

Testing the P-Map: Lenition and Position

Chad Hall

Michigan State University

1 Introduction

P-Map is a theory originally proposed by Steriade (2001) that is concerned with projecting the correspondence constraints involved in Optimality Theory (Prince & Smolensky 1993). Steriade claims that the perceptibility of pairs of sounds depends on the context the sounds are in, and that the knowledge of these relative perceptibilities in different contexts (the “P-Map”) is used to obtain a hierarchy of correspondence constraints; these ranked correspondence constraints are used to explain phonological typology. For example, devoicing is common pre-consonantly and word-finally, but not in pre-vocalic environments. To explain this, she claims, based on some premises of the distribution of auditory cues, that the difference between [p] and [b] is perceived better before vowels than consonants or word-finally (e.g. [apa]/[aba] vs. [apta]/[abta]). However, Steriade provides little direct evidence for her claim, as there is little experimental work testing the relative perceptibility of contrasts in different phonological contexts. For the modest number of papers that do test the relative perceptibility, there is no perceptual testing of the same contrast in different phonological contexts: only various contrasts in the same position.¹

This paper seeks to give a more comprehensive testing of the context-sensitive relative similarity of contrasts that Steriade (2001) presumes. The limited prior experimental work that exists has supported the claim that the relative similarity of different contrasts parallels attested typological tendencies of phonological patterns. But no previous study has tested whether these relative similarity rankings hold across distinct phonological contexts: a key claim of Steriade’s. The particular phonological process in focus is lenition in the form of spirantization and devoicing.²

Using a similarity rating task, this paper will test the perceptual similarity of voicing and continuancy contrast pairs in intervocalic and word-final positions, exploring if ratings of similarity change with phonological context in a way that reflects attested phonological processes. The perceptibility of contrast related to devoicing and spirantization has been tested in specific phonological contexts (Kaplan 2010; Kawahara & Garvey, 2010). Both papers yielded results supporting Steriade’s claim that the relative perceptibilities of contrasts parallel the cross-linguistic typology of phonological patterns. However, there have been no experiments directly testing if the perception of these contrasts changes in a context-sensitive way. In this paper, I will show that the similarity judgements of voicing and continuancy contrasts vary depending on context, and furthermore that the voicing contrast is perceived as more similar intervocalically while the continuancy contrast is perceived as more similar word-finally. This mirrors the phonological typology for word-final devoicing and intervocalic spirantization, supporting Steriade’s notion that context plays a role in explaining the preference for certain phonological processes in specific contexts.

1.1 *The initial motivation for the P-Map* The main intuition behind Steriade’s proposal is that, typically, any phonological departure from the underlying form is minimal in order to achieve compliance

*I would like to thank Suzanne Wagner, Yen-Hwei Lin, Laura Dille, the audience at AMP 2017 and the Phonology-Phonetics group at Michigan State University for insightful comments and questions. I would like to give special thanks to Karthik Durvasula for his valuable assistance and advice in all aspects of this work. Thank you to Silvina Bongiovanni for helping in the creation of the stimuli, and thanks to Monica Nesbitt, Matt Savage and Alex Mason for their help in running the experiments.

¹ Of course, the claim that there is a link between phonological typology and perceptual asymmetries existed before the P-Map (Lindblom 1986; Ohala 1981, 1993). However, I will discuss this claim from the perspective of the P-Map.

² There is evidence to suggest that certain forms of devoicing constitute fortition (Iverson & Salmons 2007), however there are cases of devoicing which can fairly be described as lenition (see e.g. discussions of final devoicing as lenition in Honeybone 2012). In this paper, devoicing will be referred to as a form of lenition.

with phonotactic restrictions. This notion is formalized as follows: Form A is phonologically closer to form B than form C if the A-B distinction is more similar than the C-B distinction. The knowledge of this relative similarity is referred to as the P-Map.

The P-Map was originally motivated by a problem within Optimality Theory (OT). In Classic OT in which correspondence and markedness constraints interact freely, the predicted amount of typological variation, related to common markedness constraints, is greater than has been actually observed. For example, across languages, the most common output form of /tæb/, apart from [tæb], is [tæp].³ From the perspective of OT, this alternation between [tæb] and [tæp] is due to the interaction between the markedness constraint *[+VOICE]/_]_word (voiced obstruents do not occur at the end of a word) and the correspondence constraint Ident-IO [±voice] (corresponding input and output sounds should have identical values for the feature [voice]). Classic OT predicts that there should be more possible outputs such as [tæw] due to a freedom of interaction between various correspondence constraints with *[+VOICE]/_]_word. These forms are, however, rarely (if ever) attested, suggesting that the interaction between markedness and correspondence constraints is not free. This has been called the “too-many-solutions” problem.

Steriade suggested that the reason the most common output form of /tæb/ is [tæp] and not the other logically possible options is that the [tæb-tæp] pair is the most perceptually similar of all possible outputs. She proposed the P-Map as a collective term for (i) the speaker’s knowledge of perceptual similarity and (ii) the mechanism that relates the rankings of correspondence constraints to perceived differences in degrees of similarity. For example, if the P-Map identifies the pair [b]-[p] to be more perceptually similar after a vowel than [b]-[w] then the effect of P-Map on the grammar is to rank higher the correspondence constraint relative to the distinctive contrast [b]-[m]. Hence, IDENT-IO [±nasal] / V_] >> IDENT-IO [±voice] / V_]. As a consequence of this ranking, maintaining the contrast between nasal-oral, pre-vocalically, is more important than maintaining the contrast between voiced and voiceless. This in turn implies that a change in voicing is a less egregious change to remove or “repair” the markedness violation (*[+VOICE]/_]_word) than a change in nasality. In essence, the analysis makes the prediction that, in order to satisfy a high-ranked constraint that blocks voiced obstruents in a word-final position (*[+VOICE]/_]_word), changes from voiced to voiceless (devoicing) are going to be much more common than changes from oral to nasal (de-nasalization), as is observed in the attested typology of phonological patterns related to final voiced obstruents.

