# Locating perceptual category centers in multi-dimensional stimulus spaces

Eric Oglesbee
Indiana University, Department of Linguistics

Kenneth de Jong
Indiana University, Department of Linguistics

---

**Abstract**

Examining phonetic categorization in multi-dimensional stimulus spaces poses a number of practical problems. The traditional method of forced identification of an entire stimulus space becomes prohibitive when the number and size of stimulus dimensions becomes increasingly large. In response to this, Iverson and Evans (2003) proposed an adaptive tracking algorithm for finding best exemplars of vowels in a multi-dimensional space. Their algorithm converged on best exemplars in a relatively small number of trials; however, the search method took advantage of special properties of the vowel space in order to achieve rapid convergence. In this paper, a more general multi-dimensional search algorithm is proposed and analyzed for inherent biases. Results showed that there are no long-term biases in the search method, and that multiple types of useful data are generated. The proposed search method appears to be a viable approach for generating a first approximation of phonetic categorization in multi-dimensional stimulus spaces.

## I. INTRODUCTION

Perceptual experiments in phonetic categorization utilizing stimulus continua typically involve adjusting one or two acoustic parameters. For example, studies examining labial stop categorization often only manipulate Voice Onset Time (VOT) (Lisker & Abramson, 1970), while vowel space categorization experiments are usually restricted to simultaneous manipulations of two formants (Hillenbrand & Gayvert (1993), Molis (2005)). In most perceptual categorization experiments low-dimensional explorations such as these are to be expected for two reasons. First, varying a small number of acoustic parameters is often sufficient for inducing robust shifts in categorization of stimuli. Second, the common method for mapping perceptual category boundaries within low-dimensional spaces (i.e. multiple forced identifications of an entire stimulus space) becomes unwieldy when the number of stimuli to be classified becomes increasingly large. In the past, the ability of low-dimensional spaces to capture categorical contrasts makes the inadequacy of the traditional investigation method for exploring higher-dimensional stimulus spaces a moot point. However, it can be shown that there is good cause for needing to examine phonetic categorization in high-dimensional spaces, which consequently means that the methodological hurdle embodied in the second reason needs to be addressed.

A strong motivation for using high-dimensional acoustic spaces in studying phonetic categorization arises from VOT experiments conducted in languages other than English. While VOT can be used to distinguish labial stops in English in both production (Lisker & Abramson, 1970) and perception (Caramazza, Yeni-Komshian, Zurif, & Carbone, 1973), additional acoustic cues appear to be necessary in other languages. In Korean, production studies have shown substantial overlap in the VOT distribution of fortis, lenis, and aspirated labial stops (Lisker &

Abramson, 1964), suggesting that VOT alone is not sufficient for resolving the 3-way contrast. In a study of VOT in French by Caramazza et. al. (1973), although VOT production distributions were separable for labial stops, a peculiarity was observed in the classification portion of the experiment. Unlike the native English speaking participants, the native French speakers had a relatively flat categorization function for a prolonged stretch in the crossover region that contained a pronounced "dip" at short lag VOT values. These features of the categorization function indicate that VOT may interact with some other component of the signal with respect to labial stop classification. These two examples suggest that when looking at phonetic categorization cross-linguistically, acoustic dimensions which are sufficient for capturing a linguistic contrast within one language may be inadequate for other languages. The implication of this observation is that it now appears that the development of practical techniques for studying phonetic categorization in multi-dimensional stimulus spaces is necessary.

As mentioned above, previously used methods for studying low-dimensional spaces do not scale up to high-dimensional spaces very well. To support this claim, a description of previously used methods for one-, two-, and five-dimensional stimulus spaces is given below, along with reasons for why none of these methods are appropriate for application to general n-dimensional stimulus spaces.

In the one-dimension approach a set of stimuli is generated where a single variable (or a group of covarying variables) is manipulated in order to generate a continuum. The stimuli are randomized and each is presented to a listener multiple times. During each stimulus presentation, the listener is asked to explicitly associate the stimulus with a label. An example of this is a continuum varying VOT, where the listener has to decide whether the consonant they heard was /p/ or /b/. By having multiple judgments of each stimulus item, an identification function can be

plotted corresponding to the probability that the listener identified a particular stimulus as belonging to one category or another. A major advantage of the one-dimensional approach is that a high resolution picture of listener behavior can be obtained in relatively few trials.

There are two primary ways the categorization of two-dimensional stimulus arrays has been approached. One way is to do a forced identification of the entire stimulus space in the same manner as described above (e.g. Kitahara, 2001). Another approach (Johnson, Flemming & Wright (1993)) is to employ a listener-directed search through the space for a best exemplar. Listeners are presented with a two-dimensional grid of buttons where each button corresponds to a single stimulus, and they are given a label and told to find the best instance of the label in the two-dimensional space. Calling this approach a "Method of Adjustment (MOA)" task, Harnsberger, J.D., Svirsky, M.A., Kaiser A.R., Pisoni D.B., Wright R., & Meyer T.A. (2001) used this strategy to study perceptual vowel spaces of cochlear implant users. One substantial advantage of this approach is that a large two-dimensional space can be tested in a short period of time; however, since the search is completely listener driven there is substantial loss of control by the experimenter.

Anything above a two-dimensional space presents a number of difficulties for the two methods discussed above. First, the issue of too many stimuli for forced identification becomes compounded even further, as is estimated in Table I, assuming 10 stimuli per dimension, 30 judgments per stimulus, and 2 seconds per response.

| # Dimensions | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| # Trials | 30*10 = 300 | 30 * 10 * 10 = 3000 | 30 * 10 * 10 * 10 = 30,000 | 30 * 10 * 10 * 10 * 10 = 300,000 | 30 * 10 * 10 * 10 * 10 * 10 = 3,000,000 |
| Total Time | 10 min | 1.67 hours | 16.7 hours | 6.94 days | 69.44 days |

**Table I.** Estimates of protocol duration for multi-dimensional stimulus arrays.

Second, a two-dimensional listener-directed search has difficulty scaling up to higher dimensions given the limitation of only being able to present two variable dimensions at one time. While this may be manageable in a three-dimensional case by having listeners search through a large number of two-dimensional spaces, in most cases this approach becomes extremely cumbersome if the smallest dimension contains more than three or four steps. The situation is even worse if the number of dimensions is greater than three.

In response to these types of scalability difficulties, a novel method for higher dimensional searches in an acoustic vowel space was proposed by Iverson and Evans (2003). In their experiment, convergence on best exemplars was achieved by eliciting goodness judgments for selected stimuli located on a single vector in the space. These goodness judgments were then used as inputs to an algorithm which was designed to output the optimal point on a current search vector. Then, using the point derived by the algorithm, a new search vector was chosen containing the derived point, but which probed a different part of the multi-dimensional space. A total of seven search vectors per vowel category were used to search five dimensions.

The approach described in Iverson & Evans (2003) is essentially a hybrid of the one- and two-dimensional methods discussed earlier. The experimenter has greater control over what

stimuli are played via the search algorithm design; however, movement through the space is driven by the listener's goodness judgments. Essentially, this approach is analogous an optometrist converging on a best prescription for a patient by using the patient's responses to different subsets of lenses in order to identify the general type of correction the patient needs. The advantage of this approach is that it is possible to find best exemplars in large stimulus spaces with a relatively small number of trials. Unfortunately, the algorithm presented by Iverson & Evans (2003) lacks generality in that rapid convergence is achieved by taking advantage of special relationships between the space's dimensions; i.e. the functioning of the algorithm is intertwined with properties of the stimulus space, and therefore inappropriate for other multi-dimensional spaces investigating other categories. The implication of this is that a general methodology for practically exploring n-dimensional spaces still has not been demonstrated.

Two experiments are presented in this paper. In the first experiment, a generalized n-dimensional search method for examining phonetic categorization is proposed. Properties of the search algorithm are analyzed using computer simulations, and the types of data generated by the search method are discussed. In the second experiment, the search algorithm is implemented within a five-dimensional acoustic space designed to test labial stop categorization. In particular, /p/ and /b/ categorization by native speakers of English is examined. Results from the two experiments suggest that the proposed search method can be used to gain a first approximation of factors affecting categorization in multi-dimensional stimulus spaces.

## II. EXPERIMENT 1

### A. Methods

#### *1. Search Algorithm Design*

Although the search method used in Iverson & Evans (2003) was not general enough to be applied to other stimulus spaces, there were a number of features which were adopted for use in the algorithm proposed here. Therefore, in the following description of the proposed algorithm, differences with respect to Iverson & Evans (2003) are briefly discussed. The proposed search algorithm operates using six basic principles:

1. *Search vectors are constrained to a single dimension.*

The term "search vector" refers to the subset of the stimulus space that is actively being probed by the search algorithm. In Iverson & Evans (2003), all of the search vectors (with the exception of the one probing duration) cut through the stimulus space at an angle, meaning that multiple stimulus variables were being manipulated at the same time. In the method proposed here, search vectors are contained within a single dimension. This means that only a single stimulus variable is varied, while all other stimulus variables are held constant. For example, in the case of a two-dimensional vowel space where F1 and F2 are stimulus variables, the search vector would vary either F1 or F2, but not both simultaneously.

