Proc Ling Soc Amer Volume 1, Article 24:1–15
doi:http://dx.doi.org/10.3765/plsa.v1i0.3725

Abstraction of phonological representations in adult nonnative speakers

Alia Lancaster & Kira Gor*

Abstract. Perception of nonnative contrasts by adult second language (L2) learners is affected by native language phonology. The current study contrasted predictions from two models of L2 phonological acquisition that focus on different representational levels as the origin of native language transfer: the abstract categorization level from the Perceptual Assimilation Model for L2 learners (PAM-L2; Best & Tyler, 2007) and the phonetic level from the Automatic Selective Perception model (ASP; Strange, 2011). The target phonemes were pairs of Arabic consonants that were equally similar on the abstract categorization level but unequally similar on the phonetic level—voiced and voiceless pharyngeal fricatives /ʕ/, /ħ/ and uvular fricatives /χ/, /ʁ/. Twenty intermediate-level English-speaking Arabic L2 learners and 10 Arabic native speakers (NS) completed auditory identification and discrimination tasks. We first conducted a discriminant analysis (DA) to quantify ASP predictions based on phonetic variables. L2 learners were generally more accurate when perceiving the pharyngeal consonants compared to the uvulars and when perceiving the voiced phonemes compared to the voiceless. These findings, and L2 learners' perceptual variation across contexts, predicted by the DA, suggest that L2 speakers were able to track phonetic cues during L2 perception and thus favor the ASP. These results support the interpretation that L2 learners attend to the phonetic detail in nonnative segments; however, they do not build nativelike phonological representations for the segments with weaker phonetic cues. This ability to process low-level phonetic cues opens the possibility for learners to create more robust phonological representations.

Keywords. L2 acquisition; phonological contrasts; perception; nonnative phonology; phonetic cues

1. Introduction. Phonological representations of L2 speech segments are different from the first language (L1) in several ways. The most salient for the purposes of this study is that L2 representations are inexorably linked to L1 phonological representations. In adult L2 learners, the target population in the current study, L1 may interfere at the phonetic as well as at the phonological level. Therefore, it is not surprising that there are multiple theories and models of L2 phonological acquisition which differ in the relative importance of the levels of representation. The phonetic level of representation encodes information to categorize phonemes. Nonnative processing at the level of phonetic cue identification has implications for representations and processing at the level of phonological categorization. A breakdown at this level translates to an inability to interpret, categorize, or reproduce a sound segment, such as the inability to discriminate two different speech segments in isolation (e.g., Lukyanchenko & Gor, 2011). Phonemes are used at the lexical level to encode words. Allophones are also abstractions that make it possible to interpret phonetic cues in relation to the context, i.e., a semi-voiceless consonant will be categorized as voiced word-finally because word-final voiced consonants are expected to be

* Authors: Alia Lancaster, University of Maryland (abiller@umd.edu) & Kira Gor, University of Maryland (kiragor@umd.edu)

devoiced to a certain degree. Allophones are not as abstract as phonemes, but are more abstract than information represented at the phonetic level. Therefore, allophones are represented at a post-phonetic or pre-phonological level.

The current study compares the predictions of two models to the perception of native speakers of English learning Arabic. The PAM-L2 (Best & Tyler, 2007) focuses on L2 categorization that involves the phonological level of representation, while the ASP (Strange, 2011) focuses on the processing of phonetic cues that involves the phonetic level of representation. These two models were chosen because they focus on different levels of representation and also generate clear and testable predictions for L2 learners' perception of nonnative segments.

According to the PAM-L2, segments in the L2, e.g., L2A and L2B, can be assimilated to L1 categories in four different ways, each of which has implications for L2 perceptual categorization of the segments. The first type of assimilation relevant for the current study, Two Category (TC), occurs when either L2A or L2B are assimilated to an L1 category. Relatively accurate discrimination results from this type of assimilation, and the phonetic parameters may or may not change depending on the goodness of fit to the existing L1 category. The second type of assimilation, Uncategorized, occurs when neither L2A nor L2B are assimilated to an L1 category. In this case, "one or two new L2 phonological categories may be relatively easy to learn perceptually," leading to relatively accurate L2 discrimination (Best & Tyler, 2007, p.28). The other two types of assimilation do not apply to the current study. Although the information L2 learners encode, according to this theory, is articulatory gestures, recent work investigating the predictions of the PAM-L2 on tones and vowels has led the authors to suggest that phonetic and/or acoustic information may be useful in determining assimilation types (e.g., Bundgaard-Nielsen, Best, & Tyler, 2011; Reid et al., 2015).

The ASP (Strange, 2011) model's scope is slightly different from the PAM-L2—it is meant to describe how adult L2 learners identify the speech of NS online and in the context of other segments and words. The ASP encoding unit is the selective perception routine (SPR), which is a learned set of relevant phonetic properties that best characterize a segment. SPRs are used automatically by NSs, and learning and automatizing the SPRs of an L2 is a learner's goal. L1 interference manifests as the use of an L1 SPR for an L2 speech segment, with gradual retuning occurring using bottom-up information.

Studies measuring acoustic characteristics of L1 and L2 have been conducted by Strange (Strange et al., 2004), and others (e.g., Escudero, Simon, & Mitterer, 2012) in order to determine the reliability of the ASP as a model. Some experiments focus on naïve learners, similar to the original PAM (Best, 1994), while others test L2 learners. Because the model focuses on the phonetic level, DA of acoustic information on L2 segments is used as a tool to predict identification results. While the DA predict the perceptual abilities of naïve listeners (Strange, Bohn, Nishi, & Trent, 2005; Strange et al., 2004), with some exceptions, this type of analysis strongly predicts the perceptual abilities of L2 listeners (Escudero et al., 2012; Gilichinskaya & Strange, 2010).

The research question is related to the predictions of the two models: does the source of difficulty in perception of nonnative segments for adult L2 speakers stem from nonnative representations at the phonetic or phonological level? Arabic phonology contains fricative consonants that are different from English consonants both phonemically and phonetically, which makes it a suitable language to investigate ASP and PAM-L2 predictions. The goal was to compare the predictions to perception of the voiceless uvular /χ/, voiced uvular /ʁ/, voiceless pharyngeal /ħ/, and voiced pharyngeal /ʕ/ by L2 learners of Arabic who are NSs of American

English. Perception was assessed in both identification and discrimination tasks. In the identification task, accurate perception was classified as correctly labeling a target phoneme within a nonce word (e.g., /χ/ in /uχu/). In the discrimination task, participants indicated if pairs of nonce words containing target phonemes were the same or different. Perception in this task was calculated using the number of correct and incorrect hits and misses to determining sensitivity to the consonants within a pair of nonce words (e.g., /uχu/-/uʁu/).

