Specific Aims

How does the auditory system produce invariant perception of a phoneme despite highly variable acoustic information? Phonemes cannot be identified by any invariant acoustic features, and this “non-invariance problem” poses a fundamental challenge in speech perception. Here we focus on an important test case, the categorization of consonants in different vowel contexts — for example, /g/ in "gah" or "gee". This case is widely considered to be a litmus test for the non-invariance problem [1]. Coarticulation causes the physical sound of a consonant (such as /g/) to be remarkably different depending on the following vowel, yet still be perceived as an invariant phoneme. Because individual acoustic features are not sufficient to correctly classify consonants, the brain must make use of multiple imperfect cues. Competing auditory speech perception theories for how this problem might be solved are largely unconstrained by neurobiological data. A critical barrier has been that fMRI and EEG methods for measuring human brain activity are simply not up to the task, because speech is too fast and the neural circuitry involved is too small. There is therefore a critical need for high-resolution measurement and manipulation of neural activity during discrimination of speech sounds in order to determine how actual neural circuits solve the non-invariance problem in real time [2]. Identifying the brain mechanisms of speech processing and perception is a fundamental goal of speech, language, and hearing research.

Although early speech theorists proposed that "speech is special" [3], today many argue that early stages of speech processing in the human brain are likely implemented by evolutionarily-conserved auditory processing mechanisms found in any mammalian auditory system [1, 4-6]. Here we propose to train mice to discriminate speech sounds, and then use optogenetics, GCaMP calcium imaging, and tetrode recordings to understand the underlying neural mechanisms they use to do so. Our long-range objective is to develop and test neurobiologically plausible models of human speech perception based on well-understood neural sound processing systems in mice. The objective of this proposal is to determine how actual auditory neural circuitry can solve the non-invariance problem for stop consonant categorization. By training mice to generalize stop consonant categories across vowel contexts, we have established a model system in which we can apply the powerful set of tools available in the mouse. We have an extensive track record of using these approaches to elucidate the cellular and synaptic mechanisms of coding transformations and complex sound processing, making us well-positioned to achieve this objective [7-10].

Aim 1. Train mice to discriminate stop consonants in different vowel contexts. Pilot studies demonstrate that mice learn to categorize consonants and rapidly generalize to novel vowel contexts, speakers, tokens (individual utterances), and speaking rates, providing a strong test of non-invariant categorization. Pilot optogenetic silencing demonstrates that auditory cortex is required.

Aim 2. Determine the cortical representation of stop consonants. We will use widefield calcium imaging to map phoneme selectivity across all auditory cortical fields and surrounding cortex in trained GCaMP6-expressing mice. We expect to find local selectivity for distinctive features such as place of articulation (e.g. labial, velar) and manner of articulation (e.g. stop, fricative), as has been seen with intracranial recordings in humans [11]. Alternatively, the representation of phoneme identity could be dense and distributed, such that phoneme selectivity is seen only with broad multi-pixel pattern analysis.

Aim 3. Determine how cortical neurons encode phoneme identity. We will use tetrode recordings in behaving speech-trained mice, targeted to auditory cortex and other areas identified in Aim 2. We will measure selectivity for intact speech sounds, as well as components (such as formants, formant transitions, and stop releases) presented individually and in combination. We expect to find combination-sensitive neurons that preferentially respond to specific combinations of features, as predicted by prominent theories of speech perception [1, 6]. Alternative theories predict the existence of template-matching cells, selectivity for manner and/or place of articulation [11], or a distributed population code [2, 12].

Impact: The proposed research will establish a powerful model system in which to tackle a broad array of outstanding problems in speech processing. We expect to learn the nature of the neural code for stop consonants, which will strongly constrain possible neurobiological models of speech processing. This project is innovative because it combines a new paradigm for the non-invariance problem in speech perception with the powerful tools now available in mice, such as optogenetics, calcium imaging, and electrophysiology. This will lay the foundation for layer-specific and cell-type-specific circuit analysis of how the auditory system overcomes non-invariance in consonant identification, as well as other hard non-invariance problems in speech processing, such as /r/ vs. /l/ discrimination, vowel identification, and categorical perception.

Significance

To listen to speech is to be fooled much of the time. Sounds that are actually different are perceived as the same, such that perception is invariant even though the stimulus is non-invariant. This so-called non-invariance problem remains a fundamental challenge for theories of speech perception. Despite the fact that speech perception is one of the most important cognitive operations performed by the human brain, we know surprisingly little about the neural computations underlying speech processing. As a result, competing models for how speech is processed by the brain are largely unconstrained by neurobiological data.

A core problem arising from phonetic context is coarticulation, in which the actual sound of a speech segment like /g/ or /d/ depends greatly on neighboring segments such as vowels (Fig. 1a). For example, say the words *goose* and *geese*. Notice that your lips are rounded in goose even before you speak, but not with geese. This is coarticulation, which causes the acoustic features of /g/ to be quite different depending on the following vowel. Yet we are unaware of this difference, and perceive both sounds as the same phoneme, /g/. The acoustic consequences of coarticulation are that the speech signal, at any moment in time, is 'colored' by the speech uttered before and after. As a result, no unique set of acoustic attributes can identify /g/, or any consonant. Categorizing stop consonants (that is, /t/, /d/, /k/, /g/, /p/, or /b/) in different vowel contexts is widely considered to be the critical test case for non-invariance, a problem hard enough to stymie most theories of speech perception yet one that is effortlessly solved by a two-year old.

