INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Modeling categorical perception with the receptive fields of auditory neurons

Chris Neufeld[1]

[1]University of Maryland
neufeldc@umd.edu

Abstract

This paper demonstrates that a low-level, linear description of the response properties of auditory neurons can exhibit some of the high-level properties of the categorical perception of human speech. In particular, it is shown that the non-linearities observed in the human perception of speech sounds which span a categorical boundaries can be understood as arising rather naturally from a low-level statistical description of phonemic contrasts in the time-frequency plane, understood here as the receptive field of auditory neurons. The TIMIT database was used to train a model auditory neuron which discriminates between /s/ and /sh/, and a computer simulation was conducted which demonstrates that the neuron responds categorically to a linear continuum of synthetic fricative sounds which span the /s/-/sh/ boundary. The response of the model provides a good fit to human labeling behavior, and in addition, is able to account for asymmetries in reaction time across the two categories.
Index Terms: speech perception, categorical perception

1. Introduction

Phonetic categories exert a powerful effect on the perception of speech sounds. Auditory percepts seem to be warped by the positions of phonetic category centers and boundaries in acoustical space, with acoustical neighbors becoming indistinguishable near category centers, and highly discriminable across category boundaries. And human labeling behavior is generally non-linear with respect to acoustic continua – there are large patches of acoustical space which humans confidently, and unequivocally label as belonging to some category or another, and only very narrow slices of truly ambiguous acoustical space, where human judgement is unreliable, variable and equivocal. Categorical perception (CP) has received a great deal of attention since the initial discoveries at Haskins Labs ([1], *et. seq.*). This study provides a proof-of-concept that some of the essential properties of categorical perception can be simply and straightforwardly accounted for in a biologically realistic framework whose computational primitives are the receptive fields of auditory neurons.

Much of the theoretical work on speech perception has focused its attention at what David Marr would have called the 'computational level' of description. So for instance, early (and unresolved) debates revolved around whether perceptual invariance lay at the auditory or the motor level [2, 3, 4]. And more recent research has found descriptive and predictive success in a statistical idiom, with, for example, Bayesian approaches [5] demonstrating that the effects of CP fall naturally out of a principled application of Bayes' rule. And the quickly growing literature on the neuroscience of categorical perception has focused broadly on where and when categorical effects can be observed in cortex – in other words, those questions best suited to contemporary neuroimaging techniques (e.g., [6, 7, 8, 9]).

But what has so far been absent is a concerted attempt to understand phonetic category perception in terms of the basic functional neurobiology of mammalian auditory cortex. This is a component of what Poeppel has alternately called 'the granularity mismatch problem' [10] or the 'mapping problem' [11], which describes a fundamental incongruity between the theoretical objects at our computational levels of description (formants, distinctive features, covariance matrices, principal dimensions, gestural scores, etc...), and the underlying functional neurobiology such as we understand it from invasive animal studies (cortical columns, delay lines, network dynamics, etc...). In Poeppel's view we are in need of 'linking hypotheses' between the primitives of cognition and neurobiology. So for instance, there are no known descriptions of the neural circuits which would be necessary to perform the Bayesian computations of the sort outlined in [12], or of the acoustic cue-weighting calculus detailed in [13], or of the multi-dimensional scaling deployed in [14].

Under Bayesian accounts, category perception is formalized as rational inference under uncertainty – so for instance, [12] formalize the problem of category identification by reciting Bayes' rule:

$$p(c|s) = \frac{p(s|c)p(c)}{\sum_s p(s|c)p(c)} \quad (1)$$

And identifying c as the set of phonetic categories, and s as the set of speech sounds. So long as the probability distributions are the right shape, the identification curve $p(c|s)$ takes its familiar sigmoid shape, where c is held constant, and s is allowed to vary linearly. Kronrod *et al* successfully show, not only that such an approach works, but that it may be a general solution to the problem of speech perception, demonstrating a good fit for both consonant and vowel identification. Bayesian modelers make no claim to be modeling the relevant neurophysiology, and the approach has a number of obvious benefits, the most impressive being that probability distributions can act as an extremely flexible lingua franca between different types of mental representations and beliefs over different data types.

