Investigating the neural code for dynamic speech and the effects of signal degradation

Mark Steadman, BEng.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

July 2015

Abstract

It is common practice in psychophysical studies to investigate speech processing by manipulating or reducing spectral and temporal information in the input signal. Such investigations, along with the often surprising performance of modern cochlear implants, have highlighted the robustness of the auditory system to severe degradations and suggest that the ability to discriminate speech sounds is fundamentally limited by the complexity of the input signal. It is not clear, however, how and to what extent this is underpinned by neural processing mechanisms.

To date, electrophysiological investigations into the neural processing of speech have predominantly utilised natural sounds. However, interpretation of these data is often difficult as it necessitates assumptions made about salient features of a stimulus set. Difficulties are further compounded when electrical hearing is taken into consideration, whereby many of the cues often assumed to be important are absent yet in many cases listeners perform speech recognition without difficulty. Human psychophysical studies using parametrically degraded speech sounds constitute a powerful toolset to explore the neural bases of speech perception.

The approach taken in this thesis was to examine the effect on the neural representation of reducing spectral and temporal information in the signal. A stimulus set from an existing psychophysical study was emulated, comprising a set of 16 vowel-consonant-vowel phoneme sequences (VCVs) each produced by multiple talkers, which were parametrically degraded using a noise-vocoder. Neuronal representations were simulated using a published computational model of the auditory nerve. Representations were also recorded in the inferior colliculus (IC) and auditory cortex (AC) of anaesthetised guinea pigs. Their discriminability was quantified using a novel nearest-neighbour classifier.

The VCVs elicited highly discriminable representations at each of the brain regions. Commensurate with human perception, this discriminability is robust to severe degradation of the stimulus. The discriminability of representations in the auditory nerve and the IC is more affected by the envelope bandwidth of the stimulus than is observed in human listeners. High rate envelope cues (>16 Hz) increased the discriminability of neural representations in the auditory nerve such that near-perfect discrimination of the 16 medial consonants was possible, even where only one spectral channel was used in the vocoder. Human listeners receive little benefit from such cues, and rely on the presence of spectral information for speech recognition performance comparable to that for natural speech. This is reflected in cortical representations of heavily spectrally degraded speech, which do not become significantly more discriminable with high rate envelope cues. Conversely, the classifier was not able to benefit to a large extent from the availability of rich spectral cues in peripheral auditory regions.

It appears then, that consistent with investigations using simple stimuli, high rate envelope modulations are represented in the auditory nerve and midbrain. It is demonstrated here that the representations of such cues could benefit human listeners in a closed-set speech recognition task where appropriate decoding mechanisms are available. Optimal encoding windows for speech discrimination increase from of the order of 1 millisecond in the auditory nerve to 10s of milliseconds in the IC and the AC. Recent publications suggest that millisecond-precise neuronal activity is important for speech recognition. It is demonstrated here that the relevance of millisecond-precise responses in this context is highly dependent on the brain region, the nature of the speech recognition task and the complexity of the stimulus set.

Acknowledgments

I would first like to thank my supervisor, Chris Sumner, for allowing me the opportunity to complete this project and facilitating the precipitous traverse from hardware and software to wetware. I would also like to thank Alan Palmer, Trevor Shackleton and Mark Wallace who designed, built and maintain the equipment that made the experiments herein possible and have provided invaluable advice along the way.

I would also like to thank my peers and colleagues who made my time in Nottingham an enjoyable one, including those who have now gone to various corners of this country and others: David Green, Mark Fletcher, Helen Nuttall, Heather Gilbert, Pete Jones, Kate Molloy, Jess Monaghan, Joe Sollini, & Ian Wiggins. Kudos in particular to those who were around whilst I was compiling this thesis and still put up with me now: Toby Wells, Joel Berger, Chris Scholes, Gemma Hutchinson & Colin Horne.

I would like to thank both of my parents for all manner of support through what seems now like an inordinately long time not going and getting a “proper job”. Special thanks to Kari Vyas for being generally amazing.

Contents

Glossary

AC	Auditory cortex
AI	Primary auditory cortex
AVCN	Anteroventral cochlear nucleus
BF	Best frequency
BMF	Best modulation frequency
CF	Characteristic frequency
CI	Cochlear implant
CIS	Continuous interleaved ampling
CN	Cochlear nucleus
CNIC	Central nucleus of the inferior colliculus
CSF	Cerebrospinal fluid
DCN	Dorsal cochlear nucleus
DNLL	Dorsal nucleus of the lateral lemniscus
DRNL	Dual resonance non-linear
FRA	Frequency response area
IC	Inferior colliculus
LSO	Lateral superior olive

MGBd Dorsal nucleus of the medial geniculate body

MGBm Medial nucleus of the medial geniculate body

MGBv Ventral nucleus of the medial geniculate body

MNTB Medial nucleus of the trapezoid body

MSO Medial superior olive

MTF Modulation transfer function

NLL Nucleus of the lateral lemniscus

PSTH Peri-stimulus time histogram

PVCN Posterventral cochlear nucleus

RLF Rate-level function

rMTF Rate modulation transfer function

RP Receptor potential

SAM Sinusoidally amplitude modulated

SNHL Sensory-neural hearing loss

SNR Signal to noise ratio

SOC Superior olivary complex

SR Spontaneous rate

TFS Temporal fine structure

tMTF Temporal modulation transfer function

VCV Vowel-consonant-vowel phoneme sequence (e.g. */aca/)*

VNLL Ventral nucleus of the lateral lemniscus

Chapter 1
General Introduction

Communication can be defined as the transmission of "meaningful" information from sender to receiver (Bradbury and Vehrencamp, 1998). In order to achieve this, a communications system must employ an encoding scheme that utilises a number of symbols that can be both reliably produced and robustly discriminated after transmission. In spoken language, these symbols are constructed from phonemes; fundamental units of speech sounds that can be concatenated together to form lexemes containing semantic value.

The production of speech sounds involves complex interactions between auditory, motor and somatosensory systems. A vast diversity of phonemes can be produced by, for example, altering the frequency response characteristics of the vocal tract by manipulations of the tongue or restriction of air flow by the teeth and lips. These are, in the simplest case, transmitted through the air as a series of compressions and rarefactions to the recipient's ears where they are transformed into spatio-temporal patterns of electrical activity in the auditory system, which are then decoded.

Speech is a spectrally and temporally rich acoustical signal, but the limitations of cochlear implants show that the auditory system appears to be remarkably robust to severe degradations in either domain (Wilson and Dorman, 2007). However, there is a high degree of variability in the outcomes of these devices (Dorman and Spahr, 2006), which is often attributed to the particular aetiology of deafness or the efficaciousness of the implantation procedure (Holden et al., 2013) leading to further degradation of the input signal. One of the major challenges that designers of auditory prostheses face is understanding how degradation affects neural representations of complex sounds and how this influences speech perception.

This thesis aims to investigate the underlying neural bases of speech discrimination that are robust to signal degradations analogous to those inherent in contemporary auditory prostheses. The mammalian auditory system, and what is known about how it processes complex sounds, will first be discussed. Common causes of signal degradation at the auditory periphery, auditory prostheses, and their effect on speech perception will also be introduced.

1.1 The mammalian auditory system

1.1.1 The outer and middle ear

Acoustic signals enter the ear via the pinna and external auditory meatus, or ear canal, where they stimulate the tympanic membrane. The respective shapes of these structures result in a frequency-specific gain to these signals,

which is also direction-dependent, thereby providing an important spectral cue for source localisation (Wightman and Kistler, 1993).

The tympanic membrane converts the pressure fluctuations into a mechanical vibration, which is transmitted via three small bones (the ossicles) to the oval window of the inner ear. This mechanism allows for efficient transmission of acoustic energy from the air to the fluid-filled cochlear, which would otherwise transmit very little energy due to the mismatched acoustic impedance of the gaseous and liquid media. This apparatus also provides a means of controlling input gain at lower frequencies (below 1-2 kHz) via the middle ear muscles (Pickles, 1988).

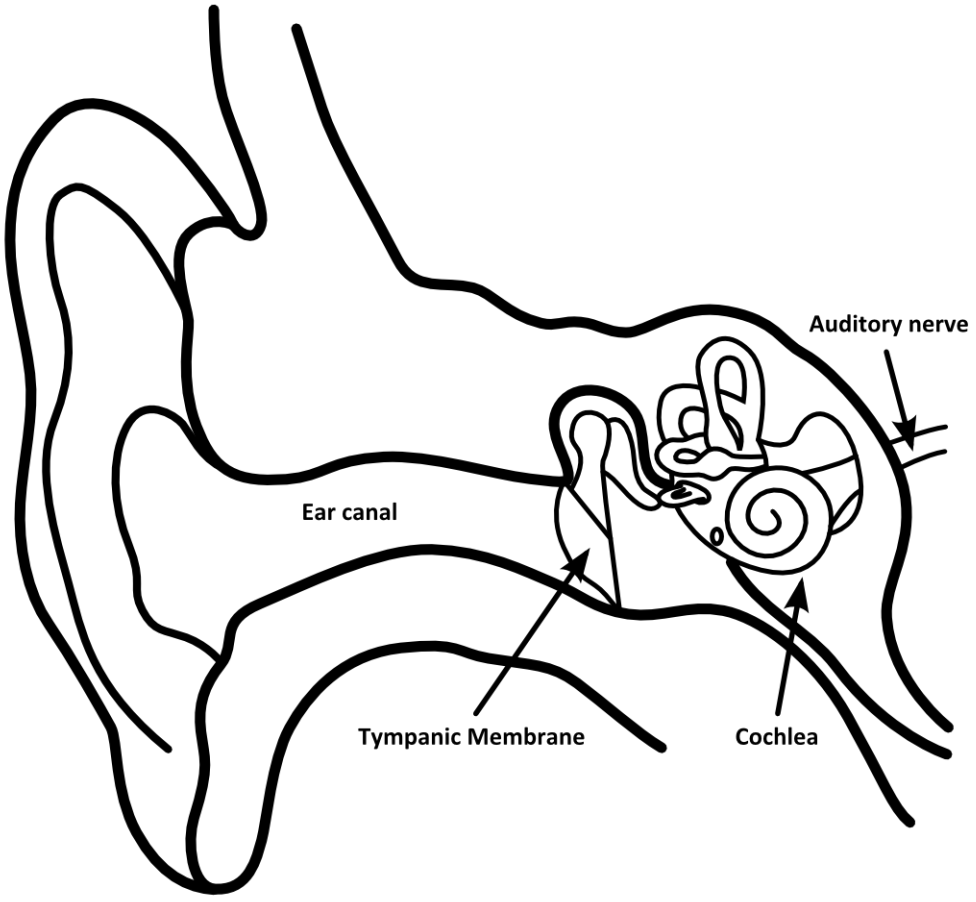

Figure 1.1: The human peripheral auditory system.

1.1.2 The cochlea

The cochlea performs the transduction of this mechanical energy to electrical signals that convey auditory information to the central nervous system via the auditory nerve. It also performs a mechanical spectral decomposition of the acoustic signal, which provides the basis of the spatial representation of sound frequency, or tonotopicity, in the auditory system. Early experiments in cochlear physiology demonstrated that the passive mechanics of the cochlea result in a travelling wave of acoustic energy that propagates through the cochlear fluids away from the oval window and peaks in a frequency-specific region (von Békésy, 1947).

The cochlear is divided internally by the Reissner's and basilar membranes. The Reissner's (or vestibular) membrane separates the perilymph-filled scala vestibuli from the endolymphatic scala media. This is in turn separated from the scala tympani by the basilar membrane (see Figure 1.2 A). The basilar membrane tapers from its narrowest at the basal end of the cochlear spiral to its broadest at the apical end, the extra space being taken up by the spiral lamina; a bony projection from the central axis of the cochlea.

The transduction of acoustic energy to electrical signals is performed by the organ of Corti situated on the basilar membrane (see Figure 1.2 B). The auditory receptor cells, known as hair cells, are contained within this organ, which is covered by the tectorial membrane. Motion of the basilar membrane, initiated by pressure differences between the scala media and the scala vestibuli created

by acoustic energy passing along the cochlea, causes a shearing motion between the tectorial membrane and these hair cells.

The inner hair cells are situated in the organ of Corti towards the central axis of the cochlea. It is the deflection of the “hairs”, or stereocilia, by movement of the surrounding endolymphatic fluid that results in the modulation of ion flow and therefore transduction currents and receptor potentials in the hair cells (Robles and Ruggero, 2001; see 2.1.1.3).

The outer hair cells are situated laterally to the inner hair cells in rows of between three and five, separated from them by the tunnel of Corti; a gap between supporting pillar cells. They are thought to be responsible for the discrepancy between early basilar membrane recordings in cadavers and frequency selectivity measured psychophysically, which is more acute. They are thought to do this by introducing an active component into the cochlear mechanics via a process referred to as reverse tranduction; the conversion of electrical signals to a mechanical output (Fettiplace and Hackney, 2006).

Approximately 30,000 afferent neurons innervate the human cochlea. About 90% of these are myelinated, type I, neurons, which terminate on a single inner hair cell in the organ of Corti. Each inner hair cell synapses with several such fibres, the number of which varies from as low as 3-4 in the basal and apical turns and as many as 15 on average elsewhere. The remaining 10% are unmyelinated (type II) and form synapses with several (10-20) outer hair cells (Spoendlin and Schrott, 1989).

The cochlear is also innervated by centrifugal neural processes via the olivocochlear bundle originating in the superior olivary complex in the brainstem. These efferent fibres form synapses with afferent fibres as well as the outer hair cells, and have been shown to modulate afferent signalling (Warren III and Liberman, 1989), although the implications of this functionality are still poorly understood and are outside of the scope of this thesis. Subsequent discussion will therefore be concerned only with afferent responses.

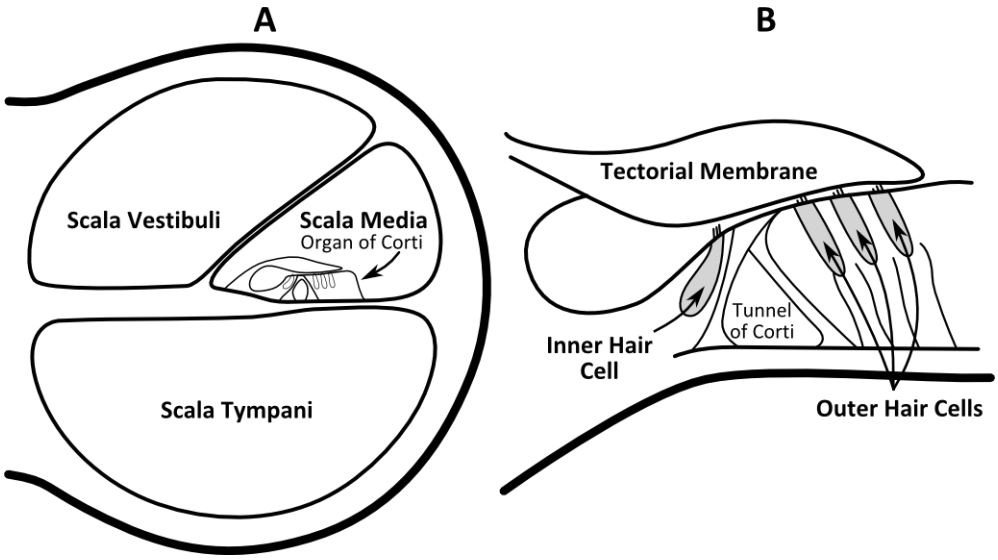

Figure 1.2: Cross-section of the mammalian cochlea (**A**) and an expanded view of the organ of Corti (**B**).

1.1.3 Auditory nerve

For over half a century, it has been possible to record from the cell bodies and fibres of single neurons in the auditory nerve using microelectrodes (Tasaki, 1954). Early recordings showed that irregular, spontaneous spiking activity occurs in the absence of acoustic stimulation. The distribution of spontaneous

firing rates is essentially bimodal (Kiang, 1965), though fibres are commonly divided into three groups with “low-”, “medium-” and “high-” spontaneous rates (SR) (Liberman, 1978) corresponding to fewer than 0.5, 0.5 to 18 and more than 18 spikes per second respectively. The distinction between low and medium SR fibres is only apparent in the context of responses to auditory stimuli as described below.

The simplest form of acoustic stimulus is the pure tone; a sinusoidal acoustic waveform. These stimuli have ostensibly a single frequency component (although in practice there is always a degree of spectral splatter due to stimulus onset and offset), and so are a useful tool for investigating fundamental properties of auditory responses. When presented alone, tonal stimuli are always excitatory in the auditory nerve. The stimulus level at which a fibre’s firing rate is measurably elevated is its threshold. Fibres with firing rates below 0.5 have much higher and more heterogeneous thresholds than those with a higher SR.

The frequency selectivity of the basilar membrane is also reflected in the responses of auditory nerve fibres to tones. The frequency selectivity of an auditory nerve fibre can be characterised by its threshold as a function of stimulus frequency. The resulting function is known as a tuning curve from which a characteristic frequency (CF), the frequency at which the pure tone threshold is lowest, and bandwidth can be estimated. Fibres are distributed in such a way that CFs are ordered in the auditory nerve from high in the periphery to low in the centre (Kiang, 1965). These then synapse with brainstem nuclei in

an orderly way, thus preserving tonotopicity in subsequent nuclei in the auditory processing pathway.

The “sharpness” of an auditory tuning curve is often expressed using its Q-Factor; a dimensionless quantity that is the ratio of the bandwidth to the centre frequency. Bandwidths are often calculated at 10dB above threshold, but other values, such as 40dB (Liberman, 1978) are also used (denoted Q_{10} or Q_{40} respectively). Studies have shown that Q tends to increase with frequency, such that high frequency fibres are broader relative to their centre frequency. This is demonstrated in Figure 1.3, which shows tuning curves produced from the auditory nerve of the cat (note the logarithmic frequency scale).

Tuning can also be estimated using a technique called reverse correlation in which the portions of a broadband stimulus immediately preceding each elicited spike are averaged together. The broadband stimulus may be thought of as being composed of a number of frequency components each of varying amplitude. When the amplitude of a frequency component that a given auditory nerve fibre is sensitive to becomes supra-threshold, the fibre will tend to fire at a particular phase of it, at least for low frequencies. Since the stimulus is broadband, other frequency components are also present, but their phases will be random and so when the waveforms preceding a number of spikes are averaged together, these will cancel out. The Fourier transform of the resulting average waveform can then be taken to estimate the frequency response of the nerve fibre. In the auditory nerve, this produces similar estimates of frequency tuning to those produced using the tuning curve method. Similar

estimates of tuning are obtained using noise, complex, or tonal stimuli. This is not necessarily the case in more central regions (Woolley et al., 2006, Christianson et al., 2008).

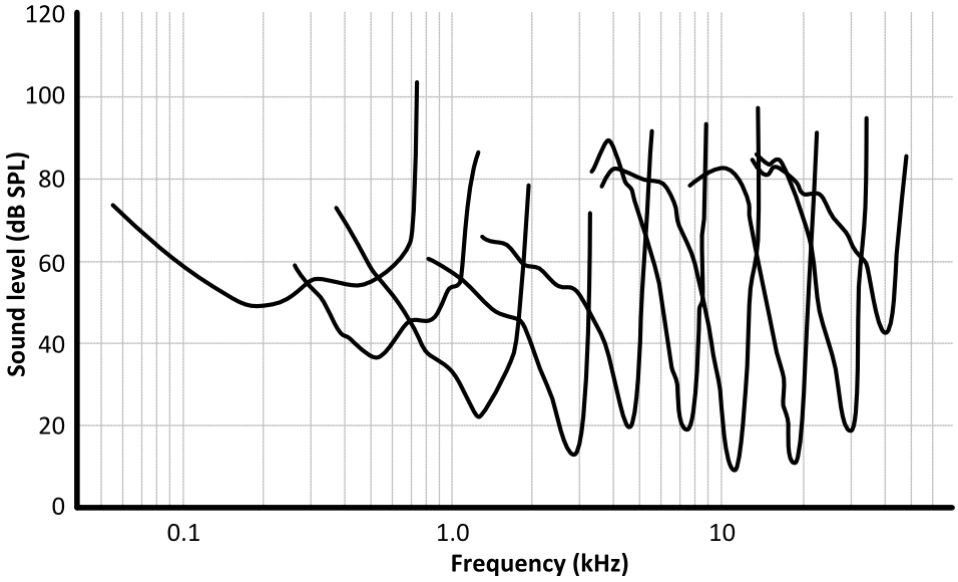

Figure 1.3: Tuning curves of 10 fibres in the auditory nerve of the anaesthetised cat. Reproduced from Palmer (1987).

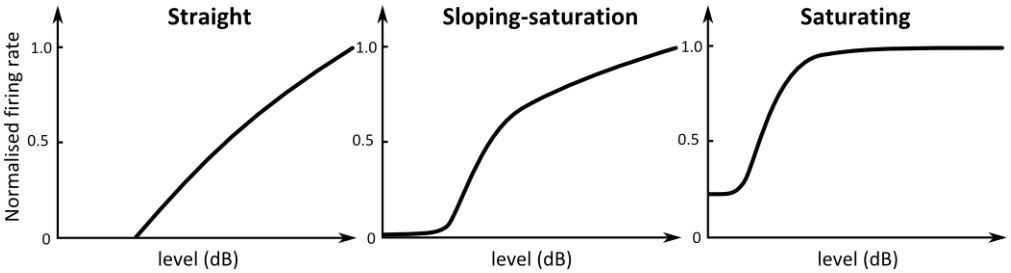

Figure 1.4: Stereotypical auditory nerve rate-level functions showing normalised spike rate as a function of pure-tone stimulus intensity.

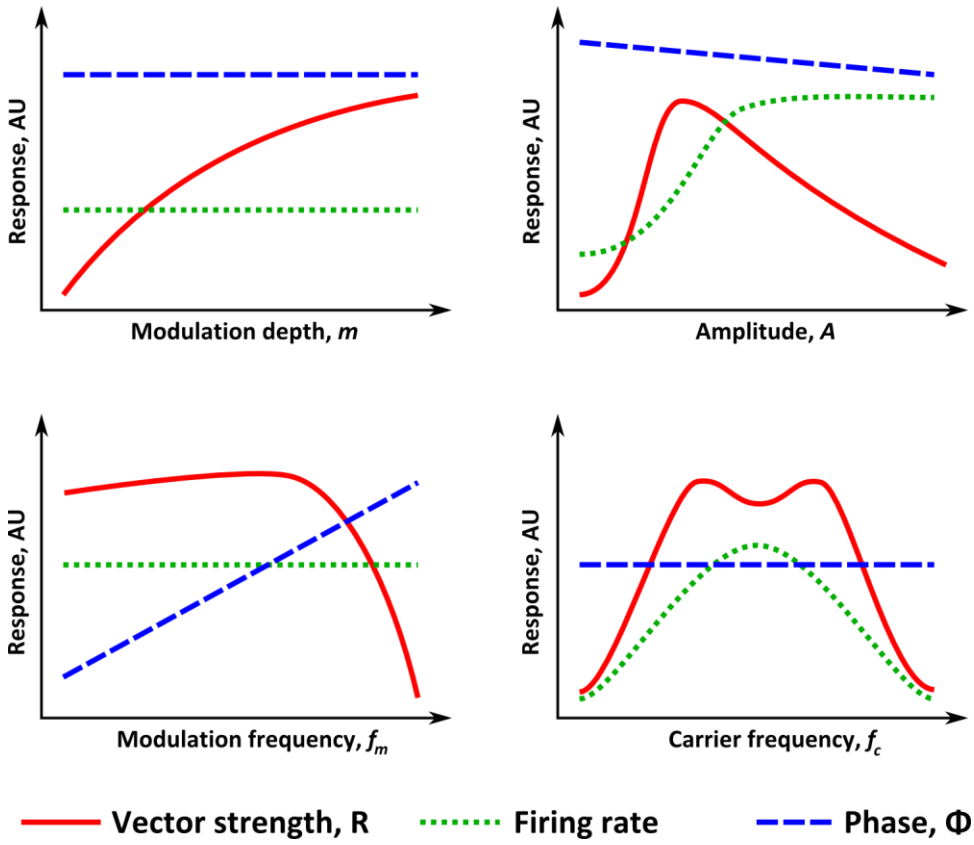

Figure 1.5: The effect of SAM tone parameters on responses in the auditory nerve. Each graph shows the effect of varying a single parameter while others are held constant on the vector strength (red), firing rate (green) and modulator phase (blue). Reproduced from Joris et al. (2004).

Another characterising feature of auditory nerve fibre responses is the way that the firing rate increases as a function of stimulus level. This is known as the rate-level function (RLF). The shape of this function describes the dynamic range of a fibre; the range of sound levels over which sound level is uniquely encoded by the firing rate of the neuron. Figure 1.4 shows stereotypical auditory nerve RLFs. Low SR fibres tend to have straight RLFs, where the increase in firing rate is approximately proportional to the decibel increase in

sound pressure. High SR fibres tend to show a saturating RLF, where the slope is steep and reaches saturation within approximately 30 dB of threshold, and medium SR fibres tend to show a sloping saturation (Winter et al., 1990).

The auditory nerve not only carries information about spectral content in the firing rate and position of its fibres, but also carries temporal information in the precise timing of action potentials within it. The simplest form of temporal code is known as phase-locking where neurons respond preferentially to a specific phase of the input signal (Rose et al., 1967, Kettner et al., 1985, Palmer and Russell, 1986). This phenomenon underlies the use of inter-aural timing cues for sound localisation, but is also thought to be used to encode stimulus periodicity and therefore pitch (Oxenham, 2012).

The extent to which neurons phase lock is often measured using the vector strength. Each spike is considered to be a vector of unit length with an angle, θ_i, corresponding to the phase of a periodic stimulus at the time it was initiated. The vector strength, r, is then given by

$$r = \frac{\sqrt{(\sum \cos \theta_i)^2 + (\sum \sin \theta_i)^2}}{N} \quad (1)$$

Where N is the total number of recorded spikes. This results in a value between zero and one, where zero indicates a uniform distribution of spikes over stimulus phase and one indicates that all spikes were initiated at the same phase (modulo 2π). Phase-locking has been shown in the auditory nerve of many species, although direct measurements in human have not been possible due to the invasive nature of the procedure. The upper frequency limit of phase locking in the auditory nerve varies across different species, but it has been

demonstrated in the guinea pig up to ≈3.5 kHz (Palmer and Russell, 1986) and > 5 kHz in cats and squirrel monkeys (Johnson, 1980, Rose et al., 1967).

This might suggest that sounds are represented using only firing rates for high CF neurons. However, phase-locking also occurs to the *envelopes* of sinusoidally amplitude modulated (SAM) tones. The envelope of a waveform can be defined as the variation in amplitude over time and is often considered separately from the fine structure, which is more closely associated with the spectral content. Envelope phase-locking can be quantified in much the same way. SAM tones have four parameters: signal level A, modulation depth m, carrier frequency f_c and modulation frequency f_m. An SAM tone is then described by:

$$x(t) = A(1 + m\sin(2\pi f_m t))\sin(2\pi f_c t) \quad (1)$$

The way that these parameters affect auditory nerve responses is summarised in Figure 1.5. With increasing modulation depth, phase-locking increases but with decreasing gain. Synchronisation as a function of level is non-monotonic and is maximal where the signal level variations correspond to the steepest part of the RLF. The function that relates synchronisation to modulation frequency is usually called the temporal modulation transfer function (tMTF) to disambiguate it from the function relating firing rate to modulation frequency, the rate modulation transfer function (rMTF). In the auditory nerve, the tMTF is a low-pass function, whereas the rMTF is generally flat. The upper limit of envelope phase-locking is limited by the bandpass filtering of the cochlea at low CFs as the modulation sidebands become increasingly attenuated as the modulation frequency increases. At high CFs, however, other physiological

processes appear to limit envelope phase-locking to ≈1 kHz at CFs above 10 kHz (Joris and Yin, 1992). This is supposed to be a reflection of the temporal response properties of hair cells, comprising the relationship between receptor potential and calcium current, calcium current and calcium concentration and the electrical properties of the cell membrane itself (Kidd and Weiss, 1990).

1.1.4 Auditory brainstem nuclei

The auditory brainstem nuclei comprise the cochlear nucleus (CN), the superior olivary complex (SOC) and the nuclei of the lateral lemniscus (NLL). The CN may be broadly divided based on cell morphology and physiology into three regions: the dorsal (DCN), posteroventral (PVCN) and anteroventral (AVCN) cochlear nuclei. Fibres of the auditory nerve spiral towards the CN where they bifurcate into descending projections innervating the DCN and PVCN and ascending projections innervating the AVCN, maintaining topological organisation in each region. The cytoarchitecture of the CN as a whole is highly heterogeneous, giving rise to much more diverse response properties than in the auditory nerve and ostensibly forming the basis of parallel processing of auditory features. For example, spherical and globular bushy neurons found primarily in the AVCN form large synapses with auditory nerve endbulbs of Held, and thereby relay spike timing information with a high degree of accuracy to the SOC, providing important cues for sound localisation. Conversely, pyramidal cells in the DCN are characterised by an extensive dendritic tree and show complex frequency response areas (FRAs) resulting from integration of inputs of diverse origins.

Frequency selectivity is, in general, preserved or even enhanced in the CN, which is attributed to the role of lateral inhibition (Rhode and Greenberg, 1994). Subpopulations of CN cells also show enhanced phase-locking to low frequency tones (Joris et al., 1994). Others show selectivity to other acoustic stimulus dimensions such as modulation frequency, as demonstrated by bandpass MTFs, or stimulus intensity, as demonstrated by the non-monotonic RLFs. As CN responses are not presented in this thesis, a detailed overview of the diverse response properties of the CN will not be presented here. For a more comprehensive overview see Rhode and Greenberg (1992) or Winer and Schreiner (2005).

Afferent projections from the CN project to the ipsi- and contralateral SOC, NLL and directly to the IC (see 1.1.5). There are also commissural projections to the contralateral CN. The SOC is the first site of integration of binaural information and comprises three main nuclei that can be reliably identified across a number of species, across which a high degree of structural diversity exists: the medial superior olive (MSO), lateral superior olive (LSO) and the medial nucleus of the trapezoid body (MNTB). The MSO is predominantly responsive to low frequency tones, and is thought to play a prominent role in the processing of interaural timing cues for sound localisation (Malmierca and Hackett, 2010). It receives bilateral ascending input from the bushy cells of the AVCN, which convey temporally precise responses from the auditory nerve. The LSO, conversely, responds to frequencies across the entire audible range and demonstrates responses that are selective to interaural level differences as a result of receiving excitatory signals from the ipsilateral AVCN and inhibitory

ones via the contralateral MNTB. Much of the focus of physiological studies in the SOC has been on binaural processing, but it is also the origin of centrifugal projections to the cochlea. However, relatively few studies have examined the processing of modulated sounds (Kuwada and Batra, 1999, Grothe, 1994), although it has been implicated in the conversion of spike time-based to rate-based encoding of amplitude modulation (Joris et al., 2004).

Projections from the SOC extend both directly to the central nucleus of the IC and to the NLL. The NLL is commonly divided into two regions, the dorsal (DNLL) and ventral (VNLL) nuclei, which are associated with binaural and monaural processing respectively. There have also been relatively few studies on the response properties of these nuclei, although it has been shown that there is tonotopic organisation (Aitkin et al., 1970) and amplitude modulation selectivity for both temporal and rate codes (Joris et al., 2004). It has also been shown that the NLL of the big brown bat (*Eptesicus fuscus*) has very short integration times, broad tuning curves and precise temporal response properties suggestive of a role in the representation of predominantly temporal properties of sound. Ascending projections from these nuclei extent bilaterally to the IC.

1.1.5 Inferior colliculus

The IC is an almost obligatory relay in the auditory processing pathway (Aitkin and Phillips, 1984), where contra- and ipsilateral signals from the CN, olivary complex and NLL converge. It also receives descending input from the auditory cortex. It can be broadly divided morphologically into and central nucleus

(CNIC) and surrounding cortical areas. The CNIC is characterised by the presence of structural laminae comprising a dense arrangement of disc-shaped cells with parallel dendritic trees interspersed with stellate cells. Cortical regions are often further subdivided into dorsal and lateral, or external, and rostral cortices, however the functional bases of these subdivisions are not well defined (Oliver, 2005). The cortical regions receive extensive non-auditory input, whereas the central nucleus is exclusively auditory (Winer and Schreiner, 2005) and will be the focus of this discussion.

Tonotopicity is maintained in the CNIC with responses most sensitive to low frequency tonal stimulation in the dorsolateral laminae and high frequencies in the ventromedial creating a tonotopic gradient orthogonal to the cellular laminae. However, FRAs show a much higher diversity of shapes than those in the auditory nerve. These are often classified subjectively based on stereotypical shapes, such primary-like V-shaped, narrow I-shaped or closed O-shaped, which have been proposed to reflect discrete origins of afferent input (Ramachandran et al., 1999). However, a recent study demonstrated that FRA shapes more likely form continua along multiple dimensions as opposed to distinct clusters (Palmer et al., 2013).

Encoding of stimulus intensity also deviates from observed mechanisms in the auditory nerve. Minimum thresholds have been shown to vary systematically in the CNIC with lower thresholds centrally and higher thresholds in peripheral regions (Stiebler, 1986). This arrangement could contribute to a place-code for stimulus intensity and subsequently the percept of loudness represented by a

centrifugal shift in response with increasing stimulus intensity. Some FRAs in the IC also show a non-monotonic relationship between pure-tone intensity and firing rate.

Representation of stimulus features in the timing of neural spikes is also evident in the IC. Phase-locked responses to pure tones have been observed (Liu et al., 2006) but with a lower limit (<1 kHz) than in those in the auditory nerve. This progressive reduction in the upper limit of pure-tone phase locking can be observed throughout the ascending auditory pathway. Phase-locking to stimulus envelopes has also been observed in the IC. Modulation gains based on synchronicity measures are higher in the IC than those measured in the auditory nerve or in the brainstem nuclei. Similarly to pure-tone responses, synchronised responses to envelopes are also restricted to lower frequencies than in more peripheral nuclei, although there is considerable variability across species. In general, maximum modulation frequencies at which phase-locking can be observed in the IC fall below ≈200 Hz. For a review, see Joris et al. (2004).

The IC also shows much greater heterogeneity of rMTFs than more peripheral nuclei. rMTFs represent the change in average firing rate as a function of modulation frequency and provide a means to measure the ability of a neuron to encode amplitude modulation in its firing rate. A striking difference is the emergence of more strongly tuned, bandpass rMTFs in which the neuron fires most rapidly in response to a narrow range of modulation frequencies. The modal value of these functions typically lies between 30 and 100 Hz, which is

consistent across a number of species, and the upper limits similar to those of the tMTFs.

Ascending projections from the IC extend to the Medial Geniculate Body (MGB) of the thalamus from which they extend to the cortex. Descending projections also feed back to the SOC and CN (Malmierca and Hackett, 2010).

1.1.6 Medial geniculate body

The medial geniculate body (MGB) is part of the thalamus and constitutes an obligatory relay for information passing from the midbrain to the cortex. It has ascending and descending connections primarily with the IC and the auditory cortex. It is commonly divided into three divisions (Winer, 1992): the ventral (MGBv), dorsal (MGBd) and medial (MGBm), although earlier accounts identified only core and belt regions (Jones et al., 1985) and others further divide these regions on the basis of histology and physiological mapping (Anderson et al., 2007). The lack of interconnectivity has also led to the description of the regions as part of parallel auditory processing pathways (Calford and Aitkin, 1983).

