ABSTRACT

Title of Document:	Representation of speech in the primary auditory cortex and its implications for robust speech processing.
	Nima Mesgarani, Ph.D., 2008
Directed By:	Professor Shihab Shamma, Electrical and Computer Engineering Department

Speech has evolved as a primary form of communication between humans. This most used means of communication has been the subject of intense study for years, but there is still a lot that we do not know about it. It is an oft repeated fact, that even the performance of the best speech processing algorithms still lags far behind that of the average human, It seems inescapable that unless we know more about the way the brain performs this task, our machines can not go much further. This thesis focuses on the question of speech representation in the brain, both from a physiological and technological perspective. We explore the representation of speech through the encoding of its smallest elements – phonemic features - in the primary auditory cortex. We report on how population of neurons with diverse tuning properties respond discriminately to phonemes resulting in explicit encoding of their parameters. Next, we show that this sparse encoding of the phonemic features is a simple consequence of the linear spectro-temporal properties of the auditory cortical neurons and that a Spectro-Temporal receptive field model can predict similar patterns of activation. This is an important step toward the realization of systems that operate based on the same principles as the cortex. Using an inverse method of reconstruction, we shall also explore the extent to which phonemic features are preserved in the cortical representation of noisy speech. The results suggest that the cortical responses are more robust to noise and that the important features of phonemes are preserved in the cortical representation even in noise. Finally, we explain how a model of this cortical representation can be used for speech processing and enhancement applications to improve their robustness and performance.

Representation of speech in the primary auditory cortex and its implications for robust speech processing

By

Nima Mesgarani

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor Of Philosophy
2008

Advisory Committee:
Professor Shihab Shamma, Chair
Professor Carol Espy-Wilson
Professor David Poeppel
Associate Professor Jonathan Simon
Research Scientist Jonathan Fritz

Acknowledgements

I would like to thank my Ph.D. advisor, Professor Shihab Shamma, for his scientific advice and knowledge and many insightful discussions. He has always been very supportive and has given me the freedom to pursue any random thoughts I have had and always valued my curiosity and creativity. I can not express how grateful I am for his understanding of my difficult times during these past years. I also would like to thank my Ph.D. committee, Doctors Carol Espy-Wilson, Jonathan Simon, Jonathan Fritz and David Poeppel. Special thanks go to Dr. Jonathan Fritz, for teaching me so much about the world of neuroscience and animal behavior and for all the constructive criticisms of my work.

I would like to thank the present and past members of Neural Systems Laboratory, for their friendship and support. To Kevin, Ling, Serin, Mounya, Sridhar, Tai-Chih, Mai, Majid, Dan, Pingbo, and everybody else: thank you for helping me in so many ways to get through the hardship of graduate school. I am especially thankful to Dr. Stephen David for all the insightful discussions and fun collaborations we have had, and also for being such a good friend.

I also thank my friends (too many to list here but you know who you are) for providing support and friendship that I needed. With you, my learning journey expanded way beyond the academics of the graduate school.

Finally, I am forever indebted to my mom and Shabnam for their understanding, endless patience, acceptance and encouragement when it was most required. This would not have been possible without your help.

College Park, August 2008
Nima

Table of Contents

List of Figures

Chapter 1:

Introduction

Speech has evolved as a primary form of communication between humans. This most used means of communication has been the subject of intense study for more than 150 years, but there is still a lot that we do not know about it. It is for instance a mystery how a thought expressed in the form of an acoustic wave organized as a string of sound segments, is perceived as the same thought in a listener. Unless this process is well understood, it will be difficult to imagine man-made machines that can communicate with humans at a comparable level of performance. It is an oft repeated fact, that even the performance of the best speech processing algorithms still lags far behind that of the average human [1]. It seems inescapable that unless we know more about the way the brain performs this task, our machines can not go much further [2].

Understanding speech processing in the brain can also benefit those with hearing impairments. For example, peripheral hearing impairment results in a distorted cortical representation of speech that reduces the reliability and efficacy of its reception. Knowing the normal representation, one can ask how we can preprocess the speech to correct the representation.

This thesis focuses on the question of speech representation in the brain, both from a physiological and technological perspective. We start by presenting an overview of basic concepts used throughout the thesis. The first is an overview of the auditory pathway, starting from the external ear and ending in the primary auditory cortex where our neural recordings are performed. The second introductory topic is on the characteristics of speech and its phonemic categorization based on their articulatory features. In the second chapter, we explore the representation of speech through the encoding of its smallest elements – phonetic features - in the primary auditory cortex. We report on how a population of neurons with diverse tuning properties respond discriminately to phonemes resulting in an explicit encoding of their parameters. Next, we show that this sparse encoding of the phonetic features is a simple consequence of the linear spectro-temporal properties of the auditory cortical neurons and that a Spectro-Temporal receptive field model can predict similar patterns of activation. This is an important step toward realization of systems that operate based on the same principles as the cortex. In chapter 3, we look at the encoding from a different angle where we discuss an inverse model that maps the population of neural responses to the sound spectrogram. This inverse model allows us to investigate more readily the features of speech encoded by the neural population. Using this method, we shall also explore the extent to which phonemic features are preserved in the cortical representation of noisy speech. The results suggest that the cortical responses are more robust to noise and that the important features of phonemes are preserved in the cortical representation even in noise. In the last chapter, we explain how a model of this cortical representation can be used for speech processing and enhancement applications. We describe specifically two systems that use the cortical model for 1) speech discrimination and 2) noise suppression. The goal of the first task is to discriminate between speech and any other sound such as music, animal vocalizations, and various types of noise. This algorithm has outperformed all other state-of-the-art systems, and has even been adapted for application in robotics (Honda

Research), tracking (Advance Acoustic Concepts, Inc) and security (US government). In the second task, we describe a noise suppression algorithm that uses the diversity of tunings seen in the cortex to separate the noise from speech in a way that is not possible at the level of the spectrogram. By suppressing the noise in this cortical representation and reconstructing the sound, we can enhance the speech part. Finally, we shall discuss future efforts to further understand the analysis of speech in the primary auditory cortex. For example, our current linear models of cortical processing are incapable of explaining the robustness of speech representation in severe noisy environments, and hence it is essential to develop nonlinear models to handle these situations. It is also critical that training (and its role in enriching and stabilizing speech representations) be addressed and incorporated in future systems.

Chapter 2

2.1 Review of auditory pathway

The auditory system in humans and other vertebrates is responsible for hearing, enables them to perceive sound by detecting vibration in the air. These vibrations are detected by the ear and transduced into nerve impulses that are perceived by the brain. In this section, we review part of auditory system in mammals that starts from the external ear and ends in the primary auditory cortex.

2.1.1 The ear

The ear has three functional parts: external ear, middle ear and internal ear. To hear, our ears must capture the mechanical energy (sound), transmit it to the ear's receptive organ, and transduce it into electrical signals suitable for analysis by the nervous system. These three tasks are the functions of the external, middle and the inner ear (Figure 1). The external ear, especially the prominent auricle, focuses sound into the external auditory meatus. Alternating increases and decreases in air pressure vibrate the tympanum. These vibrations are conveyed across the air-filled middle ear by three tiny, lined bones: the malleus, the incus, and the stapes. Vibration of the stapes stimulates the cochlea, the hearing organ of the inner ear.

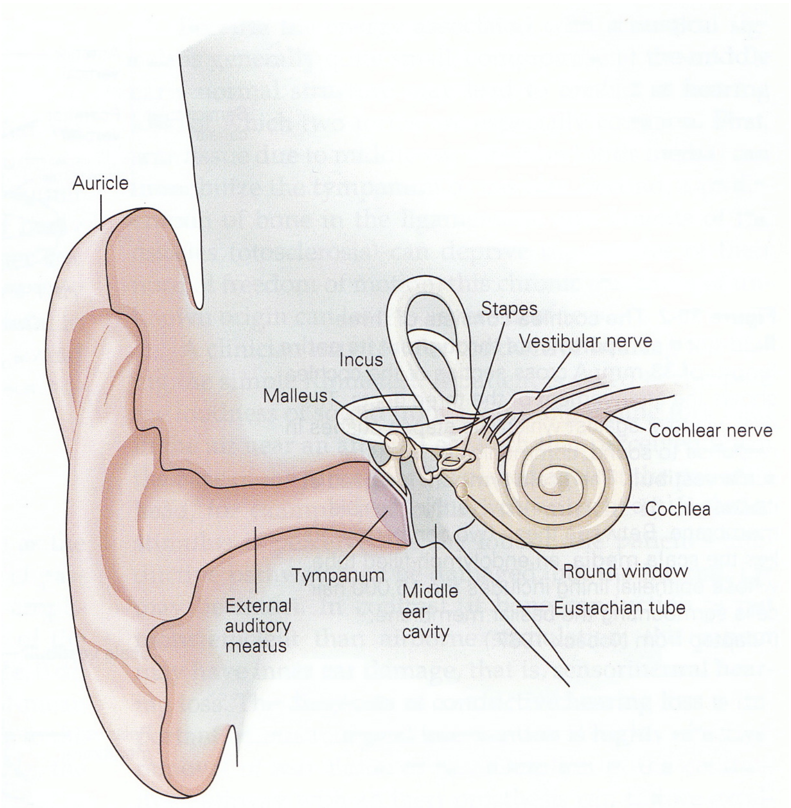

Figure 1. The structure of the human ear (external, middle and inner ear) (Adapted from Noback 1967).

The cochlea (Figure 2) in the inner ear consists of three fluid-filled compartments throughout its entire length of 33 mm. A cross section of the cochlea shows the arrangement of the three ducts. The oval window, against which the stapes pushes in response to sound, communicates with the scala vestibuli. The scala tympani is closed at its base by the round window, a tick, flexible membrane. Between these two compartments lies the scala media, an endolymph-filled tube whose epithelial lining includes the 16,000 hair cells surrounding the basilar membrane.

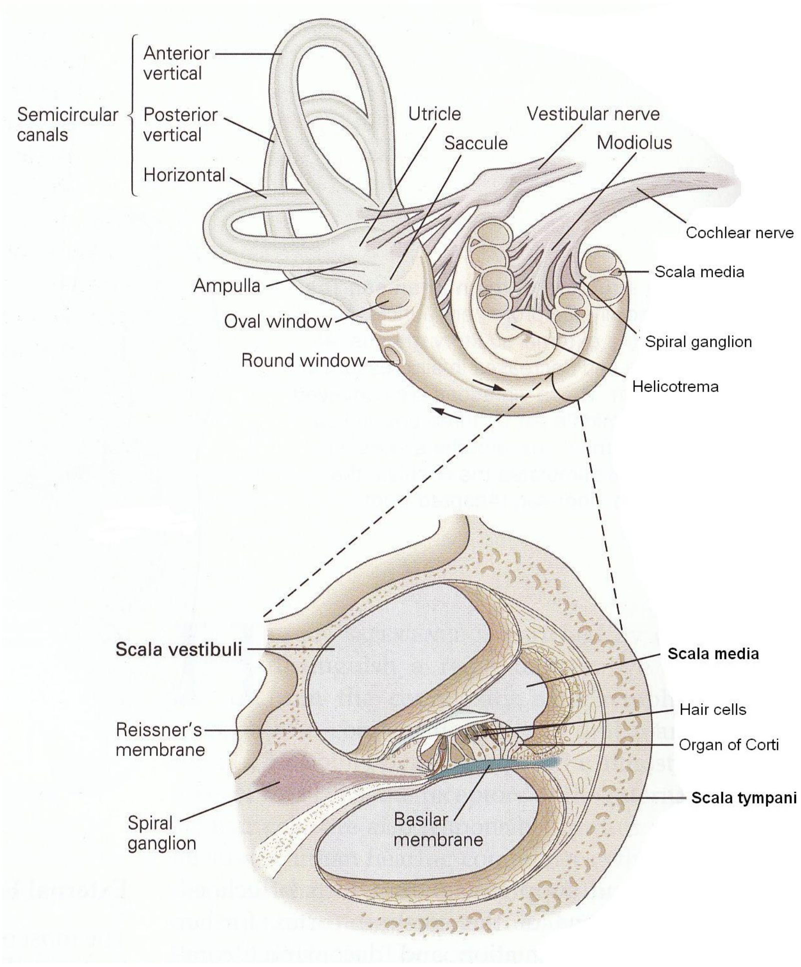

Figure 2. The structure of the cochlea

2.1.2 Functional anatomy of the Cochlea

The basilar membrane (**Figure 3**) is a mechanical analyzer of sound frequency. The mechanical properties of the basilar membrane are key to the cochlea's operation. In brief, the membrane is tapered and it is stiffer at one end than at the other. The dispersion of fluid waves causes sound input of a certain frequency to vibrate some locations of the membrane more than the other locations. As shown in experiments by Nobel Prize laureate George von Bekesy, high frequencies lead to maximum vibrations at the basal end of the cochlear coil (narrow, stiff membrane), and low frequencies lead to maximum vibrations at the apical end of the cochlear coil (wide, more compliant membrane).

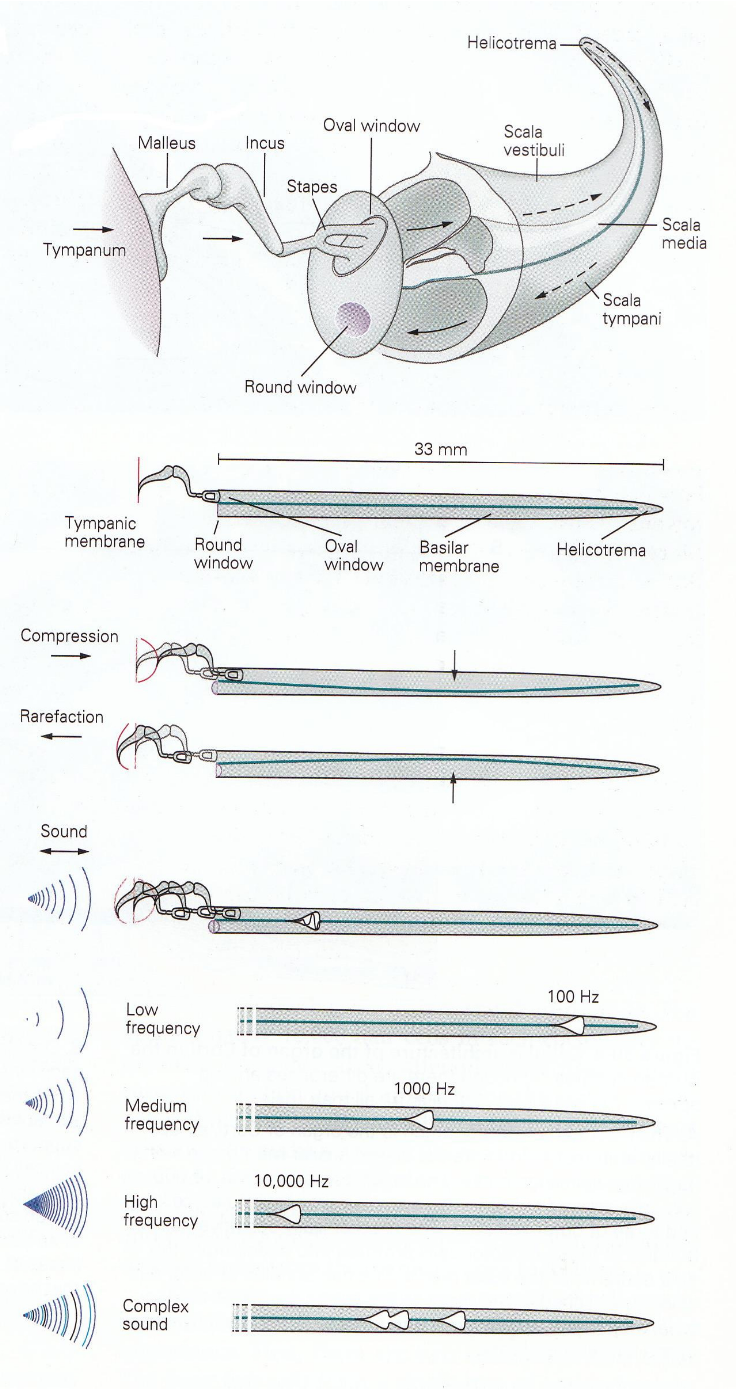

Figure 3. Motion of the basilar membrane.

2.1.3 Cellular architecture of the organ of Corti

The organ of Corti (Figure 4) is the organ in the inner ear of mammals that contains auditory sensory cells, or hair cells. The organ contains some 16,000 hair cells arrayed in four rows: a single row of inner hair cells and three of outer hair cells. The mechanically sensitive hair bundles of these receptor cells protrude into endolymph, the fluid contents of the scala media. The hair bundles of outer hair cells are attached at their tops to the lower surface of the tectorial membrane, a gelatinous shelf that extends the full length of the basilar membrane. The basic architecture of the organ of Corti is similar for all mammals.

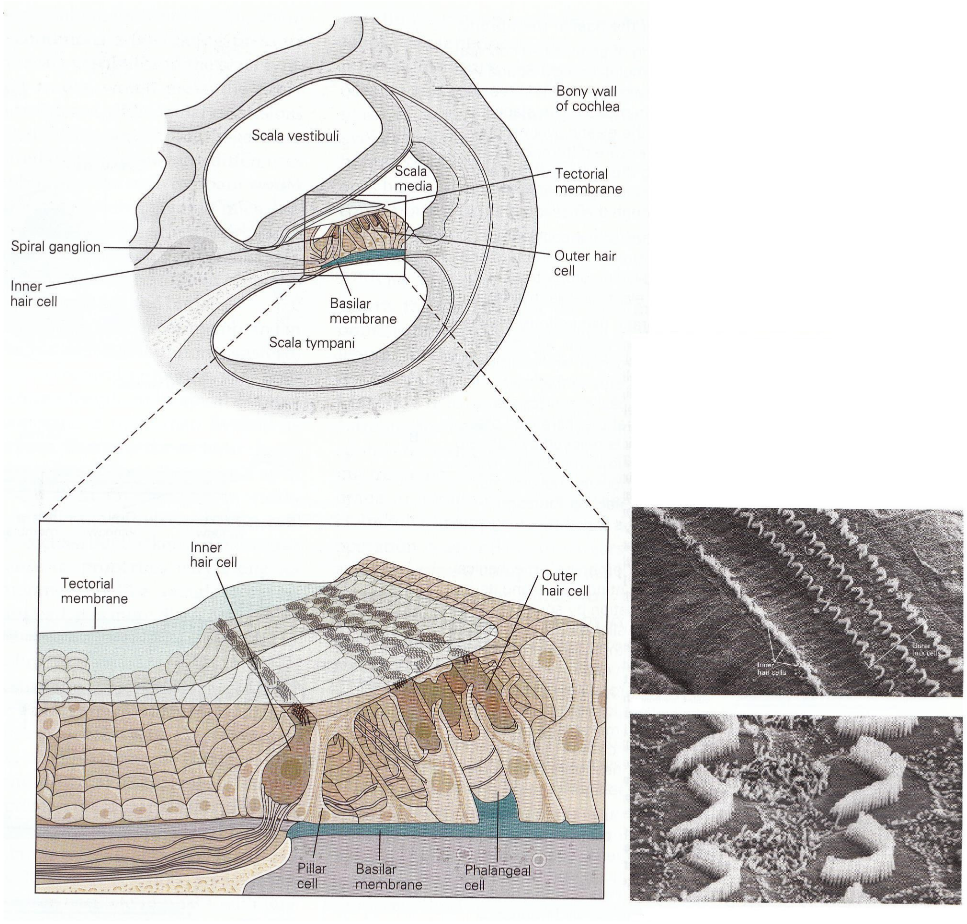

Figure 4. Cellar structure of the organ of Corti

Hair cells in the cochlea are stimulated when the basilar membrane is driven up and down by differences in the fluid pressure between the scala vestibuli and scala tympani. Because this motion is accompanied by shearing motion between the tectorial membrane and organ of Corti, the hair bundles that link the two are deflected. This deflection initiates mechanoelectrical transduction of the stimulus. When the basilar membrane is driven upward, shear between the hair cells and the tectorial membrane deflects hair bundles in the excitatory direction, toward their tall edge. At the midpoint of an oscillation the hair bundles resume their resting position. When the basilar membrane moves downward, the hair bundles are driven in the inhibitory direction (Figure 5).

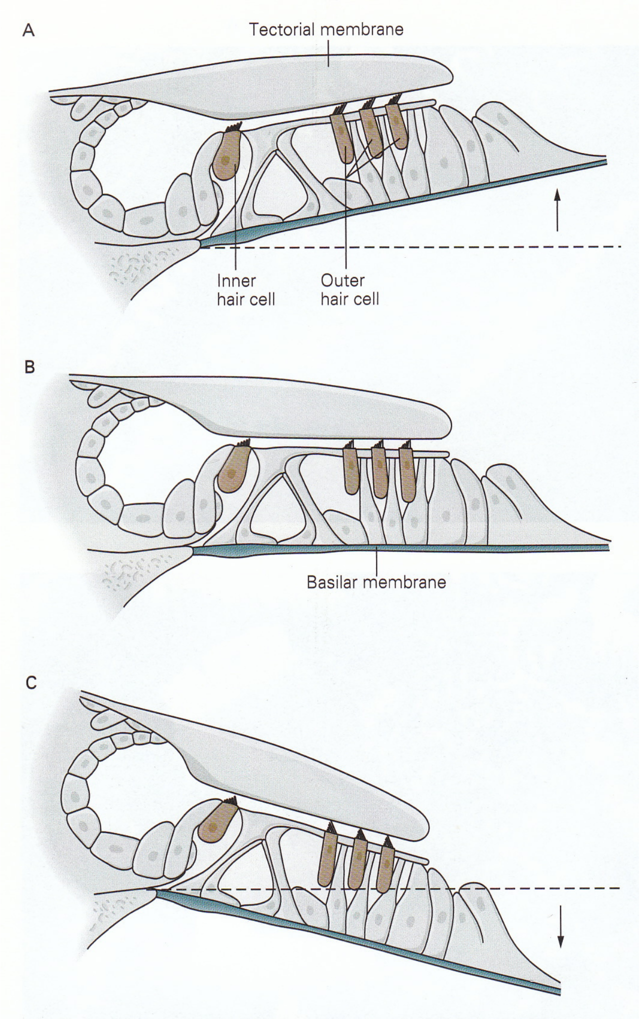

Figure 5. Stimulation of hair cells when the basilar membrane is driven up and down.

In mammalian outer hair cells (Figure 6.), the receptor potential triggers active vibrations of the cell body. Outer hair cells evolved only in mammals. They have not improved hearing sensitivity, but they have extended the hearing range and frequency selectivity which is of particular benefit for humans, because it enables sophisticated speech and music.

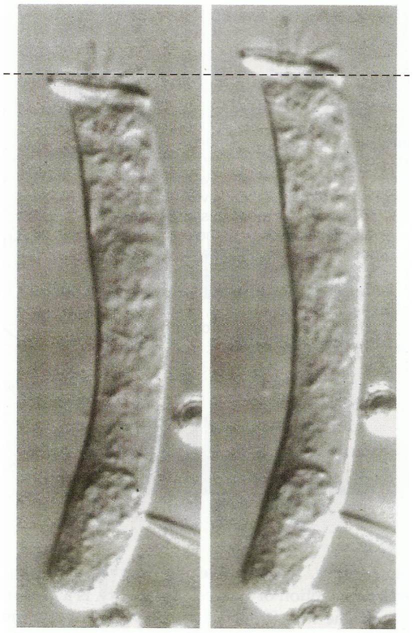

Figure 6. Outer hair cells

2.1.4 Structure of the inner hair cells

The structure of the hair cells is shown in Figure 7. The cylindrical hair cell is joined to the adjacent supporting cells by a junctional complex around its apical perimeter. From the cells apical surface extends the hair bundle, the mechanically sensitive organelle. Afferent and efferent synapses occur upon the basolateral surface of the plasma membrane. The bundle comprises some 60 stereocilia, each a cylinder with a tapered base, arranged in stepped rows of varying length. Deflection of the hair bundle to the right, the positive stimulus direction, depolarizes the hair cell; movement in the opposite direction elicits a hyperpolarization.

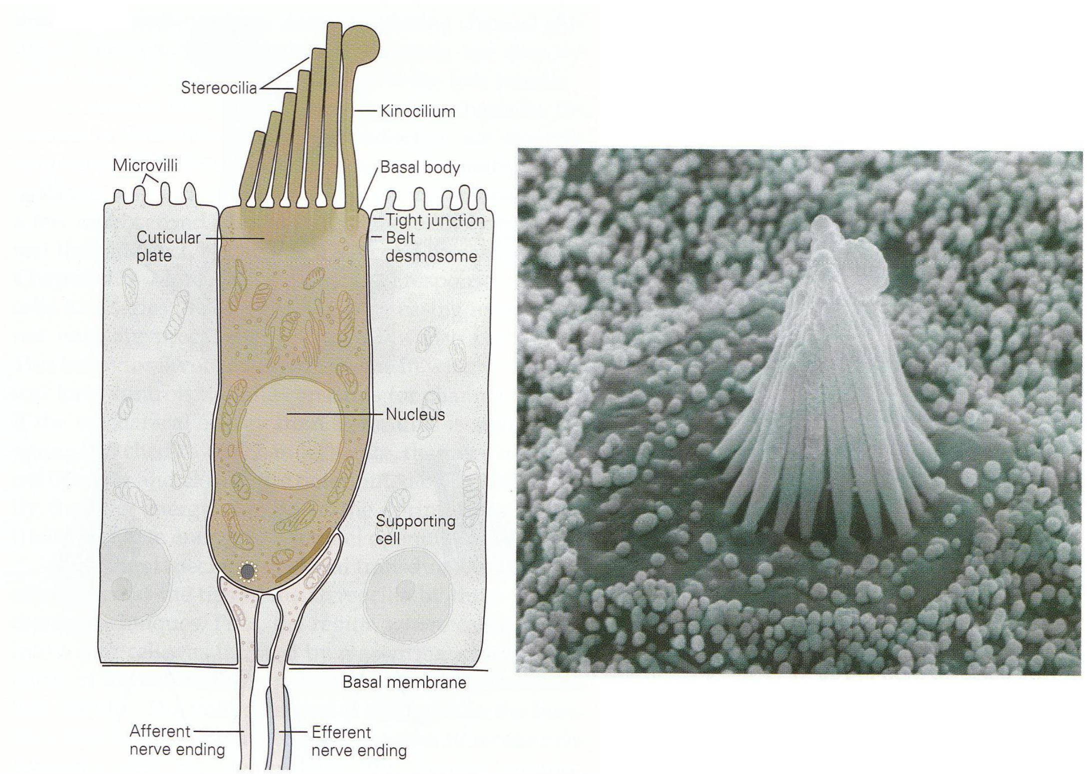

Figure 7. Structure of a vertebrate hair cell.

2.1.5 Hair cells transform mechanical energy into neural signals

Deflection of the hair bundle initiates mechanoelectrical transduction. This involves a mechanism for gating of ion channels that is fundamentally different from those employed in such electrical signals as the action potential or postsynaptic potential. The opening and closing of transduction channels is regulated by the tension in the elastic structure within the hair bundle (Figure 8). The ion channels that participate in mechanoelectrical transduction in hair cells are gated by elastic structures in the hair bundle. The channel is assumed to be a membrane-spanning protein with a cation-selective pore. When the hair bundle is at rest, each transduction channel clatters between closed and open states, spending most of its time shut. Displacement of the bundle in the positive direction increases the tension in the gating spring, here assume to be a tip link attached to each channel's molecular gate. The enhanced tension promotes channel opening and the influx of cations, thereby producing a depolarizing receptor potential.

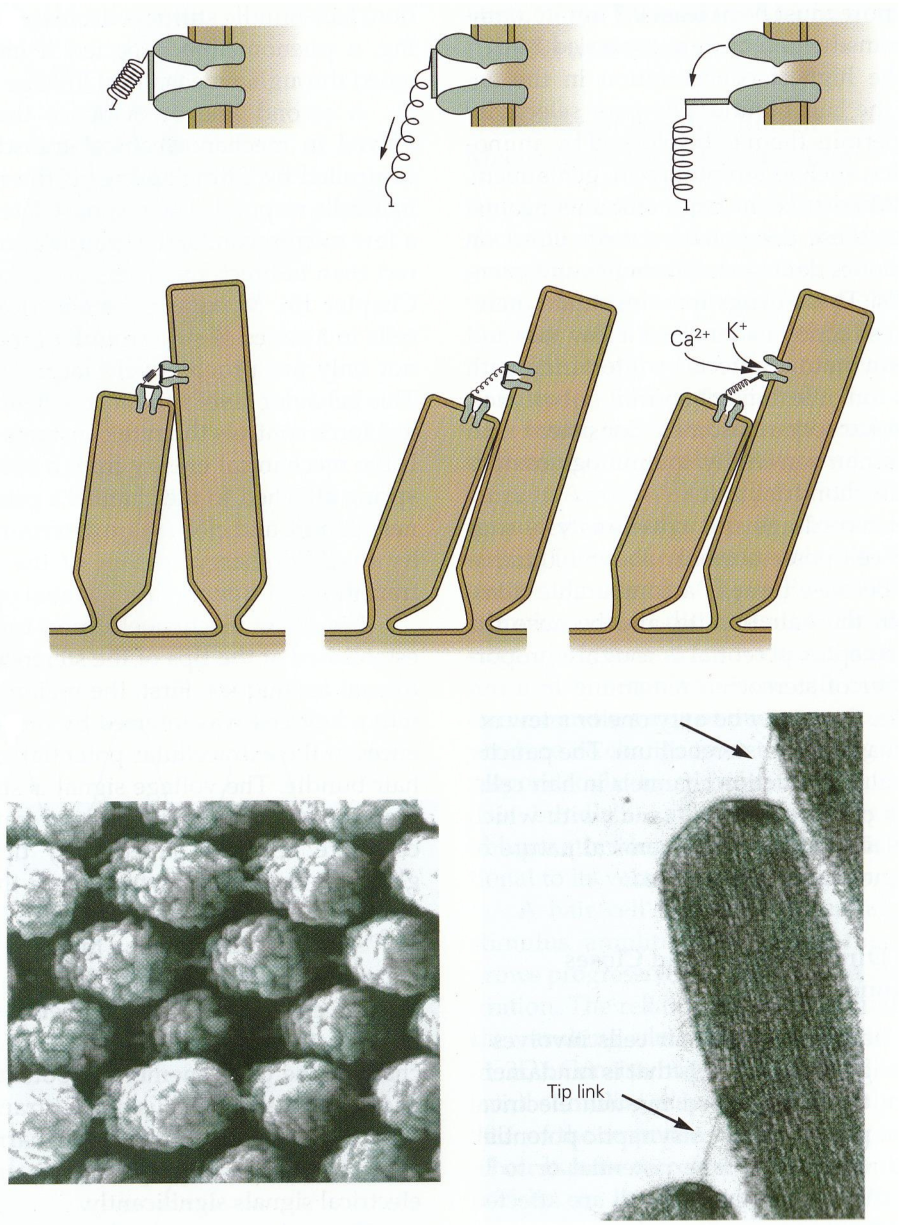

Figure 8. A model for the mechanism of the mechano-electrical transduction by hair cells.

2.1.6 Innervation of the organ of Corti.

The great majority of afferent axons end on inner hair cells, each of which constitutes the sole terminus for an average of 10 axons. A few afferent axons of small caliber provide diffuse innervation to the outer hair cells. Efferent axons largely innervate outer hair cells, and do so directly. In contrast, efferent innervation of inner hair cells is sparse and is predominantly axoaxonic, at the ending of afferent nerve fibers (Figure 9).