Because individual acoustic features are not sufficient to correctly classify stop consonants, the brain must make use of multiple imperfect cues. This is a general problem faced by all sensory systems, which excel at using multiple sources of inconsistent or noisy data to solve perceptual problems. There are numerous models for how this non-invariance problem might be solved in speech perception, which all perform about as well (or as poorly) as each other [1, 13-15]. As a result we have no way to distinguish among most theories of speech perception. A critical barrier has been that fMRI and EEG methods for measuring human brain activity are simply not up to the task, because speech is too fast and the neural circuitry involved is too small. There is therefore a critical need for high-resolution measurement and manipulation of neural activity during discrimination of speech sounds in order to determine how actual neural circuits solve the non-invariance problem in real time.

Premise: Here we propose to achieve this by using the mouse as a model system, allowing us to use optogenetics, calcium imaging, electrophysiology, and quantitative behavior to uncover how actual neural circuitry solves the non-invariance problem. A key advance that makes this possible is our behavioral paradigm, in which mice learn to categorize stop consonants and generalize across different vowel contexts, speakers, and tokens (Fig. 2), a novel and important finding. Mice can thus solve the non-invariance problem; the proposed research seeks to determine how. Although early theorists proposed that "speech is special" [3], today many argue that early stages of speech processing in the human brain are likely implemented by evolutionarily-conserved auditory processing mechanisms, such as those found in any mammalian auditory system [1, 4-6]. The proposed research will reveal how auditory neural circuitry can solve the non-invariance problem in speech perception.

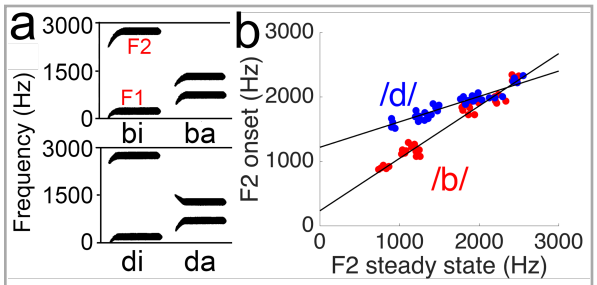

Fig. 1. The non-invariance problem. **A**: Simplified spectrograms of /b/ and /d/ in different CV contexts. The spectrotemporal features of each consonant differ markedly across contexts, but are perceived invariantly. **B**: The F2 formant transition contains combinatorial information that can partly distinguish consonants. Each dot is a token from a different CV context. Data replotted from [1]. Neurons sensitive to these feature combinations could contribute to consonant categorization.

A broad limitation of this premise is that mice could use completely different neuronal mechanisms to solve the non-invariance problem than humans do. Since the neuronal mechanisms underlying speech processing in humans are largely unknown, successful completion of these Aims will not reveal whether mice and humans use similar mechanisms. The proposed research would thus generate predictions that could be tested in humans, demonstrating impact on the field.

We consider 3 models for how neurons might solve the non-invariance problem: **(1)** The Combination-Sensitivity (CS) model proposes that combination-sensitive neurons selectively respond to specific combinations of acoustic features [1, 6, 16, 17], much like ITD-selective neurons in the inferior colliculus, or echo-pulse delay-tuned neurons in bats.

For example, stop consonants are at least partly distinguishable by distinct linear relationships between onset and steady-state second formant (F2) frequencies (Fig. 1b). Indeed, the F2 transition has been called the single most important cue in speech perception [1, 18]. The CS model predicts the existence of neurons selective for such F2 feature combinations. The CS model acknowledges that binary combinations are ambiguous in some cases (for example, where red and blue dots overlap in Fig. 1b), and that additional features (such as F3 or plosive burst spectra) are likely involved. An extended CS model therefore predicts that neurons are sensitive to multiple cues. **(2)** The Template-Matching (TM) model is based on the relative shapes of the short-term onset spectra of stop consonants [13]. Like the CS model, the TM model is a relational model because consonants cannot be distinguished by any single acoustic feature. The TM model proposes that consonants can be categorized by matching to an appropriately scaled and shifted template. **(3)** The Distributed Coding (DC) model is based on the observation that neurons in primary auditory cortex (A1) and neighboring tonotopic cortical fields tend to respond broadly to most speech sounds, but with enough variation in temporal structure to form a dense and distributed spatiotemporal code for phoneme identity (Fig. 5, [12, 19, 20]). The DC model proposes that the distributed population code is sufficient by itself, without the need for explicit combination-sensitive or template-matching neurons. In contrast, the CS and TM models propose explicit coding of phoneme information, possibly organized hierarchically in higher-order cortical areas [2]. Distinguishing between these neurobiological theories of speech perception will require measuring neural activity with cellular resolution in several auditory cortical fields during phoneme discrimination [2].

