A Comparison of Three Metrics of Perceptual Similarity in Cross-Language Speech Perception
Author(s): James D. Harnsberger
Proceedings of the 25th Annual Meeting of the Berkeley Linguistics Society (2000), pp. 157-168

A comparison of three metrics of perceptual similarity in cross-language speech perception

JAMES D. HARNSBERGER
Indiana University, Bloomington

1. Introduction

Cross-language speech perception models, such as the Perceptual Assimilation Model (PAM) and the Native Language Magnet (NLM) model, are highly dependent on assumptions concerning the mapping of non-native sounds onto native categories, as well as the internal structure of native categories, in predicting the discriminability of non-native contrasts (Best 1995, Kuhl 1991, Kuhl and Iverson 1995). These models differ from one another in cases in which both stimuli of a non-native stimulus pair are identified as exemplars of a single category (hereafter *Within-Category* assimilations). PAM predicts that the discriminability of two non-native stimuli is a function of their difference in category goodness, or difference in distance from the center of the category. In contrast, NLM predicts that the discriminability of two non-native stimuli is a function of their psychoacoustic difference, weighted by the proximity of the pair to the center of the category, or the category's *prototype*. This proximity effect is termed the *perceptual magnet effect.* Due to the perceptual magnet effect, the discriminability of a stimulus pair is predicted to decrease as the stimuli approach the prototype. In studies of within-category (WC) assimilations, internal category structure has been commonly quantified with goodness ratings, though some studies have used the frequency of a stimulus' identification as an exemplar of a particular category (Sussman and Lauckner-Morano 1995, Sussman and Gekas 1997). The evidence supporting NLM's model of category structure has been mixed (Kuhl 1991; Lively and Pisoni 1997; Lotto, Kluender, and Holt 1998; Kewley-Port and Neel 1998), while this aspect of PAM has not been directly tested.

While both models make predictions concerning the discriminability of WC assimilations, only PAM makes explicit predictions concerning the discriminability of non-native contrasts that correspond to a native contrast in the listener's experience (hereafter *Two-Category*, or TC, assimilations), namely that they would be highly discriminable. And neither model adequately accounts for non-native stimulus pairs which are not mapped consistently onto one or two categories (hereafter *uncategorizable* assimilations). NLM is silent on such pairs as it is a model solely of internal category structure. PAM makes a distinction between a pair of uncategorizable sounds (a *Both Uncategorizable*, or UU, assimilation) and a pair consisting of one uncategorizable sound and one sound that is consistently mapped onto a native sound (an *Uncategorizable-Categorizable*, or UC, assimilation). PAM does predict that UC assimilations should be as discriminable as the most discriminable within-category assimilations, though it is silent concerning the range of variability one might expect to find in the discriminability of such assimilations. However, for UU assimilations, discriminability is said to be a function of two factors, interstimulus difference (gesturally-based, given PAM's basis in direct realism), and the position of each stimulus in the overall perceptual space. Given that no further detail is provided in terms of integrating these two kinds of information, PAM is unable to predict the discriminability of a UU contrast.

The purpose of this study was to test the portions of PAM and NLM that are most amenable to cross-language validation, namely the relationship between a non-native contrast's identification pattern and its discriminability, as formalized in similarity metrics, specifically for TC and WC assimilations. In addition, the problem of predicting the discriminability of uncategorizable assimilations was addressed in this study by proposing and testing a relatively simple hypothesis concerning the identification-discrimination relationship, which will be referred to as the overlap hypothesis. It simply states that the discriminability of a stimulus pair is a function in the degree of overlap in the identification patterns of each stimulus; stimulus pairs whose identification patterns show less overlap are predicted to be more discriminable. A formalization of PAM, NLM, and the overlap hypothesis for two-category, within-category, and uncategorizable discriminations (when applicable) appear in section 3.