The MGBv, corresponding to a “core” region in alternative nomenclatures, is characterised by a cytoarchitecture similar to that of the IC with a laminar organisation of neurons with polarised dendritic fields, which preserve tonotopic organisation. Frequency selectivity is also preserved with the best pure tone tuning curve bandwidths in the MGBv comparable to those of the IC (Calford, 1983). The main source of ascending information to the MGBv originates in the ipsilateral IC and projections extend primarily to the primary

auditory cortex (see 1.1.7), where tonotopic organisation is maintained. A small proportion of neurons in this region also show moderate phase-locked responses to tones up to 1 kHz in the cat (Rouiller et al., 1979). Selectivity to amplitude modulation rates up to around 100 Hz in both rMTF and tMTFs have also been observed (Joris et al., 2004).

The two other regions of the MGB do not show tonotopic organisation. The MGBd is characterised by broad frequency tuning and receives its inputs primarily from the external cortex of the IC. Its projections synapse mainly in belt areas of the auditory cortex (see 1.1.7), where tonotopy is also generally weak or absent. The MGBm has highly variable response properties commensurate with its heterogeneous cytoarchitecture and there is evidence of multimodal integration (Jones, 2007).

1.1.7 Auditory cortex

The identification of auditory cortical areas based on both anatomy and physiology varies dramatically between species. However, across species the cortex is commonly divided into one or more “core” regions and other “belt” regions. As it is the animal model under study in this thesis, the primary focus of this review will be on the guinea pig. The first detailed map of the guinea pig auditory cortex was produced by Redies et al. (1989), who delimited auditory fields based on electrophysiological response properties to pure tones. In this study, two main tonotopically organised regions were identified: the anterior field (A) and the dorsocaudal field (DC). The two regions are delimited by a reversal in the tonotopic gradient, although response latencies and frequency

selectivity were similar across the whole region. A third region was identified situated rostrally to field A, although this field is much smaller and was accordingly designated the small field (S). Auditory responses were also reported in regions caudal to A and DC, although these did not show tonotopic organisation and were characterised by longer response latency, broad tuning and a comparative insensitivity to tonal stimulation. This map was later adjusted by Wallace et al. (2000) using both histological and electrophysiological methods. These methods led to the parcellation of the anterior field A into a primary area, AI, and another, tonotopically organised belt region, the ventrorostral belt (VRB, see Figure 1.6). In this thesis, responses were recorded exclusively from AI, which was located using anatomical landmarks, and so the focus of this and subsequent discussions (see 1.3) will be on this core area.

The auditory cortex, like all neocortex, has a laminar structure, which comprises 6 layers. The main ascending input from the thalamus originates in the MGBv and synapses topographically in layers III and IV, while layers V and VI project to many upstream auditory nuclei (Winer and Schreiner, 2005). It is also characterised by highly complex and extensive intrinsic connectivity both between and across layers.

Responses in the auditory cortex show a lower limit of phase locking than more peripheral auditory nuclei, with most neurons only firing at regular phases of periodic stimuli with frequencies less than 100 Hz, a limit that is consistent when measured using AM tones or click trains, which is incongruous with

perceivable acoustic features. Many neurons show sensitivity to specific modulation frequencies, but the vast majority of best modulation frequencies (BMFs) derived from tMTFs are below 20 Hz (Joris et al., 2004). These rates correspond only to rhythmic or syllabic features of communication sounds and are too slow to constitute a viable encoding strategy for other salient cues such as fundamental frequency. BMFs derived from rMTFs have been shown to extend to frequencies up to an octave higher than those from tMTFs, however this is still substantially lower than the upper limit for the perception of periodicity pitch (≈800 Hz; Chung and Colavita, 1976).

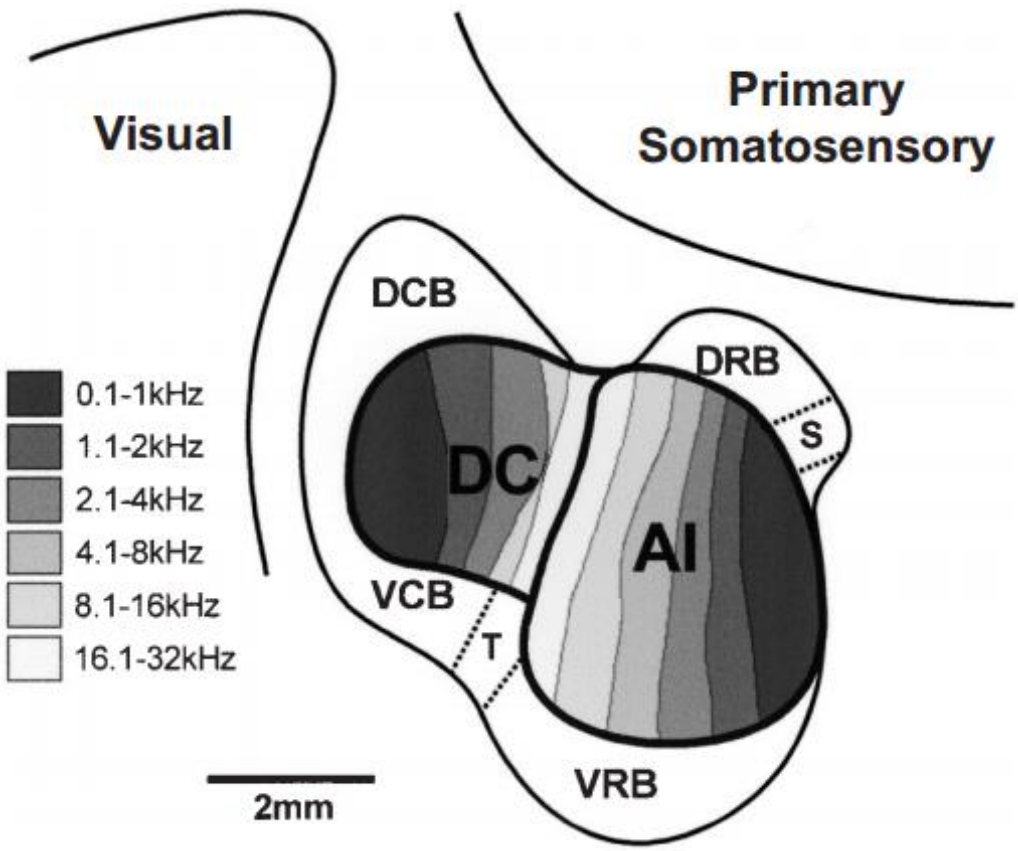

Figure 1.6: Divisions of the right auditory cortex of the guinea pig as determined using histological and electrophysiological methods. Tonotopic organisation is indicated by the shaded bands, which indicate regions with similar BFs. From Wallace et al., 2000.

1.1.8 Summary

This section has described briefly the anatomy and interconnections of the nuclei of the auditory system, as well as a functional description of each of these regions as probed using tonal and simple modulated stimuli, such as SAM tones. This kind of analysis has been crucial for the identification of anatomical locations of brain regions associated with processing various aspects of sound. However, the success of models based on these findings to predict responses to the salient features of communicative sounds is generally poor, due to the nonlinear nature of sound analysis by the auditory system (Escabí and Schreiner, 2002; Machens et al., 2004).

It is also unclear how the measures, such as the MTFs or vector strength measures of temporal coding, which are traditionally used to characterise neuronal responses relate to perception. For example, the apparent discrepancy between predictions based on phase-locked or rate-based codes for amplitude modulation and behavioural performance at the cortical level is poorly understood and suggests that these stimulus features may be encoded in some other way.

One approach to elucidating encoding strategies and signal transformations at varies stages in the auditory pathway is to record responses to complex, natural stimuli and to then perform various manipulations on these representations to examine the effect that this has on a neural decoder. This is the approach taken in this thesis, which is concerned with the representations of speech sounds and their degraded counterparts. In order to reconcile any findings using this

strategy with the extensive body of knowledge based on simpler stimuli, an understanding of the acoustic structure of speech is necessary and is discussed in the next section.

1.2 The acoustical properties of speech

In order to interpret neural responses to speech sounds, it is useful to understand the acoustic nature of speech, and how it arises. Speech sounds are produced by the movement of air from the lungs through the trachea via the larynx and out through the nose or mouth. There are two modes of excitation for the air passing though this system: random or quasi-periodic. The mode of excitation is controlled by the vocal folds (sometimes misleadingly referred to as vocal cords), which comprise part of the larynx. These muscular and cartilaginous folds are able to occlude, or partially occlude the airway thereby affecting airflow and acoustic excitation of the vocal tract.

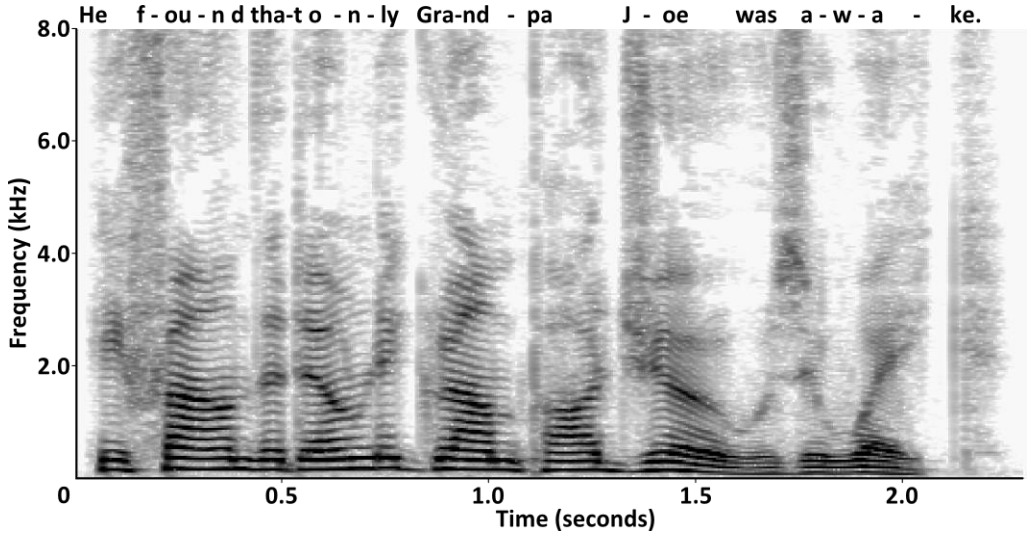

Figure 1.7: Spectrogram of a sentence produced by a male talker in quiet. The syllables are written above the corresponding position on the time axis.

When the gap between the vocal folds, known as the *glottis*, is partially closed, aerodynamic and muscular interactions cause air to be released in bursts eliciting the quasi-periodic mode of excitation. This is also known as phonation or voicing. When the glottis is open and air flow is sufficiently fast, the air flow becomes turbulent initiating random, or noisy, excitation (Laver, 1994). These two modes form the basis of voiced and voiceless phonemes. Voiced sounds have a spectrum comprising a fundamental frequency (f_0) component and other components, known as harmonics, at integer multiples of this with amplitudes that decrease with increasing frequency. Voiceless sounds have a broadband spectrum. Figure 1.7 shows a spectrogram, a spectro-temporal visual representation of sound, of a sentence produced by a male talker. The closely spaced dark, horizontal striations are clearly visible, representing individual harmonics during voicing. Broadband, noise-like sections are also visible, particularly for the */f/* in "found" or */ʒ/* in "Joe".

Phonemes are broadly divided into two categories; *vowels* and *consonants*. Vowels are sounds produced with no audible constriction of the vocal tract above the glottis and, during normal speech, are all voiced sounds discernible by their spectral content. Steady-state vowels can be discriminated based on the position of formants. These are spectral bands that are exaggerated by the resonant frequencies of the vocal tract, of which only the first two are required for steady-state vowel discrimination (Bladon and Fant, 1978). These can also be seen in Figure 1.7 as the frequency modulated bands of darker harmonics.

Consonants comprise a much more diverse group of sounds, but are broadly defined as the sounds produced with partial or complete constriction of the vocal tract. They can be either voiced or unvoiced. For example, the bilabial stop consonant */b/* as in **b**ath is discriminable from the bilabial stop consonant */p/* as in **p**ath partly on the basis of the presence and absence of voicing, respectively. The manner in which consonants are produced also forms the basis of discriminating between them. Fricatives, such as the voiceless */s/* in day**s** or the voiced */z/* in da**z**e, are produced by generating turbulent airflow by forming a narrow constriction. Stop consonants, such */d/* of */t/*, require complete closure of the vocal tract introducing a brief silence into an utterance. Affricates may be thought of as a stop followed by a fricative, such as */tʃ/* in **ch**at. The other consonant sounds, nasals and approximants, are more similar to vowels. Nasals involve the diversion of air flow through the nasal passages, such as */m/* or */n/* in **m**a**n**. Approximants, such as */j/* and */r/* in **y**ea**r**, require a lesser degree of constriction of the vocal tract. Consonants are also often described in terms of the place of articulation, such that a full description of the consonant */g/* as in **g**ap would be "voiced velar stop", since it is articulated with the back part of the tongue against the velum (the back part of the roof of the mouth). In the literature, classification schemes for consonant place may be broad such as "front", "middle" and "back", or more specific (e.g. "bilabial", "alveolar").

Discrimination of consonant sounds, then, utilises a variety of cues, both spectral and temporal. *Voice-onset time* (VOT), for example, provides a temporal cue from distinguishing */b, d, g/* from */p, t, k/* respectively, particularly when they are initial consonants, as opposed to medial ones.

1.3 The neural representation of speech sounds

So far, only the neural representation of simple and artificial stimuli have been discussed. These are important tools for understanding the fundamental processing capabilities of the auditory system, but cannot necessarily be used to predict responses to natural sounds, particularly at more central stages, which show highly nonlinear response characteristics (Ahrens et al., 2008). The neural representation of speech sounds in the auditory system has been studied for over 30 years (Young and Sachs, 1979, Delgutte, 1980, Palmer et al., 1986). A problem arises, however, in the interpretation of the recorded responses as this necessitates some assumptions about the encoding scheme utilised. At one extreme, only the number of spikes over the entire duration of the speech sounds of interest could be considered, which is referred to as a rate code. At the other, a temporal code could be assumed in which the precise timing of individual spikes are considered. Between these two extremes, encoding windows of intermediate durations could also be considered. In this context, the encoding window refers to the duration of the neural response which we assume to represent a single symbol in the neural code.

The functionality of the cochlea as a frequency analyser has been previously described (see 1.1.2). It has also been asserted that vowel discrimination is

possible on the basis of steady-state spectral cues, in particular the spectral locations of vocal tract resonances, or formants. It follows from studies using simple stimuli that these formants should correspond to peaks in firing rate at the appropriate points in the tonotopically organised auditory nerve. This is indeed the case for stimulus intensities up to approximately the level of conversational speech (≈60 dB SPL). Figure 1.8 shows a comparison between the spectrographic representation of the vowel-consonant-vowel speech token */asa/* produced by a male talker and its representation by 100 simulated auditory nerve fibers, using the computational model described in 2.1.1. The vertical striations show phase-locking to the stimulus envelope during the quasi-periodic, voiced vowel portions. There is also a clear frequency specific response, with the high frequency fibers responding preferentially to the fricative */s/* portion. Formant positions are also visible in both representations as the broader horizontal striations. It is these peaks that become indistinct at higher stimulus levels (Sachs and Young, 1979). This is due to firing rate saturation, diminishing the dynamic range of rate responses and the effect of cochlea suppression, where the first formant diminishes responses to the second as overall intensity is increased. This observation is at odds with psychophysical results, which demonstrate a robustness of intelligibility across a much broader range of stimulus intensities.

Since the auditory nerve provides the auditory input to the CNS, and we assume that information is represented in the spiking activity of its neurons, it follows that there must be some other aspect of the representations that conveys the spectral shape of a vowel-like stimulus other than in the firing rate that is robust

to intensity variations. Young and Sachs (1979) investigated the representation of steady-state vowel sounds in temporal firing patterns of the auditory nerve by examining the spectral content of responses of fibres with a range of CFs. The spatial distribution of phase-locked activity represents formant locations in a way that is more robust to stimulus intensity variation than a rate-based representation. There is also a significant amount of phase-locked activity to the fundamental frequency across all CFs. At very low frequencies this can be attributed to energy at the fundamental frequency, but at higher frequencies it is due to acoustic interactions between unresolved harmonics. Neural responses in the auditory nerve also demonstrate locking to frequencies corresponding to the resonant frequencies of the vocal tract during steady-state vowel sounds (Delgutte and Kiang, 1984). At stimulus intensities corresponding to conversational levels of speech, representations in the auditory nerve and the cochlear nucleus are broadly commensurate with those that could be predicted from responses to pure tones and the spectro-temporal properties of the signal. This extends further to more natural speech-like sounds with dynamic spectral changes (Palmer and Shamma, 2004).

Much less is known about the representation of speech sounds in more central structures compared to the auditory periphery. The existence of tonotopic organisation all the way up to primary cortical regions was determined by responses to simple stimuli and described in 1.1, as is the conversion from the 1-dimensional tonotopic axis along the basilar membrane into iso-frequency fields, which could sustain a rate-place representation of speech spectra. The ability of neurons in the central auditory to explicitly encode the temporal

structure of speech sounds reduces progressively in the ascending auditory pathway, with most neurons in the midbrain showing an upper limit of 400-600 Hz for AM stimuli, reducing to tens of Hertz in the cortex. This, however, does not preclude temporal response patterns from representing stimulus features and since speech is by nature a dynamic signal, there must be some temporal window over which neural spiking activity represents salient information. The importance of spike timing in representing speech sounds is still poorly understood, however recent studies have shown that the neural discriminability of speech sounds based on spike timing representations is best correlated with behaviour in an animal model for word-initial consonant sounds, whereas the discriminability spike rate-based representations are better correlated with behavioural discriminability of vowels (Engineer et al., 2008; Perez et al., 2013; Centanni et al., 2013).

In summary, the representation of speech in the cochlea and auditory nerve is relatively well understood and is well predicted by existing computational models of the peripheral auditory system (e.g. Holmes et al., 2004). The same cannot be said about more central structures, as it is only comparatively recently that responses in these areas to complex sounds have begun to be studied. Recent studies have shown that responses to speech sounds become increasingly diverse as the central auditory processing pathway is ascended (Ranasinghe et al., 2013) suggestive of sensitivity to higher order stimulus features and a distributed, sparse representation of complex sounds. Evidence is also accumulating for the importance of spike timing in the representations of dynamic speech stimuli.

These studies, however, have focused on tasks that involve discrimination of individual speech tokens, and it is not yet known whether these representations underlie robust discriminability in the face of non-meaningful acoustic variability, such as that between multiple exemplars of each speech sound. It has been demonstrated that in the presence of broadband noise, longer integration windows are required to maintain robust discriminability of neural representations (Shetake et al., 2011). Whether or not this extends to other non-meaningful variations, such as those between multiple exemplars of individual phonemes, remains to be investigated.

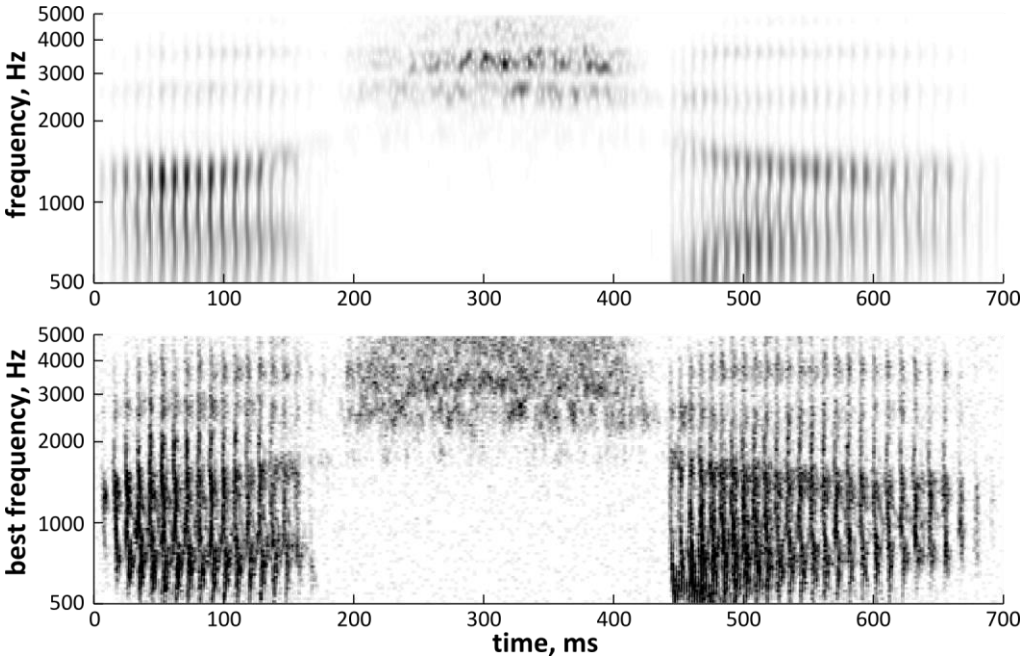

Figure 1.8: Comparison of spectrographic and auditory nerve representations of VCV */a/s/a/*. The upper panel shows a spectrogram of the speech token and the lower depicts average PSTHs from 100 simulated auditory nerve fibres with BFs logarithmically spaced between 500 and 5000 Hz.

1.4 Hearing loss and cochlear implants

Sensory-neural hearing loss (SNHL) is a type of hearing impairment resulting from damage to the cochlea, auditory nerve, central auditory nuclei or a combination of the above. Unlike conductive hearing loss, where elevated perceptual thresholds result from problems arising before the cochlear processing stage, SNHL is usually associated with deficits in frequency selectivity (Moore, 2003b) as well as problems arising from diminished temporal coding (Henry and Heinz, 2012). See 1.5 for a discussion of the effects of spectro-temporal degradation on speech recognition.

In order to compensate for elevated sensory thresholds, hearing aids amplify sounds entering the ear canal. Early hearing aids were able to do only this, but contemporary devices employ various digital signal processing techniques to choose which aspects of the signal are amplified and to what extent. For example, devices apply amplitude compression to account for loudness recruitment; the abnormal growth of perceived loudness with sound intensity experienced by hearing impaired listeners. Current processors can also apply compression in multiple frequency bands and apply de-noising algorithms to emphasize speech.

In patients with severe to profound hearing loss for whom a hearing aid is no longer efficacious for speech recognition, the function of the organ of Corti can be replaced by the array of stimulating electrodes in a cochlear implant. Experiments in electrical hearing began at the beginning of the 19th century when Alessandro Volta introduced direct current stimulation to his ears

resulting in the sensation of a “bubbling” or “boiling” sound (Volta, 1800). The first implantable electrical hearing device, however, was not realised until 1957, when an electrode was used to directly stimulate the auditory nerve, which utilised a transcutaneous induction coil link (Djourno and Eyriès, 1957). This enabled the patient to discriminate speech sounds in small closed sets, but was a long way from speech recognition using only auditory cues.

Auditory prostheses have become increasingly sophisticated ever since, with modern cochlear implants featuring multi-electrode arrays, which are distributed along the length of the cochlea and stimulate neurons of the spiral ganglia (see Figure 1.9). This multi-channel arrangement enables the devices to make use of the tonotopicity of the healthy cochlear, as frequency specific information may be presented to the appropriate tonotopically arranged neural populations of the auditory nerve. Contemporary cochlear implants utilise up to twenty-two stimulating electrodes. However, the effective number of independent frequency channels does not appear to be more than between seven and ten at most, as measured by speech discrimination performance as a function of number of active channels (Friesen et al., 2001). The reasons for this are not fully understood, but it is likely to be due to current spread to neighbouring neural populations or a mismatch between frequency-place mapping of the stimulating electrodes and that of the healthy cochlea. This results in an input to the auditory system that is profoundly spectrally degraded. A direct comparison of spectral resolution in electrical hearing to normal hearing, however, is difficult to make as it involves the estimation of the number of independent spectral channels in the healthy cochlear. Based on

estimates of auditory filter width as a function of CF and the range of frequencies important for speech recognition, though, this number is thought to be approximately twenty-eight (Moore, 2003a) .

In order to describe the nature of temporal degradation in cochlear implant systems, it is necessary to describe typical signal processing strategies used in them. Figure 1.10 shows a simplified schematic of the continuous interleaved sampling (CIS) strategy. Amplitude compression is also required to map the large range of intensities encountered in everyday listening situations to the comparatively small dynamic range of electrical hearing. This has been omitted here for simplicity.

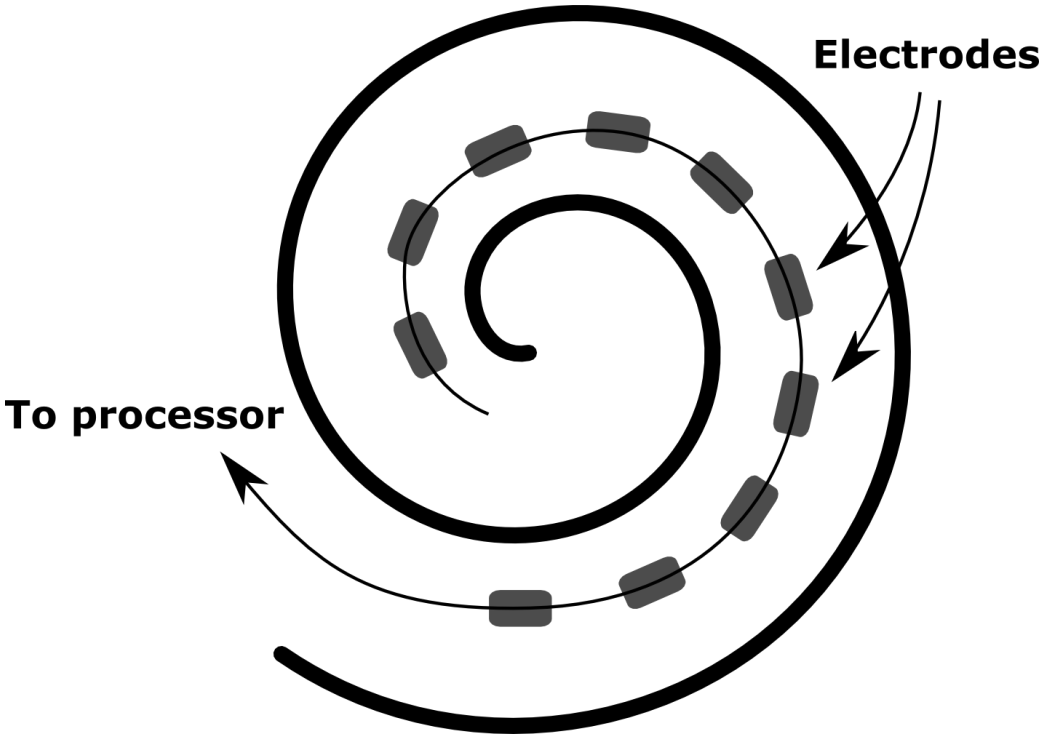

Figure 1.9: Schematic representation of a stimulating electrode array of a cochlear implant inside the cochlear spiral.

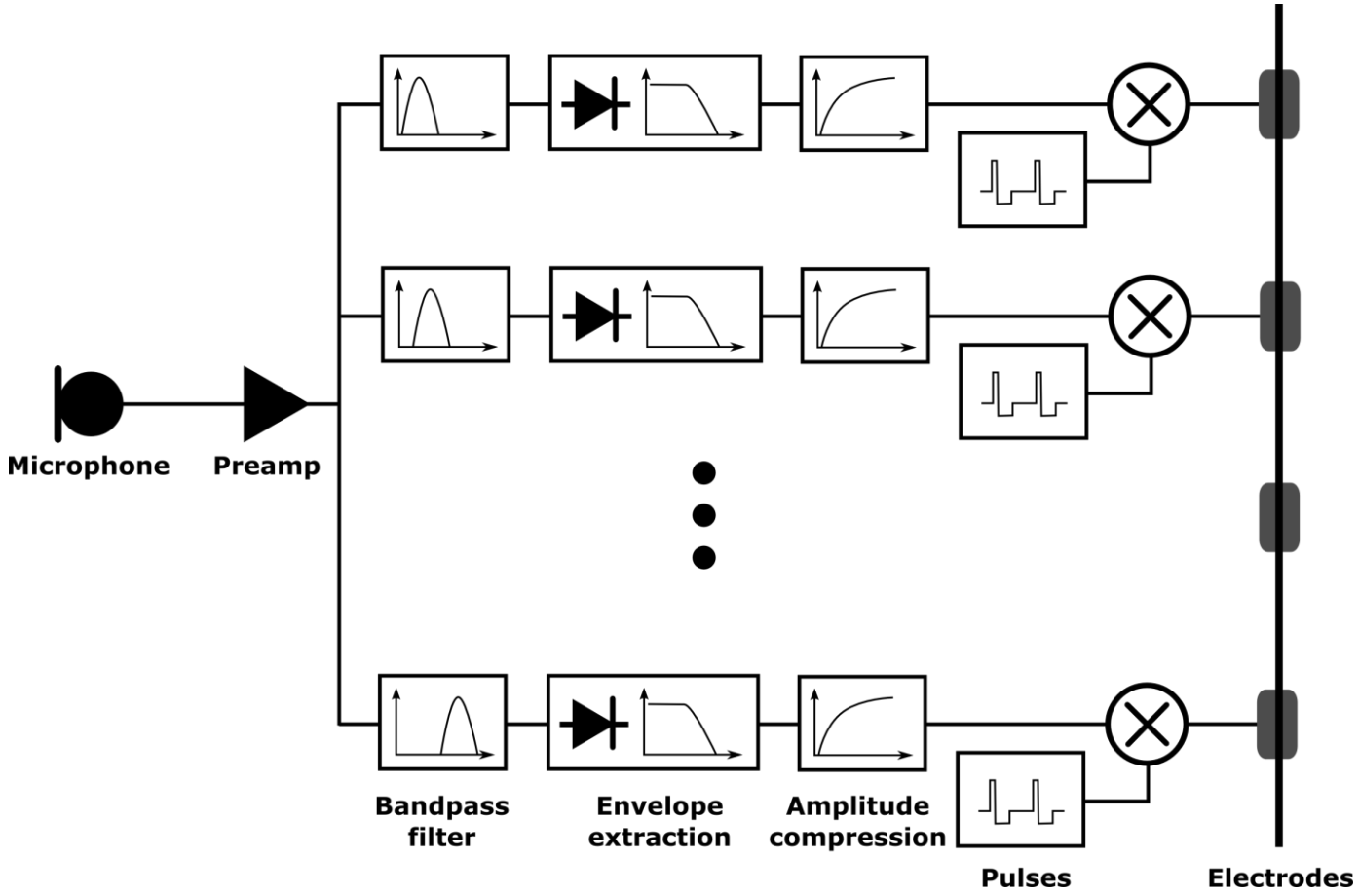

Figure 1.10: Schematic representation of the continuous interleaved sampling (CIS) cochlear implant processing algorithm.

The input signal is initially split into a number of spectral bands, or channels, by a bank of bandpass filters. Within each of these channels, the slowly varying amplitude of the signal, or envelope, is extracted and used to modulate a stream of electrical pulses, which is sent to a stimulating electrode. These pulses are interleaved in time to minimise channel interaction, and hence their timing is unaffected by the temporal fine structure (TFS) of the acoustic stimulus. This means that any salient information contained in the TFS is lost, and signal segregation and speech discrimination must be performed on the basis of the envelopes alone. The rate of envelope information that can be presented is determined partly by the cut-off of the envelope extraction filter used and partly by the frequency of pulses arriving at any one electrode, which, due to the Nyquist-Shannon sampling theorem must be at least twice the maximum frequency that can be represented.

It is worth noting here that recent studies suggest that hearing impairment is not necessarily accompanied by pathological hearing thresholds for pure-tones. The suggested mechanism involves the deafferentation of the hair cells of the cochlea and subsequent degeneration of auditory nerve fibers (Liberman and Kujawa, 2014), followed by central gain normalisation mechanisms. The effects of the reduction of afferent input from the cochlea are not well understood, but it has been proposed that this leads to a stochastic undersampling of the acoustic signal causing speech perception impairment particularly in difficult listening situations (Lopez-Poveda and Barrios, 2013).

1.5 Perceptual consequences of spectro-temporal degradation

The resilience of speech recognition to severe spectro-temporal degradation has been demonstrated in many studies involving cochlear implant users. Some of the highest performing cochlear implant users, despite receiving a comparatively crude input, are able to perform even the most difficult speech recognition tasks with similar accuracy to normal hearing listeners. However, there remains a high degree of variability, even between users of comparable systems (Wilson and Dorman, 2007). It is not yet clear what cues are most salient for these tasks and what limitations are imposed by central auditory processing mechanisms.

Early cochlear implants had only a single stimulating electrode, so speech recognition needed to be performed on the basis of temporal cues alone. Rosen et al. (1989) conducted a small study with four users of single channel cochlear implants into phonetic information provided by the devices. They found that the listeners were able to utilise envelope cues to distinguish between the presence and absence of voicing to augment intelligibility when combined with speech-reading. Listeners were also able to extract intonation cues to discriminate questions from statements, but were unable to utilise fine structure cues conveying spectral shape. This was evident in their inability to resolve confusions arising from consonant place of articulation such as those between */b/*, */d/* and */g/*. The increase in popularity of multichannel implant systems enabled experimenters to investigate the effect of the number of spectral channels on intelligibility. These studies invariably find that, despite

the high number of stimulating electrodes available (up to 22), listeners do not benefit from more than 7 or 8 for phoneme recognition tests in quiet (Fishman et al., 1997, {Friesen, 2001 #13). The number of channels required to reach asymptotic performance also depends on the test speech material, as the benefit of more channels is larger when the signal-to-noise ratio (SNR) is poor (Shannon et al., 2004).

Envelopes are represented at each electrode by a train of electrical pulses (see Figure 1.10), and the rate of amplitude envelope information that can be presented at each electrode is limited by the rate of these pulses. As the number of channels increases, however, this increases the overall rate that pulses must be presented by the processor, particularly in a CIS processing strategy, as the pulses are not presented simultaneously on any pair of electrodes. Further, each individual pulse must be shorter and must therefore be a larger amplitude to maintain loudness levels requiring more power. It is preferable in contemporary CI systems, where the power source is currently worn externally, to minimise the battery size and therefore power consumption of the system so it is of interest to CI system designers to understand the effects of pulse rate as well as the number of spectral channels on speech intelligibility. A number of studies have examined the effect of stimulation rate on speech intelligibility in cochlear implant users, but the results are often highly variable between subjects and often show little to no benefit from high stimulation rates (Friesen et al., 2005).

The use of CI implantees is potentially problematic for several reasons. Firstly, it is not possible to determine directly how stimulation at a given electrode relates to activity in the auditory nerve. Electrodes are normally adjusted to provide equal loudness for a given stimulation amplitude, but it is not clear whether decreased sensitivity is due to the presence of dead regions in the spiral ganglia, or the placement of the electrode relative to it. Secondly, the frequency-place map distortion incurred by differences in insertion depth and frequency range distortion inherent in clinical fitting procedures are not easy to characterise and compare across patients. These can have a profound effect on cochlear implant outcomes and can be a confounding factor in investigating the effect of parametric variations of processor parameters (Shannon et al., 1998) Finally, areas of the brain usually associated with primarily auditory processing may have incurred processing deficits due to being recruited for other sensory modalities after prolonged periods of deafness, for example. These differences in central processing between patients are not easily reconcilable.

