ON THE ROBUSTNESS OF SPEECH PERCEPTION

Randy L. Diehl

Department of Psychology and Center for Perceptual Systems, University of Texas, USA
diehl@psy.utexas.edu

ABSTRACT

Speech perception is remarkably robust despite many potential communicative obstacles such as background noise and talker variation. One reason for this is that speech sounds are selected to be auditorily distinctive even under adverse listening conditions. For many years, my colleagues and I have been exploring a general account of the distinctiveness of vowels and consonants referred to as the auditory enhancement hypothesis, which states that the phonetic properties of sound categories covary as they do largely because language communities tend to select properties that have mutually reinforcing auditory effects. Such relations of mutual enhancement create *intermediate perceptual properties* that serve, individually and in combination, as the basis for distinctive features. A second possible reason for the robustness of speech perception is currently being investigated in our laboratory. By comparing speech sound categorization performance of human listeners and Bayesian ideal classifiers, we have obtained evidence that human performance may be optimized with respect to the distributional properties—especially the degree of category overlap—of naturally produced speech sounds. Thus, perceptual robustness may be explained in terms of both auditory properties of speech sounds and strategies of listeners to optimize sound categorization.

Keywords: speech perception, auditory enhancement hypothesis, *intermediate perceptual properties* (IPPs), optimal speech categorization strategies

1. INTRODUCTION

It is useful to remind ourselves from time to time just how many impediments there are to successful speech communication. Environmental noise and reverberation, the common occurrence of partial hearing loss, and variation in vocal-tract properties, dialect, and idiolect all conspire to make the listener's task more challenging. As a result, phonologies or sound systems have no doubt been selected to be fairly robust signaling systems. When listening conditions are relatively favorable, talkers are apt to trade away a certain amount of auditory distinctiveness for greater ease of production [21]. However, the potential for increasing distinctiveness must be built into the signaling system, to be exploited when necessary.

2. THE DISPERSION PRINCIPLE AND THE AUDITORY ENHANCEMENT HYPOTHESIS

The idea that speech sound inventories are structured to maintain perceptual distinctiveness has a long history in linguistics [14, 26, 29]. Among the first investigators to express this idea quantitatively were Liljencrants and Lindblom [20] who showed that the structure of common vowel inventories appears to reflect a tendency toward maximal dispersion (i.e., maximal inter-vowel distances) within the available phonetic space.

The dispersion principle predicts correctly that the point vowels, /i/, /a/, and /u/, which represent acoustic and auditory extrema, should be widely attested among the world's language, and that /e/ and /o/ should also be relatively common. Although, as implemented by Liljencrants and Lindblom, the dispersion principle predicts too many high vowels for larger inventory sizes, this problem was mitigated when the phonetic space was later defined using a more realistic auditory model [2, 6]

Let us explore in more detail how talkers implement a strategy of vowel dispersion. One account is provided by the auditory enhancement hypothesis [4, 5, 16], which attempts to explain common patterns of phonetic covariation on listener-oriented grounds. The hypothesis states that the phonetic properties of sound categories co-vary as they do largely because language communities tend to select properties that have mutually reinforcing auditory effects. In the case of vowels, auditory enhancement is most often achieved by combining articulatory properties that have similar—and hence reinforcing—acoustic consequences.

The high back vowel /u/ offers a good example of how auditory enhancement works. This vowel is distinguished from lower vowels in having a low first formant frequency (F1) and from more anterior vowels in having a low second formant frequency (F2). Articulatory properties that produce a lowering of F1 and F2 thus contribute to the acoustic/auditory distinctiveness of /u/. From acoustic theory (e.g., [7]), we know that for a tube-like configuration such as the vocal tract, there are several ways to lower a resonance frequency. These are: (i) to lengthen the tube at either end, (ii) to constrict the tube at any node in the standing pressure wave corresponding to the resonance, and (iii) to dilate the tube at any antinode in the same standing pressure wave. It turns out that in carefully articulated tokens of /u/, all of these options are exploited.