But whatever a Bayesian-inference neural circuit might actually look like, it seems self-evident that it *must* supervene on a neural representation of the sound itself – i.e., a description of the receptive fields of an ensemble of auditory neurons. *If* the brain computes, say, the likelihood of a speech sound s, given a category c, it must necessarily represent the sound s. The (reductive) line of thinking pursued here is that this level of description – the neurophysiological one – this might be all that's needed for a satisfactory theory of speech perception.

And so this paper approaches the problem in reverse. We begin with a simple mathematical description of the preferred stimuli of auditory neurons – the spectro-temporal receptive field (STRF). The STRF is a real-valued function in the time-frequency plane, which describes those complex auditory objects which excite the neuron, those auditory objects which inhibit the neuron, and those auditory objects to which the neuron is ambivalent [15, 16, 17, 18, 19]. The response of the neuron, $r(t)$, is modeled as the convolution of the STRF with the input

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-1611

spectrogram $S(t, f)$, convolving across time, and integrating across frequency.

$$r(t) = \int \int STRF(\tau, f) S(t - \tau, f) d\tau df \qquad (2)$$

Since the STRF is a 2D function of both time *and* frequency, it allows for the description of neurons whose preferred (or dis-preferred) stimuli are more structurally complex – for example, chirps, or glissandos, or chords. Unlike neurons at the auditory periphery, which tend to have relatively simple, symmetrical tuning curves centered around a single frequency [20, 21], single-unit recordings from the mammalian brainstem and auditory cortex often show complex tunings in the time-frequency plane (e.g., [22, 23]). These tuning functions can be thought of linear operators which perform some computation on the input sound – so for instance, it is possible to construct STRFs which perform edge detection, or spectral band-sharpening, or acoustic feature detection, and so forth [24, 25].

This paper considers STRFs whose function is to perform binary phonemic discrimination. They are constructed rationally using the TIMIT database [26] so that they are tuned to be maximally excited by one value of the contrast, and maximally inhibited by the other value of the contrast. In the sections below, an STRF which discriminates /s/ from /sh/ is constructed, and applied to a linear continuum of synthetic fricatives which span those two categories. It is then demonstrated that this STRF exhibits a non-linear response with respect to this linear acoustic continuum, and that it provides a good match to human labeling data.

2. Method

An STRF which discriminates /s/ from /sh/ was derived from the training portion of the TIMIT corpus [26]. 700 instances each of /s/ and /sh/ were found, and a 300 ms window of speech centered around the midpoint of the fricative was extracted. For each waveform, a 128 frequency bin gammatone-like spectrogram was computed with a 10 ms window and 1 ms time resolution. At each time and frequency bin a 2-way t-test was conducted testing the hypothesis that the average amplitude at that time and frequency is higher for /s/ than for /sh/. The STRF was simply defined as the value of the t-statistic at each point in time and frequency. In other words, in regions of the time-frequency plane where /s/ and /sh/ have an equal amount of energy, the operator is 0. In regions of the time-frequency plane where /s/ *reliably* has more energy than /sh/, the operator is positive, and in regions where /sh/ reliably has more energy, the operator is negative. The STRF was then linearly tuned on held-out data (100 each /s/ and /sh/) so that, on average, both /s/ and /sh/ evoke a response of unit magnitude and opposite sign. The STRF is depicted in figure 1, and its structure in the time-frequency plane is unsurprising: it is excitatory (favors /s/) in high frequencies, inhibitory (favors /sh/) for energy in mid frequencies, and ambivalent for low frequencies.