One way to circumvent these problems is to present normal hearing listeners with an artificially degraded acoustic signal. Van Tasell et al. (1987) investigated the use of temporal cues alone for speech recognition. Envelopes of 19 VCVs in */a/*-consonant-*/a/* context were extracted by full-wave rectification and low-pass filtering, which were then used to modulate pink noise (low-pass filtered below 3 kHz). The cutoff of the low-pass envelope extraction filter was set to 20, 200 and 2000 Hz. Discrimination of the VCVs was generally poor for all conditions compared to unprocessed speech, but well above chance. This study

demonstrated that, where spectral cues are not explicitly available, envelope cues related to stimulus periodicity (those above 20 Hz), significantly increased speech discriminability. This may seem surprising in the context of the aforementioned cochlear implant studies, where temporal cues at such rates provide little to no benefit. However, in these studies spectral cues were also available, which could also be utilised to detect the presence or absence of voicing, for example, using only difference in signal power between as little as 2 spectral channels. This suggests that these high rate envelope cues only become useful where other cues are not available.

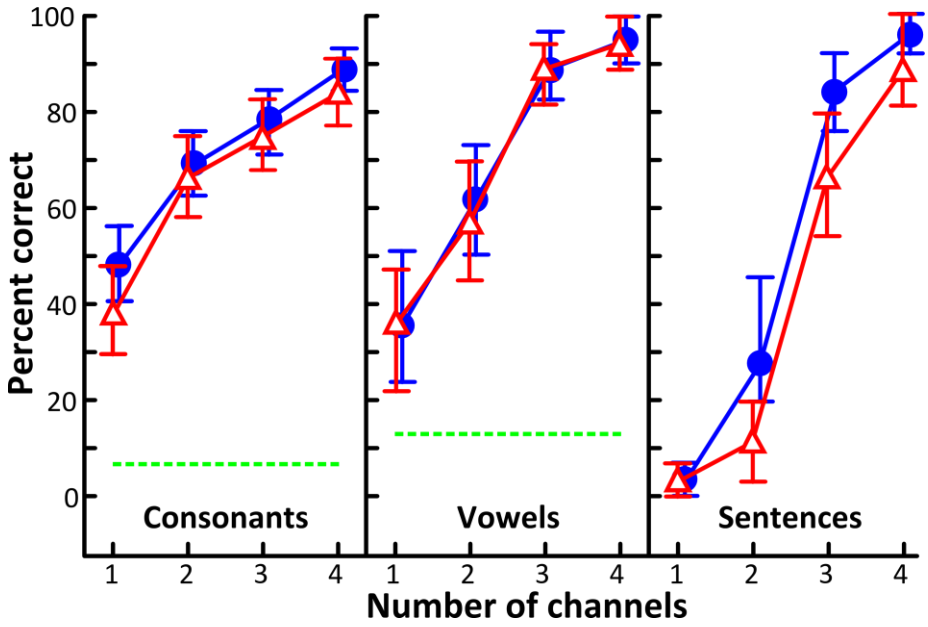

Figure 1.11: The effect of vocoding on speech recognition scores by human listeners. Data shown is for 16 Hz envelope filter cutoffs (△) and 50, 160 and 500 Hz cutoffs grouped together (●). Reproduced from Shannon et al. (1995).

Later studies utilised vocoder processing (see 2.1.2.2) to study the effects of both spectral and temporal degradation. Shannon et al. (1995) similarly degraded speech sounds, but this time extracted envelopes in 1, 2, 3 and 4 spectral channels and use them to modulate narrowband noises with the same bandwidths as the analysis channels. Envelope extraction filters were set to 16, 50, 160 and 500 Hz. They found the expected increase in performance with an increasing number of spectral channels, but no significant increase in performance when envelope cues higher than 50 Hz were not attenuated, so the data were pooled across 50, 160 and 500 Hz envelope extraction filter conditions. It should be noted, however, that in this study envelope extraction filters had slopes of a modest -6 dB per octave, so high rate cues were not severely attenuated. Further, the benefit of envelope cues higher than only 16 Hz were only very modest (see Figure 1.11).

This lack of benefit for high rate envelope cues seems surprising, as detection of the presence or absence of periodicity could feasibly provide a valuable cue for consonant discrimination (e.g. */f/* vs */v/*). As discussed previously, the relative signal power in the spectral bands can also provide cues for the presence or absence of voicing, however the benefit of these cues is not significantly greater for the 2, 3 or 4 channel conditions than it is for the 1 channel condition. These data also appear to be in contrast to the single channel data reported by Van Tasell et al. 1987, however that study presented listeners with a single exemplar of each token, whereas Shannon utilised 3 exemplars of each speech token. The lack of benefit from periodicity cues may be a consequence of them not being robust across these multiple exemplars.

Another possibility is that central processing mechanisms preclude the use of these cues for speech discrimination, but the extent to which this is the case is not yet known.

The effect of carrier type on intelligibility of vocoded speech was investigated by Dorman et al. (1997). They found that, surprisingly, there appears to be little difference in performance between sinusoidal or noise carriers regardless of the speech material. Narrowband noise carriers have intrinsic envelope fluctuations that can be thought of as adding noise to the signal envelopes, whereas tone carriers do not impose this limitation. Whitmal III et al. (2007) specifically studied the effect of carrier type on speech intelligibility and found that, contrary to the aforementioned study, tone-vocoded speech was more intelligible than noise-vocoded speech with similar vocoding parameters for sentences. They attributed this difference to differences in the spectral resolution of the vocoded stimuli in the 0.9 to 2.5 kHz range, which they asserted has been previously shown to be particularly efficacious for consonant discrimination. They also highlight the presence of a 1.2 kHz high-pass "pre-emphasis" filter in the vocoder pre-processing stages used by both Shannon et al. (1995) and Dorman et al. (1997).

1.6 Effects of degradation on neural representations of speech

Very few studies have examined the effect of the spectro-temporal degradations analogous to those inherent in cochlear implants on the neural representations of speech sounds, and hence the neural bases of the perceptual consequences of these degradations are poorly understood.

Loebach and Wickesberg (2006) investigated the effect of noise vocoding on the representation of a small set of syllables in the auditory nerve of the chinchilla, each comprising a unique word-initial stop consonant and vowel. They found that the temporal structure of the pooled, ensemble response to each of the 4 presented consonants was not significantly different to that of their 3- or 4-channel vocoded counterparts. For the 2 bilabial stop consonants */b/* and */p/*, this extended to the 2-channel vocoded condition, however in all cases the representation was degraded for 1-channel vocoded speech. This provides evidence of a neural basis for the discriminability of stop-consonants based purely on temporal cues, where voice-onset time (VOT), the temporal gap between the release of the stop and the subsequent onset of pseudo-periodic voicing, may be used to distinguish certain subsets.

It is unlikely, however, that such cues remain useful for discriminating broader sets of consonants. For example, consonants */s/* or */ʃ/* have no stop and the nasals */m/* and */n/* are voiced and so VOT may not be used to discriminate between them. A more recent study investigated the effect of vocoding on the neural representations of a more comprehensive set of speech sounds. Ranasinghe et al. (2012) presented a stimulus set comprising a single exemplar of each of 7 monosyllabic English words; */dad/, /bad/, /sad/, /tad/, /dʌd/, /diːd/* and */duːd/* produced by a female talker. These were then shifted up in frequency by an octave before being processed with a noise vocoder and presented to anaesthetised rats. Rats were also trained to discriminate pairs of these speech tokens. They measured neural discriminability between the representations of pairs of speech tokens using Euclidean distances between

response vectors comprising the spiking responses from 5 multi-unit recording sites at any one time. The response vectors contained spike counts in 1ms and 40ms bins for word-initial consonant representations, or 1ms and 400ms bins for vowel representations. They found that the behavioural discriminability of rats is best correlated with neural discriminability when 1ms spike timing was conserved for the 4 word-initial consonants, but with that for rate-based representations for the 4 vowels. They also found that neural discriminability, similarly to behavioural discriminability measured in humans, was robust to vocoding with as few as 8 spectral channels, which was also where the rats reached asymptotic performance.

The importance of precise timing in the representation of salient information in speech in the cortex, then, appears to remain even when the temporal fine structure is removed from the input. This is suggestive of the use of a temporal code, as defined by Theunissen and Miller, 1995 as the representation of stimulus properties occurring over a given epoch using response encoding windows of a shorter duration. The extent to which this fine-grained representation extends to representations that enable the formation of syllabic or phonemic categories in the face of input signal degradation remains to be explored.

1.7 The guinea pig as a model for speech processing

This thesis is concerned with the representation of sounds at the level of action potentials initiated by single, or small groups of, cells. As current non-invasive imaging technology is not able to provide this level of spatial and temporal

detail simultaneously, an animal model is required. Guinea pigs are commonly used in auditory research, one of the reasons for which is the similarity of hearing threshold as a function of pure-tone frequency at low frequencies to humans, which is certainly not the case for other common small mammals, such as the rat. Guinea pigs, like humans, are also very vocal and gregarious, with some vocalisations falling into the spectral range of human speech. Figure 1.12 shows the pure-tone threshold function for a number of species high lighting the similarity of that of the guinea pig and humans. This similarity precludes any pre-processing of the speech sounds, such as frequency shifting.

The next basic concern is the spectral and temporal fidelity of peripheral representations. Several studies have investigated neural measures of these at various auditory nuclei in the guinea pig, however the relationship between neurometric measures of auditory acuity and behaviour is still poorly understood. Evans (1992) compared behavioural and physiological frequency in the guinea pig and found a good correspondence between the two measures when the equivalent rectangular bandwidth was extracted from the neural responses using a variety of stimulus sets. This measure also allows direct comparison with human psychophysical data, and it shows that frequency selectivity is poorer than that of humans, at least for frequencies below around 10 kHz (see Figure 1.13). However, frequency selectivity in guinea pigs is still such that if the bandwidth of speech (≈0.1 – 4 kHz) were to be vocoded using 8 spectral channels, these would still be resolved. The observation that recognition of vocoded speech appears to asymptote at that point suggests that better spectral acuity is not required for a simple speech recognition task,

although this presumably becomes advantageous in more adverse listening conditions where source separation is required, for example.

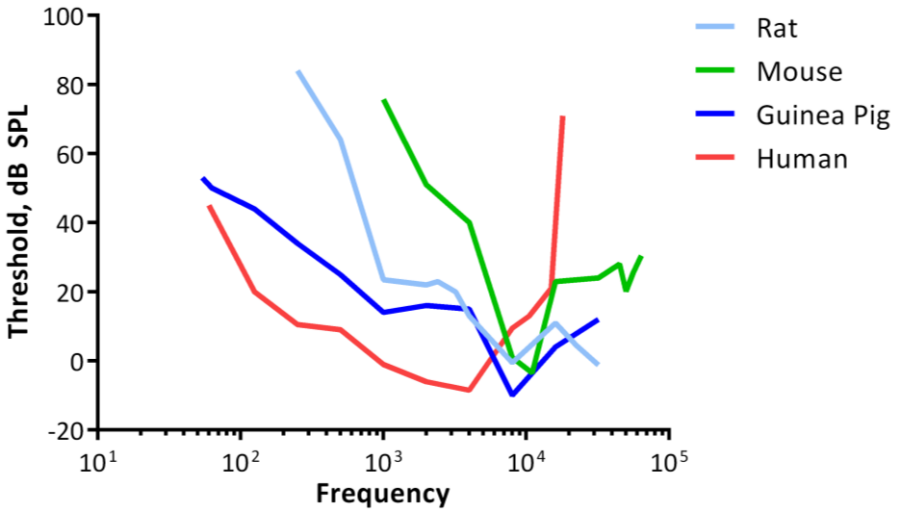

Figure 1.12: Audiograms of several mammalian species. Data reproduced from previously published articles for the Norway Rat (Heffner et al., 1994), Darwin's mouse (Heffner et al., 2001), guinea pig (Heffner et al., 1971) and human (averaged from Sivian and White, 1933 and Jackson et al., 1999).

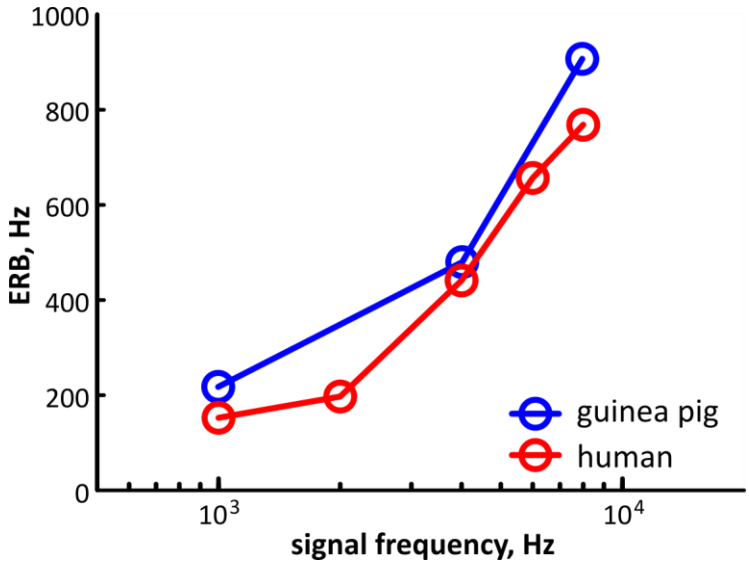

Figure 1.13: Comparison of behavioural frequency selectivity in humans and guinea pigs. Guinea pig data reproduced from Evans et al. (1992) and human data from Oxenham and Shera (2003).

1.8 Summary

Understanding how the salient features of speech sounds are represented by neural responses is an important consideration in the diagnosis of auditory pathologies as well as for designers of auditory prostheses. Of particular interest to developers of cochlear implant processing strategies is how representations are affected by degradations of the input signal.

The study of neural responses at many stages of the auditory system to both simple stimuli such as tones and more complex, speech-like stimuli has elucidated many phenomena that could be utilised by central processing mechanisms to decode spectral and temporal properties of sounds. However, it is not yet clear how the results from such investigations may be synthesised in order to understand the neural bases of complex auditory tasks such as speech recognition. For example, it can be demonstrated that both spike timing and spike rate codes carry valuable information about stimulus properties, but it is also apparent that such representations are far from robust to myriad non-salient variability, such as that caused by inter-talker differences or background noise.

In perceptual studies, it is common practice to manipulate speech stimuli by altering or removing spectral or temporal cues. Such techniques, along with the surprising success of many modern cochlear implant systems, have highlighted the surprising robustness of the auditory system to severe degradations. These studies suggest that speech recognition is fundamentally limited by spectral and temporal content in the input signal, however it is not yet clear what the

limiting role of neural processing is, and how neural representations underlie this surprising robustness. To date, the approach of many electrophysiological studies has been to examine the representations of natural speech. However, the techniques and human perceptual data from the aforementioned psychophysical literature constitutes a powerful toolset for investigating the relationship between neural mechanisms and perceptual phenomena in speech recognition.

Recent work has shown that the phenomenon of robust discrimination of heavily degraded speech sounds is possible in an animal model, and that pair-wise behavioural discriminability is well correlated with the distinctiveness of neural representations in the auditory cortex on multiple timescales, depending on the speech sounds of interest (Ranasinghe et al., 2012). It is not well understood, however, how these central representations emerge from those at the auditory periphery. Further, the nature of a perceptual speech recognition task is often quite different from the way that neural codes are examined. For example, in everyday listening situations the listener must identify speech sounds that may never have been heard before. This implies that internal representations must form categories into which the fundamental building blocks of speech must fall. As the aforementioned studies have used stimulus sets produced by a single talker, it is not clear to what extent their findings only apply to speech token discrimination tasks or if they can be extrapolated to speech recognition tasks in general.

To allow somewhat direct comparison with perceptual studies and more 'real-world' problems faced by the auditory system we will do several things. By adapting techniques similar to those used in previous studies we will predict the discriminability of speech sounds directly from the neural responses. We will also utilise a stimulus set which will enable direct comparison with a human speech recognition task. This comprises a diverse array of phonemes, with multiple exemplars of each. Such a stimulus set will not only allow us to compare neural responses to specific exemplars of an array of speech sounds, but will also enable an investigation into what extent neural representations facilitate the formation of phoneme classes at various stages of the auditory system.

This is a first step towards examining neural coding in the context of a more ecologically valid speech recognition task. Further, since representations of this stimulus set will be acquired at various stages of the auditory processing pathway, comparisons may not only made with perceptual results, albeit in a different species, but also with representations at other auditory nuclei. Thus inferences may be made about which aspects of the input signal are represented at the auditory periphery, maintained at subsequent auditory nuclei and are salient in a speech token classification context.

Chapter 2
Auditory nerve

The only conduit of information from the cochlea to the central nervous system is the auditory nerve, which thus constitutes a bottleneck to the auditory system. It contains a representation of the auditory world in the richest detail and rawest form available to the brain, from which salient information must be extracted. In order for a spoken language to convey information, then, it must comprise symbols that elicit distinct patterns of activity in the auditory nerve. Each symbol can be represented by a broad class of acoustic waveforms that are hugely variable due to the small differences between successive utterances of the same token and, to a greater extent, inter-talker differences. If these symbols are to be successfully decoded centrally, then the code used to represent them in the auditory nerve must also remain robust to these differences.

The representation of speech sounds by single neurons in the auditory nerve has been the subject of physiological and computational modelling studies spanning several decades (e.g. Kiang and Moxon, 1974; Shamma, 1985; Holmes et al., 2004), on the basis of which several encoding schemes have been

hypothesised. These vary from a rate-place based strategy, in which only the spike rate over epochs related to the perceptually relevant time course of spectro-temporal dynamics in speech are considered (Sachs and Young, 1979), to a purely temporal code in which only the precise relative timing of spikes pooled across the entire population plays a key role (Sachs and Young, 1979; Palmer, 1990).

The approach used in many of these studies was to assume a neural code and measure the extent to which it could be used as a basis to reconstruct known acoustic features of the stimulus that are ostensibly perceptually salient, based on empirical evidence from psychophysical literature. For example, the presence of peaks and troughs corresponding to formant frequency locations in the spike rate profile across CF would suggest the viability of a rate-place based code for vowel discrimination, for which representation of the position of the first 2 formants is sufficient (Peterson and Barney, 1952). Alternatively the Fourier transform of ensemble peri-stimulus time histograms (PSTHs) may also be examined to identify the presence of peaks indicating phase-locking to particular spectral components of a speech signal.

In this chapter, the approach taken is to investigate the efficacy of various hypothetical neural codes for the formation of intrinsic phoneme classes in population responses of simulated neural fibres. The use of a computational model not only removes the necessity for the use of extraneous animal models, but enables control over various factors, such as the distribution of fibre CFs, that would otherwise be problematic and would potentially require further

assumptions in the development of normalisation techniques, for example (Loebach and Wickesberg, 2006). The ability of a similar model to reproduce physiological responses to single and double vowel sounds has previously been demonstrated (Holmes et al., 2004), however an investigation into its representation of a range of consonant sounds is not yet extant, to the knowledge of the author.

Another key difference between this and other previous studies on peripheral representations of natural speech sounds is that multiple exemplars of each speech token are used, such that a distinction can be made between putative neural encoding strategies for speech token recognition and those for phoneme recognition. Of particular interest in the present study is how these putative neural codes are affected by degradations analogous to those inherent in auditory prosthetics. One previous study examined the effect of such degradations on the ensemble response of neurons in the auditory nerve of the chinchilla (Loebach and Wickesberg, 2006). They investigated the effect of vocoding on the representations of the stop consonants */b/*, */d/*, */t/* and */p/*. The approach taken was to directly compare representations of the degraded speech tokens to their naturally produced counterparts. In this chapter, the group of consonants is extended further to cover the other categories including fricatives, nasals and approximants. These sounds can rely on a much broader variety of cues for discrimination and enable more general conclusions to be drawn regarding the encoding strategies useful for speech recognition. A larger set of vocoder processing parameters may be used, as time constraints associated with *in vivo* preparations are not applicable, and representations of

the degraded tokens will be investigated with respect to representations of other tokens degraded in a similar way as opposed to those of their naturally produced counterparts.

2.1 Methods

2.1.1 A computational model of the guinea pig auditory nerve

An existing computational model of the auditory periphery was used to simulate responses in the auditory nerve of the guinea pig (Sumner et al., 2002). The use of computational models is feasible in peripheral regions as many properties of individual nerve fibres can be reproduced without the need for the use of animals. The model emulates the frequency response of the external auditory meatus, non-linear filtering by the basilar membrane, transduction by the inner hair cell and adaptation in the auditory nerve. The inner hair cell model is based on biophysical processes, as far as they are known, and includes a simulation of a synapse with an auditory nerve fibre leading to spike generation. Each stage is described in more detail below.

2.1.1.1 Outer and middle ear filtering

In order to reproduce the frequency response of the pinna and external auditory meatus measured by Evans (1972), two bandpass filters are required; one to reflect the overall frequency sensitivity and another to reflect the sharper cutoff at very high frequencies. The first is a second order Butterworth filter cutoffs of 4 and 25 kHz and the second is third order with 0.7 and 30 kHz cutoffs. A constant gain is also applied to the output of these filters such that

the magnitude is representative of stapes velocities in ms^{-1} when the input is sound pressure in μPa (Nuttall and Dolan, 1996).

2.1.1.2 Basilar membrane filtering: A dual-resonance nonlinear filterbank

The conversion of stapes velocity to basilar membrane displacement is modelled by a dual resonance non-linear (DRNL) filterbank (Meddis et al., 2001). Each filter comprises two parallel pathways, one of which applies only linear filtering operations and the other includes a compressive nonlinearity. The output is the summation of these two pathways. A schematic of this processing scheme can be seen in Figure 2.1. The linear pathway consists of a gain stage, a bandpass filter and a lowpass filter. The bandpass filter is implemented as a cascade of gammatone filters. The gammatone filter itself is a model of experimentally derived impulse response functions for auditory nerve fibres (De Boer and De Jongh, 1978), and its output $g(t)$ is realised by the following:

$$g(t) = t^{n-1}e^{-bt}\cos(\omega t)u(t) \quad (1)$$

Where n is the order of the filter, b is related to the filter bandwidth, ω is the centre frequency (in radians) and $u(t)$ is the Heaviside step function (0 for $t < 0$, 1 otherwise). This is followed by a cascade of lowpass filters (4 2nd order Butterworth). The nonlinear pathway comprises a cascade of 3 gammatone filters followed by a compressive nonlinearity. The output of the compressor, $y(t)$ is given by

$$y(t) = \mathrm{sgn}(x(t)) \times \min(a|x(t)|, b|x(t)|^{v}) \quad (2)$$

Where a, b and v are parameters used to fit response characteristics to physiological data. The compressor is followed by a second cascade of gammatone filters and a cascade of lowpass filters similar to that of the linear pathway.

The parameters of the filterbank also vary as a function of frequency (see Appendix for details). The effect of the two parallel pathways is such that the nonlinear pathway dominates the output at low stimulus levels, with the linear pathway increasingly dominating responses as stimulus levels increase and the gain of the nonlinear pathway moves into the compressive region. The centre frequency of the linear pathway also differs from that of the nonlinear pathway and thus the CF of the system is level-dependent reflecting empirical observations. This also results in a decrease in filter sharpness, as measured by both the Q_{10} and filter slopes, with increasing stimulus intensity ({Sumner, 2003 #34}).

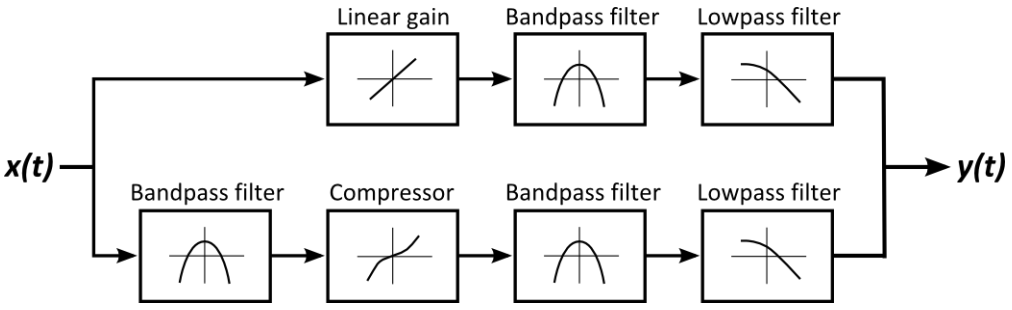

Figure 2.1: Schematic representation of a dual-resonance nonlinear filter. Input $x(t)$ is stapes velocity and output $y(t)$ is basilar membrane velocity. The output is the sum of the linear (upper) and the nonlinear (lower) pathways.

2.1.1.3 Inner hair cell transduction

Transduction by the inner cell is modelled in three stages. The first stage emulates the viscous coupling of the inner hair cell to the basilar membrane. This stage of the model was proposed by Shamma et al., (1986) and describes the relationship between basilar membrane displacement, ω, and cilia displacement, u, using the following equation.

$$\tau_c \frac{du}{dx} + u = \tau_c C \frac{d\omega}{dx} \quad (3)$$

Both C and τ_c are constants. This function has a high-pass characteristic, and also incorporates the empirically observed observation that inner hair cell displacement, and thus receptor potential (RP), appears to be driven primarily by basilar membrane velocity at very low frequencies and by displacement at high frequencies (Sellick and Russell, 1980).

Hair cell displacement causes a modulation of the flow of potassium ions in the surrounding endolymph across the cell membrane leading to depolarisation of the cell. This, in turn, causes more voltage-gated calcium channels on the basal surface of the cell to open initiating an influx of calcium ions. The probability of the release of neurotransmitter vesicles into the synapse is approximately proportional to the cube of the calcium concentration, which is derived from the lowpass-filtered calcium current and based on the relationship between presynaptic calcium current and transmitter release in the squid (Augustine et al., 1985).

The transmitter release rate drives a model of adaptation based on the depletion and reuptake of transmitter vesicles in the synapse. The release of a single transmitter vesicle is determined stochastically and initiates a spike in the simulated auditory nerve fibre. Vesicles in the synaptic cleft are either lost or reprocessed by the hair cell, which maintains a finite store of transmitter quanta. For a detailed description of this process, see Sumner et al., (2002).

2.1.2 Stimuli

2.1.2.1 Speech recordings

Sixteen vowel-consonant-vowel phoneme sequences (VCVs), each spoken by 3 male talkers, were obtained from the speech corpus described by Shannon et al., (1999). Briefly, the recordings were made in a double-walled sound-treated booth with a sample rate of 44.1 kHz and were stored in an uncompressed, 16-bit format. The selected talkers all had standard American Midwest dialect. The consonants used were */b, d, f, g, k, l, m, n, p, s, ʃ, t, ð, v, y* and *z/*, commensurate with the stimulus set used by Shannon et al., (1995), and these were always in an */a/*-consonant-*/a/* context, where */a/* is the open back unrounded vowel as in p<u>al</u>m. A single exemplar of each token produced by each of the randomly selected male talkers was selected from the corpus. All recordings were initially band-limited to between 0.1 and 4 kHz. Each recording was then aligned visually using spectrographic representations generated in Audacity® such that the medial consonant was approximately centred on the point 300 ms from stimulus onset. The recordings were then cropped such that they were 700 ms

in duration and a 10 ms raised cosine ramp was applied to both the onset and offset to reduce spectral splatter.

The level of each stimulus was set to 70 dB SPL, such that it was representative of conversational speech. Since the amount of sound energy in each consonant is dramatically different for those with and those without a stop, for example, setting levels by taking average values over the duration of each VCV could result in providing intensity cues for stimulus discrimination that do not exist in connected speech. Instead, only the vowel portions of each sound were used to determine stimulus intensity, as these were assumed to be relatively consistent. Due to coarticulation, objective determination of phonemic boundaries remains a problem for designers of speech recognition systems, and is outside the scope of this thesis. The vowel portions of each stimulus were therefore approximated visually using spectrographic representations calculated using 512 point analysis window and a 50% overlap.

2.1.2.2 Vocoder

The degree of spectral and temporal degradation was parametrically varied using a noise-vocoder (Shannon et al., 1995) implemented in Matlab®. Figure 2.2 shows a schematic representation of vocoder processing. The raw speech was divided into a number of spectral channels using a bank of bandpass filters (3rd order Butterworth). Each filter overlapped at the point where attenuation was 3 dB below that in the pass band. The speech envelopes were extracted in each channel by lowpass filtering the full-wave rectified narrowband signal in each channel (3rd order Butterworth). The extracted envelopes were used to

modulate narrowband noises produced using identical filters to those in the initial analysis stage. A single noise token was used to produce every speech sound, such that the cues for discrimination could only arise from differences in the input speech signal, and not the noise tokens themselves. The modulated noises were then processed through these filters a second time following modulation in order to attenuate signal energy falling outside of the channel frequency band as a result of the modulation process.

The degree of spectral degradation was varied by altering the number of spectral channels, which was set to 1, 2, 4, and 8. Temporal degradation was varied by changing the cutoff of the lowpass envelope extraction filter, which was set to 16 and 500 Hz. Single channel 50 and 160 Hz envelope conditions were also included. It should be noted, however, that for the 4 and 8 channel conditions envelope modulation is further restricted in low frequency channels by the filtering stage following modulation. The same analysis band filter cutoff frequencies as those used in Shannon et al. (1995) were used for the 1, 2, and 4 channel conditions. For the 8 channel condition, the channels defined for the 4 channel condition were simply bisected on a logarithmic scale, as this condition was not included in the aforementioned study (see Table 2.1).

$N_{channels}$	f_c1	f_c2	f_c3	f_c4	f_c5	f_c6	f_c7	f_c8	f_c9
1	100	4000							
2	100	1500	4000						
4	100	800	1500	2500	4000				
8	100	283	800	1095	1500	1937	2500	3162	4000

Table 2.1: Cutoffs used for bandpass filters in the noise vocoder. The same cutoffs were used in the initial analysis stage as those used to produce the narrowband noise carriers, as well as post-modulation filtering.

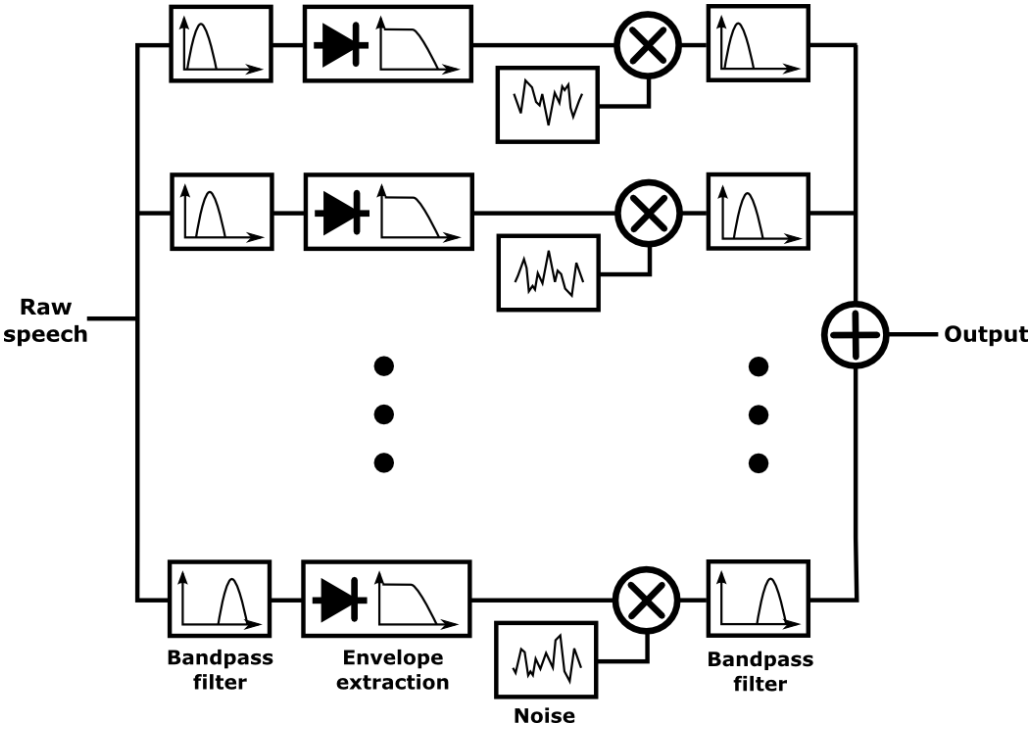

Figure 2.2: Schematic representation of noise vocoder processing. Noise carriers are produced by filtering a broadband noise token through the same bandpass filterbank as the raw speech.

2.1.3 Neural classifier

In order to quantify the discriminability of the auditory nerve representations of the speech tokens, a neural classifier was developed in Matlab®. The classifier was based on the nearest neighbour classifier described by Shetake et al., (2011). Based on the neural representation of a single presentation of a given speech sound, the classifier identifies the medial consonant produced by the talker. For each stimulus presentation, PSTHs were produced for each simulated fibre using 1ms bin widths. These PSTHs were combined together to form a pictorial representation of neural activity, or neurogram, in which each row corresponds to a single nerve fibre, each column corresponds to a single temporal bin and the value at each pixel represents the number of spikes. Figure 1.8 (lower panel) shows such a neurogram comprising simulated auditory nerve responses from 100 medium spontaneous rate fibres responding to natural speech.

The classifier training set comprised either the average neurograms for each stimulus for a single talker task, or these average neurograms combined across each of the talkers, producing a single response pattern for each consonant. If representations were simply averaged across talkers, this would imply that any putative decoding system, in this case represented by the classifier, has access to the absolute onset time of each stimulus. Since the brain must perform this task based on internal representations alone, responses were allowed to shift temporally relative to each other such that the total Euclidean distances between them were minimised. Temporal shifting may be thought of as

allowing the neurograms to slide left and right. The relative shift τ of neurogram x relative to neurogram y is given by the following:

$$\underset{\tau\in[-T,T]}{\operatorname{argmin}} \sum_{n=1}^{N}\sum_{m=1}^{M}\left[x_{n,m+\tau}-y_{n,m}\right]^2 \tag{4}$$

Where T is the maximum permissible relative temporal shift, N is the number of simulated fibres and M is the number of bins over which the distance is calculated. Since nerve fibres fire spontaneously, responses could not be zero-padded to allow for long duration temporal shifts without introducing spurious inhibitory responses. For this reason, the maximum shift T was set to 100ms. As each stimulus was 700ms long, only the central 500 bins were used to calculate the Euclidean distance. Each of the three average responses patterns for a given consonant was selected in turn, and the other two were allowed to shift relative to it. A 3×3 matrix of minimum Euclidean distances was computed, in which each column represented minimum Euclidean distance with a particular response pattern anchored in time. From this, the anchored neurogram was determined to be the one corresponding to the column with the minimum summation. Each neurogram in the training set was normalised by the number of presentations and the number of talkers.

The training set was produced with a single presentation of each stimulus removed from the data set. The responses to this single presentation comprised the test set. Each neurogram produced from this single presentation of a single stimulus was compared to each of the average neurograms using the same temporal shifting procedure as described above. The neurogram was classified as the consonant whose average neurogram it was closest to given

any relative temporal shift from -100 to 100 ms. As the consonant that evoked each test neurogram was known *a priori*, discriminability in units of percent correct could then be calculated. This training and testing procedure was repeated for each presentation, generating 10 values of percent correct for each vocoding condition.