Vocal tract lengthening can be achieved by protruding the lips, a characteristic component of the lip-rounding gesture that occurs during the production of /u/. Lengthening is also achieved by lowering the larynx, and this also has been observed in /u/ production [24, 32]. Both F1 and F2 are lowered owing to these gestures. The two primary vocal tract constrictions occur at the lips (constriction of the lip opening is another component of the lip-rounding gesture) and in the velar region near the junction between the oral and pharyngeal cavities. Because the lip orifice is located at the nodes in the standing pressure waves corresponding to the first and second resonances, lip constriction lowers both F1 and F2. The velar constriction occurs near another node in the standing wave corresponding to the second resonance, which contributes additional F2 lowering. Finally, vocal tract dilations in the vicinity of the mid-palate and the lower pharynx (both near second resonance antinodes) produce yet more lowering of F2. In summary, the shape of the vocal tract during the production of a carefully articulated /u/ is optimally suited for enhancing distinctiveness. Similar arguments apply to the other point vowels /i/ and /a/.

3. AUDITORY ENHANCEMENT IN THE PRODUCTION OF THE [VOICE] DISTINCTION

It is known from early studies of perceptual confusions in noise (e.g., [27]) that the English [voice] contrast is perceptually very robust. Here we consider three reasons why this is so.

(1) Typically, there are multiple, auditorily independent correlates of a phonological distinction. These correlates correspond to what my colleagues and I have referred to as *intermediate perceptual properties* (IPPs). In the case of the [voice] distinction, one important IPP is the presence of low-frequency energy or periodicity during or near the consonant constriction interval [16, 34]. This IPP has been labeled the "low-frequency property." In languages such as English, another IPP associated with the [voice] distinction (especially in initial position) is the presence or absence of significant aspiration. Still another IPP, important for medial and final stops in English and many other languages, is the ratio of consonant duration to preceding vowel duration, with [+voice] consonants having significantly smaller C/V duration ratios than [-voice] consonants [30]. The presence or absence of the low-frequency property, the presence or absence of aspiration, and the C/V duration ratio each corresponds to an auditorily independent perceptual variable, and collectively they provide redundant specification of the [voice] contrast. Although not every such variable is exploited in every language or in every utterance position, there are typically more than one IPP specifying whether a consonant is [+voice] or [-voice]. Thus, in English initial consonants, the [voice] distinction is signaled both by the presence or absence of the low-frequency property and by the presence or absence of aspiration, whereas in medial consonants the contrast is signaled by the presence or absence of the low-frequency property and by the C/V duration ratio.

(2) Each of the auditorily independent correlates (that is, the IPPs) of a phonological distinction may typically be analyzed into multiple sub-properties that are mutually enhancing in the sense that they all contribute to the same IPP. (This is a restatement of the auditory enhancement hypothesis as it applies, in particular, to consonants.) Consider, for example, the low-frequency property of [+voice] consonants. As Stevens and Blumstein [34] pointed out, this property can be analyzed into at least three phonetically separate sub-properties—voicing during the consonant constriction interval, a low F1 near the constriction interval, and a low fundamental frequency (F0) in the same region. All three of these sub-properties are typical correlates of [+voice] consonants, and each has been found to cue or enhance the perception of the

[+voice] category. It is reasonable to hypothesize, following [34], that the three sub-properties form a single integrated acoustic or perceptual property, or IPP. A similar kind of analysis may also be applied to the C/V duration ratio. It is obvious that the difference between [+voice] and [-voice] consonants in duration ratio may be enhanced by varying the consonant duration, the vowel duration, or both. Just as voicing during the constriction interval, a low F1 and a low F0 are mutually enhancing in that they all contribute to the low-frequency property, a short consonant and a long preceding vowel both contribute to the relative small C/V duration ratio characteristic of [+voice] consonants. Relative to either durational cue in isolation, a ratio of the two durations permits a much wider range of variation and hence greater potential distinctiveness.