To test these formalizations cross-linguistically, nasal contrasts combining bilabial, dental, alveolar, and retroflex nasals from three languages (Malayalam, Marathi, and Oriya), varying in talker, syllabic position, and vowel context, were presented in an AXB discrimination test and an identification test with category-goodness ratings to seven listener groups varying in their coronal nasal consonant inventory: Malayalam (bilabial-dental-alveolar-retroflex), Marathi and Punjabi (bilabial-dental-retroflex), Tamil and Oriya (bilabial-alveolar-retroflex) and Bengali and American English (bilabial-alveolar). Multiple listener groups and non-native contrasts were used to increase the likelihood that the results would generalize to other groups and contrasts. Nasals varying in place of articulation constituted a stimulus set that had not been examined in previous cross-language speech perception studies. Moreover, this set was chosen to maximize the potential for varying degrees of perceptual performance by non-native listeners, given the potential difficulty that non-native place contrasts are often shown to present to listeners (Werker, Gilbert, Humphrey, and Tees 1981). The goal was to generate a sufficient number of each assimilation type, particularly within-category and uncategorizable assimilations, for the purposes of testing the formalizations of PAM, NLM, and the overlap hypothesis. Section 2 describes these experiments in detail; section 3 gives the formalizations, and thus predictions, of each model; section 4 reports the results of the experiments in terms of the formalizations of PAM, NLM, and the overlap hypothesis; and section 5 concludes with a discussion of the results and suggestions for future research.

2. Methods

2.1. Stimuli

Six talkers, two each of Malayalam, Marathi, and Oriya, were recorded reading from a list of real and nonsense words from their native language, in both isolation and a sentence frame, in five repetitions each for a total of ten repetitions. Nonsense words were read in cases where the lexicon of the language did not provide a word composed of a necessary sequence of vowel(s) and nasal(s). The nasals of interest appeared in all syllable positions allowable by the individual languages, in an [a], [i], or [u] vocalic context. Some relevant characteristics of the speakers who produced the stimuli are listed in (1).

(1) Demographics of talkers. NL = native language, either Malayalam (ML), Marathi (MR), or Oriya (OR). "Home" = home city or district within India. "Years" = years outside of an environment in which the native language is widely spoken.

Name	NL	Sex	Age	Home	Years	Other Languages Spoken
YM	ML	m	58	Malabar	29	English, Hindi
YS	ML	f	47	Malabar	26	English, Hindi, Tamil
MS	MR	m	26	Mumbai	1	English, Hindi
MV	MR	f	35	Mumbai	6.5	English, Hindi, Gujarati
OC	OR	f	35	Cuttack	5	English, Hindi, Marathi, Bengali
OS	OR	f	30	Bhubaneswar	9	English, Telegu, Hindi

The recording took place in a sound-attenuated chamber in the University of Michigan Phonetics Laboratory using a Panasonic SV3500 DAT recorder. The stimuli were digitized at 44.1 kHz (filtered at 22 kHz), randomized, and played with a 3.5s intertrial interval (ITI) to native speakers of the respective languages in an identification test in order to exclude any stimuli from use in the experiment which might be poor exemplars. Of the stimuli which were recorded and evaluated, four exemplars, two from each talker, of 18 types of stimuli were used in the discrimination and identification tests. The stimulus types are listed in (2).

(2) Stimuli and their source languages. The vocalic context was [a] for all but underlined stimuli. Underlining indicates that the stimulus appeared in [i] as well as [a] contexts. The dental nasal of Malayalam talker YM was produced as interdental.

		Nasal			
Language	Syllable	[m]	[n̪]	[n]	[ɳ]
Malayalam	VCV			✓	✓
Marathi			✓		✓
Oriya				✓	✓
Malayalam	VC:V	✓	✓	✓	✓
Marathi			✓		✓
Oriya					
Malayalam	VC				
Marathi			✓		✓
Oriya					

All of the nasal stimuli appeared in an [a] context, with the exception of the Malayalam medial geminate series, which appeared in both the [a] and [i] contexts. Pilot tests had shown that the perception by non-native listeners of talker YM's dental stimuli, which were actually produced as interdentals ([n̪]), differed due to vocalic context. In an [a] context, YM's [n̪] stimuli were identified as bilabial exemplars by the non-native listeners, with a subsequent effect on their discrimination of YM's dental contrasts. In an [i] context, YM's [n̪] received the more expected label of dental or alveolar, depending on the non-native listener group in question. Both vocalic contexts were preserved in this case, since four contrasts produced by talker YM could be expected to elicit within-category assimilations from a subset of the listener groups: [am:a]-[an̪:a], [in̪:i]-[in:i], [in̪:i]-[in̪:i], and [in:i]-[in̪:i]. The 18 stimulus types were combined to make 34 contrasts, with

contrast defined in terms of: the place of articulation of each member of a contrast, the talker who produced the contrast, the vocalic context the nasal stimuli appeared in, and the syllable type the nasal stimuli were embedded in.