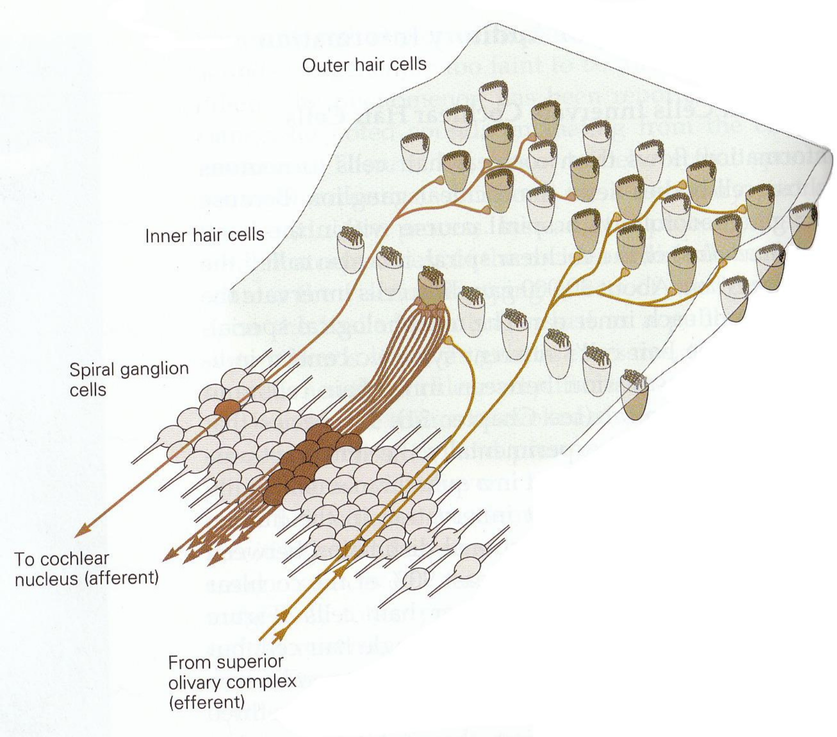
Outer hair cells
Inner hair cells
Spiral ganglion cells
To cochlear nucleus (afferent)
From superior olivary complex (efferent)

Figure 9. Innervation of the organ of Corti.

2.1.7 The central auditory pathway

The central auditory pathways extend from the cochlear nucleus to the auditory cortex. Postsynaptic neurons in the cochlear nucleus send their axons to other centers in the brain via three main pathways: the dorsal acoustic stria, the intermediate acoustic stria, and the trapezoid body. The first binaural interactions occur in the superior olivary nucleus, which receives input via the trapezoid body. In particular, the medial and lateral divisions of the superior olivary nucleus, along with axons from the cochlear nuclei, project to the inferior colliculus in the midbrain via the lateral lemniscus. Each lateral lemniscus contains axons relaying input from both ears. Cells in the colliculus send their axons to the medial geniculate nucleus of the thalamus. The geniculate axons terminate in the primary auditory cortex, a part of the superior temporal gyrus (Figure 10). Information flows from cochlear hair cell to neurons whose cell bodies lie in the cochlear ganglion. The pattern of afferent innervations in the human cochlea emphasizes the functional distinction between inner and outer hair cells. At least 90% of the cochlear ganglion cells terminate on inner hair cells. Each axon innervates only a single hair cell, but each inner hair cell directs its output to several nerve fibers, on average nearly 10. The output of each inner hair cell is sampled by many nerve fibers, which independently encode information about the frequency and intensity of sound. The tonotopic organization of the auditory neural pathways begins at the earliest possible site, immediately postsynaptic to inner hair cells.

The acoustical sensitivity of axons in the cochlear nerve mirrors the innervation pattern of spiral ganglion cells. Each axon is most responsive to stimulation at a particular frequency of sound, its characteristic frequency. Stimuli of lower or higher frequency also evoke responses, but only when presented at greater intensities. The relation between sound-pressure level and firing rate in each fiber of the cochlear nerve is approximately linear. Difference in neuronal responsiveness originate at the synapses between inner hair cells and afferent nerve fibers. Nerve terminals on the surface of a hair cell nearest the axis of the cochlear spiral belong to the afferent neurons of lowest sensitivity and spontaneous activity. The multiple innervations of each inner hair cell are therefore not completely redundant. Instead, because of systematic differences in the rate of transmitter release or in postsynaptic responsiveness (or both), the output from a given hair cell is directed into several parallel channels of differing sensitivity and dynamic range.

Three important general principles emerge from connections in the brain stem. First, acoustical information is processed in parallel pathways, each of which is dedicated to the analysis of a particular feature of auditory information. Second, the various cell types of the cochlear nuclei project to specific relay nuclei, so that the separation of information streams commence within the cochlear nuclei. Finally, there is extensive interaction between auditory structures on the two sides of the brain stem. The medial superior olive performs a specific function in a readily intelligible way. The ability to localize sound sources along the azimuthal axis stems in part from the processing of information about auditory delays.

The inferior colliculus (IC) is divisible into two major components. Because it contains many neurons sensitive to interaural timing or intensity differences, the IC is apparently involved in sound localization. The medial geniculate body (MGN) constitutes the thalamic relay of the auditory system. This nuclear complex comprises at least three subdivisions of which the principal nucleus is the best understood. Most neurons in MGN are sharply tuned to specific stimulus frequencies, and most are responsive to stimulation through either ear.

The ascending auditory pathway terminates in the cerebral cortex, where several distinct auditory areas occur on the dorsal surface of the temporal lobe. The most prominent projection from the ventral nucleus of the MGN extends to the primary auditory cortex (A1).

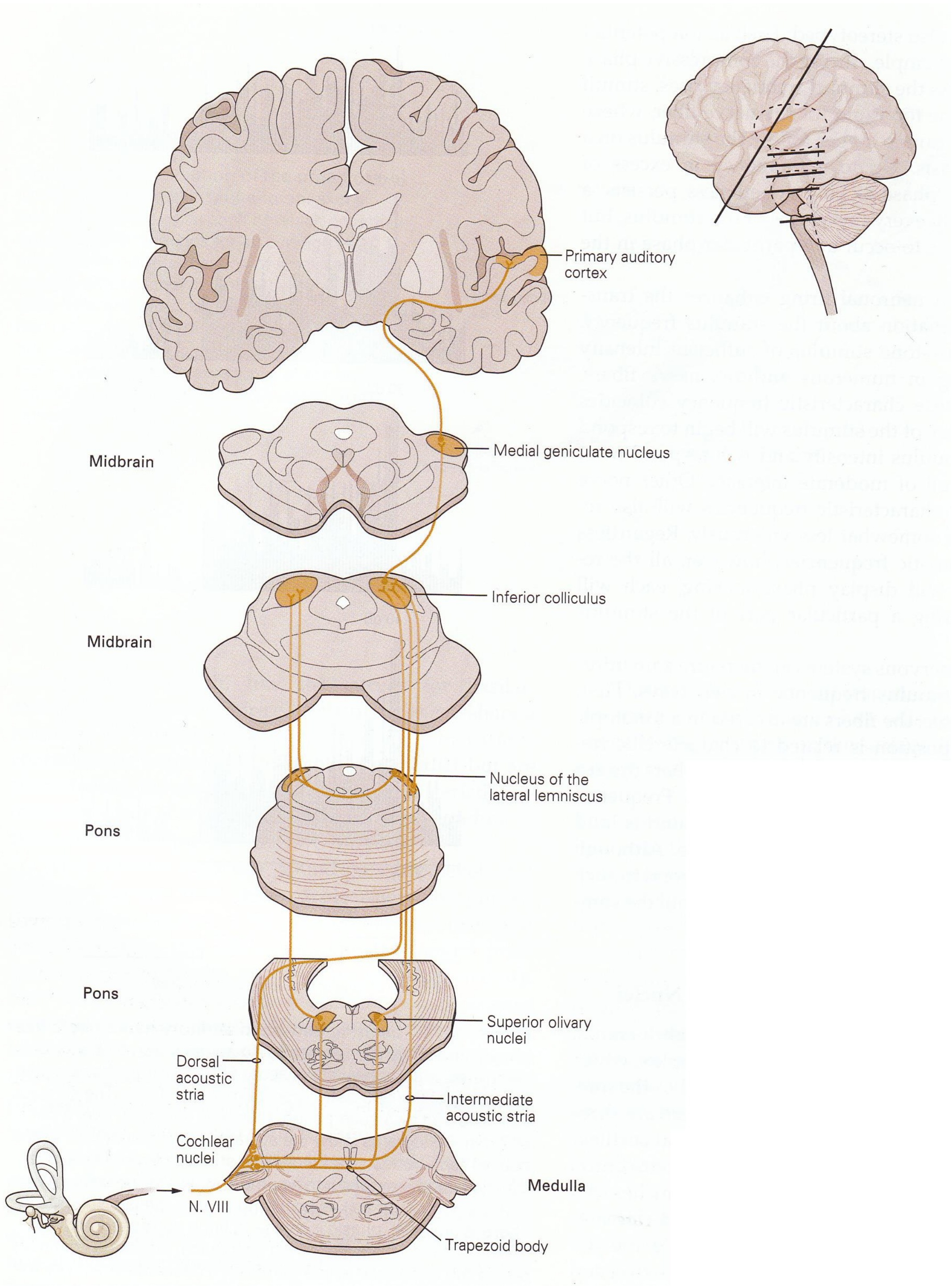

Figure 10. The central auditory pathway: Cochlear nuclei, Superior olivary nuclei, lateral lemniscus, inferior coliculus, medial geniculate nuclei, and finally primary auditory cortex

2.2 Phonemes as elements of language

Phonemes are the fundamental distinctive units of a language. The phoneme is a speech sound class that differentiate words of a language. To emphasize the distinction between the concept of a phoneme and sounds that convey a phoneme, the speech scientists uses the term phone to mean a particular instantiation of a phoneme.

Different languages contain different phoneme sets. Syllables contain one or more phonemes, while words are formed with one or more syllables. Words are concatenated to form phrases and sentences. One broad phoneme classification for English is in terms of vowels, consonants, diphthongs, fricatives and semi-vowels [3] [4].

The phoneme arises from a combination of vocal fold and vocal tract articulatory features. Articulatory features, corresponding to the first two descriptor above, include the vocal fold state (whether the vocal folds are vibrating or open); the tongue position and height (whether it is in the front, central, or back along the palate) and whether its constriction is partial or complete; and the velum state (whether a sound is nasal or not). A particular set of speech muscles is responsible for "activating" each feature with certain relative timing. In English, the combinations of features are such to give 40 phonemes.

.

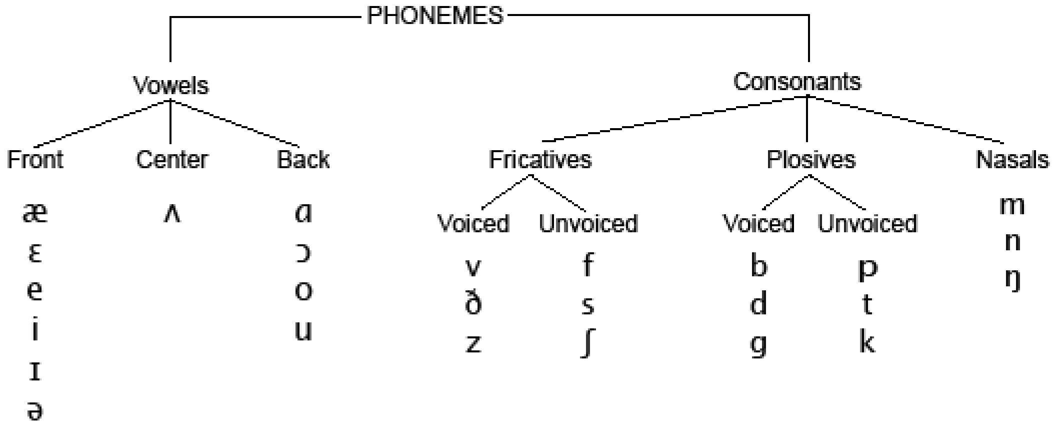

Figure 11. Categorization of English phonemes

2.2.1 Vowels

The largest phoneme group is that of vowels. Vowels contain three subgroups: front, central and back which are defined by the tongue hump position [3]. Each vowel phoneme corresponds to a different vocal tract configuration. The vocal tract shape is a function of the tongue, the jaw, the lips and the velum which is closed in non-nasalized vowels. The degree of constriction by the tongue is another shape determinant, which can be open (like in the vowel /a/) or closed (like in the vowel /i/). The particular shape of the vocal tract determines its resonance structure. The shape of the vocal tract in the production of different vowels is shown in Figure 12.

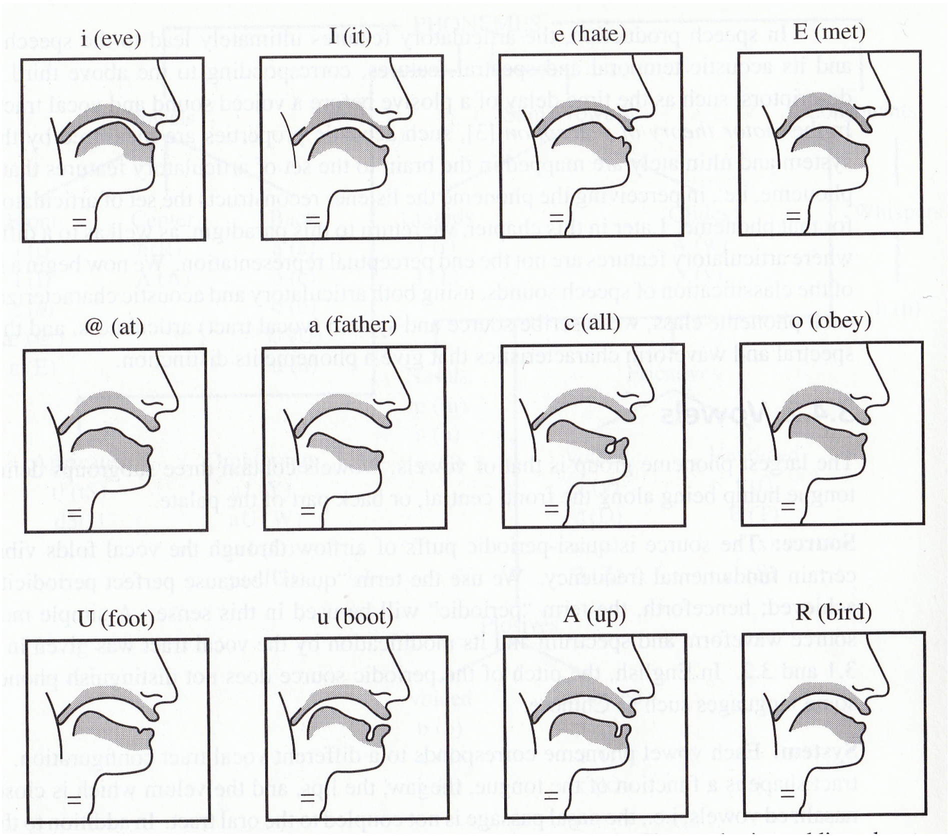

Figure 12. Production of vowels

2.2.2 Fricatives

Fricative consonants can be specified in two classes: voiced and unvoiced. In unvoiced fricatives, the vocal folds are relaxed and not vibrating. Noise is generated by turbulent airflow at the point of constriction along the oral tract. The location of the constriction by the tongue at the back, center, or front of the oral tract, as well as at the teeth or lips, influences which fricative sound is produced. The constriction separates the oral tract into front and back cavities with the sound radiated from the front cavity. Voiced fricatives have a similar noise source and system characteristics, the difference is that for voiced fricatives the vocal fold usually vibrate simultaneously with noise generation at the constriction.

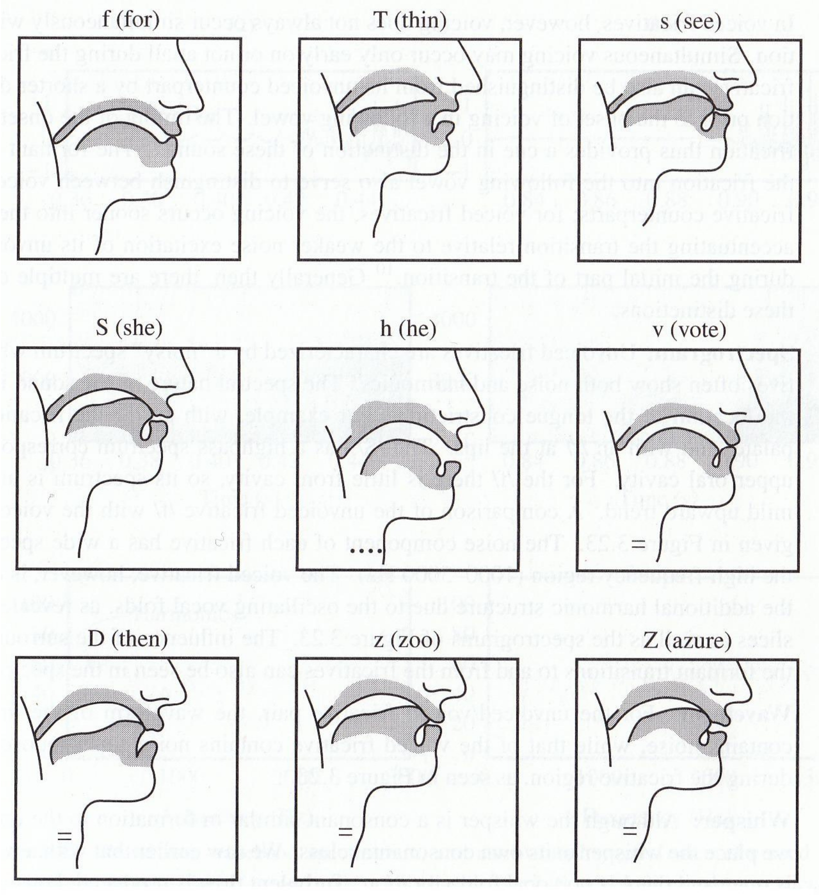

Figure 13. Production of fricatives

The spectral nature of the sound is determined by the location of the tongue constriction. For example, with an /S/, the frication occurs at the palate, and with an /f/ at the lips. The /S/ has a highpass spectrum corresponding to a short upper oral cavity. The location of the constriction in the vocal tract for different fricatives is shown in Figure 13.

2.2.3 Plosives

As with fricatives, plosives are both unvoiced and voiced. With unvoiced plosives, a “burst” is generated at the release of the buildup of pressure behind a total constriction in the oral tract. The constriction can happen at the front, center or back of the palate. The sequence of production of a plosive starts with a complete

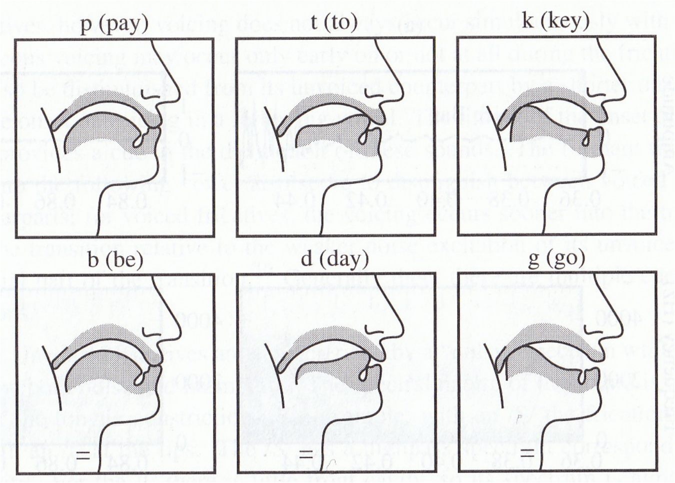

Figure 14. Production of plosives

closure of the oral tract and buildup of air pressure behind closure; followed by releases of air pressure and generation of turbulence over a very short-time duration (burst). With the voiced fricatives, there is a buildup of pressure behind an oral tract constriction, but the vocal folds can also vibrate. When this vibration occurs, although the oral tract is closed, we hear a low-frequency vibration due to its propagation through the walls of the throat. The voiced onset time is the difference between the time of the burst and the onset of voicing in the following vowel. The length of the voice onset time and the place of constriction vary with the plosive consonants.

2.2.4 Nasals

Nasals are the closest to the vowels [3]. In their production, the velum is lowered and the air flows mainly through the nostrils. The nasal consonants are distinguished by the place along the oral tract at which the tongue makes a constriction. The spectrogram of a nasal is dominated by the low resonance of the large volume of the nasal cavity. The closed oral cavity acts as a side branch with its own resonance that changes with the place of constriction of the tongue. Theses resonances absorb acoustic energy and thus are anti-resonances (zeros) of the vocal tract. Figure 15 shows the shape of the oral tract for different nasal phonemes.

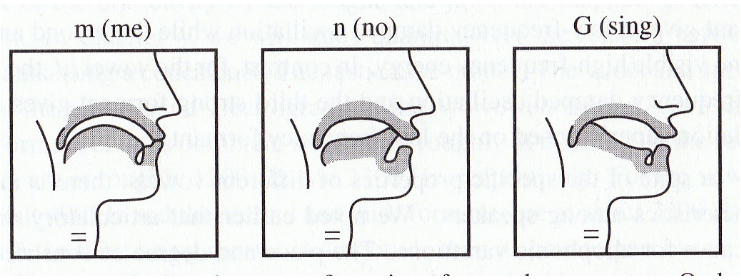

Figure 15. Production of nasals

Chapter 3

Phoneme presentation and classification in primary auditory cortex

Humans reliably identify many phonemes and discriminate them categorically, despite considerable natural variability across speakers and distortions in noisy and reverberant environments that limit the performance of even the best speech recognition algorithms [1] [2]. Trained animals have also been shown to discriminate phoneme pairs categorically and to generalize to novel situations [5] [6] [7] [8] [9] [10] [11] [12]. The neurophysiological basis of these perceptual abilities in humans and animals remains uncertain. However, there is experimental evidence for cortical encoding of phonetic acoustic features regarded as critical for distinguishing classes of consonant-vowel (CV) syllables, such as voice-onset-time [13] [14] [15] [16]. Key questions include the nature and location of the neural representations of different phonemes and, more specifically, whether the neural responses of the primary auditory cortex (A1) are sufficiently rich to support the phonetic discriminations observed in humans and animals.

The general issue of the neural representation of complex patterns is common to all neuroscience and has been investigated in many sensory modalities. In the visual system, recent studies have shown that responses of approximately 100 cells in the inferior temporal cortex are sufficient to account for the robust identification and categorization of several object categories [17]. In the auditory system, a recent study has shown that neurometric functions derived from single unit recordings in the ferret primary auditory cortex closely parallel human psychometric functions for complex sound discrimination [18]. An important aspect of our approach in the present study is the inclusion of temporal features of the response in the analysis. This is crucial because phonemes are *spectro-temporal* patterns, and hence analyzing their neural representation at a single cell or ensemble level requires consideration of the interactions between the stimuli and the intrinsic dynamics of individual neurons.

In the present study, we recorded responses of A1 neurons to a large number of American English phonemes in a variety of phonemic contexts and derived from many speakers. Our results demonstrate that (I) time-varying responses from a relatively small population of primary auditory cortical neurons (< 100) can account for distinctive aspects of phoneme identification observed in humans [19], and that (II) well known acoustic features of phonemes are indeed explicitly encoded in the population responses in A1 [20] [21].

The analysis of the categorical representation of phonemes across a neuronal population presented in this paper remains largely model-independent in that only relatively raw response measures (e.g., peri-stimulus time histograms, PSTHs) are used in the computations and illustrations. The one key departure from this rule is necessitated by the desire to organize the display of the population responses according to their best frequency, spectral scale, and temporal dynamics. These response properties are quantified using the measured spectro-temporal receptive field (STRF) model of the neurons [22] [23].

3.1 Experimental Procedures

The protocol for all surgical and experimental procedures was approved by the IACUC at the University of Maryland and consistent with NIH Guidelines

3.1.1 Surgery

Four young adult, female ferrets were used in the neurophysiological recordings reported here. To secure stability of the recordings, a stainless steel head post was surgically implanted on the skull. During implant surgery, the ferrets were anesthetized with Nembutal (40 mg/kg) and Halothane (1-2%). Using sterile procedures, the skull was exposed and a headpost was mounted using bone cement, leaving clear access to primary auditory cortex in both hemispheres. Antibiotics and analgesics were administered as needed.

3.1.2 Neurophysiological recording

Experiments were conducted with awake head-restrained ferrets. The animals were habituated to this setup over a period of several weeks, and usually remained relaxed and relatively motionless throughout recording sessions that may last 2-4 hrs. Recordings were conducted in a double-walled acoustic chamber. Small craniotomies (~1-2 mm in diameter) were made over primary auditory cortex before recording sessions. Physiological recordings were made using tungsten microelectrodes (4-8 MΩ, FHC). Electrical signals were amplified and stored using an integrated data acquisition system (Alpha Omega). Spike sorting of the raw neural traces was done off-line using a custom PCA clustering algorithm. Our requirements for single unit isolation of stable waveforms included (1) that the waveform and spike rate remained stable throughout the recording, and (2) that the inter-spike interval for each neuron was distributed exponentially with a minimum latency of 1 ms.

3.1.3 Speech Stimuli and data analysis

Stimuli were phonetically transcribed continuous speech from the TIMIT database [24]. Thirty different sentences (3 seconds, 16 KHz sampling) spoken by different speakers (15 male and 15 female) were used to sample a variety of speakers and contexts. Each sentence was presented five times during recordings. For a subset of neurons, 90 sentences spoken by 45 male and 45 female speakers were used.

3.1.4 Mean phoneme representation

The TIMIT phonetic transcriptions were used to align the responses of each neuron to all the instances of a given phoneme and then averaged to compute the peri-stimulus time histogram (PSTH) response to that phoneme, as illustrated in **Figure 16** (10 ms time bins). We did not attempt to compensate for the relatively short latency of neural responses in the ferret (15-20 ms). We also computed the auditory spectrogram of each phoneme using the following procedure: Let *S*(*t*,*f*) be the auditory spectrogram of the speech stimulus computed using a model of cochlear frequency analysis [25], and let *r*(*t*) be the corresponding neural response. For phoneme *k*, which occurs at times $t_{k_1}, t_{k_2}, \dots, t_{k_n}$, the average spectrogram is

$$\hat{S}_k\left(t, f\right) = \frac{1}{n}\sum_{i=1}^{n} S\left(t_{k_i} + t, f\right)$$

and the average neural response is

$$\hat{r}_k\left(t\right) = \frac{1}{n}\sum_{i=1}^{n} r\left(t_{k_i} + t\right). \qquad (1)$$

The total number of occurrences of each phoneme, *n*, ranged from 7 (e.g. /g/) to 72 (e.g., /ɨ/) in the chosen sentences.

3.1.5 Measurement of neuronal tuning properties

We characterized each neuron by its spectro-temporal receptive field (STRF), estimated by normalized reverse correlation of the neuron’s response to the auditory spectrogram of the speech stimulus [22]. Although methods such as normalized reverse correlation can produce unbiased STRF estimates in theory, practical implementation require some form of regularization to prevent overfitting to noise along the low-variance dimensions. This in effect imposes a smoothness constraint on the STRF. The regression parameters were adjusted using a jackknife validation set to maximize the correlation between actual and predicted responses [26]. **Figure 16B** illustrates the STRF of one such neuron. We measured several tuning properties from each STRF: Best frequency (BF) was defined as the largest positive peak value of the STRF along its frequency dimension. The STRF scale and rate were estimated from the 2-D modulation transfer function (MTF) (**Figure 16B**). The MTF is the 2-D Fourier transform of the STRF that is then collapsed along its temporal or spectral dimensions (known also as the *rate* and *scale)* to obtain the purely *spectral (sMTF)* or *temporal (tMTF)* modulation transfer functions (**Figure 16B**). The *best scale* (related to the inverse bandwidth) of an STRF is defined as the centroid of the sMTF (in "cycles/octave"), whereas "speed" or *best rate* of the STRF is defined as the centroid of the tMTF (in Hz), as illustrated in **Figure 16B**. To display the neural *population responses* for each phoneme, we generated two-dimensional "topographic" plots in which each row contained the average PSTH response of one neuron, sorted according to neural BF, scale or rate. The distribution of these three tuning properties in our sample was fairly broad, covering most BFs, best scales, and best rates (Figure 17). However, because the parameters were not distributed exactly uniformly, we interpolated the vertical axis of the smoothed PSTH (2-D disk filter: 60ms * 6 neurons) to have uniform spacing and then smoothed the PSTH display with the same

2-D filter. We characterized each phoneme according to the *locus* of maximal response within the neural population along the BF, scale and rate dimensions. For example, to find the locus along the BF dimension, we determined the position of the maximum PSTH responses over time for neurons ordered along the BF axis. The same procedure was repeated for PSTHs ordered along the scale and rate axes to obtain the three coordinates of the locus.

3.1.6 Phoneme classification and confusions

To examine the separation or overlap among the representations of different phonemes, we trained linear binary classifiers to discriminate each phoneme from all the others based on the neuronal population response. Formally, the neurons project the phoneme acoustic signals into a high dimensional space (i.e., the total number of neurons X the number of samples in each PSTH = 90 X 22). Because of the different selectivity of each neuron, different phonemes fall in specific sub-regions of this space.

A Linear Support Vector Machine (LSVM [27]) was trained to find the optimal hyperplanes for each phoneme, such that the hyperplane has the maximum distance (or "margin") to the closest data points (or "support vectors") in the two classes it separates. Using linear hyperplanes is intuitively appealing because the classifier's output is a weighted sum of the neural responses that can be interpreted easily. The output of each classifier is a scalar value indicating the distance of the data point to the hyperplane. Novel sounds are identified by choosing the classifier that produces the maximum distance to the boundary. We should emphasize that the order of the neural responses is not important in any way for classification.

3.1.7 Statistical analysis

The significance of correlations between the pattern of phoneme confusion predicted by the neural classifier and confusion observed for human perception [28] was ascertained by a randomized *t*-test. Random correlations were computed between neural and perceptual confusion matrices after randomly shuffling phoneme identity (20,000 shuffles). The significance of the correlation between the actual confusion matrices was taken as the probability that such a correlation could be produced by the randomly shuffled matrices.

3.1.8 Measuring the acoustic distance among phonemes

The average auditory spectrogram of each phoneme was computed as described above [25]. The acoustic similarity between any pair of phonemes was then defined as the Euclidean distance between their average spectrograms.