Evidence supporting the explicit coding of phoneme information comes from intracranial field potential recordings from the superior temporal gyrus (STG) in epilepsy patients, which reveal topographic selectivity for specific groups of phonemes [11]. Individual STG electrodes were selective for manner of articulation (e.g., stops vs. fricatives) and for place of articulation (e.g. labial (/b/) vs. velar (/g/). This finding raises several important questions: Does phoneme selectivity extend to cellular resolution? Do neighboring neurons have similar phoneme selectivity? Would similar phoneme selectivity arise in a generic mammalian auditory system performing phoneme categorization? Multi-unit recordings in auditory cortex in trained, pentobarbital-anesthetized rats indicate a distributed code, rather than explicit phoneme selectivity [12]. Does this finding extend to behaving animals? Are speech responses more specific in awake animals, and in single-neuron recordings? Do cortical regions beyond core auditory cortex show explicit coding in trained animals, as seen in human STG? Are there intermediate representations that progressively encode speech more sparsely and efficiently along a cortical hierarchy? The proposed research addresses each of these questions.

Impact: The non-invariance problem has dominated theoretical debate in speech research for at least 50 years [1, 21-23]. Establishing a system in which the powerful tools of modern systems neuroscience can be brought to bear on this and related problems will have a transformative impact on the field, because it has the potential to resolve this longstanding debate. Indeed, bridging the interface between auditory neuroscience and linguistics will have a significant impact on both fields. Moreover, identifying the brain mechanisms of speech processing and perception is a fundamental goal of speech, language, and hearing research. The proposed research is broadly significant because the neuronal mechanisms involved in speech perception underly the pathogenesis of a wide array of acquired and developmental language disabilities, including aphasia, specific language impairment, autism, and dyslexia. A deeper understanding of these mechanisms will have direct implications for prevention, diagnosis and treatment of these disorders.

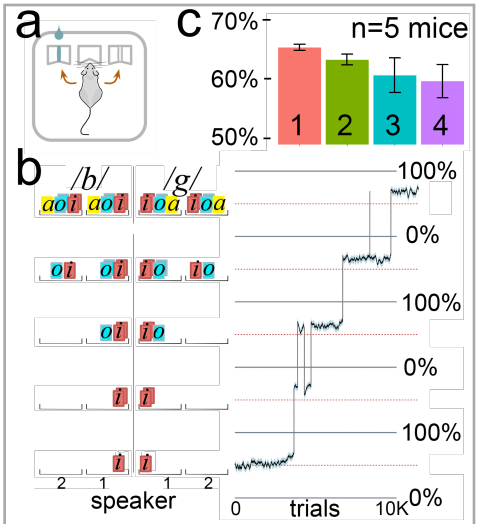

Fig. 2. Behavioral paradigm. **A:** 2-alternative choice task. Mice nose-poke for water rewards. **B:** Example mouse progressing through training steps. Stimuli at left (colored boxes: vowels and tokens). New vowels and speakers are introduced in steps. Finally, generalization is tested with completely novel speakers and vowel contexts on 20% of trials. C: Pilot data. N=5 mice are able to generalize at 60-65% 1: highest training difficulty. 2: More tokens. 3: More speakers. 4: More vowels. *Saunders & Wehr, unpublished data*.

Innovation

1. *Critical barrier:* Although rats have been trained to discriminate speech sounds [12, 24-27], and quail have even been shown to generalize to novel vowel contexts [14, 28], only single speakers and single tokens have been tested. The hard version of the non-invariance problem is whether and how the auditory system

can generalize to novel speakers (both male and female), novel tokens, and different speaking rates, in addition to novel vowel contexts. *Solution*: We show for the first time that mice rapidly generalize to novel speakers (both male and female), tokens, and speaking rates, and vowel contexts (Fig. 1), allowing direct tests of the hard version of the non-invariance problem.

2. *Critical barrier:* Existing animal models (such as rats or quail) do not permit cell-type-specific, reversible manipulations of activity during behavior in order to test causal models of circuit function. *Solution*: Optogenetic cortical suppression in transgenic Cre lines of mice permits temporally precise, reversible, and cell-type-specific circuit analysis during behavior.
3. *Critical barrier:* Non-invasive fMRI or EEG methods in humans have inadequate spatiotemporal resolution. Furthermore, serial recordings in animals are laborious and inefficient for physiologically mapping multiple widespread cortical regions. *Solution*: We will use widefield calcium imaging in GCaMP6-expressing mice to measure cortical activity over nearly the entire temporal lobe with high spatial and temporal resolution (Fig. 4).
4. *Critical barrier:* Electrophysiology in trained rats has been performed under anesthesia (e.g., pentobarbital), which reduces activity to mainly onset responses, and prevents the trial-by-trial comparison of neuronal responses with behavioral performance. Furthermore, recordings have been largely limited to local field potentials or multi-unit activity, which may not be fine-grained enough to reveal highly selective response properties like combination sensitivity. *Solution*: We will record single-neuron activity in awake behaving mice during consonant discrimination for the first time.
5. *Critical barrier:* Electrophysiology in humans has required intracranial ECoG recordings in intractable epilepsy patients, greatly limiting both the available subject pool and its accessibility to researchers. *Solution*: Our paradigm allows high-throughout single-unit electrophysiology and imaging in mice performing consonant categorization.

This will lay the foundation for future layer-specific and cell-type-specific circuit analysis of how the auditory system overcomes non-invariance in stop consonant identification, as well as other hard non-invariance problems in speech processing.

