Title: The Frame Problem in Speech Communication: Defining the Dimensional Space for Phonetic Categorization

Authors: Andrew J. Lotto, Associate Professor, University of Arizona
& Lori L. Holt, Professor, Carnegie Mellon University

Abstract: Much theoretical and empirical work has been focused on how language learners/users parse multi-dimensional auditory spaces into the phonetic categories of a native or second language. A more fundamental question is how the listener determines the relevant dimensions for the perceptual space in the first place. Harvey Sussman has offered one of the only principled theoretical accounts for the existence of particular perceptual dimensions – neural columnar encoding that leads to relative dimensions. The framework of Sussman's theory – interactions of neural processing constraints with statistics of the input – provides insights into potential answers to the more general questions about how listeners can solve the "frame problem" of which dimensions are most relevant to an auditory categorization task.

Phonetic Perception as Categorization

In the last several decades there has been a move in the field of speech perception away from discussions of detection of invariant features and simple match-to-sample approaches toward modeling phonemic acquisition and perception as a complex categorization task occurring over a multi-dimensional perceptual space (Holt & Lotto, 2010). It is typical in the categorization approach to presume that language learners parse perceptual spaces littered with exemplars of particular speech sounds into phonemic bins based on the distributions of these experienced exemplars. Subsequent phoneme perception then occurs with the target exemplar being mapped into this space and the most likely phoneme category chosen. An example of such an hypothesized space is presented in Figure 1.

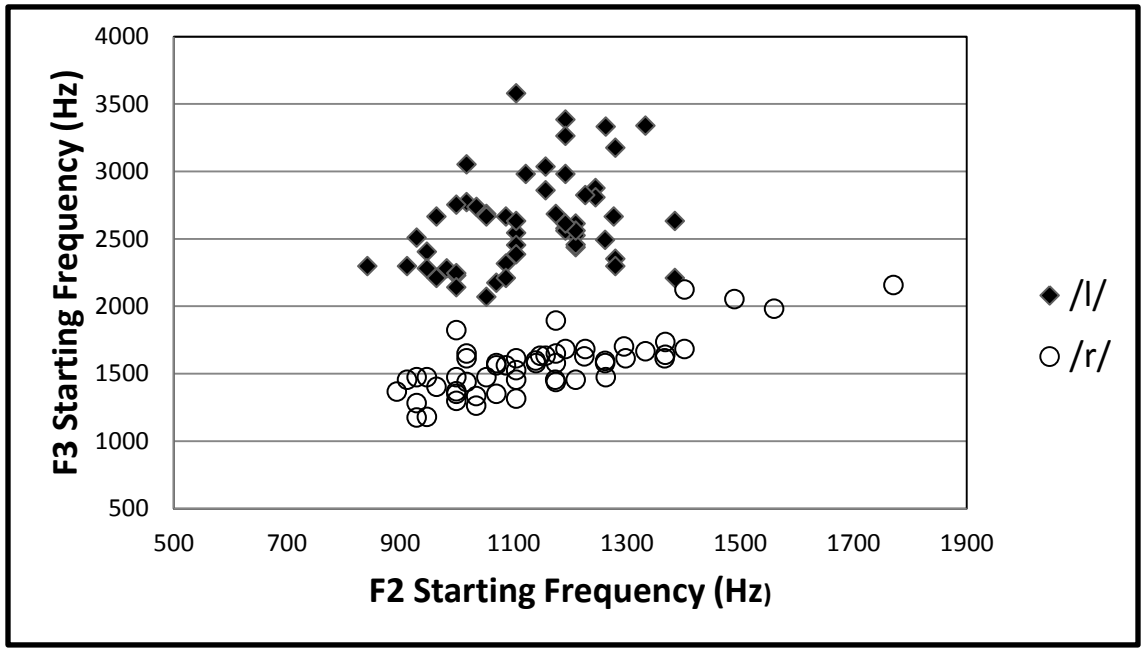

Figure 1. *Map of English productions of /l/ and/r/ (syllable initial) in a space based on the starting frequencies of the second and third formant (modified from Lotto, Sato & Diehl (2006)).*

Each point in this graph represents the measurement of the starting frequencies for the second and third formant from native English productions of /l/ and /r/ from a number of speakers producing two syllable words in isolation from Lotto, Sato and Diehl (2006). If one takes this as an estimate of the typical distribution of these two measures for this English contrast, then one can see that the two categories are fairly well distinguished by the value of the F3 starting frequency but that one would also need to consider the starting frequency of the second formant in order to perform optimally. In fact, one could readily determine the optimal boundary for separating these two categories of sounds in this space (the line would be nearly perpendicular to the F3 axis with a slight shallow positive slope). The categorization approach that is now in vogue in speech perception research presumes that native language learners form some representation of these distributional characteristics across this space and then determine the best category boundary in this space for subsequent categorization. The distributions in Figure 1 should result in native speakers making categorization decisions primarily based on the more informative dimension – the starting frequency of F3 – which is exactly what adult English speakers do (Yomada & Tohkura, 1991).