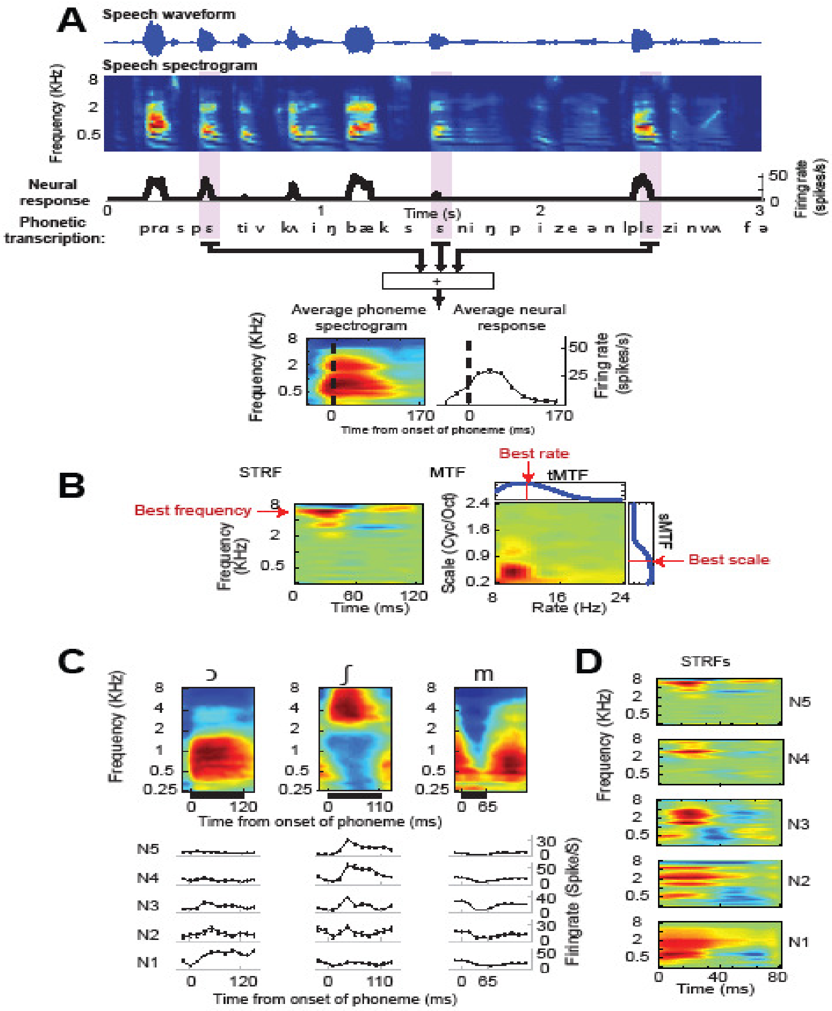

Figure 16. Neuronal responses to phoneme in continuous speech. **(A)** The spectrograms of all /ɛ/ vowel exemplars are extracted and averaged to obtain one grand average auditory spectrogram (bottom left). Red areas indicate regions of higher than average energy and blue regions indicate weaker than average energy. The corresponding PSTH response to /ɛ/ is computed by averaging neural spike rates over the same time windows (bottom right). **(B)** The spectro-temporal receptive field (STRF) of a neuron as measured by normalized reverse correlation. Red areas indicate stimulus frequencies and time lags correlated with an increased response, and blue areas indicate stimulus features correlated with a decreased response. The neuron's BF is defined to be the excitatory peak of the STRF (red arrow). The

modulation transfer function (MTF) is computed by taking the absolute value of the 2-D Fourier transform of the STRF. We then collapse along the temporal or spectral dimensions (known also as the *rate* and *scale*) to obtain the purely *spectral (sMTF)* or *temporal (tMTF)* modulation transfer functions. The *best scale* (proportional to the inverse of bandwidth) of an STRF is defined as the centroid of the sMTF (in "cycles/octave"), whereas "speed" or *best rate* of the STRF is defined as the centroid of the tMTF (in Hz). The choice of *centroid* for best-scale parameter results in a compressed range but it does not affect the ordering of neurons along this dimension. **(C)** Average auditory spectra of three phonemes (/ɔ/, /ʃ/, /m/). Below each spectrogram is the PSTH response of 5 example neurons (labeled N1-N5). **(D)** The STRFs of these neurons indicate a diversity of spectro-temporal tuning properties.

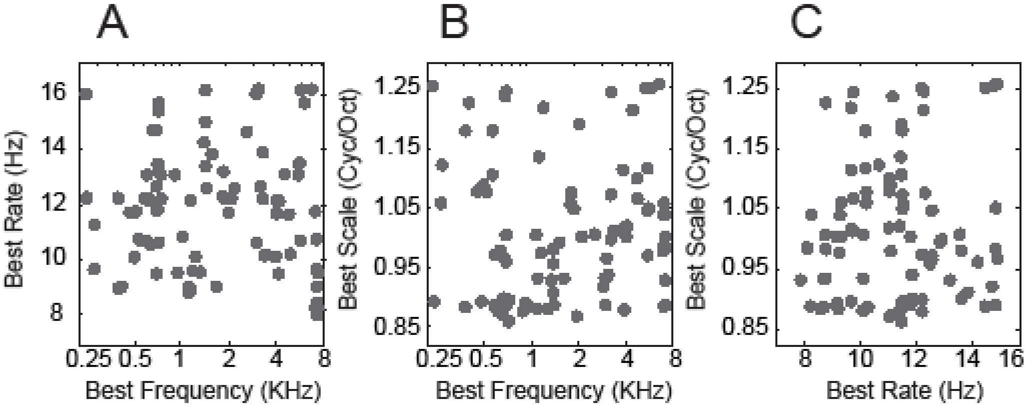

Figure 17. Joint distribution of neural parameters. Joint distributions of best frequency, best rate (A), best frequency, best scale (B) and best rate, best scale (C) of 90 neurons

3.2 Results

3.2.1 Diversity of single-unit responses to phonemes

Physiological responses were recorded from 90 single units in A1 of 4 ferrets (*Mustela putorius*) during the monaural presentation of continuous speech stimuli (see **Figure 16A**). The recorded neurons were broadly distributed in their spectral tuning and dynamic response properties as shown by population range of their best frequency (BF), best scale, and best rate (documented in the scatter plots in Figure

17). These neural tuning properties are based on measurements of the spectro-temporal receptive fields of the neurons (STRFs) as depicted in **Figure 16B** and described in detail earlier in Section **II**. **Figure 16C** illustrates the PSTH responses of 5 single units (N1-N5) to 3 different phonemes (vowel /ɔ/, fricative /ʃ/ and nasal /m/) whose average auditory spectra are depicted in **Figure 16C.** The spectro-temporal receptive fields (STRFs) of the 5 selected neurons are shown in **Figure 16D**.

Each phoneme activates these 5 neurons differentially, depending on the match between the neuron's STRF and the spectro-temporal structure of the stimulus. For instance, the vowel /ɔ/ drives N1 very effectively because of the low BF of the neuron (~ 700 Hz). By contrast, the fricative /ʃ/ maximally activates N4 and N5, which have the highest BF's (~3 KHz and ~7 KHz, respectively). Finally, the response pattern of the nasal /m/ is unique in that it causes a depression of responses in N2 and N3, reflecting the energy dip midway through the phoneme over all frequencies, but especially in the middle frequencies (~0.5 - 4 KHz) [20][21]. In this manner, each phoneme evokes a unique response pattern across the population of A1 cells that differs from the evoked responses elicited by other phonemes.

3.2.2 Population responses to phoneme classes

To appreciate the unique response patterns evoked by different phonemes and, in particular, in order to highlight the acoustic features enhanced in the neural representation, it is best to view the ordered activity of the entire population simultaneously. This ordering depends entirely on the neuronal tuning properties to be emphasized. For instance, inspired by the tonotopic organization of the auditory pathway, the most common way to organize neural PSTHs has been by frequency according to the BF of the units [29] [30]. However, unlike the receptive fields of fibers in the auditory nerve, A1 neurons exhibit systematic variations of tuning along a myriad of feature axes, including bandwidth, asymmetry, and temporal dynamics [16] [31] [32].

Here we consider the ordered representation of phoneme responses along BF and two other dimensions derived from the STRF: best scale and best rate (see Section on Experimental Procedures above and **Figure 16**B). Best scale is inversely proportional to bandwidth and indicates how wide a range of sound frequencies are integrated into the neural response. Best rate indicates the dynamic agility of a neuron's responses and hence reflects the temporal modulation of the stimulus spectrum that best drives the neuron. The coordinates of each cell along these three dimensions can be estimated using a variety of techniques and stimuli. The most common techniques include tuning curves or iso-response functions measured from tone [32] and STRFs measured from ripples [33]. We use the speech-based STRFs to estimate these parameters for each cell [22].

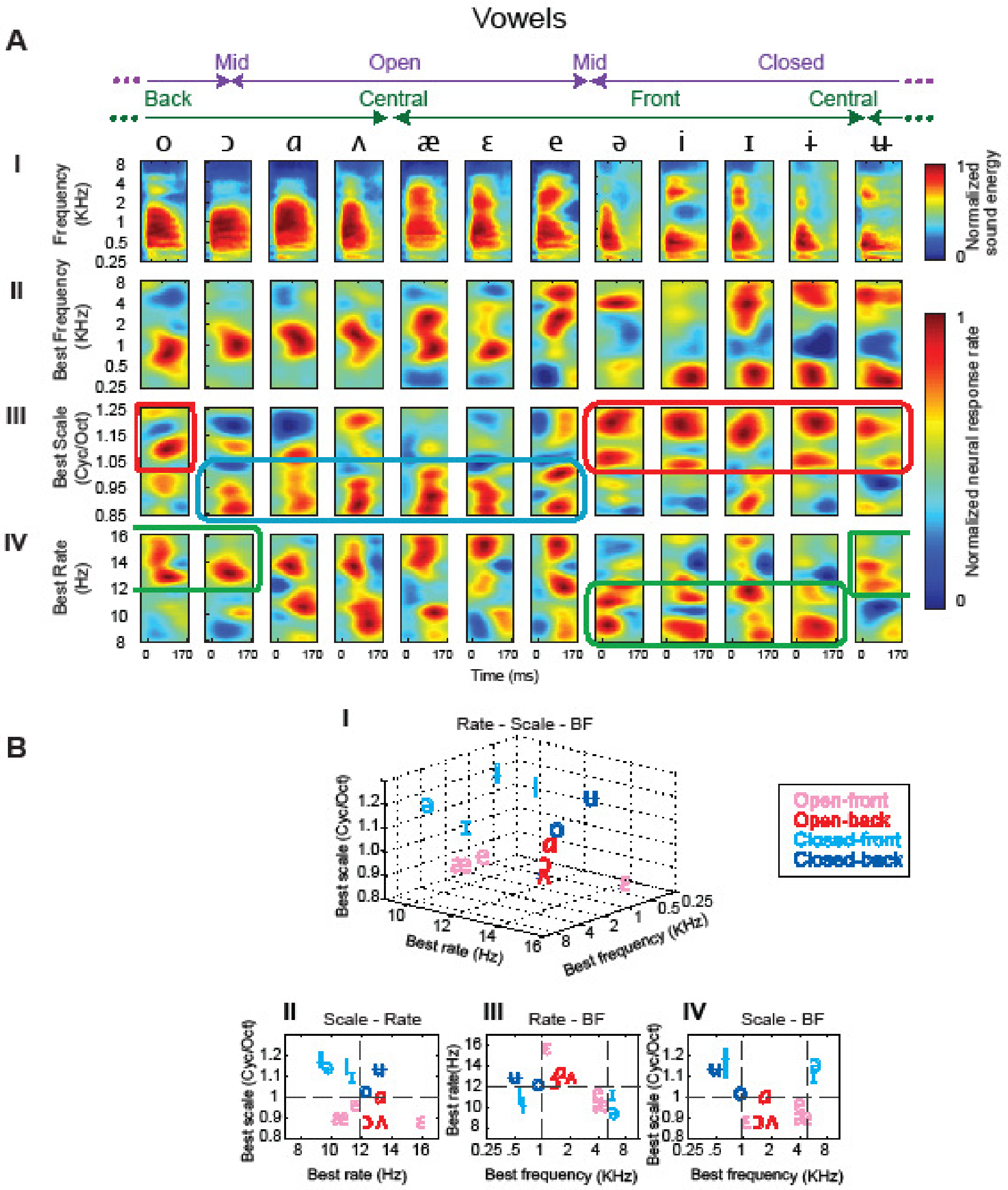

Figure 18: Population response to vowels (A) I. Average auditory spectrogram of 12 vowels organized approximately according to their open-closed and front-back articulatory features. The arrows at top indicate the *degree* of these features, with arrow "tips" representing minima (mid or central) and midpoints representing maxima. For example /ʌ/ is maximally open, but is neutral (central) on the front/back axis. Note also that the axes are

presumed to loop around the page from right to left (dashed ends joining) creating a circular representation **(II, III, IV)**: Average PSTH responses of 90 neurons to each vowel. Within each heat map, each row indicates the average response of a single neuron to the corresponding phoneme. Red regions indicate strong responses, and blue regions indicate weak responses. The average PSTH responses are sorted by neurons' best frequency (**II**), best scale (**III**) and best rate (**IV**) to emphasize the role of that parameter in the encoding of each vowel. (Details of the analysis and generation of these plots are given in Section **II**). (**B**) **I.** Each vowel is plotted at the centroid frequency, rate and scale of its average neuronal population response. The centroid values are calculated from the average PSTH responses sorted by the corresponding parameter (**2A**). Open vowels are shown in red, Closed vowels in blue, Front vowels with *hollow* font, and Back vowels with *solid*. To visualize the contribution of each tuning property to vowel discrimination, the location of each vowel is also shown collapsed in 2-D plots of (**II**) rate-scale, (**III**) rate-frequency and (**IV**) scale-frequency. All other details of the analysis and generation of these plots are given in Experimental Procedures.

3.2.3 Encoding of vowels

Population responses to 12 American-English vowels are summarized in **Figure 18**. Panels in the top row (**Figure 18A-I**) display the average auditory spectrogram of each vowel computed from all of its samples encountered in the speech database (see Section **II** for details). The vowels are organized according to their articulatory configurations along the Open/Closed and Front/Back axes [3], as illustrated at the top of **Figure 18** : /o/, /ɔ/, /ɑ/, /ʌ/, /æ/, /ɛ/, /e/, /ə/, /i/, /ɪ/, /ɨ/, /ʉ/. The three middle vowels (/ɛ/, /e/, /ə/) are tightly clustered near the midpoint of the Front/Back and Open/Closed axes, and are difficult to order accurately along this 1-dimensional representation of the vowels.

The averaged spectra (top row) reveal that Mid/Back vowels (/o/, /ɔ/, /ɑ/, and /ʌ/) have relatively concentrated activity at low to medium frequencies (~0.4 - 2 KHz), whereas Front vowels sometimes have two peaks spaced over a larger frequency range (~0.3 and ~4 KHz). This is consistent with the known distribution of the three formants (F1, F2, and F3) in these vowels [3], namely, that they have F1 and F2 that are closely spaced, creating compact single broad peak spectra at intermediate frequencies (reminiscent of the center-of-gravity hypothesis of Chistovich and Lublinskaya [34]). As the vowels become more "Front"ed, the single peak broadens and splits (/æ/ to /ə/). Continuing this trend, Front/Closed vowels (/i/, /ɪ/, /ɨ/, /ʉ/) exhibit relatively narrow and well separated formant peaks with F1 at low and F2 at high frequencies.

These averaged phoneme spectra are broadly reflected in the response distributions ordered along the BF axis; neurons with BFs matching regions of high energy in a phoneme spectrum tend to give strong responses to that phoneme (**Figure 18A-II**). However, notable differences of unknown significance exist such as the relative weakness of the low BF peaks in /e/ and /ə/, and of the high BF peak in /i/). More striking, however, are the response distributions along the best scale axis, which roughly indicates the *inverse* of the vowels' spectral bandwidths (**Figure 18A-III**). Here, consistent with the bandwidths of the spectral peaks discussed earlier, Central/Open vowels tend to evoke maximal responses in broadly tuned cells commensurate with their broad spectra (low scales < 1 Cyc/Oct) while Closed vowels evoke maximal responses in narrowly tuned cells (scales > 1 Cyc/Oct), as indicated by the blue and red boxes in **Figure 18A-III**, respectively[1]. Response distributions in the best rate panels (**Figure 18A-IV**) reveal a trend in the dynamics of the vowels as one moves along the Front/Back axis. Specifically, Front vowels (/ə/, /i/, /ɪ/, /ɨ/) evoke relatively stronger responses in the slower cells (with best rates <~ 12 Hz), as compared to the more Back vowels (/ʉ/, /o/, /ɔ/) as highlighted by the green boxes in **Figure 18A-IV**. The remaining more Central vowels (/ɑ/, /ʌ/, /æ/, /ɛ/, /e/) exhibit all dynamics. This response pattern may reflect the longer durations required to complete the articulatory excursions toward or away from Closed vowels towards the front of the vocal tract.

Figure 18B provides a compact summary of the population response to vowels. Each vowel is placed at the *locus* of maximum response in the neural population along the BF, best scale, and best rate axes. To highlight more clearly which of the three features best segregates them, the 3-D display is projected onto each of the three marginal planes (**Figure 18B-II** and **Figure 18B-IV**)). It is readily evident in these displays that the Open and Closed vowels separate along the scale axis above and below 1 Cyc/Oct (horizontal dashed lines in **Figure 18B-II** and **Figure 18B-IV**)). They are also distinguished by BF, with the Open vowels clustering in the range 1.0 – 4.5 KHz (vertical dashed lines in **Figure 18B-III**). Finally, the best rate axis segregates the Front/Back vowels (as discussed earlier), with Central and Back vowels located at high rates (> 12 Hz), and Front vowels below it. It remains to be

[1] We emphasize that this response pattern is unlikely to be due to a non-uniform sampling of the scale and frequency variables, since no such bias in the joint distribution of the scale-frequency is evident in Figure 17. Furthermore, note that high scale neurons can be driven well by spectra with low frequencies as in phoneme /o/. The opposite is true for vowel /e/ where low scale units are driven well by high frequency energy.

confirmed, however, whether these locations which reflect the vowels' overall spectro-temporal similarity, can explain the perceptual confusion among them [35].

3.2.4 Encoding of consonants

Population responses to 15 consonants are shown in **Figure 19** in the same format already described for vowels. Three properties are commonly used to organize and classify consonants: place of articulation, manner of articulation, and voicing [3] [21]. Here we examined how these three properties are encoded in the responses of the neuron population.

The distributions of the responses to the consonants sorted along the BF axis (**Figure 19A-II**) approximates the features of their averaged spectra (**Figure 19A-I**), which in turn are known to be closely related to place of articulation cues. For instance, the difference between the more forward places of constriction for /s/ compared to /ʃ/ is mirrored by the downward shift of the highpass spectral edge. Similarly the high-frequency noise burst at the onset of the forwardly-constricted /t/ contrasts with the lower-frequency distribution of the other plosives (/p/, and /k/). However, there are also some notable differences in detail between the two sets of plots. There is generally a slight delay of about 20 milliseconds in the neural responses relative to the spectrograms (presumably due to the latency of cortical responses). In addition, however, there are substantial differences between the responses and spectrograms in certain phonemes. For example, high BF responses to /f/ in **Figure 19A-II** are strong despite their relative weakness in the spectrograms. Similarly, the low BF responses to /v/ are not consistent with the spectrogram. In other consonants, there are differences in the "timing" of certain frequency regions such as the rapid onset of high frequencies in the spectrogram of /t/ relative to its more delayed response, or in the continuity of the spectral regions in /ʃ/, /d/ and /ŋ/. The origin of all these differences is unclear and may reflect the nonlinearity of neural responses or our limited sampling of the neural population (90 neurons).

Response distributions along the best scale and best rate axes (**Figure 19A-III** and **IV**) capture well the essential *manner of articulation* cues that supply the information necessary to discriminate plosives, fricatives, and nasals in continuous speech. For example, the broad distinction between "plosives" and "continuants" (e.g. /p/, /t/, /k/, /b/, /d/, /g/ versus /s/, /ʃ/, /z/, /n/, /m/, /ŋ/) is evident in the distribution of responses along the scale and rate axes (**Figure 19A-III** and **IV**). Thus, plosives with their sudden and spectrally broad onsets display relatively strong activation in broadly tuned (low scales < 1.1 cyc/oct) and fast (rates > 12 Hz) cells (regions outlined in red in **Figure 19A-III** and **IV**) compared to the more suppressed responses to longer duration unvoiced fricatives and nasals (outlined in blue in **Figure 19A-IV**). Note also the brief suppressed response preceding the onset of all plosives due to the (silent) voice-onset-time (VOT) in all panels within the red box (**Figure 19A-III** and **IV**).

Finally, the third cue of voicing is associated with the harmonic structure of voiced spectra near the low to mid-frequency range (0.2 to 1 KHz), and to a lesser extent the weak energy at low BFs near the fundamental of the voicing. Only this latter cue seems to distinguish consistently the voiced (/b/, /d/, /g/, /v/, /ð/, /z/, /m/, /n/, /ŋ/) from unvoiced (/p/, /t/, /k/, /f/, /s/, /ʃ/) consonants in our data as indicated by the green outlined region of **Figure 19A-II**. However, such a strong low BF response as an indicator of "voicing" is missing in many of the vowel responses discussed earlier (e.g., the Open/Back vowels in **Figure 19A-II**). Instead, its presence seems to correlate with the low F1 of the Closed vowels there. Therefore, our data suggest that the low frequency voicing is reliably represented only in consonant responses, and perhaps in vowels where the F1 is low enough to amplify it [36]; however, there may well be a different and separate representation of voicing in the auditory cortex, for example in terms of the pitch it evokes, or the harmonicity of its spectral components [37].

Figure 19B illustrates the locus of the population response to each consonant in a plot of best frequency, best rate and best scale similar to that used with vowels earlier. The lower panels of **Figure 19B** are projections of the 3-D plot onto its three marginal planes. Members of the three groups of consonants - plosives (red), fricatives (blue), and nasals (green) - are located roughly close together in this parameter space. For instance plosives tend to drive broadly tuned (scale < 0.9 Cyc/Oct) and fast (rates > 12 Hz) cells (**Figure 19B-II**). Rate is also a distinguishing feature between plosives on the one hand, and nasals and (most) fricatives on the other (above and below 12 Hz, respectively). Similarly, phoneme groups roughly segregate along the BF axis, with unvoiced fricatives occupying the highest frequencies (> 4KHz), unvoiced plosives falling between 2-4 KHz, and other voiced phonemes falling below 2 KHz (**Figure 19B-III** and **IV**). As with vowels, this plot of the neural loci of consonants reveals the relative distances among them and perhaps explains the pattern of perceptual confusion observed among them, as we shall elaborate next.

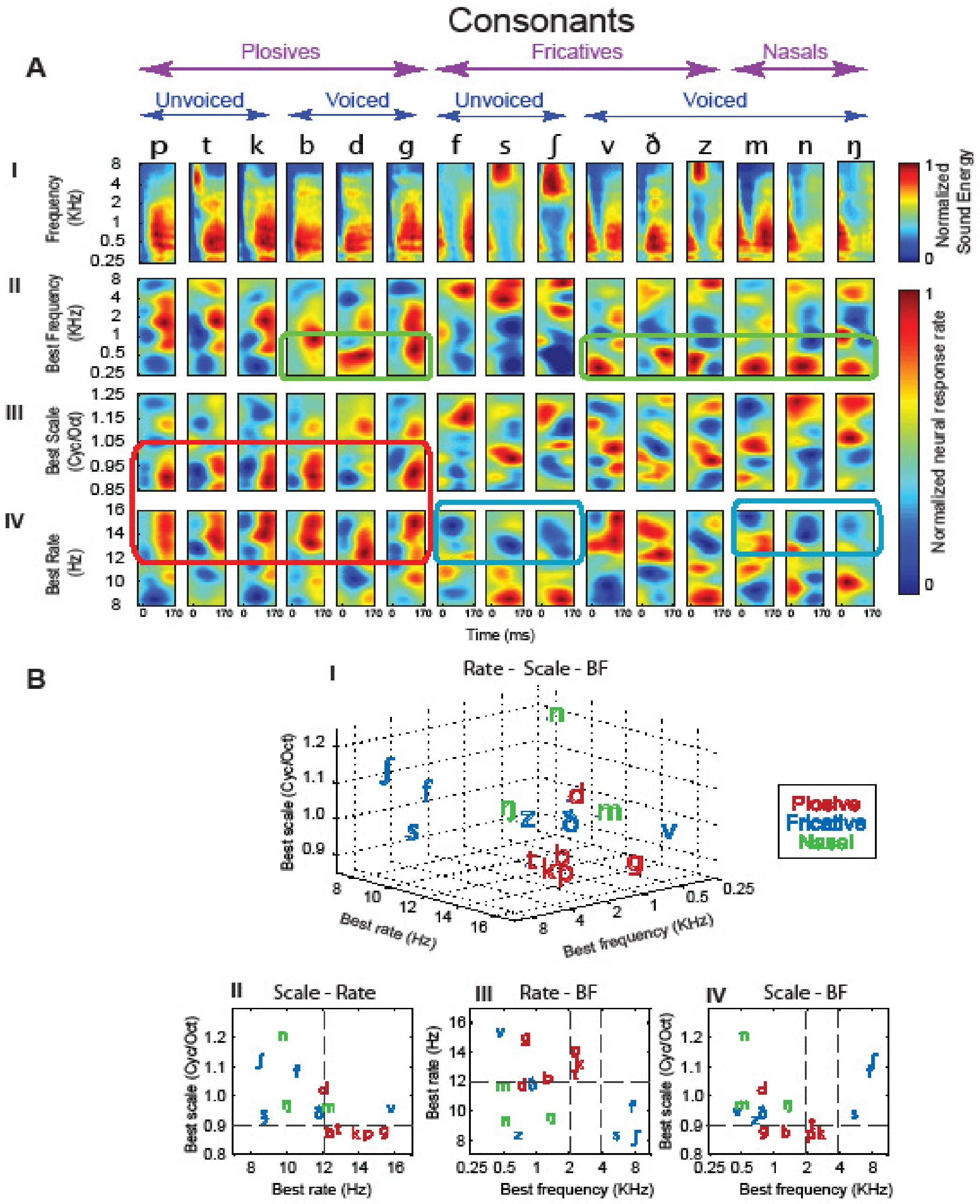

Figure 19. Population response to consonants (A) I. Average spectrogram of 15 consonants phonemes grouped as 6 plosives, 6 fricatives and 3 nasals. Each of the plosive and fricative groups contains 3 voiced and 3 unvoiced phonemes (see arrows at top). (**II, III, IV**) Average PSTH responses of the

neural population to each consonant, plotted as in **Figure 18A**. The average PSTH responses are sorted by neurons' best frequency (**II**), best scale (**II**) and best rate (**IV**) to emphasize the role of that parameter in the encoding of consonants. (All other details of the analysis and generation of these plots are given in Section **II**). (**B**) Each consonant is placed at the centroid frequency, rate and scale of its neuronal population response, measured from the corresponding PSTH responses (**Figure 19A**). Plosive phonemes are plotted in red, fricatives in blue and nasals in green. The locus of each consonant is also shown collapsed in 2-D plots of (**II**) rate-scale, (**III**) rate-frequency and (**IV**) scale-frequency. (All other details of the analysis and generation of these plots are given in Section **II**).

3.2.5 Phoneme confusions

Average phoneme responses give useful insights into the mean representation of each phoneme, but they fail to indicate how well the neural population can discriminate phonemes, given the natural acoustic variability among samples of the same phoneme during continuous speech. To delineate perceptual boundaries implied by the responses to the phonemes, we trained a linear classifier for each phoneme to separate it from all others, based on the PSTHs of the neural population. To determine the identity of a novel phoneme, the population response was applied to all the classifiers, each computing the likelihood of its designated phoneme. The classifier indicating the maximum likelihood was taken as the identity of the input phoneme. To train and test the classifiers, we divided the speech data into 100 train and test subsets. In each subset, 90% of the data was randomly chosen for training and the remaining 10% was used for testing. The classification accuracy and the confusion matrices reported here are the average results of the 100 subsets.

Once trained, each linear classifier can be viewed as a mask that selects, by multiplication with the population response, the neurons and response latencies that most effectively distinguish the associated phoneme from all others. **Figure 20** displays the masks computed for the unvoiced consonants /p/, /t/, /k/, /f/, /s/, /ʃ/. The masks are ordered in the same way as the PSTHs in **Figure 19A** (i.e., by BF, best scale, and best rate). In the masks, black regions signify neurons and response latencies for which a strong response provides evidence for the phoneme, and white regions signify strong responses that provide evidence against that phoneme. The masks in **Figure 20** differ from the mean neural responses in **Figure 19A** in that they emphasize the *unique* features of each phoneme. For example, the mean responses to /ʃ/ (**Figure 19A-II**) indicate strong responses in high and medium BF neurons, but in the masks the mid-BF neurons (2 KHz) are given higher weights. This differential weighting reflects the fact that both /ʃ/ and /s/ evoke strong responses from high BF neurons, but only /ʃ/ evokes responses from the mid-BF neurons. Similarly, the /p/, /t/, /k/ masks reflect only the features that distinguish these phonemes from each other. The BF masks (**Figure 20A**), emphasize the low (750 Hz), high (> 2 KHz), and medium (0.3-1.5 KHz) spectral regions for the /p/, /t/, /k/ bursts, respectively. Note also how the rate masks (**Figure 20C**) distinguish plosives /p/, /t/, /k/ from the long fricatives /s/, /ʃ/ by enhancing the regions outlined in the rectangle, namely the slow rates of the fricatives (< 11 Hz) relative to the faster rates of the plosives. It should be noted that the classifier performance does not depend in any way on the *order* of the neural responses, which is solely used for analysis and display purposes.

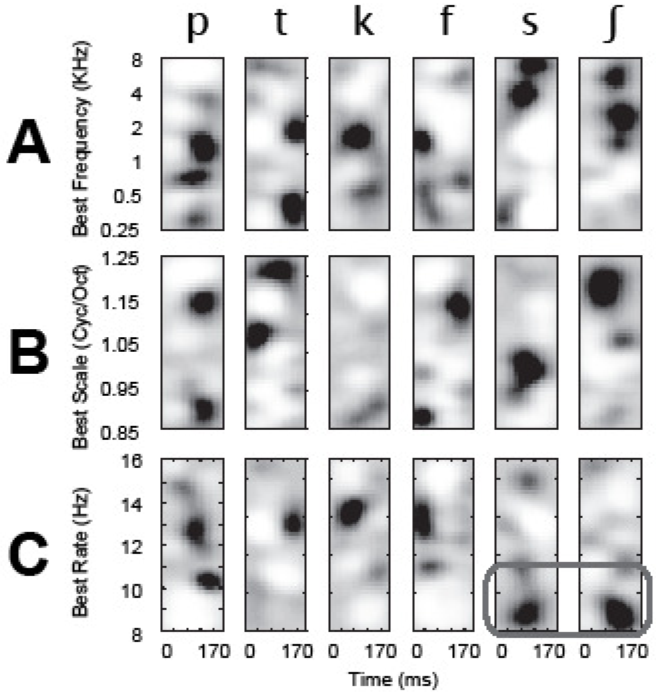

Figure 20. Phoneme classification based on the population response Classification masks for 3 unvoiced plosives (/p/, /t/, /k/) and 3 unvoiced fricatives (/f/, /s/, /ʃ/) sorted by neurons' best frequency (**A**), best rate (**B**) and best scale (**C**). Grey scale indicates the importance of the presence (*black* regions) or absence (*white* regions) of neural response for the classification of that phoneme. The output of each phoneme classifier is a scalar, computed as the sum of the population PSTH multiplied by the mask. Thus the order of the mask/PSTH is irrelevant to the output of the classifier.

The extent to which the neural phoneme representations can account for the perception of *individual* phoneme exemplars can be assessed by studying the pattern of pair-wise confusions by the classifier. **Figure 21A** shows the confusion matrix measured from classifications of the neural data. Labels along each row indicate the phoneme presented, and columns report the probability of the phoneme output by the classifier [28] [38]. The classifier was trained on two sets of data. In a small set of 20 neurons, we succeeded in measuring responses to 330 seconds of speech (90 sentences) to be used in the training; these are shown in **Figure 21**. In an ideal case that all phonemes are accurately identifiable, we would expect to see a diagonal confusion matrix. Off-diagonal values represent misidentification. The phonemes are arranged based on voiced-unvoiced and plosive, fricative, nasal consonant categories to facilitate comparison with a previous study of human perception [19] [38] (replicated in **Figure 21B**). The dashed boxes delineate the 3 major phoneme categories: plosives, fricatives, and nasals. In both neural and perceptual data, phonemes within each category—plosives (/p/, /t/, /k/), fricatives (/f/, /s/, /ʃ/), and nasals (/m/, /n/)—tend to be more confusable within the group than across categories. The correlation coefficient between the complete neural and perceptual matrices is 0.78 (*p*=0.0002, randomized *t*-test). Ignoring the confusions between voiced and unvoiced consonants improves the similarity to 0.86, with a correlation of 0.95 for only the unvoiced consonants and 0.71 for their voiced counterparts. At least some of the difference between confusion matrices reflects noise due to limited sampling of neural responses, or limited data for training the phoneme classifiers.