The PAM-L2 predicts equal, relatively accurate perception of both the uvulars and the pharyngeals. Neither of the uvular phonemes exist in L1 or are similar to an L1 category, suggesting they form an Uncategorized assimilation type. Regarding the pharyngeals, the Arabic (L2) voiceless pharyngeal, /ħ/, may assimilate to the English (L1) voiceless glottal, /h/. The Arabic voiced pharyngeal, /ʕ/, is often produced much more like an approximant than a fricative (Bin-Muqbil, 2006) and may assimilate to the English vowel /a/. The pharyngeals, therefore, form a TC assimilation situation. Recall that TC assimilation predicts accurate perception of the phonemes involved, as was demonstrated with Arabic phonemes by Tyler and Fenwick (2012). The PAM-L2 also predicts accurate perception of the uvulars, since, being Uncategorized, both would be free to create novel L2 categories.

The ASP model, on the other hand, predicts different perception patterns. While both the uvulars and the pharyngeals differ only in voicing, the voiced pharyngeal is often produced with no frication noise (Bin-Muqbil, 2006). Specifically, the lack of any frication noise in the voiced pharyngeal causes it to be distinct from the voiceless pharyngeal, which does display aperiodic noise (Bin-Muqbil, 2006; Ghazeli, 1997). Both uvulars have frication noise, making the distinction between the voiceless and voiced uvular less salient than the difference between the voiced and voiceless pharyngeals. The distinction between the pharyngeals entails not only a difference in voicing, but also a difference in sonority since approximants are more sonorous than fricatives and the voiced pharyngeal often displays phonetic characteristics of an approximant. Thus, the ASP predicts that the pharyngeals will be more accurately perceived than the uvulars, and, if there is any acoustic contextual variation, learners will be sensitive to it. The predictions of the ASP will be quantified by measuring the acoustic variables thought to influence perception in NS productions and submitting these measurements to a DA. The results of such an analysis have been shown to accurately predict L2 learners' perception of non-native vowels (Escudero et al., 2012; Gilichinskaya & Strange, 2010) in a manner that takes into account all the relevant acoustic variables and is context-dependent.

Thus far, these phonemes have only been discussed in terms of comparisons within a place of articulation—pharyngeals and uvulars. However, it is possible to create pairs that are the same in voicing but differ in place of articulation, which also tests the predictions of both models. The voiced set of phonemes—the voiced uvular and voiced pharyngeal—represent a TC pair for the PAM-L2 (L2 /ʕ/ is assimilated to L1 /a/), and a pair that is phonetically distant from one another for the ASP. Both models would predict that learners would be relatively accurate perceiving this pair of phonemes. The voiceless pair, containing the voiceless uvular and voiceless pharyngeal, represents a TC assimilation for the PAM-L2 (L2 /ħ/ is assimilated to L1 /h/), and a pair that is phonetically similar, and therefore, difficult to discriminate, for the ASP. The PAM-L2 predicts that responses to the voiceless pair would be relatively accurate, while the ASP predicts that the responses would be relatively less accurate compared to the voiced pair. Therefore, if learners follow the predictions of the ASP as quantified in the acoustic analysis, this will be evidence that learners represent L2 segments at the phonetic level. If learners follow the predictions of the PAM-L2, this will be evidence that learners represent L2 segments at the phonological level.

2. Acoustic analysis. Following the procedure in other studies investigating the ASP (e.g., Gilichinskaya & Strange, 2010), we preformed both within- and cross-language DA on productions by NSs of Arabic and English. The goal was to operationalize the predictions of the ASP by using acoustic measurements to simulate the perception process of an L2 learner if they were attending to phonetic cues. One strength of this method is that the predictions will be based on the same productions that were also used as stimuli in the behavioral tasks completed by L2 learners. If learners perform similarly to the analyses below, we can infer that they attend to the same cues.

2.1. PARTICIPANTS. Participants were recruited from a large state university. Sixteen participants completed the study, 10 NSs of English (6 male) and six NSs of Arabic (4 male). The data from one NS of English was discarded due to recording error, reducing the number to 9 native English-speaking participants. The average age of native English speakers was 24.1 (*SD*=8.82), and the average age of native Arabic speakers was 26.83 (*SD*=3.37).

2.2. MATERIALS. Two sets of nonce words were created, one for English and one for Arabic, with CVC and VCV structures. Target consonants were placed in three different contexts—the beginning of a nonce word (initial), between two vowels (medial), and the end of a nonce word (final). Vowels surrounding the consonants were either /ɑ/ or /u/, so chosen because these vowels are similar in both languages and are used in previous studies with speakers of both languages (Flege & Port, 1981; Tyler & Fenwick, 2012). However, due to large variability in the production of /ɑ/ observed in the data, only the results of nonce words containing /u/ are reported for both the acoustic analyses and the behavioral tasks.

The phonemes for English recordings were /t/, /d/, /s/, /z/, /h/, and the phonemes for Arabic recordings were /t/, /d/, /s/, /z/, /h/, /q/, /χ/, /ʁ/, /ħ/, /ʕ/. In nonce words containing the target phoneme in the initial or final position (i.e., CVC structure), a filler phoneme was used in the other consonant slot. The voiced bilabial stop /b/ was chosen because it exists in both languages.

2.3. PROCEDURE. Participants were given a printed list of the nonce words to record. Each list contained two instances of each nonce word in their native language, randomized for each participant. Participants recorded the stimuli in a soundproof booth with a headset and microphone using Praat software (Boersma & Weenink, 2014). Before recording, participants were instructed on vowel production and asked to produce each nonce word with a falling intonation. After recording, participants were allowed to play back each nonce word to determine if the production was appropriate. Three practice items were recorded with the experimenter in the booth. Consent form and procedures were approved by the university's Institutional Review Board.