Modeling phonemic acquisition and perception in terms of parsing multidimensional spaces has been quite successful in providing explanations and predictions in a wide variety of speech perception phenomena in children (see Kuhl, 2000), native-speaking adults and adult second language (L2) learners. As an example of how this sort of approach can further understanding of (and/or contribute to practical applications for) L2 acquisition, Lotto et al. showed that native Japanese speakers make more of a distinction in F2 onset frequency than native English speakers when producing the English /l/-/r/ contrast. This heavier reliance on F2 is reflected in their perception of the contrast, resulting in non-native categorization patterns. Subsequently, Lim and Holt used a categorization approach, varying the distribution of stimuli on the F2 and F3 dimensions, as the basis for a training regimen that improved the perceptual categorization of /l/ and /r/ by native Japanese-speaking learners of English.

One of the appeals of this approach is that the proposed categorization processes are quite general and not specific to speech. In fact, much of the research on auditory categorization has been co-opted from the better developed categorization work in vision (Ashby & Maddox, 2005) and the models of speech categorization are applicable to a wide-variety of perceptual tasks (e.g., Massaro, 1987; McMurray, Aslin & Toscano, 2009). Recently, Chandrasekaran, Yi, and Maddox (2013) began applying a dual-systems approach to categorization that has been well-validated for vision to auditory categorization tasks including speech perception. In this newer model, perceptual spaces such as the one in Figure 1 are still parsed based on experienced distributions, but the neural mechanisms depend (at least partly) on how these distributions align in the dimensional space. For example, the distributions in Figure 1 are most compatible with the "reflective" system that involves the prefrontal cortex and anterior caudate nucleus. If one were to rotate the distributions in space so that the optimal boundary would lie more along the dimensional diagonal (as one sees in the distributions from Japanese speakers) then the "reflexive" striatum-based learning system would be more appropriate.

Early in our work, we tried to take seriously the claim that speech perception was the result of general processes of auditory categorization (Holt, Lotto, & Kluender, 1998; Lotto, 2000). We reasoned that the best way to understand speech sound categorization was to understand auditory categorization. We undertook a large number of experiments using sounds varying on non-speech dimensions. The benefits of using novel non-speech sounds is that it allowed us precise control over the distributional properties of the sounds and it provided us complete control over the listeners' exposure history. In order to tackle the questions we felt were most compelling, we developed a number of stimulus sets with a variety of relevant stimulus dimensions. Examples of relatively straight forward dimensions included the center frequency of a narrow noise band (Sullivan, Tanji, Lotto & Diehl, 2011; Kittleson, Diehl & Lotto, 2012), the center frequency of a tone, the frequency modulation rate of a tone (Holt & Lotto, 2006), the duration of amplitude ramps on noise bursts (Mirman, Holt & McClelland,

2004), and the onset asynchrony of two tones at different frequencies (Holt, Lotto & Diehl, 2004). Other dimensions were a little less obvious. In several sets of studies we varied the center frequencies of two stop-band filters applied to white noise. That is, we created two independently positioned troughs in the spectrum of the noise (Janega & Lotto, 2000; Mirman et al., 2004). Wade and Holt (2005) created a complicated set of stimuli that included an invariant square-wave portion and a simultaneous sawtooth wave that had an initial frequency sweep followed by a steady state frequency (these sounds were designed to have characteristics similar to seen in consonant-vowel syllables varying in place of articulation and vowel). The relevant perceptual space for these sound distributions was the frequency of the steady state portion and the difference between the steady state frequency and the starting frequency. This is a subset of the dimensions we have used and does not include dimensions developed by other researchers interested in auditory categorization (for example, Christensen & Humes, 1997 varied the center frequency and frequency modulation of some frication noise along with the duration of a silent gap). The most easily summarized finding from all of this non-speech work is that categories defined in these dimensional spaces are readily learned by listeners. We also learned a lot about the *process* of auditory category formation that appears to be quite applicable to speech perception. For example, the non-speech work on cue weighting by Holt and Lotto (2006) was the inspiration for the successful L2 training work of Lim and Holt (2011). Similarly, the results using two-tone asynchrony by Holt et al. (2004) provides some basis for explaining the use of the voice-onset time dimension across languages.