2.2. Participants

Twelve to eighteen speakers each of Malayalam (N=18), Oriya (N=16), Marathi (N=14), Punjabi (N=13), Tamil (N=12), Bengali (N=15), and American English (N=18) were recruited and tested. All but the American English listeners were tested in India, in order to recruit subjects who varied little in terms of age, dialect spoken, and overall linguistic experience. All subjects from India were recruited by posting flyers on the campuses of local universities. American English listeners were recruited through introductory linguistics classes at the University of Michigan. Malayalam, Oriya, Tamil, and Marathi listeners were students attending national universities in India's capital, New Delhi. Bengali listeners were university students who were recruited and tested in Calcutta, the capital of West Bengal, where Bengali is primarily spoken. Punjabi listeners were recruited and tested in Amritsar, the cultural center of the Punjab state in northwestern India. None of the listeners who participated in the study had any experience with another language that would afford them additional nasal contrasts not found in their native language.

The Malayalam, Marathi, and Oriya listener groups were recruited to serve as controls. In addition, six listener groups represented three types of listener groups with particular coronal nasal consonant inventories: dental-retroflex (Marathi, Punjabi), alveolar-retroflex (Oriya, Tamil), and alveolar (Bengali, American English), all of whom could be expected to perceive a subset of the non-native contrasts as within-category or uncategorizable assimilations. The relevant nasal consonant inventory for each listener group is listed in (3), assuming that the category centers, or "prototypes", of these native categories are best represented by position-dependent allophonic variants of the phonemes associated with each category (Strange 1995). The predictions for a particular non-native contrast's assimilation to the native category or categories of a given listener group were difficult to make precisely, given the absence of a model of the similarity between one type of stimulus to another, or between one type of stimulus and one type of category. The absence of such a model has been cited as an important gap in cross-language speech perception models (Best 1995; Flege, Bohn, and Jang 1997). Thus, for this study, the predictions for the assimilation patterns were based on three assumptions or sources of information. First, it was assumed that a non-native stimulus would map onto a native category if both shared a common descriptive label, a position-dependent allophonic label. For example, the Malayalam dental nasal stimuli were predicted to map onto the Marathi and Punjabi listeners' native dental nasal category. Second, in the case of mismatches between a stimulus' label and any available native categories, past cross-language research shaped the prediction. For instance, in the case of the Malayalam and Marathi dental and retroflex nasals and native American English listeners, it was assumed that these non-native stimuli would map onto the English alveolar nasal, given prior work on English listeners with oral dental-retroflex stop contrasts (Werker et al. 1981, Polka 1991). Finally, all remaining predictions were guided by piloting work done with 3-6 speakers of Punjabi, Tamil, Bengali, and American English. These predictions appear in (4). These predictions could only be treated as tentative, and represented an attempt to elicit the range of assimilation types necessary for testing PAM, NLM, and the overlap hypothesis. The actual mapping of specific non-native sounds to specific

native categories was not crucial to the validation of these models, so long as a sufficient number of each assimilation type was elicited.

(3) Relevant (to the stimuli) perceptual category inventories for the six non-native listener groups. Syll = the type of syllable the nasal appears in. * = Oriya allows/disallows final consonants, depending on the dialect spoken (see Harnsberger 1998 for a summary).