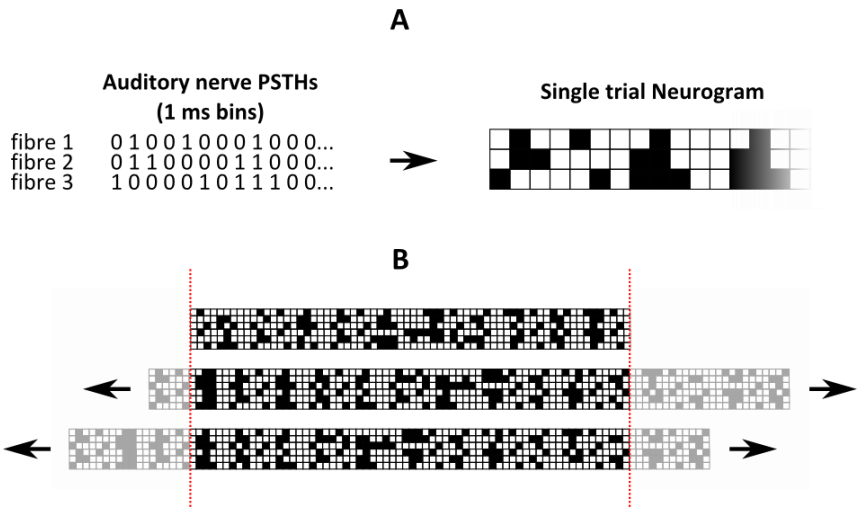

Figure 2.3: Schematic of neurogram construction. PSTHs are produced using a 1ms bin width and averaged across presentations. The resulting neurograms are then combined across talkers with relative temporal shifts such that the distances between them are minimised.

2.1.4 Information transfer analysis

Confusion matrices were also produced from the classifier responses. In order to asses which aspects of the phonemes were successfully transmitted for each vocoding condition, information transfer analysis (Miller and Nicely, 1955) was carried out on these confusion matrices. Information transfer is the ratio of successfully transmitted information to input entropy. Entropy in this context

refers to Shannon entropy H (Shannon, 1948), which can be thought of as a measure of uncertainty of a random variable X and is given by:

$$\mathrm{H(X)} = -\sum_{i} P(x_i) \log_2 P(x_i) \tag{5}$$

Where $[x_1 \cdots x_n]$ are all possible values of X. When the logarithm is in base two, entropy is measured in bits and corresponds to the average number of binary bits required to communicate one symbol of X. Given that there are 16 stimuli each with an equal probably of occurring, the entropy corresponding to the stimulus set can be expressed as 4 bits. The number of bits successfully transmitted is given by:

$$\mathrm{I} = -\sum_{x,y} p_{xy} \log_2 \frac{p_x p_y}{p_{xy}} \tag{6}$$

Where p_{xy} is the probability of observing x given input y. Neither this value nor the probability of observing a given output is known, and must be estimated from the confusion matrices. In practice, information transfer can be calculated using the following equation.

$$\mathrm{IT} = \frac{-\sum_x \frac{n_x}{N} \log_2 \frac{n_x}{N} - \sum_y \frac{n_y}{N} \log_2 \frac{n_y}{N} + \sum_x \sum_y \frac{n_{xy}}{N} \log_2 \frac{n_{xy}}{N}}{\sum_x p_x \log_2 p_x} \tag{7}$$

Where n_x is the number of times stimulus x was presented, n_y is the number of times classifier output y was observed, n_{xy} is the number of times input x and output y were observed together and N is the total number of observations, which, in this context, corresponds to the total sum of the confusion matrix. Bespoke information transfer analysis code was implemented in Matlab®.

2.2 Results

2.2.1 Natural speech

Responses to 16 VCVs each produced by 3 male talkers were simulated in 100 auditory nerve fibres with CFs logarithmically spaced between 0.1 and 5 kHz. The model incorporates stochastic spike generation to emulate the stochastic nature of the conversion of membrane potentials to action potentials in the inner hair cell to auditory nerve synapse, so 10 responses were simulated for each nerve fibre. Figure 2.4 shows the auditory nerve neurograms produced from each of the 100 simulated fibres and averaged across the 10 repetitions. The neurograms depict the entire spatio-temporal pattern of activity across each of the 100 sites using 1 ms temporal bins. The ensemble temporal representation is shown above each neurogram in red and the total spike count at each site is shown by the blue bars. The similarity of the neurograms to spectrographic representations of the stimuli is reflected in the similarity between the ensemble PSTHs and stimulus envelope. Differences in responses are also apparent in the spike count profile histograms. For example, the fricative */aʃa/* elicits proportionally higher spike counts in fibres with high CFs compared with the nasal */ama/*. The effect of spectral filtering by the outer and middle ear is also apparent, as fibres with the lowest CFs fire proportionally far less than those with higher CFs.

The classifier was initially trained and tested separately for each individual talker. Each of the three representation types depicted in Figure 2.4 were used as inputs; the ensemble PSTH representing a purely temporal code, the spike

count histograms a rate-place code and each neurogram a combination of the two. For both the neurogram and the ensemble PSTH, a 1 ms bin size was used. The classifier training set for each case comprised the aforementioned representations averaged across 9 of the simulated responses with the remaining responses forming the test set. The classifier was able utilise each of these representations to identify the correct consonant with a 100% success rate. While this high performance may seem surprising, this token recognition task is a trivial one for humans to perform and this exemplifies the highly redundant nature of the representation of speech sounds in the auditory nerve for this task. A multiplicity of codes could be utilised, and the stochastic nature of spike generation is not such that it offsets the differences between the responses to each individual token.

The results discussed so far pertain to a token discrimination task, in which classifier performance is limited purely by the ability of the auditory nerve fibres to faithfully represent stimulus features in a way that is consistent across multiple presentations. In the real world, listeners are rarely presented with the same stimulus more than once, and therefore phonemic class boundaries must be defined such that classification is robust to variation across multiple exemplars and listening conditions. A classifier training set was produced that comprised neural representations that had been combined across multiple talkers (see 2.1.3). The test set still comprised responses to single presentations of individual tokens. This process was also applied for all combinations of 2 out of the 3 male talkers

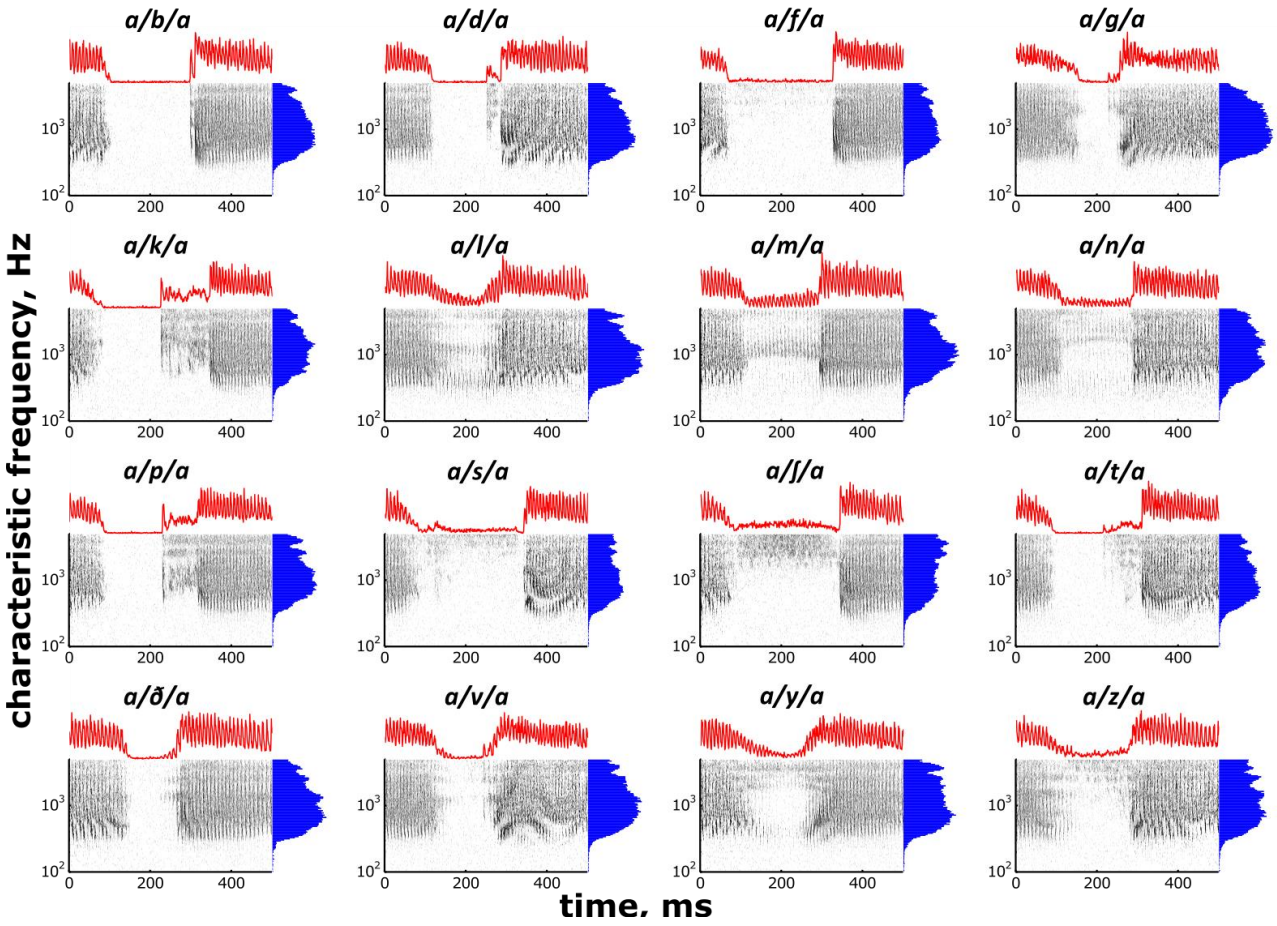

Figure 2.4: Average neurograms produced from 100 simulated auditory nerve fibres in response to the 16 VCVs produced by a single talker. Red lines show the ensemble rate profile and blue histograms show the spike count for each individual fibre.

The effect of number of talkers on the discriminability of these representations is shown in Figure 2.5 A. Both representations that incorporate temporal cues are robustly discriminable, even when combined across each of the 3 talkers. The discriminability of rate-place representations, however, is heavily compromised by additional talkers. Where the classifier was trained using representations combined across pairs of talkers, there is a high degree of variability in the discriminability of population-based representations, showing that the discriminability depends on the particular pair of talkers used. This is probably due to differential magnitude of inter-talker differences between various acoustic features such as fundamental frequency or syllabic rate for example. This demonstrates that, for natural speech, the ensemble temporal representation alone is sufficient for high performance on this consonant discrimination task.

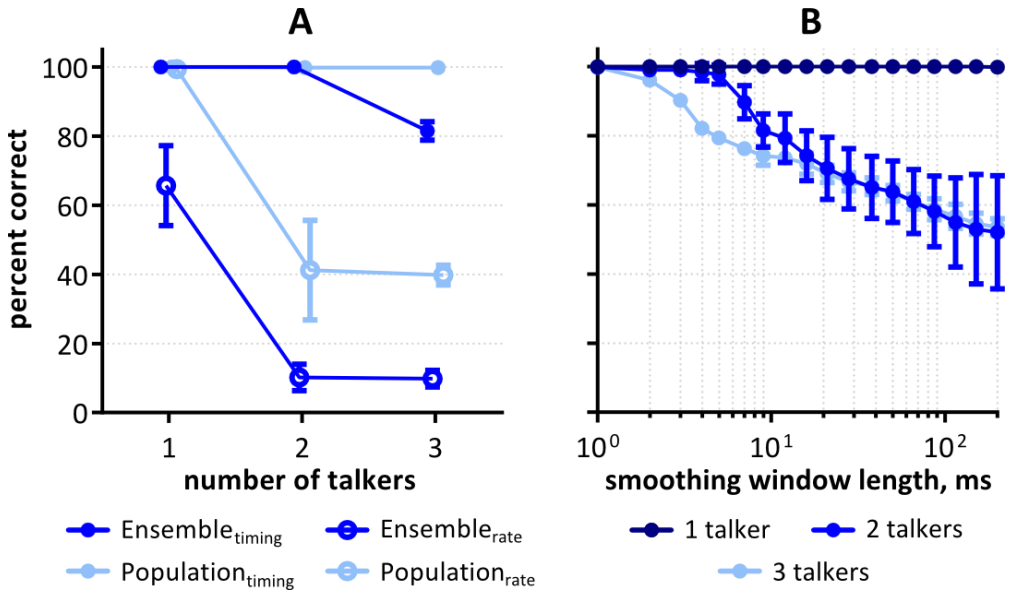

Figure 2.5: The effect of number of talkers on discriminability of simulated auditory nerve representations of consonants (**A**) and the effect of temporal smoothing on the discriminability of spatio-temporal representations (**B**).

Unlike vowel sounds, many consonants are inherently dynamic stimuli characterised by spectro-temporal changes over time. In this context, it is not surprising that long encoding windows are suboptimal for discriminable representations at this peripheral level. To examine the relationship between encoding window duration and discriminability of neural representations, the spatio-temporal response patterns were temporally smoothed by convolution with a rectangular window of varying duration. The classifier was then trained and tested using these degraded representations, and the results are shown in Figure 2.5 B. For the single talker paradigm, discriminability is not affected by temporal degradation of the response. Thus, even the rate-place pattern of mean firing rates is sufficient for recognition of individual speech tokens. For multi-talker paradigms, temporal smoothing decreases discriminability monotonically, although with fewer talkers, discriminability is more robust to small amounts of degradation of the order of a few milliseconds. These data show that should central mechanisms have access to millisecond precise spike timing information, phonemic class boundaries may be defined in the resulting high dimensional representation space that are robust across inter-talker differences within a small group of the same gender. It should also be noted that this is the case even with a simulated neural population that is very small compared to the population observed in the animal.

Each simulated nerve fibre is optimally sensitive to a range of frequencies that is comparatively narrow with respect to the bandwidth of the speech stimuli. It may be that fibres with particular ranges of characteristic frequencies produce more salient responses than others for discriminating this set. Although

psychophysical procedures may highlight spectral regions particularly salient for speech discrimination, this approach does not easily translate to regions in the cochlea, for which an examination of neural representations is more efficacious, and an important consideration for cochlear implant placement and processing strategies. It may also be the case that auditory nerve representations have a high degree of redundancy such that only very small populations of fibres are required to produce discriminable representations, with the extra information only becoming salient in more challenging listening conditions. The way that discriminability is affected by such degradation of the simulated neural representations would further understanding of the role that cochlear de-afferentation plays in speech recognition.

In order to test this, the ability of individual nerve fibres to discriminate speech tokens was also examined. Classification was performed in the same way as before, except the training set comprised PSTHs averaged across $N-1$ simulated responses (where N is the total number of stimulus presentations) of single nerve fibres using a 1ms bin size and the test set comprised individual spike trains. In order to investigate a number of putative optimal encoding windows in each frequency range, spike trains were parametrically smoothed prior to classifier training and testing by convolution with a rectangular window of varying duration.

The discriminability of single fibre representations of speech tokens was initially evaluated for each of the multiple talkers individually and the results were

averaged across them. Discriminability as a function of fibre CF is shown in Figure 2.6, which shows the maximum values for any given smoothing window.

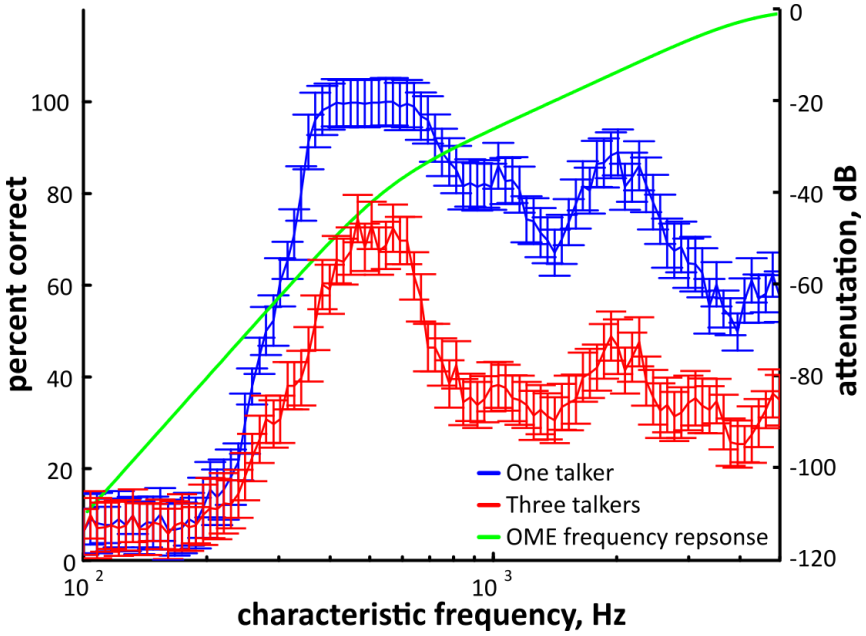

Figure 2.6: Discriminability of auditory nerve representations of consonants as a function of fibre CF for single talker (blue) and multi-talker (red) paradigms. Outer-middle ear frequency response shown in green.

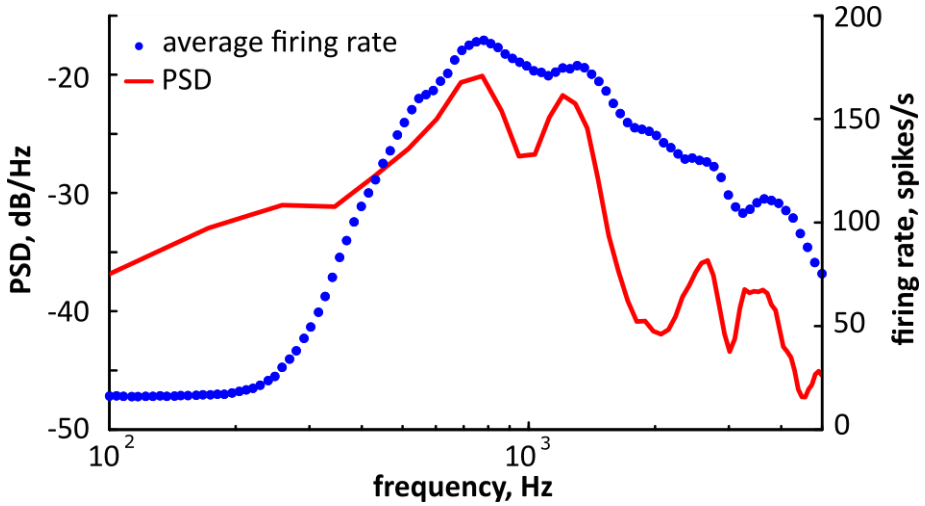

Figure 2.7: Average firing rates of simulated auditory nerve fibres (●) and average power spectral density of stimulus set (—).

The classifier was unable to utilise the responses of fibres with CFs below approximately 250 Hz as these sites were not responsive to the stimuli due to the frequency response of the outer and middle ear (OME) filters overlaid in green and did not produce firing rates that differed significantly from spontaneous activity. For those fibres that were responsive, the ones producing the most discriminable responses were maximally sensitive to pure tone frequencies of approximately 400 - 600 Hz, with which the classifier was still able to perfectly discriminate the speech tokens. Optimal discriminability then decreased for sites with higher CFs, excepting a secondary, smaller peak corresponding to CFs of approximately 2 kHz. These two maxima correspond to peaks in the average power spectrum for the stimulus set. For the multi-talker paradigm, no individual fibre was able to perfectly discriminate between the consonants, although optimal discriminability as a function of fibre CF followed a similar pattern, with CFs of approximately 500 – 600 Hz being particularly efficacious for the task. Figure 2.7 shows the average power spectrum of the stimulus set and the average firing rate of each fibre. While the average spectrum is reflected by the average firing rates, this does not fully account for the presence of peaks in Figure 2.6, which suggests that these correspond to particularly informative spectral regions for this speech corpus.

The fact that some individual fibres are able to perfectly discriminate speech tokens produced by a single talker may be indicative of a representation of stimulus fine structure, which becomes less salient when average responses are combined across multiple talkers explaining the reduction in performance for the multi-talker set. However, ensemble representations may be used to

perfectly discriminate the consonant sounds. It is not clear from the above analysis, however, whether this is due to the benefit of within channel or across channel integration. In other words, it is not clear whether stochastic undersampling within narrow spectral regions is alleviated by the addition of multiple fibres with similar CFs that effects the recovery of discriminability for the multi-talker task, or if representations from fibres with disparate CFs are critical. If the former were true, this would suggest that temporal cues alone would be sufficient to perfectly discriminate between the stimuli from peripheral representations. If the latter is true, then this suggests that spectral cues are critical for formation of consonant classes. The effect of number of fibres on discriminability for CFs closest to 500 Hz is shown below in Figure 2.8. This shows that, by combining the responses of only a small number of fibres with similar CFs, discriminability reaches asymptotic performance.

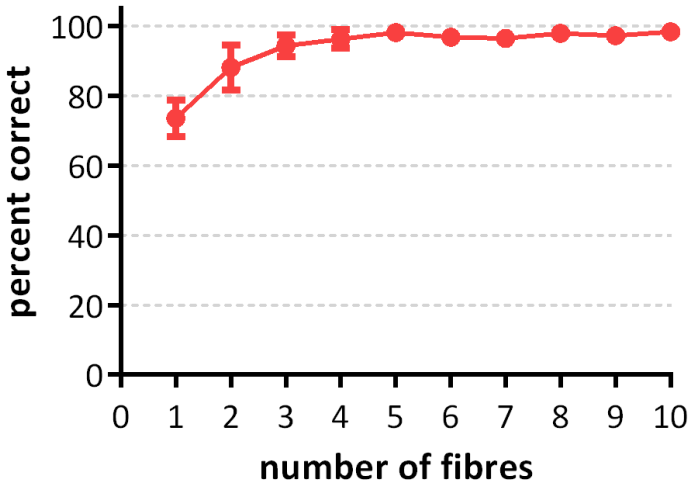

Figure 2.8: The effect of number of fibres on discriminability of auditory nerve representations of consonants (multi-talker paradigm). Fibres were selected from those with CFs closest to 500 Hz, which ranged from 415 to 592 Hz.

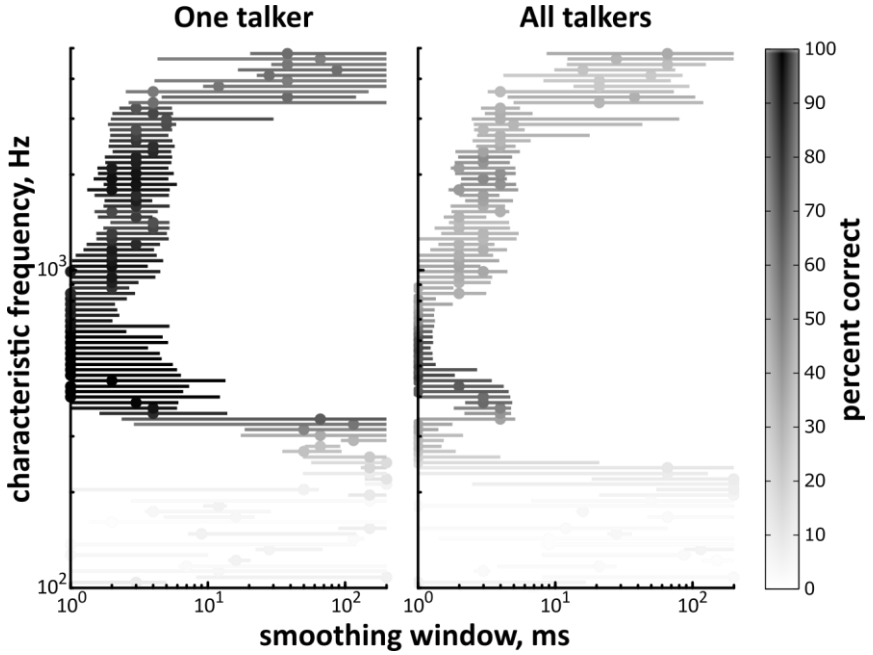

Figure 2.9: Optimal smoothing windows for discriminability of auditory nerve representations of consonants. Circular makers show the optimal smoothing window and bars represent the range over which discriminability is greater than 90% of the maximum.

The optimal smoothing window also varied as a function of fibre CF. Figure 2.9 shows the discriminability of responses for individual nerve fibres as a function of both fibre CF and smoothing window length. The peak discriminability of responses of fibres with CFs of around 600 Hz can clearly be seen, which corresponds to smoothing windows with durations of only a few milliseconds. For fibres with CFs above approximately 3 kHz, discriminability is optimal for the longest smoothing windows. This suggests that the salient stimulus features in these different frequency ranges are encoded in auditory nerve responses at different timescales. This shows that the high CF fibres do not precisely lock to salient signal components and therefore only signal the presence or absence of high frequency signal energy in their firing rates.

In this section we have demonstrated the application of a novel classifier for the automatic alignment and classification of responses to medial consonants in an */a/*-consonant-*/a/* context. A distinction was made between potential neural decoding strategies for token and for consonant recognition; primarily that the former is predominantly related to the reliability of responses across multiple presentations given an assumed neural code and the latter requires these representations to form classes in a high dimensional representation space. It was demonstrated that the discriminability of simulated auditory nerve representations of a set of 16 speech tokens was robust to heavy temporal degradation of the spike trains, and individual nerve fibres carry enough information about acoustic features of the stimuli for the classifier to discriminate between them, despite not being provided with the absolute stimulus onset time.

Representations that maintain the millisecond precise timing of action potentials constitute the optimal code for discriminability, even when the classifier training set was combined across multiple talkers. This may be indicative of the representation of temporal fine structure in the stimulus that remains a viable cue despite the combined multi-talker representations. However, it is important to draw a distinction between the timescale of the neural code and the stimulus features it represents; it may be that this is indicative of the representation of slower envelope fluctuations in the precise timing of neural spikes. In the following section, a noise vocoder is used to parametrically degrade the speech sounds. This removes the fine structure of the speech and uses frozen noise tokens such that representations of the fine

structure of the noise carriers do not provide spurious cues for discrimination. This also enables the temporal and spectral fidelity of the acoustic signal to be manipulated. The effect of the manipulations on the performance of the neural classifier for the aforementioned putative neural codes will be compared with existing psychophysics literature to distinguish between those which are optimal for a template matching classifier that has access to the full simulated representations and those that are accessible perceptually. Further, it has been demonstrated in this section that the representation of natural speech tokens in the auditory nerve is highly redundant. Vocoding allows the extent to which this redundancy is inherent in the speech tokens themselves to be explored.

2.2.2 Vocoded speech

Natural speech is a complex and highly redundant acoustic signal. Determining the nature of the redundancies in the context of auditory processing remains a problem, and is of interest to designers of auditory prostheses where logistical limitations, such as electrode size and current spread, as well as technical trade-offs, such as battery life vs. processing power, exist. One approach taken in the psychophysics literature is to parametrically degrade speech signals using a vocoder to examine the effect on perceived discriminability. To what extent speech recognition performance is limited by information contained in the speech waveforms themselves, the acuity of their peripheral representations or processing in the central auditory system is not understood.

Section 2.2.1 demonstrated the application of a novel neural classifier to quantify the discriminability of simulated auditory nerve representations of

medial consonants in units of percent correct. The classifier was able to identify the correct consonant from single trial activity patterns for both single- and multi-talker paradigms. For consonants produced by a single talker, neural codes on multiple timescales could be used for perfect discrimination of these speech tokens. Where the classifier training set comprised representations that had been combined across multiple talkers, the maintenance of spike timing in the neural responses was critical for robust discriminability, however it is not clear whether this is representative of the importance of fine structure or high rate envelope cues in the stimulus, or indicative of a temporal code for the representation of acoustic features over larger epochs. To investigate this, the stimulus set was parametrically degraded using a noise vocoder (see 2.1.2.2). The number of vocoder channels was set to 1, 2, 4 and 8 and the envelope extraction filter was set to 16 and 500 Hz producing 8 vocoded speech conditions. The same neural classifier was used to determine neural discriminability in units of percent correct.

The discriminability of the neurograms with no temporal smoothing and spike rate representations for each of the vocoding conditions for both a single talker and multi talker classification paradigm is shown in Figure 2.10. The results from the single-talker paradigm will be discussed first. Despite the high degree of spectral and temporal degradation, discriminability is very high in all conditions. For all vocoding conditions, representations that preserve spike timing are more discriminable than those that only preserve spike counts over the entire duration of the consonant. For rate-based representations (upper panel, dark grey), an increase in discriminability between the 1 and 2 channel

conditions is apparent for both the 16 and 500 Hz envelope conditions (one-sided, paired-samples, t(29)=, p<0.001; t(29)=, p<0.001) but there is no effect of envelope bandwidth. There is also no increase when spectral resolution is increased from 2 to 4 and from 4 to 8 channels, which could be attributable to a ceiling effect.

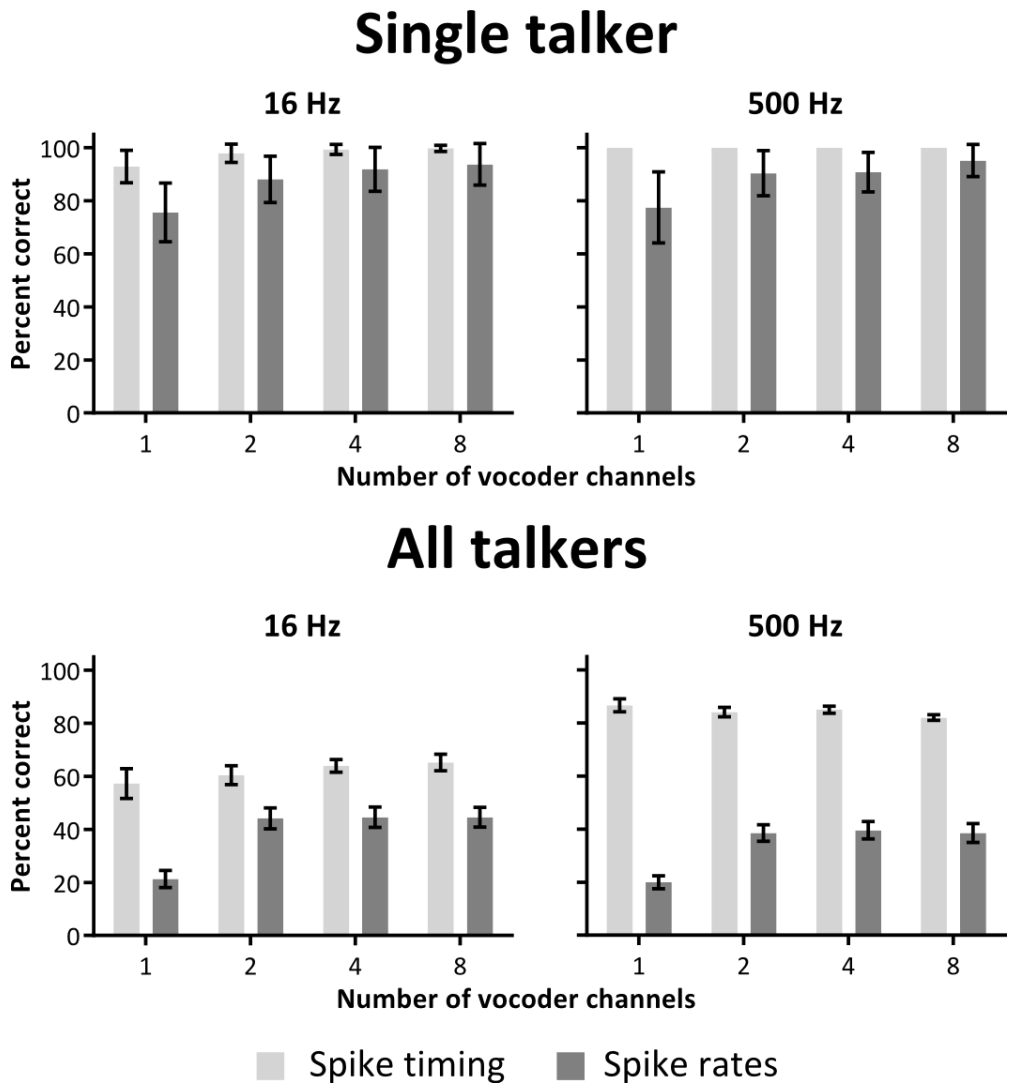

Figure 2.10: Discriminability of auditory nerve neurograms constructed using millisecond precise spike timing ( ) or spike rates only ( ). Representation comprised 100 simulated auditory nerve fibres with CFs logarithmically spaced between 0.1 and 5 kHz. Classifier was trained on responses averaged across 9 repetitions and tested on 1.

The high discriminability of these speech tokens may seem to be in conflict with speech recognition performance for these types of degradations in humans. Psychophysical studies using noise vocoding have demonstrated that, at least where only 1 or 2 spectral channels are available, speech recognition performance is poor (Shannon et al., 1995; Xu et al., 2005). However, it has also been shown that consonant sounds that can result in perceptual confusions when degraded elicit distinct response patterns in the auditory nerve (Loebach and Wickesberg, 2006) showing that these confusions are not a result of insufficient acuity at the auditory periphery. In such a token recognition paradigm, deficits in perceptual discrimination more likely result from central processing mechanisms. If we assume that temporal processing in the human auditory nerve is not very poor compared to that of this model, perceptual difficulties may arise partially from decreased acuity of stored internal representations.

It is also common in human speech discrimination tasks to ask listeners to identify the corresponding natural speech token from its degraded counterpart. This is distinct at least conceptually from a task that requires the listeners to form new internal representations for each of the degraded sounds as it requires a mapping from the vocoded speech tokens to internal representations of classes of spectro-temporally complex natural speech sounds. This mapping process may also be a bottleneck in performing discrimination of these degraded sounds, which would not be a problem faced by cochlear implant users who were implanted prelingually.

The lower axes of Figure 2.10 show the same results as those discussed above but for the multi-talker classifier training and testing paradigm. There is a reduction in discriminability across all conditions relative to the single talker data. These results demonstrate that both millisecond precise spike-timing and spike-rate representations are robust across multiple presentations of single speech tokens, but do not remain robust across multiple talkers. This is suggestive that, in order to reach the levels of discriminability shown perceptually, further processing is required to form representations of cues that might be robust across talkers such as relative amplitudes across channels.

The discriminability of timing-based representations does not increase with the addition of more than 1 spectral channel. The discriminability of rate-based representations increases when the number of spectral channels is increased from 1 to 2, but not beyond. There is also an increase in discriminability across all vocoding conditions when the envelope extraction filter cutoff is increased from 16 to 500 Hz, but only for those representations that preserve spike timing information.

The lack of increase in discriminability with increasing spectral resolution is in contrast to measured perceptual effects even for a multi-talker task. However, mean discriminability values calculated from rate-based representations are not dissimilar from those for natural speech with as few as 2 spectral channels. This may be indicative of a ceiling effect inherent in the use of rate representations alone, which may be as a result of the long (500 ms) window over which spikes are summed. This window contains a significant overlap with

the vowel context, which is typically more intense than the medial consonant and may dominate the response. In order to investigate the efficacy of intermediate encoding windows for discriminability of degraded consonants, a similar technique to that described in 2.2.1 was used, and the results of this analysis is shown in Figure 2.11. Similarly to the natural speech stimulus set, neural discriminability is robust to temporal degradation of the spike trains for the single talker paradigm regardless of the degradation of the stimulus. Even where speech envelope modulations above 16 Hz were attenuated, representations using millisecond precision were still perfectly discriminable.