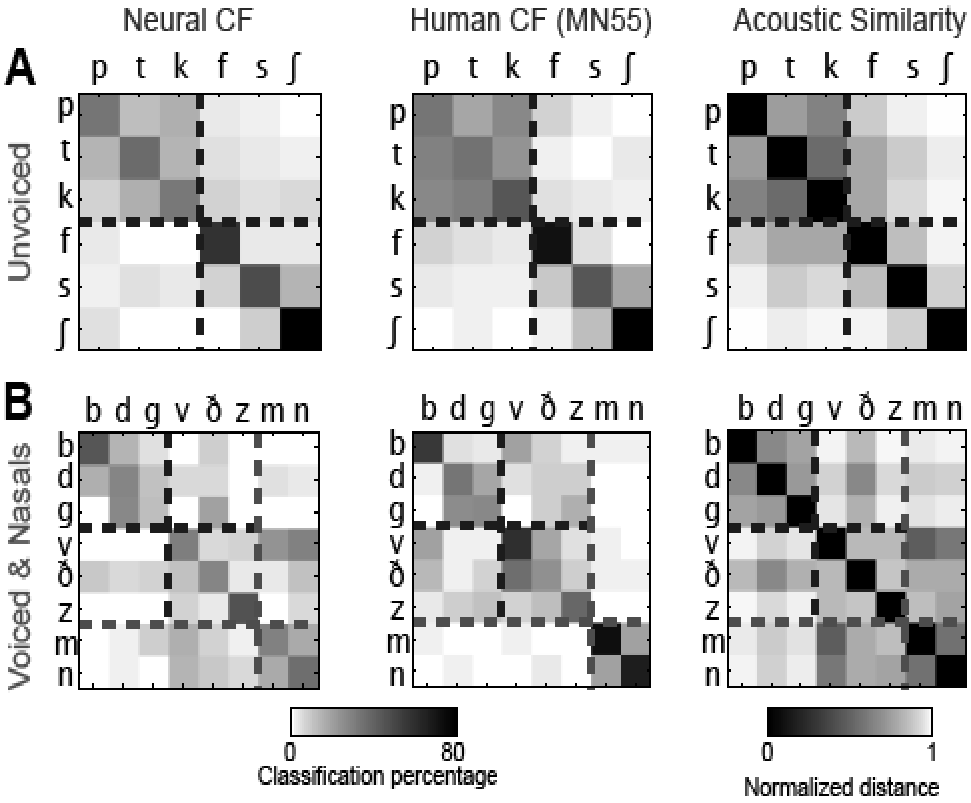

Figure 21. Neural and human phoneme confusions, and phonemes acoustic similarity. Consonant confusion matrices from neural phoneme classifiers (**left panels**) and human psychoacoustic studies [28] (**middle panels**). Grayscale indicates the probability of reporting a particular phoneme (column) for an input phoneme (row). (**Right panels**) The acoustic similarity between phoneme pairs defined as the Euclidian distance between their average auditory spectrograms. (**A**) Confusion matrices and phonemic distances for unvoiced consonants. Blue lines separate the plosives /p/, /t/, /k/ from fricatives /f/, /s/, /ʃ/. (**B**) Confusion matrices and phonemic distances for voiced consonants. Blue lines separate the plosives /b/, /d/, /g/ from fricatives /v/, /ð/, /z/ and the red lines distinguish the nasal consonants /m/ and /n/ from the rest.

Alternatively, we explored the sensitivity of the classification in **Figure 21** to the number of neurons included (using the same training material). As expected, the results indicate that percentage of correct classification (averaged across all consonant phonemes) improves as the number of randomly selected neurons is increased (**Figure 22**). More detailed exploration of this issue should take into account the differential contribution of specific neurons to different phonemes, e.g., high BF neurons to the classification of /s/ and /ʃ/.

Finally, we also explored the extent to which both the neural and human confusion matrices are a reflection of the acoustic similarity (or "distances") among the phonemes at the level of the auditory spectrograms [25]. **Figure 21** illustrates that such a phoneme "similarity matrix" fundamentally resembles the human and neural confusion matrices (with correlation coefficients of 0.66 and 0.93, respectively). In fact, the neural matrix encodes remarkably well details of the phoneme acoustic similarity, such as the confusions between /v/ and the nasals /m/, /n/, and also between /ð/ and the voiced consonants /b/, /d/, /g/.

3.3 Discussion

Neuronal responses to continuous speech in the primary auditory cortex of the naive ferret reveal an explicit multidimensional representation that is sufficiently rich to support the discrimination of many American English phonemes. This representation is made possible by the wide range of spectro-temporal tuning in A1 to stimulus frequency, scale and rate. The great advantage of such diversity is that there is always a unique sub-population of neurons that responds well to the distinctive acoustic features of a given phoneme and hence encodes that phoneme in a high-dimensional space.

As an example, consider the perception of the plosive consonant /k/ in a CV syllable, which is identified by a conjunction of several acoustic features: an initial silent voice-onset-time (VOT), an onset burst of spectrally broad noise, and the direction of the following formant transitions. Each of these features can be encoded in the cortical responses along different dimensions. Thus, neurons selective for broad spectra respond selectively to the noise burst. Rapid neurons respond well following the VOT, whereas directional neurons selectively encode the vowel formant transitions. In this manner, /k/ is encoded *robustly* by a rich pattern of activation that varies in time across the neural population. This neuronal activation pattern constitutes the phoneme representation in A1 and presumably forms the input to a set of neural "phoneme classifiers" in higher auditory areas. If one acoustic feature is distorted or absent, the pattern along the other dimensions (and hence the percept) remains stable.

We have focused here on describing a few prominent features of the response distributions that correspond to well-known distinctive acoustic features of the consonants considered [21]. There are clearly many other aspects and more details of the responses that reflect intricate articulatory gestures, contextual effects, or speaker-dependent variability that can only be reliably considered with a much larger sample of responses. One example is the distribution of the *directionality index* of the responses in the neighborhood of a consonant [39], an attribute that would indicate whether the formants are upward or downward sweeping, or if they are converging towards or diverging away from a locus frequency.

Humans confuse the phonemes of their native tongue when placed in unusual or noisy contexts. Typically, phonemes that share some acoustic features tend to be more confusable than those that do not. This was confirmed by the similarity we found between the acoustic distance and the human confusion matrices. Similarly, since A1 responses in our naive ferrets also preserve the relative acoustic distances between the phonemes (as they would presumably for other complex sounds), we are led to the conjecture that human phoneme perception can (in principle) be explained in large measure by basic auditory representations such as the auditory spectrogram and the cortical spectrotemporal analysis common to many mammalian (and also avian) species [8] [11] [12] [40] [41].

The representation of phonemic features across a population of filters tuned to BF, scale and rate suggests a strategy for improved speech recognition systems, and further study may reveal additional strategies for speech processing. However, many questions about the neural representation of phonemes still remain unclear; for example, how can one extrapolate from such neurophysiological findings to the human perceptual ability to perceive phonemes categorically (also found in monkeys [13], cats [10], chinchillas[5], birds [11] and rats [42]), and to shift categorical boundaries arbitrarily between phoneme pairs? In summary, humans' ability to discriminate perfectly their native phonemes is the result of years of training. Naïve ferrets lack such a history, and hence their perception of clean phonemes is more akin to that of humans listening to noisy phonemes. In both cases, confusion patterns would reflect the acoustic distances between the phonemes. However, if ferrets are trained to actively discriminate phonemes, it is likely that dimensions useful for this specific discrimination would be emphasized, creating the heightened sensitivity necessary to perform the task. This is presumably what happens in humans as they learn their phonemes, and what the classifier essentially simulates in our analysis when it learns the masks and boundaries that enable robust phoneme discriminations. Therefore, from a neural perspective, one may view the masks as either a subsequent layer of synaptic weights *or* as pattern of behaviorally-driven plasticity of AI receptive fields - the end-result of perceptual learning in which neurons adapt their tuning along the dimensions appropriate for the phoneme discrimination task. This same general principle would apply to any complex sound, using additional cortical response dimensions, such as pitch, spatial location, and loudness.

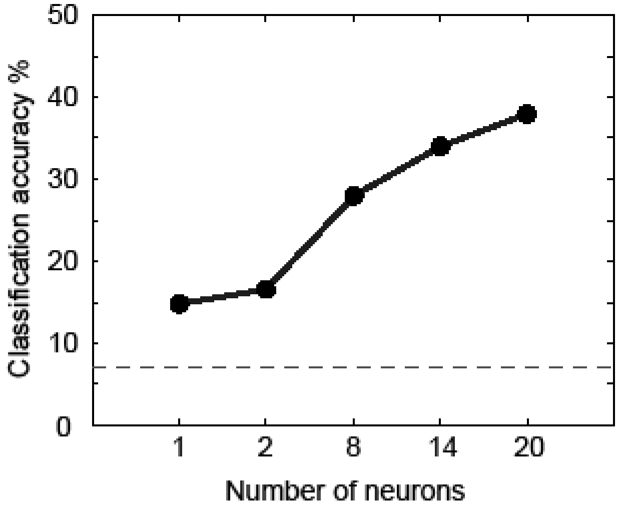

Figure 22. Dependence of phoneme classification accuracy to the number of neurons. Classification accuracy as a function of the number of neurons used by the classifier. The red line indicates chance performance (7% for 14 phonemes) (see Section **II** for details).

Chapter 4

Prediction of speech representation in the primary auditory cortex

4.1 Introduction

We showed in the previous section that a large population of neurons in the primary auditory cortex encodes the perceptually important features of phonemes along various dimensions, including frequency, spectral and temporal modulations. These parameters for the neurons in the auditory cortex can be estimated from a linear receptive field model of auditory neurons, STRF as we showed in the previous chapter. Here, we demonstrate that the predicted responses from a much larger population of neurons in the auditory cortex results in the same multidiemsional representation as the actual responses to phonemes in the auditory cortex. The predicted responses preserve all the categorical distinctions between phoneme groups along different tuning dimensions. Having a large set of receptive fields provide a more complete coverage of the parameter space and let us investigate the role of different neural tuning in encoding the perceptually and categorically important features of speech. We will examine the role of several tuning characteristics in the encoding: frequency, temporal (rate) and spectral (scale) modulations and directional selectivity in the actual and predicted responses.

4.2 Methods

The physiological experiments were executed in a manner similar to what we described in the previous chapter. We have already shown how the best frequency, rate and scale are estimated from the STRFs of the neurons. We also used the MTF function to measure the directionality preference of the neurons for downward and upward going modulations. The directionality index was estimated by finding the ration of the power in the first quadrant (upward) to the second (downward) of the MTF. As in the previous chapter, to highlight the role of a specific parameter in the encoding of the phonemic categories, we sorted the average neural responses to that phoneme according to the parameter of interest.

For the predicted responses, we used STRFs of 2500 primary auditory cortex neurons to predict the response of the population to speech stimuli using STRF equation:

$$r(t)=\sum_{f}\int s(\tau,f)\,h(\tau-t,f)\,d\tau$$

We used the same characteristic parameters as the real data to estimate them from the 2500 STRFs and we used the sorted predicted responses along different parameters to highlight the categorical encoding of phonemes along that parameter.

4.3 Encoding of vowels

The average neural and predicted responses to different vowels are shown in **Figure 23**. When the responses are sorted by the best frequency of neurons (**Figure 23a**), the predicted responses (second row) match the actual ones with good accuracy. For example, the gradual increase in the frequency of the neurons responding to open vowels or the two peaks (two dominant formants) of closed vowels that can be seen in both the actual and predicted responses. **Figure 23b** shows the sorted responses according to the scale of the neurons in which a close match can be seen between actual and predicted responses. Closed vowels activate the high scale neurons better (black box in **Figure 23b**) and open vowels activate the low scale neurons which happens both in the actual and predicted responses (**Figure 23b** black and blue boxes). Similarly, in the panels that show the actual and predicted responses sorted by rates (**Figure 23c**), we see that the different responses of fast and slow neurons to open and close vowels is evident both in the actual and predicted responses. Finally, when we sort the responses by the direction preference of the neuron (as described in the methods), we see that the neurons that respond to upward going modulations are activated more strongly by open-back vowels (**Figure 23d** bottom row).

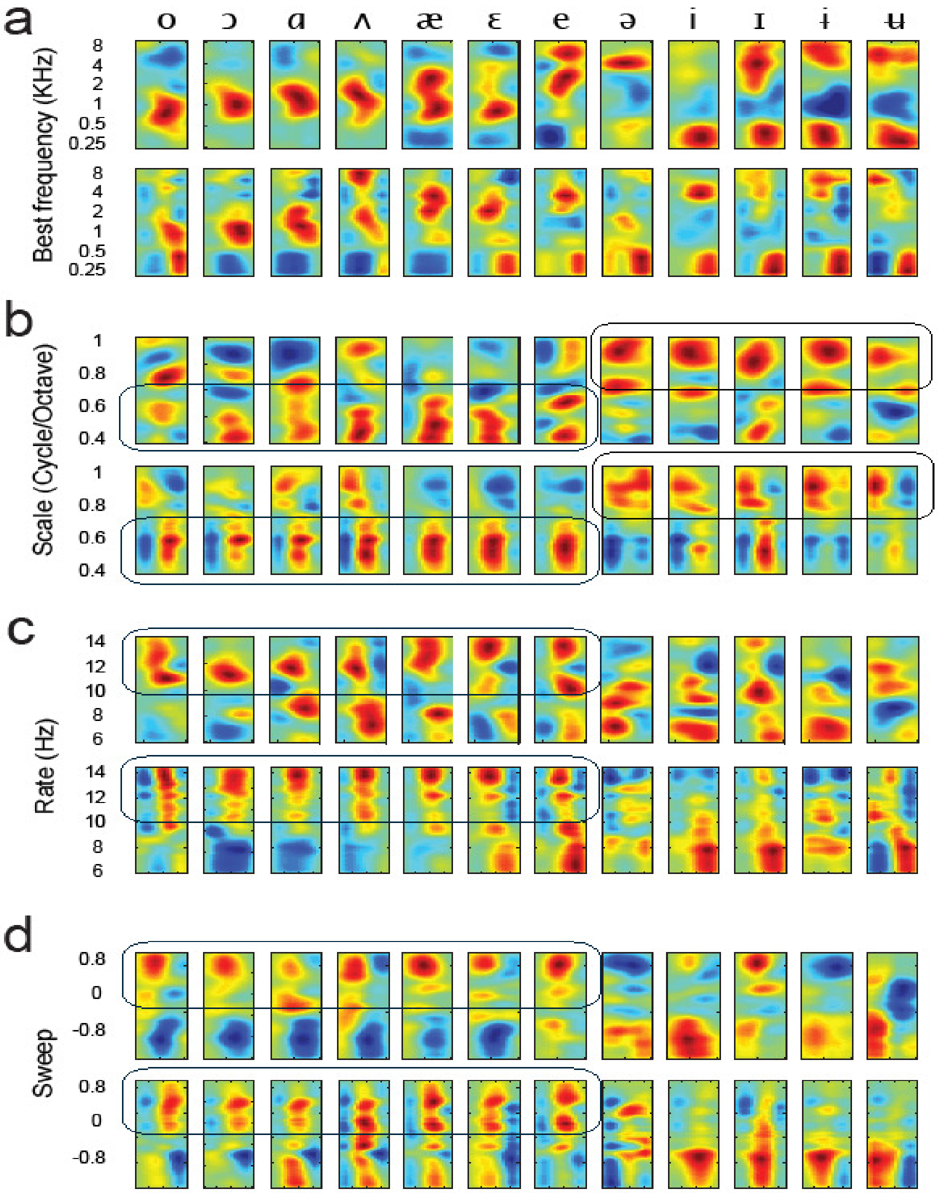

Figure 23. Actual and predicted average vowel neural responses. (a), (b), (c) and (d) show the average actual (top rows) and predicted (bottom rows) of neural responses to vowels sorted by best frequency, best rate, best scale and best direction correspondingly. In all cases, the actual and predicted responses show similar categorically distinct patterns (highlighted by boxes).

4.4 Encoding of consonants

Different population of neurons respond distinctively to different consonants based on how they match the spectro-temporal tuning of neurons as we described in the previous section. Here, to see if this distinct pattern of response can be captured by the linear tuning model of neurons, we used the predicted neural responses to different consonants and compared it to the actual one as shown in **Figure 24**. It is evident that the neurons with best frequency tuning that respond to certain phonemes show the same pattern in their predicted responses. For example, the population that respond to fricative /s/ are mostly high frequency cells, and the predicted responses of high frequency neurons are also very strong to /s/ (**Figure 24a)**. When we sort the responses according to their best scale and rate, we see that the patterns described in the previous section are also observed in the predicted ones: plosives activate low scale and high rate neurons, and fricatives and nasals activate high scale ones. In addition, when we consider the sorted responses along the directionality preference of the neurons, we see that neurons with diverse directionality tuning respond selectively to different consonants. For example, neurons that are tuned to upward spectrotemporal modulations respond well to plosive /p/, the ones with downward preference are active when /t/ is spoken, and /k/ activates the neurons with less directionality preference. In addition, the high frequency fricatives seem to activate upward population better (green boxes in **Figure 24d**, top row). This pattern is captured well by the predicted responses as well (**Figure 24d,** bottom row).

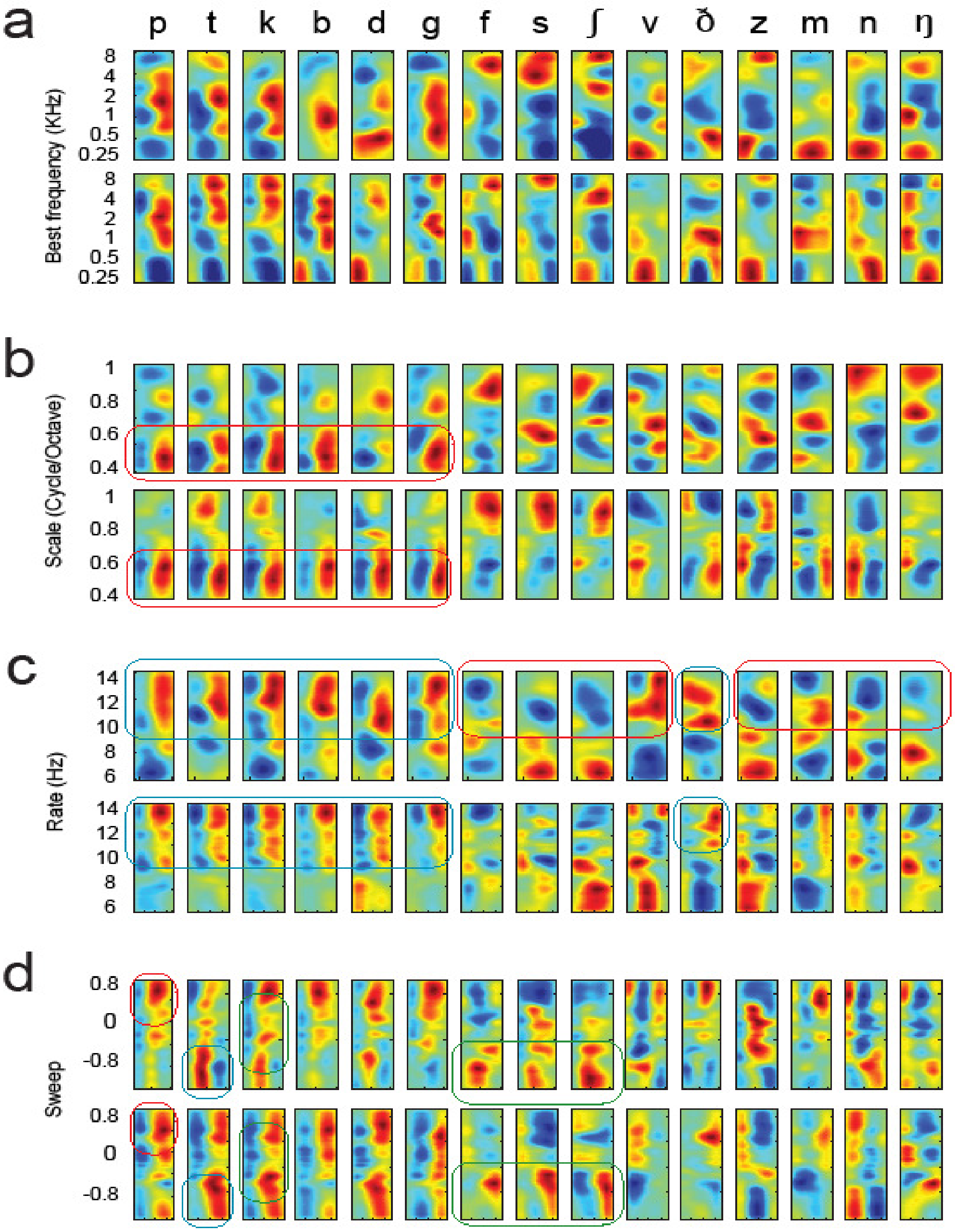

Figure 24. . Actual and predicted average consonant neural responses. (a), (b), (c) and (d) show the average actual (top rows) and predicted (bottom rows) of neural responses to consonants sorted by best frequency, best rate, best scale and best direction correspondingly. In all cases, the actual and predicted responses show similar categorically distinct patterns (highlighted by boxes).

Finally, to do a more quantitative analysis on the predicted and actual responses, we estimated the correlation coefficients between the average actual and predicted responses to phonemes. The correlation coefficients for different phonemes are shown in **Figure 25**. For most phonemes, the correlation is quite good validating our hypothesis that the distinct patterns observed in the precious section are the result of neural tunings that can be explained by a linear STRF model.

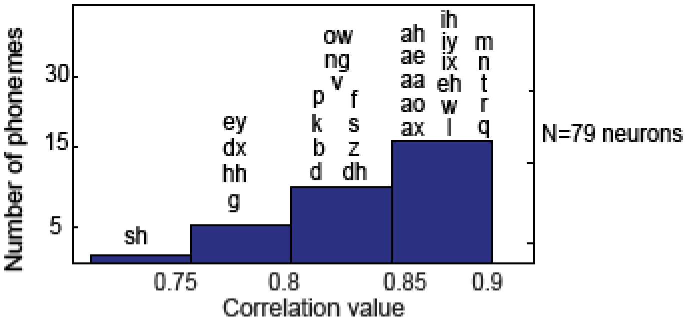

Figure 25. Correlation coefficients between average predicted and actual neural responses to phonemes. Vowels show the highest correlations (more than 0.85) while most fricatives and plosives have a correlation higher than 0.8.

Chapter 5

Reconstruction of speech from population responses in auditory cortex

5.1 Introduction

Population responses of cortical sensory neurons encode considerable details about their stimulus structure, detail that is often difficult to discern because of the complexity and diversity of cortical receptive fields. The typical approach used for understanding how stimuli are encoded across a neural population is to examine the distribution of some tuning property measured for a large set of neurons. By examining the distribution of tuning, one can infer that ranges spanned by a large number of neurons reflect stimulus features that are encoded with greater fidelity than ranges spanned by fewer neurons. Such an approach has been used to develop fundamental descriptions of most sensory systems (e.g. [43] [44]).

A more complete understanding of population codes can be developed by visualizing responses to several different stimuli after sorting the neurons according to basic tuning properties. This approach is most commonly applied to the auditory system with the "neurogram," in which neural responses are sorted by best frequency [45]. More generally, it is possible to organize the population responses along *ordered axes* of any parameter derived from the receptive field or response sensitivities of each neuron. Examples include distributions along a "rate axis" that reflects the dynamics of each neuron, a "scale axis" that encodes spectral selectivity (or bandwidth) [46] or an "AM or FM" index that captures the best modulation rate sensitivity [47]. This approach can provide a more complete understanding of the population code, but it is still limited, particularly in systems where neurons are characterized by several different tuning properties that vary randomly from neuron to neuron.

A different approach to tackle this question is to reconstruct the stimulus from the response of the neural population. This method of *reverse reconstruction* [48] finds the best approximation of the input stimulus, which can then be compared to the original to discover which features are preserved or enhanced and to assess the accuracy of their encoding. This method was developed for studies of the fly visual system [48] [49] [50] but has since been employed successfully in measuring the encoding of the fine temporal structure of visual stimuli by motion–sensitive neurons in macaque MT area [51], quantifying the effect of natural sound structure in neural coding in frog auditory nerve [52] and reconstruction of a visual stimuli by the collective activity of many retinal ganglion cells [53] and LGN [54].

An important issue in reconstructing natural and complex stimuli is the presence of strong statistical regularities in the stimulus. Knowledge of these correlations allows for the application of priors to the reconstruction procedure that can substantially benefit reconstruction in noise or when sampling of the neural population is limited. Although reconstruction may be possible with a relatively small neural population by taking advantage of these priors, it is unclear whether actual neural systems benefit from the same prior information in decoding sensory representations.

Another issue with stimulus reconstruction is effect of neuronal plasticity on reconstruction. If the response properties of a neuron change, then its contribution to an optimal stimulus reconstruction should also change. However, allowing a neuron's

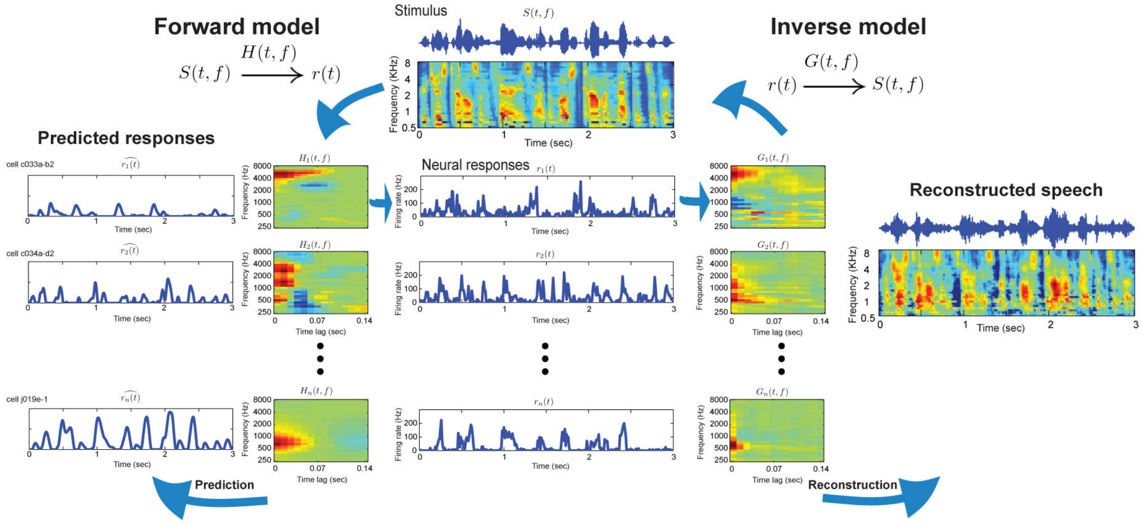

Figure 26. The forward and inverse models. The forward model (left columns) is the mapping from the acoustic spectrogram to the neural response found using reverse correlation technique (see Methods). The forward model describes neural response properties with the STRF, which can predict the response to a novel stimulus, as displayed in the left panels for three neurons. The inverse model of a neuron (right column) is the mapping from a population of neuronal responses back to the sound spectrogram. Using the inverse model, one can reconstruct the spectrogram of a sound from its population responses (right panel).

role in the reconstruction procedure to change implies that the downstream system that interprets the activity of that neuron also changes to match the plasticity. We report here on how we applied this computational framework and the associated experimental procedures to study the representation of complex auditory noise and natural speech stimuli [24] in primary auditory cortex (A1) of the ferret. Specifically, we assessed the dynamics and spectral selectivity of a large set of A1 neurons as they responded to these stimuli. We also explored the influence of prior knowledge of stimulus statistics on the reconstruction accuracy, both in neural data and in simulation. Finally, we explored how such reconstructions can be utilized to discover and interpret response modulations and receptive field plasticity induced during behavior [55].

5.2 Methods

The protocol for all surgical and experimental procedures was approved by the IACUC at the University of Maryland and consistent with NIH Guidelines.

5.2.1 Surgery

Four young adult, female ferrets were used in the neurophysiological recordings reported here. To secure stability of the recordings, a stainless steel head post was surgically implanted on the skull. During implant surgery, the ferrets were anesthetized with Nembutal (40 mg/kg) and Halothane (1-2%). Using sterile procedures, the skull was exposed and a headpost was mounted using bone cement, leaving clear access to primary auditory cortex in both hemispheres. Antibiotics and analgesics were administered as needed.

5.2.2 Neurophysiological recording

Experiments were conducted with awake head-restrained ferrets. The animals were habituated to this setup over a period of several weeks, and usually remained relaxed and relatively motionless throughout recording sessions that may last 2-4 hrs. Recordings were conducted in a double-walled acoustic chamber. Small craniotomies (~1-2 mm in diameter) were made over primary auditory cortex before recording sessions. Physiological recordings were made using tungsten microelectrodes (4-8 MΩ, FHC). Electrical signals were amplified and stored using an integrated data acquisition system (Alpha Omega). Spike sorting of the raw neural traces was done off-line using a custom PCA clustering algorithm. Our requirements for single unit isolation of stable waveforms included (1) that the waveform and spike rate remained stable throughout the recording, and (2) that the inter-spike interval for each neuron was distributed exponentially with a minimum latency of 1ms. The number of neurons used for each analysis varied. The analysis of spectro-temporally modulated noise used 256 neurons; the speech reconstruction analysis used 250 neurons; and the analysis of behavior-induced plasticity used 11 or 23 neurons.

5.2.3 Auditory stimuli and analysis

Experiments and simulations described in this report include spectro-temporally modulated noise and speech. The spectro-temporally modulated noise consisted of 30 Temporally Orthogonal Ripple Combinations (TORCs) [56]. Each TORC was a broadband noise with a dynamic spectral profile that was the superposition of the envelopes of six ripples (depicted in **Figure 27a**). A single ripple has a sinusoidal spectral profile, with peaks equally spaced at 0 (flat) to 1.4 peaks per octave; the envelope drifted temporally up or down the logarithmic frequency axis at a constant velocity of up to 48 Hz [56]. It was constructed by adding the envelopes of 6 ripples, where each ripple is a broadband noise with a sinusodially-modulated envelope along the frequency dimension (spectral density in cycles/octave), and spectral peaks that drift at a constant velocity along the time dimension (rate in Hz). All ripples in a TORC were of equal level and the same spectral density, but spanned a range of rates from -48 to 48 Hz (i.e., drifting up and down the frequency axis) with specially selected phases [56]. Therefore, the two-dimensional Fourier transform of each TORC envelope was a line that we refer to as a modulation spectrum (MS). We constructed 30 TORCs that spanned spectral modulations from 0 to 1.4 cyc/oct in steps of 0.2 cyc/oct (each consisting of ripples at 6 different rates from -48 to 48 Hz), and with two variants for each TORC with opposite polarities to minimize bias from the spike threshold in neural responses on measurements of spectro-temporal tuning, as detailed in [56]. To estimate the reconstruction error, we first subtracted the reconstructed TORC spectrograms from the original ones. The normalized error for different spectral and temporal modulations is then defined as the magnitude of the normalized 2-D Fourier transform of subtracted spectrograms (error spectrograms) at the corresponding modulation values.

Speech stimuli were phonetically transcribed continuous speech from the TIMIT database [24]. Thirty different sentences (3 seconds, 16 KHz sampling) spoken by different speakers (15 male and 15 female) were used to sample a variety of speakers and contexts. Each sentence was presented five times during recordings. To compute the average spectrogram representation of a given phoneme, the TIMIT phonetic transcriptions were used to align the auditory spectrograms of all the instances of that phoneme and then averaged across different exemplars as described in details in [57].