1.2 The unanswered questions of the P-Map The P-Map has been invoked to explain a wide range of phonological phenomena (Adler 2006; Côté 2004; Jun 2004; Kawahara 2006, 2007; Kawahara & Shinohara 2009; Kenstowicz 2007; Kwon 2005; McCarthy 2011; Steriade 2001a, 2003; Wilson 2006; Zuraw 2007). Whilst the theory of P-Map seems to be a sound approach in accounting for attested and unattested phonological patterns across languages, as well as potentially solving the too-many-solutions problem of OT, there are important questions that are yet to be answered.

The first major issue is that despite P-Map being a perceptually grounded theory, little effort has been made to run perceptual experiments to test the P-Map. Instead, it has been assumed that relevant perceptual similarity scales parallel the cross-linguistic phonological patterns, without clear experimental evidence that justifies the presence of such perceptual scales for the different contrasts. The second issue, tied into the first problem, is that testing this supposed universal knowledge of perception is difficult, since language-specific phonetic and phonological knowledge affects speech perception (Dupoux et al. 1999; Massaro & Cohen 1983; Moreton 2002; Pitt 1998), though Kawahara & Garvey (2010) do point out that there can be universal perceptual patterns, sometimes observed in the perception patterns of non-speech sounds. Nevertheless, the following question remains: how can we test a universal scale of perceptual similarity when speech perception becomes language-specific very quickly?

In Section 2, I provide a critical review of two rare papers in the sense that they report on perceptual experiments in order to test the P-Map: Kawahara and Garvey’s (2010) study of coda devoicing, and Kaplan’s (2010) study of intervocalic lenition, also noting problems in their methodologies.⁴ Then, after a brief

³ Other rarer (non-phonetically proximate) attested outputs such as [tæm] in particular languages are typically due to more than one historical language change (Merrill 2015). Capturing such alternations synchronically is an issue for the P-Map; however, I will set aside such patterns in the current paper.

⁴ The second issue raised above is rather difficult to address, and I will address only briefly in the final discussion of my paper.

discussion of spirantization and devoicing, I will present the results of a new set of experiments that seek to better proceed in providing evidence for the P-Map and show that continuancy contrasts are perceived as more similar intervocalically than word-finally while voicing contrasts are perceived as more similar word-finally than intervocalically. This supports the typology that devoicing is preferred more in word-final position than intervocalic while spirantization is preferred more in intervocalic position than word-final position.

2 Background: P-Map Experimental Work

2.1 Kawahara & Garvey (2010) Kawahara and Garvey (2010) look at the phenomenon Steriade originally used to explain her P-Map theory: coda devoicing. They ran four orthography-based experiments and two auditory similarity judgement experiments to test the preference across languages for coda devoicing (/tab/ → [tap]) rather than other input-output pairs such as /tab/ → [tam] or /tab/ → [ta].

The first experiment was an orthography-based forced-choice similarity judgement task run online on English speakers. Participants were initially presented with an orthographic form that contained a coda voiced stop (e.g. [ab]) and then other orthographic forms which represented devoicing ([ap]), nasalization ([am]), deletion ([a]) and epenthesis ([apa]). They were then asked to choose the form that was most similar to the original form. The participants overwhelmingly preferred the devoiced form to the others. Experiment two was virtually the same orthography-based experiment but the choice given to the participants was binary instead of multiple-choice. For example, instead of the participant being given a choice of four options like in the first experiment, they were presented with just two choices such as [ap] and [am]. The results further confirmed the findings of experiment one while also suggesting that epenthesis may be the second most similar form to the original form.

The third and fourth experiments were orthography-based similarity rating tasks where speakers were asked to judge the differences between pairs of the same stimuli from experiments one and two on a 5-point scale. The results again supported the P-Map's assumption that devoicing yields a form that is most perceptually similar to the form with a coda voiced stop. However, since the results were based on orthographic representation only, two further experiments were run in order to see whether the results extended to auditory perception.

Therefore, the fifth and sixth experiments were auditory similarity judgement tasks. In the fifth experiment, the coda voiced stop target stimuli were [ab], [ad], [ag], [itab], [ikad] and [itag] and for each main stimulus, Kawahara and Garvey prepared devoiced forms, nasalized forms, forms with deletion and forms with a final [i]. The stimuli were created by recording two native English speakers from New Jersey. Participants were then presented with the coda voiced stop, followed by the four alternative options, all in auditory format, and were asked to decide which alternative was the most similar to the coda voiced stop. Though devoicing was preferred, the skew was less extreme than in the orthography-based experiments. In the sixth experiment, the vowel-final candidate had a schwa instead of [i]. Additionally, in order to prevent reduction of word-final consonants, speakers were encouraged to release all the word-final consonants when creating the stimuli. The releases were then spliced off post-recording. The results showed that the vowel-final candidate was now chosen by listeners as the most similar to the original form.

Kawahara and Garvey find mixed results in testing P-Map. They conclude from their experiments that there is good support for the P-Map hypothesis since devoicing was seen as the most appropriate repair for final voiced stops in five out of six experiments. But they also challenge the theory, in the sense that their Experiment VI suggested that final schwa insertion is the most perceptually similar to a final voiced stop, which contradicts the fact that languages prefer to devoice final voiced stops rather than add a final schwa.