2. *Best exemplars on each dimension are derived from goodness judgments taken from three points contained in the search vector (middle, left endpoint, right endpoint)*.

Once the search vector is defined for a particular dimension, listeners are played three stimuli from the search vector in the following order: middle point, left endpoint, and right endpoint. After each stimulus is played, listeners move a slider bar to indicate the goodness level

of the stimulus relative to the target category. The slider bar records a goodness value between .01 and 1.0. These three goodness judgments are then used to estimate where on the search vector the "best" exemplar should be located. In principle there are any number of possible decision rules which can be employed to select the "best" exemplar. In the experiments reported here the ratio of the two highest goodness judgments (highest over second highest), modulated by a decreasing exponential (shown in Fig. 1), determined where the "best" exemplar should be located. This location was expressed as a proportion of the distance between the two stimuli with the highest goodness judgments as measured from the highest rated stimulus. Modulating the goodness judgment ratio by the decreasing exponential in Fig. 1 has a major advantage. It effectively weights the choice of the "best" exemplar towards the stimulus with the highest goodness judgment in a way that is stronger than a simple ratio. For example, if a straight ratio was being used and the goodness ratings were in a 2:1 relationship, the location of the derived "best" point would be 33% of the distance from the highest rated point to the second highest rated point. However, using the function in Fig. 1 to modulate the ratio, the derived "best" point is located at 10% of the distance. This weighting is advantageous because it biases the selection of the derived point towards a location that the listener has already identified as being the best of the three stimuli that have been presented, as well as encourages movement through the space.

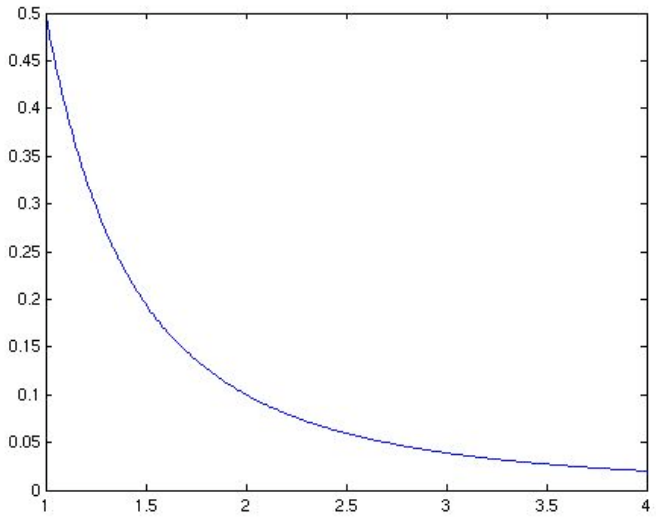


**Fig. 1.** Decreasing exponential function used to modulate goodness ratios. The function is $X(r) = .5 * r^{-2.3219}$, where $r = g_1/g_2$ , $g_1$ is the highest goodness judgment, and $g_2$ is the second-highest goodness judgment. The goodness ratio "r" is shown on the horizontal axis, and the distance of the "best" exemplar from the stimulus associated with goodness judgment $g_1$ is given as a percentage of the distance between the stimuli associated with $g_1$ and $g_2$.

*3. The slider bar used for goodness judgments is not reset between stimulus presentations.*

This is an important feature for anchoring the goodness judgments of listeners since an absolute scale is not used. The implications of this design element are given later in discussion sections 2c and 2d.

4. *Only a subset of a stimulus dimension is searched at a given time.*

Iverson & Evans (2003) allowed search vectors to span the full length of the stimulus dimensions. In contrast to this, search vectors in the method described here contain at most 2/3 of a stimulus dimension. In some cases, however, less than 2/3 of a dimension is probed. This occurs when the middle point of a search vector lies close enough to the endpoint of a dimension

that the search vector becomes compressed. Because the proposed search algorithm is iterative in nature (as described below), searching only 2/3 of a stimulus dimension results in two desirable outcomes. First, listeners are exposed to a large variety of stimuli. This is accomplished since it is highly likely that when a dimension is tested multiple times the search vector will be shifted to different locations, resulting in listeners being exposed to a large number of different stimuli. Second, keeping the search vector fairly large without spanning the full length of a stimulus dimension helps guarantee that there will be a dynamic range in the goodness judgments of the three test stimuli. For example, if the length of a search vector was small relative to the stimulus dimension (i.e. 30% to 40%), there is a good chance that the three points tested on the search vector may not completely leave the category being tested. On the other hand, if the search vector was equal to the length of the entire stimulus dimension, then two of the three search vector points would be the same on every iteration (i.e. the stimulus dimension endpoints). Having an intermediate search vector length gives good dynamic range in goodness judgments, while also making it possible for a number of different stimuli to be judged by the listener.

5. *All dimensions are probed before a dimension is probed again.*

Every dimension is given the chance to be "tuned" before it is tested again. Whether or not dimensions are probed in the same fixed order is optional.

6. *Multiple iterations of the search process are used to achieve convergence*.

The primary principle behind the proposed search method is that incremental progress made in each individual dimension aids global convergence on a best exemplar. By iterating through all of the dimensions multiple times, listeners are given multiple opportunities to steer

the algorithm into the proper portion of the space. Although more efficient approaches are possible, the transparency and robustness of the proposed method are believed to offset inefficiencies. Figure 2 contains a graphical representation of what the search algorithm would look like in a two-dimensional stimulus space.

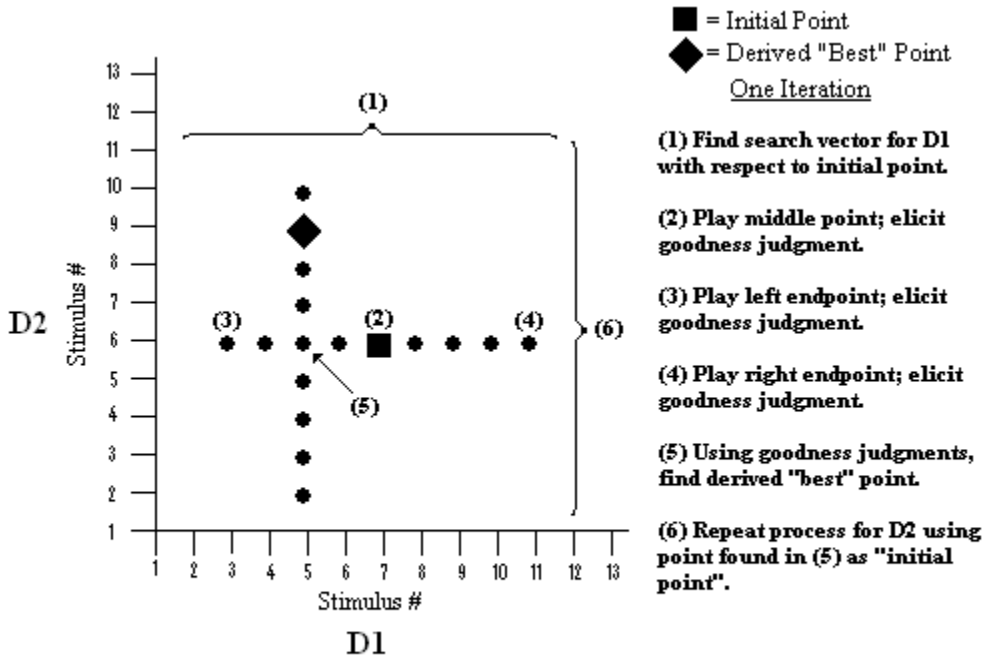


**Fig. 2.** Example of a single iteration of the search algorithm procedure in a two-dimensional space. The initial point is shown with a filled square, and the destination of the search algorithm after sampling dimensions D1 and D2 is shown with a filled diamond.

*2. Algorithm Test Procedure*

An important question with any automated procedure for sampling a stimulus space is whether or not there are features of the search procedure which generate a systematic bias towards or away from specific points in the space. The answer to this question has ramifications for deciding whether or not effects observed in real data are a product of listener or task properties.

One way to explore this concern with the proposed search method is to feed randomly generated goodness ratings into the algorithm and observe the probability distribution of the outputs. This approach makes it possible to locate potential "dead" regions (i.e. regions that have a systematically low probability of being final destination points), as well as attractors within the space. Although the search method proposed here is intended to be applied to multi-dimensional searches, given the independence of random goodness judgments from each other, we can use single dimension searches as a basis upon which to evaluate the algorithm. This is possible because in essence, the n-dimensional case is simply the product of n single dimension searches.