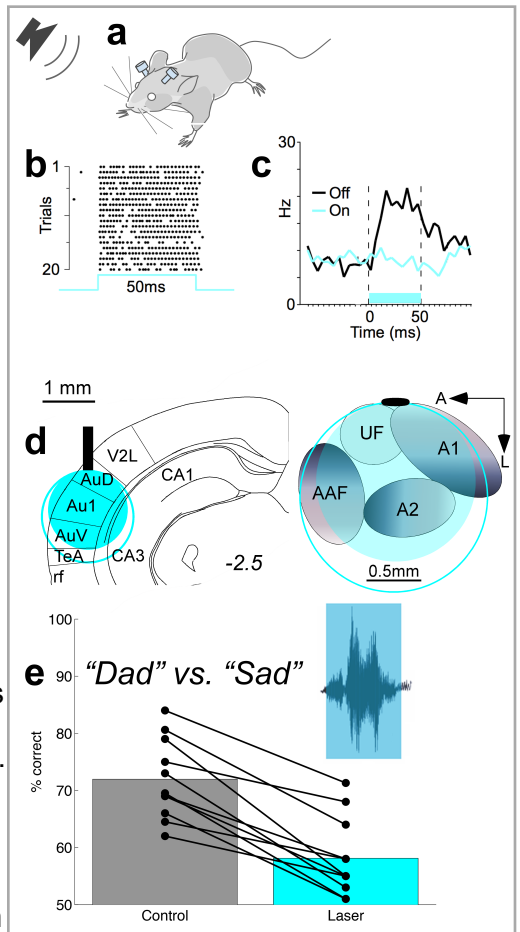

Fig. 3. Optogenetic shutdown of auditory cortex impairs phoneme discrimination. **A:** Bilaterally implanted fibers in PV-Cre mice. **B-C:** Laser strongly activates PV+ inhibitory neurons (B), which strongly suppresses all non-PV neurons (C) [9]. **D:** We calibrate laser intensity to control the spread of suppression (blue shaded region). **E:** Embedding phonemes in a laser pulse (inset) impairs discrimination (n=11 mice, p=0.0005). *O'Sullivan & Wehr, unpublished data*.

Approach

Aim 1. Train mice to discriminate stop consonants in different vowel contexts.

Hypothesis: The hard version of the non-invariance problem can be solved by the circuitry in any generic mammalian auditory system.

Rationale: The hard version of the non-invariance problem is whether and how the auditory system can generalize to novel vowel contexts, speakers (both male and female), tokens, and different speaking rates. It's hard because individual acoustic features are not sufficient to correctly classify stop consonants. We have successfully trained mice to do this, laying the foundation for identifying the neural circuitry involved in Aims 2 and 3. Optogenetic shutdown demonstrates that A1 is necessary for the task, motivating us here to go beyond A1 and test the necessity of other auditory cortical fields and higher cortical areas.

Experimental approach: Mice perform a 2-alternative choice task for a water reward (Fig. 2). We use a shaping procedure in which an easy pure-tone discrimination task is incrementally paired with and ultimately replaced by phoneme discrimination. Speech sounds are consonant-vowel (CV) pairs. Mice discriminate consonants (/b/ vs /g/) in different vowel contexts. We pitch-shift speech sounds up into the mouse hearing range. Mice progress through the introduction of additional speakers (2 males and 2

females), 2 tokens (utterances), 6 vowel contexts, and 2 speaking rates (fast or slow). In the final stage they are challenged on 20% of trials with completely novel speakers, tokens, and vowel contexts. Preliminary data: N=5 mice generalize at 60-65% correct (Fig 2).

Optogenetic shutdown: We implant chronic optical fibers bilaterally over A1 or other auditory cortical fields, in transgenic PV-ChR2 mice, as we have done previously [9, 10]. Activation of PV+ inhibitory neurons evokes profound suppression of all excitatory neurons throughout the depth of cortex (Fig. 3). We calibrate laser intensity to restrict suppression to the target area (Fig. 3). We have previously validated this suppression method using both tungsten microelectrodes and chronic tetrodes [9, 10]. We quantify effects on task performance with ROC analysis. Suppression of A1 significantly impairs phoneme discrimination (Fig. 2, p=0.0005, n=11 mice). We will use multi-fiber arrays to test the effects of suppression targeted to A1, anterior cortical fields (AAF, UF), lateral fields (A2), as well as additional areas identified as phoneme-responsive in Aim 2. We will also test for lateralization by individually suppressing right or left cortical regions. All of these methods are already in use [9, 10], so we do not anticipate any technical challenges.

Interpretation: Based on our preliminary data (Fig. 2), we predict that mice will learn to categorize the consonants /g/ and /d/ in different vowel contexts, and will generalize to novel speakers, tokens, vowels, and speaking rates. This would support our hypothesis that mice can use their generic mammalian auditory system to solve the hard non-invariance problem. Our preliminary suppression data (Fig. 3) provides a very strong basis for expecting that A1 suppression will significantly impair discrimination. The effects of suppression of cortical fields beyond A1 are completely unknown, but based on the phoneme selectivity seen in human STG, we predict that suppression of anterior cortical fields will impair phoneme discrimination. Our interpretation here would be that A1 is necessary, and that downstream higher-order areas are also necessary for phoneme discrimination. Alternatively, if we find that A1 is the only necessary region, this would suggest that a distributed representation in A1 supports categorization without requiring explicit coding for combinations of features in higher areas.