2.4. RESULTS. Various measurements were documented using Praat software (Boersma & Weenink, 2014). Measurements for the non-filler consonants and all adjacent vowels were duration, mean intensity, mean fundamental frequency, mean first formant, mean second formant, mean third formant, and percent unvoiced (consonants only). Percent unvoiced represents the percentage of locally unvoiced pitch frames as determined by the voice report in Praat and is a continuous and more objective measure of voicing than a binary judgement. Mean formant measures were calculated by averaging formants at 25% duration, 50% duration, and 75% duration during a segment, and were subsequently transformed into the bark scale, which is psychoacoustically relevant given that the results of the analyses will be compared to learners' perception. All the measurements were obtained for each production of each nonce word by each participant. The data were then split by context (initial, medial, final) and speaker gender for separate analyses. For the acoustic analyses as well as the analysis of the behavioral tasks, a subset

of phonemes were examined: /s/, /z/, /h/ for English and /s/, /z/, /h/,/ʕ/ /χ/, /ʁ/, /ħ/ for Arabic. Stops were initially included in order to mask the study purpose from participants—namely, to examine fricative production.

A DA within each language was conducted, which determines variates, or dependent variable combinations, that best discriminate the consonants from one another (Field, Miles, & Field, 2012). The dependent variables for the within-language DA were all of the acoustic variables measured on the NS productions. Using the linear discriminant coefficients, or the weight given to each measurement variable in determining the consonant, the consonant predicted by the model was saved and compared to the original (i.e., when the original consonant was /s/, how often did the analysis predict an /s/?). The objective of the within-language DA was to evaluate the predictive ability of the linear discriminant coefficients. If the predictive ability is high (i.e., many phonemes are correctly classified), it is evidence that the acoustic variables measured are similar to those NSs use to discriminate phonemes. Rates of correct classification range from 63% to 100% (e.g., Strange et al., 2004). Averaged across the separate analyses by gender, classification rates of the Arabic within-language DA were all above 80% (initial: 89%, medial: 94.5%, final: 81%). Classification rates of the English within-language DA were all above 95% (initial: 100%, medial: 100%, final: 96.5%). The high percentage of correct classifications indicates that the acoustic variables entered into the models were sufficient to distinguish consonants within each language from one another.

DA also have the ability to predict the consonant given new acoustic measurements which were not used to create the original linear discriminant coefficients. In order to simulate the response behavior of L2 learners and quantify ASP predictions, discriminant function weights from the within-language English DA were used to predict consonants in the Arabic dataset based on Arabic measurement variables. This process is meant to mirror a native English speaker using L1 SPRs to classify L2 (Arabic) segments. Because separate analyses were again conducted for each context, the results were informative about the acoustic similarities and differences in consonants specific to each context. The aim in these analyses was to record classification patterns among the target consonants. Since English was used as the data to create the coefficients, the model was only able to predict consonants that existed within that data: /s/, /z/, or /h/. The most common classification is therefore a reflection of the English consonant that the model determined as closest, based on the acoustic variables entered, to the Arabic consonant.

As seen in Table 1, the target phonemes (/ʕ/, /χ/, /ʁ/, /ħ/) in the initial context except for the voiced pharyngeal were most often categorized by the models as /s/; the voiced pharyngeal's most common classification is split evenly between /h/ and /s/. An inference from this finding is that, for an Arabic learner with a native English background, the voiced pharyngeal is acoustically different from the other target phonemes. The same pattern is also seen in the final context. In the medial context, however, all target phonemes are most often categorized as /h/.

	Phoneme produced						
Context	/χ/	/ʁ/	/ħ/	/ʕ/	/s/	/z/	/h/
Initial	/s/ (58%)	/s/ (50%)	/s/ (58%)	/h/ (50%), /s/ (50%)	/s/ (50%)	/s/ (50%)	/h/ (54%)
Medial	/h/ (83%)	/h/ (83%)	/h/ (83%)	/h/ (83%)	/h/ (71%)	/h/ (54%)	/h/ (83%)
Final	/s/ (67%)	/s/ (42%)	/s/ (46%)	/h/ (50%)	/s/ (54%)	/s/ (75%)	/h/ (50%)

Table 1: Most common classification of the Arabic consonant productions by the between-language DA averaged across gender. Percentages represent percentage of observations that were classified as the consonants listed in each cell.

These results operationalize and specify the ASP predictions based on acoustic information from productions by NSs of Arabic and English. Most of the predictions based on previous literature were correct—namely, that the pharyngeals are more acoustically distinct from one another than the uvulars (and voiced more distinct than voiceless). The analyses provided more detailed predictions with regard to the context of the target phonemes. Pharyngeals appear to be more distinct from uvulars in the initial and final contexts, but not the medial. By the same token, the voiced phonemes appear to be more distinct from the voiceless in the initial and final contexts, but not the medial.

3. Behavioral tasks.

3.1. PARTICIPANTS. Participants were recruited from a large state university. Thirty-three participants completed the study: 21 L2 learners of Arabic (5 male), and 12 NSs of Arabic (7 male). The average age of L2 learners was 19.81 (*SD*=1.63), and the average age of NSs was 25.25 (*SD*=4.27).

3.2. MATERIALS. Recordings from one male and one female native Arabic speaker were used as the stimuli for the discrimination and identification tasks. The identification task consisted of four trials for each nonce word—two in the male voice and two in the female, with a of a total of 96 trials randomized for each participant in terms of voice, consonant, and context. During a discrimination task trial, participants heard two nonce words, one produced by a male native Arabic speaker and one produced by a female native Arabic speaker. Each nonce word was paired with every other nonce word in each context to create different pairs (e.g., /uʕu/-/uħu/). In order to balance response type (yes/no), an equal number of pairs with the same consonant were also presented (e.g., /uʕu/-/uʕu/). Order and voice gender were balanced, leading to a total of 336 pairs (112 per context). The number of stimuli was doubled to 672 pairs so that each participant would hear each different pair four times. Two lists were created in order to counterbalance gender of the voice. Within a list, there were an equal number of male-female and female-male voice pairs.

3.3. PROCEDURE. Participants completed the discrimination task, identification task, cloze test (if an L2 leaner), and a language history questionnaire in a fixed order. The entire experimental session lasted approximately two hours. All tasks occurred in an isolated room with a closed door on a personal computer, and those requiring listening used noise-cancelling headphones. Both the discrimination and identification task were presented using DMDX (Forster & Forster, 2003).

During the identification task, participants were instructed to press one of eight keys, which were labeled with the eight target Arabic letters in alphabetical order from right to left, to indicate the consonant they heard in the nonce word. As a filler, the consonant *b* was ignored and was not among the choices of Arabic letters labels. Each trial began with 300ms of silence followed by the nonce word. Participants were given three seconds in which to respond as quickly and as accurately as possible. The next trial began immediately after the feedback (correct/incorrect) to the response. After concluding a set of five practice trials, participants responded to nonce words in two blocks of 46 nonce words each with untimed breaks between each block. The blocks and items within the blocks were randomized for each participant. This task took approximately 20 minutes to complete.