There are three main factors which should affect the performance characteristics of the algorithm:

**(1) Dimension size** (how many steps are used for each stimulus variable)

**(2) Initial starting point** (which member of a stimulus dimension is presented first)

**(3) Number of iterations** (how many times each dimension is sampled)

Factor (1) is particularly important for analyzing the impact of edge effects, as well as the internal structure of the space. Factors (2) and (3) deal with the interaction between short-term and long-term biases within the space. If the number of iterations is very low, the location of the initial starting point should have an effect on the output probability distribution; however, as the number of iterations increases, the effects of the initial starting point should give way to the long-term effects of the algorithm itself.

A fourth variable which is important for this analysis, but which is not connected to the functioning of the algorithm itself, is the number of random trials used to generate the probability distribution. Essentially, the resolution of the probability distribution becomes more refined with

each additional random trial. Based on piloting, a sufficient number of random trials were chosen in order to precisely depict the contours of the probability space.

Two sets of simulations were run. The first set examined the effects of the three main factors given above, and the parameters for these simulations are given in Table II. The second set of simulations examined the shape of the probability distributions in a multi-dimensional setting using parameters (Table III) taken from an experiment designed to test the proposed multi-dimensional search method.

| | **Small Dimension Test** | **Large Dimension Test** |
|---|---|---|
| **Dimension size (number of steps)** | 5 | 17 |
| **Initial starting points** | 2,3,4 | 5,9,13 |
| **Number of iterations** | 1,3,5,10,20 | |
| # Trials | 50,000 | 17,000 |

**Table II.** Parameters for testing the properties of the search algorithm in a one-dimensional space.

| | |
|---|---|
| **# Dimensions** | 5 |
| **Dimension Sizes** | 4,5,11,14,17 |
| **Initial Points** | IP1 = [3 3 4 7 6]<br>IP2 = [3 3 4 7 11]<br>IP3 = [3 3 8 7 6]<br>IP4 = [3 3 8 7 11] |
| **# Iterations** | 3 |
| # Trials | 50,000 |

**Table III.** Parameters for testing the properties of the search algorithm in a five-dimensional space. These parameters correspond to a pilot experiment that was run to test the effectiveness of proposed search method.

B. Results

*1. One-dimension Simulations*

Figure 3 contains the results of the one-dimensional simulations run using the parameters from Table II. Each point represents the proportion of times that a particular stimulus number was the final destination point of the algorithm. The "expected" curve indicates the proportion level that should be observed if every stimulus was equally likely to be a destination point.

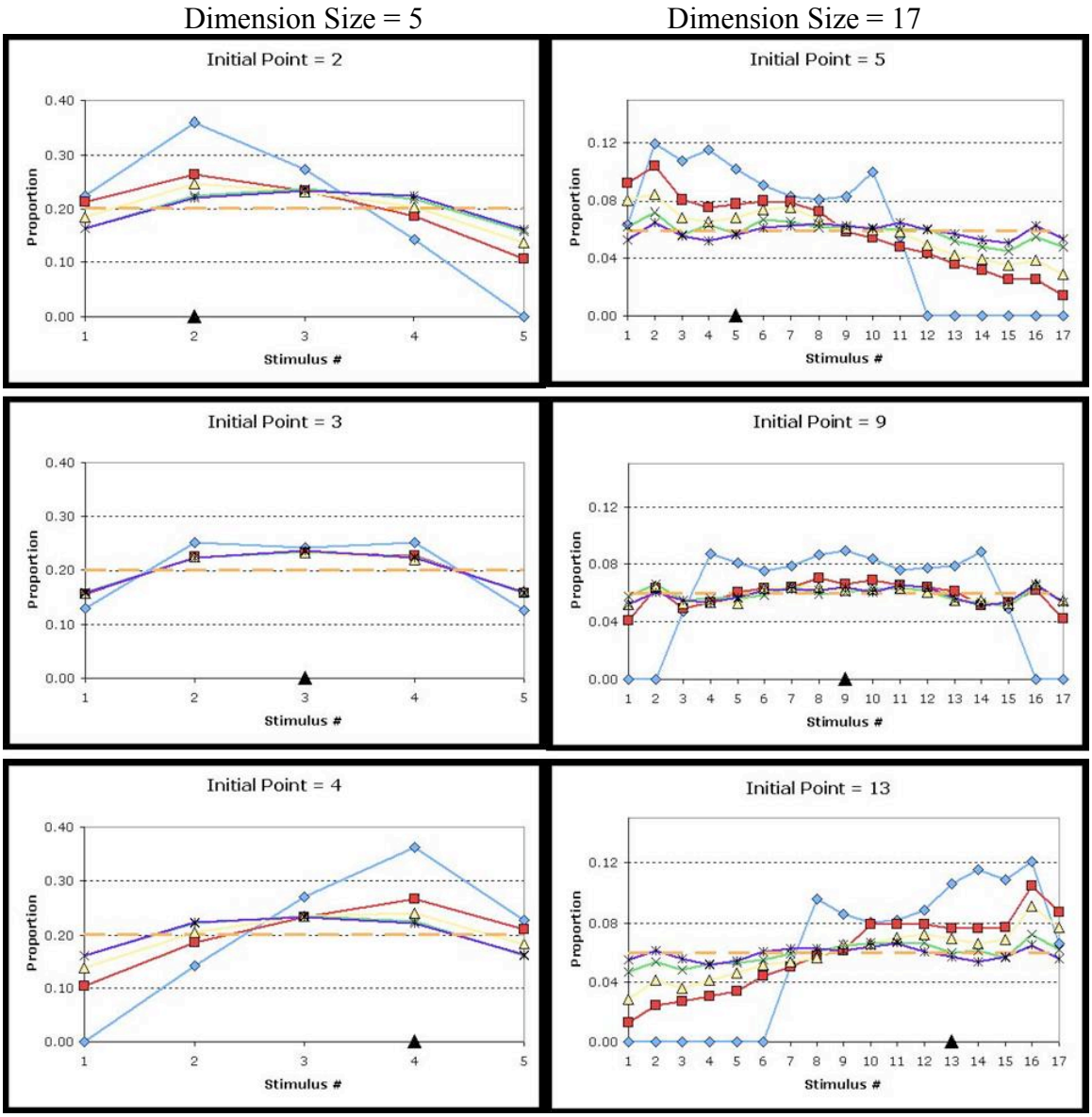


**Figure 3. Probability distribution through multiple iterations.** Final stimuli chosen by the algorithm using random goodness judgments. Panels in the left column correspond to a dimension size of five, while the panels in the right column correspond to a dimension size of 17. Panels in the top row have initial points towards the left end of the dimension, panels in the middle row have initial points in the middle, and panels in the bottom row have initial points towards the right end of the dimension. Each curve corresponds to a different number of iterations. The curve labeled "expected" illustrates what the distribution would look like if each stimulus was equally likely as a destination point.

*1a. Effect of dimension size*

The effect of dimension size is most noticeable in regards to boundary effects. Regardless of dimension size, when looking at the 20 iteration case the endpoints exhibit depressed probabilities relative to most other points in the space. If we view the "expected" curves in Fig. 3 as being the standard by which we measure the level of depression of the endpoint probabilities, we see that the depressed probability effect is more pronounced when the dimension size is small, as is summarized in Table IV. For example, the probability of ending up on the far left endpoint (stimulus #1) for the dimension with five elements is .16232. If of the stimuli on the dimension were equally likely, we would expect that the probability of ending up on the left endpoint of the dimension would be .2. If we take expected (.2) minus observed (.16232) and divide by the expected (.2), we find that the left endpoint has a destination probability that is 18.84% lower than what we would expect. Compare this with the 9.80% value obtained for the larger dimension.

| *20 Iterations* | **Left Endpoint** | | | **Right Endpoint** | | |
|---|---|---|---|---|---|---|
| | Left IP | Mid IP | Right IP | Left IP | Mid IP | Right IP |
| **Dimension Size = 5** | 18.84% | 20.45% | 20.21% | 19.90% | 20.20% | 18.82% |
| **Dimension Size= 17** | 9.80% | 11.10% | 6.70% | 8.20% | 9.20% | 5.70% |

**Table IV.** Relative percentage below the random probability for left and right endpoints. "IP" stands for "Initial Point".

*1b. Effect of initial point*

As expected, if the number of iterations is small, the location of the initial point has a profound effect on the probability distribution. Given that a search can span at most 67% of the space, some points are simply inaccessible during the first iteration.

*1c. Effect of number of iterations*

As the number of iterations increases, the probability functions begin to flatten out, regardless of the initial point. After 20 iterations, although a very small ripple is still present, there are not any true dead spaces or extremely strong attractors. In the cases where the initial point is located at the center of the space, convergence to the long-term shape of the probability distribution occurs after only three iterations. In the cases where the initial points are offset from the center, convergence does not appear to be occurring until after approximately 10 iterations. Very little difference is observed between 10 and 20 iterations.