(3) The sub-properties that contribute to one IPP of a phonological distinction often contribute to other IPPs of the same distinction. That is, the sub-properties have more than one perceptual role and often these roles are auditorily independent. Lisker [22] showed that variation in closure duration is sufficient to cue the distinction between medial [+voice] and [-voice] consonants. He later found that the presence of voicing during the closure yielded an increase in [+voice] labeling responses [23]. The latter is an example of what is often referred as a *trading relations* experiment. In such an experiment, listeners' categorization of stimuli varying in the value of one acoustic property is shown to depend on the value of a second acoustic property that the experimenter varies orthogonally. The second property is said to "trade" with the first if the category boundary with respect to the first property shifts as a function of the value of the second property.

Parker, et al. [28] conducted a modified version of Lisker's trading relations experiment [23]. Two stimulus series ranging perceptually from /aba/ to /apa/ were created by varying the closure duration of the consonant. The two series differed only with respect to the presence or absence of low-frequency pulsing (simulating voicing) during the closure. As in the study by [22], variation in closure duration was sufficient to cue the /b/-/p/ distinction. Also, as expected on the basis of [23], the presence of pulsing during closure shifted the /b/-/p/ labeling boundary toward larger values of closure duration (i.e., there were more [+voice] responses). This boundary shift is, of course, predicted by the fact that pulsing contributes to the low-frequency property characteristic of [+voice] consonants. However, [28] hypothesized that the presence of a pulsing segment has another effect as well, namely, it reduces the *perceived* closure duration, making the stimulus appear even more strongly [+voice].

To test this hypothesis, [28] also prepared three sets of non-speech stimuli that mimicked the temporal and peak amplitude properties of the /aba/-/apa/ stimuli in both the pulsing and no-pulsing conditions. In one set every stimulus consisted of two *steady-state* square-wave segments equal in duration to the pre- and post-closure segments of the /aba/-/apa/ stimuli. One stimulus series in this set had silent intervals of varying duration corresponding to the silent closure intervals of the speech no-pulsing condition. The other series in this set was identical, except that the medial gaps contained the same segments of pulsing used in the corresponding speech condition. A second set of square-wave stimuli was the same as the first, except that F0 fell over the final 40 ms of the first square-wave segment and then rose again over the initial 40 ms of the second square-wave segment. These are referred to as *fall-rise* stimuli. Finally, in a third set of square-wave stimuli, called the *rise-fall* stimuli, the direction of F0 change was reversed in the vicinity of the medial gap. For each of the three non-speech stimulus conditions, listeners identified both the pulsing and no-pulsing series. First, they were presented a random sequence of the two series-endpoint stimuli (i.e., the stimuli with the shortest and longest gap durations) and were required to learn by means of feedback lights which of two response keys corresponded to each endpoint stimulus. The listeners then identified the entire stimulus series with the instructions to press the key corresponding to the training stimulus to which item sounded most similar.

The presence of pulsing during the medial gap produced a significant boundary shift only in the fall-rise condition. The shift was in the same direction as that for the /aba/-/apa/ stimuli but only about 1/3 the magnitude. Thus, at least in the fall-rise condition, the presence of pulsing made the gap between the square-wave segments appear smaller in length. This is consistent with the hypothesis of [28] that one effect of voicing is to enhance the closure duration cue for [+voice] stops.

How are we to account for the differences across the various speech and non-speech conditions in the size of the boundary shift induced

by the presence of pulsing? The relatively large boundary shift in the /aba/-/apa/ condition can be explained on the assumption that voicing serves at least two independent roles in specifying the [+voice] category. First, it is a major contributor to the low-frequency property, a main IPP of [+voice] consonants. Second, it enhances the closure-duration cue (or C/V duration ratio IPP) by making brief closures seem even shorter. However, in the square-wave conditions, pulsing during the medial gap serves, at most, only the later of the two roles. This is because the non-speech training categories were defined on the basis of gap duration alone and were uncorrelated with the presence or absence of pulsing. Thus, for the square-wave categories, the low-frequency property has no distinctive role, and the effect of pulsing is limited to altering the apparent duration of the medial gap.