However, in the midst of all this success was a lingering question with no clear answer. We had good models of how listeners parse the perceptual space given a set of distributions, but…how can we be sure what the perceptual space is?

From whence come thy dimensions?

The question of whether the perceptual space that listeners are utilizing is the same as that presumed by the researcher is important for modeling categorization and phonetic perception. As mentioned above, the relevant system for categorization according to a dual-systems approach, a la Maddox and Ashby, depends on the relationship of the distributional structure to the dimensional axes of the perceptual space. The predictions of this model to this point have been made presuming that the perceptual space maps onto the stimulus space being manipulated by the experimenter. If the listener/learner is using a different "space" to represent the experienced exemplars, then these predictions would be difficult to make.

This is not just a concern about psychophysically scaling the dimensions to barks or mels or sones, it is a question of whether the listener is actually representing the exemplars using the same aspects of the signal that the researcher is manipulating. The presumption, of course, is that the stimulus space and perceptual spaces are similar. If this is the case (and given the effectiveness of auditory modeling efforts this appears to be the case), then one is left to wonder how it is that listeners manage to so readily determine the relevant dimensions for a task. This is really a fundamental question for understanding auditory learning/categorization. In these laboratory tasks, listeners often hear one sound on each trial and then make a response often followed by feedback. From this coarse sampling, they must determine which aspects of the sound are relevant. Does this mean that listeners are monitoring all possible relevant dimensions of a stimulus – duration, onset amplitude slope, frequency and rate of change of resonances, fundamental frequency, etc., etc. – and determining which are most useful for the current task?

A typical explanation would indeed be that listeners choose dimensions initially based on where the greatest variance lies (as a sign of potential information) and then pares these dimensions down to the working space on the basis of whether the variance in these dimensions are predictive of the

feedback. That seems reasonable enough except that it still requires listeners to calculate relative variance across a large number of potential dimensions. The list of dimensions used in the non-speech tasks above provides some sense of the potential size of this set. For example, the categorization task of Wade and Holt (2005) was solvable on a space with the dimensions – (1) steady state frequency of the resonance on the higher frequency square wave and (2) that steady state frequency minus the onset frequency of that resonance. Why would these be readily available dimensions for listeners? To make matters more complicated, Wade and Holt (2005) did not provide explicit feedback for auditory categorization nor were participants instructed to categorize the sounds. The learning – apparently based on something like these dimensions – was implicit as part of playing an action video game.

The above argument is not to say that it is a miracle that listeners come up with the relevant dimensions, it is a call for a theoretical investigation of how listeners come up with these dimensions. This basic first step is in many ways more interesting than the modeling of how listeners parse this space once they have it. Considering the importance of dimensions leads to a number of interesting questions. Is there a constrained set of possible dimensions? What would constrain the set – other than something uninteresting such as the variance in that dimension not being robustly encoded in the auditory periphery or beyond a temporal window of possible integration? That is, are there categorizations that cannot be learned even if they could be potentially linearly separated in a perceptual space? If the set is not constrained, then is there a cost to adding novel dimensions and how does a listener monitor them all?

In addition to these theoretically interesting questions that arise when we focus on the dimensions, there are also some troubling concerns that emerge for those who investigate auditory categorization. Above, we described non-speech dimensions based on the center frequency of troughs of energy in the spectrum of white noise. Of course, it is unlikely that listeners are in fact using that center frequency. It is more likely that they are using the higher or lower cutoff frequencies (which will be released from inhibition by the trough of energy). A listener may also use a center of spectral gravity dimension or some other aspect of the shape of the spectral envelope. Listeners may differ in which one of these dimensions they use to perform the task. In the case of these laboratory experiments, this lack of precision in defining the dimensions is not too detrimental because these dimensions would be so highly correlated and would all provide sufficient resolving power to show categorization. However, it is easy to construct scenarios in which these differences *will* matter and, in particular, the importance of properly defining dimensions becomes much more obvious for speech categories.