*2. Five-dimension simulations*

Examining the probability distribution of a multi-dimensional space presents some difficulties. Once the number of dimensions is greater than two, it is not possible to efficiently display interactions between all of the dimensions. Therefore, for this paper, interaction effects in multi-dimensional spaces will be put aside, and only marginal probabilities will be reported. In order to calculate these probabilities, each dimension is viewed individually with data collapsed across all of the other dimensions. The result of this analysis is shown in Fig. 4. It is immediately apparent that these results are nearly identical to those found in the one-dimensional simulations reported above.

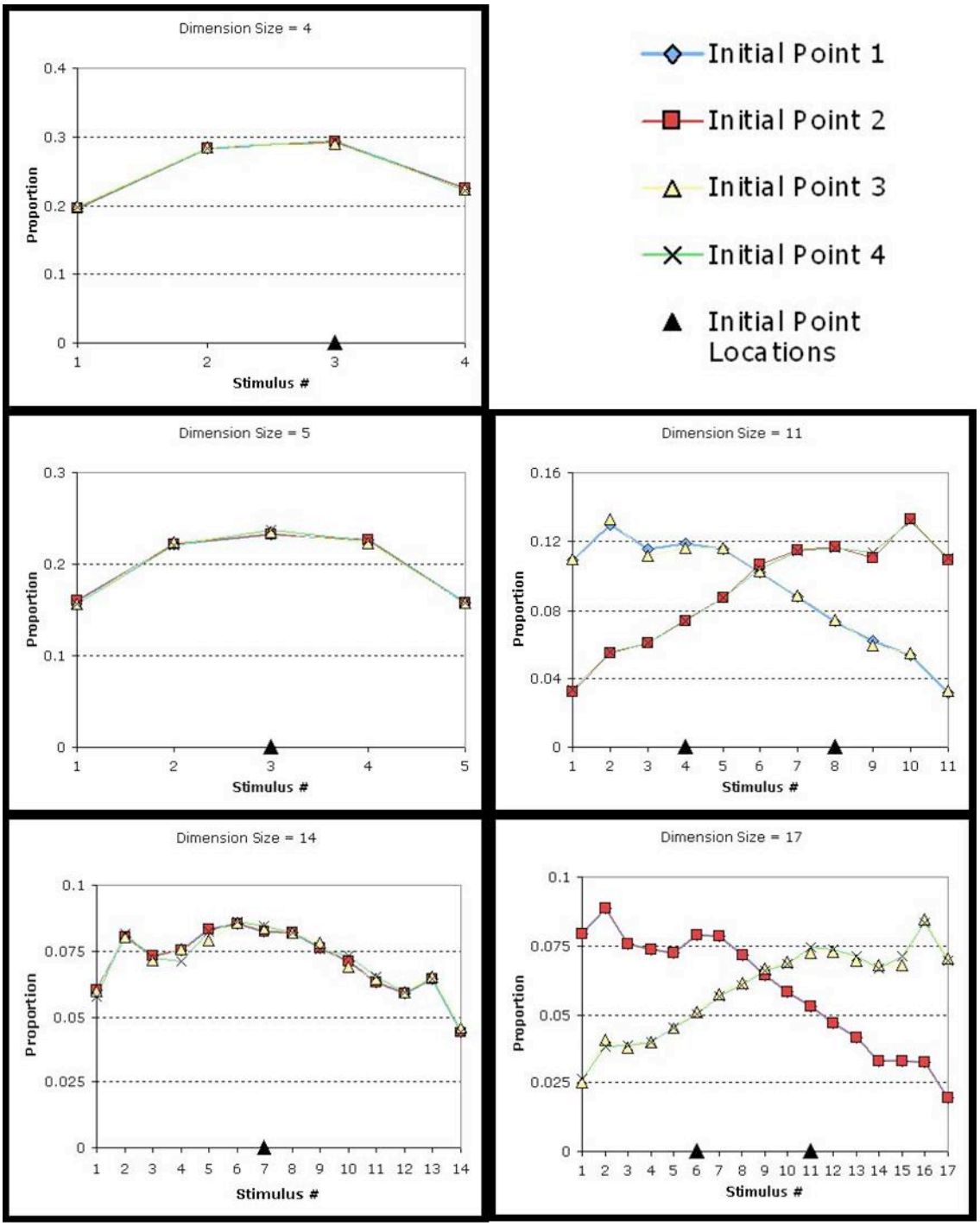


**Figure 4**. Probabilities of stimuli being the destination point of the search algorithm using random goodness judgment in a five-dimensional space. Each curve corresponds to a different initial point (see Table III). Each dimension is shown with results collapsed across all other dimensions.

## C. Discussion

### *1. Search algorithm biases*

Three types of bias were observed in the results section: depressed dimension endpoint probabilities, ripple in the probability function, and initial point effects. Although the endpoints of the stimulus dimensions have slightly depressed probabilities relative to the other stimuli, the difference is not great enough to cause concern. In fact, slightly depressed endpoint probabilities are a desirable feature in that listeners are less likely to get caught within the edges of the stimulus space. The ripple in the probability function was also fairly small, and consequently will not be discussed further. Depending on the number of iterations, the location of the initial starting point of the search process appears to be the most substantial source of bias in the algorithm.

It is not surprising that different initial points may bias the outcome of the search algorithm. When the number of iterations is relatively low, the area surrounding the initial point has a heightened probability; however, as the number of iterations increases the impact of the initial starting point is lessened. According to the simulations, the most straightforward way to minimize initial point bias is to have the initial starting point of the search be at the center of every dimension. In cases where it is not desirable to always start from the same initial point each time the algorithm is run, the next best solution is to pick initial points that are symmetrical about the center of each dimension. By doing this, it should be possible to tease apart real subject effects from effects that are due to initial point bias.

Initial point bias does not have to be a confound; it can actually be used as a part of a research protocol. An advantage to initial point bias is that it can be used as a way to test the robustness of effects within a particular dimension. By having an initial point on the opposite end

of where best exemplars are expected, convergence effects should be magnified. Essentially, if listeners consistently converge to a low-probability region in the stimulus space, then it is reasonable to trust that the observed convergence is not due to the algorithm.

*2. Types of data generated by the algorithm*

*2a. Destination point of the multi-dimensional search*

After all of the prescribed iterations have taken place, the final "best" point as derived by the search algorithm should be located near the center of the category being tested, assuming the stimulus space was appropriately designed and sufficient iterations through the stimulus dimensions were run. Given that this point was derived from an adaptive tracking procedure that relied on listener-provided goodness judgments, results from a single block should not be considered definitive. In order to allow for noisiness in listener behavior, as well as listener interactions with the search algorithm, the results from multiple blocks should be used to develop a candidate region for best exemplars.

*2b. Multi-dimensional trajectories*

In the cases where the number of iterations is greater than one, it is possible to examine how listeners move through the stimulus space. This is accomplished by plotting the location of the current "best" point that has been derived after a full iteration has taken place. For example, in a protocol containing three iterations there would be four points that could be used to examine the listener's trajectory in the stimulus space: initial point, point derived after first iteration, point derived after second iteration, and point derived after third iteration (i.e. destination point). This provides more information than a simple destination point; by looking across iterations it is

possible to see if the algorithm has locked into a stable orbit about a point within the stimulus space, or if the movements through the space appear erratic and nonsensical. One difficulty though with processing this sort of data is that multi-dimensional trajectories are difficult to efficiently display and analyze.

*2c. Algorithm tracking performance*

Given the design of the algorithm, it is possible to determine if the adaptive track is steering listeners into regions of the stimulus space where "best" exemplars are likely to be found. When a "best" point has been identified after probing a dimension (see point (5) in Fig.2), the first point that is probed on the next search dimension is this same "best" point (see (6) in Fig.2). Since the slider bar used to record goodness judgments is not reset between stimulus presentations (see the third point in the "Search Algorithm Design" section), a direct comparison of the goodness judgments of these two points is possible. If the search algorithm is steering the listener into better exemplars, the second of these two points (i.e. the derived "best" point) should have a goodness judgment that is either the same or higher than the previously tested point. By pooling these results over a large number of trials, it is possible to see the degree to which the search algorithm promotes convergence towards best exemplars. If the number of blocks is large enough, this approach could also be used to test whether the algorithm is tracking properly within individual dimensions. This data could be used to help decide whether specific dimensions are actually relevant for capturing a specific linguistic contrast.

*2d. Dimension response tendencies*

Another type of available data from the algorithm is which end of a dimension listeners seem to prefer with respect to the linguistic category being probed. During the search process, the endpoints of the search vector are played one right after the other. Once again, since the slider bar is not reset after each stimulus presentation, this allows for a direct comparison of the search vector endpoint stimuli. By looking across multiple trials, it is possible to see which end of the search vector listener's appeared to prefer when probing a specific linguistic category. If there is no endpoint preference for a given dimension (i.e. roughly half the time one endpoint was preferred as compared to the other), this would suggest that the dimension may not be important for the linguistic category being tested.