During the discrimination task, participants were instructed to press the right shift key if the pair of syllables they heard were exactly the same, labeled *yes*, and to press the left shift key if the pair was not exactly the same, labeled *no*. Each trial began with 300ms of silence, followed by the first nonce word, 100 ms of silence, followed by the second nonce word. Participants were

given two seconds in which to respond as quickly and accurately as possible. The next trial began immediately after feedback (correct/incorrect) to the response. After concluding a set of five practice trials, participants responded to pairs in eight blocks of 79 pairs each with untimed breaks between each block. The blocks and items within the blocks were randomized for each participant. This task took approximately an hour to complete.

L2 participants also completed an Arabic cloze test, which consisted of five paragraphs of text with 36 blanks. Next to each blank were three words (or variants of one word) from which to choose. All participants completed an online language history questionnaire which asked about multiple facets of L2 learning and usage (Li, Zhang, Tsai, & Puls, 2013). Consent form and procedures were approved by the university's Institutional Review Board.

3.4. RESULTS. We utilized multilevel modeling due to its advantages over multiple regression or ANOVA methods (Linck & Cunnings, 2015). The multilevel models were conducted with the lme4 package version 1.1-12 (Bates, Maechler, Bolker, & Walker, 2016) in R version 3.2.5 (R Core Team, 2015). Before submitting the data to multilevel models, we trimmed the data to exclude responses that were 0ms in length and that were 3000ms for the identification task and 2000ms for the discrimination task. These exclusion criteria resulted in 1.93% of observations dropped for the identification task analysis and 1.33% for the discrimination task analyses. All models were run as forced entry models for fixed effects and subject random intercepts. Random slopes were tested one-by-one via likelihood ratio tests, and only random slopes that significantly improved model fit and resulted in converging models were retained (Baayen, Davidson, & Bates, 2008).

To examine the data with respect to the research question and the predictions of the two competing theories, we first conducted analyses with the L2 learners as the baseline group, which are reported in tables below. In order to aid interpretation of the other fixed effects with respect to the effects of speaker, the same models were run with NSs as the reference, the results of which are reported in-text. The cloze test was administered in order to account for variation in L2 learners due to proficiency in the analyses. However, scores did not differ from chance ($t(19)=1.15$, $p=.132$), indicating that it was too advanced for the learners. Accordingly, cloze scores did not significantly improve the model fit for any of the analyses reported, thus it was excluded. Self-reported proficiency on four skills (listening, speaking, reading, writing) gathered in the language history questionnaire was averaged for an alternative proficiency score. According to this measure, Arabic learners were on average of intermediate proficiency, ranging from 2 to 5.25 on a scale of 1 (very poor) to 7 (native-like; $M=3.72$, $SD=0.96$), while NSs rated their Arabic abilities higher, ranging from 3 to 7 ($M=5.72$, $SD=1.54$). However, due to missing data for 28% of participants for the self-report measure, it was not used in the analyses.

The logistic multilevel model for the identification task was run using the "bobyqua" optimizer. Item random intercepts were not appropriate for the identification task analysis because the term was very closely related to one of the fixed effects – target phoneme. The remaining variance in items was voice of the speaker (male, female) and was captured by including Voice random intercepts. Unfortunately, the model did not converge with the random effect of Voice, so it was dropped. The linear multilevel model for the discrimination task was fit using restricted maximum likelihood estimation. Item random intercepts were not appropriate for the discrimination task analyses because the dependent variable, d-prime, was calculated across items. Due to the unreliable nature of producing *p*-values using MCMC sampling for linear multilevel models, the output provides *t*-values, not *p*-values (Bates et al., 2016). Thus *t*-values were considered

marginal if greater than an absolute value of 1.65 and significant at the p <.05 level if greater than an absolute value of 2.00 (Gelman & Hill, 2007; Linck & Cunnings, 2015).

3.4.1 IDENTIFICATION TASK. The average accuracy for correctly identifying each phoneme is presented in Figure 1, separated by speaker and by context. Note that while mean accuracies are informative for descriptive purposes, the model outlined below examined not mean accuracy, but the probability of a correct or incorrect response on an item given the predictors in the model.

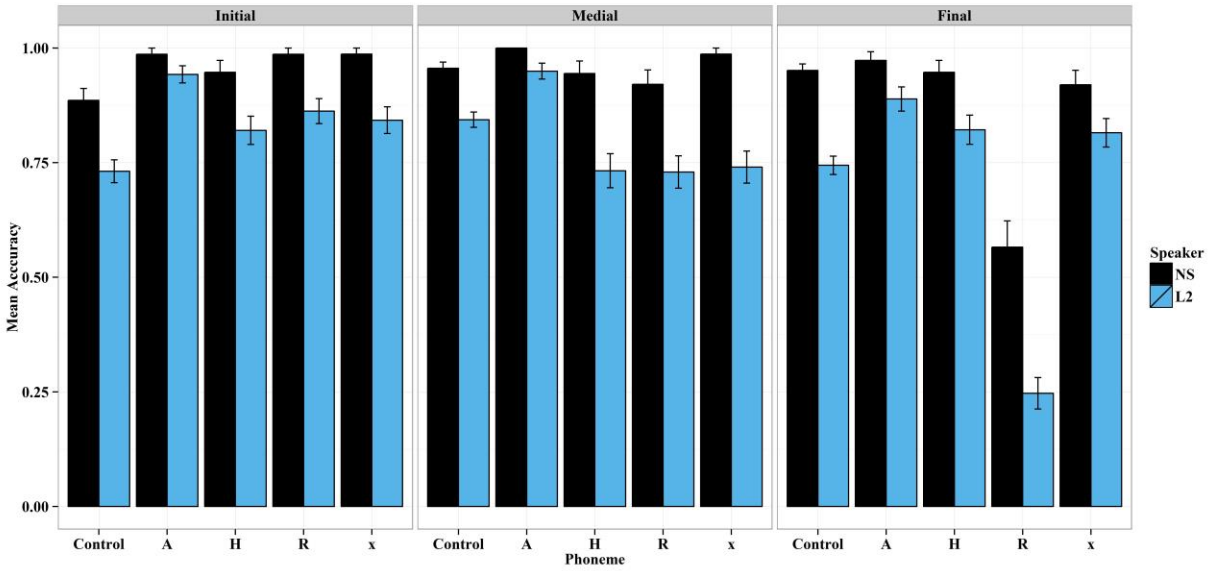

Figure 1: Accuracy results of identification task. Labels along the x-axis represent the target phonemes (A=voiced pharyngeal, H=voiceless pharyngeal, R=voiced uvular, x=voiceless uvular)

Accuracy results for the identification task were submitted to a logistic multilevel model. The dependent variable was accuracy in identifying the target phoneme in a nonce word (0, 1). Fixed effects included phoneme presented in a nonce word (voiced pharyngeal, voiceless pharyngeal, voiced uvular, voiceless uvular, or control), the phoneme context in a nonce word (initial, medial, or final), speaker (L2, native), and interactions between phoneme, context, and speaker. The voiced pharyngeal phoneme in the initial condition acted as the baseline, thus all effects in the model are interpreted with respect to this baseline. The voiced pharyngeal phoneme was chosen as the baseline instead of control phonemes because the model would more readily compare performance between it and other target phonemes, which is more directly relevant to the research question. Control phonemes (/s/, /z/, /h/) were grouped into one category for the purpose of analysis.