There are some issues with Kawahara and Garvey's paper in terms of methodology and conclusions drawn from the experiments. First, four of the six experiments were orthography-based. Their aim was to test the P-Map, and therefore perceptual awareness, but it is unclear how orthographic judgements are related to perceptual (auditory) similarity. The participants had to recall their experience or knowledge related to that sound instead of being presented with the stimulus itself, which does not seem appropriate for testing perceptual similarity. Second, for the sixth experiment, speakers were forced to produce sounds that were unnatural based on their native language, such as releasing coda voiced stops, which were then clipped. This was done because Kawahara and Garvey found in a pilot study that when the stimuli contained a heavily released final voiced stop, it was often confused with the vowel-final option. Regardless, the stimuli produced

in the sixth experiment may not have reflected natural language, which could have played a part in producing unexpected results.

Third, although a range of possible surface mappings are tested for coda final voiced stops, a major point of the P-Map is that the context of the phonological process is vital in perceptual similarity. Yet all of the mappings are tested in word-final/coda position. For example, devoicing should have been tested not just word-finally but in word-medial positions too.⁵ Only when this is done will we get a more comprehensive view of the accuracy of P-Map. Finally, it would have been useful to look at other common phonological phenomena apart from the one that Steriade referenced most frequently when initially proposing P-Map.

The results of Kawahara and Garvey's study are encouraging enough in the sense that devoicing appears to be the most preferred solution to final coda voiced stops as the typology suggests. However, there is certainly room for improvement in terms of testing the P-Map. The next paper that will be discussed does not have some of the methodological issues as Kawahara and Garvey's paper though it too does not investigate the perception of contrasts across contexts.

2.2 Kaplan (2010) The second paper of interest is Abby Kaplan's dissertation (Kaplan 2010), which tests the P-Map on the phenomenon of intervocalic lenition. Kaplan highlights how for intervocalic stops, certain phonological processes are preferred to others. For example, for intervocalic voiced stops, the typology shows that spirantization occurs (e.g., /d/ → [z]) but devoicing is an unnatural repair (e.g., /d/ → [t]). For intervocalic voiceless stops on the other hand, spirantization (e.g. /t/ → [s]) is attested and so is voicing (e.g., /t/ → [d]) (Grijzenhout 2000; Gurevich 2004; Kirchner 1998; Lavoie 2001). Kaplan suggests that the P-Map can explain this pattern.

Kaplan's first experiment was designed to test the relative perceptibility of the two potential repairs for intervocalic voiced stops—devoicing and spirantization—with the latter occurring far more often in natural language while devoicing is rarely attested (Beguš 2016). She compared voiced stops at each of the three major places of articulation ([b], [d], [g]) with regard to their mutual confusability with their voiceless counterparts ([p], [t], [k]) and fricative counterparts ([β], [ð], [ɣ]). The stimuli for the experiment consisted of each of the nine consonants recorded in the environment [a__a]. The tokens were recorded by two native Spanish speakers and three linguistically-trained native English speakers. Certain recorded stimuli were left out of the final experiment: for the voiceless stops recorded by the English speakers, the stimuli selected had the shortest period of aspiration, since the aspiration contrast makes English voiced and voiceless stops easier to discriminate than the voicing contrast in a language that does not use aspiration to the same extent. For the voiced fricatives, all tokens from both the English and Spanish speakers were rated for naturalness by two native Spanish speakers. This was important especially for the tokens from the English speakers since [aβa] and [aɣa] are not phonotactically legal in English. For the voiced stop stimuli, the tokens chosen were closest to the voiced fricative and voiceless stop stimuli with respect to the F1 and F2 of the vowels on either side of the consonant. This was to ensure that sensitivity to relevant pairs was grounded solely in differences among the consonants, not differences among the vowels. Twenty-four native speakers of English participated in the experiment: a same-different task where subjects were presented with a pair of tokens and asked to indicate whether they had heard two different words or two repetitions of the same word. The stimulus presentation was purely auditory. The results show that listeners are less sensitive to the distinction between voiced stops and fricatives intervocalically than they are to the distinction between voiced stops and voiceless stops intervocalically, explaining the preference for spirantization over devoicing of intervocalic voiced stops in natural language.

Kaplan ran a second experiment focussed on the two attested phonological repairs for intervocalic voiceless stops, voicing and spirantization. The structure of the experiment was the same as that of the previous experiment, except different stimuli were used: the voiceless stops in the three main places of articulation ([p], [t], [k]), their voiced counterparts ([b], [d], [g]) and their fricative counterparts ([f], [θ], [x]). Three talkers were recorded for the creation of the stimuli: A native Bulgarian speaker, a native German speaker and a native English speaker. 24 native speakers of English participated in the experiment. The results

⁵Note that I did not mention testing word-initial position since lenition is uncommon in any form word-initially (Béres & Honeybone 2012).

of the experiment show that listeners did not show a large difference in perceptibility between voiced and voiceless stops compared to voiceless stops and fricatives, intervocalically. This finding is also compatible with the P-Map: Voiceless stops may undergo lenition either by voicing or by spirantization because neither alternation is less perceptible than the other.

Unlike Kawahara and Garvey's experiments, all of Kaplan's experiments were auditory stimuli, not orthographic, focussing more on the perceptual aspect of the task. Kaplan also reduced potential phonetic implementation issues that Kawahara and Garvey faced, not by forcing unnatural speech and then modifying it, but simply by leaving out stimuli not deemed natural enough for the experiment.

There is, however, a key issue still not resolved by Kaplan with regards to testing P-Map: While she appears to give a strong account that the P-Map can explain what is occurring in terms of intervocalic lenition across languages, a major aspect of the P-Map is that context is significant. While spirantization for stops is common intervocalically, it is a process that is less common in word-final position (Kirchner 1998). Inversely, while devoicing is rarely attested intervocalically, it is common word-finally. In order to get stronger evidence in favor of the P-Map, one must show not just why spirantization occurs frequently intervocalically, but why it occurs less frequently word-finally, with devoicing favoured in that context instead. If context is significant as Steriade suggests, then spirantization in word-final position must be more perceptible than intervocalic spirantization. Furthermore, devoicing in intervocalic position must be more perceptible than word-final devoicing.