## III. EXPERIMENT 2

A. Method

*1. Stimuli*

A five-dimensional acoustic space was generated by varying properties known to affect the voicing categorization of initial stops. The following properties of recorded utterances of the syllable [ba] were systematically varied to produce the five-dimensional acoustic space:

(1) *Fundamental Frequency (5 levels)*

The syllable [ba] was produced by the author with multiple F0 levels. Five tokens with similar lengths were chosen for digital manipulation. The F0 values at the midpoints of the

tokens were 120, 141, 149, 153, and 163 Hz. All five tokens were normalized in amplitude according to the peak amplitude of the vowel.

(2) *Formant Transitions (4 levels)*

Previous research has shown that the property of F1-cutback is related to voicing identification, reflecting that fact that voiceless (aspirated) stops in English have part of the consonant-to-vowel transition removed when aspiration is removed from the signal. In order to adjust the amount of formant transition information present, individual pitch pulses from [ba] were removed at the beginning of the vowels using Wavesurfer. The number of pitch pulses removed ranged from one to four.

(3) *Vowel Onset Ramping (11 levels)*

The amplitude envelope of the vowel onset was modified using artificial contours modeled off of productions of the vowel [a] by native speakers of English, Korean, and Japanese. As with all of the dimensions, the goal was to develop a range of stimuli that would span phonetic categories across languages. The method for creating different levels of onset rampings was as follows. First, a normalized amplitude envelope was created for each token by calculating the proportion of the amplitude in the first half of the vowel relative to the vowel midpoint. Then, this information was used to "stamp" artificial contours onto the vowel. The original amplitude envelope of the vowel was rescaled using the following process. A 20 ms Hamming window was used to find the rms amplitude at 1 ms intervals beginning with the onset of the vowel and ending with the vowel midpoint. For each window, the obtained rms value was scaled to some proportion of the midpoint rms, where the proportion was determined by piecewise

functions like those in Fig. 5. The particular functions in Fig. 5 were determined as two-piece approximations of the amplitude profile found in a variety of [p] and [b] productions by Japanese, Korean, and English native speakers. It should be noted that the sharp elbow in the joint between the two functions is not evident in the final signal due to the use of overlapping analysis windows in the stamping procedure. The 11 levels were indexed by the ratio of the onset amplitude relative to the vowel midpoint amplitude, and ranged from 0 to 1 in increments of 0.1.

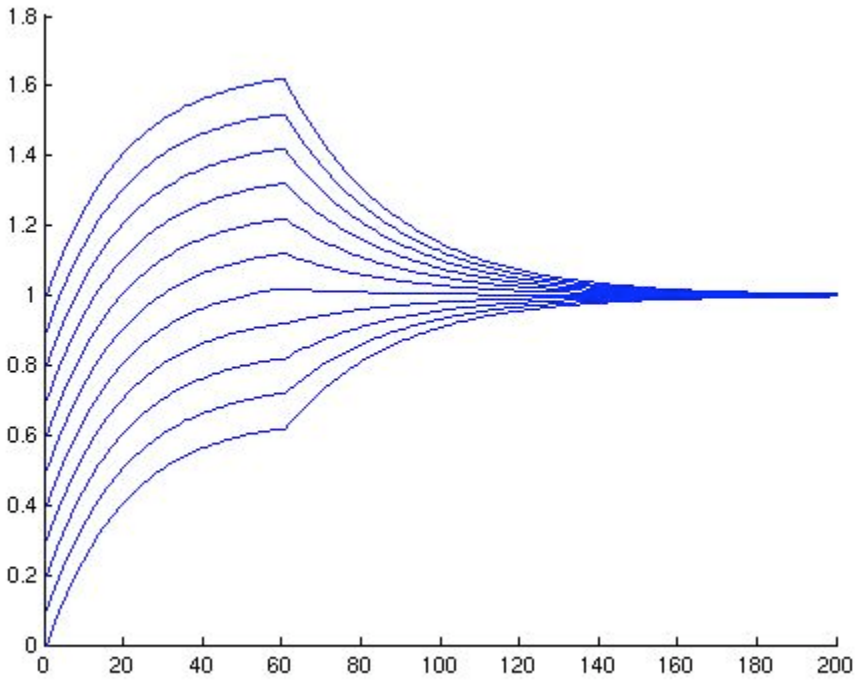


**Figure 5.** 11 amplitude profiles used to generate the onset ramp dimension of the stimulus space. These contours would have been used for the first half of a vowel (vowel onset is at "0"; vowel midpoint is at "200"). The horizontal axis indicates time, and the vertical axis indicates amplitude proportional to the vowel midpoint (i.e. in the top curve the amplitude at time = "0" is identical to the amplitude at the vowel midpoint).

(4) *Voice Onset Time (VOT) (17 levels)*

VOT was varied from -40 ms to 40 ms in steps of 5 ms.

(a) *Prevoicing*: All of the prevoicing portions were built off of recordings of the author producing a prevoiced [b] at approximately 100 Hz. A single pulse was isolated and appropriately amplitude scaled. This pulse was copied multiple times in order to get the desired prevoicing lengths. Since the production frequency was roughly 100 Hz, the cuts needed to make 5 ms increments occurred very close to a zero crossing.

(b) *Aspiration*: All of the aspiration was generated using white noise that had been spectrally shaped to match the quality of the vowel. This allowed for accurate control of aspiration lengths in creating the continua. The amplitude of the white noise was scaled to 33 dB (digital) prior to shaping.

After generating these components, they were either added into the signal (prevoicing) or spliced in following the burst release (aspiration) using Matlab.

(5) *Burst Strength (14 levels)*

The amplitude of the burst release was scaled according to the amplitude of the vowel midpoint in 14 equal steps. The range of values was 5% - 70%, with step sizes of 5%.

The parameter values for each dimension are summarized in Table V. The total number of stimuli in the space was 5*4*11*17*14 = 52,360 stimuli.

| Stim # | F0 (Hz) | Cutback (# pulses) | Onset Ramp (prop. V-mid) | VOT (ms) | Burst Amp. (prop. V-mid) |
|---|---|---|---|---|---|
| **1** | 120 | 1 | .00 | -40 | .05 |
| **2** | 141 | 2 | .10 | -35 | .10 |
| **3** | 149* | 3* | .20 | -30 | .15 |
| **4** | 153 | 4 | .30* | -25 | .20 |
| **5** | 163 | | .40 | -20 | .25 |
| **6** | | | .50 | -15* | .30 |
| **7** | | | .60 | -10 | .35* |
| **8** | | | .70* | -5 | .40 |
| **9** | | | .80 | 0 | .45 |
| **10** | | | .90 | 5 | .50 |
| **11** | | | 1.00 | 10* | .55 |
| **12** | | | | 15 | .60 |
| **13** | | | | 20 | .65 |
| **14** | | | | 25 | .70 |
| **15** | | | | 30 | |
| **16** | | | | 35 | |
| **17** | | | | 40 | |

**Table V.** Parameter values for each level of the five dimensions in the stimulus space. Entries with a "*" indicate values used in the four initial starting points.

*2. Search algorithm*

The search algorithm used in experiment was nearly identical to the algorithm reported in Experiment 1; however, following data collection, it was discovered that there was a bias towards left endpoints that was being caused by the use of a floor function when rounding was necessary. This error was not present in the code used to generate the simulation results in Experiment 1. Unfortunately, the error was not caught before all of the listeners in Experiment 2 were run. In order to allow for a fair comparison between simulation results and the human listener data, the simulations from Experiment 1 were rerun with the rounding error included. These simulation results are presented in the results section of this experiment, along with the destination point data from the human listeners.

*3. Testing procedure*

A graphical user interface (GUI) was developed in Matlab which allowed the listeners to interact with the algorithm. In each trial, a sound file was played, and the question "How good of a __ was this?" was displayed (the blank contained either "b" or "p" depending on the phonetic category being tested). Listeners would then have the option to adjust a slider bar to indicate their goodness judgment. The slider bar recorded values between .01 and 1 with a step size of .01. A "repeat" button was available so that listeners could hear the current stimulus as many times as they liked. Once a goodness rating was decided upon, listeners had to check a separate box in order to enable the button which would allow them to move on to the next sound file. The slider bar position did not reset between stimulus presentations.

The experiment consisted of four blocks with target "b", and 4 blocks with target "p". The "b" and "p" blocks were alternated. Depending on listener response latency, the experiment was completed in a single experimental session lasting less than 45 minutes.