Dental-Retroflex Group

	Marathi				Punjabi			
Syll	[m]	[n̪]	[n]	[ɳ]	[m]	[n̪]	[n]	[ɳ]
CV	[m]	[n̪]			[m]	[n̪]		
VCV	[m]	[n̪]		[ɳ]	[m]	[n̪]		[ɳ]
VC:V	[m]	[n̪]		[ɳ]	[m]	[n̪]		
VC	[m]	[n̪]		[ɳ]	[m]	[n̪]		[ɳ]

Alveolar-Retroflex Group

	Oriya				Tamil			
Syll	[m]	[n̪]	[n]	[ɳ]	[m]	[n̪]	[n]	[ɳ]
CV	[m]		[n]		[m]	[n̪]		
VCV	[m]		[n]	[ɳ]	[m]		[n]	[ɳ]
VC:V					[m]		[n]	[ɳ]
VC	[m] *		[n] *	[ɳ] *				

Alveolar Group

	Bengali				American English			
Syll	[m]	[n̪]	[n]	[ɳ]	[m]	[n̪]	[n]	[ɳ]
CV	[m]	[n̪]			[m]		[n]	
VCV	[m]		[n]		[m]		[n]	
VC:V	[m]		[n]					
VC	[m]		[n]		[m]		[n]	

2.3. Procedure

Two perceptual tests were administered to the subjects, a discrimination and an identification test. For both tests, stimuli were presented binaurally over Sony MDR-7506 headphones connected to a Sony TCD-D8 portable DAT recorder. Responses were made on photocopied answer sheets. The categorial AXB discrimination test consisted of 544 trials, 16 trials each of the 34 different types of contrasts. In order to ensure that the results were not dependent on the intelligibility of a single exemplar, two exemplars of each member of the 34 contrasts were used.

The contrasts appeared in four possible orders, AAB, ABB, BAA, BBA. A and B were always from the same talker, and all stimuli that were paired together were selected to minimize acoustic differences that were assumed not relevant to the identity of the stimulus, such as the duration or the fundamental frequency pattern of a stimulus. The ISI for the discrimination test was 1s, the ITI was 3s, and the IBI was 6s, with 20 trials per block. The total time for the discrimination test was 58.5

minutes. Subjects were told to indicate whether the nasal consonant in the first or last word was the same as the nasal consonant in the middle word by circling a number on the answer sheet. The term "nasal consonant" was defined through the use of simple examples in which nasals appeared in different syllable positions and vocalic contexts. A, X, or B were not physically identical, so listeners made categorial matches. Listeners were instructed to ignore "irrelevant differences" of duration, tone, or voice quality. A familiarization set of 20 trials was presented before the AXB test, with particular trials selected to reinforce the instructions. The identification test consisted of a labeling task in which listeners used response sets based on native phonemic categories. Listeners were also instructed to provide category-goodness judgments on a five-point scale. The ratings were necessary in evaluating PAM and NLM given these models' reference to category-goodness and prototypicality. Listeners identified and rated two exemplars each of the eighteen types of stimuli, produced by two talkers each, for a total of 72 stimuli. The identification test included two repetitions of this set, presented in random order, for a total of 144 trials. The test employed a 6s ITI and a 6s IBI, with 10 trials per block (total test time: 16.25 minutes). Listeners were again instructed to ignore "irrelevant differences" of duration, tone, or voice quality. Before the test began, ten trials were presented to familiarize listeners with the time allotted for labeling and rating a stimulus, with no feedback from the investigator.

(4) Predictions of possible assimilations. TC = two-category assimilation, U = uncategorizable assimilation, WC = within-category assimilation. * = predictions involving dental contrasts differed when Malayalam talker YM's contrasts were involved, given this talker's realization of dental nasals as interdental. YM's [m]-[n̪] in an [a] context were predicted to be WC for all non-native groups, while his [n̪]-[n] and [n̪]-[ɳ] in an [a] context were predicted to be TC for all non-native groups. These exceptional predictions were made based on earlier pilot studies.