This could be indicative of a temporal code for representing the slower envelope modulations, or of a representation of intrinsic envelope fluctuations in the noise. Since the same noise token was used for each speech token, this would only be possible with the classifiers ability to temporally shift representations. If this shifting mechanism is driven primarily by the slowly changing intensity of the speech envelopes, then the fine structure and high rate envelopes of the noise would then provide an additional cue for discrimination. For the multi-talker paradigm, maximum discriminability is still less than that in the single-talker paradigm, but representations are no longer robust to a high degree of temporal degradation. For 16 Hz envelope conditions, discriminability is relatively constant until smoothing windows become greater than approximately 60 – 80 ms. This is commensurate with the reciprocal of the 16 Hz envelope bandwidth (64 ms), and reflects the results from psychophysical studies, which indicate that intelligibility of speech sounds begin to degrade where only envelope cues at modulation rates less than 16 Hz

are available (see 1.5). For long smoothing windows, there is also no benefit from the addition of high rate envelope cues. These do not elicit differential firing rates in response to each consonant sound that are robust across the multiple talkers. However, discriminability is increased relative to the 16 Hz conditions when spike timing is conserved.

Optimal discriminability for each of the vocoding conditions for any smoothing window duration is shown in Figure 2.12, along with the data from Shannon et al (1995), who presented an analogous stimulus set to human normal hearing listeners in a 16-alternative forced choice paradigm. As suggested previously, neural discriminability increases where the number of spectral channels increases from 1 to 2 for the 16 Hz conditions. However, no further increases are observed. One possible explanation for this is that as the number of channels is increased, the number of fibres per vocoder channel decreases. For the single channel condition, all 100 fibres represent the envelope of the broadband speech signal such that the degraded nature of the neural representation due to the stochastic nature of spike generation is offset by multiple samples. In the 8 channel condition, however, only 12 fibres have CFs that fall within each frequency band on average in the 8 channel condition. There is also a disproportionate number of fibres that fall into the low frequency channels due to the way the spectrum was split by Shannon et al. (1995), which was reproduced here (see Table 2.1).

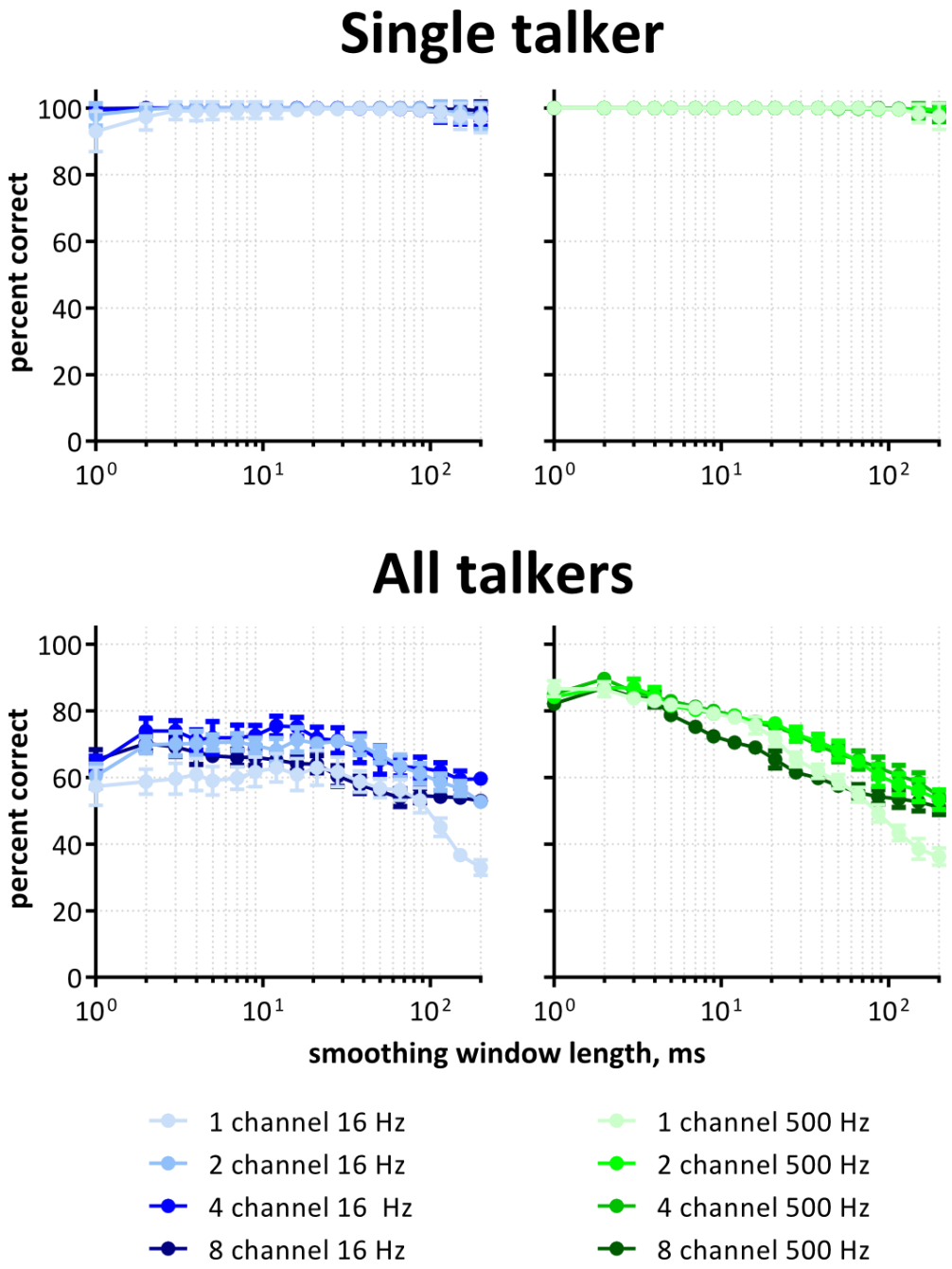

Figure 2.11: The effect of temporal smoothing on the neural discriminability of auditory nerve representations of 16 VCVs.

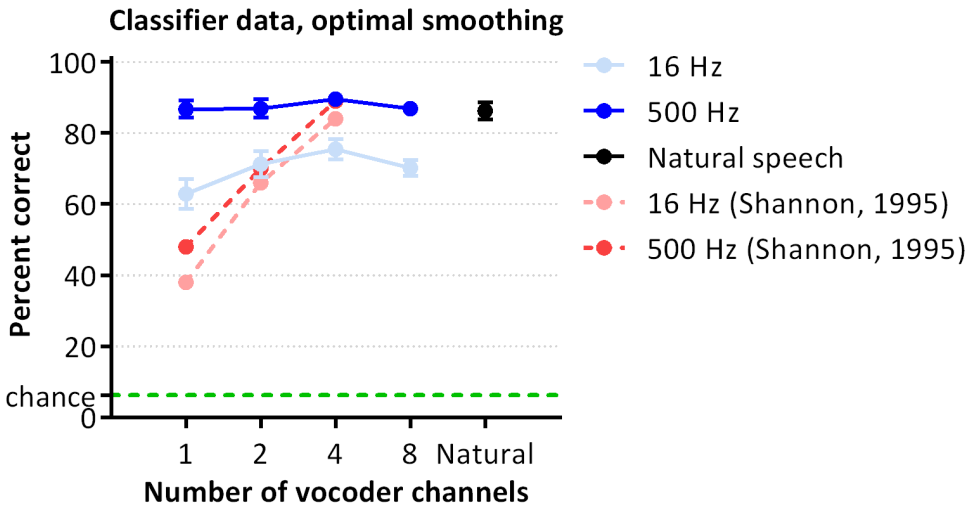

Figure 2.12: Neural discriminability of auditory nerve representations of vocoded speech. Data shown corresponds to the peak discriminability for any smoothing window duration for each vocoding condition.

Altering the number of spectral channels also has secondary effects on stimulus properties; as channel bandwidth decreases, the intrinsic envelope fluctuations of the noise carrier increase in intensity. The signal is, of course, bandpass filtered by the cochlear filterbank in all conditions, so these are still present even in the 1 channel condition. However, as the bandwidth of the vocoder channels approaches that of the cochlear filters the noise envelope begins to become increasingly dominant thereby inhibiting discriminability based on auditory nerve representations alone. Such a condition could be alleviated by subsequent cross-channel integration in upstream nuclei, for example.

Where high rate envelope cues are not attenuated, there is a large increase in neural discriminability such that there is no difference between that for the single channel vocoded condition and natural speech, demonstrating that high rate envelope cues increase the discriminability of auditory nerve representations of the 16 VCVs even after combining across multiple talkers. These same cues, however, do not appear to increase perceived discriminability to the same extent. Although the presence of high rate envelope cues does increase perceived discriminability robustly (Van Tasell et al., 1992; Shannon et al., 1995; Xu et al., 2005), none of these show a recovery of speech discrimination performance to that of natural speech on the basis of temporal cues alone.

Earlier analysis of simulated auditory nerve responses to natural speech sounds showed that, for a single talker paradigm, individual fibres generated responses that were sufficiently reliable to allow perfect discrimination performance by the classifier provided a high degree of temporal precision was preserved in the responses (see Figure 2.9). This could be indicative of two possibilities. Firstly, this could indicate the importance of stimulus features occurring over short timescales such as high rate envelope cues or temporal fine structure. Secondly, this could indicate a temporal code whereby the encoding window for optimal discriminability was short relative to the "integration window"; the length of time preceding a given point in a neuron's response pattern over which changes in the input may meaningfully affect spiking responses (Theunissen and Miller, 1995).

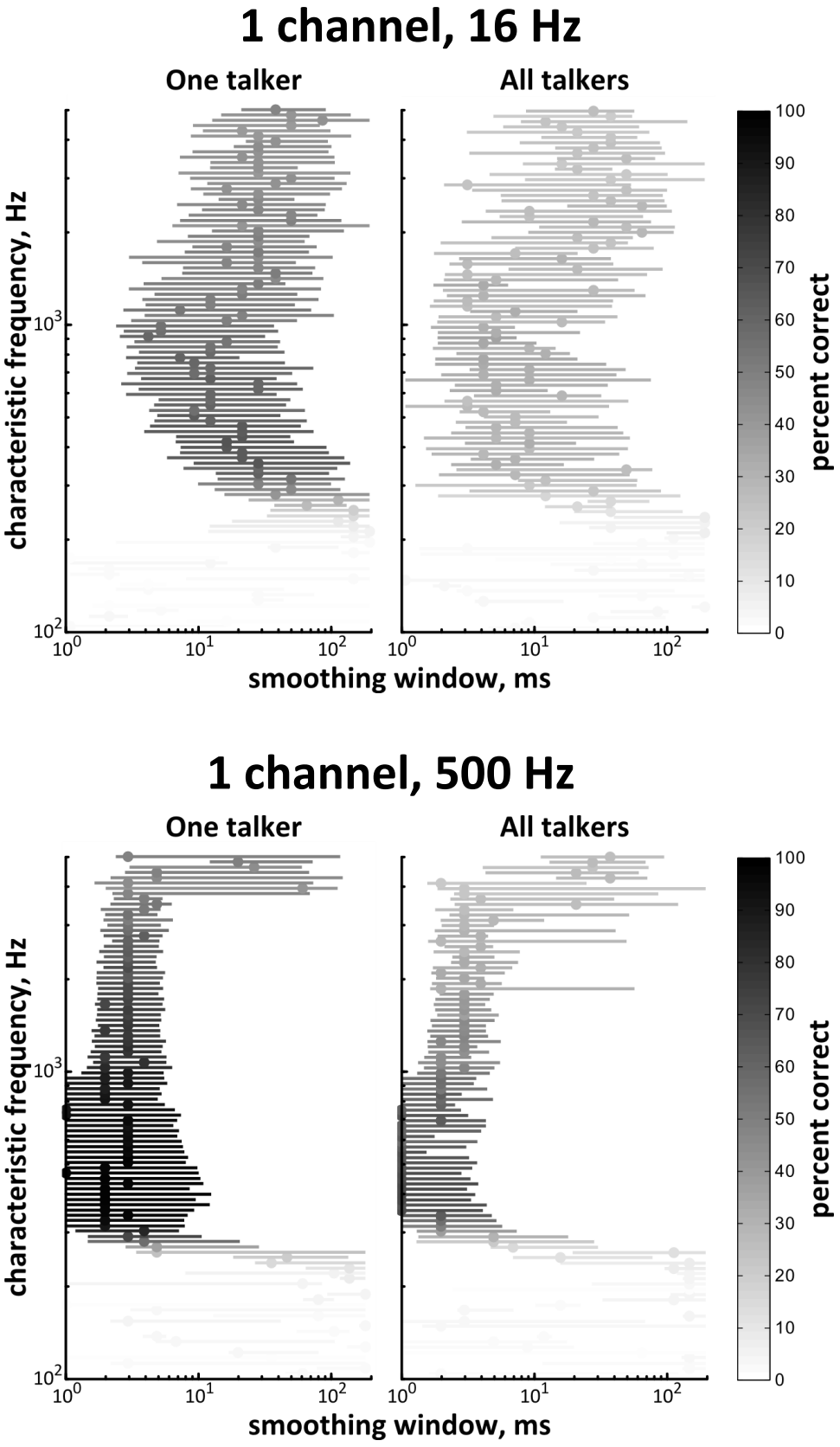

Figure 2.13: Optimal smoothing windows for discriminability of single fibre representations of single channel vocoded speech.

A similar analysis was carried out on the responses to single channel vocoded speech and the results are shown in Figure 2.13. In these speech sounds, all fine structure cues are removed and high rate envelope cues are either attenuated (upper axes) or not (lower axes). Where high rate envelope cues are attenuated (16 Hz condition), no individual fibre produces responses that may be used to reliably discriminate between the consonant sounds. Where they are not attenuated, the picture looks very similar to that for natural speech where all fibres with CFs within the range where the stimulus is attenuated by neither the outer-middle ear filters nor the 0.1 to 4 kHz pre-processing bandpass filter enable near perfect discriminability. This shows that temporal fine structure cues are not integral to classifier performance, but also that high rate envelope cues increase spike timing reliability across presentations. The fact that the discriminability of the population responses to the 1 channel, 16 Hz speech does reach 100% despite high rate envelope cues being attenuated shows that the reduced reliability of spike timing is alleviated by the presence of multiple fibres within the same channel.

Information transmission analysis (Miller and Nicely, 1955) was carried out the confusion matrices resulting from classification. The phoneme classes used are shown in Table 2.2, and the resulting information transmission for each of the three features is shown in Figure 2.14. The confusions resulting from the auditory nerve model are qualitatively different to those arising perceptually. In the psychophysical study, for example, nearly all confusions arising from a failure to determine the presence or absence of voicing are resolved when there at least 2 spectral channels, regardless of whether or not periodicity cues

are attenuated. High rate envelope cues do help to resolve these confusions in the single channel case, however. For the model, detection of the presence or absence of voicing does not vary as the number of channels is increased, although the presence of high rate envelope cues does improve the transmission of voicing information for all channel conditions. The transmission of information regarding the manner of articulation of each consonant is similarly under-predicted by the model where the number of spectral channels is 2 or more, although the acoustic cues underlying classification of this feature are more complex and related to the other two features. The percentage of information transmitted regarding place of articulation, on the other hand, is over-predicted by the model for all conditions. This is a surprising result too as consonant place is typically regarded to be associated with spectral cues. However, this could be indicative that, while spectral shape may be represented by phase-locked responses to formant frequencies, for example, this is not accessible by more central processing mechanisms concerned with speech discrimination and an epiphenomenon emerging from the need for temporally precise responses at the periphery for sound localisation tasks.

These results show that the pattern of confusions generated by the classifier are not a good predictor of perceptual confusions, and that the relative locations of consonants in a perceptual feature space are not reflective of auditory nerve representations alone.

Consonant	Voicing	Manner	Place
b	1	0	0
d	1	0	1
f	0	2	0
g	1	0	2
k	0	0	2
l	1	3	1
m	1	1	0
n	1	1	1
p	0	0	0
s	0	2	1
ʃ	0	2	2
t	0	0	1
ð	1	2	1
v	1	2	0
j	1	3	1
z	1	2	1

Table 2.2: Consonant class groupings used in information transmission analysis for voicing, manner and place.

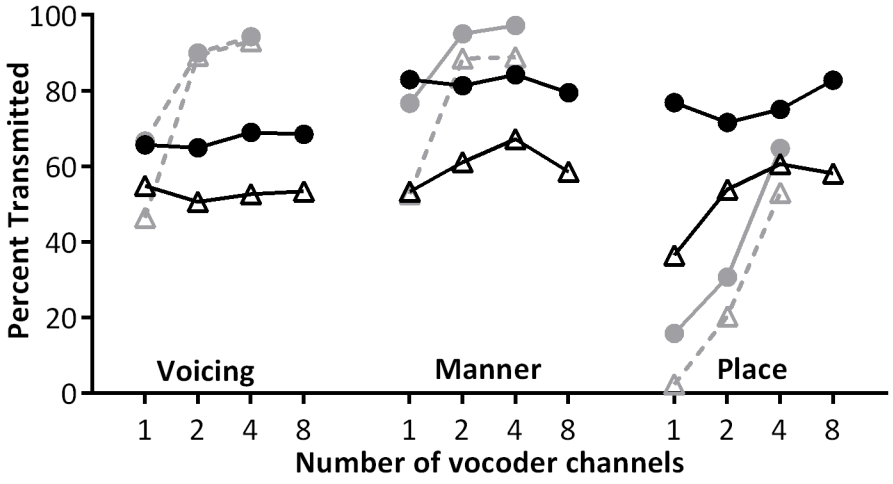

Figure 2.14: Information transmission analyses of classifier confusion matrices for consonant voicing, manner and place. Data shown for 16 (**Δ**) and 500 (●) Hz envelope conditions for confusions produced by the neural classifier (black) and reproduced from Shannon et al. (1995, grey).

2.3 Discussion

In this chapter a neural classifier was presented that was able to identify speech tokens based on single trial activity patterns of populations of simulated auditory nerve fibres of the guinea pig. The classifier was based on a nearest neighbour classifier, which measured Euclidean distances between response vectors whilst allowing response patterns to temporally shift relative to each other such that the position of representations of the medial consonant sounds did not explicitly need to be provided, which would be an extraneous external temporal reference that the brain does not have access to.

It has previously been demonstrated that the auditory nerve of the guinea pig represents acoustic features of vowel sounds in the temporal discharge patterns of individual fibres and fibre ensembles (Palmer et al., 1986; Palmer, 1990). These discharge patterns are reproduced by the same computational model presented here (Holmes et al., 2004). To the knowledge of the author, no previous studies have examined the representation of consonant sounds in this animal model. The representations of small groups of consonants in the auditory nerve has been examined in the cat (Sinex and Geisler, 1983) and the chinchilla (Loebach and Wickesberg, 2006), which found temporal representations of various cues such as voice onset time as well as formant transitions. This is the first study to use a classifier to quantify discriminability of auditory nerve representations of a large group of consonants produced by multiple talkers for various putative neural codes, in which no prior assumptions were made regarding salient acoustic cues.

The ability of the classifier to discriminate natural speech sounds was tested using both single and multiple talker paradigms. For the single talker paradigm, the discriminability of neural responses was remarkably robust to temporal smoothing. It was demonstrated that individual fibres carry enough information to discriminate the tokens provided that spike timing was preserved. Where only spike rates were made available to the classifier, responses from multiple nerve fibres were necessary to reliably identify the correct speech token. The addition of multiple talkers made discrimination performance based on rate cues alone decrease although perfect discrimination was possible provided sufficient temporal resolution was preserved in the auditory nerve responses.

Speech sounds were parametrically degraded using a noise vocoder. For the single talker paradigm, similarly to the natural speech sounds, a multiplicity of neural codes could be utilised to produce reliably discriminable responses regardless of the degree of spectral and temporal degradation. For human listeners, speech recognition tasks using similar spectral and temporal degradations are very difficult even in a single talker paradigm (Van Tasell et al., 1992), regardless of training. The data presented here supports the assertion that discriminability of these tokens is not limited by their peripheral representations, but by cognitive factors exacerbated by how the task is presented to the listener. For example, listeners are often required to map the modulated noise tokens to internal representations of natural speech, and this mapping may be the cause of this limitation as very different cues are available in the natural speech tokens and their degraded counterparts.

The classifier was also trained using representations that were combined across the three talkers and then tested using single trial response patterns. There was a clear increase in discriminability between vocoding conditions in which high rate envelope cues were attenuated and those in which they were not. The benefit of high rate envelope cues was more pronounced than that which is observed in human psychophysics, suggesting that these cues are not available to central processing systems for speech discrimination per se, but these have been shown to be useful in noisy situations or where there are competing talkers, for example (Stone et al., 2008). Human psychophysical studies also demonstrate a trade-off between spectral and temporal information; where there are limited spectral cues available, temporal cues become more salient and vice versa (Xu et al., 2005).

For the 16 Hz vocoding conditions, the classifier was not able to benefit from increasing the number of spectral channels beyond 2. Since a fixed number of fibres was used for each vocoding condition, as the number of channels increases the number of fibres per channel decreases. It was demonstrated that multiple fibres are required to produce robustly discriminable response classes in a multi-talker paradigm, and it may be that the increase in spectral resolution is offset here by undersampling within each spectral channel. This could also be a factor to consider in cochlear implant patients, in whom auditory nerve degeneration leading to stochastic undersampling of speech envelopes may be a concern depending on the etiology of deafness, a problem which would be exacerbated by increasing the number of spectral channels.

As the vocoder analysis filter bandwidth decreases, so do the intrinsic envelope fluctuations of the noise carrier leading to further degradation of the speech envelope. This should not significantly affect auditory nerve representations of vocoded speech until the bandwidths of the channels approach that of the cochlear filters, as the vocoded speech is processed through the cochlear filterbank. This could be remedied by integrating inputs across multiple fibres with different CFs that lie within a given vocoded channel, as these noise fluctuations should effectively cancel out leaving a representation of the speech envelope. If this is the case, then the benefit of increased spectral resolution should become apparent at subsequent auditory nuclei after such integration has taken place.

This model of cochlear nerve responses appears to under-predict phoneme discriminability in humans for conditions where rich temporal cues are not available and over-predict where there is minimal spectral information, particularly where high rate envelope cues are available. As such, auditory nerve representations in this model are a poor predictor for human perception. However, studies involving simpler stimuli show that representations are transformed by subsequent neural processes. Subsequent chapters use a similar paradigm to investigate how these transformations manifest in the context of the neural codes for speech discrimination in upstream auditory nuclei.

Chapter 3
Inferior colliculus

The inferior colliculus is a site of major convergence of parallel pathways of ascending and descending auditory information in which almost all ascending inputs to the primary auditory cortex synapse (see 1.1.5). It has been shown in previous studies that different speech tokens elicit distinct spatio-temporal patterns of activity in the IC (Ranasinghe et al., 2012, Perez et al., 2013), and that if a temporal code is assumed with spike-timing precise to 1-10 ms, neural discriminability is well correlated with behavioural performance in a consonant discrimination task under certain conditions. The extrapolation of these results to speech coding in general, however, is limited by the fact that the representations were compared pairwise, so as to reflect the behavioural task performed by the animal. Further only a single exemplar of each speech token was used to generate phonemic classes. In reality, these phonemic classes must be robust across a variety of different acoustic variations. Inter-talker differences is a prominent source of variability so it is important to examine whether discriminability of neural representations at these timescales is robust to these variations.

The aim of this chapter is to investigate various putative neural coding strategies for discrimination of speech sounds. It extends previous studies by investigating how these putative codes affect the formation of phonemic classes that are robust to inter-talker variations. The number of phonemes compared at any one time is also increased such that results may be compared to results from human psychophysics literature.

3.1 Methods

3.1.1 Subjects

Experiments were carried out on 18 tricolour guinea pigs (Cavia porecellus), which were bred in house (11 male, 7 female). All animals weighed between 378 and 1342g (μ=663, σ=256) and experiments were carried out in accordance with UK Home Office regulations.

3.1.2 Surgical preparation

Animals were anaesthetised with an intraperitoneal injection of urethane (0.5 g kg^{-1} in 20% solution; Sigma), and supplemental doses of Hypnorm (0.2 ml) were administered intramuscularly every 1 to 2 hours such that forepaw areflexia was maintained. Bronchial secretions were suppressed using a premedication of atropine sulphate (0.2 ml) delivered subcutaneously. Upon achieving surgical anaesthesia, as determined by the cessation of a forepaw pedal withdrawal reflex, the head, neck and ears were shaved. A small incision was made in the throat and the trachea was exposed by blunt dissection. An incision was made in it so that it could be cannulated using polythene tubing, which was secured in place using nylon thread. Parts of both tragi were

resected in order to expose the external auditory meatus and the condition of the tympanic membranes was checked for abnormalities under a microscope. Any detritus occluding the membranes was also removed.

The animal was placed in a stereotaxic frame inside a sound attenuating booth with the head secured in place by a bite bar and aural specula. A mid-sagittal incision was made in the scalp and cranial periosteum from the rostral edge of the orbit to the base of the neck on the dorsal side and the skin on the dorsal surface was resected using blunt dissection. In order to maintain pressure equalization in the inner ear, the temporalis muscles were partially resected in order to expose the surfaces of the auditory bullae in which small holes were made. Long (0.5 mm diameter) polythene tubes were then inserted into each bulla and the hole was sealed using petroleum jelly. The muscles at the back of the neck were also resected such that a small incision could be made in the tissue over the foramen magnum, which relieved cerebrospinal fluid (CSF) pressure increasing the stability of recordings.

A 5 mm by 5 mm craniotomy was made over the right inferior colliculus, which was centred 11.3 mm posterior to the intersection of the sagittal and coronal sutures (bregma) rostro-caudally and 2.5 mm lateral to the sagittal suture medio-laterally. Figure 3.1 shows the location of the craniotomy in relation to anatomical landmarks on the skull. The visible dura was resected and the exposed cortex was covered with agar (1.5% agar in 0.9% normal saline) in order to prevent desiccation.

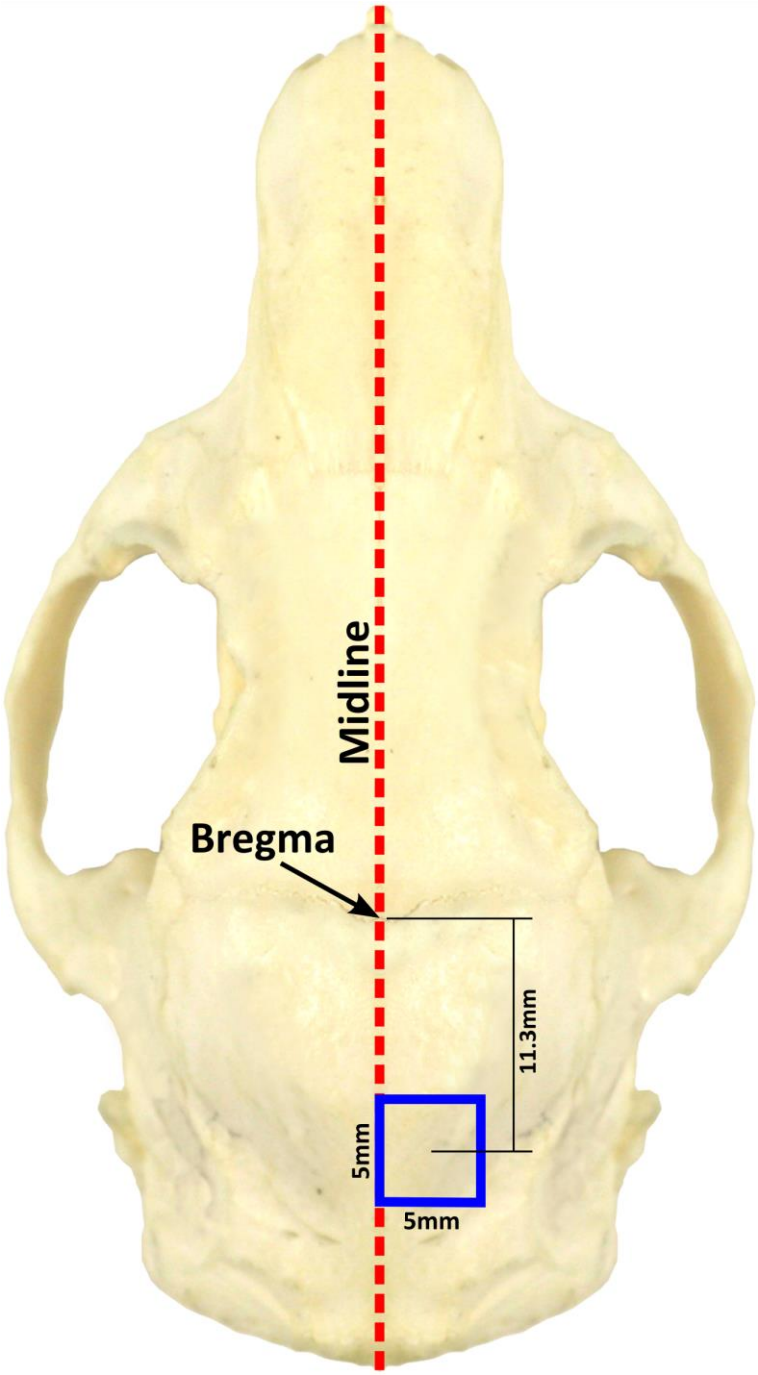

Figure 3.1: Stereotaxic positioning of a craniotomy for electrophysiological recording in the right CNIC. Image is a top view of a guinea pig skull showing the craniotomy location (in blue) relative to major anatomical landmarks.

A respiratory pump delivered 100% oxygen via the tracheal cannula. Core body temperature was monitored using a rectal probe and maintained at 38°C ± 0.5°C using a homeothermic blanket. End tidal carbon dioxide level and heart function were monitored using a vital signs monitor (VetSpecs). End tidal CO_2 was maintained within normal range (28-38 mmHg) by modulation of volume and rate of the respiratory pump. Stimuli were presented diotically via speakers (modified RadioShack 40-1377) inserted into the hollow aural specula.

3.1.3 Stimuli

3.1.3.1 Speech stimuli

The same stimulus set as that described in 2.1.2 was used. Briefly, 16 VCVs, each produced by 3 male talkers, were selected from the corpus described in Shannon et al., (1999) and parametrically degraded using a noise vocoder (Shannon et al., 1995). A custom DSP program was designed using RPvdsEx (TDT) to run on an RX8 processor (TDT, System 3). Vocoding was performed offline, and stimuli were loaded into a buffer on the hardware. The timing and order of stimulus presentation was controlled from a PC using Jan Schnupp's BrainWare. Due to buffer size limitations, the vocoding conditions were presented sequentially. For each vocoding condition, 10 repetitions of each speech token were presented diotically at a rate of 1 per second. Digital to analogue conversion was performed by the RX8, after which the signal was amplified and delivered via the speakers.

Stimulus level was again set using only the vowel portions of each stimuli, as determined by visual inspection of the spectrograms. The frequency response

of the stimulus presentation system was estimated at the start of each experiment using a condenser microphone (Brüel and Kjaer; 4134) inserted into each aural speculum, and used to set the level of each speech token. Each speech token was initially normalised, and the Fourier transform of the vowel portions was calculated. The frequency domain representation of the vowel portions was then weighted using the estimated frequency response of the system. The RMS was calculated from the weighted frequency domain representation using Parseval's theorem, which states that:

$$\sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X[k]|^2 \quad (1)$$

It follows that the RMS of the weighted vowel portions is given by the following.

$$RMS = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} |x[n]|^2} = \sqrt{\frac{1}{N^2} \sum_{k=0}^{N-1} |X[k]|^2} \quad (2)$$

A full amplitude tone corresponded to a signal level at the speaker of 120 dB SPL. The level in dB SPL of a given speech token with no attenuation is then given by:

$$level_{dB\ SPL} = 120 - 20 \log_{10} \left(\frac{RMS_{speech}}{RMS_{tone}} \right) \quad (3)$$

The required attenuation was then the difference between this value and the desired output level, which was set to 70 dB SPL. This was then converted to a scalar value, which was used to multiply the speech signal before it was presented. This process was completed separately for both the left and right

channels and all stimuli were ramped on and off using a 10ms raised cosine ramp.

3.1.3.2 Frequency response area

The frequency response areas were estimated at each site. Tones of 50ms duration were generated using the same RX8 TDT hardware and presented through the system described above. A 10 ms onset and offset ramp was applied to each tone. Tone frequencies began at 200 Hz, and were separated by quarter octave steps for 7 octaves, such that the highest frequency was 25.6 kHz. The frequency response of the system was used to set the level of each tone such that tones at each frequency were played at levels between 0 and 80 dB SPL in 5 dB steps. Tone stimuli were presented once every 200 ms in randomised order for 10 repetitions.

3.1.4 Data acquisition and signal processing

3.1.4.1 Extracellular recordings

Extracellular potentials were recorded from individual or small groups of neurons. Recordings were classified as either single- (SU) or multi-unit (MU; see 3.1.4.2). SU recordings were made using tungsten-in-glass microelectrodes, which were manufactured in house by the experimenter using the methods described by Bullock et al. (1988). The exposed conductive tips were between 2 and 12 µm in length. For a given experiment, between 1 and 8 individual electrodes were mounted onto a circuit board with a tip spacing of approximately 500 µm. MU recordings were made using 16 channel multi-electrode arrays (Neuronexus) in a 1×16 configuration with an electrode

spacing of 100 µm and conductive site areas of 177 μm^2. Example voltage traces produced using the two electrode types are shown in Figure 3.2.

Electrodes were mounted onto an inchworm motor (Burleigh), which was positioned over the centre of the craniotomy by hand using a micromanipulator arm (Kopf), such that any major blood vessels were avoided. A ground wire was placed under the skull in a small hole over the motor cortex and attached to the electrode board. Electrodes were oriented such that inchworm motor movement was along a dorsal-ventral axis. Electrodes were initially advanced to ≈7mm below the cortical surface. Signals from the electrodes were bandpass filtered between 0.3 and 6 kHz and amplified such that they were audible when played through a speaker. Speech tokens were selected from the stimulus set at random and delivered diotically and electrodes were advanced in 3 µm steps until spiking activity was audible on one or more electrode channels. If the 1×16 electrode array was being used, the array was advanced until evoked signals could be detected on the dorsal-most sites. Voltage traces of the responses to the speech sounds were also visualised. For SU recordings, electrodes were repositioned until a high signal-to-noise ratio was achieved on at least one of the electrode channels.

Once the electrodes were positioned, the frequency response area stimuli were presented followed by the speech stimuli. For recording, only a low pass filter with a cutoff of 10 kHz was used so that the local field potentials (LFPs) were retained. Signal acquisition and analogue to digital conversion were performed

by an RX6 multi-channel processor (TDT). Signals were digitised at 25 kHz, and stored in a 16 bit, non-compressed format.