Four different initial points were used. For every initial point, F0, cutback, and burst amplitude were set to points near the middle of their dimensions. In contrast, VOT and onset ramp were fully crossed for low and high initial values since these dimensions were expected to be the most important for categorization. The steps used as initial values are marked with "*" in Table V. The five dimensions were iterated three times, using the following order:

(1) VOT

(2) Cutback (Formant transition information)

(3) Ramp

(4) F0

(5) Burst Strength

The above ordering was selected to allow for rapid convergence by having dimensions believed to be the most important for categorization presented first. In principle, there is no reason why the ordering needs to be fixed on each iteration.

*4. Listeners*

Five native speakers of English (four female; one male) between the ages of 18 - 30 were recruited from the Indiana University community and paid for their participation.

B. Results

*1. Listener behavior*

Figure 6 contains a graphical representation of the goodness judgments from one listener for four blocks where the target category was /b/. Each curve in Fig. 6 represents a different initial starting point, and each panel indicates a different iteration. Stimulus numbers 1,4,7,10, and 13 are search vector midpoints, numbers 2,5,8,11, and 14 are search vector left endpoints, and numbers 3,6,9,12, and 15 are search vector right endpoints. The five stimulus dimensions are indicated at the top of the figure in the order they were presented (VOT, cutback, onset ramp, F0, and burst strength). Two features of these tracks should be noted. First, although the listener appears to recalibrate their use of the rating scale throughout the blocks, this listener typically utilizes a fairly small range of goodness judgments. Second, the iterative nature of the search algorithm appears to help the listener navigate out of areas of the space with depressed goodness ratings.

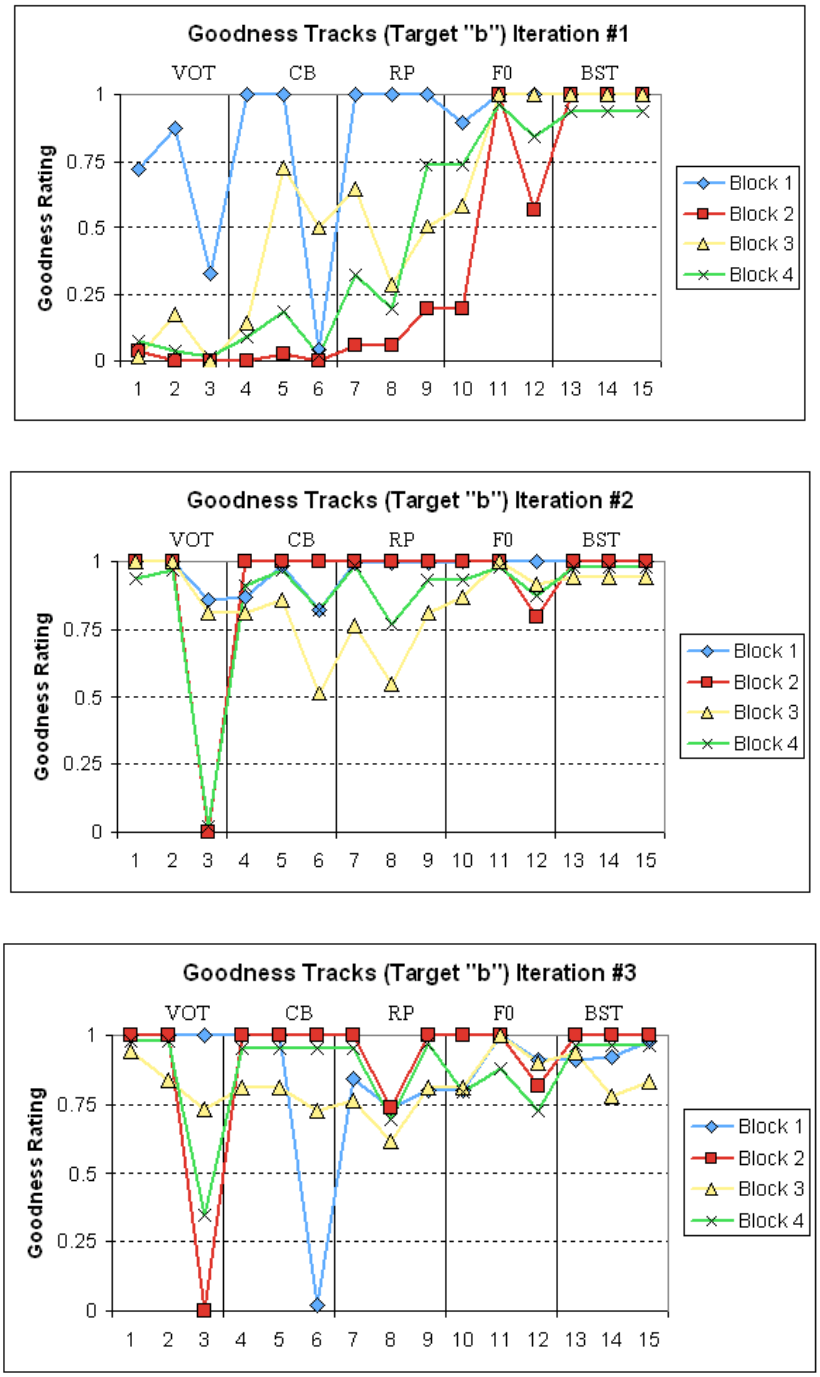


**Figure 6.** Goodness judgments for one listener's target /b/ blocks. Each panel corresponds to a different iteration (top panel = iteration 1, middle panel = iteration 2, bottom panel = iteration 3). Different curves correspond to different initial starting points. The horizontal axis indicates which point of a search vector is being tested (1,4,7,10, and 13 are search vector midpoints, 2,5,8,11, and 14 are search vector left endpoints, and 3,6,9,12,15 are search vector right endpoints), and the vertical axis indicates goodness ratings.

*2. Algorithm effectiveness*

As was discussed in experiment 1, the ability of the algorithm to track into regions of the multi-dimensional space containing best exemplars can be tested by comparing the goodness rating of a stimulus derived by the algorithm by the goodness rating of the stimulus that immediately preceded it. In Fig. 6 this would correspond to comparing the goodness ratings across stimuli 3-4, 6-7, 9-10, 12-13, and 15-1 (this occurs when moving from one iteration to the next). If the algorithm is working properly, the second stimulus in each of these pairs should have a goodness rating that is greater than or equal to the first stimulus in the pair. Although in principle it would be better to see if the derived point had the highest goodness rating relative to all three test points, this is not a fair comparison given the manner in which goodness judgments are obtained. Because the slider bar is not reset between stimulus presentations, only goodness judgments from adjacent stimuli can be fairly compared. Figure 7 depicts the proportion of times across all listeners where the goodness rating of the derived point increased, decreased, or stayed the same with respect to the preceding stimulus. Overall, 91.8% of the time, the derived point had a goodness rating that was greater than or equal to the preceding stimulus, and the derived point resulted in a decreased goodness rating only 8.2% of the time. These results suggest that the algorithm successfully tracks into regions of the acoustic space where good exemplars are located.

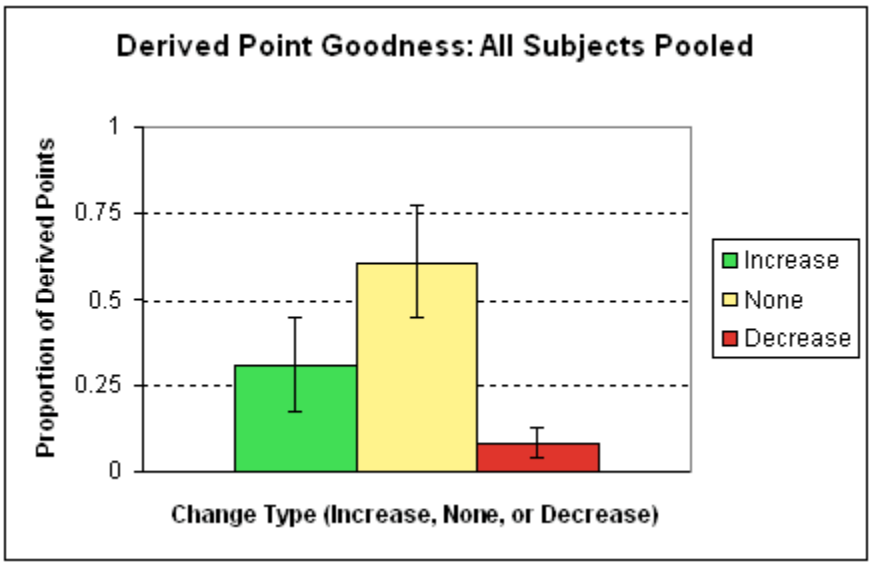


**Figure 7.** Proportion of cases where the goodness rating of the stimulus derived by the search algorithm increased, stayed the same, or decreased with respect to the preceding stimulus. Error bars indicate one standard deviation.