Table 2 presents the model of accuracy in the identification task. In the initial position (i.e., baseline), L2 speakers were significantly more likely to accurately identify the pharyngeal phoneme (/ʕub/) than all other target phonemes ($p_{\text{control}} < .001$, $p_{\text{voiceless pharyngeal}} < .001$, $p_{\text{voiceless uvular}}=.003$, $p_{\text{voiced uvular}}=.012$). This pattern applies to the other contexts (medial and final) given the remaining non-significant terms. The exception is for the voiced uvular phoneme; Arabic learners were significantly less likely to correctly label it than the voiced pharyngeal in the medial context (p=.050) and in the final context (p < .001).

When the same model was run with NSs as the baseline, NSs were not significantly different in their probability of correctly labeling any of the other target phonemes compared to the voiced

pharyngeal in any context. There are two exceptions. The first is that NSs were more likely to correctly identify the voiced pharyngeal than the control phonemes in the initial context (*b*=-2.39, *SE*=1.07, *p*=.026). Secondly, like learners, NSs displayed a drop in accuracy when identifying the voiced uvular in the final position. This shift from the pattern in the initial context (i.e., the voiced uvular is less often correctly identified compared to the voiced pharyngeal) was significant (*b*=-4.03, *SE*=1.62, *p*=.013).There was a significant three-way interaction between target phoneme, context, and speaker ($\chi^2(8)=14.11$, *p*=.079).

Fixed effects	***b***	***SE***	***z*-value**	***p*-value**
Intercept (voiced pharyngeal, initial context,L2)	3.27	0.44	7.35	<.001
Target phoneme: Control	-2.18	0.41	-5.32	<.001*
Voiceless pharyngeal	-1.62	0.45	-3.63	<.001*
Voiceless uvular	-1.36	0.45	-2.99	.003*
Voiced uvular	-1.16	0.46	-2.51	.012*
Context: Medial	0.23	0.56	0.42	.674
Final	-0.84	0.49	-1.72	.086^
Native speaker	1.78	1.17	1.52	.128
Target phoneme x context: Control x medial	0.60	0.59	1.02	.310
Voiceless pharyngeal x medial	-0.89	0.63	-1.40	.161
Voiceless uvular x medial	-1.00	0.64	-1.57	.116
Voiced uvular x medial	-1.26	0.64	-1.96	.050*
Control x final	0.92	0.52	1.77	.077^'
Voiceless pharyngeal x final	0.77	0.59	1.31	.191
Voiceless uvular x final	0.60	0.59	1.02	.310
Voiced uvular x final	-2.58	0.58	-4.42	<.001*
Target phoneme x speaker: Control x native	-0.21	1.14	-0.18	.855
Voiceless pharyngeal x native	0.18	1.25	0.15	.884
Voiceless uvular x native	1.42	1.53	0.93	.351
Voiced uvular x native	1.22	1.53	0.80	.424
Context x speaker: Medial x native	11.96	426.17	0.03	.978
Final x native	0.16	1.36	0.12	.906
Target phoneme x context x speaker: Control x medial x native	-11.68	426.18	-0.03	.978
Voiceless pharyngeal x medial x native	-11.46	426.18	-0.03	.979
Voiceless uvular x medial x native	-11.19	426.18	-0.03	.979
Voiced uvular x medial x native	-12.92	426.18	-0.03	.976
Control x final x native	0.77	1.44	0.54	.592
Voiceless pharyngeal x final x native	-0.09	1.59	-0.06	.956
Voiceless uvular x final x native	-1.93	1.80	-1.07	.284
Voiced uvular x final x native	-1.45	1.76	-0.82	.410
Random effects	**Variance**	***SD***		
Intercepts-Subject	.98	.99		

Note: *Significant at $p < .05$; ^Marginal at $p < .10$.

Table 2: Identification task results of logistic multilevel model for accuracy

When the same model was run with NSs as the baseline, NSs were not significantly different in their probability of correctly labeling any of the other target phonemes compared to the voiced

pharyngeal in any context. There are two exceptions. The first is that NSs were more likely to correctly identify the voiced pharyngeal than the control phonemes in the initial context (*b*=-2.39, *SE*=1.07, *p*=.026). Secondly, like learners, NSs displayed a drop in accuracy when identifying the voiced uvular in the final position. This shift from the pattern in the initial context (i.e., the voiced uvular is less often correctly identified compared to the voiced pharyngeal) was significant (*b*=-4.03, *SE*=1.62, *p*=.013).There was a significant three-way interaction between target phoneme, context, and speaker ($\chi^2(8)=14.11$, *p*=.079).

The voiced uvular in the final position is an interesting exception to the overall high accuracy seen in NSs. One advantage of an identification task is that it allows for investigation into the patterns of incorrectly labeled items. Arabic learners on average correctly labeled the voiced uvular in the final position for 27.9% of the items, and on average mislabeled the voiced uvular as the voiceless uvular for 57.9% of the items. NSs followed a similar pattern, to a lesser degree, on average correctly labeling the voiced uvular in the final position for 67.7% of the items and on average mislabeled the voiced uvular as the voiceless uvular for 34% of the items.