A case of dimensional imprecision: Voice Onset Time

One of the most common variables in speech acoustics and phoneme categorization research has been Voice Onset Time (VOT, Lisker & Abramson, 1968). In essence VOT is the time between the release of the occlusion for a stop consonant and the onset of vocal fold vibration (we will focus here on the classic word-initial stop consonant voicing contrast). The first obvious problem with the use of VOT as a perceptual dimension is that it is an articulatory definition and not an auditory one. With due respect (whatever that may be) to the Motor Theory of Speech Perception (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967) or the direct realism of Fowler (1986), it would be far more satisfactory to have perceptual dimensions that cohere to our current understanding of what is encoded by the auditory system. The problem is that the articulatory definition provided above corresponds to a large number of acoustic changes that could serve as perceptual dimensions (Lisker, 1986). (We understand that it is this many-to-one mapping that provides some of the motivation for the theories mentioned above but we choose to ignore this debate here as we have engaged in it in other places, see Diehl, Lotto & Holt, 2004).

When VOT is discussed in terms of an acoustic dimension, it appears to usually stand-in for the duration of the aspiration noise prior to the onset of periodic energy. Figure 2 presents a pseudo-spectrogram of a consonant-vowel.

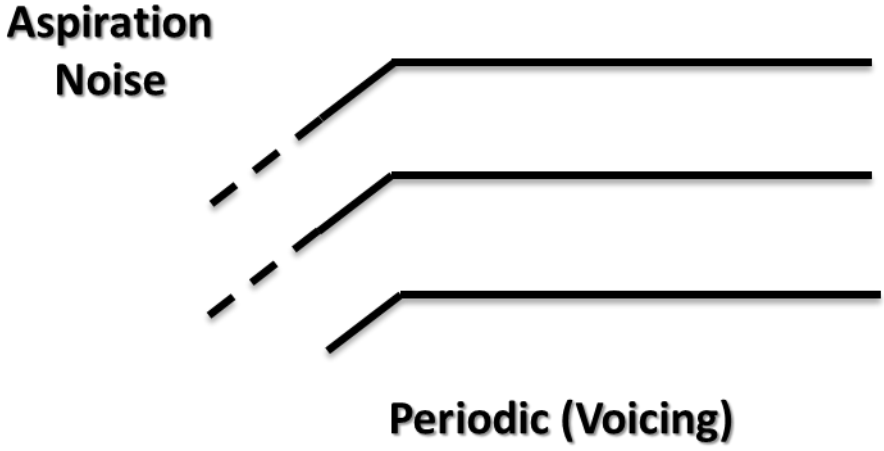

Figure 2

The dashed lines represent formants being excited by aspiration energy and solid lines represent formants excited by periodic energy. The duration of the aspirated portion could be a dimension that listeners use to categorize a voicing contrast (e.g., /b/ vs. /p/). Of course, one can quickly come up with other acoustic measures that would serve the same purpose – e.g., the duration of the periodic portion or the duration of formant transitions that are excited by periodic energy. One could also rely on the loudness of the aspiration noise given that loudness is a function of the amplitude of the energy AND the duration of the energy (Bloch's Law).

One might also note that there is little aspiration energy in the region of the first formant (no dashed lines in the figure). This lack of energy is due to an anti-resonance caused primarily by an anti-resonance due to tracheal coupling with the vocal tract. This feature results in several other potential acoustic dimensions that relate to voicing categorizations. For instance, the starting frequency of the first formant would be higher for voiceless stops (since the formant transitions for these stops are always increasing in frequency and one would be delaying the onset, see Kluender & Lotto, 1994). In addition, instead of measuring the time of aspiration noise, one could measure the time from stimulus onset to F1 onset. F1-onset-delay or F1-onset-frequency may seem like unlikely perceptual dimensions in this case but Lotto and Kluender (2002) demonstrated that reasonable voicing categorizations can be made even for stimuli varying only in the "cutback" of F1 with no aspirated portion. That is, a good exemplar of /pa/ can be created by filtering energy near F1 even when the entire stimulus is excited by a periodic source (equivalent to Figure 2 with no dashed lines; see also, Lisker, 1975).

The above discussion of the many acoustic dimensions related to voicing is nothing new. The use of multiple cues to arrive at a phonemic decision is the basis of the work on trading relations (see Repp, 1984 for a review). In fact, the effectiveness of so many dimensions for phonemic perception was not considered a problem for the listener because all of these cues arise from the same articulatory actions. If listeners were perceiving speech in terms of these articulatory actions (or the planning of those actions) in accordance with Motor Theory, then the problems of unconstrained sets of dimensions go away. What is important is the perception of the voicing gesture and not the individual acoustic dimensions. For those of us who are not satisfied with such an explanation, the problem of being explicit about the dimensional space for phonetic categorization remains. Despite its ubiquity, VOT is not an adequate dimension for models of auditory categorization. Even after decades of empirical and theoretical work in speech perception, there are a lack of theories on listeners uncover the dimensions underlying their categorization of speech sounds. However, there is at least one attempt to provide

such an explanation that provides a template for future endeavors and that comes from the work of Harvey Sussman.