Contrast	Assimilation Type	Listener Group(s)
[m]-[n̪]	TC/WC*	all
[n̪]-[n]	WC/TC*	all
[n̪]-[ɳ]	TC	Marathi, Punjabi
	U/TC*	Tamil, Oriya
	WC/TC*	Bengali, AE
[n]-[ɳ]	TC	Tamil, Oriya
	U	Marathi, Punjabi
	WC	Bengali, AE

3. Predictions

A set of specific predictions for the discriminability of non-native contrasts was generated from PAM, NLM, and the overlap hypothesis. They are listed below by three general assimilation types: two-category (TC), within-category (WC), and uncategorizable (U). A stimulus pair was classified as a TC assimilation if the two stimuli were identified as exemplars of different native categories and if each stimulus was identified with its corresponding category in 90% or more of responses. If one or both stimuli were identified by a listener group with a single native category in fewer than 90% of responses, then such a contrast was classified

as a U assimilation. A contrast was classified as a WC assimilation for a given listener group if both stimuli were identified as exemplars of a single-category in 90% of responses.

3.1. Two-Category (TC) assimilations

PAM predicts that these assimilations should elicit the highest scores of any assimilation type. In addition, for this experimental paradigm, PAM predicts that TC discrimination scores should fall between 90-100%. NLM makes no direct predictions concerning TC assimilations.

3.2. Uncategorizable (U) assimilations

NLM makes no predictions concerning stimulus pairs that are not within-category. PAM predicts only that, for two stimuli that fall outside of a native category, "discrimination is expected to range from poor to very good, depending on their proximity to each other and to native categories" (Best 1995:195). Two metrics are included here – one referring to interstimulus proximity, which also can be thought of as psychophysical salience, and a second referring to the proximity of each stimulus to one or more native categories (category proximity). Two problems emerge here in trying to predict the discrimination scores for an uncategorizable contrast. First, it is unclear how many category proximity distances must be computed for a given listener group. Should one include the proximity of a stimulus to the nearest native category, to the two nearest, or perhaps to all native categories in the same natural class? The second problem concerns the final formula combining all of the interstimulus and category distances: how are they weighted or combined in the final computation of the predicted discrimination score or range of scores?

To address this shortcoming of PAM, a very simple similarity metric was devised based on the overlap hypothesis described earlier, a *categorization* metric, which took into account the degree to which a non-native contrast mapped onto a native contrast. The assumption, based on the classic definition of Categorical Perception, was that listeners would be able to discriminate non-native contrasts that map onto a native contrast, with a corresponding decrement in discrimination ability with non-native contrasts that were less consistently categorized into two distinct categories. The metric was calculated solely on the basis of the frequency with which the non-native stimuli composing a contrast were mapped onto the various native categories possessed by a listener group. Specifically, the score was the sum of the differences in percent-identification of a stimulus to each native category of the listener group, as shown in (5).

$$(5)\quad C = \sum_{i=1}^{n} | A_i - B_i |$$

where C = categorization score, n = the number of native categories, A_i = percent identification of stimulus A as an exemplar of category i, and B_i = percent identification of stimulus B as an exemplar of category i. In essence, the categorization metric is a measure of the degree to which two stimuli differ with one another in their degree of identification with multiple categories.

3.3. Within-Category (WC) assimilations

Within PAM, within-category stimulus pairs can either be classified as single-category (SC) or category-goodness (CG) assimilations (Best 1995). The

distinction between these two assimilation types lies in the degree to which the two stimuli forming a contrast differ in their similarity to center of the category: a stimulus pair in which both stimuli are equally good, or equally poor, exemplars of a single category is classified as an SC assimilation, while a stimulus pair in which both stimuli differ in their similarity to the category center is classified as a CG assimilation. A common similarity metric underlies this distinction, that the degree of difference in category-goodness between two stimuli determines their discriminability. Such a metric can be formalized as:

(6) $D = |R_A - R_B|$

where D = discriminability of a hypothetical /A-B/ non-native contrast, and R_A and R_B = the mean category-goodness rating of stimulus A and stimulus B, respectively.