3.4.2 DISCRIMINATION TASK. D-prime scores were calculated based on accuracy for both same and different pairs of nonce words, which better reflect sensitivity to binary contrasts in discrimination tasks requiring yes/no answers. D-prime scores were calculated for each phoneme pair in each context for each participant, with higher d-prime indicating higher sensitivity to the difference between the phonemes in the nonce word pair presented (i.e., better discrimination ability). Separate analyses were conducted to compare d-prime scores for pairs that were the same in place of articulation (pharyngeal vs. uvular) and the same in voicing (voiced vs. voiceless). For instance, in the place of articulation analysis, the pharyngeal pairs were those that, regardless of order, contained the pharyngeal voiced fricative /ʕ/ and the pharyngeal voiceless fricative /ħ/ (i.e., initial: /ʕub/-/ħub/, medial: /uʕu/-/uħu/, and final: /buʕ/-/buħ/). In the voicing analyses, for example, the voiced pairs were those that, regardless of order, contained the uvular voiced fricative /ʁ/ and the pharyngeal voiced fricative /ʕ/. The control pairs used in the analyses were all pairs that were not the target. For instance, the d-prime scores entered in the place of articulation analysis as controls were those that were calculated from pairs that did not contain only pharyngeal or only uvulars (e.g., /usu/-/uzu/). The average d-prime scores for each comparison (place of articulation or voicing) can be seen in Figures 2 and 3.

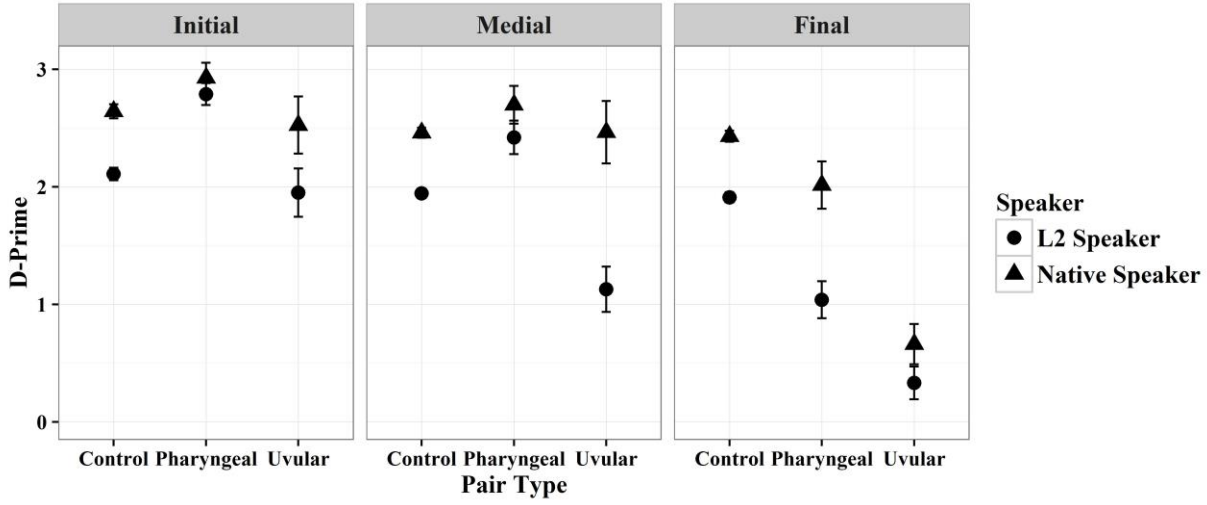

Figure 2: Discrimination d-prime for pairs similar in place of articulation

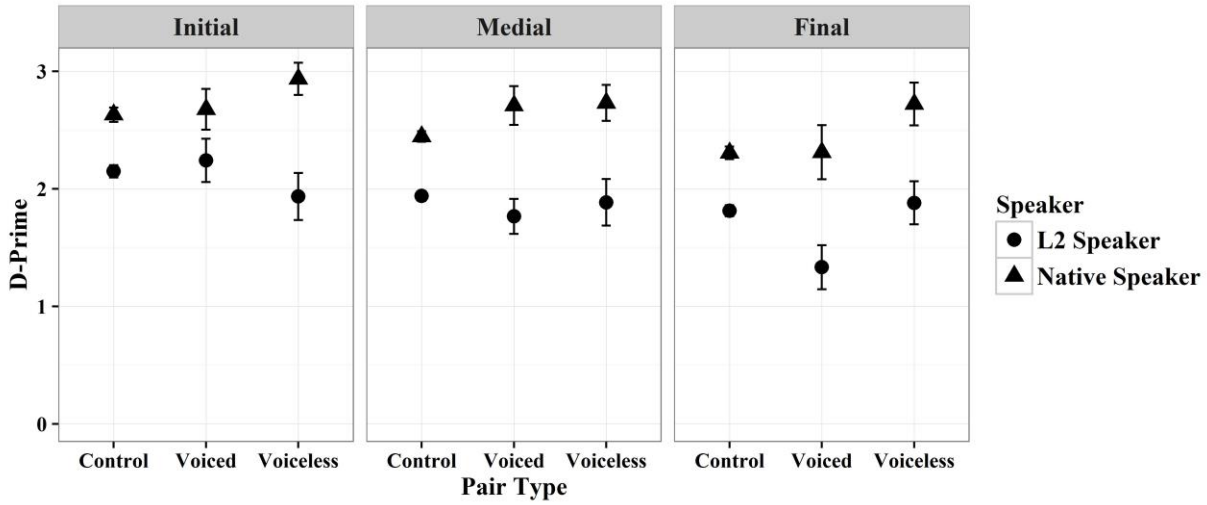

Figure 3: Discrimination d-prime for pairs similar in voicing

Log d-prime for the discrimination task were submitted to linear multilevel models as a continuous dependent variable. Visual inspection of residual and quantile-quantile plots revealed large deviations in normality for raw d-prime, which were rectified by the log transformation. To begin with the analyses comparing pairs of nonce words similar in place of articulation, fixed effects included pair type (pharyngeal, uvular, or control), phoneme context in the nonce words (initial, medial, or final), speaker (L2, native), and interactions between pair type, context, and speaker. Scores for the pharyngeal pair in the initial context acted as the baseline. By-subject random slopes for context significantly improved model fit ($\chi^2(6)=20.79$, $p=.002$).

Table 3 presents the final model of log d-prime scores for the discrimination task comparing pairs of nonce words similar in place of articulation and the final model for nonce words similar in voicing. In the initial context, L2 speakers were less sensitive to control (e.g., /sub/-/zub/; $t=4.42$) and uvular (/ʁub/-/χub/; $t=4.02$) phoneme pairs than they were to pharyngeal phoneme pairs (/ʕub/-/ħub/). In the medial context, L2 speakers maintained the significant difference in sensitivity between pharyngeal and uvular pairs and between pharyngeal and control pairs. L2 learners were significantly less sensitive to the pharyngeal phoneme pair in the final context than in the initial context ($t=7.27$), which led to learners being significantly less sensitive when discriminating pharyngeal pairs compared to control pairs in this context ($t=-6.50$). However, the difference in sensitivity between the pharyngeal and uvular pairs was maintained in the final context, although to a lesser degree than in the initial context. Overall, L2 speakers were more sensitive to nonce word pairs containing pharyngeals than they were to those containing uvulars.