The present study tests these hypotheses using experimental methods that build on but also improve upon those used by Kawahara & Garvey (2010) and Kaplan (2010). The experiment is described in Sections 3 and 4.

3 Aim and Hypotheses

The aim of this paper is to test the P-Map hypothesis that attested phonological patterns vary depending on phonological context due to differences in the perceptual similarity of two phonological contrasts in different contexts. Specifically, the phenomena in focus will be spirantization and devoicing. Kaplan (2010) showed that the attested lenition processes for voiced stops in intervocalic position appear to be related to perception. This paper will focus on the same processes as Kaplan, but it will not only investigate whether the attested patterns for voiced stops in intervocalic position are correlated with perceptual distance, but it will also investigate whether the attested patterns for voiced stops in word-final position are correlated with perceptual distance too.

As discussed already, spirantization of voiced stops is very common intervocalically while devoicing is very uncommon. In contrast, devoicing of voiced stops is very common word-finally while spirantization is less common. According to the P-Map, this is because the perceptual similarity between voiced stops and their fricative counterparts is more between two vowels than at the end of a word, and inversely, the perceptual similarity between voiced stops and their voiceless counterparts is more word-finally than between two vowels. This is putatively what accounts for these attested patterns.

With this in mind, if the P-Map is correct, then in accordance with the cross-linguistically common patterns, I hypothesize that:

- 1) A continuancy contrast will be rated as more similar word-finally than intervocalically.
- 2) A voicing contrast will be a more dissimilar intervocalically than word-finally.

If these hypotheses are confirmed, we will have stronger evidence in favour of the P-Map than the two studies described in Section 2 were able to provide.

To address the aim of this paper, a similarity judgement task with a within-subject design was run across two conditions. Condition I of the experiment was similar to Kaplan's experiments, in that it tested the perception of lenition intervocalically. Condition II tested the perception of these same contrasts word-finally. Native English speakers were presented with pairs of stimuli. Some of these were pairs of words with a voiced stop-voiced fricative contrast while others were pairs of words with a voiced stop-voiceless stop contrast. All listeners were presented with the same pairs – In Condition I, the contrast was intervocalic and in Condition II, the contrast was word-final. Listeners were asked to rank how similar these pairs were. The words were nonsense words to prevent any lexical bias, and both the intervocalic and word-final stimuli were spliced

from the same recordings to prevent any chance of external factors outside of the consonant contrast affecting perception.

As per the P-Map, a pair of words that show a contrast in continuancy should be perceived as more similar when the contrast is intervocalic than word-final. Additionally, a pair that shows a contrast in voicing should be perceived as more similar when the contrast is word-final than intervocalic. This is found to be the case in this study.

4 Method

4.1 Participants Nineteen participants took part in this experiment.⁶ Seven were males and 12 were females. They were all native Michigan English speakers aged between 19-21 years old. All experiments were run in the Phonetics and Phonology Laboratory at Michigan State University. Participants were compensated with extra credit in a language-related course.

4.2 Stimuli For this experiment, the participants listened to pairs of stimuli that differed only in one consonant. The difference was between either a voiced-voiceless stop pair, or a voiced stop-voiced fricative pair at three places of articulation, bilabial, alveolar and velar. Thus, the differences in these tokens were between the consonants [b]-[p], [d]-[t], [g]-[k], [b]-[β], [d]-[ð] and [g]-[ɣ], the same as the consonants used by Kaplan (2010). In order for these consonant pairs to be in either an intervocalic environment or a word-final environment, a phonetically-trained Spanish speaker was recorded producing the token [əzʊ ə] where each consonant mentioned above was placed between the [ʊ] and the [ə]. This particular environment was chosen since it involves vowels that do not vary much in pronunciation and are less common in American English, reducing the chances of lexical interference. In addition, the reason that the stress was put on the vowel preceding the differing consonant, [ʊ], was to make the phonological environment for lenition as strong as possible. In other words, it was to create an environment where lenition was most likely to occur in natural language (Bérces & Honeybone 2012). Recordings were made using a Samson Meteor Mic, which is a USB condenser microphone with a flat response frequency of 20Hz-20kHz and a 16-bit, 44.1/48kHz resolution. For each token, several recordings were made at a sampling rate of 44.1kHz and converted into .wav files on Audacity (Audacity Team 2017) with the best three versions of each token used for the experiment.

Once the best recordings were chosen, each token was spliced in two different ways on Audacity. To create bisyllabic intervocalic tokens, the initial schwa was deleted to give [zʊ ə] tokens; and to create bisyllabic word-final tokens, the final schwa was deleted from the same recordings to give [əzʊ] tokens. This is a novel method, not one used by Kaplan (2010) or Kawahara and Garvey (2010). A silence of 50ms was added to the start and end of each token using Praat (Boersma & Weenink 2019). The above procedure was followed to create three unique recordings each of intervocalic stimuli ([zʊbə], [zʊpə], [zʊβə], [zʊdə], [zʊtə], [zʊðə], [zʊgə], [zʊkə], [zʊɣə]) and word-final stimuli ([əzʊb], [əzʊp], [əzʊβ], [əzʊd], [əzʊt], [əzʊð], [əzʊg], [əzʊk], [əzʊɣ]). Then, in order to make a sound file containing a pair of these words, two separate recordings were combined into one single audio file on Audacity. Since the focus of this experiment was to test the similarity of consonants contrasting in voicing or continuancy in intervocalic and word-final contexts, three versions of the following 12 pairs were made: [zʊbə]-[zʊpə], [əzʊb]-[əzʊp], [zʊbə]-[zʊβə], [əzʊb]-[əzʊβ], [zʊdə]-[zʊtə], [əzʊd]-[əzʊt], [zʊdə]-[zʊðə], [əzʊd]-[əzʊð], [zʊgə]-[zʊkə], [əzʊg]-[əzʊk], [zʊgə]-[zʊɣə] and [əzʊg]-[əzʊɣ]. In total, 36 sound files were created and used for the experiment. Table 1 summarizes the place of articulation and (POA), the phonological contrast and phonological context represented by each pair.