Within the framework of NLM, within-category assimilations can conceivably be classified in terms of three assimilation types for the purposes of comparison with PAM: prototypical (both stimuli are prototypical), nonprototypical-prototypical, and nonprototypical (both stimuli are nonprototypical). NLM predicts that prototypical assimilations should elicit lower discrimination scores than nonprototypical ones, holding constant the psychoacoustic, or interstimulus, differences involved (Kuhl and Iverson 1995). No predictions can be made directly by NLM for nonprototypical-prototypical assimilations, though one would expect them to elicit higher scores than prototypical ones, since they involve a stimulus that is perceptually distinct from a prototypical stimulus. The perceptual similarity metric underlying all three assimilation types is more complex than (6). It involves two measures: of the similarity of each stimulus to the category center, or prototype, and the interstimulus similarity, or its acoustic robustness. The first measure is one typically taken by, for instance, category-goodness ratings; the latter is simple to quantify for synthetic stimuli. However, for natural stimuli, quantifying acoustic robustness involves assumptions concerning what aspects of the acoustic signal are relevant for speech, and how to combine disparate types of information in the signal (frequency, duration, amplitude). For the purposes of this study, the following formalization of the NLM was used:

(7) $D = AR / R_X$

where D = discriminability of contrast /A-B/, AR = acoustic robustness (defined below), R_X = the mean category-goodness rating of either stimulus A or B (lowest).

In essence, the formalization in (7) states that the discriminability of a stimulus pair is a function of its acoustic robustness, weighted by the proximity of the stimuli to the category center. Proximity here is defined in terms of just one of the two stimuli, namely the one furthest from the prototype, the one eliciting the lowest goodness rating in this study. This metric assumes that only one stimulus out of the pair needs to be at the periphery of the category for the perceptual magnet effect to be minimal. Thus, nonprototypical-prototypical stimulus pairs were assumed to be as discriminable as nonprototypical stimulus pairs, holding acoustic robustness (AR) constant, since the former involves a stimulus that is perceptually distinct from a prototypical stimulus. AR was calculated for this stimulus set by summing the F2 and F3 transition magnitude differences (calculated in Bark) between the two stimuli forming the contrast. Formant transition differences have been frequently cited as

important acoustic cues in differentiating nasal consonants differing by place (Larkey, Wald, and Strange 1978; Dart 1991). The transition magnitude difference for a given formant was calculated as follows: the difference was taken between the formant's center value (measured in Hz using narrowband FFTs and then converted to Bark) at vowel offset (12.5ms before the vowel/murmur boundary), and its value at vowel temporal midpoint. Thus, rising transitions into the murmur had positive transition magnitude values, while falling transitions had negative values. In cases of stimuli with medial nasals, and thus two vowel/murmur transitions, the transition magnitudes for a given formant were averaged (i.e. the F2 transition magnitudes of the preceding and following vowel were averaged for a given stimulus). The transition magnitude difference between two stimuli could then be calculated for a given formant (F2, F3). The difference scores for each formant could then be summed for the AR measure in (7). Unfortunately, AR could only capture some important *spectral* differences in contrasts. In some contrasts, significant *temporal* differences between the murmurs were observed, as well as differences in the *spectral* structure of the murmurs. Contrasts which exhibited such differences were excluded from the test of metric (7).

4. Results

The control groups performed as expected for all native contrasts, with one exception: Malayalam listeners' mean percent correct discrimination score for Malayalam talker YS' [aɳ:a]-[aɳ;a] contrast was 82%. Of the 238 assimilations (34 types of contrasts * 7 listener groups) collected in this study, the majority (157) were uncategorizable; 47 were two-category, while only 34 involved within-category stimulus pairs.

4.1. Two-category assimilations

All of the two-category assimilation scores fell within a high range, 87-100%, averaging 97% ($s = 3\%$), which matches closely with PAM's predictions. Only two assimilations failed to reach the 90% criterion, the identification of Malayalam talker YS' [am:a]-[aɳ:a] and [ana]-[aɳa] by speakers of Oriya.