When the same model was run with NSs as the baseline, NS sensitivity to pharyngeal pairs in the initial context was not significantly different than of uvular pairs ($b=.20$, $SE=.13$, $t=1.59$) nor control pairs($b=.13$, $SE=.09$, $t=1.48$). The same patterns appeared for NSs in the medial context. In the final context, like L2 speakers, NSs were significantly less sensitive to the pharyngeal pair than they were in the initial context ($b=.48$, $SE=.13$, $t=3.67$), which led them to be significantly less sensitive when discriminating pharyngeal pairs compared to control pairs in the final context ($b=-.35$, $SE=.13$, $t=-2.73$). The three-way interaction between pair type, context, and speaker was significant ($\chi^2(4)=13.88$, $p=.008$), driven by the fact that NSs were more sensitive to most pair types than L2 speakers, and the difference in sensitivity between pharyngeal and uvular pair types that was significant for L2 speakers was not significant for NSs.

	Place of articulation			Voicing			
Fixed effects	***b***	**SE**	***t*-value**	***b***	**SE**	***t*-value**	
Intercept (pharyngeal/voiced pair, initial context, L2)	.44	.08	5.55	.69	.08	8.39	
Pair type: Control	.31	.07	4.42*	.04	.07	0.58	
Uvular/voiceless	.39	.10	4.02*	.15	.10	1.50	
Context: Medial	.19	.10	1.93^	.24	.10	2.32*	
Final	.74	.10	7.27*	.39	.11	3.66*	
Native speaker	-.10	.13	-0.75	-.21	.14	-1.53	
Pair type x context: Control x medial	-.11	.10	-1.06	-.14	.10	-1.31	
Uvular/voiceless x medial	.13	.14	0.92	-.23	.14	-1.61	
Control x final	-.65	.10	-6.50*	-.24	.10	-2.30*	
Uvular/voiceless x final	-.18	.14	-1.35	-.36	.14	-2.53*	
Pair type x speaker: Control x native	-.18	.12	-1.54	-.05	.12	-0.38	
Uvular/voiceless x native	-.19	.16	-1.20	-.30	.16	-1.83^	
Context x speaker: Medial x native	-.06	.16	-0.34	-.26	.17	-1.51	
Final x native	-.26	.17	-1.55	-.21	.17	-1.23	
Pair type x context x speaker: Control x medial x native	.10	.16	0.59	.28	.17	1.65^	
Uvular/voiceless x medial x native	-.24	.22	-1.10	.37	.23	1.61	
Control x final x native	.30	.16	1.83	.25	.17	1.46	
Uvular/voiceless x final x native	.46	.22	2.07*	.31	.23	1.32	
Random effects	**Variance**	***SD***	**Correlation**	**Variance**	***SD***	**Correlation**	
Intercepts-Subject	<.001	<.001		<.001	<.001		
Slopes-Subject by context							
Medial	.02	.12	-.57	.01	.12	-.57	
Final	.02	.15	-.65	.02	.15	-.65	.99
Residual	.09	.30		.10	.32		

Note: *Significant at $p < .05$; ^Marginal at $p < .10$.

Table 3: Discrimination task results of logistic multilevel models for log d-prime for pairs similar in place of articulation on the right and pairs similar in voicing on the left. Intercept for place of articulation analysis was the pharyngeal pair type and was the voiced pair type for voicing analysis, as indicated by the "/" in the fixed effects descriptions.

Moving onto the analyses comparing pairs of nonce words similar in voicing, fixed effects were the same as those described for the place of articulation model, with the pair type baseline instead being voiced nonce word pairs. As indicated in Table 3, the pair type comparison was voiced vs. voiceless instead of pharyngeal vs. uvular. By-subject random slopes for context significantly improved model fit ($\chi^2(6)=17.66$, $p=.007$). L2 speakers were not significantly different in sensitivity to control (e.g., /sub/-/zub/) and voiceless (e.g., /χub/-/ħub/) pairs than they were to

voiced pairs (e.g., /ʁub/-/ʕub/), and the same pattern was observed in the medial context. Sensitivity to the voiced pair significantly changed from initial to medial (t=2.32) and from initial to final contexts (t=3.66). Due to L2 learners' dip in sensitivity to the voiced pair in the final context, learners were significantly less sensitive to the voiced pair compared to the control (t=-2.30) and voiceless pairs (t=-2.53). This finding was not predicted in either model but is consistent with the decrease in accuracy in the final context as detected in the identification task.

When the same model was run with NSs as the baseline, NSs generally did not show much variation in sensitivity. In the initial context, like L2 speakers, NSs' sensitivity to the voiced pair was not significantly different from that of the control (b=-.002, SE=.10, t=-0.03) or voiceless pairs (b=-.15, SE=.13, t=-1.16), as was also true in the medial and final contexts. The three-way interaction between pair type, context, and speaker was not significant ($\chi^2(4)=3.66$, p=.453), nor was the two-way interaction between speaker and context ($\chi^2(2)=0.13$, p=.936). However, the two-way interaction between speaker and pair type was significant ($\chi^2(2)=11.66$, p=.003), suggesting that NSs displayed less differences in sensitivity between pair types than L2 speakers.

4. Discussion. The current study examined the degree of abstraction in Arabic L2 learners' representations of the Arabic pharyngeal (/ʕ/, /ħ/) and uvular (/ʁ/, /χ/) phonemes. The PAM-L2 states that L2 learners attend to the phonological level, positing that the L2 learners in the current study would have relatively equal perception among the target phonemes. The ASP states that learners attend to the phonetic level, which includes more detail such as the variation in phonetic cues induced by different contexts. The predictions of the ASP based on previous literature were quantified for the current stimuli by measuring productions by NSs of Arabic and English and conducting both within- and cross-language DA. The updated ASP predictions based on these analyses were that L2 learners would more accurately perceive the pharyngeals compared to uvulars and the voiced compared to voiceless phonemes in the initial and final contexts. In the medial context, the cross-language DA indicated that L2 learners would be equally accurate in their perception of all phonemes. Although previous knowledge of the acoustic characteristics of the target Arabic phonemes was useful in generating ASP predictions, the cross-language DA were able to simulate the task of a leaner. For instance, it was previously known that both consonant duration and vowel second formant frequency following a pharyngeal or uvular differ (Bin-Muqbil, 2006), but measuring both of these characteristics in the same productions and submitting them to a DA allowed for a clearer picture of how the different phonetic cues work together to impact not only native language perception, but non-native perception as well.