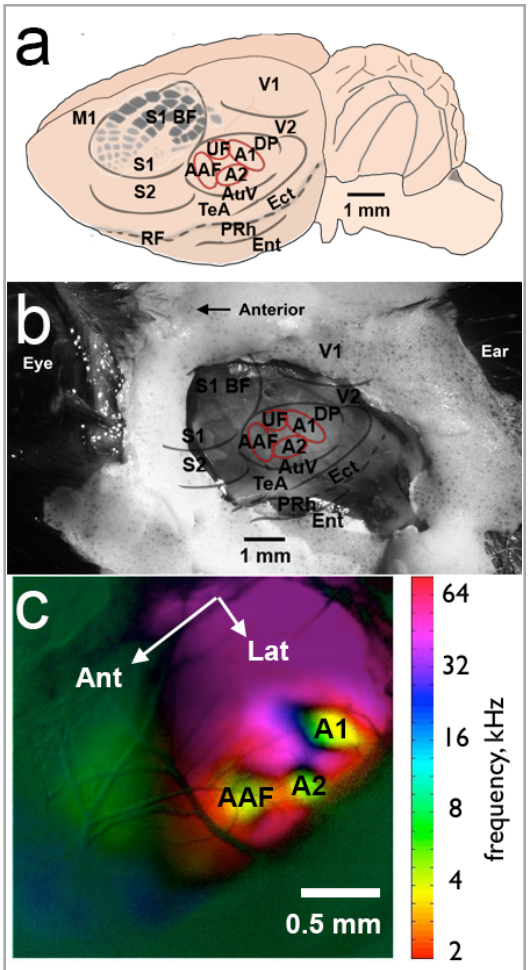

Fig. 4. Widefield GCaMP6 imaging. **A:** Lateral surface of the mouse brain. S1 BF: barrel cortex. AuV: ventral auditory cortex. PRh: perirhinal cortex. RF: rhinal fissure, Ent: entorhinal cortex. Ect: ectorhinal cortex. TeA: temporal association cortex. **B:** Clear-skull imaging preparation with ~5 mm window surrounded with white acrylic. **C**: Pilot widefield imaging of tone responses in a small window (~2 mm) reveals tonotopic regions A1, A2, AAF in anesthetized mouse. Pixel size 26 µm. *O'Sullivan, Saunders, & Wehr, unpublished*.

Potential pitfalls: Mice might learn the set of all individual stimulus-response pairs during training, rather than the consonants *per se*. The fact that they rapidly generalize to new speakers, tokens, and vowel contexts argues against this possibility.

Aim 2. Determine the cortical representation of stop consonants.

Hypothesis: Phonemes are encoded with distributed representations in A1, but with local selectivity for feature combinations in higher auditory cortical areas.

Rationale: Distributed coding for phonemes has been seen in early auditory cortical fields (A1, AAF, SRAF) using multi-unit recordings in anesthetized rats [12, 19, 20, 24]. Local selectivity for higher-order acoustic features of phonemes (manner and place of articulation) has been seen in higher auditory cortical areas in human STG [11]. Do different CVs activate overlapping regions of auditory cortex, or distinct regions? Do mice that can generalize to consonant categorization in novel contexts show explicit coding for higher-order features in cortical regions beyond early auditory cortex, or is a distributed population code sufficient for solving the task? Candidate higher cortical areas within our field of view include anterior auditory fields, the ventral auditory pathway (implicated in auditory object recognition [29]) and perirhinal cortex (implicated in recognition memory and feature conjunction selectivity [30]).

Experimental approach: We will use widefield calcium imaging to map phoneme selectivity across all auditory cortical fields and surrounding cortex in trained mice (Fig. 4). Our CamK2-GCaMP6 mice express GCaMP6 broadly in excitatory neurons throughout the cerebral cortex, providing high signal-to-noise (routinely 20% ΔF/F) [31]. We will chronically image in a window 5-10 mm in diameter, encompassing the entire lateral face of the temporal lobe and adjacent portions of frontal, parietal, and occipital lobes, using a custom tandem-lens fluorescence macroscope (Fig. 4). Mice are trained as described in Aim 1, but are head-fixed and passively listening to speech sounds during imaging. We will present intact CVs, as well as their component acoustic features (such as formants, formant transitions, and stop releases) presented individually and in combination. We will also present tones and white noise to relate to known topographical organization (Fig. 4). This method provides a global picture of cortical activity but does not provide cellular resolution, which will be pursued in future studies using 2-photon microscopy. Developing a head-fixed version of the discrimination task is another future direction. Preliminary data in anesthetized mice reveal multiple tonotopic fields (A1, A2, AAF), demonstrating the feasibility of this approach (Fig. 4c). To test for local phoneme selectivity, we will compute the Phoneme Selectivity Index (PSI, [11]) of each pixel. The PSI (which ranges from 0 to 1) is high for pixels that respond selectively to one or a few phonemes, and is low for pixels that respond unselectively to most phonemes. To test for phoneme information that is represented by a distributed code, we have adapted the multi-voxel pattern analysis (MVPA) technique from fMRI. In this approach we train a multivariate classifier (support vector machine) on a subset of data, and then test on remaining data. Successful classification would indicate that information is contained in the distributed activity pattern across pixels. This analysis also identifies which pixels are most informative. Note that a pixel can be informative even if it doesn't have high selectivity, which would indicate that it participates in a distributed code.