In order to test the response properties of this toy neuron, a synthetic 11-point fricative continuum was generated which spans the boundary between /s/ and /sh/. The fricatives were modeled as 150 ms long broadband noise. The transfer functions of the fricatives were modeled as gammatone filters, with a constant bandwidth (300 Hz), and logarithmically spaced center frequencies ranging from approximately 2600-4700 Hz. The center frequencies are shown as black notches overlaying the STRF in figure 1, and the transfer functions are plotted in figure 2A. A synthetic 3-formant vowel [i] was generated using the

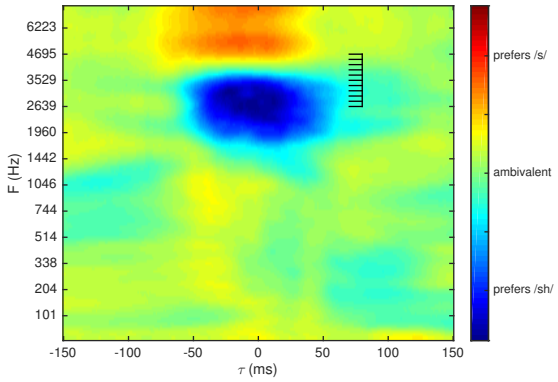

Figure 1: *The spectro-temporal field of a model neuron which discriminates /s/ from /sh/. Blue regions in the time frequency plane are inhibitory – sounds with energy in these regions of the time-frequency plane lower the firing rate of the neuron (an /sh/ response) Red regions in the time frequency plane are excitatory – sounds with energy in these regions increase the firing rate of the neuron (an /s/ response). Green regions are ambivalent – sounds with energy in these regions leave the firing rate at baseline. The black notches indicate the center frequencies of the synthetic fricatives used in the identification experiment. (see figure 2 for a more detailed picture of the experimental stimuli)*

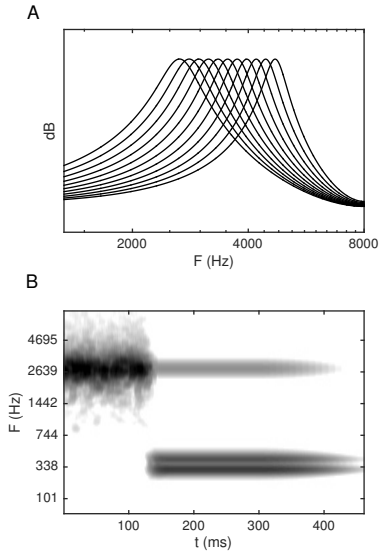

Figure 2: ***A**: Transfer functions of the synthetic fricative continuum used in the discrimination experiment. Each transfer function is a gammatone filter, and their center frequencies are logarithmically spaced, and the bandwidth held constant in Hz.* ***B**: Time frequency representation of an example word used in the discrimination experiment. A 3-formant synthetic /i/ is appended to the fricative creating the percept of the words 'see' or 'she'.*

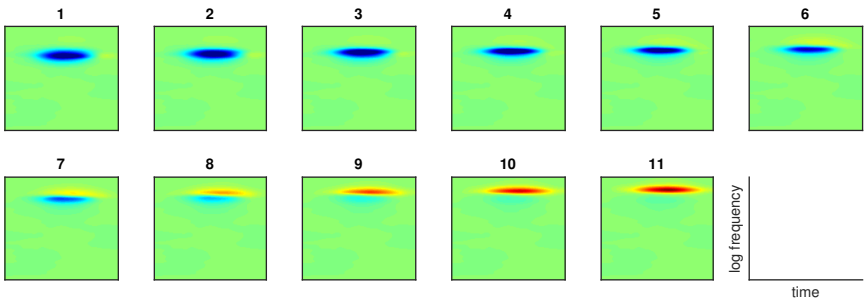

Figure 3: *Result of convolving the /s/-/sh/ STRF with the time-frequency representations of the synthetic fricative continuum. The energy distribution of the fricatives 1-6 fall predominantly in the inhibitory (/sh/-preferring) patch of the STRF, and for 9-11, into the excitatory (/s/-preferring) patch. Stimuli 7-8 straddle the category boundary, evoking both an excitatory and inhibitory response from the model neuron.*

Klatt synthesizer implemented in Praat [27], and was appended to each fricative with a 10 ms cross-fade, creating an (acoustically) continuous cline between the words 'she' and 'see'. A time-frequency representation of one of the stimuli is shown in figure 2B.