What remains to be explained is why, among the three square-wave conditions, pulsing had a significant effect only in the fall-rise condition. A tentative explanation is that a falling pattern before the gap and a rising pattern after the gap makes the pulsing more continuous with the flanking sounds, and that this continuity is necessary for pulsing to be integrated with the rest of the signal so as to influence the perceived duration of the medial gap. It is noteworthy that a fall-rise spectral pattern is characteristic of vowels flanking [+voice] stops in natural speech. As described earlier, both F1 and F0 are lower in the vicinity of [+voice] stop closures than [-voice] ones, and these differences have been shown to influence [voice] judgments. It was earlier argued that a low F1 and a low F0 near the consonant closure affect [voice] judgments by enhancing the low-frequency property characteristic of [+voice] consonants. However, the square-wave results suggest that they may also influence medial [voice] judgments in a quite independent way, namely, by creating a higher degree of spectral continuity between the closure pulsing and the flanking vowels, hence contributing to the perceived shortening of the closure.

The interlocking network of mutual enhancement relations apparently does not end there. Javkin [15] reported that when listeners were asked to vary the duration of a tone to match that of a vowel, the presence of voicing in the following consonant yielded tone duration settings that were reliably longer. This means that voicing contributes to a more distinctive C/V duration ratio in two ways: by shortening the perceived duration of the consonant and by lengthening the apparent duration of the vowel. These effects, together with the contribution of voicing to the low-frequency property, help to explain the perceptual robustness of the [voice] contrast.

We can now understand the senses in which an IPP is an *intermediate* perceptual property. It is intermediate because it lies between the raw, measurable phonetic properties of the speech signal and the distinctive feature values, because the phonetic properties may contribute to more than one IPP, and because each distinctive feature value may be determined by the values of more than one IPP.

4. THE ADAPTED GARNER PARADIGM: REVISING THE LOW-FREQUENCY PROPERTY HYPOTHESIS

An adapted version of the Garner paradigm [11] was used to evaluate further the claim that voicing during the consonant closure and low values of F1 and F0 all contribute to the same IPP—the low-frequency property [3, 17]. This paradigm permits a more direct test of perceptual integration of phonetic sub-properties than does the trading relations design used by [28]. In the adapted Garner paradigm, a fixed classification design is used to test whether stimuli varying orthogonally in two acoustic properties are more discriminable when the properties are combined in the same way they are combined in natural speech. For example, are stimuli that pit a low F1 and a long voicing continuation against a high F1 and a short voicing continuation more discriminable than stimuli that pit the opposite combinations of values for those acoustic properties against one another? If the answer is yes, then we can conclude that a low F1 and a long voicing continuation integrate perceptually, that is, they contribute to the same IPP. These fixed classification experiments measure sensitivity to stimulus differences rather than response biases as in the trading relations experiments.

[3] and [17] tested the hypothesis that F1, F0, and closure voicing co-vary between intervocalic stops contrasting for [voice] because they integrate perceptually. The IPP produced by the integration of these acoustic properties was assumed to be the low-frequency property. Both a low F1 and a low F0 at the edges of vowels flanking the stop were found to integrate perceptually with the continuation of voicing into the stop, but not with

each other. These results suggest that our original definition of the low-frequency property is not strictly correct. The perceptually relevant property is the continuation of low-frequency energy across the vowel-consonant boundary and not simply the amount of low-frequency energy near the consonant constriction. [17] also found that neither F1 nor F0 at the edge of the vowel integrate with closure duration, which shows that only auditorily similar properties integrate and not any two properties that reliably co-vary. Finally, parallel experiments using non-speech (viz., single formant) analogs showed that the above acoustic properties either integrate perceptually (or fail to) in the same way as in the original speech stimuli. This result indicates that integration arises from the auditory similarity of certain acoustic correlates of the [voice] distinction and does not depend on speech experience per se.

The trading relations experiments of [28] and the fixed classification experiments of [3] and [17] underscore the important role of auditory enhancement in maintaining a robust speech signaling system.

5. DO LISTENERS USE OPTIMAL SPEECH CATEGORIZATION STRATEGIES: COMPARISONS WITH IDEAL OBSERVERS

Whereas the representation of speech signals in the auditory periphery is now fairly well understood, models purporting to describe knowledge-based components of human speech perception and lexical access remain at an early stage of development. What has often been lacking in the analysis of human speech processing tasks is what David Marr [25] referred to as a "computational theory," that is, an abstract and formally rigorous characterization of the information processing problem to be solved, of constraints on possible solutions, and of optimal or at least efficient solutions given these constraints.