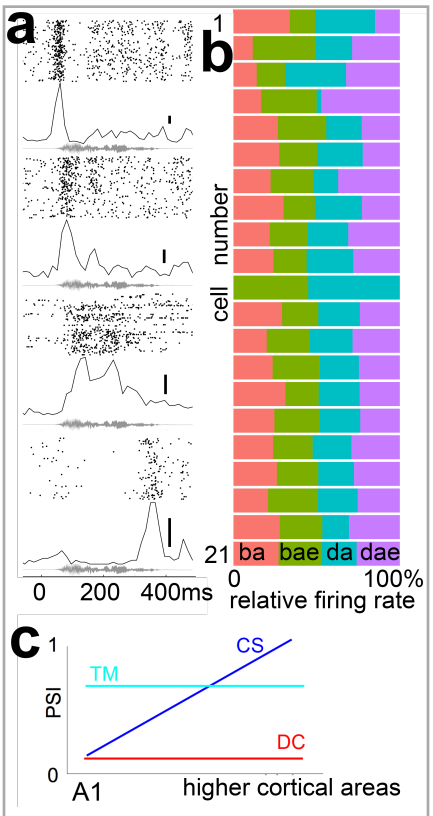

Fig. 5. Distributed phoneme coding in A1. **A:** Responses of 4 different A1 neurons to the CV "ba." Scale bar, 10 Hz firing rate. Tetrode recordings in awake mice. **B:** Relative evoked firing rates across 21 A1 neurons and 4 phonemes. **C:** PSI predicted by each model across cortical hierarchy. *Yavorska, Saunders, & Wehr, unpublished*.

Interpretation: In higher auditory cortical areas, we expect to find high PSI values, indicating local selectivity for distinctive features such as place of articulation (e.g. labial, velar) and manner of articulation (e.g. stop, fricative), as has been seen with intracranial recordings in humans [11]. Observing this in a non-human brain would be completely novel and highly significant because it would demonstrate similar brain mechanisms for phonetic processing across species. In this case, phoneme-selective regions will be targeted for tetrode recordings in Aim 3. In A1 and early auditory cortical areas (A2, AAF), we expect the representation of phoneme identity to be dense and distributed, such that phoneme-selective information is seen only with multi-pixel pattern analysis. Alternatively, we could find only distributed coding in all brain areas, which would be inconsistent with our hypothesis. In this case, regions that are highly informative for phoneme classification will be targeted for tetrode recordings in Aim 3. Given that phonemes evoke multi-unit responses broadly across multiple cortical regions [12, 19, 20, 24], the sensitivity of MVPA methods makes it highly likely that we will identify informative regions, indicating a strong likelihood of success for this Aim. Finding local selectivity in early auditory fields such as A1 would be unexpected (and disprove our hypothesis) but novel and exciting.

Potential pitfalls: If we find only distributed coding, this does not rule out the possibility that explicit coding could exist in a brain structure inaccessible to imaging (i.e. deep structures such as IC, MGB, or striatum). Because our cortical optogenetic shutdown indicates that auditory cortex is involved, this possibility seem unlikely, but in this case we could use the electrophysiology methods of Aim 3 to target deep structures.

Aim 3. Determine how cortical neurons encode phoneme identity.

Hypothesis: Single-neuron phoneme selectivity shows hierarchical organization, with single-feature selectivity in A1 and combination sensitivity in higher cortical areas (i.e., the CS model).

Rationale: In Aim 3 we will record with cellular resolution while mice categorize consonants to directly test competing models of speech perception. These models make distinct predictions about which features are used and how they are combined to identify consonants.

Experimental approach: We will use tetrode recordings in behaving speech-trained mice, targeted to auditory cortex and other areas identified in Aim 2. We implant mice with 8 tetrodes for chronic recordings (typically ~8 high-quality neurons per day, for multiple weeks), as we have done previously [9, 10]. Thus this Aim is not as ambitious as it might seem, because we will be able to rapidly collect a large data set of single-neuron recordings in multiple areas within a 2-year time frame. Tetrode placements are verified histologically. We will measure selectivity for intact speech sounds, as well as components (such as formants, formant transitions, and stop releases) presented individually and in combination. Neuronal selectivity will be quantified using the Phoneme Selectivity Index (PSI, [11]), analogous to the PSI for pixels in Aim 2. We will also present tones and white noise to characterize standard neuronal response properties (e.g., STRFS). We will test whether STRFs derived from tones can predict the responses to intact phonemes and their component features, as we have done previously for natural sound stimuli [32].