⁶ Originally I had 25 participants but six were omitted due to either excessively long reaction times throughout the task or because they gave a very small range of scores during the task.

Contrast	POA	Intervocalic	Word-final
Voicing	Bilabial	[zóbə]-[zópə]	[əzób]-[əzóp]
	Alveolar	[zódə]-[zótə]	[əzúd]-[əzót]
	Velar	[zógə]-[zókə]	[əzóg]-[əzók]
Continuancy	Bilabial	[zóbə]-[zóbə]	[əzób]-[əzób]
	Alveolar	[zódə]-[zódə]	[əzúd]-[əzúd]
	Velar	[zógə]-[zógə]	[əzóg]-[əzóg]

Table 1: A Summary of the POA, phonological contrast and phonological context represented by each pair of stimuli.

4.3 Procedure Each participant was seated in a sound-attenuated booth in front of a keyboard and computer with which they would complete the task. Participants wore headphones. The headphones were the Plantronics Audio .355, a 3.5mm analog stereo headset with a 40mm diameter speaker driver size and a speaker frequency response of 20Hz-20kHz.

The experiment itself involved a similarity judgement task presented on the program PsychoPy (Peirce 2007). The aim of the task was to rate the similarity of pairs of tokens on a scale of 1-7, with 7 meaning very similar/identical and 1 meaning very dissimilar. Participants were told to respond to each stimulus as quickly and as accurately as possible. Each participant gave a similarity score for each pair they heard. After a single pair was heard, a scale between 1-7 would appear on the screen and the participant would use a mouse to click the number on a visual scale to give their answer. There were no visual cues to the stimuli, i.e. no orthographic forms. Participants were allowed to take a short break between blocks.

The experiment was presented in four blocks. The first two blocks contained the intervocalic voicing and continuancy pairs for Condition I ([zóbə]-[zópə], [zóbə]-[zóbə], [zódə]-[zótə], [zódə]-[zódə], [zógə]-[zókə] and [zógə]-[zógə]). The three versions of each pair were played twice in each block in random order, meaning there were 36 responses per block (6x3x2). The first check was a practice trial for the participant to familiarize themselves with the task and the second block was the main trial. Blocks three and four contained the word-final voicing and continuancy pairs for Condition II ([əzób]-[əzóp], [əzób]-[əzób], [əzúd]-[əzót], [əzúd]-[əzúd], [əzóg]-[əzók] and [əzóg]-[əzóg]). Again, the three versions of each pair were played twice in each block in random order. The third block was a practice trail for familiarization with the word-final stimuli and the final block was the main trial.

4.4 Analysis The responses and reaction times of each participant were recorded and saved to a .csv file. The .csv files for the individual participants were then concatenated into one file using R (R Development Core Team 2008). Responses in which the reaction time exceeded three seconds were removed from the data as these were not deemed fast enough. Then the mean average score from each participant for each pair was calculated. Outliers from individual mean scores were removed from the data on SPSS (IBM Corp 2017). Values that are 1.5xIQR below Quartile 1 and above Quartile 3 were deemed as outliers by SPSS. A paired samples t-test and Wilcoxon test were then run, comparing the scores between the intervocalic and word-final pairs of the same phonological contrast to see if there was a significant difference in the perceived similarity of the same phonological process in different contexts. A 3-way ANOVA was also run to check if there was an interaction effect between POA (alveolar vs. bilabial vs. velar), Contrast (voicing vs. continuancy) and Position (word-medial vs. word-final). The significance threshold was set at .05. The pairs whose results were compared were [zóbə]-[zópə] vs. [əzób]-[əzóp], [zódə]-[zótə] vs. [əzúd]-[əzót], [zógə]-[zókə] vs. [əzóg]-[əzók], [zóbə]-[zóbə] vs. [əzób]-[əzób], [zódə]-[zódə] vs. [əzúd]-[əzúd], [zógə]-[zógə] vs. [əzóg]-[əzóg]. The results are as follows.

5 Results

Table 2 shows a summary of the overall mean scores for each pair of tokens from the experiment. Figures 1 and 2 summarize the same results in a cluster bar chart with error bars included representing one standard error.

Contrast	POA	Intervocalic	Word-final
Voicing	Bilabial	3.33	6.14
	Alveolar	3.38	3.41
	Velar	3.06	4.45
Continuancy	Bilabial	4.81	4.13
	Alveolar	5.16	5.11
	Velar	5.76	3.51

Table 2: A Summary of the mean scores for each pair of tokens, on scale from 1 (least similar) to 7 (most similar).