Harvey Sussman: Solving the frame problem

One of the best examples of the need to get the perceptual dimensions "right" when trying to understand auditory categorization comes from the categorization of stop consonants by place of articulation. If one looks at the acoustics of /ba/, /da/ and /ga/ spoken by a typical speaker, there is a salient difference in the onset frequency of the second formant. F2-onset frequency appears to be a good dimension for categorizing these stops. However, the distinctiveness of these stops on F2-onset disappears as one varies the following vowel. The extreme context-dependence of the acoustics of these stops has been one of the principal/classic demonstrations of the lack of invariance in speech (Liberman et al. 1967; Stevens & Blumstein, 1978; Blumstein & Stevens, 1979). However, as Sussman and colleagues have shown (Sussman, McCaffery & Matthews, 1991), a tremendous amount of orderly structure can be witnessed by plotting exemplars in an F2-onset-frequency by F2-vowel-midpoint-frequency space. What appears to be a nearly impossible categorization problem becomes less mystical when one sees the structure inherent in a different acoustic space.

The stimulus distributions in the non-speech experiment by Wade and Holt (2005) had a similar structure to what one sees for place of articulation in speech – the categories were separable in an acoustic space defined by the onset and steady-state frequencies of the second resonance filter applied to the stimulus. In fact, this stimulus space was inspired by Sussman's work in categorization of speech. As discussed above, the problem that arises with these complex stimulus spaces is the lack of an explanation of how listeners determine the relevant space. The question is analogous to the Frame Problem in artificial intelligence and philosophy (McCarthy & Hayes, 1969). In AI, the problem comes from determining which aspects of representation of the environment are relevant for any action – this is very difficult to determine *a priori*. In the case of auditory categorization research, the problem is determining which acoustic dimensions are relevant for a task – very difficult to do *a priori*. How does a listener come to appreciate that the key to the categorization of place of articulation is the linear relations between F2 onset and mid-vowel frequencies, if that is in fact the correct dimension?

Sussman (e.g., Sussman, Fruchter, Hilbert, & Sirosh, 1998; Sussman, 2013) has provided one of the only attempts to answer the frame problem for audition. He has taken a neuroethological approach and pointed out that the type of relational processing required for place of articulation categorization is seen in auditory processing across a number of species (with well-described neural models of this processing). Based on this cross-species work, he proposes that humans are biologically endowed with neural processes that would favor the perception of relational cues in audition. Further, the importance of these types of cues for processing speech sounds is not a happy coincidence but is the result of evolving a communication system that takes advantage of these biases in perceptual processing.

The most beneficial aspect of Sussman's work is not that it provides a basis for understanding how humans distinguish /b/, /d/, and /g/, but that it offers some guidelines and hope for theories about the possible set of auditory dimensions and the development of perceptual spaces. The key is first to admit that there are important unanswered questions about perceptual spaces and then to go back to first principles and think about what the auditory representation and what kinds of processing the auditory system does well.

A thought-experiment: Formants

Probably the most popular stimulus/perceptual space in speech research is the F1-frequency x F2-frequency space for plotting vowel exemplars. Since the classic plot from Peterson and Barney (1952),

the F1 x F2 space has been utilized to explain acquisition of vowel categories in native language (e.g., Kuhl, 1991), in a second language (e.g., Bosch, Costa, & Sebastian-Galles, 2000) and as the basis for computational modeling of speech categorization (e.g., Kluender, Lotto, Holt & Bloedel, 1998; Maddox, Diehl, & Molis, 2002). It is widely accepted that vowel perception is strongly related to, if not strictly dependent on, the frequency of the first two formants.

Despite the centrality of formant frequencies to the science of speech perception, it is difficult to derive these as basic perceptual dimensions if one goes back to first principles. Formants are resonances of the vocal tract. They often result in spectral peaks, but not always – especially for higher fundamental frequencies. It is often difficult to measure the frequencies of these measures from spectra (as anyone who has done extensive acoustic analyses knows) and we often use methods such as linear predictive coding to estimate what the resonant frequencies are based on a model of speech production/acoustics. We are again left with dimensions that are inherently related to speech production but may be acoustically obscure. To make matters more confusing, we treat the formant frequencies as if they vary independently, but they do not.