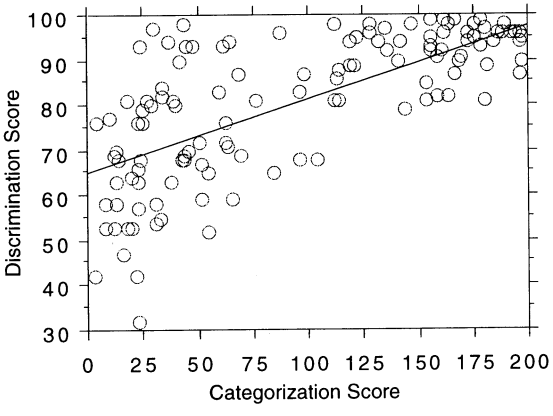
4.2. Uncategorizable assimilations

The scores for the uncategorizable contrasts varied greatly, from 32-99%, with the majority of scores falling within an upper range (80-100%). For each uncategorizable assimilation, a categorization score was generated. The categorization scores for all the identification results for non-native contrasts were paired with the discrimination scores in a simple regression analysis. If discrimination performance could be explained simply by the identification patterns, as quantified by the categorization scores, then we would expect to see a very high correlation across the entire distribution of the discrimination scores. The results of the regression analysis are presented in (8).

The discrimination and categorization scores were significantly correlated ($F[1] = 209$, $p < 0.0001$), $r = 0.76$. However, at lower categorization scores, there were more outliers from the regression line. The lower range of categorization scores corresponded, in most instances in this study, to identification patterns that were more like within-category assimilations, in which both stimuli composing a contrast were consistently judged as exemplars of the same category. One explanation for the variance at lower categorization scores would be that individual categories possess some degree of phonetic detail, allowing for a certain degree of listener

sensitivity to fine-grained, nonphonemic information. Intracategory detail cannot be modeled with this score, since it is based on the degree to which stimuli overlap in different perceptual categories.

(8) Categorization scores plotted against discrimination scores for uncategorizable assimilations.



4.3. Within-category assimilations

The capacity to test PAM and NLM predictions concerning intracategory pairs (single-category vs. category-goodness, prototypical vs. nonprototypical and nonprototypical-prototypical) was limited by the small number of within-category assimilations elicited in this experiment. However, many of the uncategorizable contrasts were within-category assimilations if only the top labeling choice of a particular listener group was considered. To generate more test cases for NLM and PAM predictions, the discrimination scores for within-category (WC) and uncategorizable near-WC assimilations were recalculated to represent only those subjects who consistently (in 90% of more of responses) labeled both stimuli of a contrast with the same label. When inconsistent labelers were excluded from the analysis, in some cases, so few consistent labelers remained that the corresponding mean discrimination score was an average over a much smaller group of listeners. In cases where the remaining pool of labelers dropped below 8, the assimilation was excluded from the analysis.

Using only the "consistent" subjects, a total of 64 within-category assimilations were generated. Of those, all 64 were used to test PAM's similarity metric (see (6)), while 50 were used to test the formalization of NLM appearing in (7); 14 assimilations were excluded in testing (7) due to the limitations of the acoustic robustness portion of the metric, outlined in section 3.3. The scores generated by both metrics, based on the identification test results, were included with the corresponding discrimination scores in a Spearman rank correlation analysis. Both the PAM and the NLM metrics showed significant, though low correlations ($z = 2.2$, $p = 0.03$, $r = 0.32$; $z = 4.7$, $p = 0.0001$, $r = 0.68$; respectively). While the NLM metric may have appeared to have fared better than the PAM metric, in fact, it was no improvement over a correlation between the acoustic robustness and discrimination test scores ($z = 4.7$, $p = 0.0001$, $r = 0.69$). Given that the NLM metric incorporates

acoustic robustness, the weighting due to the particular formalization of the perceptual magnet effect in (7) was expected to have produced scores that showed a greater correlation with the discrimination scores than the measures of acoustic robustness alone.

5. Discussion and Conclusion

The results obtained in this experiment revealed an unexpected range of discrimination scores for different types of within-category assimilations, and a large proportion of uncategorizable contrasts, raising a number of problems for both PAM and NLM. First, it was clear that neither model could account for the majority of assimilations elicited in this study, uncategorizable ones, involving one or more stimuli that fell between native categories. The proposed categorization score was a somewhat successful attempt at accounting for variation in the discriminability of uncategorizable contrasts, though other factors are apparently at work in within-category assimilations. The categorization score's advantage was in its ability to capture less categorical, more graded identification responses that were so prevalent in this study, due no doubt to the use of several listener groups with more than one relevant perceptual category for the stimuli in question.