Another way of capturing whether or not the algorithm is exhibiting a global convergence towards better exemplars is to look at the average goodness rating recorded during each iteration. This measure should be handled gingerly given that it (a) contains goodness values compared across listeners, and (b) listeners are constantly recalibrating their use of the goodness rating mechanism. With these caveats in mind, Fig. 8 indicates that the algorithm does seem to be guiding subjects into regions with better exemplars, since their average judgments tend to increase in subsequent iterations.

One item that is worth noting is that initial starting position appears to have an effect on goodness ratings in some instances. Looking at the average goodness plot for target "p" in Fig. 8, the curve corresponding to initial point "p3" exhibits odd behavior relative to the other curves. Although curve "p3" is steadily increasing, the average goodness rating is substantially lower than was observed with the other initial points. Looking at the individual data, it appears for three out of five listeners that the stimuli being played were given extremely low goodness

ratings. Detailed analyses of these listener's tracks through the space are planned; however, regardless of those results, it does appear that initial starting point can have an influence on goodness ratings when only three iterations are employed. Fortunately, this result was isolated to one initial point in one target condition, and was not universal across all subjects.

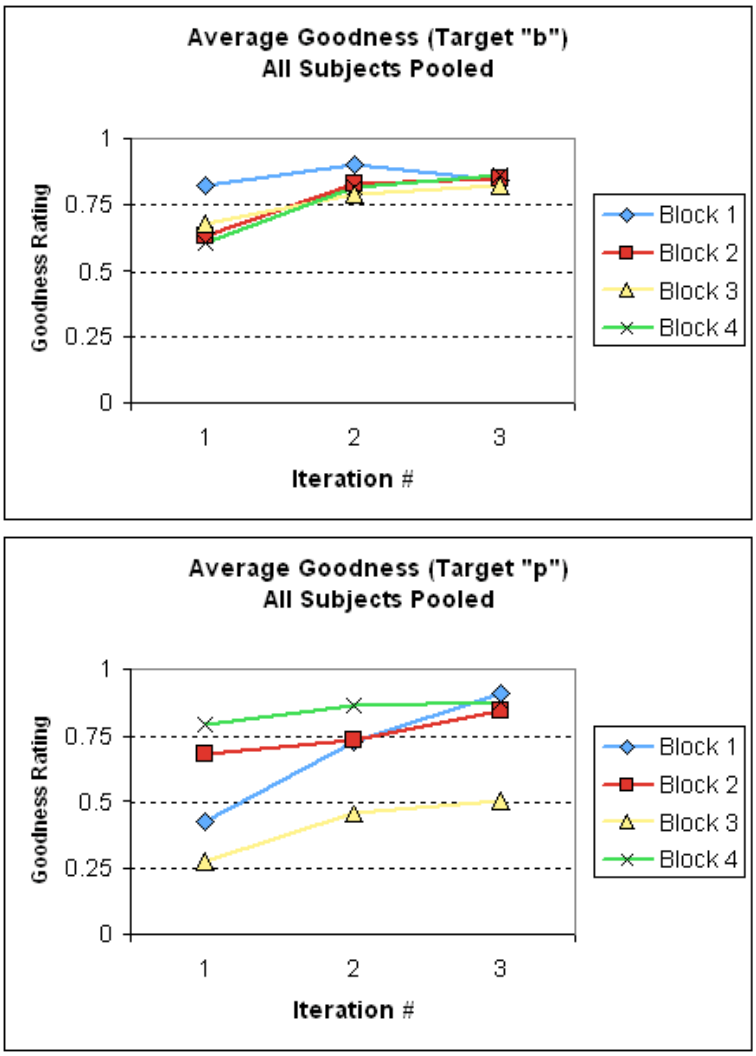


**Figure 8.** Mean goodness ratings for each iteration for each trial of /b/ and /p/. Different curves indicate different initial starting points.

*3. /p/ and /b/ category data*

In experiment 1, three types of data generated by the algorithm were discussed for their potential to examine phonetic categorization patterns: destination point of the multi-dimensional search, response tendencies for search vector endpoints, and multi-dimensional trajectories. Due to difficulties associated with representing multi-dimensional trajectories, only destination point results and search vector endpoint tendencies are presented here

*3.a. Destination points*

The final point derived by the search algorithm is theoretically an approximation of the best exemplar of the category being tested. A certain amount of noise influencing the location of this point is to be expected given the number of listener and task variables that are a part of the search process. Therefore, in order to examine destination points it is necessary to pool results across multiple blocks, and in the case of the current experiment, multiple listeners. By taking this approach, a picture of the probability distribution of destination points of the algorithm can be used to identify general regions where best exemplars should be located. One inherent difficulty of examining the destination points in this experiment is that they are situated within a five-dimensional space, and consequently it is impossible to practically display every dimension in a single figure. Therefore, in Fig. 9, each dimension is displayed individually, with data collapsed across all other dimensions (e.g. in the VOT plot, results from the other dimensions are not represented). Given the bias that was discovered in the algorithm following data collection, probability distributions for simulations using random goodness ratings are also displayed. The plots in Fig. 9 are essentially "main effect" distributions for each dimension, and consequently interaction effects between dimensions can not be seen.

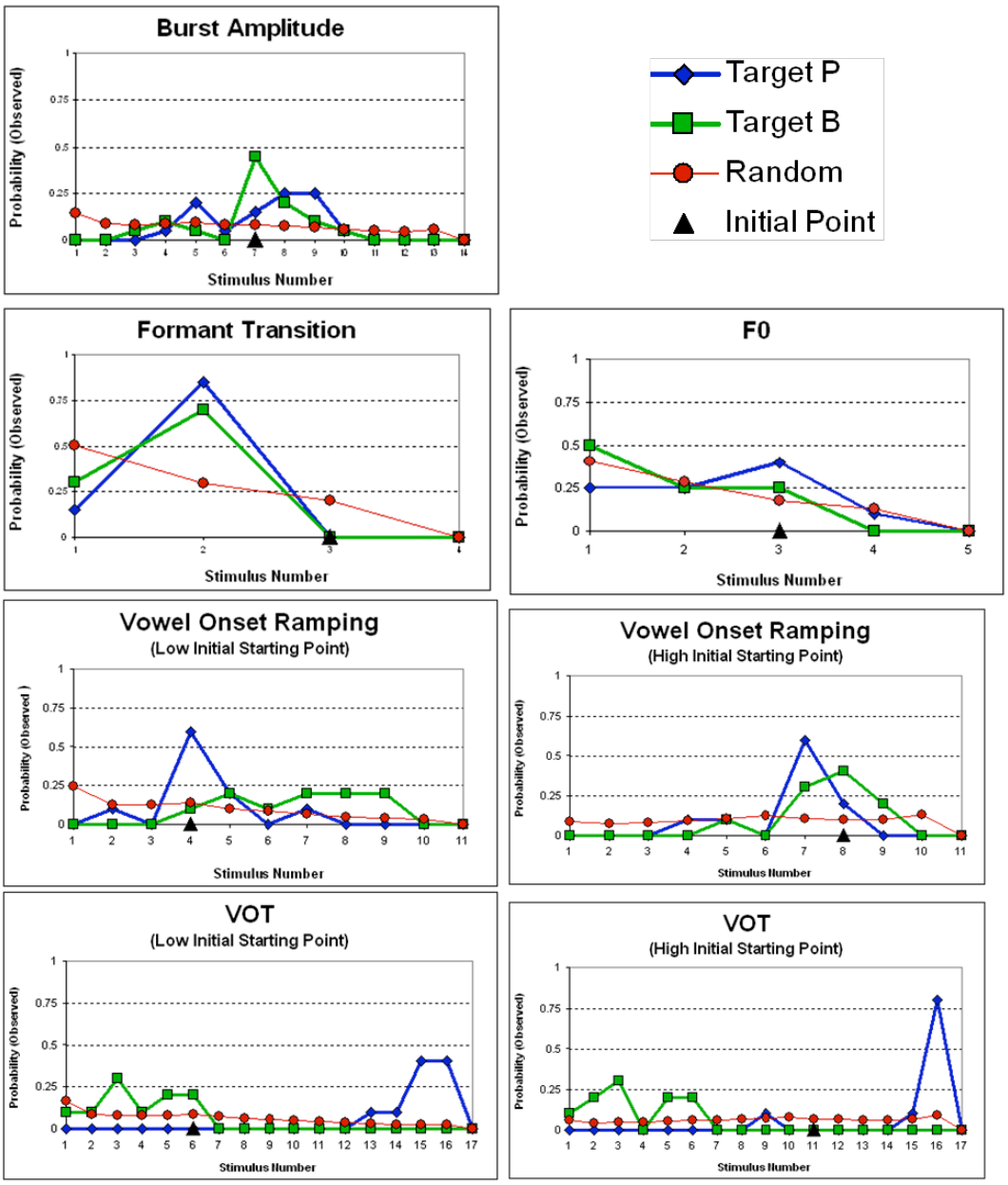


**Figure 9.** Probability distributions of destination points for /b/ and /p/ after three iterations. Each solid curve corresponds to a different target category (blue for /p/; green for /b/). A red solid curve indicates the long-term shape of the probability space according to simulations using random goodness judgments. Initial points are indicated by black triangles on the horizontal axis, and parameter values can be obtained by referencing Table V. In cases where a dimension had two possible initial values (VOT and Onset Ramp), results from each initial point are given in separate panels.