[12, 13] argued that in the domain of perception, an appropriate formalization of the notion of "computational theory" is provided by the concepts of Bayesian statistical decision theory, and, more particularly, of ideal observer theory. The aim of ideal observer theory is to determine optimal performance for a perceptual task given the physical and statistical properties of the relevant stimulus set. Humans and other organisms generally do not perform optimally on perceptual tasks, and therefore ideal observers are not, in most cases, to be construed as models of actual performance. Nevertheless, derivation of an ideal observer for a perceptual task is extremely useful because (1) it provides a precise measure of the information available to perform the task, (2) it serves as a computational theory of how to perform the task given the constraints that apply, (3) it offers an appropriate benchmark for evaluating the performance of real observers, and (4) it can be restricted in various ways, making it a useful starting point for developing testable models of actual performance. Ideal observer theory was first developed within the context of information theory and signal detection theory. It has been applied to a wide variety of perceptual tasks including sensory coding, detection and discrimination, categorization, parameter estimation, object recognition, and spatial navigation. It has even been used to characterize optimal and actual pathways in the evolution of perceptual systems [12, 13].

Najemnik and Diehl (unpublished) have recently developed a set of techniques for approximating ideal classifiers for natural speech, applying these techniques to a subset of American English phoneme categories sampled from the TIMIT Continuous Speech Corpus. We sampled 256 male tokens of each of six phonemes that form contrastive pairs: the stop consonants /b/ versus /p/, the fricative consonants /s/ versus /ʃ/, and the vowels /ɪ/ versus /ɛ/. Given the 16 kHz sampling rate used and the length of each sample (e.g., a 100-ms sound segment corresponds to an 800 element vector), the dimensionality of the space in which the tokens are described must be greatly reduced if the problem is to be computationally tractable. We lowered the dimensionality of the space by (1) down-sampling the waveforms to 8 kHz, (2) converting the waveforms to spectrograms, eliminating most phase information, and, most importantly, (3) performing a principal components analysis on the entire set of spectrograms corresponding to the tokens of each contrastive phoneme pair. For the purposes of data reduction, we retained in the analyses only those principal components that collectively explained 90% of the variance for each of the three pair-wise contrasts. These numbered 22 for the stop consonants, 17 for the vowels, and 13 for the fricatives. Each phoneme category was described as a multivariate Gaussian distribution within the relatively low dimensional PCA space, with the means and covariance matrices of these

distributions being fitted to the data points according to a maximum likelihood criterion.

Our implementation of the ideal classifiers for the three contrastive phoneme pairs was constrained to resemble biological systems in having internal noise. The question then became: what distribution of internal noise across the perceptual space is optimal for correct phoneme categorization of the speech tokens in our sample? It can be shown that categorization performance tends to improve if the perceptual space is stretched in the vicinity of the category boundary and contracted elsewhere, leaving overall perceptual distance the same. Such a warping of the perceptual space was achieved in our simulations by reducing the internal noise level between contrastive phoneme categories and raising it within categories, leaving the total amount of internal noise constant. Even with the reduced dimensionality of the PCA space, searching for the optimal internal noise profile for accurate categorization performance would pose intractable problems without additional constraints. Accordingly, for each contrastive phoneme pair and for each principal component of the derived perceptual space, we assumed that there was exactly one localized dip in the internal noise and that the shape of the dip was an inverted Gaussian. The search task was thus reduced to finding, for all principal components, the positions (means) and widths (variances) of the noise dips that maximized correct phoneme categorization.

To estimate the performance of the classifier for a given internal noise profile, the following steps were carried out. First, we assumed that the Gaussian distributions fitted to the phoneme categories were equivalent to the actual population distributions in American English. This allowed us to sample a much larger number of tokens from each phoneme category than we had originally sampled from the real speech data of the TIMIT corpus. In the simulations, we randomly selected 1000 new pseudo-samples—called Reference Samples—from the Gaussian distribution of each phoneme category. We next randomly selected another 1000 pseudo-samples—called Test Samples—from each phoneme category, and added an internal noise to each sample according to its location in the PCA space (i.e., an internal nose with lower variance was added to Test Samples that lie close to the dip in the internal noise profile). Each Test Sample was then assigned to a phoneme category on the basis of the category identities of the three nearest Reference Samples. This procedure allowed us to estimate, for a given profile of internal noise, the confusion matrix for the phoneme categorization task and hence the expected proportion of correct classifications.