Time-frequency representations for each synthetic fricative were estimated using the same parameters as those used to derive the STRF, and the response of the discriminative neuron was calculated for each fricative in the continuum according to equation 2. The responses were then thresholded using the *tanh* function.

In order to assess the accuracy of this model, a simple identification experiment was performed with a single native-English speaking subject. Each synthetic word was presented 30 times, in a random order (330 trials total), and the subject was instructed to identify each word as either 'see' or 'she'. There was a 5s timeout, and 1000 ms of silence $\pm$100 ms between the response and the next stimulus.

3. Results

Figure 3 shows the result of convolving the STRF against the synthetic fricatives, but before integration across the frequency axis. It can be seen that continuum elements 1-6 evoke a negative (/sh/-biased) response, and elements 9-11 evoke a positive (/s/-biased) response. Elements 7-8 – particularly 8 – straddle the category boundary, and their center frequencies are in the 'ambivalent' region of the STRF in between the strongly inhibitory (lower frequency) and strongly excitatory (higher frequency) regions of the STRF. The sidebands of these fricatives can be seen to evoke a simultaneously inhibitory response from the /sh/-patch of the STRF, and an excitatory response from the higher-frequency /s/-patch of the STRF.

Figure 4A shows the result of integrating the outputs of the STRF across frequency bands for all synthetic fricatives in the continuum, and then thresholding with *tanh* to simulate the model neuron's response as a function of time. For all the stimuli the response begins at rest, and by 100ms the responses begin to diverge. By around 200ms, the responses are driven generally to either a strongly excitatory (positive) or inhibitory (response). Figure 4B shows the response of the model neuron extracted at $t = 220$ms as a function of the continuum step. The response shows the classic sigmoid shaped curve typical of categorical perception. Figure 4B also shows the percentage of /s/ responses from the experiment in blue. It can be seen that the data and model match quite well (Pearson's r=0.992), with

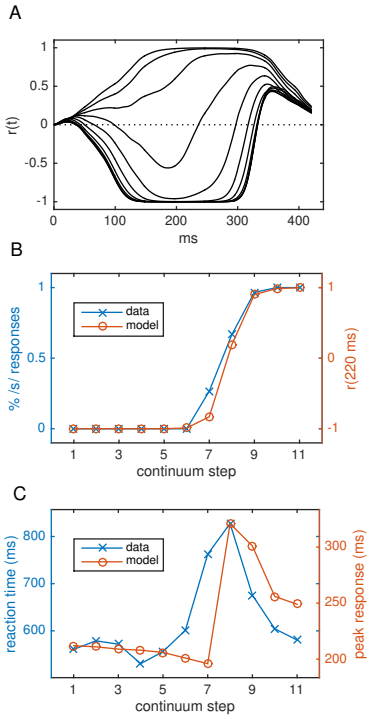

Figure 4: ***A*** *Response functions of the model neuron applied to each of the 11 synthetic fricatives.* ***B*** *Depiction of the response of the model (in red) as a function of continuum step. In blue is plotted the percentage of /s/ responses from a human subject.* ***C*** *Temporal data. The blue trace shows the human reaction time from the identification experiment as a function of the acoustic continuum. The red trace shows the peak response of the model neuron output.*

the category boundary somewhere between stimuli 7 and 8.