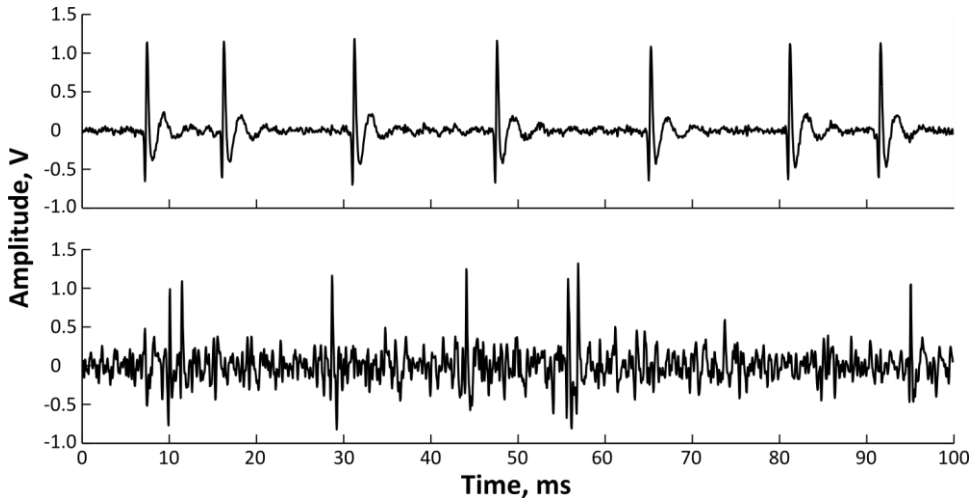

Figure 3.2: Example 100 ms voltage traces recorded using a tungsten micro-electrode (upper) and Neuronexus electrode array (lower). Both recordings were made in the right inferior colliculus, and bandpass filtered between 0.3 and 6 kHz (4th order Butterworth).

3.1.4.2 Spike detection and recording classification

A graphical user interface (GUI) was developed in Matlab for offline spike detection, and was interfaced with existing Matlab-based spike sorting software developed in-house. Recordings were initially bandpass filtered between 0.3 and 6 kHz (4th order, Butterworth). An initial spike threshold, θ_{spike}, was set using a multiple of an estimate of the standard deviation of the background noise that has been shown to be more robust to changes in neuronal firing rates than the standard deviation of the entire signal (Quiroga et al., 2004).

Initially:

$$\theta_{spike} = 4\sigma_n \tag{4}$$

Where

$$\sigma_n = median\left(\frac{|x|}{0.6745}\right) \tag{5}$$

Crossings of this threshold were detected and 1.5 ms supra-threshold "events" centred on each of them were extracted. A minimum gap of 0.75 ms between crossings was enforced. A histogram of the peak value of each supra-threshold event, or supra-threshold event peak histogram (STEPH), was computed and visualised using 50 bins equally spaced between the threshold and the maximum peak value. For recordings where spike events were well isolated from background noise, an improved threshold could be set by inspection of the filtered waveform and the STEPH. Figure 3.3 shows a screen capture of the threshold editor GUI. The upper panel shows the filtered voltage trace from a single electrode. The lower panel shows the STEPH calculated from spikes detected using the initial threshold calculated using equation 4. The resulting histogram has a characteristic shape, which is the summation of two or more distributions. The first is effected by erroneous detection of spikes in the noise, and subsequent distributions arise from peak values of genuine spike events. In Figure 3.3, the distribution corresponding to spike peaks has a mean of approximately 2.4. This shape suggests that the threshold could be increased to the location of the red dashed line to decrease the number of false detections, while still detecting the majority of genuine spike events. In order

to quantify how effective the detection threshold was, the probabilities of false positive and false negative spike detection were also estimated by fitting a Gaussian function to both the distribution of spike peaks and the distribution of peaks of 1.5ms epochs randomly selected from the residual signal once spike events were removed. Figure 3.4 shows a diagram of this fitting process. Bins to the left of the threshold represent the histogram of peak values of randomly selected epochs of the residual signal and to the right is the STEPH.

The Gaussian function took the form:

$$y = ae^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

The Matlab function `fminsearch` was used to find the values of a, μ and σ for which the squared difference between y and the histogram was minimised. Initial parameters were set to the maximum value of the STEPH, the mean peak value and the standard deviation of all peak values respectively. Differences were calculated only for the values corresponding to $x > \mu$ for the STEPH fit, and $x < \mu$ for the residual fit such that fits were not skewed by the truncation of the histogram at θ_{spike}. The probability of false negative detection, $P(FN)$, is then given by:

$$P(FN) = \Phi\left(\frac{\theta_{spike} - \mu_{STEPH}}{\sigma_{STEPH}}\right) \tag{7}$$

Where Φ is the cumulative density function of the standard normal distribution. The probability of false positive detection, $P(FP)$ is given by:

$$P(FP) = 1 - \Phi\left(\frac{\theta_{spike} - \mu_{residual}}{\sigma_{residual}}\right) \tag{8}$$

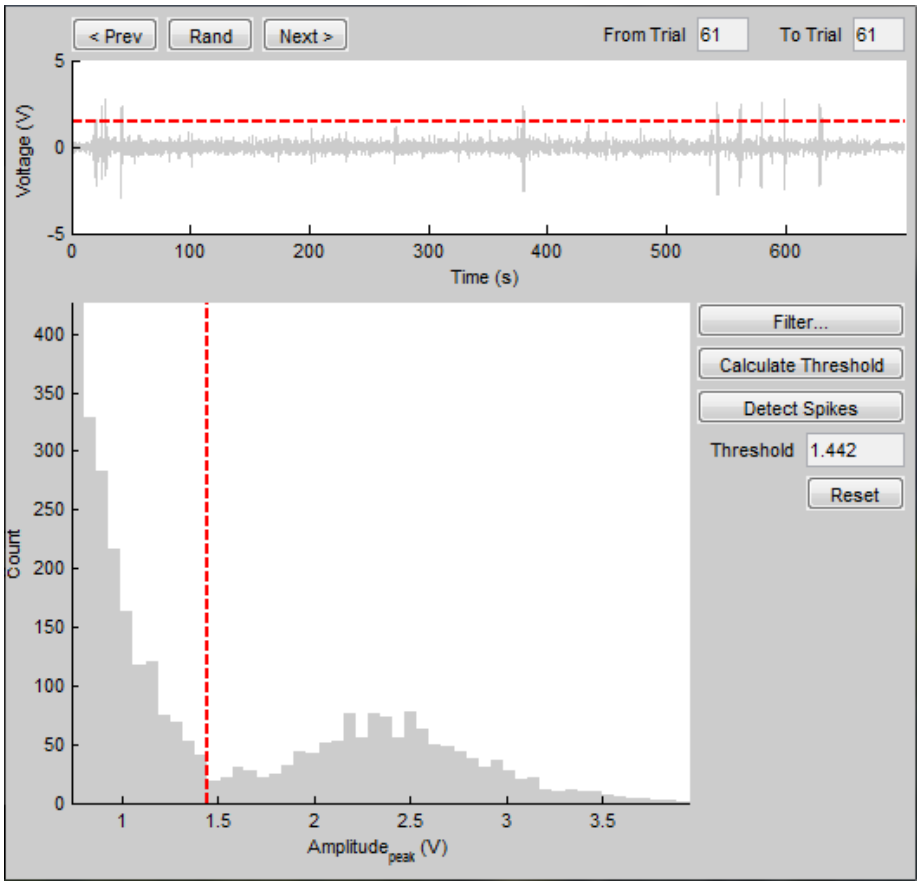

Figure 3.3: Screenshot of the Matlab GUI for offline filtering and automatic calculation and manual adjustment of spike detection thresholds. The dotted red lines are interactive and can be dragged either vertically on the voltage trace axis or horizontally on the histogram.

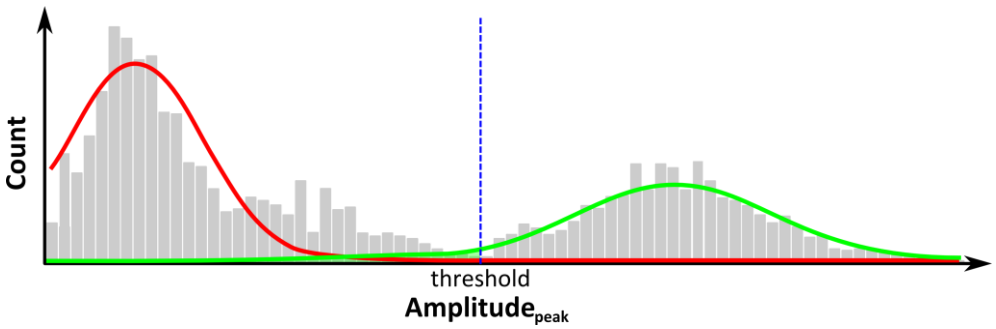

Figure 3.4: Histogram of peak values of randomly sampled epochs of the residual signal and a Gaussian fit (red) and the STEPH with its Gaussian fit (green). The spike detection threshold is also shown.

Principle components analysis (PCA) was used to identify artefacts and anomalous spikes. The extracted spikes were initially aligned by their maxima. Each spike waveform was allowed to shift by a maximum of 5 samples, corresponding to approximately 0.2ms, and the edges were truncated. PCA is a commonly used method of dimensionality reduction in multi-dimensional data sets, which is a common pre-processing step in clustering algorithms. Essentially, PCA transforms a data set comprising N dimensions, in this case corresponding to each sample of a spike waveform, to a new orthogonal N-dimensional coordinate system in which the first axis (PC1) is such that the data is maximally variable when projected onto it. Subsequent principle components are defined in the same way with the constraint that they are orthogonal. In this way, spike shapes may be visualised in 2-dimensional space by projection onto the first 2 principle components, thereby representing the largest degree of variability possible.

A GUI was used to manually identify clusters in projections onto principle component space. Electrode drift could cause variations in spike shape over time. This could, for example, cause projections onto principle components that are separate over any given comparatively short epoch to overlap when collapsed across time. For this reason, projections were also visualised using a temporal axis. This had the advantage of enabling the distinction between two distinct and simultaneous spike shapes, most likely arising from separate neural sources and a single source that changed rapidly at some stage during recording due to electrode movement. If a cluster could be defined that corresponded to a stereotypical spike shape, then the time stamps of each of the spikes relative

to the onset of each sweep were extracted and assigned to a single unit (SU). If, on the other hand, there were contemporary overlapping clusters or the spread was such that a stereotypical shape was not discernible then the recording was designated multi-unit (MU).

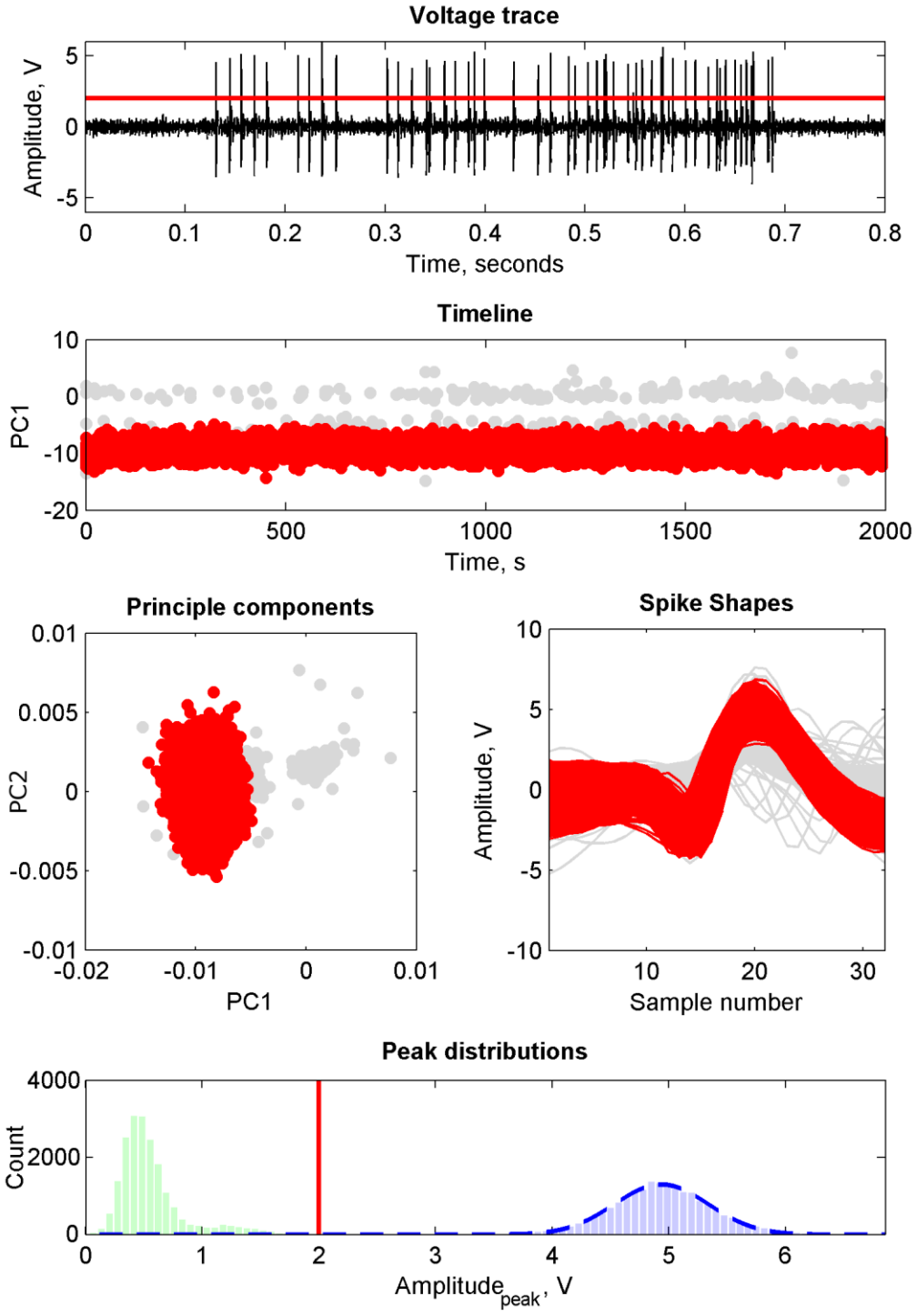

Figure 3.5: Visualisations of extracellular activity recorded in the inferior colliculus. The upper panel shows the voltage trace of 800 ms of the recording, with the threshold marked in red. The axes below show the variation of each spikes projection onto PC1 over time, with those attributed to a single source shown in red. Beneath is the projection of all spikes onto PC1-PC2 space (left) and a plot of all detected supra-threshold events (right). The lower axes shows the histogram and Gaussian fit of the STEPH (blue) and a surrogate spike set from the residual (green).

Due to the larger conductive surface in the Neuronexus probes, many of these recordings comprised signals from a high number of sources. In some of these recordings, spiking activity often overlapped and it wasn't meaningful to extract individual spike events in the way described above. For this reason, these recordings were processed separately using neurograms produced from instantaneous RMS values evaluated over sliding windows of varying lengths. As many recordings can be produced easily using multi-electrode arrays, it was not feasible to examine all recordings manually. Instead, an automated method was devised to determine which recording sites were responsive to auditory stimuli. Recordings from the row of the FRA corresponding to 0 dB SPL were initially bandpass filtered between 0.3 and 3 kHz (6^{th} order Butterworth). The RMS of the response to each tone was then calculated and averaged across presentations. The same procedure was then carried out for the row of the FRA corresponding to the maximum signal level presented.

A similar method was used to determine which of the recording sites demonstrated auditory-evoked local field potentials (LFPs). In this case, recordings were initially lowpass filtered below 0.3 kHz (6^{th} order Butterworth).

3.1.4.3 Frequency response area analysis

The process used to automatically extract measures of tuning from the frequency responses areas was the same as that described in Palmer et al., (2013). Frequency response areas were generated in response to 50 ms pure tones at various levels (see 3.1.3.2). The average spike count in response to each stimulus was used to produce an $N \times M$ matrix, where N is the number of stimulus levels and M is the number of stimulus frequencies. The first row of this matrix corresponded to average spike counts in response to the lowest stimulus level presented and was used as a baseline measurement from which a threshold was determined. A threshold was defined for each frequency step as the sound level at which spike rate exceeded the mean plus 4 standard deviations of the baseline rate or the baseline rate plus 0.15 times the range from minimum to maximum rates, whichever was greater. In order to minimise the effect of artefacts and to account for the small sample of frequencies and levels that were presented, the matrix was initially upsampled by a factor of 2 using the Matlab built-in function `interp2`, which applies a low-pass filter that preserves that values of original data and interpolates between them, assuming that signal is band-limited to half of the original sampling rate.

Linear interpolation was used to calculate thresholds, and it was required that the threshold criterion was also exceeded 10 dB above threshold at any given frequency. A similar process was carried out on the inverted FRA matrices to trace an upper edge of the response area, if one was present. The CF was then defined as the frequency at which the threshold corresponded to the lowest

sound level. Q-values were also extracted at 10 and 40 dB above threshold. These values, denoted Q_{10} and Q_{40} were defined as the CF divided by the width of the threshold contour at these two signal levels.

3.2 Results

Reponses were recorded at 302 sites in the inferior colliculus, of which 36 were classified as single unit (SU). Of the remaining recording sites, 106 demonstrated responses in the 0.3 to 6 kHz bandwidth that were modulated by auditory stimuli and these sites were used in MU analyses.

3.2.1 Pure tone responses

Pure-tone response characteristics of neurons in the inferior colliculus have been examined in several mammalian species including the cat (Ramachandran et al., 1999), the mouse (Egorova et al., 2001) and the guinea pig (Syka et al., 2000; Palmer et al., 2013). Pure-tone frequency response areas (FRAs) were also produced in the present study with several motives. Firstly, the analysis of pure-tone tuning may be used in conjunction with published studies to complement histological methods for electrode placement validation. Secondly, inferences may be made regarding the efficacy of central processing mechanisms for producing responses conducive to complex signal discrimination by correlating parameters such as sharpness of tuning with the neural discriminability of responses to speech sounds. Finally, the specification of characteristic frequency (CF) could enable identification of frequency regions in the neural representations, if any, that are particularly salient for speech discrimination. This has implications for placement of auditory prostheses.

Of the 36 SU recording sites, FRAs were recorded from 25 of which 22 demonstrated firing rates that were modulated by pure-tone stimuli. These FRAs are all shown in Figure 3.7. Of the 106 MU recording sites, 80 showed pure-tone modulated responses. Each of these 102 FRAs were fitted with a single iso-rate contour by the algorithm described in 3.1.4.3. An influential publication on response area shapes suggested that they could be broadly classified broadly as type-I, type-V of type-O (Ramachandran et al., 1999), which each correspond to one of three afferent pathways. Type-I response areas are characterised by an excitatory response to only a narrow range of frequencies, the width of which is invariant to sound level, with inhibitory sidebands. Type-V excitatory regions become broader as sound level increases in a similar manner to those found in the auditory nerve and show no inhibitory sidebands. Type-O response areas are characterised by showing selectivity to both frequency and level, resulting in an enclosed area of excitation.

Across each of the 102 FRAs recorded in the IC, only 2 showed enclosed, type-O responses. By inspection, it was not clear that remaining FRAs could be classified as type-V vs type-I. Figure 3.6 shows a summary of pure-tone response statistics, including the distribution of Q_{10} and Q_{40} values. In this parameter space, type-V and non-type-V FRAs would be discriminable as separate clusters with type-V below unity, corresponding to decreased Q for higher stimulus levels relative to threshold, and type-I or type-O at unity or above. This clustering does not appear to be apparent in the present dataset. A recent study used objective classification methods to examine the existence of inherent classes for pure-tone FRAs (Palmer et al., 2013). In this large scale

study several more responses classes were subjectively identified. They found that, whilst stereotypical, extreme FRA shapes were apparent, many intermediate FRA shapes existed, which could be identified using objective methods. FRA shapes therefore formed continua and not discrete classes. These data support this finding, and for this reason FRAs were not classified into discrete groups in subsequent analyses.

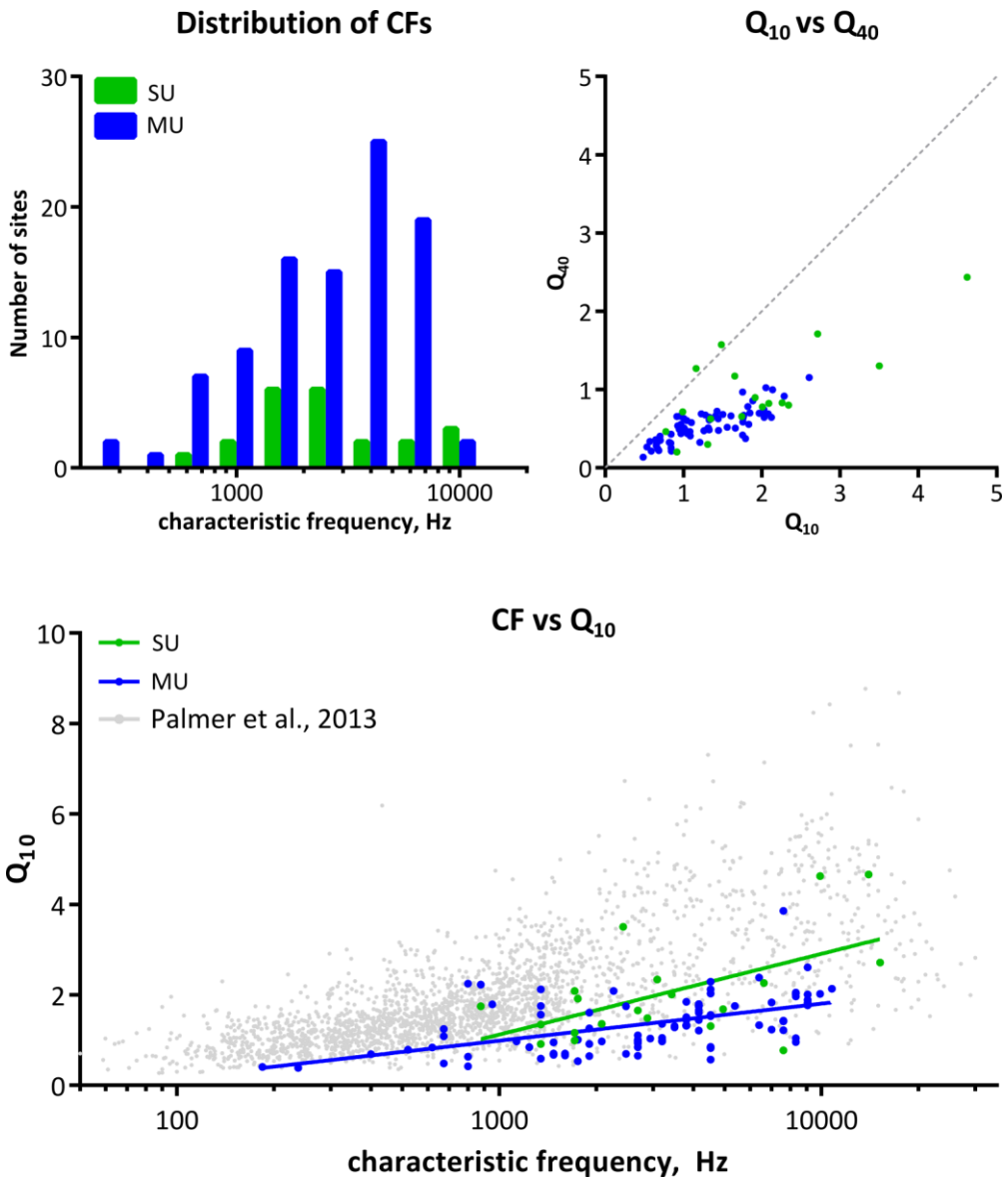

Figure 3.6: Summary of pure-tone response characteristics for SU and MU recording sites in the inferior colliculus of anaesthetised guinea pigs. Data from (Palmer et al., 2013) included for comparison.

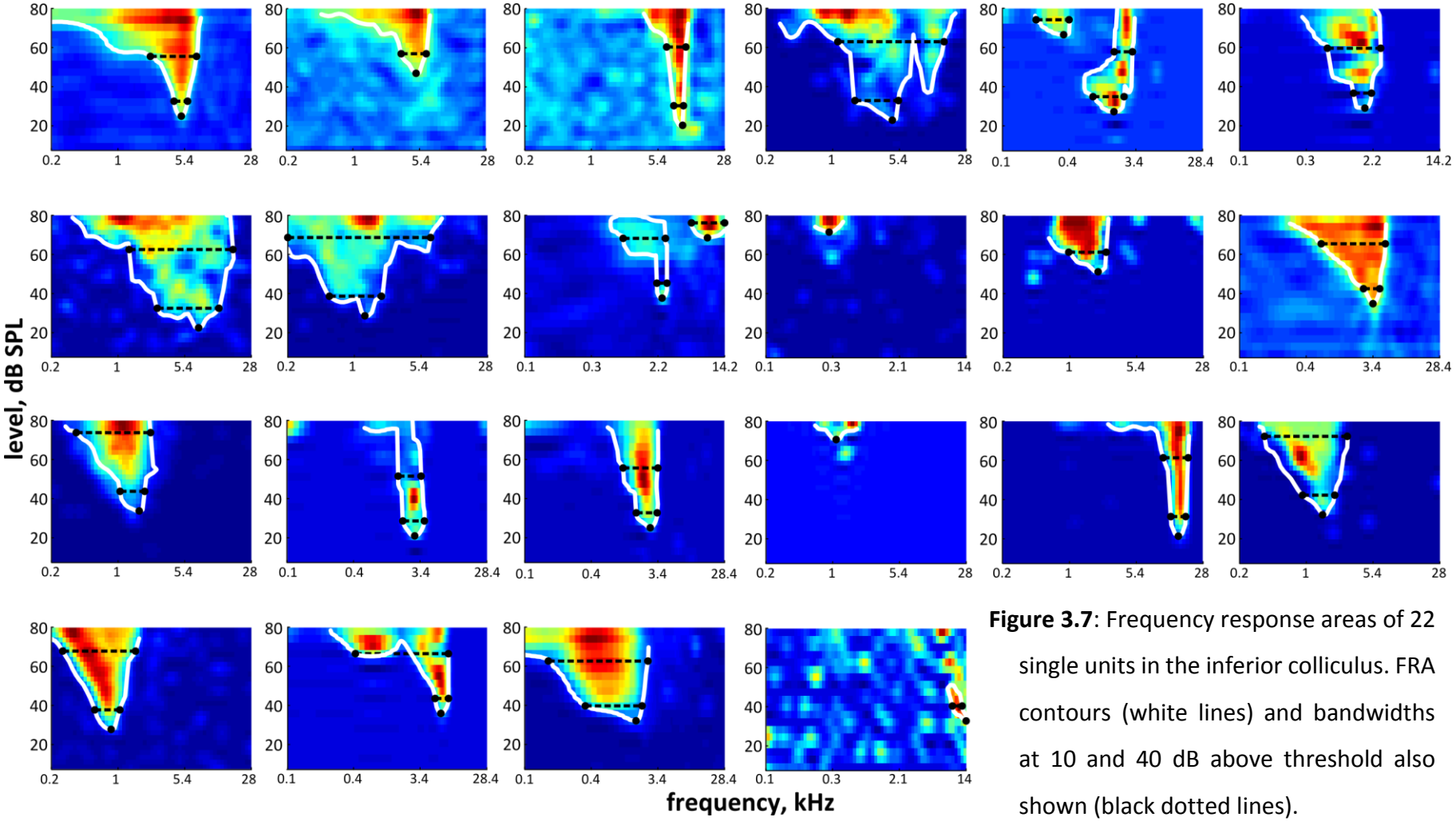

Figure 3.7: Frequency response areas of 22 single units in the inferior colliculus. FRA contours (white lines) and bandwidths at 10 and 40 dB above threshold also shown (black dotted lines).

The distribution of CFs of each of the SU and MU recording sites is shown in the upper left panel of Figure 3.6. These cover the most of audible range of pure-tone stimuli in the guinea pig (approximately .086 to 46.5 kHz at 50 dB SPL; Heffner et al., 1971), excluding one octave at both the upper and lower ends although the majority of recording sites have CFs between 1 and 10 kHz (81.8% SU and 83.3% MU respectively). This means that low frequency recording sites are underrepresented in the neural population and in particular in the SU set, the lowest CF of which was 775 Hz. The implications of this for the representation of vocoded speech are discussed in 5.2.

The sharpness of the FRAs as measured by the Q_{10} increases linearly with characteristic frequency (on a logarithmic scale), commensurate with previous studies on frequency selectivity for both SU and MU ($R^2 = 0.32$ and $R^2 = 0.28$ respectively). The Q_{10} values of the SU recordings were also significantly higher than those of the MU recordings when the effect of CF on Q_{10} is taken into account. This is unsurprising, as the MU FRAs are presumably a combination of FRAs from several SUs and so would be unlikely to become narrower. Neither the MU nor the SU data differed significantly from the more comprehensive set of SU recordings reported by Palmer et al. (2013) showing that the pure-tone response characteristics of these recording sites are representative of those in the CNIC.

3.2.2 Natural speech

The neural classifier was initially trained separately for each of the three talkers. Responses to at least six presentations of each stimulus were recorded from 34 of the SU sites. For the remaining SU sites, the unit was lost before data acquisition was completed so these were not included in these analyses. Figure 3.8 shows the SU average population responses of the population to a single exemplar of each VCV produced by one of the talkers. When the classifier was trained and tested using the spike timing representations, it was able to correctly identify the medial consonant with a success rate of 69.8 % (σ=13.1). For spike count representations, classifier performance significantly increased (μ=85.1, σ=17.4). These values are significantly above chance (6.25 %), so these results show that representations comprising both the precise spike timing and only the spike count over longer epochs carry useful information for a speech token discrimination task, but temporal integration increases discriminability. The implication is that overall spike count in SU responses is more robust across numerous presentations of a given speech stimulus than the precise time of each spike.

Responses to 10 presentations of each stimulus were recorded at each of the 106 MU recording sites that demonstrated responses to auditory stimuli. Figure 3.9 shows the MU representations of each of the 16 VCVs produced by a single talker. In this case, each row represents instantaneous signal power in the bandpass filtered signal (between 0.3 and 3 kHz) at a given recording site sampled at 8 kHz and arranged by CF, where it was available from the FRA fit.

Unlike the SU responses, the spectro-temporal structure of the stimuli is clearly represented by these neurograms, which resemble spectrographic representations. For example, */ada/* contrasts with */aba/* by the presence of increased activity in the high frequency population. Similarly, the marked difference in activity between the low and high frequency populations during the fricative, unvoiced, medial portion of VCV */aʃa/* is contrasted by the much more uniform activation during the same portion of */ana/*. Notably, elevated responses to formant frequencies are not clearly apparent as they are in the auditory nerve neurograms in Figure 2.4. This may, however, be predominantly due to the fact that recording site BFs are not uniformly distributed across frequency (see Figure 3.6).

These population responses were temporally degraded by convolution with a rectangular window of varying length. Figure 3.10 shows the effect that parametric temporal degradation had on neural discriminability. For both the SU and MU representations, neurograms are perfectly discriminable for a broad range of smoothing window durations from a few milliseconds up to hundreds of milliseconds. This is, perhaps, not too surprising, as the speech tokens differ acoustically in many respects on several timescales and the task is clearly a trivial one for human listeners to perform.

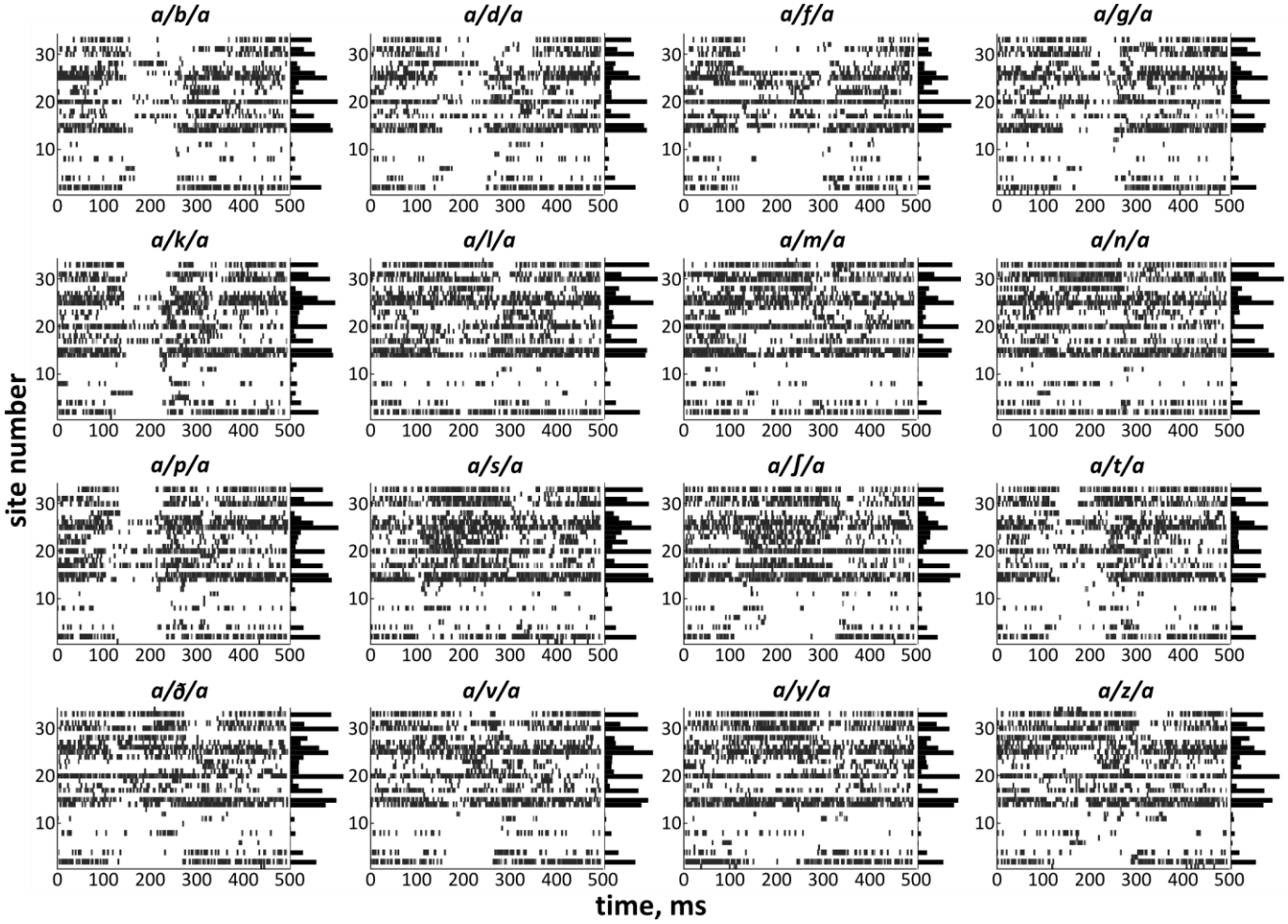

Figure 3.8: Neurograms depicting SU responses to a single exemplar of each of the 16 VCVs produced by a single talker. PSTHs were generated using 1ms bins and averaged across 6 presentations of each stimulus. Horizontal bars represent the total spike count in each row.