Although the ideal classifier is not necessarily a model of the real observer, it is nevertheless worth asking whether one of the central assumptions behind our derivation of the classifier is biologically realistic. Concerning the use of principal components analysis to define the perceptual space, a number of theorists have argued that the demand for efficient neural coding of signals suggests that some form of factor analysis (e.g., PCA or independent components analysis) may be implicit in the design of sensory/perceptual systems [8, 18, 33].

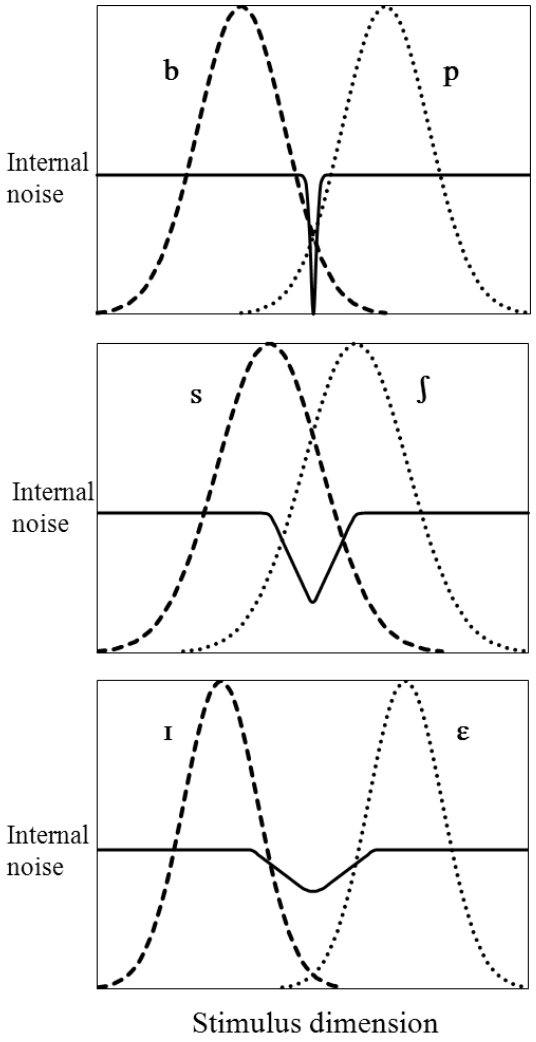
The search for the optimal internal noise profile for each pair-wise categorization task was conducted using simulated annealing, a probabilistic algorithm for global optimization problems [1]. In order to realize this algorithm, we first defined a cost function for the possible solutions, -log (PC), where PC is the expected proportion of correct classifications. The algorithm searches for the solution that minimizes the cost. Simulated annealing is a form of gradient descent with the added feature of "cooling." The algorithm attempts to move downhill along the cost function but is forced at the same time to jump around randomly, especially in the early stages of the process when the temperature is "hot." The "heat" allows the search process to jump out of local minimum traps in order to find the global minimum. In our applications, the complexity of the annealing algorithm scales linearly with the number of Gaussian dips used to locally stretch the perceptual space and with the number of principal components used to represent the phonemes. In order to reduce the computation time, we ran the simulations in stages, first finding the optimal noise profiles for each of the three phoneme contrasts separately, and then combining the dimensions of these three spaces to form another, higher-dimensional, space (which was re-orthogonalized using the Graham-Schmidt technique). By combining the spaces in this way, we were able to apply a common metric in comparing the optimal internal noise dip parameters (location and width) for the three pair-wise contrasts.