It is not clear that the frequencies of resonances of the vocal tract are salient features of the auditory input of vowels. Here is a quick thought experiment: if you consider the encoding of vowel stimuli in the auditory periphery, what is the likelihood that the dimensions that will be considered most relevant will be F1-F2 frequencies (presuming one does not have a special module or innate knowledge of vocal tracts and the physics of speech production)? We can even start with simplifying obviously-wrong starting assumptions such as – we only receive steady-state vowels and the auditory system provides something like a spectral envelope of the incoming stimulus. One may imagine that during vowel learning, the initial dimensions arise from tracking the spectral variance and coming up with dimensions that account for most of that variance. This would be similar to computing a principal components analysis (PCA) across the experienced exemplars. Figure 3 shows what the first two components of such an analysis would look like. In this case, we simply took the spectral envelopes (log-scaled in amplitude) for 10 English vowels and performed a PCA.

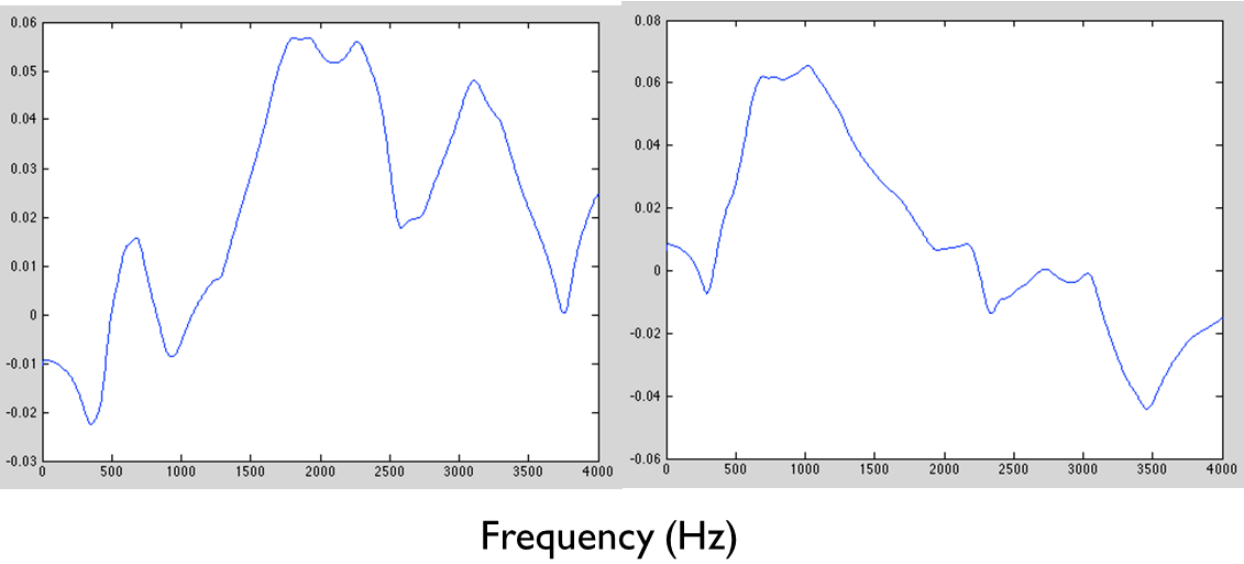

Figure 3: *First two components of principal component analysis from spectral envelopes of 10 English vowels.*

These two components are independent spectral shapes and one can reconstruct the 10 vowels by weighting each of these spectral shapes in a combination (accounts for 78% of the variance in the original vowel spectra). In essence the weights represent dimensions for a perceptual space that is related to peripheral auditory encoding. Note that there is no reference here to formants or tube models of speech production. Figure 4 is a plot of the original 10 vowels represented within this new space.