5.2.4 Reconstructing sound spectrograms using the inverse model

The *inverse* model is a linear mapping between the response of a population of neurons and the original stimulus [48] [54]. We represent the response of each neuron, *n*, as a function of time, $r_n(t)$. Because neurons in auditory cortex are not phase-locked to the modulations in the original sound pressure waveform, we represent the stimulus as its spectrogram $S(t, f)$, which has a more linear relationship to responses in A1 [58]. The inverse function, $G(t, f)$, is then defined as follows:

$$S(t, f) = \sum_{n}\sum_{\tau} G_n(t-\tau, f)\, r_n(\tau)$$

The function G is estimated by minimizing the mean-squared error between actual and reconstructed stimulus. Solving this analytically results in normalized reverse correlation [48] [54]:

$$\min\ e = \sum_{f}\sum_{t}\left(\hat{S}(t, f) - S(t, f)\right)^2 \ \rightarrow\ G = C_{rr}^{-1} C_{rs} \qquad (1)$$

5.2.5 Reconstruction of the sound using the forward model of auditory neurons

The *forward* model of a neuron maps the sound spectrogram to the neural response. We characterized each neuron by its spectro-temporal receptive field (STRF), estimated by normalized reverse correlation of the neuron's response to the auditory spectrogram of the speech stimulus. Although methods such as normalized reverse correlation can produce unbiased STRF estimates in theory, practical implementation require some form of regularization to prevent overfitting to noise along the low-variance dimensions [22]. This in effect imposes a smoothness constraint on the STRF. The regression parameters were adjusted using a jackknife validation set to maximize the correlation between actual and predicted responses. Having the STRFs of the neurons, we now describe the reconstruction using the forward models:

The STRF, $h(t, f)$, is a mapping from the sound spectrogram $s(t, f)$ to the neural response $r(t)$:

$$r(t) = \sum_f \int h(\tau, f) s(t-\tau, f) d\tau \tag{2}$$

It is not possible to invert this equation to find $s(t,f)$ from $r(t)$ because of the ambiguity of the frequency dimension (response is a function of time and not frequency). However, we can recover the frequency dimension provided we have enough neurons to construct an invertible system of linear equations (full coverage of the frequency space). To do so, we rewrite equation 1 in the matrix form:

$$r = hS$$

The STRFs are estimated using normalized reverse correlation:

$$h = C_{ss}^{-1} C_{sr} \tag{3}$$

Assuming we have the response of n neurons to the same sound, we construct the following system of linear equations:

$$\begin{matrix} r_1 = h_1 S \\ \vdots \\ r_n = h_n S \end{matrix} \quad \rightarrow \quad R = HS \tag{4}$$

Assuming the H matrix has a pseudo-inverse, which intuitively means the set of STRFs cover the whole frequency space, we invert equation 4 to find S:

$$S = (H^T H)^{-1} H^T R$$

5.2.6 The relation between forward and inverse model of reconstructions

We described two methods for reconstructing the input spectrograms from the neural responses. They are illustrated schematically in **Figure 26**. In the first method, we used the neuron's STRFs to construct a series of linear equations mapping the sound spectrogram to the neural responses and then solving for the input. In the second method, we directly estimate the mapping from neural responses to the spectrograms using the inverse model of neurons. The main difference between the two methods is inclusion and exclusion of known statistical structure of the input in the reconstruction. The first method does not use any prior knowledge about the statistics of the stimuli since they are taken out in the estimation of STRFs. The second method, by contrast, imposes that prior knowledge into the reconstruction. From equations 1 and 3, the forward model (H) and inverse model (G) are related by the following equation:

$$H = C_{ss}^{-1} C_{rr} \, G$$

5.2.7 Effect of STRF changes on reconstructions

During behavior, the functional relationship between stimulus and neural response can change to facilitate behavior [55], e.g., through top-down attentional influences that change the gain or shape of the receptive field H, e.g., during performance of a behavioral task. To understand how these changes affect reconstruction, consider the receptive field formulation:

$$S \, H = R$$

where S is the stimulus, H the neural receptive fields, and R is the population response. Because the model is linear, if there is a change in the receptive field, ΔH , it results in a change, ΔR , in the neural response:

$$S \, \Delta H = \Delta R$$

Rather than modeling the change in the receptive field, we assume that the system has not changed and instead find the effective stimulus change that produces the observed change in the neural response:

$$\Delta S \, H = \Delta R$$

The effective stimulus change (ΔS) can be found using the inverse reconstruction method:

$$\Delta S = G \Delta R\,,$$

where G is the inverse transformation of the system measured *before* the change. In effect, this approach enables us to project response changes (ΔR) to the stimulus domain where it is often more intuitive to interpret. Minor variations on this formulation include examining transformations of the reconstructed stimulus (such as its Fourier transform), or computing the stimulus change as $\Delta S = G' \Delta R$, where G' is the inverse transformation computed from the responses collected *after* the change in receptive fields.

5.3 Results

5.3.1 Reconstruction of spectro-temporally modulated noise spectrograms

To study the fidelity of auditory encoding by neurons in primary auditory cortex (AI), we reconstructed spectrograms of specially designed spectro-temporally modulated broadband noise stimuli ("temporally orthogonal ripple combinations," TORCs) from the responses of 256 AI neurons (see Methods and [56]). Each of the 30 TORCs contains a range of spectral and temporal modulations as illustrated in **Figure 27**. The fidelity of their reconstruction reveals the extent to which these modulations are preserved in cortical responses. Such modulations are the key carriers of information in complex signals such as speech and music. Hence it is important to determine if AI responses could encode them and whether the range encoded matches the observed perceptual capabilities of the ferret.

Figure 27b illustrates the result of reconstructing all 30 TORCs, which had with an average correlation of 0.5. The representation of the combined modulation spectrum of the TORCs (originally flat in the range +/- 48 Hz, and 0-1.4 cyc/octaves) had a bandpass structure that could be visualized more clearly when collapsed along its temporal and spectral dimensions (**Figure 27b,** right and top panels). The temporal modulation spectrum was most sensitive around 10-14Hz, while the spectral modulation spectrum showed a low pass characteristic. These neural spectral and temporal modulation spectra are consistent with the perceptual modulation sensitivity measured in ferrets [59].

5.3.2 Effect of number of neurons on reconstruction accuracy

Neurons in AI vary substantially in their spectro-temporal tuning properties, each of which matches only a narrow range of the spectro-temporal patterns in TORCs [60]. It was therefore expected that many neurons would be required to achieve a full coverage of the stimulus and that increasing the number of neurons used for reconstruction would improve its accuracy. To test this hypothesis, we varied the number of neurons used for reconstruction and measured the corresponding normalized reconstruction error (see Methods). For each number of neurons used in reconstruction, *N*, we selected 10 random subsets and averaged the reconstructed spectrograms across the subsets.

Figure 27c shows the normalized error for different temporal modulation ranges as a function of *N*. As *N* increases from very small values, the reconstruction error improves for all temporal modulation ranges. However, the error for slower

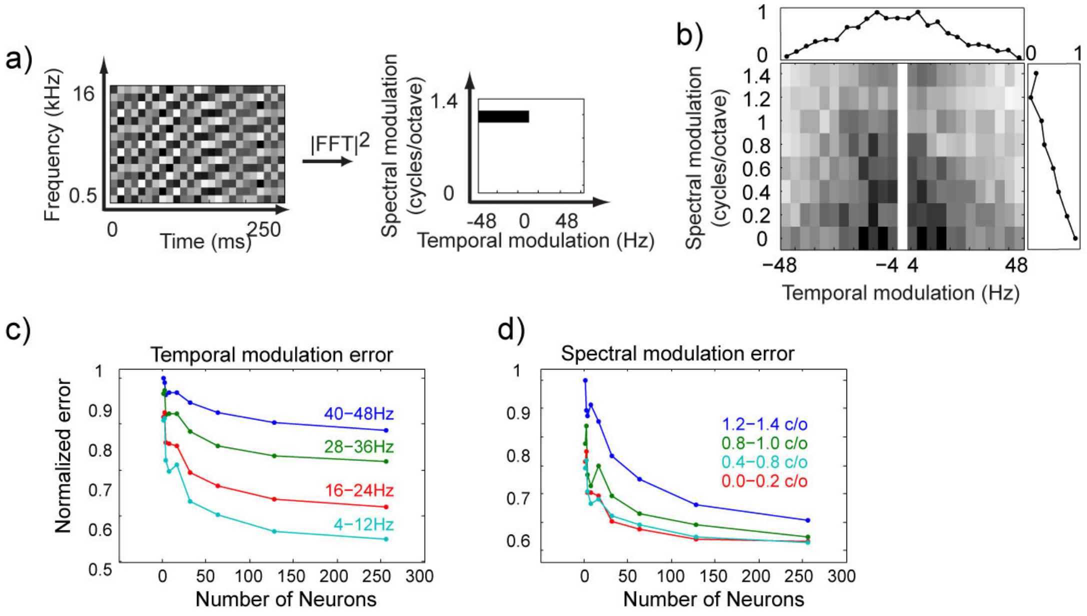

Figure 27. Reconstruction of broadband noise stimuli (TORC). Temporally Orthogonal Ripple Combinations (TORCs) are specially constructed broadband modulated noise stimuli. In each of 30 different TORCs, the stimulus consists of the superposition of the envelopes of six ripples, all at the same ripple density (0.2, 0.4 … 1.2 or 1.4 cyc/octave) and ripple velocities from 4 to 48 Hz and -4 to -48 Hz. Therefore **(a)** the two dimensional power spectrum of each TORC (modulation spectrum, MS) is a line at a particular spectral modulation density. **(b)** The average MS of the TORCs reconstructed from all neurons appears bandpass compared to the original flat MS, illustrated in the collapsed MS along temporal and spectral modulations (right and top, respectively). **(c)** Normalized reconstruction error (see Methods) vs. number of neurons for different temporal modulation ranges. The reconstruction error converges to lower values for lower temporal modulations. **(d)** Normalized reconstruction error for different spectral modulations. Higher scales show larger errors.

modulation rates (4-12Hz) remains consistently lower than for faster modulation rates (40-48Hz). Based on analysis, the reconstruction error converges to lower values for 4-12 Hz modulations and to higher for 40-48 Hz.

Figure 27d shows the normalized error for different spectral modulation ranges as a function of *N*. As for temporal modulation, we observe and improvement in error with increasing *N*. However, the difference low and high spectral modulations is less extreme. For 0-0.2 cycle/octave modulations, the reconstruction error converges to lower values than the error for 1.2-1.4 cycle/octave modulations.

A possible reason for the lower bounds is the loss of information at higher rates and scale as cortical neurons fail to phase-lock to these faster modulations. Another possible factor is that increasing the number of neurons adds to the number of parameters that have to be estimated from limited data, and hence constrains the quality of the reconstruction.

5.3.3 Comparison of forward and inverse models for speech reconstruction

Because TORCs are designed to have minimal spectro-temporal correlations, reconstructing TORC spectrograms cannot take advantage of prior knowledge of stimulus statistics to improve reconstruction accuracy. However, for natural stimuli such as continuous speech, which do contain strong correlations, this information can be used to infer features in the stimulus that are not explicitly coded in the neural responses. Regions of the stimulus spectrogram that do not excite a neuron directly can still be recovered if they are highly correlated with another feature of the stimulus that is encoded by the neuron. As defined in the Methods, the *inverse* model for reconstruction utilizes information about stimulus correlations, and the *forward* model does not.

We examined how the inverse model benefited from prior knowledge of input correlations to reconstruct continuous speech spectrograms from simulated responses of a sparse sampling of neurons. We simulated the responses of 8 neurons that were spectrally narrowly tuned and were spaced in such a way as to leave parts of the spectrogram unseen by the population. We then reconstructed the spectrogram of one speech sample using the two methods as illustrated in **Figure 28b, c**. The forward model (**Figure 28b**) resulted in a sparse reconstruction with no data in the unseen parts of the spectrogram. In this case, the correlation of the reconstructed and original spectrograms was 0.70. Using the inverse model (**Figure 28c**), however, improved the reconstruction accuracy to 0.83 and resulted in less sparse reconstruction because speech is a highly correlated signal.

To contrast the forward and inverse models using experimental data, we reconstructed the spectrograms of speech from actual responses of 250 AI neurons (all the neurons were presented the same stimuli, but they were not recorded simultaneously). For each neuron, responses to speech stimuli over 90 seconds were collected and used to estimate the forward and inverse models (H and G in Figure 26). **Figure 28** shows an example of one speech sentence (**Figure 28a**) and its reconstructions using the forward (**Figure 28e**) and inverse (**Figure 28f**) models. Using exactly the same data, the reconstruction by the inverse model was superior to that of the forward model, as judged by its higher correlation with the original speech

(0.78 versus 0.42), demonstrating the benefit of prior knowledge in the

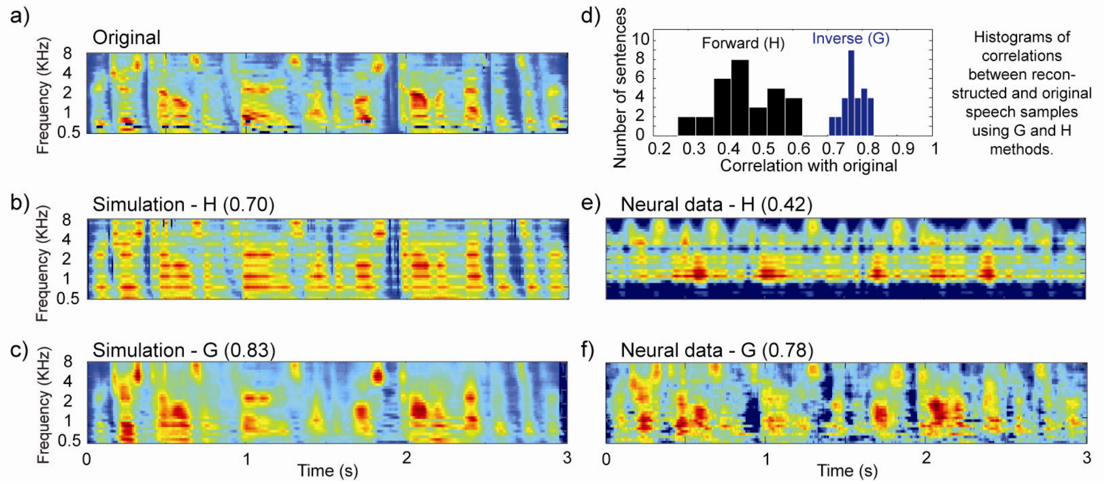

Figure 28. Comparison of the forward and inverse reconstruction methods. Reconstruction of a speech spectrogram using forward and inverse models in simulations and with physiological data. **(a)** Spectrogram of a speech sample. **(b)** Forward reconstruction of the speech spectrogram from simulated responses of 8 neurons that were spectrally narrowly tuned and spaced in such a way as to leave parts of the spectrogram unseen by the population, resulting in a sparse spectrogram. (**c**) Reconstruction using inverse model from same neurons in (b) that results in a better correlation with the original spectrogram. (**d**) Comparison of the reconstruction accuracy of the inverse and forward models for 30 different sentences. The distribution of the correlation coefficients for the forward model was considerably lower than that for the inverse model. (**e**) Forward reconstruction from physiological responses to 250 neurons. (**f**) Inverse reconstruction of the speech produces a better match to the original spectrogram.

reconstruction. Similar results are summarized in **Figure 28d** which depicts the correlation coefficients between original and reconstructed speech for 30 additional sentences using both methods.

To compare the perceptibility of these reconstructions, we inverted the spectrograms to generate the best approximation of corresponding acoustic signals using a convex projection method described in (61). Audio examples of reconstructions using the inverse model were noticeably more intelligible than those from the forward model.

5.3.4 Reconstructed phonemes from neural population responses

How and to what extent do responses of cortical neurons in the ferret encode phonemes with enough fidelity to account for their perception in humans? This question implicitly tests the hypothesis that auditory processing mechanisms up to the level of the primary auditory cortex, and that are common across other mammals like ferrets, are sufficient to account for the robust perception of speech [15] [2]. Previous analyses of A1 responses in mammals have been consistent with this point of view [46] [62]. Here we took a different approach to shed more light on this issue and examine in particular the pattern of errors observed in the perception of various phonemes.

We first analyzed the encoding of the *average* features of each phoneme in the population response **Figure 29(a-d)** (top rows) illustrates the average phoneme spectrograms of four groups of phonemes (Plosives, Fricatives, Nasals, and Vowels) extracted from using the methods detailed in [46]. The corresponding panels in the bottom rows of **Figure 29(a-d)** depict the average spectrograms of the same phonemes but from reconstructions using the inverse model. The strong similarity between the two sets of spectrograms (average correlation coefficient of 0.88) indicates that average responses of AI neurons have the dynamics and spectral selectivity to capture linearly most details of the average spectrotemporal features of phonemes.

Comparing only the average phoneme spectrograms may improve apparent performance by averaging out differences between phoneme exemplars. To make a more critical assessment of the results, we examined the accuracy of reconstructions for each phoneme exemplar separately. **Figure 29e** plots the average of such correlations across all instances of each phoneme. Some phonemes, such as the high frequency fricatives (s, ʃ) [3], display excellent reconstruction accuracy even at the level of individual exemplars. Most plosives (p, b, t, d, k, g) were encoded with an intermediate level of accuracy (average correlation 0.6). Nasals (m, n, ŋ) were the least accurately decoded perhaps due to their weak acoustic energy.

The sometimes low and variable accuracy of the reconstructions for individual phoneme exemplars stands in striking contrast to the highly accurate encoding of the average features **Figure 29**(**a-d**). Averaging of spectrograms across all instances of a phoneme preserved only features that were common across all syllabic contexts and hence not affected by co-articulatory factors. These common features were generic enough to be captured well by the linear spectro-temporal response models in AI. By contrast, the unique features of individual phoneme samples were sometimes not well described by the reconstruction (as with the Nasals in **Figure 29e**).

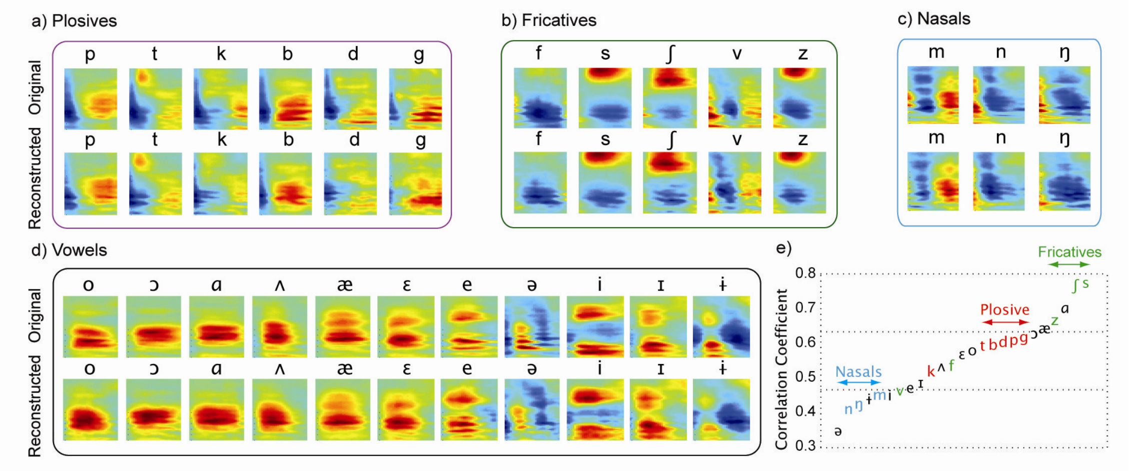

Figure 29. Average phoneme spectrograms from original and reconstructed phonemes (a-d) (Top panels) The average phoneme spectrograms of four groups of phonemes (plosives, fricatives, nasals, and vowels). (Bottom panels) The corresponding panels depicting the average phoneme spectrograms from reconstructed spectra using the inverse model. The original and reconstructed spectrograms are quite similar and have an average correlation coefficient of 0.88. (**e**) The correspondence between reconstructions and actual spectrograms for each phoneme exemplar, averaged across all instances of each phoneme.

5.3.5 Stimulus reconstruction and neural plasticity

Receptive field properties in AI can rapidly change during task performance in accordance with specific task demands and salient sensory cues [55]. Such plasticity may reflect attentional demands, task difficulty and performance. One approach for inferring the functional significance of these changes is to examine the receptive fields of single neurons and to extrapolate the consequences of their changes on neural encoding [59]. In tasks requiring simple discrimination (e.g., tone versus noise), changes in spectro-temporal tuning have been shown to enhance overall cortical responsiveness to a foreground (or target) sound while suppressing the background (or reference) sound, presumably increasing the likelihood of detecting the attended target.

However, as the complexity of auditory tasks increases (e.g. when discriminating among phonemes, tonal sequences, or musical timbres), receptive field changes are likely to become more elaborate and hence more challenging to relate to the acoustical properties of the stimuli. Another limitation of the traditional approach of examining the plasticity of neurons in isolation is that it does not benefit from multielectrode recordings. The complexity of the analysis increases proportionately with the number of electrodes, but the same amount of data is required from each neuron in order to identify significant changes in tuning. Both of these challenges can be addressed by a reformulation of the problem to make use of the inverse model to reconstruct the stimuli.

We demonstrate this approach using both simulated and actual changes that were recorded in behavioral physiology experiments. TORC responses were used to measure the spectro-temporal receptive fields (STRFs) and hence the adaptive changes that they exhibit. Therefore, all reconstructions considered here were those of the TORC stimuli *before* and *during* the STRF changes.

A simulation of the effects of STRF plasticity is shown in **Figure 30a.** The simulated STRF changes were based on observations in tone detection experiments reported in [55]. Here the STRFs constituted a bank of tonotopically distributed filters, illustrated by the three STRFs on the left centered at the 7th, 9th, and 11th channels (**Figure 30a**, **left column**). During behavior, a target tone was played at the center frequency of the 8th channel, which caused the nearby STRFs (at the 7th and 9th channels) to expand towards the target, or become more sensitive to it (**Figure 30a**, **right column**), inducing channel 8 responses to become more correlated with its neighboring channels. **Figure 30b** illustrates the reconstructed TORC stimuli *before* the task (top panel) and *during* it (middle panel). The changes were best seen in the *difference* (ΔS) between two reconstructions (bottom panel). When we measured the cross-correlation between each pair of channels in the reconstructed stimulus (**Figure 30c**), we found significant non-zero values at off-diagonal entries between the 8th channel and each of the 7th and 9th channels emerging only *during* the task when the STRF was changing.

To illustrate how this method can be applied to real physiological data, we analyzed the data previously published in [55] that showed STRF changes similar to the simulations above. **Figure 31a** illustrates the difference between the reconstructed TORC spectrograms from responses prior to the behavior (passive) and during the behavior (active) for data collected at two different target frequencies, channel 13 (top panel, 11 neurons) and 12 (bottom panel, 23 neurons). The difference between passive and active TORC reconstructions showed a noticeable change at the frequency of the target in both data sets. This difference is more directly indicated after the collapsing over time (**Figure 31b**) and in the matrices of correlations between spectral channels (**Figure 31c**), which showed an enhancement at the target frequency. In both cases the target channel also showed a weak negative correlation with adjacent frequencies (light blue regions in **Figure 31c**), consistent with a decrease in response to frequencies close to the target [55].

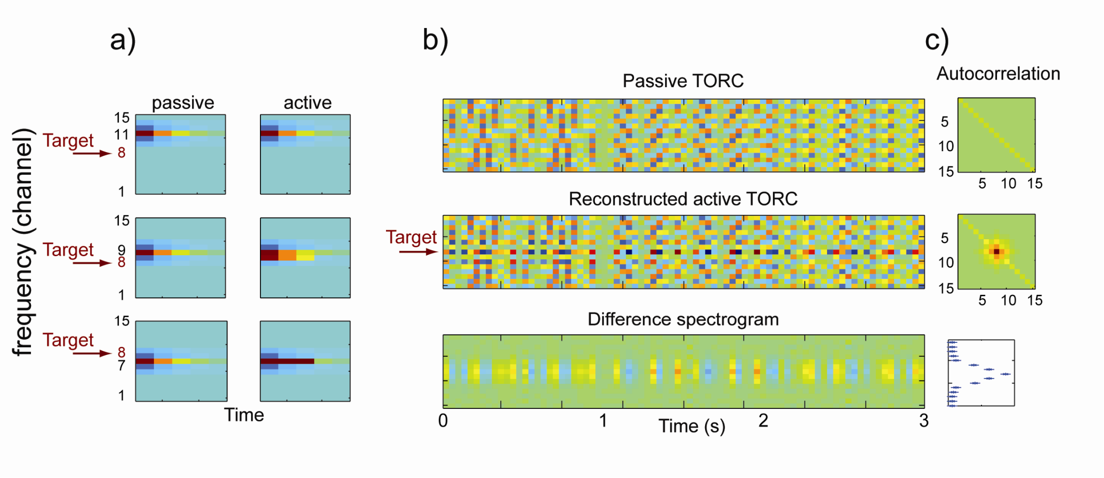

Figure 30. Detecting STRF changes in simulated data. (a) Simulated STRFs displaying changes that might occur during behavioral experiments. **Left column** ("passive")**:** Before the behavioral experiment, three STRFs are tuned to different frequencies centered at the 7th, 9th, and 11th channels. The target tone during the experiment is at the frequency of the 8th channel. **Right column** ("active")**:** During behavior, the STRF closest to the target tone becomes more sensitive to the target tone frequency by broadening its excitatory field toward the target tone frequency. **(b)** Reconstructed TORCs using passive (top panel) and active (middle panel) responses. The change in the STRF at the 8th channel causes the TORC reconstruction to change locally, which can be detected by subtracting the two TORC spectrograms (bottom panel). **(c):** Detecting and quantifying the changes through cross-correlation of spectral channels (top and middle panels)

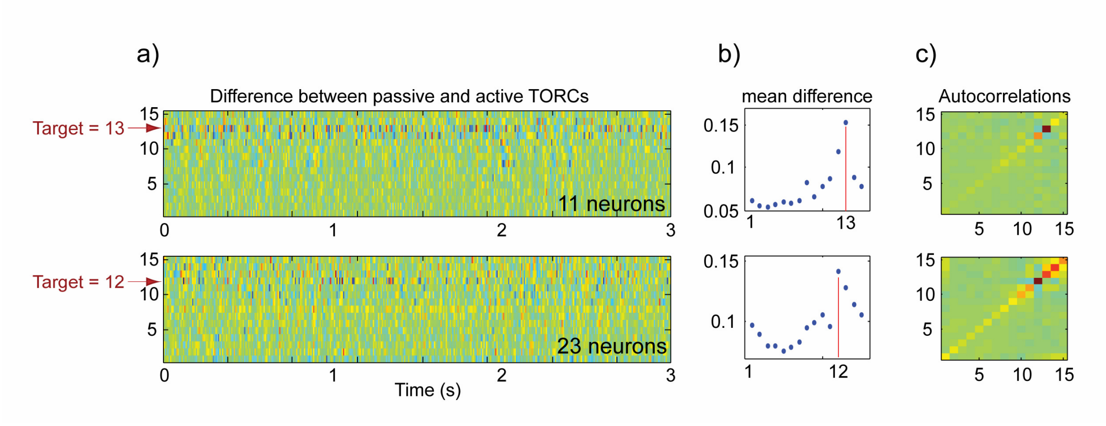

Figure 31. Detecting STRF changes in actual physiological data. Difference between reconstructed TORCs before and during a tone detection task. Top panel shows the difference for data from 23 neurons when the target tone was in channel 12. Bottom panel shows the difference for reconstruction from 11 neurons when the target tone was in channel 13 **(b)** In each case, averaging the mean difference between the reconstructed spectrograms over time shows a peak at the target frequency. **(c)** The matrix of correlations between spectral channels also signals the change at the target tones.

5.4 Cortical representation of speech in white noise

It is well known that humans can robustly perceive phonemes despite of the variability across speakers, context, and natural distortions like noise and reverberation. This robustness is attributed to a rich and invariant representation of perceptually important features of speech. Here, we use the method of reconstruction described in this section to study the issue of noise robust representation of speech in the primary auditory cortex. To investigate this issue, we recorded the responses of 100 neurons in the primary auditory cortex to clean, 6dB and 0dB Signal To Noise (SNR) speech in white noise. The method of reconstruction then is used to go back to the spectrogram representation and to compare the reconstruction with the noisy one.

The speech samples used in the noise study were the same as the previous sections,: 90 seconds speech from Timit [24], half male and half female speakers. The only difference was that in two additional conditions, white Gaussian noise was added to the clean speech at two SNR levels: 0 and 6dB.

Figure 32 shows the spectrogram of one such sentence in clean and white noise (**Figure 32**, left column). The effect of white noise is more evident in higher frequencies because of the increasing bandwidth of the filters. We used the responses of the neurons to clean speech to estimate the inverse transformation (G functions) for the population. This G then was applied to the responses of neurons to noisy speech to reconstruct back the speech spectrograms. The result of the reconstruction can be seen in **Figure 32** right column. As one can see, the noise distortion is reduced in the reconstructed signal compare to the original noisy spectrograms.

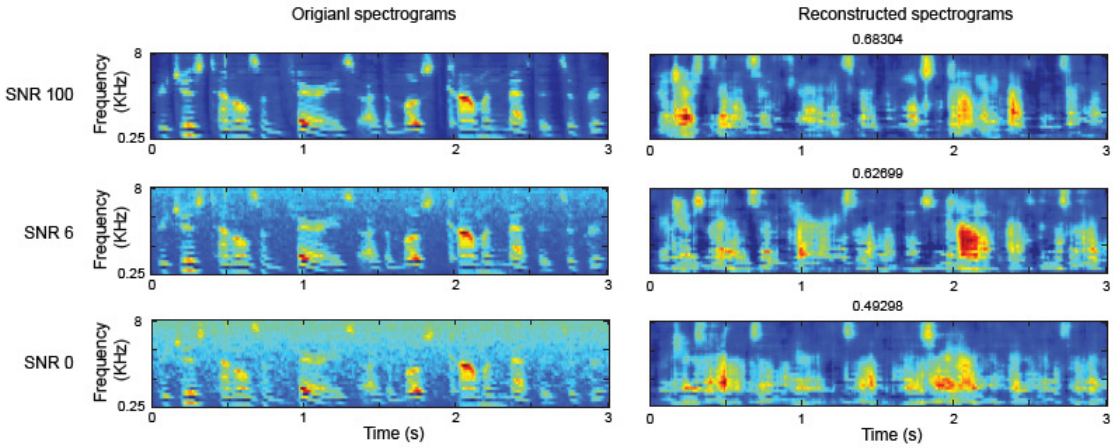

Figure 32. Original and reconstructed spectrograms of speech in clean and noise. Left column shows the spectrogram of a speech sample in clean (top row), 6dB (middle row) and 0dB (bottom row). The right column shows the reconstructed spectrograms from the neural responses to clean and noisy speech. The reconstructed noisy spectrograms show more similarity to the clean than is the case in the noisy original spectrograms.

To quantify this effect, we compared the correlation coefficient between the actual and reconstructed spectrograms for speech at various noise levels as shown in Figure 33. This figure shows that the reconstructed speech spectrograms are always more similar to the clean speech than the noisy ones, even when the reconstruction is from the noisy speech responses. This can be the result of the lack of the representation of noise in the neural responses that causes a poor correlation between reconstruction from noisy speech and actual noisy spectrograms (blue bars in **Figure 33**).

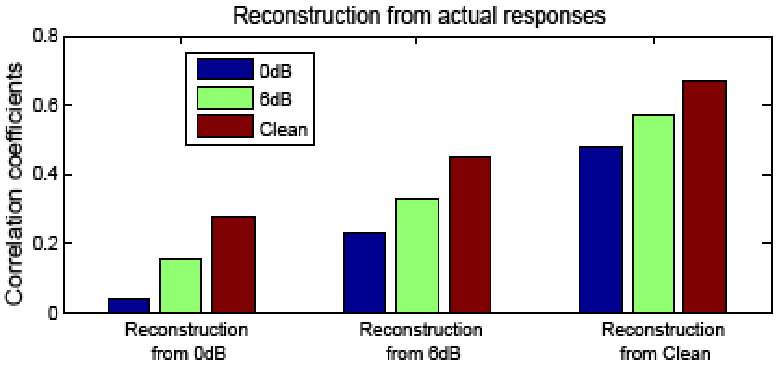

Figure 33. Correlation coefficients between original and reconstructed spectrograms from actual neural responses. Reconstructed speech always has a higher correlation with clean speech (red bars), even if the responses are from noisy speech.

To investigate whether this noise robustness observed in the neural responses can be explained by linear receptive model, we used the predicted responses of neurons to noisy speech obtained from the STRFs. When we reconstruct the spectrograms using the predicted responses to noisy and clean speech, a different pattern emerges that was different from the actual data. **Figure 34** shows the same correlation analysis in **Figure 33** but for the predicted responses. In this case, the correlation coefficient for each case is the largest between reconstructed spectrogram and corresponding original spectrogram (blue bar for 0dB, green bar for 6dB and red bar for clean). This shows the deficiency of the STRF model to capture the noise robustness of the responses, which can be due to a non-linear effect such as adaptation to background noise.