The optimal dip parameters were found to depend on the statistical properties of the phoneme

categories. Unsurprisingly, the optimal dip location was aligned with the point along each PCA dimension that maximized the separation of the contrasting categories. What is of greater interest is that the optimal dip width and depth for any dimension was strongly related to the variances of the contrasting phoneme distributions and their mean separation. Specifically, the optimal dip first became narrower and deeper (i.e., the variance of the inverted Gaussian decreased) as the degree of category overlap increased along the straight lines connecting the pair-wise category means in the combined PCA space, but as the degree of category overlap increased further beyond a certain critical point, the optimal dip became wider and shallower. Degree of category overlap along the straight lines connecting the pair-wise category means in the combined PCA space was relatively low for the vowels /ɪ/ and /ɛ/, intermediate for the stop consonants /b/ and /p/, and relatively high for the fricatives /s/ and /ʃ/. This gave rise to optimal noise dips (along the same straight lines) that were narrow and deep for the stop consonants, somewhat wider and shallower for the fricatives, and even wider and shallower for the vowels. (See Figure 1.) This means that the perceptual space of the ideal classifier was stretched more—and hence there was a greater increase in between-category discriminability relative to within-category discriminability for the stops than for the fricatives and for the fricatives than for the vowels.

The above ordering is intriguingly parallel to that reported for human listeners in studies of categorical versus continuous perception of speech sounds [9, 10, 19, 31]. A phonetic dimension that spans at least two phoneme categories is said to be perceived categorically if the phoneme labeling responses show an abrupt cross-over from one category to another and if discrimination performance for pairs of neighboring stimuli is near chance within a phoneme category but highly accurate when the stimulus pair straddles a category boundary. Perception of a phonetic dimension is described as non-categorical, or continuous, if the phoneme labeling function shows a gradual cross-over between categories and if discrimination performance for within-category stimulus pairs is comparable to that for between-category pairs. In general, human listeners tend to exhibit strong tendencies toward categorical perception for stop consonants, tendencies toward continuous perception for vowels, and intermediate results for fricatives [31].

Figure 1: Optimal noise profiles.



This variation in tendencies toward categorical perception has been explained in several different ways. For example, according to the speech mode hypothesis [19] stop consonants are among the most "encoded" speech sounds (i.e., they show a high degree of "restructuring," or context-dependent variation), vowels are among the least encoded sounds, and fricatives are somewhere in between. The special speech decoder, or speech mode, is assumed to be engaged to a greater extent for encoded than unencoded speech sounds, and because categorical perception is viewed as diagnostic of perception in the speech mode, the observed variation in tendencies toward categorical perception is expected.

In view of our simulation results, another possible account is suggested. Variation in tendencies toward categorical perception may

reflect not only intrinsic properties of speech sounds per se (e.g., encodedness) but also distributional properties of speech sounds (e.g., degree of category overlap) along with a listener strategy to try to optimize categorization performance with respect to those distributional properties.

6. CONCLUSION

The robustness of speech perception is partly explained by the tendency of language communities to combine phonetic properties that are mutually enhancing auditorily so as to achieve perceptual distinctiveness of phonological contrasts. It may also be partly the result of an implicit strategy of listeners to optimize categorization performance with respect to the statistical properties of phoneme categories. The latter hypothesis is promising and worthy of further test.

7. REFERENCES

[1] Das, A., Chakrabarti, B.K. (eds.). 2005. *Quantum Annealing and Related Optimization Methods. Lectures in Physics* vol. 679. Heidelberg: Springer.

[2] Diehl, R.L. 2008. Acoustic and auditory phonetics: The adaptive design of speech sound systems. *Phil. Trans. Royal Soc. London* B 363, 965-978.

[3] Diehl, R.L., Kingston, J., Castleman, W.A. 1995. On the internal perceptual structure of phonological features: The [voice] distinction. *J. Acoust. Soc. Am*. 97, 3333 (Abstract).

[4] Diehl, R.L., Kluender, K.R. 1989a. On the objects of speech perception. *Ecological Psychol*. 1, 121-144.

[5] Diehl, R.L., Kluender, K.R. 1989b. Reply to commentators. *Ecological Psychol*. 1, 195-225.

[6] Diehl, R.L, Lindblom, B., Creeger, C.P. 2003. Increasing realism of auditory representations yields further insights into vowel phonetics. *Proc. 15th ICPhS* Barcelona, 2, 1381-1384.