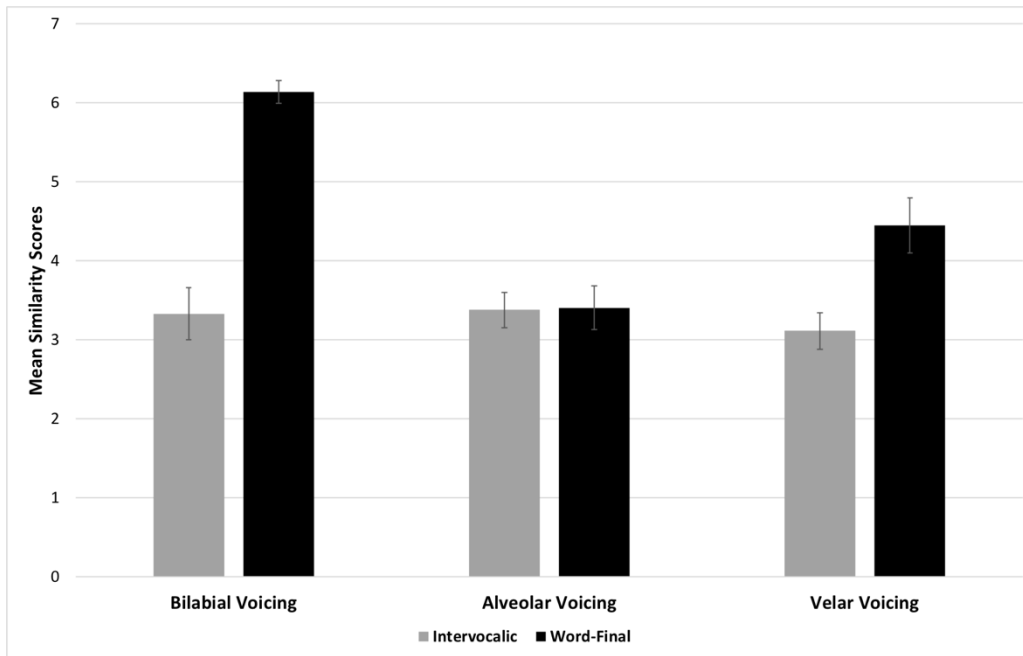


Figure 1: A summary of the mean scores for each pair of voicing tokens, on scale from 1 (least similar) to 7 (most similar).

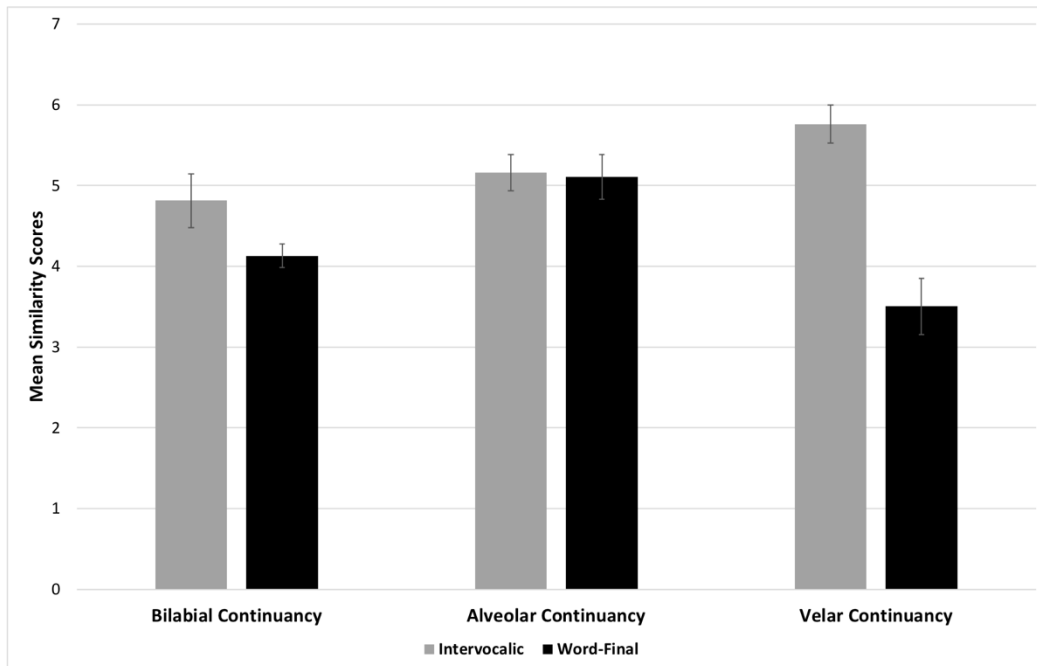


Figure 2: A summary of the mean scores for each pair of continuity tokens, on a scale from 1 (least similar) to 7 (most similar).

The ANOVA results crucially show that there was a statistically significant three-way interaction between POA, Contrast and Position, $F(2, 207) = 17.481$, $p < .001$. This suggests that the ratings for the different pairs varied by POA, Contrast and Position. It is important to bear in mind that the main predictions stemming from the P-Map that are relevant for the current paper are pair-wise comparisons contrasting position (word-final vs. word-medial). In what follows, I include a visual presentation of the results along with a discussion of the crucial pairwise comparisons.

Going from left to right in Figure 1, the results for voicing show that for the bilabial voicing contrast, the mean score for word-final position (6.14) is higher than for intervocalic position (3.33). This difference is statistically significant ($\text{Mean}_{\text{diff}} = -2.81$, $t_{17} = -7.78$, $p < .001$, BCa 95% CI [-2.04, -3.57]).⁷ A Wilcoxon test also shows that there was a significant difference between the mean scores ($Z = -3.62$, $p < .001$).⁸ For the alveolar voicing contrast, the mean score for intervocalic position (3.38) and word-final position (3.41) were not statistically different from each other ($\text{Mean}_{\text{diff}} = -0.03$, $t_{19} = -0.09$, $p > .05$, BCa 95% CI [-0.65, 0.60]). In fact, the scores were almost identical. The Wilcoxon test also shows that this difference in mean scores was not statistically significant ($Z = -0.04$, $p > .05$). For the velar voicing contrast, the mean score for word-final position (4.45) is higher than for intervocalic position (3.06). This difference is statistically significant ($\text{Mean}_{\text{diff}} = -1.39$, $t_{18} = -3.44$, $p = .003$, BCa 95% CI [-2.25, -0.54]). The Wilcoxon test also shows statistical significance ($Z = -2.77$, $p = .006$).