To demonstrate that the nonlinear response of the model is not solely driven by the application of the *tanh* thresholding function, figure 5 shows the step-wise finite difference of the model output across adjacent continuum steps, both before and after thresholding. It can be seen that even without the *tanh*

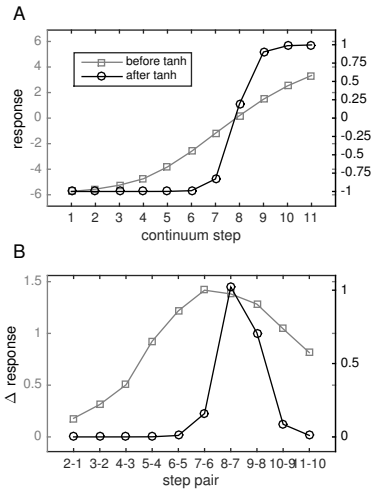

Figure 5: ***A*** *Response of the model neuron at 220 ms as a function of continuum step, both before (gray) and after (black) the application of the* tanh *thresholding function.* ***B*** *The finite difference of the response functions in* ***A****, with the difference computed across adjacent continuum steps, for both before (gray) and after (black) the application of the* tanh *thresholding function. Even before the* tanh *non-linearity, the response of the model neuron is non-linear with-respect to the linear acoustical continuum (otherwise the gray trace would be constant), and the difference peaks near the category boundary.*

non-linearity, the model neuron's response function is not constant with respect to the (linearly varying) acoustics, and peaks somewhere around the category boundary.

Somewhat surprisingly, the model also appears to account for the human reaction time (RT) data. The RT is plotted in figure 4C in blue, and shows the familiar peak at the categorically ambiguous stimulus. However, it can also be seen that the reaction times to the *unambiguous* /s/ sounds (stimuli 9-10-11, identified as /s/ 96%, 100%, 100% of the time) are a bit higher than their unambiguous /sh/ counterparts. And examining the STRF (figure 1) it can be seen that right-most edge of the inhibitory (/sh/-preferring) region of the STRF leads the right-most edge of the /s/-preferring region. Since the operator is causal, and is applied from left to right, this appears to make the prediction that /sh/ may be identified slightly earlier than /s/ since the /sh/-detecting component of the operator leads the /s/-detecting component in time. In order to test this prediction, the time of the peak of the response of the model neuron (defined as the argmax of the absolute value of the response) was calculated for each fricative in the continuum. The results are plotted as the red trace in figure 4C, overlaying the behavioral data. The model pro fits the human data reasonably well (Pearson's r=0.587), with both functions peaking on the ambiguous token.

4. Conclusion

This study illustrates the feasibility of a model of phonetic category perception formulated in terms of the computational primitives of auditory neurons, and demonstrates that such a model can provide a good fit to empirical data. The approach shares some features with extant models of speech perception. Like Bayesian accounts, it is the variation (and reliability) of previously observed speech signals which shape the percept recovered by the listener. However, unlike Bayesian accounts, there is no probabilistic calculus in the model itself. Like exemplar accounts, the model commits itself to the possibility that phonemic detail of arbitrarily high resolution may be deployed in making categorical judgements – since the operators are simply defined as real-valued functions over the time-frequency plane, the model is quite free, and permits the possibility of extremely detailed spectro-temporal objects driving perception. However, unlike exemplar theories, there is no mass of tokens which need to be stored in long-term memory, there are simply the receptive fields of auditory neurons which have been shaped by experience to optimally categorize speech sounds.

5. Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada, grant PGS-D3-443975-2013, and enriched by stimulating discussion with William Idsardi, Naomi Feldman and Ellen Lau.

6. References

[1] A. Liberman, K. Harris, H. Hoffman, and B. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, no. 5, pp. 358–368, 1957.

[2] A. Liberman, F. Cooper, D. Chankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychological Review*, vol. 74, pp. 431–461, 1967.

[3] P. Kuhl and J. Miller, "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants," *Science*, vol. 190, p. 69, 1975.

[4] A. Liberman and I. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[5] N. Feldman, T. Griffiths, and J. Morgan, "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference," *Psychological Review*, vol. 116, no. 4, pp. 752–782, 2009.

[6] J. Arsenault and B. Buchsbaum, "Distributed neural representations of phonological features during speech perception," *The Journal of Neuroscience*, vol. 35, no. 2, pp. 634–642, 2015.