[7] Fant, G. 1960. *Acoustic Theory of Speech Production*. The Hague: Mouton.

[8] Field, D.J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am*. A 4, 2379-2394.

[9] Fry, D.B., Abramson, A.S., Eimas, P.D., Liberman, A.M. 1962. The identification and discrimination of synthetic vowels. *Lang. Speech* 5, 171-189.

[10] Fujisaki, H., Kawashima, T. 1970. Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo* 29, 207-214.

[11] Garner, W.R. 1974. *The Processing of Information and Structure*. Potomac, M.D.: Erlbaum.

[12] Geisler, W.S., Diehl, R.L. 2002. Bayesian natural selection and the evolution of perceptual systems. *Phil. Trans. Royal Soc. London* B 357, 419-448.

[13] Geisler, W.S., Diehl, R.L. 2003. A Bayesian approach to the evolution of perceptual and cognitive systems. *Cogn. Sci*. 27, 379-402.

[14] Jakobson, R. 1941. *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala, Sweden: Uppsala Universitets Arsskrift.

[15] Javkin, H.R. 1976. The perceptual basis of vowel duration differences associated with the voiced/voiceless distinction. *Report of the Phonology Laboratory* 1, 78-93, University of California, Berkeley.

[16] Kingston, J., Diehl, R.L. 1994. Phonetic knowledge. *Language* 70, 419-454.

[17] Kingston, J., Diehl, R.L., Kirk, C.J., Castleman, W.A. 2008. On the internal perceptual structure of distinctive features: The [voice] contrast. *J. Phonetics* 36, 28-54.

[18] Lewicki, M.S. 2002. Efficient coding of natural sounds. *Nature Neuroscience* 5, 356-363.

[19] Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychol. Rev*. 74, 431-461.

[20] Liljencrants, J., Lindblom, B. 1972. Numerical simulation of vowel quality systems. *Language* 48, 839-862.

[21] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H & H theory. In Hardcastle, W., Marchal, A. (eds.), *Speech Production and Speech Modeling*. Dordrecht: Kluwer, 403-439.

[22] Lisker, L. 1957. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language* 33, 42-49.

[23] Lisker, L. 1978. On buzzing the English /b/. *Haskins Laboratories Status Report on Speech Research*, SR-55/56, 181-188.

[24] MacNeilage, P.R. 1969. A note on the relation between tongue elevation and glottal elevation in vowels. *Monthly Internal Memorandum*, University of California, Berkeley, 9-26.

[25] Marr, D. 1982. *Vision*. New York: W.H. Freeman.

[26] Martinet, A. 1955. *Économie des Changements Phonétiques*. Berne: Francke.

[27] Miller, G.A., Nicely, P.E. 1955. An analysis of perceptual confusions among some consonants. *J. Acoust. Soc. Am*. 27, 338-352.

[28] Parker, E.M., Diehl, R.L., Kluender, K.R. 1986. Trading relations in speech and nonspeech. *Percept. Psychophys.* 39, 129-142.

[29] Passy, P. 1890. *Études sur les Changements Phonétiques et Leurs Caractères Géneraux*. Paris: Firmin-Didot.

[30] Port, R.F., Dalby, J. 1982. Consonant/vowel ratio as a cue for voicing in English. *Percept. Psychophys*. 32, 141-152.

[31] Repp, B.H. 1984. Categorical perception: Issues, methods, and findings. In Lass, N.J. (ed.), *Speech and Language: Advances in Basic Research and Practice*. vol. 10. New York: Academic Press, 243-335.

[32] Riordan, C.J. 1977. Control of vocal-tract length in speech. *J. Acoust. Soc. Am*. 63, 998-1002.

[33] Simoncelli, E.P., Olshausen, B.A. 2001. Natural image statistics and neural representation. *Annual Review of Neuroscience* 24, 1193-1215.

[34] Stevens, K.N., Blumstein, S.E. 1981. The search for invariant acoustic correlates of phonetic features. In Eimas, P.D., Miller, J.L. (eds.), *Perspectives on the Study of Speech*. Hillsdale, N.J.: Erlbaum, 1-38.