Turning to the continuity results in Figure 2, for the bilabial continuity contrast, the mean score for intervocalic position (4.81) is higher than word-final position (4.13). This difference is statistically significant ($\text{Mean}_{\text{diff}} = 0.68$, $t_{16} = 2.97$, $p = .009$, BCa 95% CI [0.19, 1.17]). The Wilcoxon test also shows this difference is significant ($Z = -2.44$, $p = .015$). For the alveolar continuity contrast, the mean score for intervocalic position (5.16) and word-final position (5.11) were not statistically different from each other ($\text{Mean}_{\text{diff}} = 0.05$, $t_{18} = 0.18$, $p > .05$, BCa 95% CI [-0.60, 0.71]). Again, the scores were almost identical. The Wilcoxon test also shows no statistical significance ($Z = -0.20$, $p > .05$). Finally, for the velar continuity contrast, the mean score for intervocalic position (5.76) was higher than word-final position (3.51). This difference is

⁷ BCa stands for Bias-Corrected and Accelerated. CI stands for Confidence Interval.

⁸ I also present Wilcoxon tests in this paper; t-tests are not completely appropriate for likert scale ratings, as they violate the normality of errors assumption.

statistically significant ($\text{Mean}_{\text{diff}} = 2.25$, $t_{18} = 6.34$, $p < .001$, BCa 95% CI [3.01, 6.34]). The Wilcoxon test also shows the difference between the two mean scores is statistically significant ($Z = -3.67$, $p < .001$).

To sum up, word-final voicing contrasts were perceived as more similar than intervocalic contrasts. The only exception was the alveolar stimuli, where similarity scores across contexts were not statistically different. For the pairs that involved a continuancy contrast, intervocalic contrasts were perceived as more similar than word-final contrasts. Again, the only exception was the alveolar stimuli. Broadly though, the results of this experiment match the perceptual-distance hypotheses stemming from the cross-linguistic tendencies related to devoicing and spirantization patterns. These results will now be discussed in further detail.

6 Discussion

The aim of this experiment was to test a particular presumed aspect of the P-Map that had not yet been empirically supported: that attested phonological contrasts vary depending on context due to differences in the perceptual similarity of two phonological contrasts in different contexts. The experiment yields fascinating results that generally favour the P-Map. They show that for bilabial and velar consonants, the contrast of voicing is perceived as significantly more similar word-finally than intervocalically. This supports why devoicing is a process that is typologically more strongly preferred in word-final position than intervocalically. Furthermore, for the bilabial and velar consonants, the results show that the contrast of continuancy is perceived as significantly more similar intervocalically than word-finally. Again, this supports why spirantization is a process that is more strongly preferred in the intervocalic environment than word-final. These results show evidence for a part of the P-Map not yet explored, that context does in fact affect the perceptibility of a phonological contrast, explaining the natural language typology.

There is however one set of results that do not, on the surface, support this notion. This is the results for the alveolar contrasts ([d]-[t] and [d]-[ð]). Although the results are in the expected direction in that the voicing contrast is perceived as more similar word-finally and the continuancy contrast is perceived as more similar intervocalically, the difference is not close to statistical significance, nor is it of the same magnitude as the other differences. As a matter of fact, for both contrasts, the mean scores are almost identical. In other words, the voicing contrast for alveolars was perceived as equally similar in intervocalic and word-final position, and so was the continuancy contrast. This does not support attested language systems universally or Steriade's original claims. There is nothing in the literature or typology to suggest that the alveolar consonants should act differently to the bilabial and velar consonants in terms of the perception of phonological contrasts.

There is however a potential explanation for these unexpected results for the alveolar consonants. Although the Spanish speaker who recorded the stimuli was phonetically trained, it may have been difficult in particular to produce a natural alveolar [d] and [t] sound since in Spanish, these consonants are dental, not alveolar (Hualde 2005). Consequently, it may have been difficult to produce natural sounding alveolar consonants which may have affected the results. This could be addressed in future work by recording a speaker who is capable of naturally producing all of the stimuli in the correct contexts. Furthermore, the speaker might have introduced additional acoustic cues in the pronunciations, as Kawahara and Garvey (2010) speculate for their experiment as well.

7 Summary

The P-Map has been argued to be effective in accounting for phonological phenomena across languages. However, there is little empirical data testing the P-Map and, up to now, no experiments testing one of the key claims of the P-Map: that attested phonological patterns vary depending on phonological context due to differences in the perceptual similarity of two phonological contrasts in different contexts. The experiment in this paper was conducted to test this notion by testing the perceptual similarity related to spirantization and devoicing in intervocalic and word-final contexts.

Overall, despite the irregular outcomes for the alveolar stimuli, the results support the P-Map's claim. Voicing was shown to be a more perceptible contrast intervocalically than word-finally, explaining the preference for word-final devoicing over intervocalic devoicing universally, and spirantization was shown to be more a perceptible contrast word-finally than intervocalically, explaining the preference for intervocalic

spirantization over word-final devoicing across languages. The results are significant for the field of Phonology, arguing in favour of a phonetics-based approach to the field. The fact that the perception of contrasts changes depending on context, and that these changes seem to relate directly to attested phonological phenomena, shows that Phonology should look towards Phonetics more frequently as a basis of explanation for phonological patterns. Additionally, the study supports the P-Map as a credible theory for explaining variation in phonological phenomena.

