003) Salience and similarity in loanword adaptation: A case study from Fijian, ms. MIT.

Kingston, John (2005) From ears to categories: New arguments for autonomy. In *Prosodies: With Special Reference to Iberian Languages*, Sónia Frota, Marina Vigário, & Maria João Freitas, eds., Berlin: Mouton de Gruyter, 177–222.

Kingston, John & Randy Diehl (1994) Phonetic knowledge. *Language* **70**: 419–454.

Kingston, John, Shigeto Kawahara, Della Chambless, Daniel Mash, & Eve Brenner-Alsop (2009) Contextual effects on the perception of duration. *Journal of Phonetics* **37**(3): 297–320.

Kwon, Bo-Young (2005) The patterns of vowel insertion in IL phonology: The P-map account. *Studies in Phonetics, Phonology and Morphology* **11**(2): 21–49.

Lagerquist, Linnes (1980) Linguistic evidence from paronomasia. In *Proceedings of Chicago Linguistic Society 16*, Jody Kreiman & Almerindo E. Ojeda, eds., Chicago: CLS, 185–191.

Lehiste, Ilse (1970) *Suprasegmentals*. Cambridge: MIT Press.

Lindblom, Björn (1968) Vowel duration and a model of lip mandible coordination. *Speech Transmission Laboratory–Quarterly Progress and Status Report. Royal Institute of Technology, Stockholm* **4**: 1–29.

Lisker, Leigh (1986) ”Voicing” in English: A catalog of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* **29**: 3–11.

Lombardi, Linda (2001) Why place and voice are different: Constraint-specific alternations in Optimality Theory. In *Segmental Phonology in Optimality Theory*, L. Lombardi, ed., Cambridge, UK: Cambridge University Press, 13–45.

Mann, Virginia A. (1986) Distinguishing universal language-specific factors in speech perception: Evidence from Japanese listeners’ perception of /l/and /r/. *Cognition* **24**: 169–196.

Massaro, Dominic & Michael Cohen (1983) Phonological context in speech perception. *Perception & Psychophysics* **34**: 338–348.

McCarthy, John J. (2002) *A Thematic Guide to Optimality Theory*. Cambridge: Cambridge University Press.

McCarthy, John J. (2006) Restraint of analysis. In *Wondering at the Natural Fecundity of Things: Essays in Honor of Alan Prince*, Eric Bakovic, Junko Ito, & John J. McCarthy, eds., Santa Cruz: UCSC Linguistic Research Center, 213–239.

McCarthy, John J. (2008) *Doing Optimality Theory*. Oxford: Blackwell-Wiley.

McCarthy, John J. (2010) Perceptually grounded faithfulness in Harmonic Serialism, (Ms., University of Massachusetts, Amherst).

McCarthy, John J. & Alan Prince (1986) Prosodic morphology, ms. University of Massachusetts and Rutgers University.

Miller, George & Patricia Nicely (1955) An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* **27**: 338–352.

Moreton, Elliott (2002) Structural constraints in the perception of English stop-sonorant clusters. *Cognition* **84**: 55–71.

Myers, Scott (to appear) Final devoicing: Production and perception. In *Prosody Matters: Essays in honor of Elisabeth Selkirk*, Toni Borowsky, Shigeto Kawahara, Takahito Shinya, & Mariko Sugahara, eds., London: Equinox.

Parker, Steve (2002) *Quantifying the Sonority Hierarchy*. Doctoral dissertation, University of Massachusetts, Amherst.

Pater, Joe (1999) Austronesian nasal substitution and other NC effects. In *The Prosody-Morphology Interface*, Rene Kager, Harry van der Hulst, & Wim Zonneveld, eds., Cambridge: Cambridge University Press, 310–343.

Peters, Robert (1963) Dimensions of perception for consonants. *Journal of the Acoustical Society of America* **35**: 1985–1989.

Peterson, Gordon & Ilse Lehiste (1960) Duration of syllable neuclei in English. *Journal of the Acoustical Society of America* **32**: 693–703.

Pitt, Mark (1998) Phonological processes and the perception of illegal consonant clusters. *Perception & Psychophysics* **60**: 941–951.

Port, Robert & John Dalby (1982) Consonant/vowel ratio a s a cue for voicing in English. *Perception & Psychophysics* **32**: 141–152.

Postal, Paul (1968) *Aspects of Phonological Theory*. New York: Harper and Row.

Prince, Alan & Paul Smolensky (1993/2004) *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden and Oxford: Blackwell.

R Development Core Team (1993-2010) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reips, Ulf-Dietrich (2002) Standards for internet-based experimenting. *Experimental psychology* **49**(4): 243–256.

Shepard, Roger (1972) Psychological representation of speech sounds. In *Human Communication: A Unified View*, Edward E. David & Peter B. Denes, eds., New York: McGraw-Hill, 67–113.

Shinohara, Shigeko (2004) A note on the Japanese pun, dajare: Two sources of phonological similarity, ms. Laboratoire de Psycheologie Experimentale.

Silverman, Daniel (to appear) Schwa. In *The Blackwell companion to phonology*, Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, & Keren Rice, eds., Oxford: Blackwell-Wiley.

Steriade, Donca (2001a) Directional asymmetries in place assimilation: A perceptual account. In *The Role of Speech Perception in Phonology*, Elizabeth Hume & Keith Johnson, eds., New York: Academic Press, 219–250.

Steriade, Donca (2001b) What to expect from a phonological analysis. Talk presented at LSA 77, Washington, D.C.

Steriade, Donca (2001/2008) The phonology of perceptibility effects: The P-map and its consequences for constraint organization. In *The nature of the word*, Kristin Hanson & Sharon Inkelas, eds., Cambridge: MIT Press, 151–179, originally circulated in 2001.

Steriade, Donca (2003) Knowledge of similarity and narrow lexical override. In *Proceedings of the 29th annual meeting of the Berkeley Linguistics Society*, Pawel M. Nowak, Corey Yoquelet, & David Mortensen, eds., Berkeley: BLS, 583–598.

Umeda, Noriko (1975) Vowel duration in American English. *Journal of the Acoustical Society of America* **58**(2): 434–445.

Walden, Brian & Allen Montgomery (1975) Dimensions of consonant perception in normal and hearing-impaired listeners. *Journal of Speech and Hearing Research* **18**: 445–455.

Wilson, Colin (2006) Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* **30**(5): 945–982.

Yip, Moira (1987) English vowel epenthesis. *Natural Language and Linguistic Theory* **5**(4): 463–484.

Zuraw, Kie (2007) The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog infixation. *Language* **83**(2): 277–316.

Zwicky, Arnold (1976) This rock-and-roll has got to stop: Junior's head is hard as a rock. In *Proceedings of Chicago Linguistic Society 12*, S. Mufwene, C. Walker, & S. Steever, eds., Chicago: CLS, 676–697.

Zwicky, Arnold & Elizabeth Zwicky (1986) Imperfect puns, markedness, and phonological similarity: With fronds like these, who needs anemones? *Folia Linguistica* **20**: 493–503.