Second, the formalizations of PAM and NLM predictions for the identification-discrimination function were not strongly upheld by the results. This poses the greatest problem for PAM, given that the formalization in (6) represents a very straightforward translation of the principles articulated by Best (1995) for within-category assimilations (single-category plus category-goodness). These results, and the modest correlation reported between the acoustic robustness (AR) and the discrimination scores, suggest than interstimulus distance is an important factor in within-category assimilations, contrary to the predictions of PAM. The failure to validate NLM could have been due to this study's particular formalization of the perceptual magnet effect, conceived of as a weighting of AR, or it could reflect the limitations of the AR portion of the metric. AR was limited to the magnitude of transitions of the second and third formants, two cues that appeared to show the greatest cross-stimuli differences in an acoustic analysis of the stimuli[1] (Harnsberger 1998). However, only a subset of the within-category assimilations (50/64) were used in testing the NLM formalization, constituting contrasts that only appeared to be cued by F2 and F3 transition differences. If the NLM formalization was appropriate, then the results of the study are consistent with those that have failed to find a perceptual magnet effect (Lively and Pisoni 1997; Lotto, Kluender, and Holt 1998).

The results reported in this study provide a number of suggestions for future cross-language studies. First, the frequency with which uncategorizable contrasts were elicited in this study points to the importance of this assimilation type in cross-language speech perception models and the need for formal models to account for their relative discriminability. The categorization metric of the overlap hypothesis represents the beginnings of one such model, but it clearly must be augmented by some sort of model of the internal structure of categories. The model of internal category structure in PAM was clearly not supported by this experiment, and it is unclear whether or not NLM can provide an adequate model. Such a model doubtless requires more cross-language study, using both natural and synthetic stimuli, and examining both vowels and consonants, with a variety of listener groups. Such future experiments should allow us to successfully evaluate current perceptual category models and develop more predictive models of perceptual performance than have been hitherto possible.

Notes

[1] The acoustic analysis included measures of F1-F5 at the onset, midpoint, and offset of the vowel and murmur portions of each stimulus, as well as measures of overall stimulus duration and murmur duration.

References

Best, C.T. (1995). A direct realist view of cross-language speech perception. In *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, 171-203, ed. by W. Strange. Baltimore: York Press.

Dart, S. (1991). *Articulatory and Acoustic Properties of Apical and Laminal Articulations*. Ph.D. dissertation, UCLA.

Flege, J.F, Bohn, O-S, and S. Jang (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25, 437-470.

Harnsberger, J.D. (1998). *The perception of non-native nasal contrasts: A cross-linguistic perspective*. Ph.D. dissertation, The University of Michigan (unpublished).

Kewley-Port, D. and Neel, A.T. (1998). Relation between discrimination and identification of English vowels. *Journal of the Acoustical Society of America* 105, 2983.

Kuhl, P.K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50, 93-107.

Kuhl, P.K. and Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect". In *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, 121-154, ed. by W. Strange. Baltimore: York Press.

Larkey, L., Wald, J., and Strange W. (1978). Perception of synthetic nasal consonants in initial and final position. *Perception and Psychophysics* 23, 299-312.

Lively, S.E. and Pisoni, D.B. (1997). On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 23, 1665-79.

Lotto, A.J., Kluender K.R., and Holt, L.L. (1998). Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America* 103, 3648-3655.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *Journal of the Acoustical Society of America* 89, 2961-77.

Strange, W. (1995). Cross-language studies of speech perception: A historical review. In *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, 3-45, ed. by W. Strange. Baltimore: York Press.

Sussman, J. and Gekas, B. (1997). Phonetic category structure of [I]: Extent, best exemplars, and organization. *Journal of Speech and Hearing Research* 40, 1406-24.

Sussman, J. and Lauckner-Morano, V.J. (1995). Further tests of the "perceptual magnet effect" in the perception of [I]: Identification and change/no-change discrimination. *Journal of the Acoustical Society of America* 96, 539-52.

Werker, J.F., Gilbert, J.H.V., Humphrey, K., and Tees, R.C. (1981). Developmental aspects of cross-language speech perception. *Child Development* 52, 349-53.