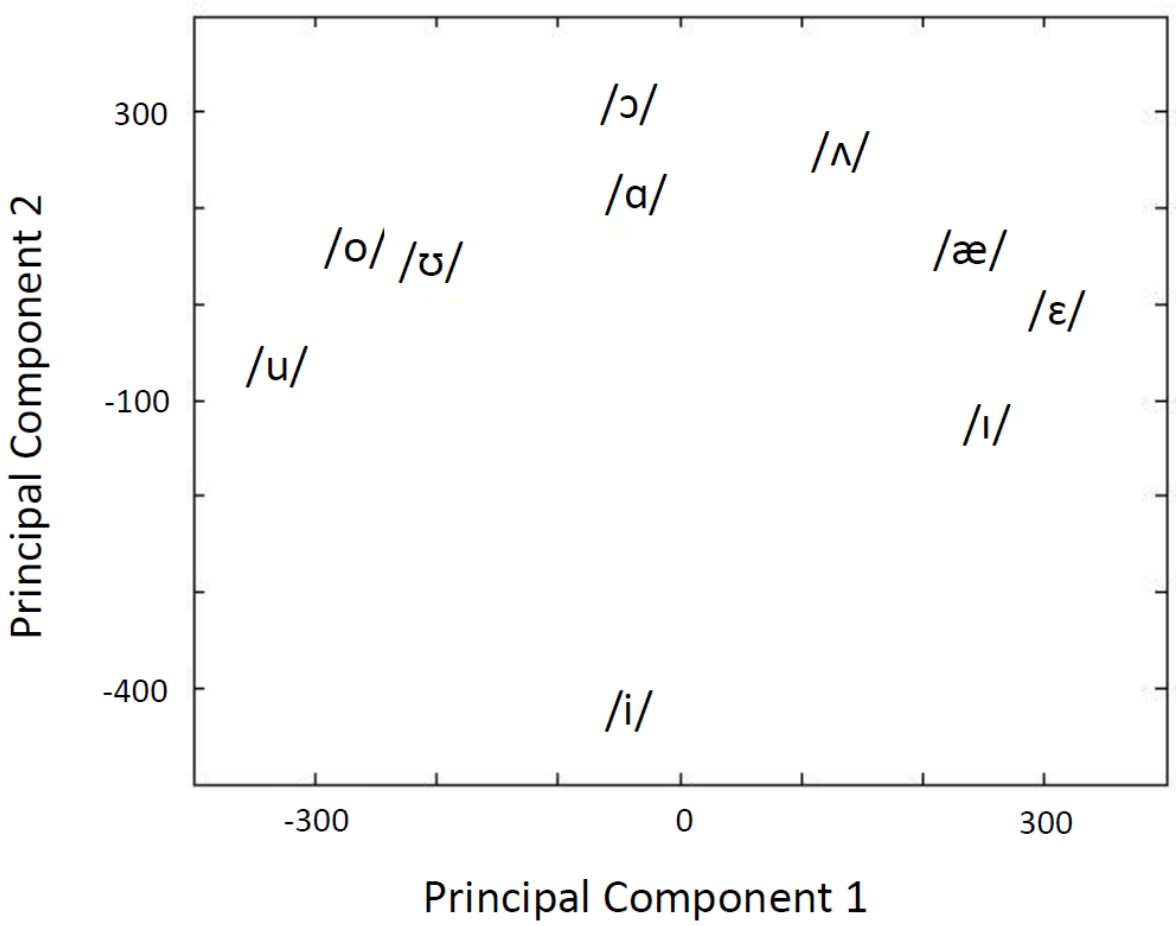

Figure 4: Vowel space across first two components.

This simple exercise is NOT intended as a model of how vowels are perceptually represented. To start with, one would want to scale the spectra either psychophysically or using a model of the auditory periphery. Also, one needs to determine whether the strict requirements of orthogonal dimensions for PCA are necessary. The real purpose of this thought experiment is to provoke the idea that we may want to derive our auditory dimensions for speech categorization instead of accepting those that were developed mainly in research on the acoustics of speech production. It is this sort of analysis from the auditory perspective that is inspired by the successful example provided by Harvey Sussman. The end result may be a solution to the dimensional Frame Problem and could lead to a better understanding of how humans acquire and use auditory categories for communication.

Acknowledgements: We wish to thank Nico Carbonell and Jessamyn Schertz for assistance with editing, proofreading and graph construction. Also, we want to express our gratitude to Harvey Sussman, who has been a supportive of us for our entire career starting in graduate school. We are honored to call him a friend. This work is supported by NIH- R01 DC004674-09 to AJL and LLH.

References

Ashby, F.G., & Maddox, T.W. (2005). Human Category Learning. *Annual Review of Psychology*. 56, 149-178.

Blumstein, S.E. & Stevens, K.N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Am*. 66, 1001.

Bosch, L. Costa, A. & Sebastian-Galles, N. (2000). First and second language vowel perception in early bilinguals. European Journal of Cognitive Psychology. 12(2), 189-221.