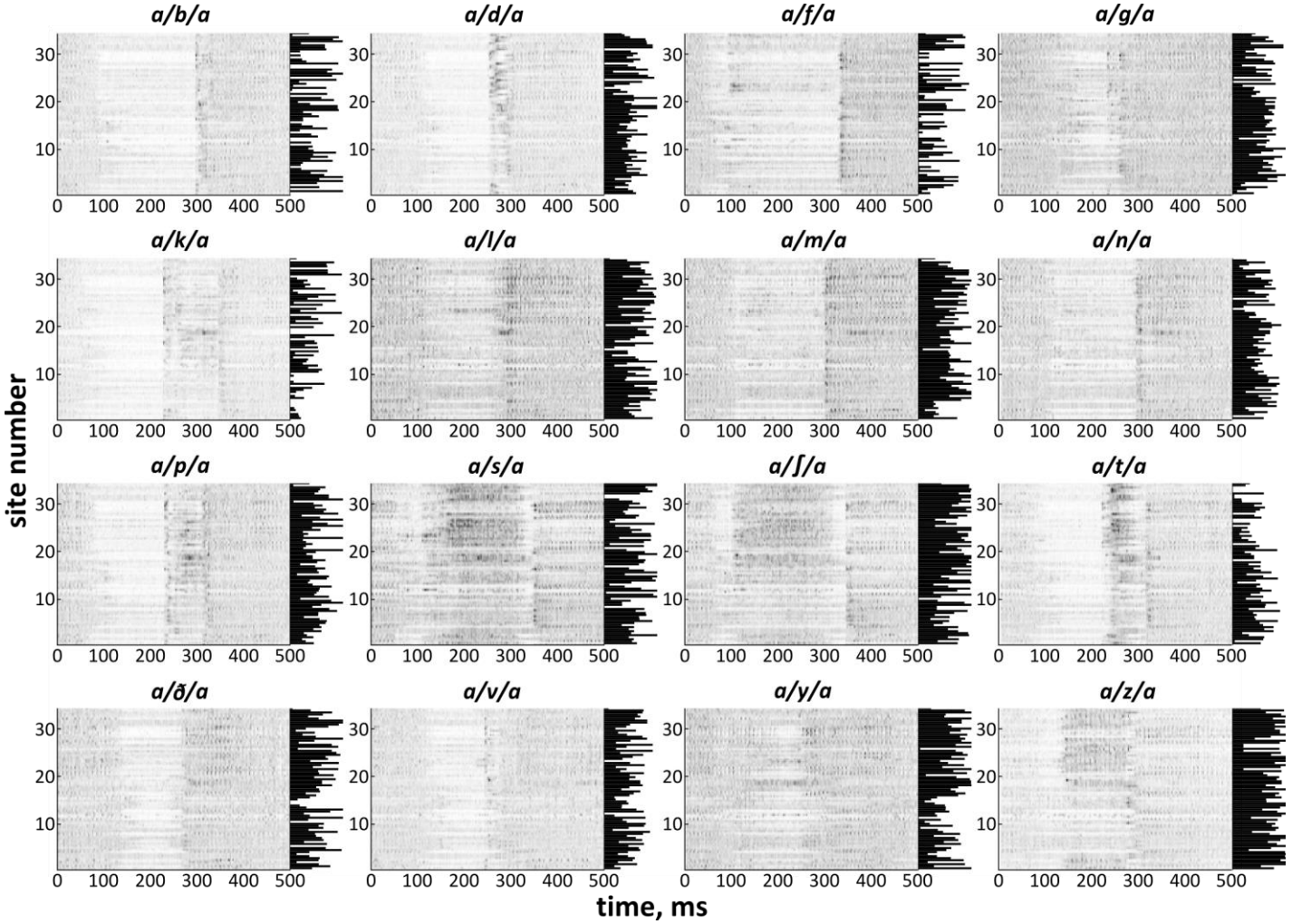

Figure 3.9: Neurograms depicting MU responses to a single exemplar of each of the 16 VCVs produced by a single talker. Each row represents signal power in the 0.3 – 3 kHz bandwidth averaged across 10 presentations of each stimulus. Horizontal bars represent the normalised RMS signal power at each recording site.

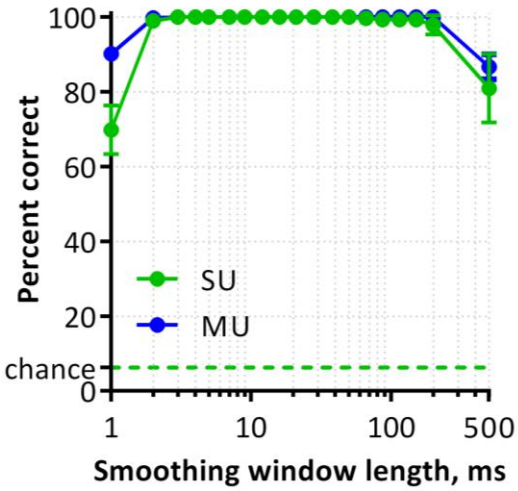

Figure 3.10: The effect of smoothing on discriminability of neural representations of speech sounds produced by a single talker. Representations comprise 34 SU and 106 MU recording sites.

The results presented so far pertain to a speech token discrimination task, as is congruent with existing studies on the neural discriminability of speech sounds (Engineer et al., 2008; Ranasinghe et al., 2012; Perez et al., 2013). Of greater interest is the aspects of these neural representations that are useful for the formation of phonemic classes that are robust to non-meaningful inter-exemplar differences, such as those between exemplars of the same phoneme produced by multiple talkers. In order to examine this, the classifier was trained on average response patterns that were combined across either two or all three of the talkers, using the relative temporal shifts that maximised similarity between them (see 2.1.3). The test set similarly comprised single trial response patterns, which were removed from the data set prior to classifier training. Figure 3.11 shows classifier performance for both SU and MU data as a function of smoothing window length for 1, 2 and 3 talkers. Both data sets show remarkable robustness to temporal degradation of the neural response, with discriminability only beginning to decrease when smoothing of the order of 100 ms is used. For both SU and MU representations, the classifier is able to perform perfect discrimination across multiple talkers when appropriate

temporal smoothing is used. For SU representations, smoothing windows of a few milliseconds duration are sufficient for perfect classification, whereas for MU representations, a window of approximately 10 ms is required.

From this analysis, it is not clear whether this is representative of a homogeneous coding strategy across the population, or representative of some mean value. It was shown previously that for simulated auditory nerve representations, different temporal resolutions are optimal for discriminability for fibres with different CFs despite having similar pure-tone response characteristics. A similar analysis was also carried out on the IC data; the classifier was trained and tested on responses from a single recording site at any one time. The results of this analysis is shown in Figure 3.12, which shows classifier performance for individual SU and MU recording sites for both a single talker and multiple talker task as a function of both smoothing window and site CF.

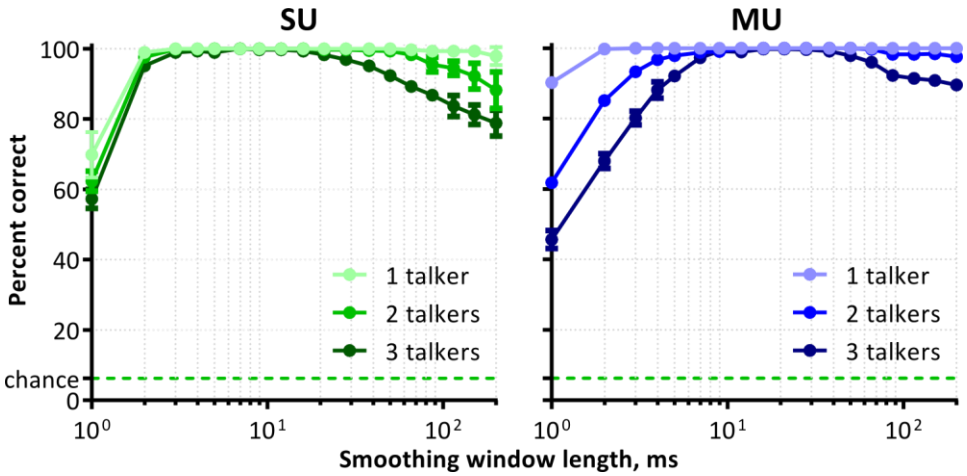

Figure 3.11: The effect of number of talkers on optimal smoothing for neural discriminability of natural speech. Data shown is for representations comprising responses from 34 SU sites (left) and 106 MU sites (right).

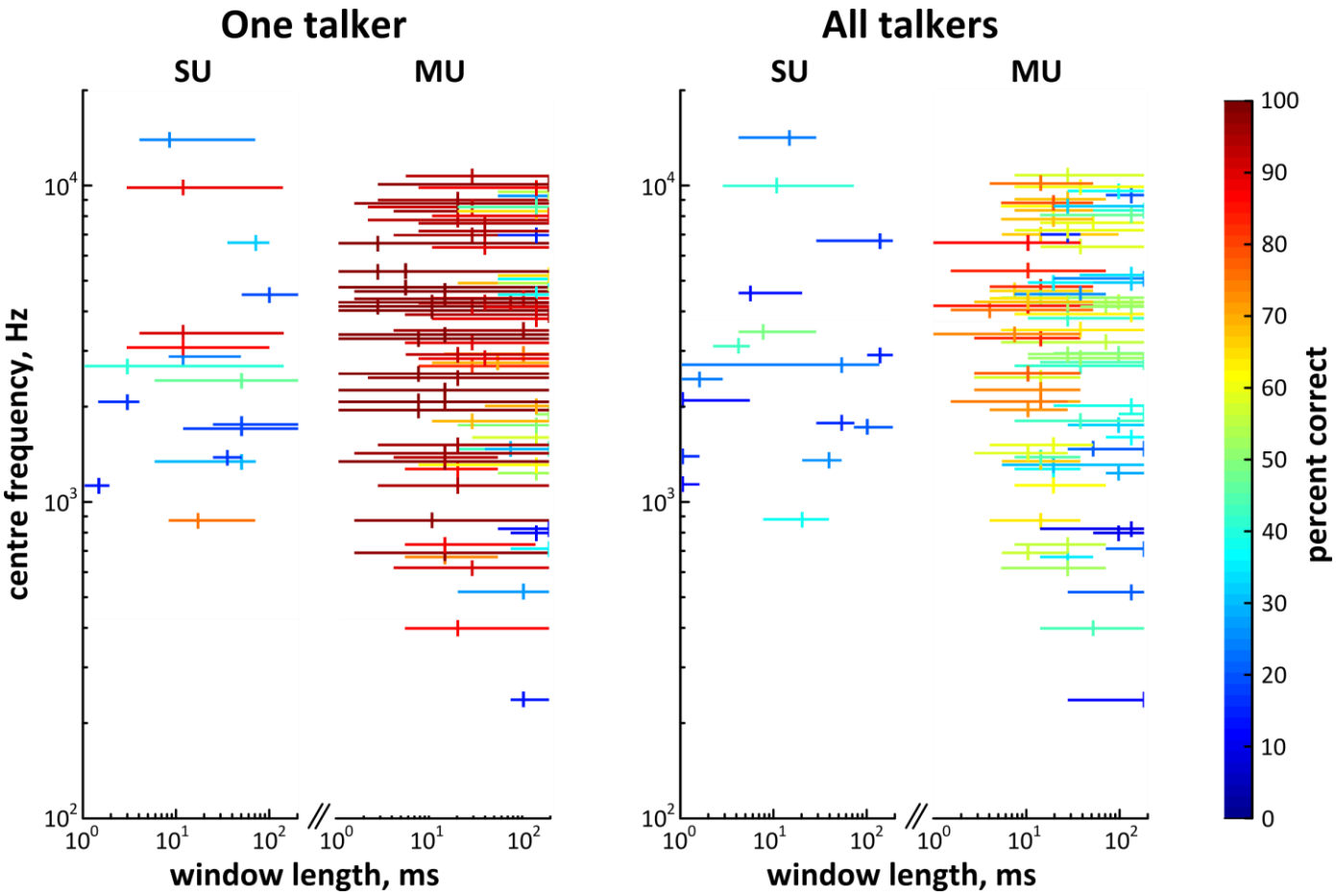

Figure 3.12: Neural discriminability of single site responses to natural speech for a single talker (left) and multiple talker (right) task. Optimum smoothing window is shown by vertical bars and values over which discriminability is over 90% of maximum is shown by horizontal bars. Maximum discriminability is shown by colour.

These plots show that for a single talker task several SU and many MU individual recording sites are able to discriminate the speech tokens with near perfect performance. This performance is also robust to severe temporal degradation of the spike trains or signal power waveform. The optimum smoothing window length for each site is highly variable. However, those with long optimal smoothing windows tend to perform poorly. A histogram of optimal smoothing windows for those sites with discrimination performance of over 80% in the single talker task is shown in Figure 3.13. The majority (91%) of high performing sites have optimal smoothing windows of between 10 and 50 ms. For the multiple talker task, individual site performance is significantly reduced and no individual recording site is able to perform perfect discrimination. However, discriminability of the population responses still reaches 100% given an appropriate choice of temporal smoothing (see Figure 3.11).

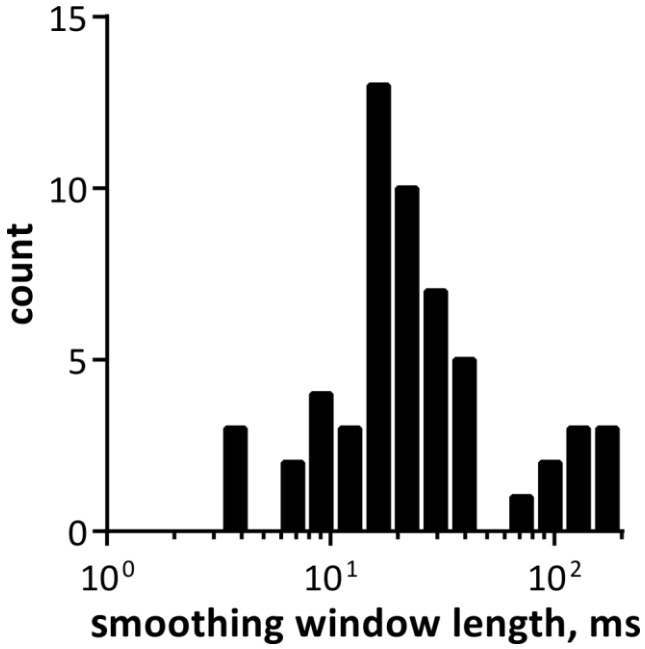

Figure 3.13: Optimal smoothing windows combined across all SU and MU recording sites with >80% discriminability for the single talker task.

These data suggest that acoustic features of these complex sounds are represented with a high enough fidelity in a subpopulation of recordings to identify the unique tokens and that these are no longer useful when phonemic classes that are robust across multiple talkers must be formed. In this case, populations of neurons must be used simultaneously, increasing the dimensionality of the feature space in which clusters may be formed. The number of recording sites required to reach maximum performance was examined by selecting random subsamples of both the SU and MU populations and performing the classification based on the corresponding neurograms. This process was repeated 100 times for each population size. The results of this analysis are shown in Figure 3.14. For low numbers of sites, the mean value of MU representation discriminability is greater than that of the SU population, although it is not significantly different due to the large degree of variability. The error bars shown represent one standard deviation, however the maxima show that high performance may be achieved with a very small number of sites, the combinations of which correspond to those sites that perform best individually.

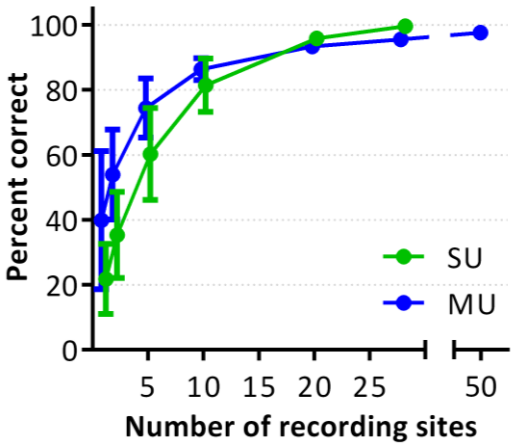

Figure 3.14: The effect of number of recording sites on the discriminability of neural representations of 16 VCVs produced by 3 male talkers.

3.2.3 Vocoded speech

The same speech sounds used in the previous section were parametrically degraded using a noise vocoder. Hardware constraints made it necessary to present each vocoding condition sequentially. Responses to at least 6 presentations of each stimulus in every vocoding condition were recorded at 28 SU sites. All 10 presentations of each stimulus were recorded at all 106 MU recording sites used in the previous analysis. The neurograms were similarly temporally degraded prior to classifier training and testing using smoothing windows of varying lengths. Classification was performed using both single talker and multi-talker paradigms and discriminability values as determined by the classifier were averaged across each of the three talkers for the single talker task.

The results for each of the vocoding conditions using optimal smoothing windows are summarised in Figure 3.15. For responses to VCVs produced by a single talker, neurograms were almost perfectly discriminable by the classifier. Only in the 1 channel 16 Hz condition was discriminability of the SU representations significantly below that for natural speech. Even when only a single talker is used and the task is simply to correctly identify previously heard tokens, human listeners perform very poorly when spectral modulations are removed (Van Tasell et al., 1987). These data suggest that the limiting factor for this type of task is not the fidelity of the sensory representation in the IC but rather the ability of the CNS to make use of it for a discrimination task.

The right-hand axes of Figure 3.15 show the optimal discriminability values of neurograms where the training set comprised average representations combined across the talkers. Also shown is the human behavioural data reproduced from Shannon et al. (1995), which investigated the behavioural discriminability of degraded speech sounds by human listeners using an analogous stimulus set. For the vocoding conditions where the envelope extraction filter cutoff was set to 16 Hz, there were no significant differences between the SU and MU datasets for 1, 2 or 4 8 channel vocoders ($t(14)$, $p = 0.011$; $p=0.74$; $p=0.92$; $p=0.21$). The mean discriminability of the representations of these sounds also increased significantly as a function of the number of channels from 1 to 2 and 2 to 4 ($p = 0.01$). The increase was not significant between 4 and 8 vocoder channels for either SU or MU representations ($t(10)$, $p = 0.50$; $t(18)$, $p=1.0$). Discriminability values are nonetheless well correlated with the behavioural measures for the 16 Hz conditions, where high rate envelope cues associated with voicing or burst envelopes, for example, are heavily attenuated. The same cannot be said for representations where these cues are not attenuated, as these are almost perfectly discriminable by the classifier regardless of the degree of spectral degradation. This is in contrast to the effect of preserving high rate envelope cues on behavioural discriminability, which provides only a comparatively modest increase in performance.

The upper panel of Figure 3.16 shows how smoothing window length affects discriminability for the single talker task. In comparison to the data for natural speech (see Figure 3.10), discriminability is much less tolerant to temporal

smoothing with very long windows, particularly where fewer spectral channels are used in the vocoder. SU discriminability is also poorer than MU discriminability for the 8 channel, 16 Hz condition. This shows that long integration windows, or a rate-based code, may be utilised for a consonant discrimination task provided that firstly there are sufficient spectral cues available and secondly that the neural population is sufficiently large to represent them. Across all conditions, performance decreased when responses precise to 1 ms are used, and optimal discriminability values were reached using a broad range of smoothing window durations. For 16 Hz envelope speech, discriminability continues to increase as smoothing windows are increased up to 10 ms except where ceiling has already been reached with shorter windows and decreases for windows greater than approximately 50ms for SU representations, however the discriminability of MU representations is robust to smoothing with windows of the order of 100 ms. For 500 Hz speech, the classifier were able to perform perfect discrimination for all conditions tested for smoothing windows between 2 and 20 ms for SU representations, and anywhere between 3 and 200 ms for MU representations. For rate-based representations, corresponding to 500 ms smoothing, representations are not more discriminable for 500 Hz envelopes than for 16 Hz ones. A multiplicity of potential decoding strategies may be utilised for discrimination of a learned set of speech tokens, none of which appear to be utilised by the CNS for a single talker speech discrimination task.

.

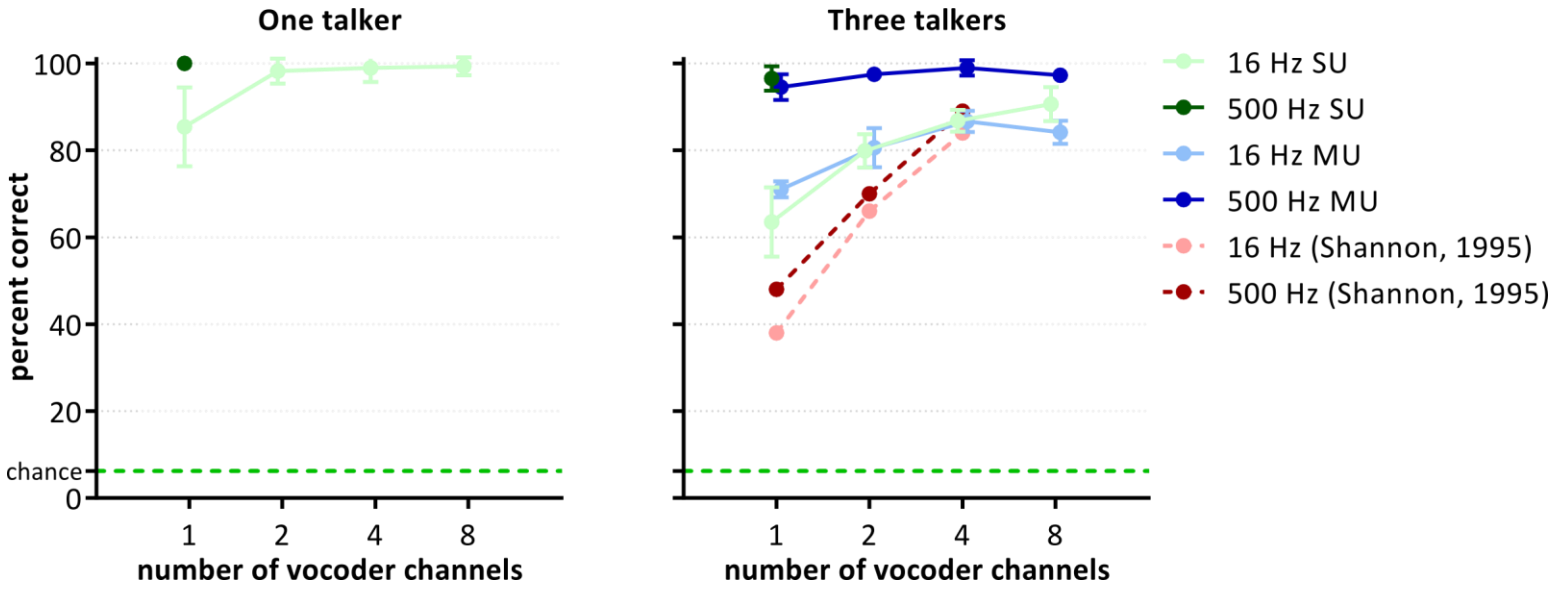

Figure 3.15: Discriminability of neural representations of vocoded speech. Classifier was trained and tested using response to 16 VCVs produced by a single talker (left) and combined across 3 talkers (right). Responses comprised 28 SU sites and 106 MU.

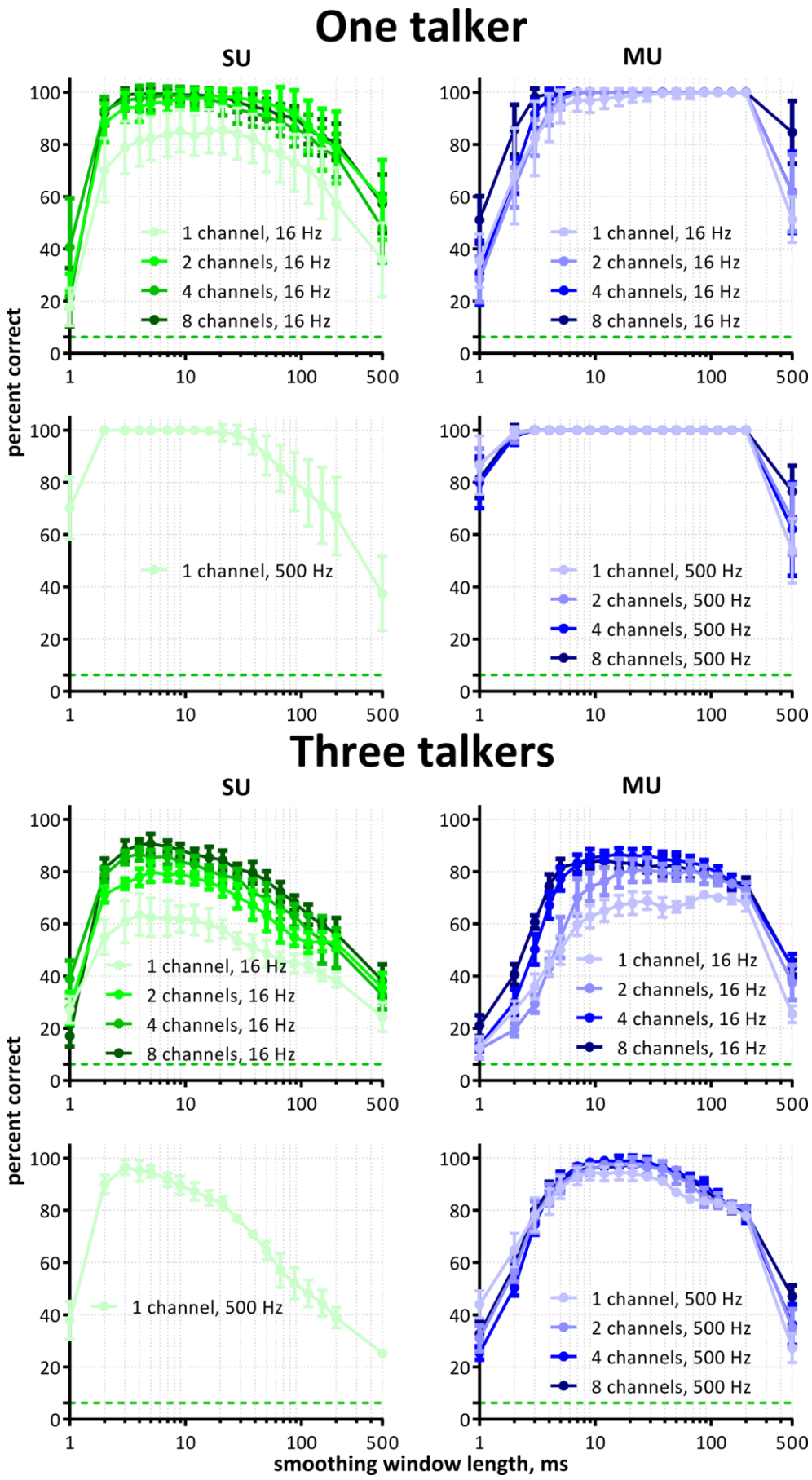
One talker
SU
MU
percent correct
1 channel, 16 Hz
2 channels, 16 Hz
4 channels, 16 Hz
8 channels, 16 Hz
1 channel, 500 Hz
2 channels, 500 Hz
4 channels, 500 Hz
8 channels, 500 Hz
Three talkers
smoothing window length, ms

Figure 3.16: The effect of temporal smoothing of the neural discriminability of vocoded VCVs. Neural responses comprised 28 SU (green) and 106 MU (blue) recording sites. Classifier was tested and trained on responses to all 16 VCVs produced by a single talker (upper) and three talkers (lower).

The lower panel of Figure 3.16 shows the effect of smoothing on discriminability of consonants combined across 3 talkers. These show a general reduction in optimal discriminability across all 16 Hz conditions, however as shown in Figure 3.15, classifier performance reaches ceiling for all 500 Hz envelope conditions. Figure 3.15 also demonstrates a trade-off between spectral and temporal fidelity in the input signal, whereby perfect discriminability may be achieved provided that either high rate envelope cues are conserved *or* there are sufficient spectral channels. Since these cues exist on different timescales, it might be assumed that they are represented using encoding windows of different durations. However, the lower panel of Figure 3.16 shows a pronounced peak in discriminability for the single channel, 500 Hz conditions corresponding to 3 ms of smoothing and 5 ms for the 8 channel, 16 Hz condition.

3.3 Discussion

This chapter demonstrated that a set of 16 VCVs produce unique responses in the inferior colliculus, which may be utilised by a nearest neighbour classifier to perfectly discriminate between them. The spatio-temporal response patterns are optimally discriminable when they are temporally smoothed. The relationship between smoothing window length and discriminability of the

representations is quite different from that in the simulated auditory nerve response patterns presented in the previous chapter. Auditory nerve representations were optimally discriminable where millisecond-precise spike timing was preserved, whereas in the inferior colliculus, optimal smoothing windows are between 10 and 100 milliseconds. This distinction, however, only becomes apparent in multi-talker paradigms, since both auditory nerve and inferior colliculus representations are perfectly discriminable when a very broad range of encoding windows are assumed for a single-talker paradigm. The main difference in this instance occurs where millisecond precise responses are preserved and there is no temporal integration. This causes the discriminability of the inferior colliculus responses to decrease, which is not the case for auditory nerve representations.

The discriminability of IC representations is also generally higher than that in the auditory nerve, and this is particularly evident in the 16 Hz envelope conditions; the discriminability of representations in the IC increases as the number of vocoder channels increases, a phenomenon not observed in the auditory nerve. In the previous chapter, it was hypothesized that this lack of increase may be due to stochastic undersampling of speech envelopes within each vocoder channel, and could be mitigated by integration at subsequent auditory nuclei. The data presented here appears to support this view, despite the IC SU representations comprising far fewer units.

In summary, this chapter presented the first evidence that the discriminability on neural representations of speech sounds in the inferior colliculus is robust to severe spectral and temporal degradation. A simple nearest neighbour classifier was able to perfectly discriminate consonant sounds provided either enough spectral cues were available or enough temporal cues. This trade-off does not reflect behavioural phenomena, where high rate-envelope cues provide only a modest boost in performance, primarily from the resolution of confusions between voiced and unvoiced consonants. In the present data, this does not appear to be the case as confusions between many consonants differing in many respects are resolved by the presence of these high rate cues.

Chapter 4
Auditory cortex

The auditory cortex is the most central predominantly unimodal stage in the auditory processing pathway, and is typically divided into primary and belt regions on the basis of both anatomical and physiological observations. While many across-species differences in exactly how the auditory cortex may be functionally divided exist, there is general agreement that there exists a core, tonotopically organised region and one or more belt regions. It is this core, or primary region that is the focus of this chapter. It has been proposed that several other functional topographies exists in this region suggesting the possibility of spatial codes for other acoustic features such as periodicity (Langner et al., 2009) however these are less well established. A robust finding is that responses are highly nonlinear (Sadagopan and Wang, 2009), locally heterogeneous (Kanold et al., 2014) and generally more transient than in more peripheral regions, particularly in anaesthetised preparations .

The neural code used to represent speech sounds in the primary auditory cortex is poorly understood. Clearly a code that utilises spike timing has the

potential for higher information throughput, but the ability of the auditory system to use such codes, or their efficacy for the formation of discriminable representations of speech sounds, is still disputed. Recent studies have examined the responses to speech sounds in the auditory cortex of the rat (Engineer et al., 2008; Perez et al., 2013; Ranasinghe et al., 2012). These studies demonstrate that the "distinctiveness" of neural responses in primary regions is well correlated with behavioural discriminability in the rat when spike timing information precise to between 1 and 50 ms is conserved, although it has been demonstrated that longer integration windows are required to enable robust discriminability when speech tokens are degraded by the addition of random noise (Shetake et al., 2011). These studies have also focussed on pairwise comparisons of speech tokens produced by a single talker. Generalisation to larger sets of speech sounds and the efficacy of these codes for the formation of response classes that are robust to inter-talker differences is yet to be investigated. Further, these explorations of putative neural codes for discriminable representations of speech sounds have been most extensively studied in the rat, necessitating an upwards shift of spectral content to better match the animal's audible range. The use of the guinea pig as an animal model makes it feasible to play speech sounds without this spectral shift, which facilitates direct comparison to human psychophysics.

The effect of degradation on neural representations of speech in the auditory cortex of the rat was also examined by Ranasinghe et al. (2012). They found that the qualitative features of neural representations of consonant sounds

were robust to noise vocoding and only began to degrade where only 4 or fewer spectral channels were used. This was mirrored by the rats' behavioural discrimination performance which asymptotes where at least 8 spectral channels were used, which is reflected in human psychophysical studies (Shannon et al., 1995; Xu et al., 2005). They also examined correlations between neural discriminability and behavioural discriminability using a neural classifier. Discriminability of consonants as a function of spectral or temporal degradation was only correlated when spike timing was conserved in the neural representations. These correlations were absent if only the spike count over the first 40 ms of the word-initial consonants was used.

In summary, there is a growing body of evidence that representations based on precise spike timing are necessary to predict behavioural discrimination of speech tokens from cortical neural representations, at least for word-initial consonants produced by a single talker. This appears to remain the case even when the temporal fine structure and high rate envelope cues available in the speech tokens are discarded as is the case in vocoded speech. It is not clear, however, to what extent this is specific to the reduced phoneme sets used in the aforementioned studies. It may also be the case that the efficacy of spike timing representations is specific to onset responses, as these are often quite distinct from sustained activity particularly in anaesthetised preparations and therefore only applies to word-initial consonants. Since consonant sounds in general are not preceded by long periods of silence, investigations into representations of word-initial consonants may not generalise well to the

representations of phonemes embedded in continuous speech. It is also not clear if such representations form phonemic classes at the level of the primary cortex that are robust to inter-talker differences. If this is the case, it would be supportive of the hypothesis that complex spatio-temporal acoustic features, such as the presence of periodicity or relative intensity across frequency bands, are re-encoded into spatio-temporal spike patterns. If not, then the importance of spike timing may be indicative of the representation of acoustic features specific to a given utterance of a given speech token. Whilst this may contain salient information, such as speaker identity, it is not necessarily related to neural mechanisms underlying phoneme classification. In this chapter, investigations into neural discriminability in the face of signal degradation are extended to multiple talkers and a larger group of phonemes. Target sounds are also embedded in a vowel context.

4.1 Methods

4.1.1.1 Subjects

Experiments were carried out on 20 tricolour guinea pigs (Cavia porecellus), which were bred in house (9 male, 11 female). All animals weighed between 378 and 1141g (μ=876, σ=176) and experiments were conducted in accordance with UK Home Office regulations.

4.1.1.2 Surgical preparation

The surgical procedures were similar to those described in 3.1.2. Briefly, animals were anaesthetised with urethane (0.5 g kg^{-1} in 20% solution; Sigma), and supplemental doses of Hypnorm (0.2 ml). Atropine sulphate (0.2 ml) was

also administered. A tracheotomy was performed, and the trachea was cannulated with polythene tubing. The external auditory meatus was exposed on both sides and the condition of the tympanic membranes was checked for abnormalities. Any detritus occluding the membranes was also removed. The surfaces of the auditory bullae were exposed and vented with long polythene tube, and the tissue covering the foramen magnum was perforated to relieve CSF pressure. The right temporalis muscle was further resected such the lateral suture was exposed and the posterior portion of the orbit was visible. A craniotomy of approximately 5mm by 5mm was positioned such that it was approximately bisected dorso-ventrally by the lateral suture and the rostral edge was aligned with bregma (see Figure 4.1). The dura under the craniotomy was resected under a microscope using sharp micro dissecting forceps and a hypodermic needle.