Interpretation: Results predicted by each model are schematized in Fig. 5c. Based on preliminary data (Fig. 5) and previous studies [19, 20, 24], we expect that neurons in A1 and neighboring tonotopic fields (A2, AAF) will respond to nearly all phonemes tested (low PSI), but will be selective for individual acoustic features as predicted by their STRFs. In higher cortical areas, we expect to find combination-sensitive neurons that preferentially respond to specific combinations of features (high PSI), as predicted by the CS model [1]. In particular, one specific prediction of the CS model is the existence of neurons sensitive to the combination of the second formant (F2) frequency at two time points during the formant transition (onset and steady-state, Fig. 1b) [1, 6]. More generally, because CS neurons respond to the combination of components but not to those components in isolation (a nonlinear property), STRFs would not predict their phoneme responses. Alternatively, the Template-Matching model [13] predicts the existence of template-matching neurons, which should be phoneme-selective (high PSI) and for which the STRF should successfully predict their phoneme responses. Comparing neuronal STRFs with the weighted average spectrogram of the phonemes they are selective to will test the extent to which phonetic selectivity results from tuning to spectrotemporal features. Another alternative outcome is that neurons in higher cortical areas could encode speech sounds based on manner and/or place of articulation, similar to field potential recordings in human STG [11] but extended to cellular resolution. In this case we would test whether neighboring neurons have similar phoneme selectivity, which would indicate topographic representation for these phoneme attributes. None of these models are mutually exclusive; it is possible that distinct subsets of neurons could provide evidence supporting different models. We would interpret this as indicating that the auditory system can use multiple mechanisms in a hybrid fashion to categorize consonants.

Potential pitfalls: We could find that phonemes are represented and categorized with a distributed population code at all levels and in all cortical areas tested, similar to the distributed code seen in A1 and AAF [2, 12]. This would be inconsistent with our hypothesis of hierarchical organization, but absence of evidence would not prove that an explicit code for combinations of features, manner, or place of articulation is not used somewhere else in the brain. However, this result would rule out an explicit code in tested cortical areas, thus disproving the cortical hierarchy proposed by the CS and TM models.

Future directions: The research proposed here will provide significant insight into the brain mechanisms underlying consonant categorization, widely considered to be the hardest test case of the non-invariance problem in speech perception. With this foundation, we will be poised to address deep questions about the neural mechanisms of speech perception in future R01 research:

- If combination-sensitive neurons exist, we will seek to identify optogenetic strategies to manipulate their activity to test their causal role in driving speech perceptual decisions.
- We will compare phoneme selectivity in naive and trained mice, and during task learning, to determine whether and how selectivity develops with learning.
- Using layer-specific and cell-type-specific Cre lines (as we have done previously [9, 10, 33, 34]), we will seek to identify the cortical circuitry involved in transforming phoneme selectivity from a distributed code to a higher-order feature-selective code.
- Using *in vivo* whole-cell recordings, we will seek to identify the cellular and synaptic mechanims that shape phoneme selectivity, as we have for other auditory neural computations [7, 8, 35, 36].
- Using 2-photon microscopy, we will image neuronal activity with cellular resolution from up to hundreds of neurons simultaneously, allowing detailed topographic analysis of the local or distributed nature of phoneme selectivity. To enable simultaneous imaging and behavior, we will develop a head-fixed version of the phoneme discrimination task.
- We will use this powerful suite of methods to tackle other hard non-invariance problems in speech perception, such as /r/ vs. /l/ discrimination, vowel identification, and categorical perception.

References cited

1. Sussman, H.M., D. Fruchter, J. Hilbert, and J. Sirosh, *Linear correlates in the speech signal: the orderly output constraint.* Behav Brain Sci, 1998. **21**(2): p. 241-59; discussion 260-99.

2. Schreiner, C.E., *Input limitations for cortical combination-sensitive neurons coding stop-consonants?* Behavioral and Brain Sciences, 1998. **21**(2): p. 284.

3. Liberman, A., *Speech: A special code.* MIT Press, 1996.

4. Kuhl, P.K., *Theoretical contributions of tests on animals to the special-mechanisms debate in speech.* Exp Biol, 1986. **45**(3): p. 233-65.

5. Poeppel, D., W.J. Idsardi, and V. van Wassenhove, *Speech perception at the interface of neurobiology and linguistics.* Philos Trans R Soc Lond B Biol Sci, 2008. **363**(1493): p. 1071-86.

6. Sussman, H.M., *Neural coding of relational invariance in speech: human language analogs to the barn owl.* Psychol Rev, 1989. **96**(4): p. 631-42.

7. Gao, X. and M. Wehr, *A coding transformation for temporally structured sounds within auditory cortical neurons.* Neuron, 2015. **86**(1): p. 292-303.

8. Scholl, B., X. Gao, and M. Wehr, *Nonoverlapping sets of synapses drive on responses and off responses in auditory cortex.* Neuron, 2010. **65**(3): p. 412-21.

9. Weible, A.P., C. Liu, C.M. Niell, and M. Wehr, *Auditory cortex is required for fear potentiation of gap detection.* J Neurosci, 2014. **34**(46): p. 15437-45.

10. Weible, A.P., A.K. Moore, C. Liu, L. DeBlander, H. Wu, C. Kentros, and M. Wehr, *Perceptual Gap Detection is Mediated by Gap Termination Responses in Auditory Cortex.* Current Biology, 2014.

11. Mesgarani, N., C. Cheung, K. Johnson, and E.F. Chang, *Phonetic feature encoding in human superior temporal gyrus.* Science, 2014. **343**(6174): p. 1006-10.