Of course, there are still more questions about the P-Map that need to be answered that lie beyond the scope of this particular study. For example, could these results be replicated by native speakers of a different language? The P-Map is argued to be universal knowledge of relative perceptibility. For this reason, the current results need to be replicated with speakers of other languages. However, it is worth noting that these results are particularly striking given that this experiment was run on native Michigan English speakers and neither devoicing nor spirantization are processes that happen in Michigan English. Therefore, there could not have been any bias in the similarity judgements of these pairs. This lends weight to the idea that these results reflect the universal perception of these phonological processes, indicating a preference for intervocalic spirantization and word-final devoicing in natural language and supporting the notion that there is a universal P-Map shared by all humans.

References

- Adler, A. (2006). Faithfulness and perception in loanword adaptation: A case study from Hawaiian. *Lingua* 116, 1021-1045.
- Audacity Team (2017). *Audacity [Computer program]*. Version 2.1.3, retrieved 8 Mar 2017 from <http://www.audacityteam.org/>.
- Beguš, G. (2016). *Post-nasal Devoicing and a Probabilistic Model of Phonological Typology*. Ms., Harvard University.
- Béres, K. B. & Honeybone, P. (2012). Splitting ‘intervocalic’: Expanding the typology of lenition environments. *Acta Linguistica Hungarica* 59, 27-48.
- Boersma, P. & Weenink, D. (2019). *Praat: doing phonetics by computer [Computer program]*. Version 6.0.2, retrieved 24 May 2017 from <http://www.praat.org/>.
- Côté, M. (2004). Syntagmatic distinctness in consonant deletion. *Phonology* 21, 1-41.
- Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C. & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* 25, 1568-1578.
- Grijzenhout, J. (2000). Voicing and devoicing in English, German and Dutch; evidence for domain-specific identity constraints. *Theory des Lexikons; Arbeiten des Sonderforschungsbereichs* 282: 1-22.
- Gurevich, N. (2004). *Lenition and Contrast: The Functional Consequences of Certain Phonologically Conditioned Sound Changes*. Outstanding Dissertations in Linguistics. Garland Publishing, New York.
- Honeybone, P. (2012). Lenition in English. In T. Nevalainen and E. Traugott (eds.) *The Oxford Handbook of the History of English*. Oxford University Press. Oxford.
- Hualde, J. I. (2005). *The Sounds of Spanish*. Cambridge University Press. New York.
- IBM Corp (2017). *IBM SPSS Statistics for Windows [Computer program]*. Version 25.0, retrieved 2 Apr 2019 from <https://www.ibm.com/us-en/marketplace/spss-statistics/>.
- Iverson, G. K. & Salmons, J. C. (2007). Domains and directionality in the evolution of German final fortition. *Phonology* 24:1, 121-145
- Jun, J. (2004). Place assimilation. In B. Hayes, R. Kirchner and D. Steriade (eds.) *Phonetically Based Phonology*. Cambridge University Press. Cambridge. 58-86.
- Kaplan, A. (2010). *Phonology Shaped by Phonetics: The Case of Intervocalic Lenition*. Doctoral dissertation, University of California at Santa Cruz, USA.
- Kawahara, S. (2006). A faithfulness ranking projected from a perceptibility scale: The case of voicing in Japanese. *Language* 82:3, 536-574.
- Kawahara, S. (2007). Half-rhymes in Japanese rap lyrics and knowledge of similarity. *Journal of East Asian Linguistics* 16:2, 113-144.
- Kawahara, S. & Garvey, K. (2010). Testing the P-map hypothesis: Coda devoicing. *Rutgers Optimality Archive* 1087.
- Kawahara, S. & Shinohara, K. (2009). The role of psychoacoustic similarity in Japanese puns: A corpus study. *Journal of Linguistics* 45:1, 111-138.
- Kenstowicz, M. (2007). Salience and similarity in loanword adaptation: A case study from Fijian. *Language Sciences* 29:2-3, 316-340.
- Kirchner, R. (1998). *An Effort-based Account of Consonant Lenition*. Doctoral dissertation, University of California at Los Angeles, USA.
- Kwon, B. (2005). The patterns of vowel insertion in IL phonology: The P-map account. *Studies in Phonetics, Phonology and Morphology* 11:2, 21-49.

- Lavoie, L. (2001). *Consonant Strength: Phonological Patterns and Phonetic Manifestations*. Outstanding Dissertations in Linguistics. Garland Publishing. New York.
- Massaro, D. & Cohen, M. (1983). Phonological context in speech perception. *Perception & Psychophysics* 34, 338-348.
- McCarthy, J. (2011). Perceptually grounded faithfulness in Harmonic Serialism. *Linguistic Inquiry* 42:1, 171-183.
- Merrill, J. (2015). Nasalization as a repair for voiced obstruent codas in Noon. In *LSA Annual Meeting Extended Abstracts* (Vol. 6, pp. 14-1).
- Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition* 84, 55-71.
- Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods* 162:1-2, 8-13.
- Pitt, M. (1998). Phonological processes and the perception of illegal consonant clusters. *Perception & Psychophysics* 60, 941-951.
- Prince, A. & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell. Malden and Oxford.
- R Development Core Team (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Steriade, D. (2001). The phonology of perceptibility effects: The P-map and its consequences for constraint organization. In K. Hanson and S. Inkelas (eds.) *The Nature of the Word*. MIT Press. 151-179.
- Steriade, D. (2001a). Directional asymmetries in place assimilation: A perceptual account. In E. Hume & K. Johnson (eds.) *The Role of Speech Perception in Phonology*. Academic Press. New York. 219-250.
- Steriade, D. (2003). Knowledge of similarity and narrow lexical override. In P. Nowak, C. Yoquelet and D. Mortensen (eds.) *Proceedings of the 29th Annual Meeting of the Berkeley Linguistics Society*. BLS. Berkeley. 583-598.
- Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30:5, 945-982.
- Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog infixation. *Language* 83:2, 277-316.