The number of tokens is too small to say much about F0, burst strength, or cutback when looking solely at destination points; however, there does appear to be something occurring in regards to vowel onset ramping. In the case of target /p/, listeners did not stray very far from the initial point (although with the high initial starting point listeners did systematically move lower). In regards to target /b/, there was a distinct preference for higher onset ramp values. It seems reasonable to hypothesize that if additional iterations had been done, the plateau in the upper ranges of the plot containing the low initial starting point would have built itself up into a peak. In regards to VOT, listeners demonstrated a systematic preference for prevoicing in target /b/, and an expected preference for high VOT values with target /p/.

*3b. Search vector endpoints*

One feature of the search algorithm is that the endpoints of the search vector are played in succession, and consequently the goodness judgments from the search vector endpoints can be directly compared. Over the course of multiple blocks this information can be used to determine if there is a general preference for one end of a dimension as opposed to another. The two panels of Fig. 10 contain histograms for target /b/ and /p/ indicating the proportion of times when the left or right endpoint of a dimension's search vector was preferred, as well as when there was no change in goodness judgment. Results were pooled across all listeners. In the case of target /b/, listeners preferred shorter VOT values, more formant transition information, higher vowel ramping, and lower F0. Burst strength tended to not elicit changes in goodness ratings. In the case of target /p/, there was an extremely strong preference for long VOT, with the other dimensions being mostly irrelevant.

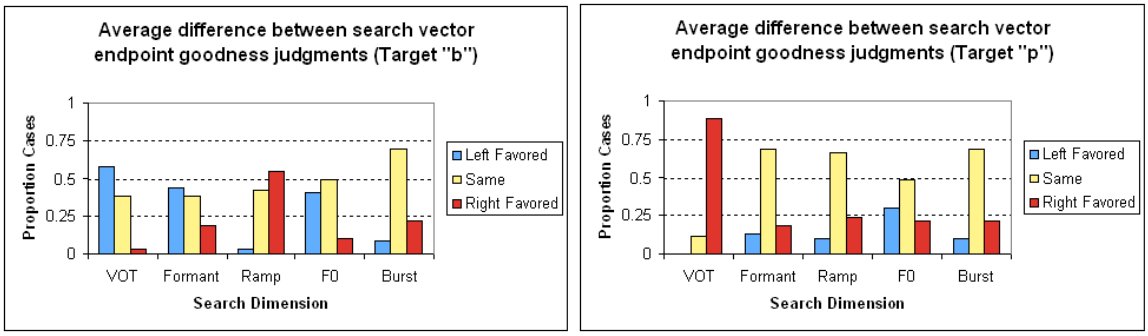


**Figure 10.** Histograms indicating search vector endpoint preferences for each dimension. Target /b/ is shown in the left panel, and target /p/ is shown on the right. Blue bars indicate a preference for "left" search vector endpoints (low stimulus #s), red bars indicate a preference for "right" search vector endpoints (high stimulus #s), and yellow bars indicate no difference in goodness judgments. The height of the bar indicates the proportion of cases for each preference.

C. Discussion

*1. Stimulus space noise*

Although the results from experiment 2 are very encouraging, there are a number of potential sources of noise that can be traced to the stimulus space. First, since the different F0 values were obtained by having the first author produce multiple utterances, this introduced a number of difficulties for stimulus calibration. Second, it is apparent that the steps in the cutback dimension were far too coarse. Also, since creation of the cutback dimension required the deletion of entire pitch pulses from the sound file, calibration was once again a difficult issue since reference points for two of the dimensions (burst strength and onset ramp) were shifted. These two observations indicate that additional methods will need to be developed that allow for more fine-grained manipulation of formant transition information, as well as minimize calibration difficulties.

*2. Number of iterations*

In experiment 2 the number of iterations was arbitrarily picked as an optimization between data collection time and what was expected to be sufficient for convergence on a best

exemplar. Based on the results from the onset ramp dimension destination points, it appears that at least four or five iterations should be used. Adding additional iterations would also make it easier to identify covarying relationships between dimensions.

## IV. GENERAL SUMMARY

### *1. Judgment of effectiveness*

Based on the results from simulations using random goodness judgments, as well as piloting on a stimulus space designed to test /b/ and /p/ categorization, it appears that the proposed multi-dimensional search method provides a good first approximation of best exemplar locations in relatively few trials. The design of the algorithm results in multiple types of data being available for both (a) tracking the search method's effectiveness, and (b) verifying the importance of different stimulus dimensions. In addition to these positives, the algorithm is also general enough to be applied to any multi-dimensional stimulus space for in which gradient goodness judgments are possible. Although all five dimensions in experiment 2 were acoustic, this was not necessary for the functioning of the search method. In principle, the generalness of the approach allows for multi-modal stimulus spaces to be explored.

### *2. Limitations*

It is important to note that the proposed search method is not intended to give a high resolution picture of phonetic categorization in multi-dimensional stimulus spaces; rather, its primary function is to provide a first approximation of the important factors involved in stimulus categorization. This search method is designed to aid in the development of lower dimensional stimulus spaces that can be explored with traditional methods by either (a) eliminating extraneous dimensions from stimulus spaces, or (b) creating single dimensions by covarying

multiple stimulus variables. Also, as mentioned above, underlying the search procedure is the idea that it is possible to elicit gradient judgments of the categories being studied. If this condition is not met the proposed method is not applicable.

*3. Future work*

The purpose of the current study was to begin to develop a methodology for exploring phonetic categorization in multi-dimensional stimulus spaces. Based on the results from experiment 2, the proposed search method at least preliminarily appears to be effective. The next step in the current project is to refine the stimulus space by addressing some of the issues raised earlier. Also, small refinements to the search procedure are also planned (i.e. randomly alternating which endpoint of a search vector is played first). Following these adjustments, the goal is to test phonetic categorization of labial stops in multiple languages (namely Korean, Japanese, and English) using the same stimulus space.

## V. CONCLUSION

In this paper, motivation for needing to explore phonetic categorization in multi-dimensional stimulus spaces was presented based on finding from VOT and vowel space studies. The inadequacy of applying traditional methods of phonetic categorization to higher dimensional stimulus spaces was discussed, as well as the non-genearlizibility of a previously proposed multi-dimensional search method. In response, a generalized multi-dimensional search algorithm was proposed and evaluated using computer simulations. The proposed algorithm was found to not contain any substantial uncontrollable biases, and the design of the algorithm was also shown to provide multiple types of data for analyzing phonetic categorization tendencies. Using a stimulus space designed to test labial stop categorization, the categorization of stimuli into /b/ and /p/

categories by native speakers of English was tested. Results showed that the search method appeared to be effective in providing a first approximation of factors deemed important by listeners for labial stop categorization. Limitations of the method were discussed, as well as plans for refinements of both the stimulus space and search method.

## V. ACKNOWLEDGMENTS

This work was supported by NSF grant BCS-04406540. We would also like to thank Diane Kewley-Port for generously providing lab space to run listeners.

**References**

Caramazza, A., G.H. Yeni-Komshian, E.B. Zurif, & E. Carbone (1973). The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals. *J. Acoust. Soc. Am., 45*, 4231-4328.

Harnsberger, J.D., Svirsky, M.A., Kaiser A.R., Pisoni D.B., Wright R., & Meyer T.A. (2001). Perceptual "vowel spaces" of cochlear implant users: implications for the study of auditory adaptation to spectral shift. *J Acoust. Soc. Am. 109*, 2135-2145.

Hillenbrand, J., & Gayvert, R.T. (1993). Identification of steady-state vowels synthesized from the Peterson and Barney measurements. *J. Acoust. Soc. Am., 94,* 668-674.

Iverson, P. & Evans, B. G. (2003). A goodness optimization method for investigating phonetic categorization. *Paper presented at the 15th International Conference of Phonetic Sciences*.

Johnson, K., Flemming, E. & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, *69*, 505-528.

Kitahara, Mafuyu (2001). "Category structure and function of pitch accent in Tokyo Japanese," Doctoral Dissertation, Indiana University.

Lisker, L. & Abramson, A. S. (1964). A cross-language study of Voicing initial stops: Acoustical measurements. *Word, 20,* 384-422.

Lisker, L., & Abramson, A. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967* (pp. 563-567). Prague: Academia.

Molis, M. R. (2005). Evaluating models of vowel perception. *J. Acoust. Soc. Am., 118*, 1062 – 1071.