12. Engineer, C.T., C.A. Perez, Y.H. Chen, R.S. Carraway, A.C. Reed, J.A. Shetake, V. Jakkamsetti, K.Q. Chang, and M.P. Kilgard, *Cortical activity patterns predict speech discrimination ability.* Nat Neurosci, 2008. **11**(5): p. 603-8.

13. Blumstein, S.E. and K.N. Stevens, *Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants.* J Acoust Soc Am, 1979. **66**(4): p. 1001-17.

14. Kluender, K.R., R.L. Diehl, and P.R. Killeen, *Japanese quail can learn phonetic categories.* Science, 1987. **237**(4819): p. 1195-7.

15. Stilp, C.E. and K.R. Kluender, *Efficient coding and statistically optimal weighting of covariance among acoustic attributes in novel sounds.* PLoS One, 2012. **7**(1): p. e30845.

16. Margoliash, D. and E.S. Fortune, *Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc.* J Neurosci, 1992. **12**(11): p. 4309-26.

17. Lewicki, M.S. and B.J. Arthur, *Hierarchical organization of auditory temporal context sensitivity.* J Neurosci, 1996. **16**(21): p. 6987-98.

18. Liberman, A.M., F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy, *Perception of the speech code.* Psychol Rev, 1967. **74**(6): p. 431-61.

19. Centanni, T.M., A.M. Sloan, A.C. Reed, C.T. Engineer, R.L. Rennaker, 2nd, and M.P. Kilgard, *Detection and identification of speech sounds using cortical activity patterns.* Neuroscience, 2014. **258**: p. 292-306.

20. Ranasinghe, K.G., W.A. Vrana, C.J. Matney, and M.P. Kilgard, *Increasing diversity of neural responses to speech sounds across the central auditory pathway.* Neuroscience, 2013. **252**: p. 80-97.

21. Liberman, A.M., P. Delattre, and F.S. Cooper, *The role of selected stimulus-variables in the perception of the unvoiced stop consonants.* Am J Psychol, 1952. **65**(4): p. 497-516.

22. Apfelbaum, K.S. and B. McMurray, *Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization.* Psychon Bull Rev, 2015. **22**(4): p. 916-43.

23. Blumstein, S.E., *The mapping from acoustic structure to the phonetic categories of speech: The invariance problem.* Behavioral and Brain Sciences, 1998. **21**(2): p. 260.

24. Centanni, T.M., C.T. Engineer, and M.P. Kilgard, *Cortical speech-evoked response patterns in multiple auditory fields are correlated with behavioral discrimination ability.* J Neurophysiol, 2013. **110**(1): p. 177-89.

25. Engineer, C.T., C.A. Perez, R.S. Carraway, K.Q. Chang, J.L. Roland, and M.P. Kilgard, *Speech training alters tone frequency tuning in rat primary auditory cortex.* Behav Brain Res, 2014. **258**: p. 166-78.

26. Engineer, C.T., K.C. Rahebi, E.P. Buell, M.K. Fink, and M.P. Kilgard, *Speech training alters consonant and vowel responses in multiple auditory cortex fields.* Behav Brain Res, 2015. **287**: p. 256-64.

27. Porter, B.A., T.R. Rosenthal, K.G. Ranasinghe, and M.P. Kilgard, *Discrimination of brief speech sounds is impaired in rats with auditory cortex lesions.* Behav Brain Res, 2011. **219**(1): p. 68-74.

28. Lotto, A.J., K.R. Kluender, and L.L. Holt, *Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica).* J Acoust Soc Am, 1997. **102**(2 Pt 1): p. 1134-40.

29. Bizley, J.K. and Y.E. Cohen, *The what, where and how of auditory-object perception.* Nat Rev Neurosci, 2013. **14**(10): p. 693-707.

30. Suzuki, W.A., *The anatomy, physiology and functions of the perirhinal cortex.* Curr Opin Neurobiol, 1996. **6**(2): p. 179-86.

31. Wekselblatt, J.B., E.D. Flister, D.M. Piscopo, and C.M. Niell, *Large-scale imaging of cortical dynamics during sensory perception and behavior.* J Neurophysiol, 2016: p. jn 01056 2015.

32. Machens, C.K., M.S. Wehr, and A.M. Zador, *Linearity of cortical receptive fields measured with natural sounds.* J Neurosci, 2004. **24**(5): p. 1089-100.

33. Moore, A.K. and M. Wehr, *Parvalbumin-expressing inhibitory interneurons in auditory cortex are well-tuned for frequency.* J Neurosci, 2013. **33**(34): p. 13713-23.

34. Moore, A.K. and M. Wehr, *A guide to in vivo single-unit recording from optogenetically identified cortical inhibitory interneurons.* J Vis Exp, 2014(93): p. e51757.

35. Scholl, B. and M. Wehr, *Disruption of balanced cortical excitation and inhibition by acoustic trauma.* J Neurophysiol, 2008. **100**(2): p. 646-56.

36. Tan, A.Y. and M. Wehr, *Balanced tone-evoked synaptic excitation and inhibition in mouse auditory cortex.* Neuroscience, 2009. **163**(4): p. 1302-15.