The recordings were then utilized as stimuli during identification and discrimination tasks completed by L2 learners and NSs of Arabic. L2 learners accurately labeled the voiced pharyngeal more often than other target phonemes in all contexts, and NSs were overall more accurate in this task. This result was predicted by the cross-language DA, with the voiced pharyngeal being phonetically distinct from the remaining target phonemes. D-prime from the discrimination task demonstrated that L2 learners were better at discriminating nonce word pairs containing pharyngeal than those containing uvulars in all contexts. The cross-language DA also predicted this result in the initial and final context, but not in the medial context. L2 learners were equally able to discriminate voiced and voiceless pairs in the initial and medial contexts, but were less sensitive to the voiced pairs in the final context. None of the results regarding the nonce word pairs similar in voicing were predicted by either model, but the difference between contexts can be accounted for under the ASP, as its scope includes contextual variation.

While the results for every context and every task did not all follow the ASP predictions, the general patterns of the L2 speakers follow the ASP predictions more than the PAM-L2 predictions. The variation in responses between contexts in the discrimination task was not predicted by the PAM-L2, as it posits a level of representation where such detail is not encoded after categorization occurs. The ASP provides a framework in which to investigate not only differences in syllable or word position, as in the current study, but also to examine effects of coarticulation or the context of the overall utterance (e.g., in isolation, in a sentence; Nishi, Strange, Akahane-Yamada, Kubo, & Trent-Brown, 2008; Strange et al., 2005; Strange et al., 2007). In the discrimination task, the general greater sensitivity to pharyngeal pairs over uvular pairs by L2 speakers was predicted by the ASP but not by the PAM-L2.

Overall, the basic patterns of behavior appear to support the greater role of phonetic information in the participants' representations. However, several limitations constrain the generalizability of the conclusions. Due to inconsistency in production of the /a/ vowel, only the nonce words including /u/ were analyzed in the DA and used for the identification and discrimination task. More vowel variation with similar outcomes would lead to greater generalizability. Moreover, the proficiency of the L2 speakers was beginning to intermediate according to self-report. It is possible that greater proficiency would lead to a greater sensitivity to phonetic cues in the final context. These results tentatively posit that when learning non-native phonemes, leaners of an L2 gradually learn to attend to the phonetic detail in segments in order to efficiently use the phonetic cues for phonemic categorization. However, the findings are not inconsistent with more recent studies investigating PAM-L2 that call for use of phonetic and acoustic information when forming assimilation types (Bundgaard-Nielsen et al., 2011; Reid et al., 2015). Attending to the phonetic level allows learners to not only form categories that are attuned to the L2 phonetics instead of the L1, but also to form categories that account for context variability. More robust categories that are able to adjust to allophonic variation due to phonetic context lead to more native-like production and perception.

References

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 309–412. http://doi.org/doi:10.1016/j.jml.2007.12.005

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2016). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-12). Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, *167*, 224.

Bin-Muqbil, M. S. (2006). *Phonetic and phonological aspects of Arabic emphatics and gutturals* (Unpublished doctoral dissertation). University of Wisconsin.

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer (Version 5.4.04). Retrieved from http://www.praat.org/

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories. *Applied Psycholinguistics*, *32*(01), 51–67. http://doi.org/10.1017/S0142716410000287

Escudero, P., Simon, E., & Mitterer, H. (2012). The perception of English front vowels by North Holland and Flemish listeners: Acoustic similarity predicts and explains cross-linguistic and

L2 perception. *Journal of Phonetics*, *40*(2), 280–288. http://doi.org/doi:10.1016/j.wocn.2011.11.004
Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: Sage Publications Ltd.
Flege, J. E., & Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech*, *24*(2), 125–146. http://doi.org/10.1177/002383098102400202
Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124.
Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/ hierarchical models*. New York: Cambridge University Press.
Ghazeli, S. (1997). *Back consonants and backing coarticulation in Arabic* (Unpublished doctoral dissertation). The University of Texas at Austin.
Gilichinskaya, Y. D., & Strange, W. (2010). Perceptual assimilation of American English vowels by inexperienced Russian listeners. *The Journal of the Acoustical Society of America*, *128*(2), EL80. http://doi.org/10.1121/1.3462988
Linck, J. A., & Cunnings, I. (2015). The Utility and Application of Mixed-Effects Models in Second Language Research: Mixed-Effects Models. *Language Learning*, *65*(S1), 185–207. http://doi.org/10.1111/lang.12117
Li, P., Zhang, F., Tsai, E., & Puls, B. (2013). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition*. http://doi.org/10.1017/S1366728913000606
Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., & Trent-Brown, S. A. (2008). Acoustic and perceptual similarity of Japanese and American English vowels. *The Journal of the Acoustical Society of America*, *124*(1), 576. http://doi.org/10.1121/1.2931949
R Core Team. (2015). *R: A language environment for statistical computing*. R Foundation for Statistical Computin. Retrieved from http://ww.R-project.org
Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N. X., & Best, C. T. (2015). Perceptual assimilation of a lexical tone: The role of language experience and visual information. *Attention Perceptual Psychophysology*, *77*, 571–591. http://doi.org/10.3758/s13414-014-0791-3
Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*(4), 456–466. http://doi.org/10.1016/j.wocn.2010.09.001
Strange, W., Bohn, O.-S., Nishi, K., & Trent, S. A. (2005). Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *The Journal of the Acoustical Society of America*, *118*(3), 1751. http://doi.org/10.1121/1.1992688
Strange, W., Bohn, O.-S., Trent, S. A., & Nishi, K. (2004). Acoustic and perceptual similarity of North German and American English vowels. *The Journal of the Acoustical Society of America*, *115*(4), 1791. http://doi.org/10.1121/1.1687832
Strange, W., Weber, A., Levy, E. S., Shafiro, V., Hisagi, M., & Nishi, K. (2007). Acoustic variability within and across German, French, and American English vowels: Phonetic context effects. *The Journal of the Acoustical Society of America*, *122*(2), 1111. http://doi.org/10.1121/1.2749716
Tyler, M., & Fenwick, S. (2012). Perceptual Assimilation of Arabic Voiceless Fricatives by English Monolinguals. In *INTERSPEECH*.