[7] E. Chang, J. Rieger, K. Johnson, M. Berger, N. Barbaro, and R. Knight, "Categorical speech representation in human superior temporal gyrus," *Nature Neuroscience*, vol. 13, pp. 1428–1432, 2010.

[8] E. Myers, S. Blumstein, E. Walsh, and J. Eliassen, "Inferior frontal regions underlie the perception of phonetic category invariance," *Psychological Science*, vol. 20, no. 7, pp. 895–903, 2009.

[9] N. Mesgarani, C. Cheung, K. Johnson, and E. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, pp. 1006–1010, 2014.

[10] D. Poeppel and D. Embick, "The relation between linguistics and neuroscience." in *Twenty-first century psycholinguistics: four cornerstones*, A. Cutler, Ed. Hillsdale, NJ: Lawrence Erlbaum Associates Inc., 2005, pp. 103–120.

[11] D. Poeppel, "The *maps* problem and the *mapping* problem: Two challenges for a cognitive neuroscience of speech and language," *Journal of Cognitive Neuropsychology*, vol. 29, no. 1-2, pp. 34–55, 2012.

[12] Y. Kronrod, E. Coppess, and N. Feldman, "A unified account of categorical effects in phonetic perception," *Psychonomic Bulletin & Review*, vol. 6, pp. 1681–1712, 2016.

[13] L. Holt and A. Lotto, "Cue weighting in auditory categorization: Implications for first and second language acquisition," *Journal of the Acoustical Society of America*, vol. 119, no. 5, 2006.

[14] P. Iverson and P. Kuhl, "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling," *Journal of the Acoustical Society of America*, vol. 99, pp. 1130–1140, 1995.

[15] A. Aertsen and P. Johannesma, "The spectro-temporal receptive field. a functional characteristic of auditory neurons," *Biological Cybernetics*, vol. 42, pp. 133–143, 1981.

[16] D. Depireux, J. Simon, D. Klein, and S. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiology*, vol. 85, no. 3, pp. 1220–34, 2001.

[17] J. Eggermont, P. Johannesma, and A. Aertsen, "Reverse-correlation methods in auditory research," *Quarterly Reviews in Biophysics*, vol. 16, pp. 341–414, 1983.

[18] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectro-temporal receptive fields in primary auditory cortex." *Nature Neuroscience*, vol. 6, pp. 1216–1223, 2003.

[19] F. Theunissen and J. Elie, "Neural processing of natural sounds," *Nature Reviews Neuroscience*, vol. 15, no. 6, pp. 355–366, 2014.

[20] N. Kiang, M. Sachs, and W. Peake, "Shapes of tuning curves for single auditory-nerve fibers," *Journal of the Acoustical Society of America*, vol. 42, no. 6, pp. 1341–1342, 1967.

[21] S. Narayan, A. Temchin, A. Recio, and M. Ruggero, "Frequency tuning of basilar membrane and auditory nerve fibers in the same cochleae," *Science*, vol. 282, pp. 1882–1884, 1998.

[22] S. Andoni, N. Lin, and G. Pollack, "Spectrotemporal receptive fields in the inferior colliculus revealing selectivity for spectral motion in conspecific vocalizations," *Journal of Neuroscience*, vol. 27, pp. 4882–4893, 2007.

[23] C. Atencio and C. Schreiner, "Spectrotemporal processing in spectral tuning modules of cat primary auditory cortex," *Plos ONE*, vol. 7, p. e31537, 2012.

[24] T. Lindeberg and A. Friberg, "Idealized computational models for auditory receptive fields," *PLoS ONE*, vol. 10, no. 3, p. e0119032, 2015.

[25] N. Mesgarani, C. Cheung, K. Johnson, and E. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 423, pp. 1006–1010, 2014.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus," 1993.

[27] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer*, 2015, version 5.4.08. [Online]. Available: http://www.praat.org/