Diehl, R.L., Lotto, A.J., & Holt, L.L. (2004). Speech perception. *Annual Review of Psychology*. 55, 149-179.

Chandrasekaran, B. Yi, H.G., & Maddox, W.T. (2013) Dual-learning systems during speech category learning. *Psychon. Bull. Rev*. 1–8.

Christensen, L. A., and Humes, L. E. (1997). Identification of multidimensional stimuli containing speech cues and the effects of training. *J. Acoust. Soc. Am*. 102, 2297–2310.

Fowler C A (1986). An Event Approach to the Study of Speech Perception from a Direct-Realist Perspective. *J of Phon.* 14(1), 3–28.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *J. Acoust. Soc. Am.* 119, 3059-3071.

Holt, L.L., Lotto, A.J., 2010. Speech perception as categorization. *Atten. Percept. Psychophys*. 72, 1218–1227.

Holt, L. L., Lotto, A. J., and Diehl, R. L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *J. Acoust. Soc. Am.* 116, 1763–1773.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (1998). Incorporating principles of general learning in theories of language acquisition. In M. C. Gruber, D. Higgins, K. S. Olson, & T. Wysocki (Eds.), *Chicago Linguistic Society—Vol. 34: The panels* (pp. 253-268). Chicago: Chicago Linguistic Society.

Janega, J., & Lotto, A.J. (2000). Explicit training of complex auditory categories". Presentation at the 72nd Meeting of the Midwestern Psychological Association, Chicago, IL.

Kittleson, M., Diehl, R.L., & Lotto, A.J. (2012). Optimal categorization of sounds varying on a single dimensions. Presentation at the 164th Meeting of the Acoustical Society of America, Kansas City.

Kluender, K.R., & Lotto, A.J. (1994). Effects of first formant onset frequency on [-voice] judgments result from general auditory processes not specific to humans. *Journal of the Acoustical Society of America*, *95*(2), 1044-1052.

Kluender, K.R., Lotto, A.J. & Bloedel, S.L. (1998). Role of experience for language-specific functional mappings of vowel sounds. *J. Acoust. Soc. Am*. 104, 3568.

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

Kuhl, P. K. (2000). Language, mind, and brain: Experience alters perception. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 99-115). Cambridge, MA: MIT Press.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*. 74, 431-461.

Lim, S. & Holt, L.L. (2011). Learning Foreign Sounds in an Alien World: Non-Native Speech Categorization. *Cognitive Science.* 35(7), 1390-1405.

Lotto, A.J. (2000). Language acquisition as complex category formation. *Phonetica* 57:189–96.

Lotto, A.J., & Kluender, K.R. (2002). Synchrony capture hypothesis fails to account for effects of amplitude on voicing *perception*. *Journal of the Acoustical Society of America*, 111, 1056-1062.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 50+ years of discoveries in speech communication* [Online conference proceedings]. Cambridge, MA: MIT.

Maddox, W.T., Molis, M.R. & Diehl, R.L. (2002). Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception & Psychophysics*. 64(4), 584-597.

Massaro, D. W. (1987). Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. *Lawrence Erlbaum Associates*, Hillsdale, NJ.

McCarthy, John and P.J. Hayes (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. D. Michie (ed.), *Machine Intelligence 4*, American Elsevier, New York, NY.

McMurray, B., Aslin, R.N., & Toscano, J.C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*. 12(3), 369-378.

Mirman, D., Holt, L.L., McClelland, J.L. (2004). Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly-changing acoustic cues. *J. Acoust. Soc. Am*. 116, 1198.

Repp, B. (1984). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language & Speech, 27*, 245–254.

Stevens, K.N. & Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am*. 64, 1358.

Sullivan, S.C., Tanji, J.A., Lotto, A.J., & Diehl, R.L. (2011). Sensitivity to changing characteristics of Guassian-shaped stimulus distributions in auditory categorization. Presentation at the 162nd Meeting of the Acoustical Society of America, San Diego, CA.

Sussman, H.M. (2013). Neuroethology in the service of neurophonetics. *Journal of Neurolinguistics*, 26, 511-595.

Sussman, H.M., Fruchter, D., Hilnert, J., Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, 21, 241-299.

Sussman, H. M., McCaffrey, H.A., & Mathews, S.A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am*. 90, 1309.

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *J. Acoust. Soc. Am.* 118, 2618-2633.

Yamada, R. A., and Tohkura, Y. (1990). Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese. *Proceedings of International Conference on Spoken Language Processing*, Kobe, Japan, 757–760.