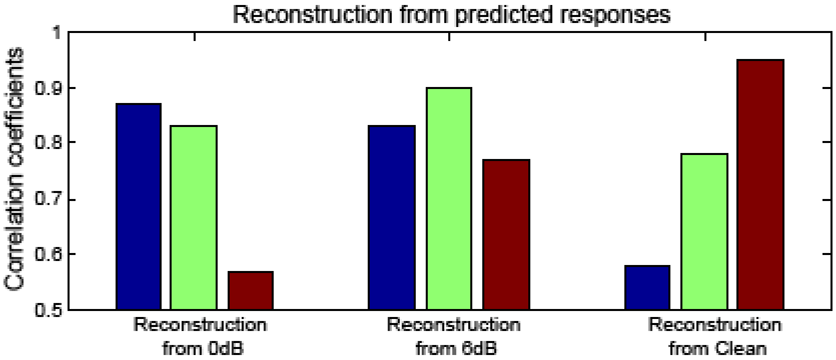

Figure 34. **Correlation coefficients between original and reconstructed spectrograms from predicted neural responses.** Reconstructed speech has a higher correlation with when its from the responses to the same noise level. This means a linear STRF model can not predict the noise robustness observed in the data.

Finally, to investigate the effect of noise on phoneme representation, we constructed the average phoneme responses from the original and reconstructed clean and noisy speech. The reconstructions are shown in **Figure 35** for vowels and **Figure 36** for the consonants. To exemplify the noise robustness of the neural representation, we consider the vowel /I/. The second formant for this vowel (red box in **Figure 35**) is masked by white noise specially at 0dB (green box in **Figure 35**). However, the representation of this formant is more evident in reconstructed spectrogram from 0dB responses (blue box in **Figure 35**). This pattern also is evident in the encoding of the consonants as shown for /t/ and /s/ in **Figure 36**. Again, the reconstructed noisy spectrograms show a stronger representation of the high frequency features of the consonants /t/ and /s/ (**Figure 36** black box) compared to the actual noisy spectrogram (**Figure 36** blue box).

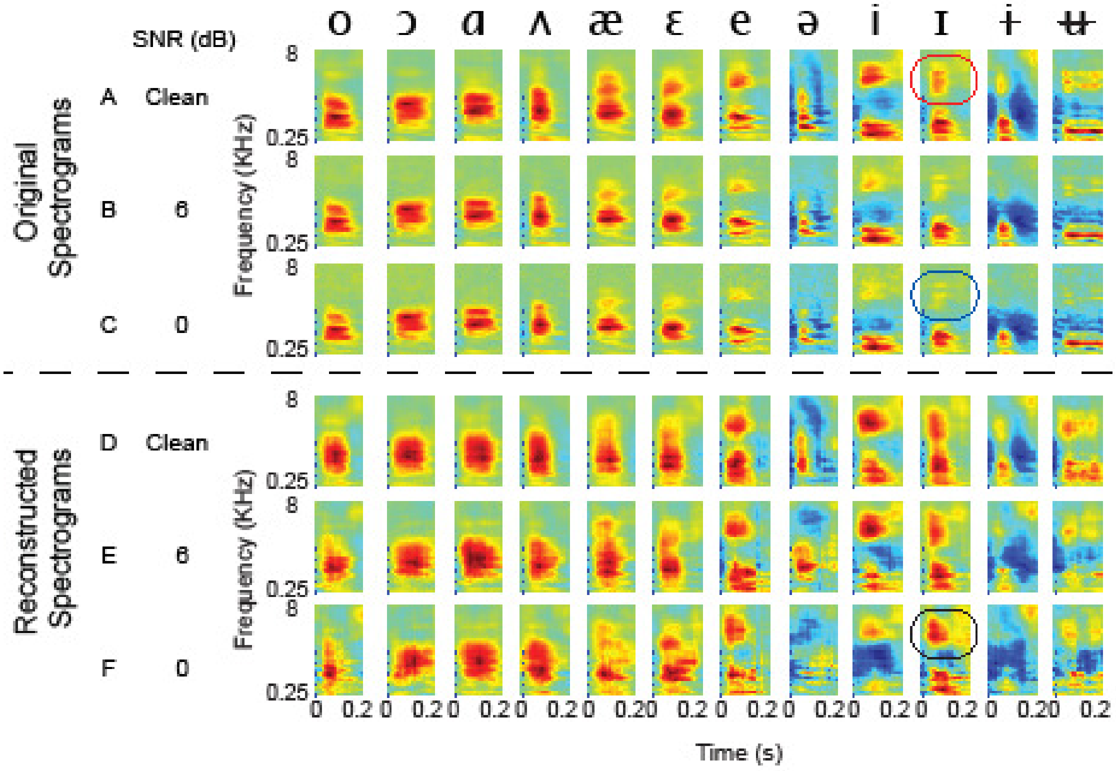

Figure 35. Average vowel representation obtained from original and reconstructed spectrograms of speech at different noise level. Top three rows are the average vowel spectrograms from the original clean and noisy speech, the bottom three rows are from reconstructed spectrograms. The highlighted boxes show how the second formant of vowel /I/ begins to disappear in the original noisy spectrograms but not in the reconstructed one.

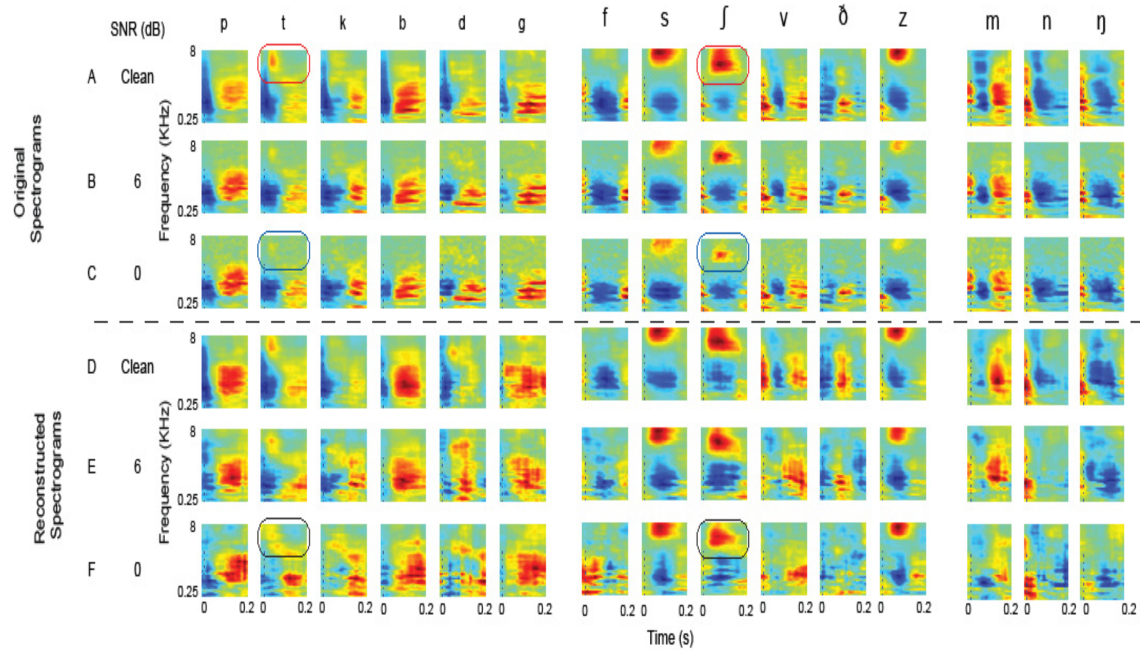

Figure 36. Average consonant representation obtained from original and reconstructed spectrograms of speech at different noise level. Top three rows are the average vowel spectrograms from the original clean and noisy speech, the bottom three rows are from reconstructed spectrograms. The highlighted boxes show how features of plosive /t/ and fricative /s/ becomes distorted in the noisy spectrograms (third row), but not as badly in the reconstructed speech from noisy responses (bottom row).

5.4.1 Phoneme discriminability in the original and reconstructed spectrograms

To get an estimate of phoneme discriminability, we measured the ratio of within-class to between class variability in the representation of phonemes. The higher the ratio is, the easier it is to separate different categories. The within and between class variability are defined as follow:

$$Sw=\frac{1}{n}\sum_{j=1}^{k}\sum_{x\in X_j}(x-c^{(j)})(x-c^{(j)})^T$$

$$S_b = \frac{1}{n}\sum_{j=1} n_j (c^{(j)} - c)(c^{(j)} - c)^T$$

Where $c^{(j)}$ is the centroid of the j_{th} class and c is the global centroid. By definition, $trace(S_w)$ measures the within-class cohesion, while $trace(S_b)$ measures the between-class separation. A good representation of phonemes is the one that keeps a high between-class separating, and at the same time, minimizes the within-class cohesion resulting in a larger separability: $\rho = \frac{trace(S_b)}{trace(S_w)}$. Here we compare this ratio for original and reconstructed spectrograms at different *SNR* levels. **Figure 37** shows this ratio for 0dB, 6dB and clean speech. The ratio from original spectrograms is shown in red bars while the reconstruction is in blue. The separability in clean is almost the same for both original and reconstruction, however, in the degraded speech, the separability drops much faster for the original representation than the reconstructed (53% compared to 26%). Thus, the separation between the representations of different phonemes is more preserved in the cortical responses than is the case in the spectrogram.

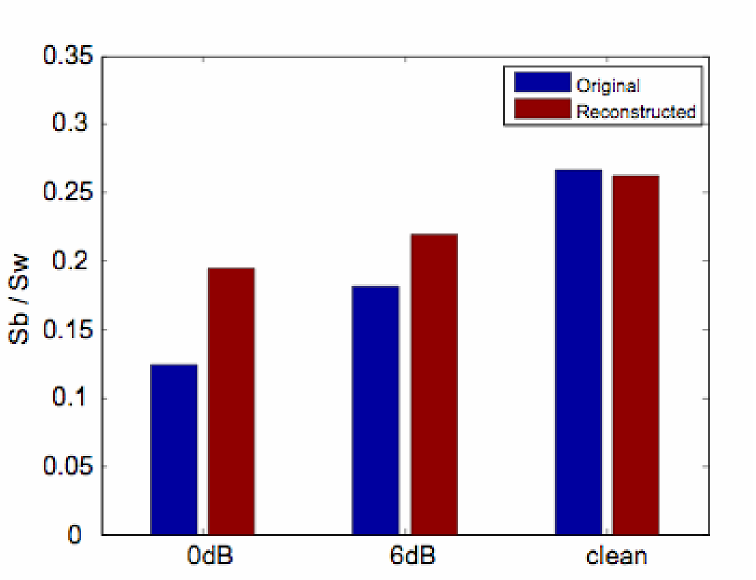

Figure 37. Fisher discriminability between different phoneme categories. In clean, the original and reconstructed spectrograms show a similar discriminability, however, when noise is added the discriminability drops much faster for the original spectrograms compare to the reconstructed.

5.5 Discussion

We have illustrated the advantages and limitations of reconstructing stimulus spectrograms from the responses of populations of neurons in primary auditory cortex (AI). The inverse model, used by the method of reverse reconstruction, was contrasted with the forward model for estimating the STRFs and using them to reconstruct the stimuli. The basic difference between the two methods is the former's utilization of the stimulus correlations in the reconstruction. In natural stimuli, including speech and music, significant correlations exist across a wide range of frequencies. Consequently, reconstructions with the inverse model making use of these priors generate far cleaner reconstructions than is possible with the forward model.

5.5.1 Characteristics of the A1 population code

Beyond their overall fidelity, stimulus reconstructions can indicate the limits, tuning, and many other properties of the cortical population response, as well as of the stimulus features preserved in them. By judicious choice of stimuli, one can interrogate the ability of cortical cells to encode their parameters. The example we presented of the encoding of TORC spectral and temporal modulations is but one that is appropriate in primary auditory cortical cells that tend to phase-lock well over a range of rates and densities. This approach can also be beneficial in pre-cortical areas that follow stimulus modulations to higher rates.

This same approach can generalize to areas outside of the auditory system. Rather than reconstructing the stimulus spectrogram, one can parameterize the stimulus in terms of other features that are correlated with neuronal responses. Such an approach is similar to methods for decoding movements from the population response in the motor system. In the visual system, reconstruction methods could be used to measure coding of stimulus orientation, spatial frequency and phase [63]. In more central areas, this approach could be used to measure information about abstract and learned stimulus features [64].

5.5.2 Encoding of Complex Features as in Speech

An appealing aspect of the inverse reconstruction method is the mapping of potentially complex acoustic features in the neural response back to the stimulus space, where they can be displayed intuitively. Speech is a prime example where much has been learned over the decades about its acoustic features almost exclusively in the spectrogram domain [2]. It is of course possible to explore the encoding of plosive bursts and voice-onset-times [60] but this exercise requires manually identifying features and provides less general insight than the reconstructed spectrograms, where these features were defined in the first place. The disparity between the inverse and forward approaches becomes bigger as more cells are recorded and as STRFs exhibit more complex shapes that cannot be reduced to simple orderly mappings [57].

5.5.3 Interpreting Adaptive STRFs

A potentially exciting deployment of the “inverse” method is in detecting and interpreting adaptive responses as illustrated with the simple example in **Figure 30**. Changes in the STRFs at the frequency of the target tone can be revealed by the trace they induce in the difference between original and reconstructed spectrograms (**Figure 30**b and **Figure 31**a). One can extrapolate from this simple example to far more intricate situations where top-down influences such as attention, expectations, or memory can substantially modify receptive field responses and shapes. For instance, changes induced in detecting an amplitude-modulated target tone or a phoneme modified on a spectral or temporal dimension might span many frequencies and temporal parameters. These changes would be expressed differently in each STRF depending on its BF and initial spectrotemporal properties. In such a case, simultaneous recordings from large assemblies of primary cortical neurons may facilitate reconstruction of the “adapted” stimulus, and hence reveal in the spectrogram the features directly targeted by the top-down influences. This possibility suggests a means to interpret large multiunit recording methods in behavioral physiology in the future.

Chapter 6

Applications of spectrotemporal modulations in speech signal processing

We have shown so far, how the brain performs a multiresolution mapping to represent speech in a high dimensional space where different categories fall in separate locations in this space. Here, we use a model of this transformation in two different speech processing tasks. In the first application, we show how speech has a very different characteristics in the cortical representation and we use this fact to discriminate between speech and other sounds. In the second task, we use this separation to part speech from noise and subsequently achieve noise suppression.

6.1 A computational model for spectrotemporal features

The computational auditory model is based on neurophysiological, biophysical, and psychoacoustical investigations at various stages of the auditory system [65]–[11]. It consists of two basic stages. An early stage models the transformation of the acoustic signal into an internal neural representation referred to as an auditory spectrogram. A central stage analyzes the spectrogram to estimate the content of its spectral and temporal modulations using a bank of modulation-selective filters mimicking those described in a model of the mammalian primary auditory cortex [65]. This stage is responsible for extracting the key features upon which the classification is based.

6.1.1 Early auditory system

The stages of the early auditory model are illustrated in **Figure 38**. The acoustic signal entering the ear produces a complex spatiotemporal pattern of vibrations along the basilar membrane of the cochlea. The maximal displacement at each cochlear

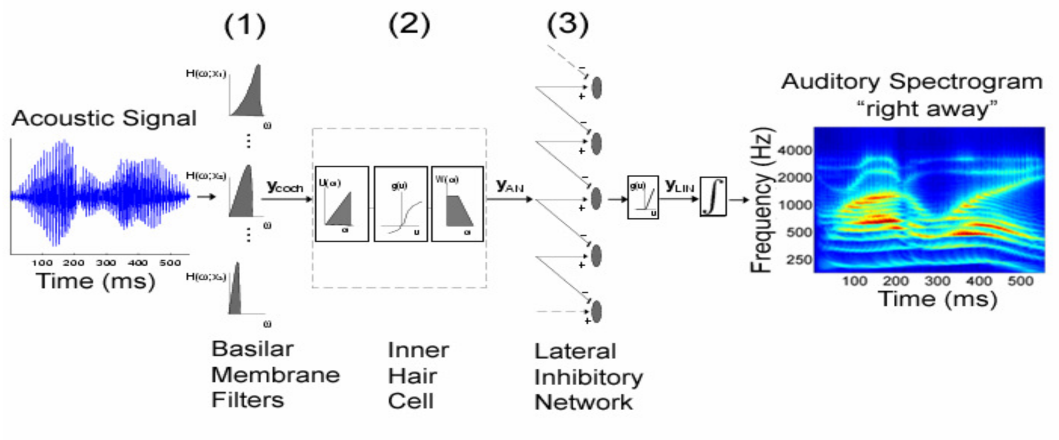

Figure 38. Schematic of the early stages of auditory processing. *(1)* Sound is analyzed by a model of the cochlea consisting of a bank of 128 constant-Q bandpass filters with center frequencies equally spaced on a logarithmic frequency axis (tonotopic axis). *(2)* Each filter output is then transduced into auditory-nerve patterns by a hair cell stage which is modeled as a 3-step operation: a highpass filter (the fluid-cilia coupling), followed by an instantaneous nonlinear compression (gated ionic channels) and then a lowpass filter (hair cell membrane leakage). *(3)* Finally, a lateral inhibitory network detects discontinuities in the responses across the tonotopic axis of the auditory nerve array by a first-order derivative with respect to the tonotopic axis and followed by a half-wave rectification. The final output of this stage (auditory spectrogram) is obtained by integrating *YLIN* over a short window, mimicking the further loss of phase-locking observed in the midbrainphase-locking observed in the midbrain.

point corresponds to a distinct tone frequency in the stimulus, creating a tonotopically-ordered response axis along the length of the cochlea. Thus, the basilar membrane can be thought of as a bank of constant- highly asymmetric bandpass filters (Q = 4) equally spaced on a logarithmic frequency axis. In brief, this operation is an affine wavelet transform of the acoustic signal. This analysis stage is implemented by a bank of 128 overlapping constant-Q (QERB = 5.88) bandpass filters with center frequencies (CF) that are uniformly distributed along a logarithmic frequency axis (f), over 5.3 octaves (24 filters/octave). The impulse response of each filter is denoted by *hcochlea* (*t*; *f*) . The cochlear filter outputs are then transduced into auditory- nerve patterns *yan*(*t, f*) by a hair cell stage which converted cochlear outputs into inner hair cell intracellular potentials. This process is modeled as three-step operation: a highpass filter (the fluid-cilia coupling), followed by an instantaneous nonlinear compression (gated ionic channels) $g_{hc}(.)$, and then a lowpass filter (hair cell membrane leakage) $\mu_{hc}(t)$. Finally, a lateral inhibitory network (LIN) detects discontinuities in the responses across the tonotopic axis of the auditory nerve array [66]. The LIN is simply approximated by a first-orderderivative with respect to the tonotopic axis and followed by a half-wave rectifier to produce $y_{LIN}(t,f)$. The final output of this stage is obtained by integrating $y_{LIN}(t,f)$ over a short window, $\mu_{midbrain}(t,\tau)$ with time constant $\tau = 8$ ms mimicking the further loss of phase-locking observed in the midbrain. This stage effectively sharpens the bandwidth of the cochlear filters from about Q = 4 to 12 [65]. The mathematical formulation for this stage can be summarized as follows:

$$\begin{aligned}
&y_{cochlea}(t,f) = s(t) * h_{cochlea}(t;f)\\
&y_{an}(t,f) = g_{hc}\left(\partial_t y_{cochlea}(t,f)\right) * \mu_{hc}(t)\\
&y_{LIN}(t,f) = \max(\partial_f y_{an}(t,f), 0)\\
&y(t,f) = y_{LIN}(t,f) * \mu_{midbrain}(t;\tau)
\end{aligned}$$

where $*$ denotes convolution in time. The above sequence of operations effectively computes a spectrogram of the speech signal (**Figure 38**, right) using a bank of constant-*Q* filters, with a bandwidth tuning *Q* of about 12 (or just under 10% of the center frequency of each filter). Dynamically, the spectrogram also encodes explicitly all temporal *envelope modulations* due to interactions between the spectral components that fall within the bandwidth of each filter. The frequencies of these modulations are naturally limited by the maximum bandwidth of the cochlear filters.

6.1.2 Central Auditory System

Higher central auditory stages (especially the primary auditory cortex) further analyze the auditory spectrum into more elaborate representations, interpret them, and separate the different cues and features associated with different sound percepts. Specifically, the auditory cortical model employed here is mathematically equivalent to a two-dimensional affine wavelet transform of the auditory spectrogram, with a spectro-temporal mother wavelet resembling a two-dimensional D spectro-temporal Gabor function. Computationally, this stage estimates the spectral and temporal modulation content of the auditory spectrogram via a bank of modulation-selective filters (the wavelets) centered at each frequency along the tonotopic axis. Each filter is tuned (Q = 1) to a range of temporal modulations, also referred to as rates or velocities (ω in hertz) and spectral modulations, also referred to as densities or scales (Ω in cycles/octave). A typical Gabor-like spectro-temporal impulse response or wavelet [usually called spectro-temporal response field (STRF)] is shown in **Figure 39**.

We assume a bank of directional selective STRFs (downward and upward) that are real functions formed by combining two complex functions of time and frequency.

This is consistent with physiological finding that most STRFs in primary auditory cortex have the quadrant separability property [33].

$$STRF_{+} = \Re\left\{ H_{rate}(t;\omega,\theta) . H_{scale}(f;\Omega,\phi) \right\}$$
$$STRF_{-} = \Re\left\{ H^{*}_{rate}(t;\omega,\theta) . H_{scale}(f;\Omega,\phi) \right\}$$

where $\Re$ denotes the real part, * the complex conjugate, ω and Ω the velocity (rate) and spectral density (scale) parameters of the filters, and θ and ϕ are characteristic phases that determine the degree of asymmetry along time and frequency respectively. Functions H_{rate} and H_{scale} are analytic signals (a signal which has no negative frequency components) obtained from h_{rate} and h_{scale}

$$H_{rate}(t;\omega,\theta) = h_{rate}(t;\omega,\theta) + j\hat{h}_{rate}(t;\omega,\theta)$$

$$H_{scale}(f;\Omega,\phi) + j\hat{h}_{scale}(f;\Omega,\phi)$$

Where $\hat{\ }$ denotes Hilbert transformation. h_{rate} and h_{scale} are temporal and spectral impulse responses defined by sinusoidally interpolating between symmetric seed function $h_r(.)$ (second derivative of a Gaussian function) and $h_s(.)$ (Gamma function), and their asymmetric Hilbert transforms

$$h_{rate}(t;\omega)\cos\theta + \hat{h}(t;\omega)\sin\theta$$

$$h_{scale}(f;\Omega,\phi) = h_s(f;\Omega)\cos\phi + \hat{h}_s(f;\Omega)\sin\phi$$

The impulse responses for different scales and rates are given by dilation

$$h_r(t;\omega) = \omega\, h_r(\omega t)$$

$$h_s(f;\Omega) = \Omega h_s(\Omega f)$$

Therefore, the spectro-temporal response for an input spectrogram is given by

$$r_{+}(t, f; \omega, \Omega; \theta, \phi) = y(t, f) *_{t,f} \; STRF_{+}(t, f; \omega, \Omega; \theta, \phi)$$

$$r_{-}(t, f; \omega, \Omega; \theta, \phi) = y(t, f) *_{t,f} \; STRF_{-}(t, f; \omega, \Omega; \theta, \phi)$$

where $*_{t,f}$ denotes convolution with respect to both t and f. It is useful to compute the spectro-temporal response $r_{\pm}(.)$ in terms of the output magnitude and phase of the downward $(+)$ and upward $(-)$ selective filters. For this, the temporal and spatial filters, h_{rate} and h_{scale} can be equivalently expressed in the wavelet-based analytical forms $h_{r\omega}(.)$ and $h_{s\omega}(.)$ as

$$h_{r\omega}(t; \omega) = h_{r}(t; \omega) + j\hat{h}_{r}(t; \omega)$$

$$h_{sw}(f; \Omega) = h_{s}(f; \Omega) + j\hat{h}_{s}(f; \Omega)$$

The complex response to downward and upward selective filters, $z_{+}(.)$ and $z_{-}(.)$, is then defined as

$$z_{+}(t, f; \Omega, \omega) = y(t, f) *_{tf} \left[h^{*}_{r\omega}(t; \omega) h_{s\omega}(f; \Omega)\right]$$

$$z_{-}(t, f; \Omega, \omega) = y(t, f) *_{tf} \left[h_{r\omega}(t; \omega) h_{s\omega}(f; \Omega)\right]$$

where * denotes the complex conjugate. The cortical response [66] [33] for all characteristic phases θ and ϕ can be easily obtained from $z_{+}(.)$ and $z_{-}(.)$ as follows:

$$r_{+}(t, f; \omega, \Omega; \theta, \phi) = \left|z_{+}\right| \cos(\angle z_{+} - \theta - \phi)$$

$$r_{-}(t, f; \omega, \Omega; \theta, \phi) = \left|z_{-}\right| \cos(\angle z_{-} + \theta - \phi)$$

where $|\cdot|$ denotes the magnitude and $\angle\cdot$ the phase. The magnitude and the phase of z_+ and z_- have a physical interpretation: at any time and for all the STRFs tuned to the same (f,ω,Ω), the ones with $\theta=\frac{\angle z_+ + \angle z_-}{2}$ and $\phi=\frac{\angle z_+ - \angle z_-}{2}$ symmetries have the maximal downward and upward responses of $|z_+|$ and $|z_-|$. These maximal responses are used for the purpose of classification. Where the spectro-temporal modulation content of the spectrogram is of particular interest, we obtain the summed output from all filters with identical modulation selectivity or STRFs to generate the rate-scale plots: [as shown in **Figure 39** for speech]

$$u_+(\omega,\Omega)=\sum_t\sum_f |z_+(t,f;\omega,\Omega)|$$

$$u_-(\omega,\Omega)=\sum_t\sum_f |z_-(t,f;\omega,\Omega)|$$

The final view that emerges is that of a continuously updated estimate of the spectral and temporal modulation content of the auditory spectrogram. All parameters of this model are derived from physiological data in animals and psychoacoustical data in human subjects as explained in detail in [40], [33], and [67]. Unlike conventional features, our auditory-based features have multiple scales of time and spectral resolution. Some respond to fast changes while others are tuned to slower modulation patterns; A subset are selective to broadband spectra, and others are more narrowly

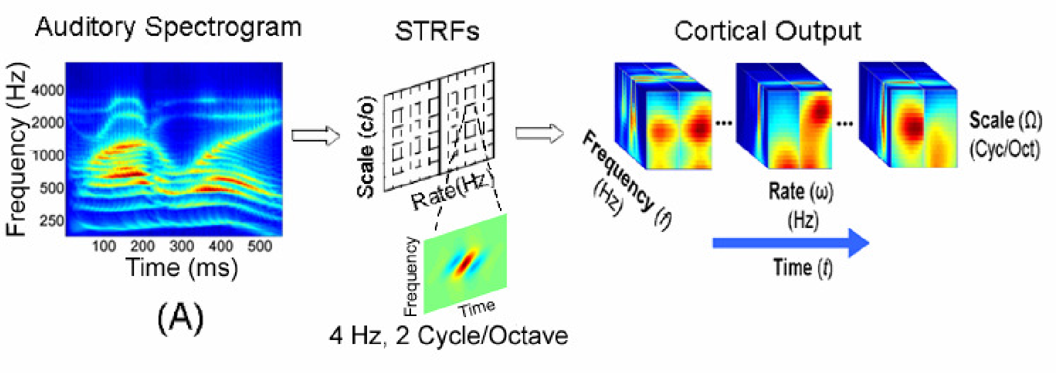

Figure 39. The cortical model of auditory pathway (A) The cortical multi-scale representation of speech. The auditory spectrogram (the output of the early stage) is analyzed by a bank of spectrotemporal modulation selective filters. The spectro-temporal response field (STRF) of one such filter is shown which corresponds to a neuron that responds well to a ripple of 4Hz rate and 0.5 cycle/octave scale. The output from such a filter is computed by convolving the STRF with the input spectrogram. The total output as a function of time from the model is therefore indexed by three parameters: scale, rate, and frequency.

tuned. For this study, temporal filters (rate) ranging from 1 to 32 Hz, and spectral filters (scale) from 0.5 to 8.00 cycle/octave, were used to represent the spectro-temporal modulations of the sound.

6.2 Speech detection

Audio segmentation and classification have important applications in audio data retrieval, archive management, modern human-computer interfaces, and in entertainment and security tasks. In speech recognition systems designed for real world conditions, a robust discrimination of speech from other sounds is a crucial step. Speech discrimination can also be used for coding or telecommunication applications where nonspeech sounds are not of interest, and, hence, bandwidth is saved by not transmitting them or by assigning them a low resolution code. Finally, as the amount of available audio data increases, manual segmentation of audio sounds has become more difficult and impractical and alternative automated procedures are much needed. Speech is a sequence of consonants and vowels, nonharmonic and harmonic sounds, and natural silences between words and phonemes. Discriminating speech from nonspeech is often complicated by the similarity of many sounds to speech, such as animal vocalizations. As with other pattern recognition tasks, the first step in this audio classification is to extract and represent the sound by its relevant features. To achieve good performance and generalize well to novel sounds, this representation should be able both to capture the discriminative properties of the sound, and to resist distortion under various noisy conditions. Research into content-based audio classification is relatively new. Among the earliest is the work of Pfeiffer *et al.* [68], where a 256 phase-compensated gammatone filter bank was used to extract audio features that mapped the sound to response probabilities. Wold *et al.* [69] adopted instead a statistical model of time-frequency measurements to represent perceptual values of the sound. A common alternative approach involves the extraction of different higher level features to classify audio, such as Mel-frequency cepstral coefficients (MFCCs) along with a vector quantizer [70], or noise frame ratios and band periodicity along with K-nearest neighbor and linear spectral pair-vector quantization [71], average zero-crossing rate and energy with a simple threshold to discriminate between speech and music [72], and an optimized dimensionality reduction using distortion discriminant analysis (DDA) [73].

Two more elaborate systems have been proposed, against which we shall compare our system. The first is proposed by Scheirer and Slaney [74] in which thirteen features in time, frequency, and cepstrum domain are used to model speech and music. Several classification techniques [e.g., maximum *a posteriori* (MAP), Gaussian mixture model (GMM), K nearest neighbor (KNN)] are then employed to achieve a robust performance. The second system is a speech/nonspeech segmentation technique [75] in which frame-by-frame maximum autocorrelation and log-energy features are measured, sorted, and then followed by linear discriminant analysis and a diagonalization transform. The novel aspect of our proposed system is a feature set inspired by investigations of various stages of the auditory system [65][40]. The features are computed using a model of the auditory cortex that maps a given sound to a high-dimensional representation of its spectro-temporal modulations. A key component that makes this approach practical is a multilinear dimensionality reduction method that by making use of multimodal characteristic of cortical representation, effectively removes redundancies in the measurements in each subspace separately, producing a compact feature vector suitable for classification.

6.2.1 Multilinear tensor analysis

The output of the auditory model is a multidimensional array in which modulations are presented along the four dimensions of time, frequency, rate, and scale. For our purpose here, the time axis is averaged over a given time window which results in a three mode tensor for each time window with each element representing the overall modulations at corresponding frequency, rate, and scale. In order to obtain a good resolution, sufficient number of filters in each mode are required. As a consequence, the dimensions of the feature space are very large (5 scale filters * 12 rate filters *128 frequency channels = 7680). Working in this feature space directly is impractical because a sizable number of training samples is required to characterize the space adequately [76]. Traditional dimensionality reduction methods like principal component analysis (PCA) are inefficient for multidimensional data because they treat all the elements of the feature space similarly without considering the varying degrees of redundancy and discriminative contribution of each mode. Instead, it is possible using multidimensional PCA to tailor the amount of reduction in each subspace independently of others based on the relative magnitude of corresponding singular values. Furthermore, it is also feasible to reduce the amount of training samples and computational load significantly since each subspace is considered separately. We shall demonstrate here the utility of a generalized method for the PCA of multidimensional data based on higher-order singular-value decomposition (HOSVD) [77].

