Modeling categorical perception with Hilbert spaces

Chris Neufeld, University of Maryland, Department of Linguistics, neufeldc@umd.edu

Here I present a novel mathematical formalization of speech perception. The speech module of the auditory system is formalized as a Hilbert space. Phonetic categories are modeled as vector subspaces, and the phonetic identity of speech signals is determined by assessing how much the speech signals project onto these categorical subspaces. A computer simulation trained with TIMIT data demonstrates that this model captures two key properties of categorical perception: nonlinearity in identification functions, and a sharp peak in discrimination between categories.

The vector space that will be considered here is the space of all discretized complex-valued functions, which represent the Fourier transform of sound waves. This space is complete, and can be equipped with the inner product $\langle x, y\rangle : \sum_{i=1}^{N} x_i \overline{y_i}$, making this a Hilbert Space [1]. Inner products automatically induce a *norm*, which is a unary operation that maps vectors to real-valued scalars: $\|x\| = \sqrt{\langle x, x\rangle}$, which in this case corresponds to the root-mean-squared amplitude of acoustic signals. A norm induces a distance metric, which is a binary operation that maps pairs of vectors to a real-valued scalar, defined as: $d(x, y) = \|x - y\|$. Here this corresponds to the amount of acoustical power in the spectral difference between two signals.

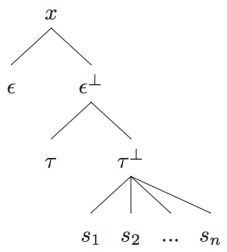

Figure 1: Schematic representation of the model. The input signal x has the nonspeech noise removed to yield, $\epsilon^{\perp}$. Then the components shared between all categories are removed to yield $\tau^{\perp}$. This vector is then projected onto all the phonetic subspaces, yielding s_1, s_2....

The core proposal is that phonetic categories are *subspaces* of this much larger Hilbert space of acoustic signals. Every vector subspace S can be associated with a projection operator P_s[1]. For any vector $x \in V$, $P_s x = y$ where $x = y + z$ and $y \in S$ and $x \in S^{\perp}$, where $S^{\perp}$ is the orthogonal complement of S. In other words, projection operators decompose arbitrary vectors into two parts, y and z, where y is a vector which is a member of the associated subspace S, and z is not a member of that subspace. For example, if S represents the subspace of all [s] signals, then P_s is the operator which allows one to decompose any acoustic signal into two parts – the [s]-component (potentially 0), and the rest. If the signal really *is* an [s], then the vector returned by P_s will be relatively 'large' compared to the the vector returned by the projection operator associated with a different phonetic category.

This model works by decomposing an input signal, x using projection operators in the following way: U is the union of all the categorical subspaces, and E is its orthogonal complement – i.e., the subspace of all vectors which are *not* speech sounds. P_{ϵ} is the projection onto this subspace. That is, $U = \bigcup_{i=1}^{N} S_i$ and $E = U^{\perp}$. Applying this operator to an input signal, x, gives us ϵ, the portion of the signal which is non-speech noise. To remove it, it is subtracted from x to get $\epsilon^{\perp}$. T is the intersection of all categorical subspaces, and P_{τ} is the projection onto this subspace, $T = \bigcap_{i=1}^{N} S_i$. Applying this operator to the denoised vector, $\epsilon^{\perp}$ gives us τ, the projection onto the subspace common to all speech sounds. Since this is uninformative for the purposes of phoneme identification, it is discarded: $\tau^{\perp} = \epsilon^{\perp} - P_{\tau}\epsilon^{\perp}$. We now have a vector, $\tau^{\perp}$ which has non-speech noise removed, and the projection onto the common subspace, τ, removed. This vector is then projected onto each categorical subspace S_i using the corresponding projection operator: $s_i = P_{s_i}\tau^{\perp}$. This cascade of decompositions is depicted in figure 1.

[1]For instance, in $\mathbb{R}^2$, the x-axis is a subspace, and there is a particular operator, P_x which projects arbitrary vectors onto the x-axis

Category membership is estimated by deriving weights from projections. If a vector comes from a particular category, the magnitude of its projection onto that subspace will be relatively larger than if it does not. Weights are computed for each category: $w_i = \left\langle \frac{s_i}{\|s_i\|}, \frac{\tau^\perp}{\|\tau^\perp\|} \right\rangle^g$, where g is a free parameter. This is the inner product of the input signal, $\tau^\perp$ and its projection onto the subspace S_i, both normalized, and raised to some power $g \geq 1$. If $\tau^\perp \in S_i$, this will be equal to 1. If $\tau^\perp \in S^\perp$, this will be equal to zero. To compute the probability that the input signal belongs to a particular category, the weights are normalized: $p_i = \frac{w_i}{\sum_{i=1}^N w_i}$. The signal is then reconstituted from its constituent projections s_i, reweighted by p_i: $\hat{x} = \sum p_i s_i$.

A simulation was created using the voiceless coronal fricatives [ʃ] and [s] from the TIMIT database. Projection operators were estimated by using Principal Components Analysis to derive a basis for each phonetic category, from which projection operators can be derived. An [ʃ] and an [s] were chosen randomly, and a linear, 21-point continuum was created by cross-fading the two signals. Figure 2 displays the identification function (derived from the weights estimated from the projection operators), and the discrimination function, operationalized here as the distance metric between adjacent pairs of reconstituted signals, $d(\hat{x}_i, \hat{x}_{i+1})$. It can be seen that the identification curve follows the familiar, sigmoid-shape, first observed in [4], and widely reported since. Second, the discrimination curve has a sharp peak in between the categories. This model exhibits the perceptual magnet effect without any notion of 'category center' or 'prototype [3], or any computations of probability distributions [2], or selection or ranking of acoustical 'cues' [5]. Perceptual warping arises out of the latent structure of phonetic categories, here modeled as vector subspaces which contain the complex spectra of speech sounds. Once the inner-product is defined, notions of length (i.e., loudness), and distance naturally arise as simple mathematical consequences. This action of the projection operators is also mathematically identical to the formal description of the action of neurons in auditory cortex with complex spectral-receptive fields – consequently, the model has a direct neural interpretation.

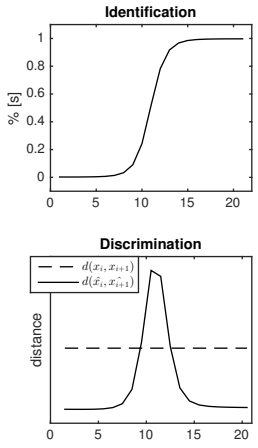

Figure 2: Simulation results. A linear continuum between [ʃ] and [s] , and the ID-function derived from the projection weights, and the discrimination function derived from the distance metric between adjacent, reconstituted signals (solid line), and distance metric between input signals (dashed line).

References

[1] L. Debnath and P. Mikusiński. *Introduction to Hilbert Spaces with Applications*. Academic Press, Inc, Boston, 1990.

[2] N. Feldman, J. Morgan, and T. Griffiths. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752–782, 2009.

[3] P. Iverson and P. Kuhl. Mapping the perceptual magnet effect for speech using signal detection. *Journal of the Acoustical Society of America*, 97(1):553–562, 1995.

[4] A. Liberman, K. Harris, H. Hoffman, and B. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358–368, 1957.

[5] J. Toscano and B. McMurray. Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34(3):434–464, 2010.