6.2.2 Basic tensor definition

Multilinear algebra is the algebra of tensors. Tensors are generalizations of scalars (no indices), vectors (single index), and matrices (two indices) to an arbitrary number of indices. They provide a natural way of representing information along many dimensions. Substantial results have already been achieved in this field. Tucker first formulated the three-mode data model [78], while Kroonenberg formulated alternating least-square (ALS) method to implement three mode factor analysis [79]. Lathauwer *et al.* established a generalization of singular value decomposition (SVD) to higher order tensors [77], and also introduced an iterative method for optimizing the best rank $(R_1, R_2, \ldots, R_N)$ approximation of tensors [80]. Tensor algebra and HOSVD have been applied successfully in wide variety of fields including higher-order-only independent component analysis (ICA) [81], face recognition [82], and selective image compression along a desired dimension [83].

A Tensor $A \in \Re^{I_1 \times I_2 \times \cdots \times I_N}$ is a multi-index array of numerical values whose elements are denoted by $a_{i_1 i_2 \cdots i_N}$. Matrix column vectors are referred to as mode-1 vectors and row vectors as mode-2 vectors. The mode-n vectors of an N_{th} order tensor A are the vectors with I_n components obtained from A by varying index I_n while keeping the other indices fixed. Matrix representation of a tensor is obtained by stacking all the columns (rows, …) of the tensor one after the other. The mode-n matrix unfolding of $A \in \Re^{I_1 \times I_2 \times \cdots \times I_N}$ denoted by $A_{(n)}$ is the $(I_n \times I_1 I_2 \cdots I_{n-1} I_{n+1} \cdots I_N)$ matrix whose columns are n-mode vectors of tensor A.

An N_{th}-order tensor A has rank – 1 when it is expressible as the outer product of N vectors

$$A = U_1 \circ U_2 \circ \ldots \circ U_N$$

The rank of an arbitrary N_{th}-order tensor A, denoted by $r = rank(A)$ is the minimal number of rank-1 tensors that yield A in a linear combination. The $n-$rank of $A \in \Re^{I_1 \times I_2 \times \cdots \times I_N}$ denoted by r_N, is defined as the dimension of the vector space generated by the mode-n vectors

$$R_n = rank_n(A) = rank(A_{(n)})$$

The n-mode product of a tensor $A \in \Re^{I_1 \times I_2 \times \cdots \times I_N}$ by a matrix $U \in \Re^{J_n \times I_n}$ denoted by $A \times_n U$ is an $(I_1 \times I_2 \times \cdots \times J_n \times \cdots \times I_N) -$ tensor given by

$$(A \times_n U)_{i_1 i_2 \cdots j_n \cdots i_N} = \sum_{i_n} a_{i_1 i_2 \cdots i_n \cdots i_N} u_{j_n i_n}$$

For all index values.

6.2.3 Multilinear SVD and PCA

Matrix singular-value decomposition orthogonalizes the space spanned by column and rows of the matrix. In general, every matrix *D* can be written as the product

$$D = U \cdot S \cdot V^T = S \times_1 U \times_2 V$$

in which *U* and *V* are unitary matrices contains the left- and right-singular vectors of *D*. *S* is a pseudodiagonal matrix with ordered singular values of *D* on the diagonal. If *D* is a data matrix in which each column represents a data sample, then the left singular vectors of *D* (matrix) are the principal axes of the data space. Keeping only the coefficients corresponding to the largest singular values of *D* (principal components or PCs) is an effective means of approximating the data in a low-dimensional subspace. To generalize this concept to multidimensional data, we consider a generalization of SVD to tensors [77]. Every -tensor can be written as the product

$$A = S \times_1 U^{(1)} \times_2 U^{(2)} \ldots \times_N U^{(N)}$$

in which $U^{(n)}$ is a unitary matrix containing left singular vectors of the *mode-n* unfolding of tensor *A*, and *S* is a $(I_1 \times I_2 \times \ldots \times I_N)$ tensor which has the properties of all-orthogonality and ordering. The matrix representation of the HOSVD can be written as

$$A_{(n)} = U^{(n)}.S_{(n)} \cdot \left(U^{(n+1)} \cdots U^{(N)} \otimes U^{(1)} \otimes U^{(2)} \cdots U^{(n-1)}\right)^T$$

in which $\otimes$ denotes the Kronecker product. The previous equation can also be expressed as

$$A_{(n)} = U^{(n)} \cdot \Sigma^{(n)} \cdot V^{(n)^T}$$

in which $\Sigma^{(n)}$ is a diagonal matrix made by singular values of $A^{(n)}$ and

$$V^{(n)} = \left(U^{(n+1)} \cdots U^{(N)} \otimes U^{(1)} \otimes U^{(2)} \cdots U^{(n-1)}\right)$$

This shows that, at matrix level, the HOSVD conditions lead to an SVD of the matrix unfolding. Lathauwer *et al.* shows [77] that the left-singular matrices of the different matrix unfolding of *A* correspond to unitary transformations that induce the HOSVD structure which in turn ensures that the HOSVD inherits all the classical space properties from the matrix SVD. HOSVD results in a new ordered orthogonal basis for representation of the data in subspaces spanned by each mode of the tensor. Dimensionality reduction in each space is obtained by projecting data samples on principal axes and keeping only the components that correspond to the largest singular values of that subspace. However, unlike the matrix case in which the best *rank-R* approximation of a given matrix is obtained from the truncated SVD, this procedure does not result in optimal approximation in the case of tensors. Instead, the optimal best $rank-(R_1, R_2, \dots, R_N)$ approximation of a tensor can be obtained by an iterative algorithm in which HOSVD provides the initial values [80].

6.2.4 Multilinear analysis of cortical representation

The auditory model transforms a sound signal to its corresponding time-varying cortical representation. Averaging over a given time window results in a cube of data in rate-scale-frequency space. Although the dimension of this space is large, its elements are highly correlated making it possible to reduce the dimension significantly using a comprehensive data set, and finding new multilinear and mutually orthogonal principal axes that approximate the real space spanned by these data. The assembled training set contains 1223 samples from speech and nonspeech classes. The resulting data tensor *D*, obtained by stacking all training tensors is a 5*12*128*1223 tensor. Next, the tensor is decomposed to its *mode-n* singular vectors

$$D = S \times_1 U_{frequency} \times_2 U_{rate} \times_3 U_{scale} \times_4 U_{samples}$$

In which $U_{frequency}$, U_{rate}, and U_{scale} are orthonormal ordered matrices containing subspace singular vectors, obtained by unfolding *D* along its corresponding modes. Tensor *S* is the core tensor with the same dimensions as *D.* Each singular matrix is then truncated by setting a predetermined threshold so as retain only the desired number of principal axes in each mode. New sound samples are first transformed to their cortical representation, *A*, and are then projected onto these truncated orthonormal axes $U'_{freq}, U'_{rate}, U'_{scale}$ (as shown in Fig. 3)

$$Z = A \times_1 {U'_{freq.}}^T \times_2 {U'_{rate}}^T \times_3 {U'_{scale}}^T$$

The resulting tensor *Z* whose dimension is equal to the total number of retained singular vectors in each mode, thus, contains the multilinear cortical principal components of the sound sample. *Z* is then vectorized and normalized by subtracting its mean and dividing by its norm to obtain a compact feature vector for classification.

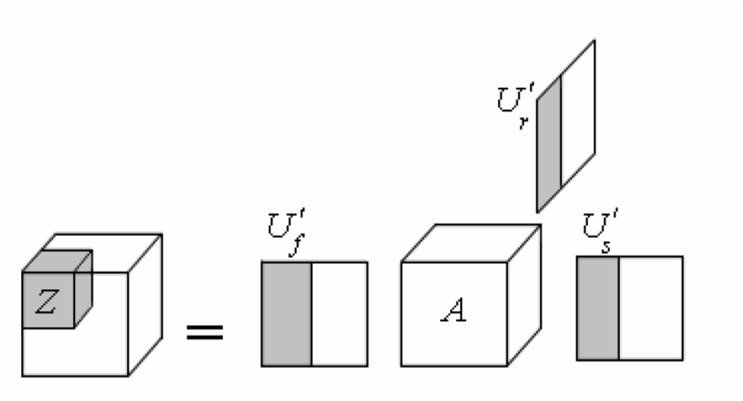

Figure 40. Illustration of Tensor dimensionality reduction

6.2.5 Classification

Classification was performed using a support vector machine (SVM) [84], [85]. SVMs find the optimal boundary that separates two classes in such a way as to maximize the margin between separating boundary and closest samples to it (support vectors). This in general results in improving generalization from training to test data [84]. Radial basis function (RBF) was used as SVM kernel.

6.2.6 Experimental results

Audio database

An audio database was assembled from five publicly available corpora. Details of the database are as follows. Speech samples were taken from TIMIT Acoustic-Phonetic Continuous Speech Corpus [24] which contains short sentences spoken by male and female native English speakers with eight dialects. Two hundred ninety-nine different sentences spoken by different speakers (male and female) were selected for training and 160 different sentences spoken by different speakers (male and female) were selected for test purpose. Sentences and speakers in training and test sets were also different. To make the nonspeech class as comprehensive as possible, sounds from animal vocalizations, music, and environmental sounds were assembled together. Animal vocalization were taken from BBC Sound Effects audio CD collection [86] (263 for training, 139 for test). Music samples that covered a large variety of musical styles were selected from RWC genre database [87] (349 for training, 185 for test). Environmental sounds were assembled from Noisex [88] and Auroa [89] databases which have stationary and nonstationary sounds including white and pink noise, factory, jets, destroyer engine, military vehicles, cars, and several speech babble recorded in different environments like restaurant, airport, and exhibition (312 for training, 167 for test). The training set included 299 speech and 924 nonspeech samples and the test set consisted of 160 speech and 491 nonspeech samples. The length of each utterance in training and test is equal to the selected time window (e.g., one 1-s sample per sound file).

Number of principal components

The number of retained PCs in each subspace is determined by analyzing the contribution of each PC to the representation of associated subspace. The contribution of j_{th} principal component of subspace S_i whose corresponding eigenvalue is $\lambda_{i,j}$ is defined as

$$\alpha_{i,j} = \frac{\lambda_{i,j}}{\sum_{k=1}^{N_i} \lambda_{i,k}}$$

where N_i denotes the dimension of S_i (128 for frequency, 12 for rate and 5 for scale). The number of PCs in each subspace then can be specified by including only the PCs whose α is larger than some threshold. The classification accuracy on a validation set was used to determine the number of PCs used in each subspace. Based on this analysis, the minimum number of principal components to achieve 100% accuracy was specified to be 7 for frequency, 5 for rate and 4 for scale subspace that includes PCs that have contribution of 3.5% or more.

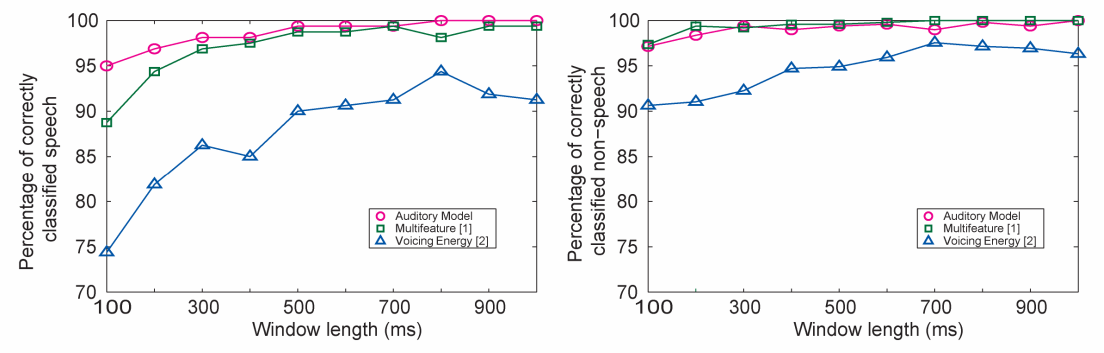

Figure 41. Effect of window length on the percentage of correctly classified speech and non-speech

6.2.7 Comparison and results

To evaluate the robustness and the ability of system to generalize to unseen noisy conditions, we conducted a comparison with two state-of-the-art studies, one from generic-audio analysis community by Scheirer and Slaney [74] and one from automatic-speech-recognition community by Kingsbery *et al.* [75]. *Multifeature [74]:*

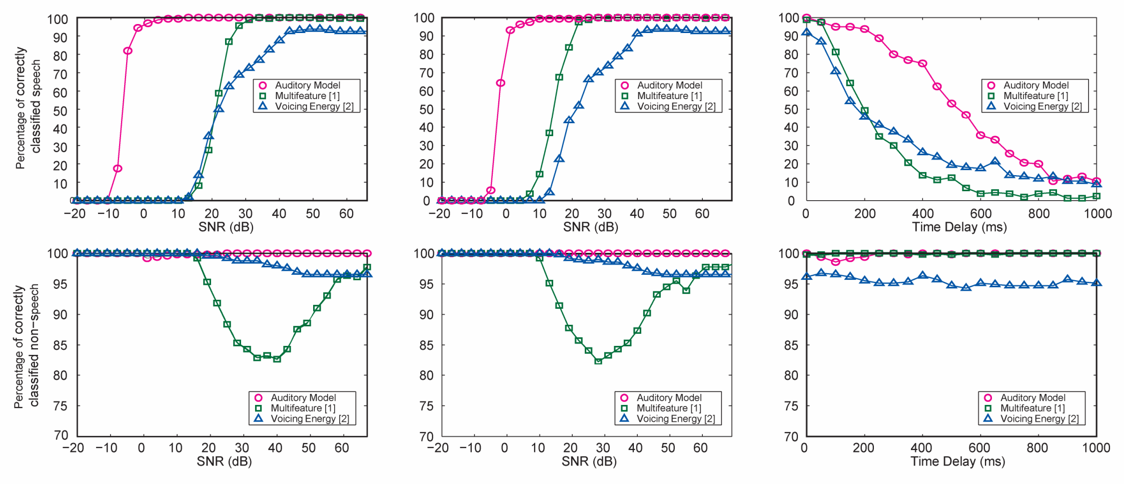

Figure 42. Percentage of correctly classified speech and non-speech in noise and reverberation. The percentages are shown for auditory model, multifeature[74] and voicing-energy[75] methods in additive white noise (left panels), additive pink noise (middle panels) and reverberation (right panels).

The first system, which was originally designed to distinguish speech from music, derived 13 features in time, frequency, and cepstrum domain to represent speech and music. The features were 4-Hz modulation energy, percentage of "low-energy" frames, spectral rolloff point, spectral centroid, spectral flux, zero-crossing rate, cepstrum resynthesis residual, and their variances. The 13th feature, pulse metric, was neglected for this comparison since its latency was too long (more than 2 s). In the original system, two models were formed for speech and music in the feature space. Classification was performed using a likelihood estimate of a given sample for each model.

To eliminate performance differences due to the use of different classifiers, an SVM with an RBF kernel was used in all comparisons. Our implementation of the system was first evaluated on the original database and similar or better results were obtained with SVM compared to the original publication [74]. *Voicing-Energy [75]:* A second system was tested that was based on an audio segmentation algorithm from the ASR work [75]. In the proposed technique, the feature vector used in the segmentation incorporated information about the degree of voicing and frame-level log-energy value. Degree of voicing is computed by finding the maximum of autocorrelation in a specified range, whereas log-energy was computed for every short frame of sound weighted with a Hanning window. Several frames of these features were then concatenated and sorted in increasing order, and the resulting feature vector was reduced to two dimensions by a linear discriminant analysis followed by diagonalizing transform. The reason for sorting the elements was to eliminate details of temporal evolutions which were not relevant for this task. Our evaluation of Kingsbury's system suggested that direct classification of the original sorted vector with an SVM classifier similar to the other two systems outperformed the one in reduced dimension. For this reason, the classification was performed in the original feature space. Our auditory model and the two benchmark algorithms from the literature were trained and tested on the same database. One of the important parameters in any such speech detection/discrimination task is the time window or duration of the signal to be classified, because it directly affects the resolution and accuracy of the system. **Figure 41** demonstrate the effect of window length on the percentage of correctly classified speech and nonspeech. In all three methods, some features may not give a meaningful measurement when the time window is too short. The accuracy of all three systems improve as the time window increases. Audio processing systems designed for realistic applications must be robust in a variety of conditions because training the systems for all possible situations is impractical. Detection of speech at very low SNR is desired in many applications such as speech enhancement in which a robust detection of nonspeech (noise) frames is crucial for accurate measurement of the noise statistics [90]. A series of tests were conducted to evaluate the generalization of the three methods to unseen noisy and reverberant sound. Classifiers were trained solely to discriminate clean speech from nonspeech and then tested in three conditions in which speech was distorted with noise or reverberation. In each test, the percentage of correctly detected speech and nonspeech was considered as the measure of performance. For the first two tests, white and pink noise were added to speech with specified signal to noise ratio (SNR). White and pink noise were not included in the training set as nonspeech samples. SNR was measured from the average power of speech and noise

$$SNR = 10\log\frac{P_s}{P_n}$$

Figure 42 left and middle column illustrate the effect of white and pink noise on the average spectro-temporal modulations of speech. The spectro-temporal representation of noisy speech preserves the speech specific features (e.g., near 4 Hz, 2 cycle/octave) even at SNR as low as 0 dB (**Figure 43**, top and middle rows). The detection results for speech in white noise (**Figure 42**, left column) demonstrate that while the three systems have comparable performance in clean conditions, the auditory features remain robust down to fairy low SNRs. This pattern is repeated with additive pink

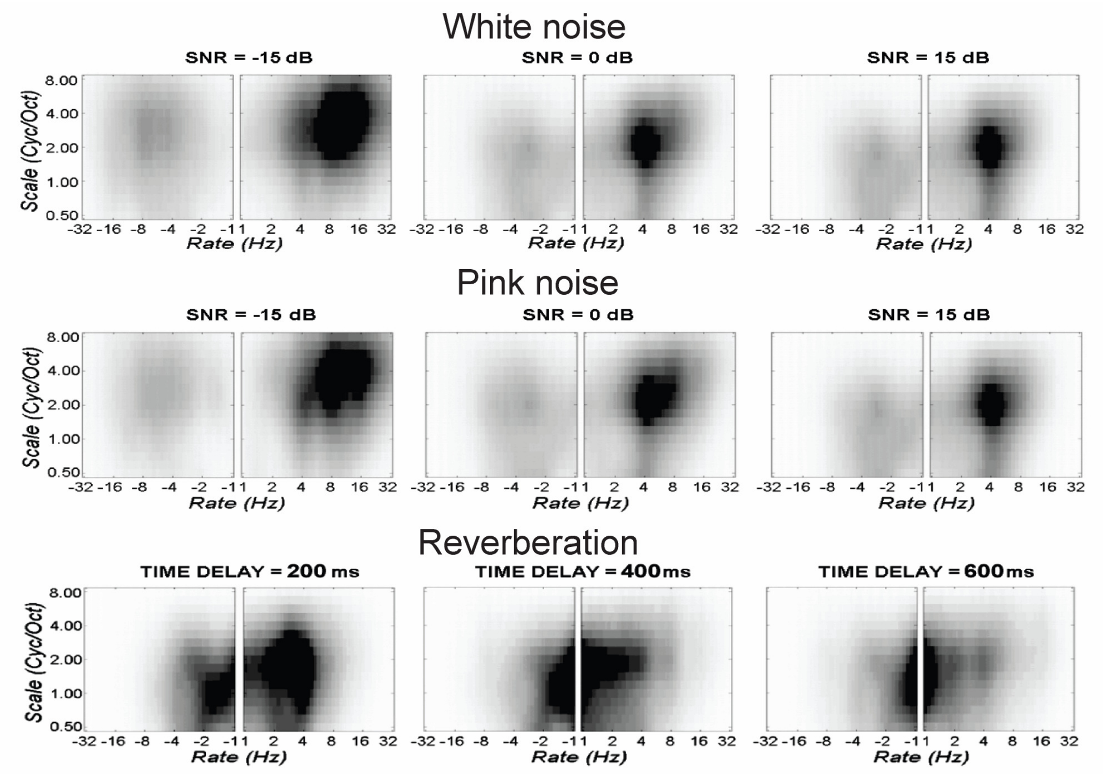

Figure 43. Effect of white and pink noise and reverberation on average rate-scale representation of speech. (top row) Effect of *white* noise on average spectro-temporal modulations of speech for SNRs −15dB, 0dB and 15dB. The spectro-temporal representation of noisy speech preserves the speech specific spectro-temporal features (e.g. near 4Hz, 2Cyc/Oct) even at SNR as low as 0 *dB*. **(middle row)** Effects of *pink* noise on average spectro-temporal modulations of speech for different SNRs −15dB, 0dB and 15dB. The speech specific spectrotemporal features (e.g. near 4Hz, 2Cyc/Oct) are preserved even at SNR as low as 0 *dB*. **(bottom row)** Effects of *reverberation* on average spectro-temporal modulations of speech for time delays 200ms, 400ms and 600ms. Increasing the time delay results in gradual loss of high-rate temporal modulations of speech.

noise although performance degradation for all systems occurs at higher SNRs (**Figure 42**, middle) because of more overlap between speech and noise energy.

Reverberation is another widely encountered distortion in realistic applications. To examine the effect of different levels of reverberation on the performance of these systems, a realistic reverberation condition was simulated by convolving the signal with a random gaussian noise with exponential decay. The effect on the average spectro-temporal modulations of speech are shown in **Figure 43**. Increasing the time delay results in gradual loss of high-rate temporal modulations of speech. **Figure 42** demonstrate the effect of reverberation on the classification accuracy.

On the whole, these tests demonstrate the significant robustness of the auditory model.

6.2.8 Conclusions

A *spectro-temporal auditory method* for audio classification and segmentation has been described, tested, and compared to two state-of-the-art alternative approaches. The method employs features extracted by a biologically inspired auditory model of auditory processing in the cortex. Unlike conventional and spectral resolution. The drawback of such a representation is its high dimensionality, and, hence, to utilize it, we applied an efficient multilinear dimensionality reduction algorithm based

on HOSVD of multimodal data. The performance of the proposed auditory system was tested in noise and reverberation and compared favorably with alternative systems, thus, demonstrating that the proposed system generalizes well to novel situations, an ability that is generally lacking in many of today's audio and speech recognition and classification systems. The success of these multiscale features for this speech detection task suggests that these features are more worth investigating for speech recognition or noise suppression than conventional approaches based on simple cepstral features. This work is but one in a series of efforts at incorporating multiscale cortical representations (and more broadly, perceptual insights) in a variety of audio and speech processing applications. For example, the deterioration of the spectro-temporal modulations of speech in noise and reverberation (**Figure 43**), or indeed under any kind of linear or nonlinear distortion, can be used as an indicator of predicted speech intelligibility [91]. Similarly, the multiscale rate-scale-frequency representation can account for the perception of complex sounds and perceptual thresholds in a variety of settings [92]. Finally, the auditory model can be adapted and expanded for a wide range of applications such as the speech enhancement [90], or the efficient encoding of speech and music [93].

6.3 Speech enhancement

Noise suppression with complex broadband signals is often employed in order to enhance quality or intelligibility in a wide range of applications including mobile communication, hearing aids, and speech recognition. In speech research, this has been an active area of research for over fifty years, mostly framed as a statistical estimation problem in which the goal is to estimate speech from its sum with other independent processes (noise). This approach requires an underlying statistical model of the signal and noise, as well as an optimization criterion. In some of the earliest work, one approach was to estimate the speech signal itself [94]. When the distortion is expressed as a minimum mean-square error, the problem reduces to the design of an optimum Wiener filter. Estimation can also be done in the frequency domain, as is the case with such methods as spectral subtraction [94], the signal subspace approach [95], and the estimation of the short-term spectral magnitude [96]. Estimation in the frequency domain is superior to the time domain as it offers better initial separation of the speech from noise, which (1) results in easier implementation of optimal/heuristic approaches, (2) simplifies the statistical models because of the decorrelation of the spectral components, and (3) facilitates integration of psychoacoustic models [97]. Recent psychoacoustic and physiological findings in mammalian auditory systems, however, suggest that the spectral decomposition is only the first stage of several interesting transformations in the representation of sound. Specifically, it is thought that neurons in the auditory cortex decompose the spectrogram further into its spectrotemporal modulation content [98]. The focus of this section is an application of this model to the problem of speech enhancement. The rationale for this approach is the finding that modulations of noise and speech have a very different character, and hence they are well separated in this multiscale representation, more than the case at the level of the spectrogram. Modulation frequencies have been used in noise suppression before (e.g., [99]), however this study is different in several ways: (1) the proposed method is based on filtering not only the temporal modulations, but the joint spectrotemporal modulations of speech; (2) modulations are not used to obtain the weights of frequency channels. Instead, the filtering itself is done in the spectrotemporal modulation domain; (3) the filtering is done only on the slow temporal modulations of speech (below 32 Hz) which are important for intelligibility. A key computational component of this approach is an *invertible* auditory model which captures the essential auditory transformations from the early stages up to the cortex, and provides an algorithm for inverting the “filtered representation” back to an acoustic signal.

6.3.1 Multiresolution representation of speech and noise

In this section, we explain how the cortical representation captures the modulation content of sound. We also demonstrate the separation between representation of speech and different kind of noise which is due to their distinct spectrotemporal patterns. The output of the cortical model described is a 4-dimensional tensor with each point indicating the amount of energy at corresponding time, frequency, rate, and scale ($z\pm(t,f,\omega,\Omega)$). One can think of each point in the spectrogram (e.g., time *tc* and frequency *fc* in **Figure 45**) as having a two-dimensional rate-scale representation ($z \pm (tc, fc,\omega,\Omega)$) that is an estimate of modulation energy at different temporal and spectral resolutions. The modulation filters with different resolutions capture local and global information about each point as shown in **Figure 45** for time *tc* and frequency *fc* of the speech spectrogram. In this example, the temporal modulation has a peak around 4Hz which is the typical temporal rate of speech. The spectral modulation, scale, on the other hand spans a wide range reflecting at its high end the harmonic structure due to voicing (2–6 Cycle/Octave) and at its low end the spectral envelope or formants (less than 2 Cycle/Octave). Another way of looking at the modulation content of a sound is to collapse the time dimension of the cortical representation resulting in an estimate of the average rate-scale-frequency modulation of the sound in that time window. This average is useful, especially when the sound is relatively stationary as is the case for many background noises and is calculated in the following way:

$$U_{\pm}(\omega,\Omega,f)=\int_{t_1}^{t_2}\left|\,z_{\pm}(\omega,\Omega,f,t)\right|dt$$

Figure 44 shows the average multiresolution representation ($U\pm$) of speech and four different kinds of noise chosen from Noisex database [100]. Top row of **Figure 44** shows the spectrogram of speech, white, jet, babble, and city noise. These four kinds of noise are different in their frequency distribution as well as in their spectrotemporal modulation pattern as demonstrated in **Figure 44**. Rows B, C, and D in **Figure 44** show the average rate-scale, scale-frequency, and rate-frequency representations of the corresponding sound calculated from the average rate-scale-frequency representation ($U\pm$) by collapsing one dimension at a time. As shown in rate-scale displays in **Figure 44b**, speech has strong slow temporal and low-scale modulation; on the other hand, speech babble shows relatively faster temporal and higher spectral modulation. Jet noise has a strong 10Hz temporal modulation which also has a high scale because of its narrow spectrum. White noise has modulation energy spread over a wide range of rates and scales. **Figure 44c** shows the average scale-frequency representation of the sounds, demonstrating how the energy is distributed along the dimensions of frequency and spectral modulation. Scale-frequency representation shows a notable difference between speech and babble noise with speech having stronger low-scale modulation energy. Finally, **Figure 44d** shows the average rate frequency representation of the sounds, that shows how energy is distributed in different frequency channels and temporal rates. Again, jet noise shows a strong 10Hz temporal modulation at frequency 2 KHz. White noise on the other hand activates most rate and frequency filters with increasing energy for higher-frequency channels reflecting the increased bandwidth of constant-Q auditory filters. Babble noise activates low and mid frequency filters better, similar to speech but at higher rates. City noise also activates wide range of filters. As **Figure 44** shows that spectrotemporal modulations of speech have very different characteristics than the four noises, which is the reason we can discriminately keep its modulation components while reducing the noise ones. The three-dimensional average noise modulation is what we used as the noise model in the speech enhancement algorithm as described in the next section.

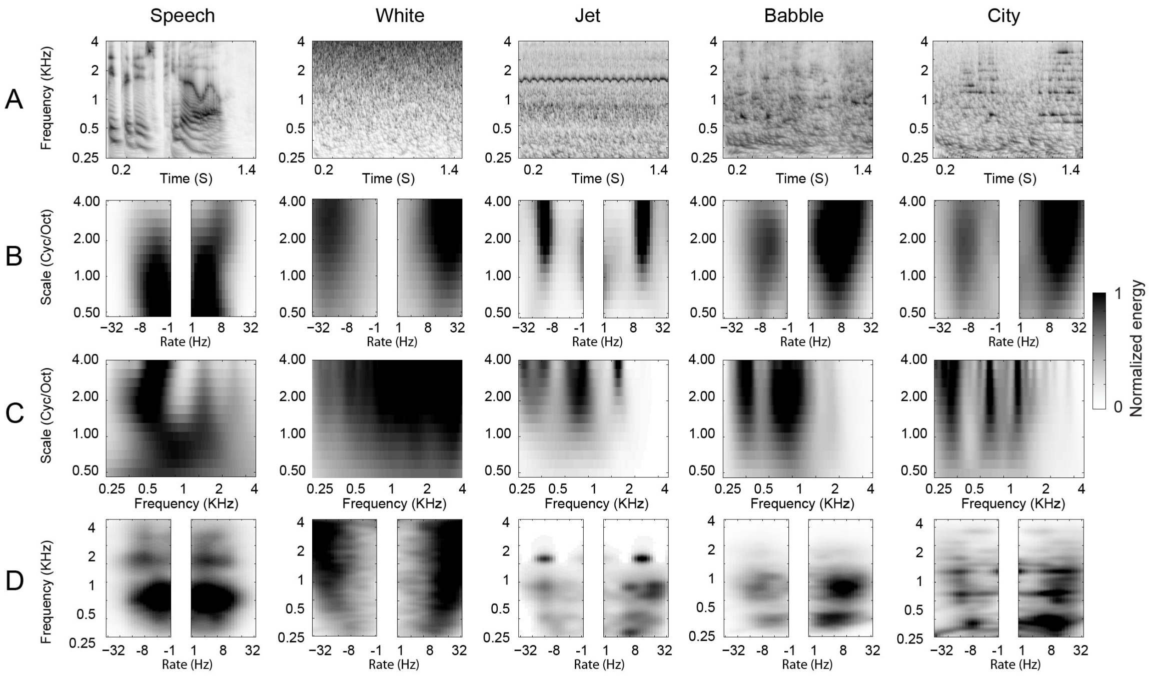

Figure 44. Auditory spectrogram and average cortical representations of speech and four different kinds of noise. Row (a): auditory spectrogram of speech, white, jet, babble, and city noise taken from Noisex database. Row (b): average rate-scale representations of sound demonstrate the distribution of energy in different temporal and spectral modulation filters. Speech is well separated from the noises in this representation. Row (c): average scale-frequency representations. jet have mostly high scales because of its narrow-band frequency distributions. Row (d): average rate-frequency representations show the energy distributions in different frequency channels and rate filters.

6.3.2 Estimation of noise modulations

A crucial factor in affecting the performance of any noise suppression technique is the quality of the background noise estimation. In spectral subtraction algorithms, several techniques have been proposed that are based on three assumptions: (1) speech and noise are statistically independent, (2) speech is not always present, and (3) the noise is more stationary than speech [97]. One of these methods is voice activity detection (VAD) that estimates the likelihood of speech at each time window and then uses the frames with low likelihood of speech to update the noise model. One of the common problems with VADs is their poor performance at low SNRs. To overcome this limitation, we employed the speech detector (also based on the cortical representation) which detected speech reliably at SNR's as low as −5dB as described in the previous section. The frames marked by the SVM as nonspeech are then added to the noise model (*N*±),which is an estimate of noise energy at each frequency, rate, and scale:

$$N_{\pm}(f,\omega,\Omega)=\int_{noise\ frames}\left|z_{\pm}(t,f,\omega,\Omega)\right|dt$$

As shown in **Figure 44**, this representation is able to capture the noise information beyond just the frequency distribution, as is the case with most spectral subtraction-based approaches. Also, as can be seen in Figure 3, speech and most kinds of noises are well separated in this domain.

6.3.3 Noise Suppression

The exact rule for suppressing noise coefficients is a determining factor in the subjective quality of the reconstructed enhanced speech, especially with regards to the reduction of musical noise [97]. Having the spectrotemporal representation of noisy sound and the model of noise average modulation energy, one can design a rule that suppresses the modulations activated by the noise and emphasize the ones that are from the speech signal. One possible way of doing this is to use a Wiener filter in the following form:

$$H_{\pm}(t,f,\omega,\Omega)=\left(\frac{SNR_{\pm}(t,f,\omega,\Omega)}{1+SNR_{\pm}(t,f,\omega,\Omega)}\right)\approx\left(1-\frac{N_{\pm}(f,\omega,\Omega)}{S_{N_{\pm}}(t,f,\omega,\Omega)}\right)$$

where $N_{\pm}$ is the noise model calculated by averaging the cortical representation of noise-only frames and *SN* is the cortical representation of noisy speech signal. The resulting gain function maintain the output of filters with high SNR values while attenuating the output of low-SNR filters:

$$\hat{z}_{\pm}(t,f,\omega,\Omega)=z_{\pm}(t,f,\omega,\Omega)\cdot H_{\pm}(t,f,\omega,\Omega)$$

$\hat{z}$ is the modified (denoised) cortical representation from which the cleaned speech is reconstructed. This idea is demonstrated in **Figure 45**. **Figure 45**A shows the spectrogram of a speech sample contaminated by jet noise and its rate scale representation at time *tc* and frequency *fc* (**Figure 45**A) which is a point in the spectrogram that noise and speech overlap. This type of noise has a strong temporally modulated tone (10 Hz) at frequency around 2 KHz. The rate-scale representation of the jet noise for the same frequency, *fc*, is shown in **Figure 45**B. Comparing the noisy speech representation with the one from noise model, it is easy to see what parts belong to noise and what parts come from the speech signal. Therefore, we can recover the clean rate-scale representation by attenuating the modulation rates and scales that show strong energy in the noise model. This intuitive idea is performed by

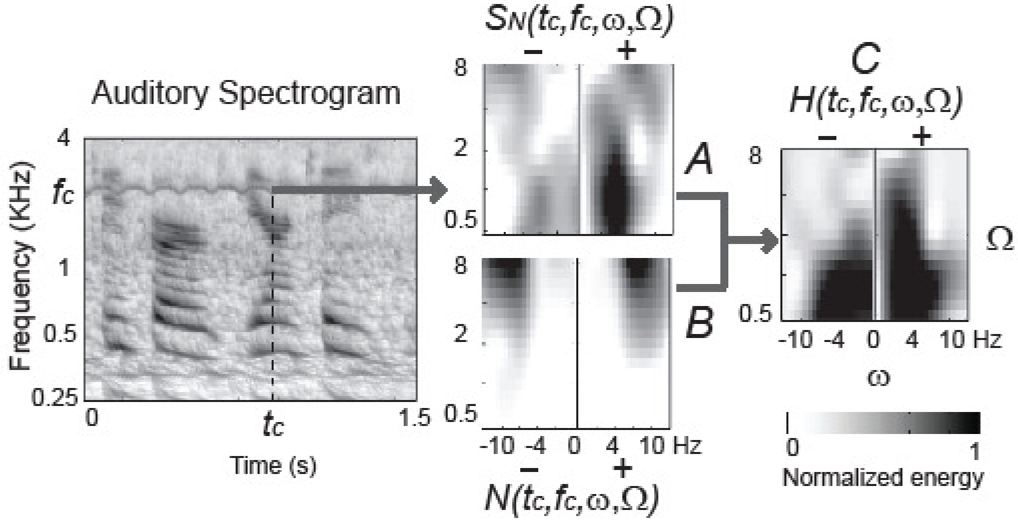

Figure 45. **Filtering the rate-scale representation**: modulations due to the noise are filtered out by weighting the rate-scale representation of noisy speech with the function *H*(*t*, *f* ,*ω*,Ω). In this example, the jet one noise from Noisex was added to clean speech at SNR 10 dB. The rate-scale representation of the signal, *rs*(*tc* , *fc* ,*ω*,Ω) and the rate-scale representation of noise, *N*(*tc* , *fc* ,*ω*,Ω) were used to obtain the necessary weighting as a function of *ω* and Ω (11). This weighting was applied to the rate-scale representation of the signal, *rs*(*tc* , *fc* ,*ω*,Ω) to restore modulations typical of clean speech. The restored modulation coefficients were then used to reconstruct the cleaned auditory spectrogram, and from it the corresponding audio signal.

the Wiener filter which for this example results in the function shown in **Figure 45**C. The *H* function has low gain for fast modulation rates and high scales that are due to the background noise (as shown in **Figure 45**B), while emphasizing the slow modulations (<5 Hz) and low scales (<2 cyc/oct) that come mostly from speech signal. Multiplication of this rate-scale-frequency gain which is a function of time, and the noisy speech representation results in denoised representation which is then used to reconstruct the spectrogram of the cleaned speech signal using the inverse cortical transformation (**Figure 46**).

6.3.4 Results from experimental evaluations

To examine the effectiveness of the noise suppression algorithm, we used subjective and objective tests to compare the quality of denoised signal with the original and a Wiener filter noise suppression method by Scalart and Filho [101] implemented in [102]. The noisy speech sentences were generated by adding four different kinds of noise: white, jet, babble, and city from Noisex [100] to eight clean speech samples from TIMIT [24]. The test material was prepared at three SNR values: 0, 6, and 12 dB. We used mean opinion score (MOS) test to evaluate the subjective quality of the denoising algorithm. In the subjective quality tests, ten subjects were asked to score

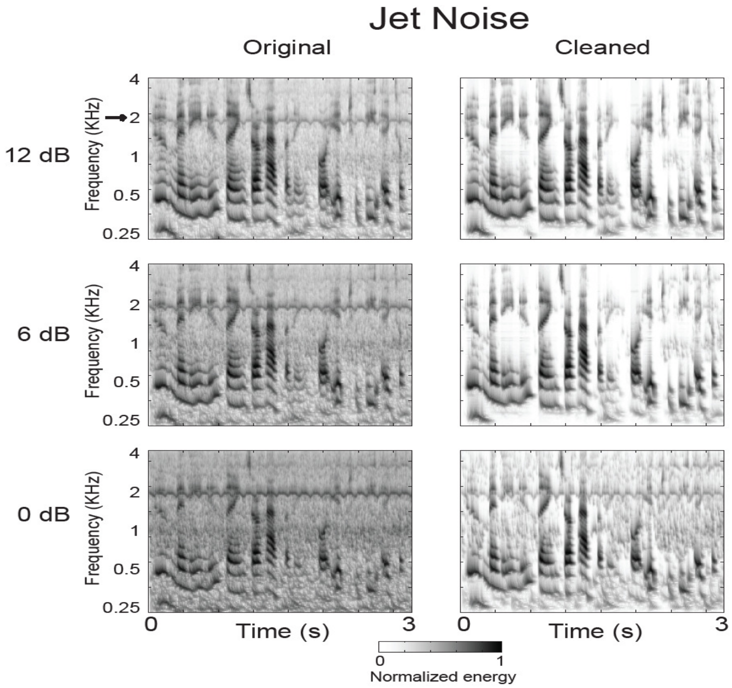

Figure 46. Examples of restored spectrograms after "filtering" of spectrotemporal modulations. Jet noise from Noisex was added to speech at SNRs 12 dB (top), 6 dB (middle) and 0 dB (bottom) panels. Left panels show the original noisy speech and right panels show the denoised ones. The clean speech spectrum has been restored although the noise has a strong temporally modulated tone (10 Hz) mixed in with the speech signal near 2 kHz (indicated by the arrow).

the quality of the original and denoised speech samples between one (bad) and five (excellent). All subjects had prior experience in psychoacoustics experiments and had self-reported normal hearing. The sounds were presented in a quiet room over headphones at a comfortable listening level (approximately 70 dB) and the responses were collected using a computer interface. **Figure 47** shows the MOS score and the errorbars for the original and denoised signals using modulation and Wiener methods. The results are shown for four types of noise and three SNR levels. In most stationary noise conditions, subjects reported the highest scores for the modulation method.

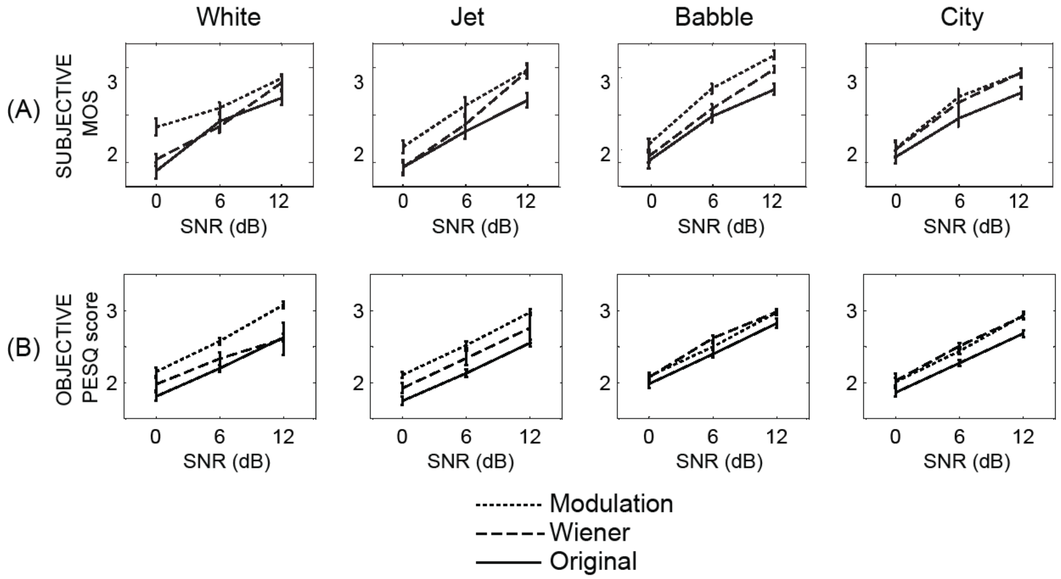

Figure 47. Subjective and objective scores on a scale of 1 to 5 for degraded and denoised speech using modulation and Wiener methods. (a): Subjective MOS scores and errorbars averaged over ten subjects for white, jet, babble, and city noise. (b): Objective scores and errorbars transformed to a scale of 1 to 5 for degraded and denoised speech using modulation and Wiener methods.

However, for the nonstationary sounds, the modulation method outperformed the Wiener methods in the babble tests, and produced comparable results for the city sounds. In addition, we conducted objective test using perceptual evaluation of speech quality (PESQ) [103] measure for the twelve conditions to obtain the objective score for each sample. The resulting scores and their errorbars are reported in **Figure 47b**. PESQ gives higher scores for the modulation method in the stationary conditions, but the performance in this measure appears comparable for the nonstationary conditions. Our method performs better for stationary noise because of its ability to model the average spectrotemporal properties of the stationary noise better. This also explains the better performance in the babble speech since the babble is relatively “stationary” in its long-term spectrotemporal behavior, especially compared to the city noise which fluctuates considerably.

6.3.5 Conclusions

We have described a new approach for the denoising of contaminated broadband complex signals such as speech. In this method, the noisy signal is first transformed to the spectrotemporal modulation domain in which the speech and noise are separated based on their distinct modulation patterns.

This allows for the possibility of suppressing noise even when it spectrally overlaps with the desired signal. The spectrotemporal representation used is based on a model of auditory processing inspired by physiological data from the mammalian primary auditory cortex. Subjective and objective tests are reported that they demonstrate the effectiveness of this method in enhancing the quality of speech without introducing artifacts or substantially deleting spectrally overlapping speech energy.

Chapter 7

Conclusions

7.1 Thesis overview

This thesis is an attempt to fill the gap between what we know about the physiology of speech and engineering models that can be applied to speech processing systems. To investigate the neural basis of speech perception, we observed how neuronal responses to continuous speech in the primary auditory cortex of the naive ferret reveal an explicit multidimensional representation that is sufficiently rich to support the discrimination of many American English phonemes. This representation is made possible by the wide range of spectro-temporal tuning in A1 to stimulus frequency, scale and rate. The great advantage of such diversity is that there is always a unique sub-population of neurons that responds well to the distinctive acoustic features of a given phoneme and hence encodes that phoneme in a high-dimensional space. In addition, using a method of stimulus reconstruction from the population of neurons in the auditory cortex, we showed how we can investigate the neural code for speech, and observed a remarkable robustness of the cortical representation to noise and distortions.

We then explored the efficacy of a simple spectro-temporal receptive field model of auditory cortical neurons and showed that this model is capable of predicting the selectivity of auditory cortical neurons to phoneme categories. This is an important step toward building systems that use this knowledge. In fact, we showed how systems that use such a representation can achieve state of the art performance operating better than many traditional methods.

7.2 Future directions

We have focused here on describing a few prominent features of the response distributions that correspond to well-known distinctive acoustic features of the consonants considered. There are clearly many other aspects and more details of the responses that reflect intricate articulatory gestures, contextual effects, or speaker-dependent variability that can only be reliably considered with a much larger sample of responses. The representation of phonemic features across a population of filters tuned to BF, scale and rate suggests a strategy for improved speech recognition systems, and further study may reveal additional strategies for speech processing. However, many questions about the neural representation of phonemes still remain unclear; for example, how can one extrapolate from such neurophysiological findings to the human perceptual ability to perceive phonemes categorically and to shift categorical boundaries arbitrarily between phoneme pairs. In addition, although we showed how the linear spectrotemporal models predict the selective representation observed in the auditory neurons, the current models are unable to explain the noise robustness of the representation observed in the cortex. In order to enhance the performance of speech processing systems, we need to understand the mechanism and the theory behind this noise robustness which is likely to be the result of nonlinear adaptation in the neurons.

Bibliography

1 Lippmann, R. P., Speech recognition by machines and humans, Speech Commun., vol. 22, pp. 1-15, (1997).

2 Greenberg, S., Ainsworth, W., Popper, A.N., & Fay, R.R., Speech Processing in the Auditory System, Springer-Verlag, New York, Volume 18, (2004).

3 Ladefoged, P., A course in phonetics. Orlando: Harcourt Brace. 5th ed. Boston: Thomson/Wadsworth (2006).

4 Quatieri, T. F., Discrete-time speech signal processing: principles and practice, Prentice Hall publication in Prentice Hall Signal Processing Series, (2002).

5 Kuhl, P.K. & Miller, J.D., Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants, Science, Vol 190, Issue 4209, pp. 69-72 (1975).

6 Kuhl, P.K & Padden, D.M., Enhanced discriminability at the phonetic boundaries for the place feature in macaques, J. Acoust Soc Am. Mar;73(3): pp. 1003-10, (1983).

7 Kuhl, P.K. & Padden, D.M., Enhanced discriminability at the phonetic boundaries for the place feature in macaques, J Acoust Soc Am. Mar;73 (3): pp. 1003-10, (1983)

8 Kluender, K.R., Lotto, A.J., Holt, L.L. & Bloedel, S.L., Role of experience for language-specific functional mappings of vowel sounds. J Acoust Soc Am. Dec;104(6): pp. 3568-82, (1998).

9 Pons, F., The effects of distributional learning on rats' sensitivity to phonetic information, J Exp Psychol Anim. Behav. Process. 32(1): pp. 97-101, (2006).

10 Hienz, R.D., Aleszczyk, C.M. & May, B.J., Vowel discrimination in cats: acquisition, effects of stimulus level, and performance in noise, J Acoust. Soc Am. 99(6): pp. 3656-68, (1996).

11 Dent, M.L., Brittan-Powell, E.F., Dooling, R.J. & Pierce, A., Perception of synthetic /ba/-/wa/ speech continuum by budgerigars (Melopsittacus undulatus), J Acoust Soc Am. Sep;102(3): pp. 1891-7, (1997).

12 Lotto, A.J., Kluender, K.R. & Holt, L.L., Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica), J Acoust Soc Am. Aug;102(2 Pt 1): pp. 1134-40, (1997).

13 Steinschneider, M., Fishman, Y.I. & Arezzo, J.C., Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey, J Acoust Soc Am. 114(1): pp. 307-21, (2003).

14 Steinschneider, M., Reser, D., Schroeder, C.E., Arezzo, J.C. Tonotopic organization of responses reflecting stop consonant place of articulation in primary cortex (A1) of the monkey.Brain Research, 674, pp. 147-152, (1995).

15 Steinschneider, M., Volkov, I.O., Fishman, Y.I., Oya, H., Arezzo, J.C., Howard, M.A., Intracortical responses in human and monkey auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter. Cerebral Cortex, 15, pp. 170-186, (2005).

16 Eggermont, J.J. & Ponton, C.W., The neurophysiology of auditory perception: from single units to evoked potentials. Audiol Neurootol. Mar-Apr; 7(2): pp. 71-99. Review, (2002).

17 Hung, C.P., Kreiman, G.K., Poggio, T., & DiCarlo, J.J., Fast readout of object identity from macaque inferior temporal cortex. Science 310, 863-866 (2005).

18 Walker, K., King, A., Ahmed, B. & Schnupp, J. W. H, Psychometric and neurometric discrimination of non-conspecific vocalizations, Abstract 430, MidWinter Meeting of Association for Research in Otolaryngology, Baltimore, (2006).

19 Miller, G. & Nicely, P., An analysis of perceptual confusions among some English consonants," J. Acoustical Society America, vol. 27, pp. 338-352, (1955).

20 Ladefoged, P., A course in phonetics. Orlando: Harcourt Brace. 5th ed. Boston: Thomson/Wadsworth (2006).

21 Stevens, K. N., Acoustic Phonetics, The MIT Press, Cambridge, MA, (1980)

22 Theunissen, F.E., David, S.V., Singh, N.C., Hsu A., Vinje, W.E. & Gallant, J.L., Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli, 1: Network. 12(3): pp. 289-316, (2001).

23 Klein, D.J., Simon, J. Z., Depireux, D. A. & Shamma, S. A., Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex, J Comput Neurosci., 20(2): pp. 111-36, (2006).

24 Seneft. S., & Zue, V, Transcription and alignment of the timit database", J. S. Garofolo, Ed. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, (1988).

25 Yang, X., Wang, K., & Shamma, S. A, Auditory representation of acoustic signals, IEEE Trans. Inf. Theory, 38 (2), pp. 824-839, (Special issue on wavelet transforms and multi-resolution signal analysis), (1992).

26 David, S. V & Gallant, J.L., Predicting neuronal responses during natural vision. Network, 16(2-3): pp. 239-60, (2005).

27 Vapnik, V. N, The Nature of Statistical Learning Theory, Springer, (1995).

28 Miller, G. & Nicely, P., An analysis of perceptual confusions among some English consonants," J. Acoustical Society America, vol. 27, pp. 338-352, (1955).

29 Shamma, S., Speech Processing in the Auditory System. Part I: The Representation of Speech Sounds in the Responses of the Auditory-Nerve, J. Acoust. Soc. Am. 78(5), pp. 1612-1621, (1985).

30 Young ED, Sachs MB (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. J Acoust Soc Am 66:1381–1403.

31 Schreiner, C. E., Read, H. L. & Sutter M.L, Modular organization of frequency integration in primary auditory cortex. Annu Rev Neurosci.;23: pp. 501-29, Review (2000).

32 Read, H.L., Winer, J.A. & Schreiner, C. E., Functional architecture of auditory cortex. Curr Opin Neurobiol., Aug; 12(4): pp. 433-40, Review (2002).

33 Depireux, D. A., Simon, J. Z., Klein, D.J & Shamma, S. A., Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex, Journal of Neurophysiology, 85, pp. 1220-1234, (2001).

34 Chistovich, L. A., and Lublinskaya, V. V. The `center of gravity effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli, Hear. Res., 185-195, (1979)

35 Klein, W., Plomp, R., & Pols, L. C., Vowel spectra, vowel spaces and vowel identification, The Journal of the acoustical society of America, 48 (4), pp. 999-1009, (1970).

36 Deshmukh, O., Espy-Wilson, C., Salomon .A and Singh, J., Use of Temporal Information: Detection of the Periodicity and Aperiodicity Profile of Speech, IEEE Transactions on Speech and Audio Processing, Vol. 13 (5), pp. 776-786, Sept. (2005)

37 Bendor, D. and Wang, The neuronal representation of pitch in primate auditory cortex, Nature 436, 1161–1165 (2005).

38 Allen, J. B., Articulation and intelligibility, Morgan & Claypool Publishers, (2005).

39 Depireux D., Simon, J. Z and Shamma S., Measuring the dynamics of neural responses in primary auditory corte, Comments in Theoretical Biology, 5(2), 89-118, (1998).

40 Kowalski N., Depireux D., Shamma S., Analysis of dynamic spectra in ferret primary auditory cortex: Prediction of single-unit responses to arbitrary dynamic spectra, J. Neurophysiology, 76(5) 3524-3534, (1996).

41 Miller, L. M., Escabi, M. A, Read, H. L, Schreiner, C. E., Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol. Jan; 87(1):516-27, (2002).

42 Novitski, C. T. et al., Program 800.18 / Poster E45, Neural coding of speech sounds in naïve and trained rat primary auditory cortex, Society for Neuroscience, Atlanta (2006).

43 De Valois RL, Yund E.W, Hepler N., The orientation and direction selectivity of cells in macaque visual cortex, Vision research, 22(5): 531-44, 1982.

44 Ferragamo, M. J., Haresign, T., Simmons, J. A., Frequency tuning, latencies, and responses to frequency-modulated sweeps in the inferior colliculus of the echolocating bat, Eptesicus fuscus, J. Com Physiology, 182(1): 65-79, Jan 1998.

45 Sachs M, Young E (1979): "Encoding of steady state vowels in the auditory nerve: Representation in terms of discharge rate", J. Acoust. Soc. Am. 66, 470–479.

46 Mesgarani N, David SV, Fritz JB, Shamma SA., Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am. 2008 Feb; 123(2):899-909.

47 Schreiner C, Langner G (1988):"Periodicity coding in the inferior colliculus of the cat. II. Topographical organization", J. Neurophysiol. 60, 1823–1840.

48 Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D, Reading a neural code, Science. 1991 Jun 28;252(5014):1854-7.

49 De Ruyter van Steveninck RR, Lewen GD, Strong SP, Koberle R, Bialek W., Reproducibility and variability in neural spike trains, Science. 1997 Mar 21;275(5307):1805-8.

50 Haag J, Borst A., Active membrane properties and signal encoding in graded potential neurons., J Neurosci. 1998 Oct 1;18(19):7972-86.

51 Buracas GT, Zador AM, DeWeese MR, Albright TD. Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. Neuron. 1998 May;20(5):959-69

52 Rieke F, Bodnar DA, Bialek W., Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents, Proc Biol Sci. 1995 Dec 22;262(1365):259-65.

53 Warland DK, Reinagel P, Meister M., Decoding visual information from a population of retinal ganglion cells. J Neurophysiol. 1997 Nov;78(5):2336-50

54 Stanley GB, Li FF, Dan Y., Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. J Neurosci. 1999 Sep 15;19(18):8036-42

55 Fritz J, Shamma S, Elhilali M, Klein D., Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex, Nat Neurosci. 2003 Nov;6(11):1216-23. Epub 2003 Oct 28

56 Klein, D.J., Depireux, D.A., Simon, J.Z. & Shamma, S.A. Robust spectro-temporal reverse correlation for the auditory system: optimizing stimulus design. J. Comput. Neurosci. 9, 85-111 (2000).

57 Mesgarani N, David SV, Fritz JB, Shamma SA., Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am. 2008 Feb; 123(2):899-909.

58 Xiaowei Yang, Kuansan Wang, Shihab A. Shamma: Auditory representations of acoustic signals. IEEE Transactions on Information Theory 38(2): 824-839 (1992)

59 Fritz J. B., Bozak D., Depireux D. A., Dobbins H., Tillman A., Shamma S. A, Measuring the ferret spectro-temporal transfer function (MTF) using a conditioned behavioral task, Abstract #802, 2002, ARO meeting, Jan 30, 2002.

60 Eggermont, J.J. & Ponton, C.W., The neurophysiology of auditory perception: from single units to evoked potentials. Audiol Neurootol. Mar-Apr; 7(2): pp. 71-99. Review, (2002).

61 Chi T, Ru P, Shamma SA., Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am. 2005 Aug; 118(2):887-906.

62 Engineer C. T., Perez C. A., Chen Y. H., Carraway R. S., Reed A. C., Shetake J. A., Jakkamsetti V., Chang K. Q., Kilgard M. P., Cortical activity patterns predict speech discrimination ability, Nature Neuroscience 11, 603 - 608 (2008)

63 James A. Mazer, William E. Vinje, Josh McDermott, Peter H. Schiller, and Jack L. Gallant, Spatial frequency and orientation tuning dynamics in area V1, Proc Natl Acad Sci U S A. 2002 February 5; 99(3): 1645–1650.

64 Wallis JD, Miller EK, Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. European Journal of Neuroscience 18:, 2003

65 K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," IEEE Trans. Speech Audio Process., vol. 3, no. 5, pp 382–395, Sep. 1995.

66 S. A. Shamma, Methods of neuronal modeling, In: Spatial and temporal processing in the auditory system, second edition, MIT press, Cambridge, MA, pp. 411-460, 1998.

67 M. Elhilali, T. Chi, and S. A. Shamma, A spectro-temporal modulation index (STMI) for assessment of speech intelligibility, Speech communication, vol. 41, pp. 331-348, 2003.

68 S. Pfeiffer, S. Fischer, and W. Efferlsberg, Automatic audio content analysis, in proc. 4th ACM International Multimedia Conference, pp.21-30, 1996.

69 E. Wold, T. Blum, and D. Keislar et al., Content-based classification, search, and retrieval of audio, IEEE multimedia, pp. 27-36, Fall 1996.

70 J. Foote, Content-based retrieval of music and audio, In C.-C. J. Kuo et al., editor, Multimedia Storage and Archiving Systems II, Proc. of SPIE, Vol. 3229, pp. 138-147, 1997.

71 L. Lu, H. Zhang, and H. Jiang, Content analysis for audio classification and segmentation, IEEE Transaction on Speech,Audio and Signal Processing, Vol. 10, No. 7, 2002.

72 J. Saunders, Real-time discrimination of broadcast speech/music, Proc. International Conference on Acoustic, Speech and Signal Processing, vol. II, Atlanta,

GA, pp. 993-996, May 1996

73 Christopher J. C. Burges, J. C. Platt, and S. Jana, Distortion Discriminant Analysis for Audio Fingerprinting, IEEE Transaction on Speech and Audio Processing, Vol. 11, No. 3, May 2003

74 E. Scheirer, and M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator, International Conference on Acoustic, Speech and Signal Processing, Munich, Germany, 1997.

75 B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system, International Conference on Acoustic, Speech and Signal Processing, vol. I, Orlando, FL, pp. 53-56, May 2002

76 R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, 1961

77 L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM Journal of Matrix Analysis and Applications, vol. 21, pp. 1253-1278, 2000.

78 L. R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika, vol. 31, pp. 279-311, 1966.

79 P. M. Kroonenberg, Three-mode principal component analysis, DSWO Press, Leiden, Netherlands, 1982.

80 L. De Lathauwer, B. De Moore, and J. Vandewalle. On the best rank-1 and rank-(R1,R2, ...,RN) approximation of higher order tensors, SIAM Journal of Matrix Analysis and Applications, vol. 21(4):1324-1342, 2000.

81 L. De Lathauwer, B. De Moor, J. Vandewalle, Dimensionality reduction in higher-order-only ICA, IEEE signal processing workshop on Higher Order Statistics, Banff, Alberta, Canada, pp.316-320, 1997.

82 M. A. O. Vasilescu and D. Terzopoulos, Multilinear analysis of image ensembles: TensorFaces, European Conference on Computer Vision, Copenhagen, Denmark. pp. 447-460, May 2002.

83 M. A. O. Vasilescu and D. Terzopoulos, Multilinear subspace analysis of image ensembles, IEEE Conference on Computer Vision and Pattern Recognition, Madison, WI, June 2003.

84 V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

85 T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning,B. Scholkopf, C. Burges and A. Smola (ed.), MIT-Press, 1999

86 BBC Sound Effects Library, Original Series, 40 Audio CD Collection, Distributed by Sound Ideas, (c) 1984

87 M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, RWC Music Database: Music Genre Database and Musical Instrument Sound Database, International Conference on Music Information Retrieval, pp. 229230, 2003.

88 A. Varga, H .J .M Steenneken, M. Tomlinson, and D. Jones, The NOISEX-92 study on the effect of additive noise on automatic speech recognition, Documentation included in the NOISEX-92 CD-ROMs, 1992.

89 H. G. Hirsch, and D. Pearce, The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions,

ISCA ITRWASR2000 Automatic Speech Recognition: Challenges for the Next Millennium;Paris, France, September 18-20, 2000.

90 N. Mesgarani and S. A. Shamma, Speech enhancement base on filtering the spectrotemporal modulations, International Conference on Acoustic, Speech and Signal processing, Philadelphia, PA, March 2005.

91 M. Elhilali, T. Chi, and S. A. Shamma, A spectro-temporal modulation index (STMI) for assessment of speech intelligibility, Speech communication, vol. 41, pp. 331-348, 2003.

92 R. P. Carlyon, S. A. Shamma, An account of monaural phase sensitivity, Journal of Acoustic Society of America, vol. 114(1), pp. 333-48, 2003.

93 L. Atlas, S. A. Shamma, Joint acoustic and modulation frequency, Eurasip Journal on Applied Signal Processing, No. 7, pp. 668-675, June 2003.

94 J. S. Lim, A. V. Oppenheim, ”Enhancement and bandwith compression of noisy speech”, Proc. IEEE, Vol 67, pp.1586-1604, Dec. 1979.

95 Y. Ephraim, H. L. Van Trees, ”A signal subspace approach for speech enhancement”, IEEE Trans. Speech and Audio Proc., Vol 3, pp.251-266, July 1995.

96 Y. Ephraim, D. Malah, ”Speech enhancement using a minimum mean square error Log-spectra amplitude estimator”, IEEE Trans. Acoust., Speech and Signal Proc., vol. ASSP-33, pp. 443-445, Apr. 1985.

97 R. Martin, ”Statistical methods for the enhancement of noisy speech”, Inter. Workshop on Acoust. Echo and Noise Control, Kyoto, Japan, Sept. 2003.

98 S. Shamma, ”Encoding sound timbre in the auditory system”, IETE J. Res. 49(2), 193-205, 2003.

99 J. Tchorz and B. Kollmeier, ”SNR estimation based on amplitude modulation analysis with application to noise suppression”, IEEE trans. Speech and Audio Proc., Vol. 11, No.3, May 2003

100 A. Varga, H .J .M Steeneken, M. Tomlinson, and D. Jones, ”The NOISEX-92 study on the effect of additive noise on automatic speech recognition”, Documentation included in the NOISEX-92 CD-ROMs, 1992.

101 P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’96), vol. 2, pp. 629–632, Atlanta, Ga, USA,May 1996.

102 E. Zavarehei, http://dea.brunel.ac.uk/cmsp/Home Esfandiar

103 “Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” ITU-T Recommendation P.862, February 2001.