4.1.2 Stimuli

4.1.2.1 Pure tone stimuli

Similarly to the IC experiments, guinea pigs were presented with a set of 50 ms pure tones, with 10 ms raised-cosine onset and offset ramps. Tone frequencies were between 0.2 and 25.6 kHz and separated in quarter octave steps and each frequency was presented at levels between 0 and 80 dB SPL in 5 dB steps. These were presented every 500 ms.

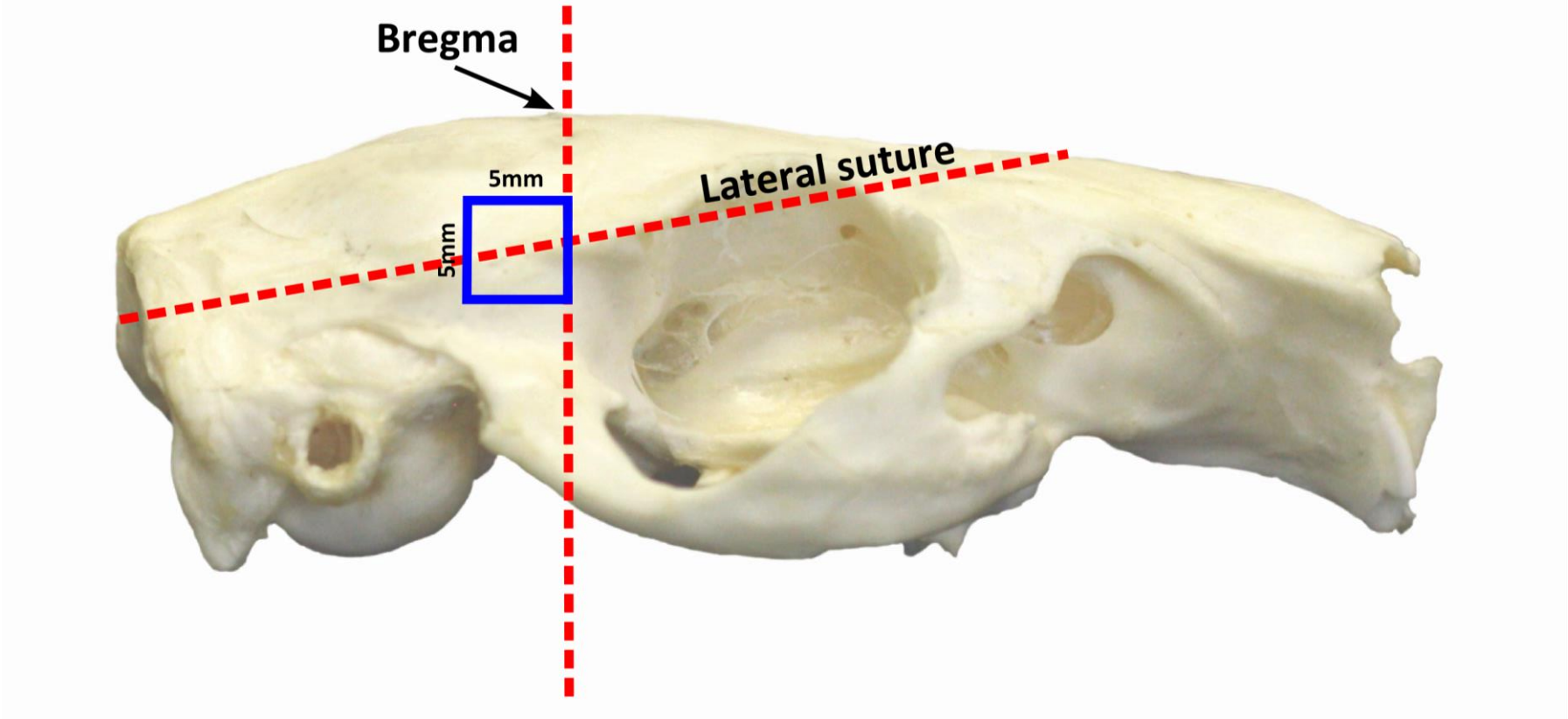

Figure 4.1: Photograph of a guinea pig skull showing the location to the craniotomy for recording in primary auditory cortex (blue) with the anatomical landmarks used shown in red.

4.1.2.2 Speech stimuli

The same speech corpus as that described in 2.1.2.1 was presented. The corpus comprised 3 exemplars of 16 VCVs produced by male talkers in which the vowel context was always */ɑ/* as in pa<u>l</u>m. These speech tokens were similarly parametrically degraded using a noise vocoder. In total, there were 11 vocoding conditions comprising natural speech, and vocoded speech with 1, 2, 4 and 8 spectral channels each with and envelope extraction filter cutoff of 16 and 500 Hz (see 2.1.2.2 for details of the vocoding process). Speech tokens were presented diotically using the same hardware as described in 3.1.3.1 with levels set such that the intensity of the pseudo-steady-state vowel portions of each stimulus was 70 dB SPL in the auditory meatus.

4.1.3 Automatic detection of auditory responsiveness

Multichannel electrode arrays allow the acquisition of large amounts of data to be collected over large areas of the cortex simultaneously. Some sites may not demonstrate responses that are modulated by auditory stimuli. An automated method to quantify the responsiveness of each recording site was used such that the effect of predominantly noisy recordings on classifier performance could also be investigated. The responses to natural speech were used for this analysis. Recordings were initially bandpass filtered between 0.3 and 3 kHz (6^{th} order Butterworth). The portion corresponding to the central 500 ms of each speech token was divided in to 50 ms bins and the RMS was calculated for each of them. The mean and standard deviation of the maxima for each recording was evaluated, as was the mean and standard deviation of RMS of the last 50

ms of the 800 ms response window, during which no auditory stimulus was being presented. The statistic used to quantify responsiveness was *d'*, which is generally used to quantify the separation of signal and noise in signal detection theory and is given by the standardised difference between the distribution means. It may be calculated using the following equation.

$$d' = \frac{\mu_S - \mu_N}{\sqrt{\frac{\sigma_S^2 + \sigma_N^2}{2}}}$$

Only recording sites that had a *d'* value greater than 1 were used in subsequent analyses.

4.2 Results

Responses were recorded at 382 sites in the auditory cortex, of which 31 were classified as SU. Of the 239 sites recorded using Neuronexus multi-electrode arrays, 208 were deemed online to be of acceptable quality to complete presentation of the speech corpus. These were later analysed using the method described in 4.1.3.

4.2.1 Pure tones

Frequency response areas of neurons in the auditory cortex have been recorded in several species including the rat (Sally and Kelly, 1988), the cat (Sutter, 2000) and the guinea pig (Redies et al., 1989; Wallace et al., 2000). These studies invariably examine frequency selectivity in neuronal responses to pure tones and tonotopic organisation in core areas. The most detailed examination of FRA shapes was presented by Sutter (2000), who noted that, while stereotypical shapes could be subjectively identified, quantitative

analysis did not suggest intrinsic discrete pure-tone response classes. Further, there was a greater prevalence of I-shaped response areas than in more peripheral nuclei indicative of increasingly intensity invariant tuning in the cortex. The aim of investigating pure tone response properties prior to the presentation of the speech stimuli is to provide a set of neural descriptors that may be later be used to investigate neural mechanisms underlying the production of discriminable responses to speech in a way that can be related to the aforementioned pure tone studies, which analysed substantially larger populations.

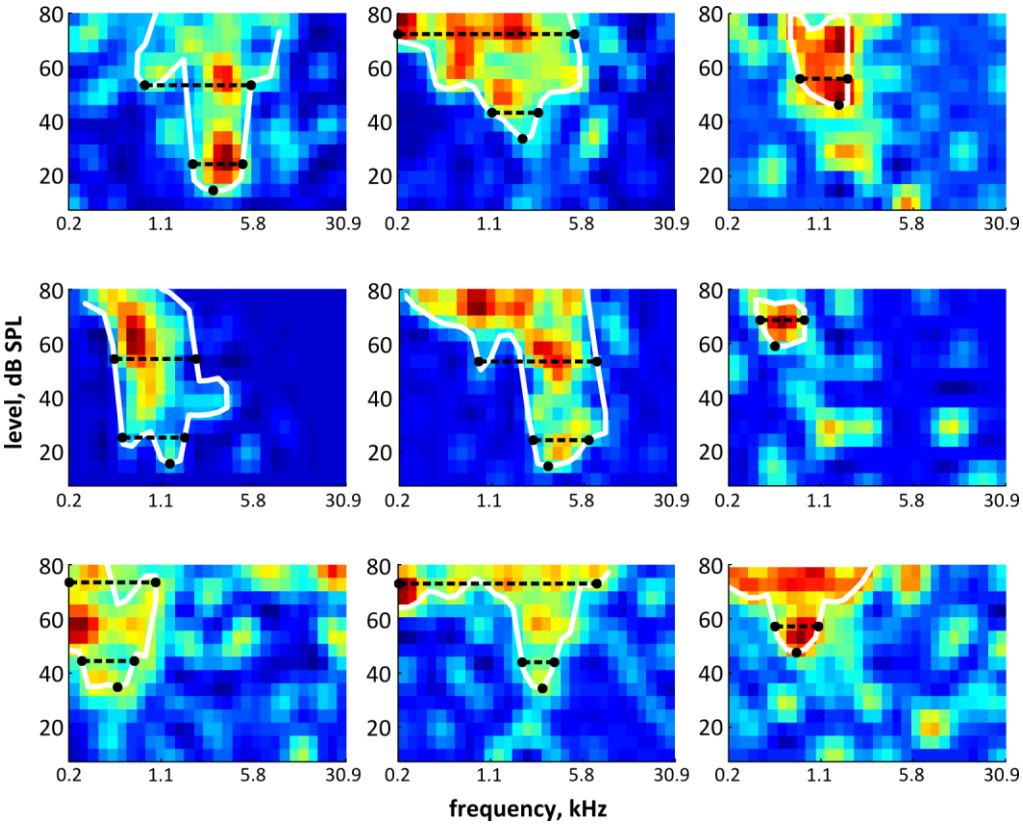

Figure 4.2: Frequency response areas of 9 single units in the auditory cortex of the guinea pig. Response area contours were fitted algorithmically (see 3.1.4.3) and shown in white. CF (●) and bandwidths at 10 and 40 dB above threshold are also indicated (---).

In the present study, FRAs were produced at 20 of the 31 SU recording sites. Of these, only 9 demonstrated responses that were modulated by pure-tone stimuli. These response areas are shown in Figure 4.2. Each of these demonstrate complex selectivity to both frequency and intensity with non-monotonic and, in several cases, multi-peaked rate-level functions at CF. It may be argued that the fragmented nature of these areas is a product of the low spontaneous firing rate, predominantly phasic response properties and the small number of repetitions. However, there is clear indication of frequency selectivity.

Of the 208 recorded MU sites, 96 had FRAs that were successfully fitted using the algorithm described in 3.1.3.2. A summary of the tuning characteristics extracted by this algorithm for both the SU FRAs shown above and the MU FRAs is shown in Figure 4.3. The upper left panel shows the distribution of characteristic frequencies. The distributions for both the SU and MU recordings are confined almost exclusively to the region between 0.3 and 5 kHz. Whilst this is clearly not representative of the audible range of the guinea pig (see 1.7), nor the range of CFs recorded in other microelectrode studies, this covers much of the bandwidth of the speech stimuli used in this study (0.1 to 4 kHz). However, frequencies lower than 400-500 Hz are under-represented in the MU dataset, with 94% of sites most sensitive to frequencies >500 Hz. This has implications for representations of vocoded speech, which are discussed in section 4.3. The distribution of CFs is also representative of a systematic bias towards regions most sensitive to frequencies within the power spectrum of the stimuli. Figure 4.4 shows an image of a craniotomy over the auditory cortex with the dura

resected and the targeted location for electrode penetrations relative to anatomical landmarks. This was intended to sample the region most sensitive to the speech stimuli, commensurate with the low frequency region of the primary auditory cortex (Wallace et al., 2000).

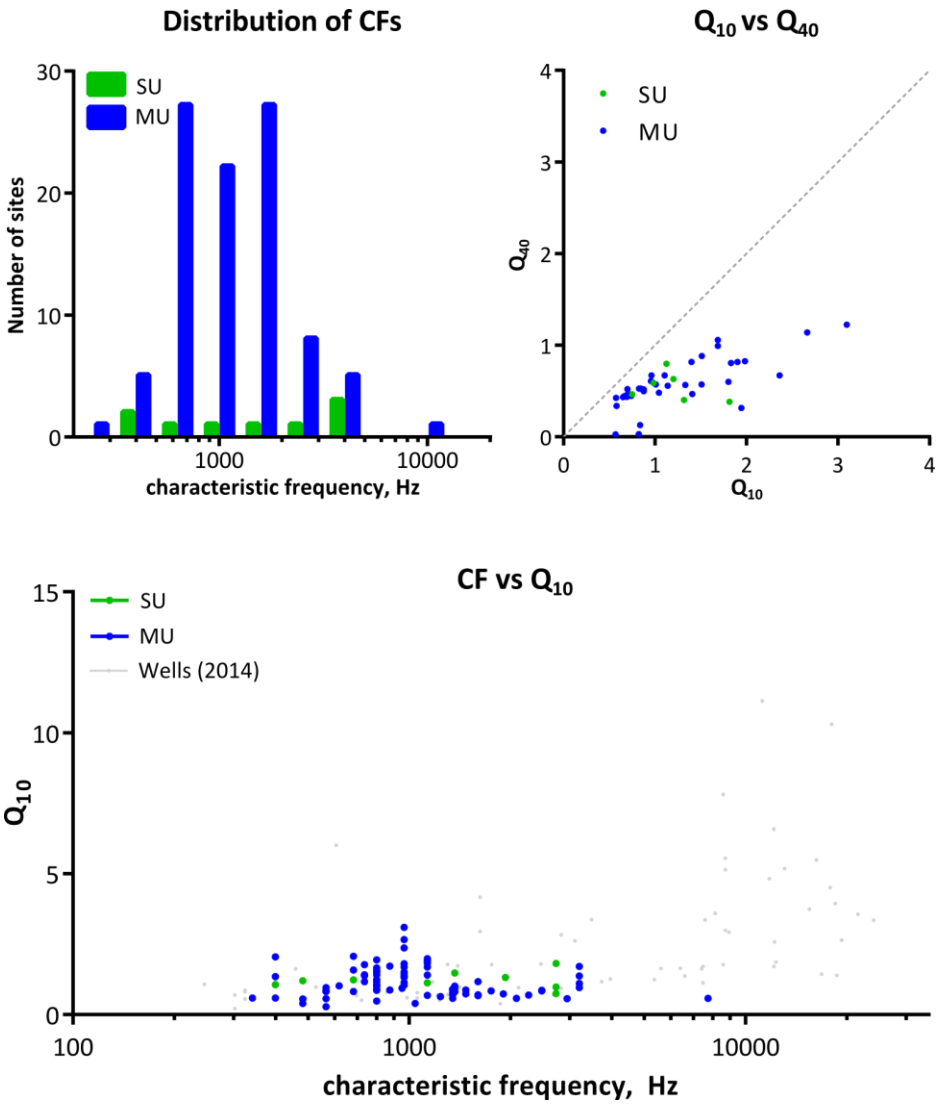

Figure 4.3: Pure-tone response characteristics in the auditory cortex of anaesthetised guinea pigs. Data included from Wells (2014) for comparison.

The present results show that recordings were predominantly made in a region of the auditory cortex that shows selectivity to pure tone frequencies. This, in conjunction with more extensive functional maps of the guinea pig auditory cortex and the electrode placement relative to anatomical landmarks support the assertion that recordings were made in the primary region (AI). There is an underrepresentation of sites that are optimally sensitive to very low frequencies of the order of a few hundred Hertz. However, this does not preclude sites with higher CFs from responding to acoustic energy in these spectral regions. There is also a lack of recording sites optimally sensitive to high frequency pure tones, however previous microelectrode studies suggest that these sites would not respond strongly to the speech stimuli.

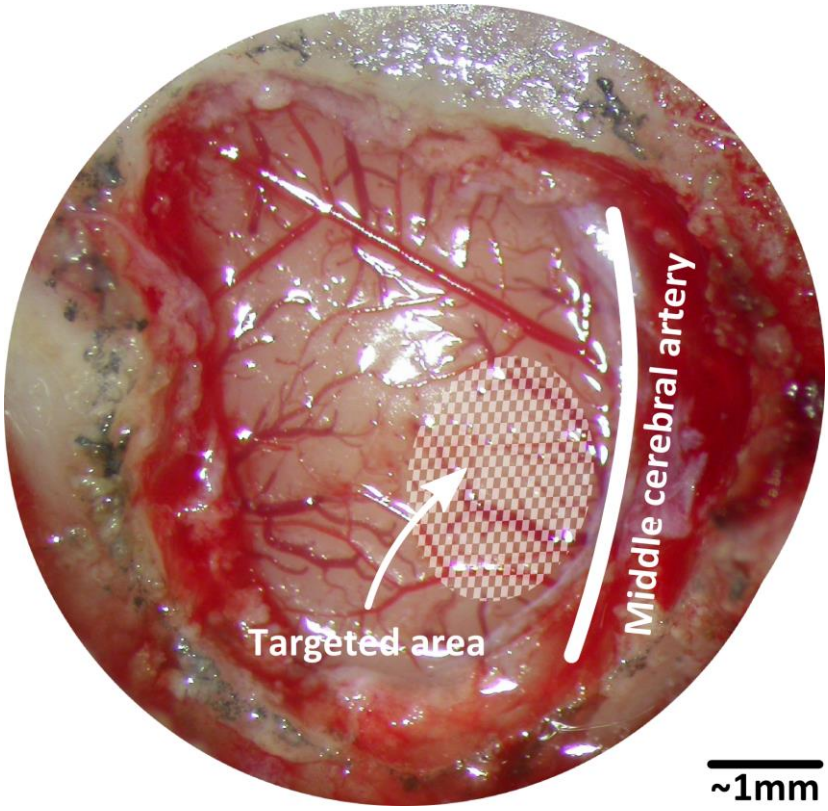

Figure 4.4: Photograph of the exposed auditory cortex showing the middle cerebral artery and the target location of electrode penetrations.

4.2.2 Natural speech

The responses to at least 6 presentations of the natural speech corpus were recorded at all 31 SU sites. The average responses of each of the SU sites to each consonant produced by 1 of the 3 talkers are shown in Figure 4.5. The SU response patterns show a visibly more sparse representation of each sound with much lower average spike rates than in auditory nerve or IC representations. Nonetheless, the PSTHs shown in red clearly show differential ensemble responses to each of the consonant sounds, with some eliciting a much greater central peak than others, for example (e.g. */ada/* vs */ama/*). Surprisingly, however, the presence of this peak does not appear to correspond to the presence or absence of a plosive and therefore is not well correlated with the overall stimulus envelope. For example, this peak is very prominent in the response to */afa/*, but not in that to */ata/*. This may be suggestive that these responses are either selective to narrow spectra, in which envelopes do not reflect the broadband stimulus envelope or that responses are selective to higher-order spectro-temporal features, such as formant transitions.

Responses to 10 presentations of each stimulus were recorded at each of the 208 responsive MU sites. Of these, 90 demonstrated pure-tone response areas that could be fitted using the algorithm described in 3.1.4.3. Neurograms constructed from these 90 sites are shown in Figure 4.6, in which sites are organised by CF, such that the higher site numbers correspond to sites most sensitive to high frequency tones and vice versa. A common feature of many of these responses is the presence of a peak in the ensemble response

corresponding to the onset of the second vowel, which varies in magnitude depending on the previous consonant. For example, a very pronounced peak occurs following the consonant */ʃ/*, but this peak is very much reduced for the sequence */aya/*. These two consonants occupy different frequency regions and also have different average intensities; the consonant */y/*, for example, is voiced and has a similar intensity to the preceding vowel sound whereas the consonant */ʃ/* is unvoiced, a-periodic and lower intensity. Frequency specific responses are also discernible in these representations. For example, strong responses can be seen in the high frequency population the medial portion of */aʃa/*, which are reduced during */ana/*, however, these representations also contain features that are not apparent in spectrographic representations, such as the multiple peaks in the response to */aka/* or the medial peak in */aʃa/*. Another feature not as apparent in auditory nerve (see Figure 2.4) or the inferior colliculus (see Figure 3.9) is the reduction of sustained activity. Responses show predominantly phasic activity, which is not maintained for the duration of the stimulus. This activity reoccurs at subsequent points, presumably corresponding to spectro-temporal dynamics in the stimulus, but activity is not continuous for steady state portions, such as the final vowel.

The classifier was initially run using a single-talker paradigm in which training and testing was carried out for each talker individually. The response patterns used were the ensemble PSTH (with 1 ms bins), spike counts at each site and the full spatio-temporal neurograms, also using 1 ms bins, corresponding to the red line, blue bars and neurogram images in Figures 4.5 and 4.6, respectively. This analysis, and all subsequent analyses, were done for the SU and MU

datasets separately. All 31 sites were used to generate SU representations. Out of the 208 MU sites, 196 met the inclusion criterion described in 4.1.3, at which responses to the vocoded versions of the speech token were also recorded (see 4.2.3). The results of these analyses are shown in Figure 4.7. For every condition, neural discriminability was significantly better than chance ($p = 0.01$), except for where SU neurograms were used. However, none of these representations were sufficient to produce a high level of discriminability even for the single-talker token discrimination task. Previous studies have shown that correlations with behavioural discriminability exist where spatio-temporal activity patterns using spike timing precise to 1 ms are used, even when they comprise as few as 5 MU sites. It may seem surprising, therefore, that a similar classifier would show such poor discriminability. A number for possible explanations exist for the poor performance of the classifier for these cortical representations. Firstly, the stimulus set differs from previous studies in being more comprehensive. It may be that 1 ms precise responses can be used to discriminate small sets of certain consonant sounds, but not others. This could be indicative of neural codes on multiple timescales being efficacious for certain consonant groups, but it may be that intermediate integration windows are required for the formation of a larger number of response classes. Secondly, it could be that intermediate integration windows are required for sounds that are not word-initial, implying that there may be differential decoding strategies are optimal for onset and medial consonant sounds.

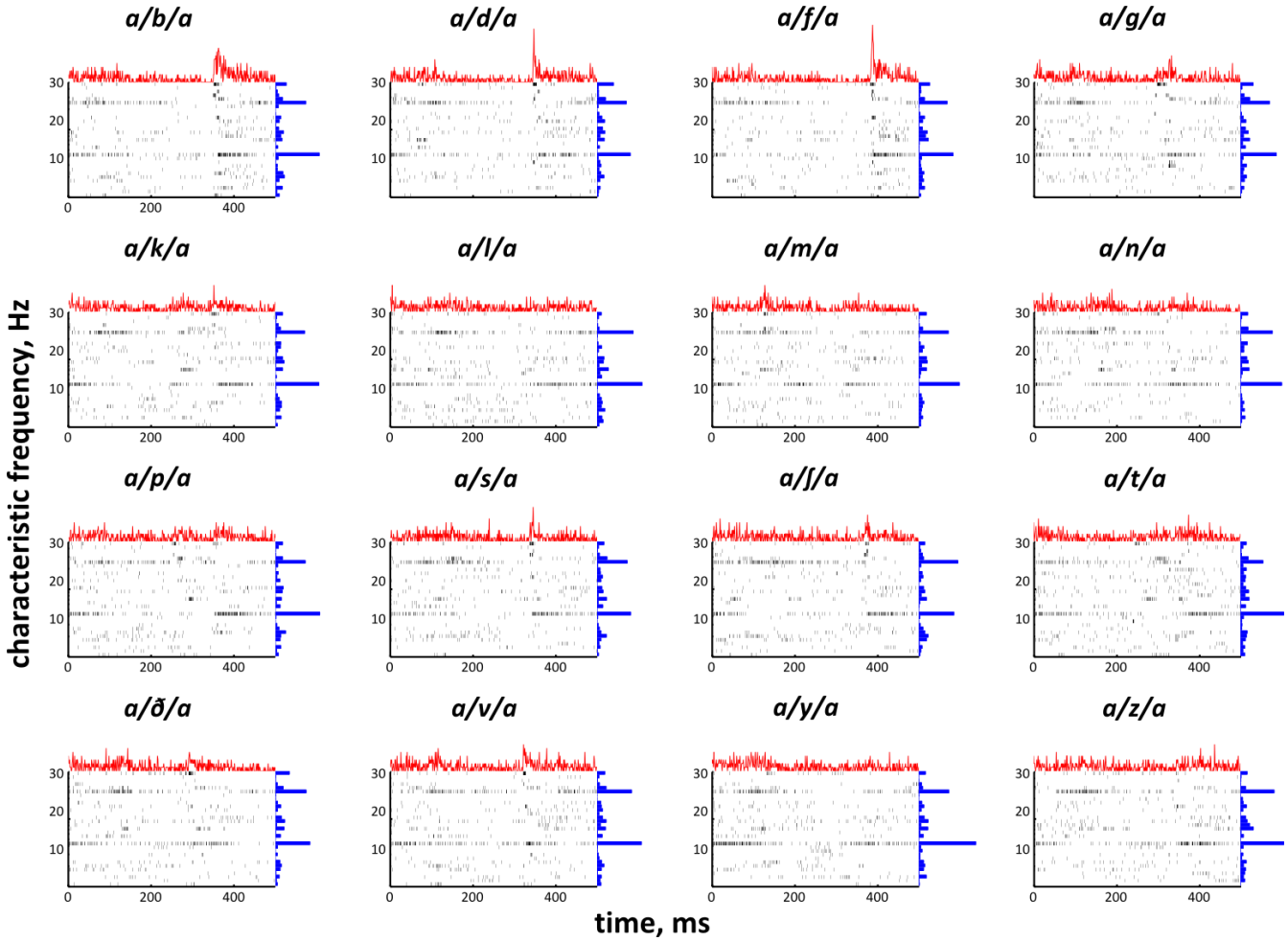

Figure 4.5: Average neurograms produced from 31 cortical SU sites in response to the 16 VCVs produced by a single talker. Red lines show the ensemble rate profile and blue histograms show the spike count for each individual unit.

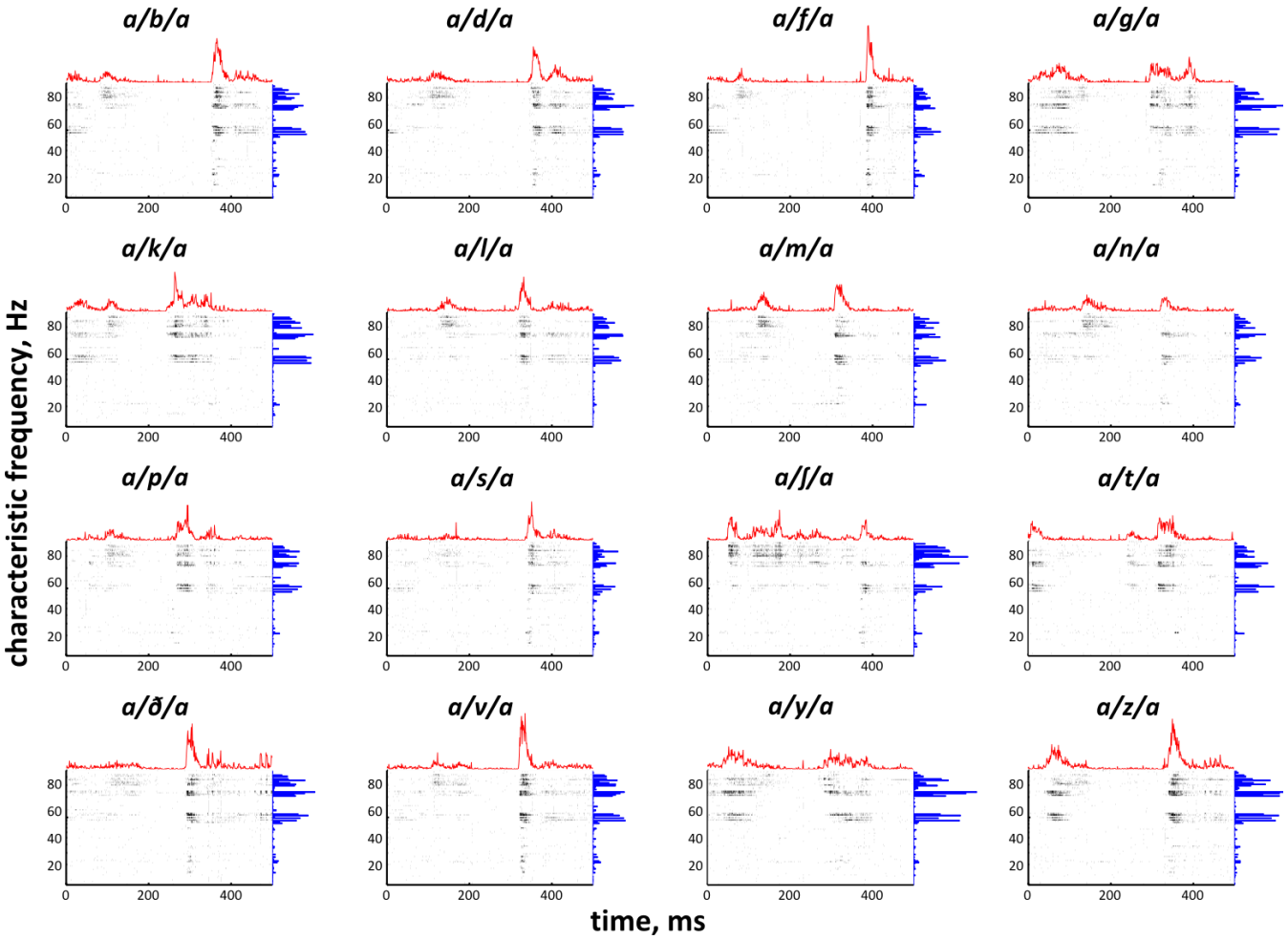

Figure 4.6: Average neurograms produced from 106 cortical MU sites in response to the 16 VCVs produced by a single talker. Red lines show the ensemble rate profile and blue histograms show the spike count for each individual unit.

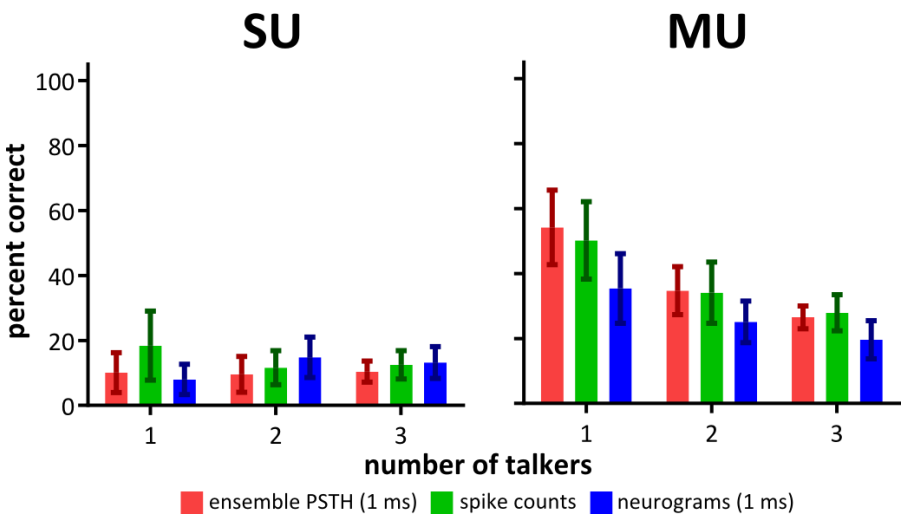

Figure 4.7: The effect of number of talkers and spatio-temporal resolution on the discriminability of cortical representations of consonants.

Classification was then performed using a multi-talker paradigm in which the training set comprised representations that were combined across each of 2 or 3 talkers. Classifier performance based on SU responses unsurprisingly remains low for the 2 and 3 talker conditions. Discriminability of MU representations decreases monotonically with the addition of more talkers. The decrease in performance is greatest when the number of talkers is increased from 1 to 2. However, this monotonic decrease could be indicative that the responses at these timescales represent fine grain acoustical properties of each token, which is essentially corrupted when average representations are combined across the talkers. Such properties can indicate perceived cues such as intonation or talker identification, for example, and may be used to detect token repetition, which the human auditory system is highly sensitive to but do not necessarily enable robust classification of phonemes for which neural codes with intermediate integration windows may be optimal.

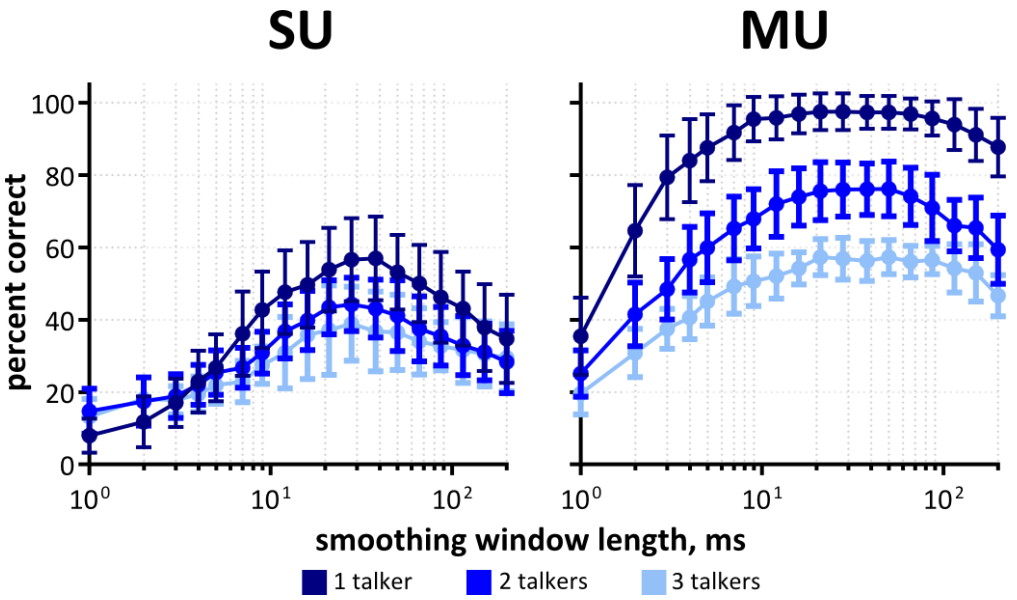

Figure 4.8: Discriminability of SU and MU neurograms as a function of temporal smoothing window duration and number of talkers.

Neurograms were parametrically temporally degraded by row-wise convolution with a rectangular window of varying length prior to classifier training and testing. This was done for both SU and MU representations. The results of this analysis for the single-talker paradigm are shown in the lower plots in Figure 4.8, represented by the darkest of the 3 curves. The relationship between neural discriminability of SU representations and temporal smoothing is non-monotonic with a clear peak corresponding to 30 – 40 ms, where discrimination reaches approximately 60%; significantly higher than that of representations using either 1 ms precise timing or the spike counts alone. The same analysis carried out on the MU dataset shows a similar shape, however the peak is less prominent, which may be attributable to a ceiling effect as a result of reaching near-perfect discriminability for temporal smoothing windows between 10 and 100 ms in duration. A similar analysis was carried out using a multi-talker paradigm. The results are also shown in Figure 4.8,

indicated by the light curves. There is a decrease in peak discriminability as the number of talkers increases. However, there is no clear shift towards longer optimal smoothing window lengths.

There is often an implied dichotomy in discussions relating to the temporal precision of the neural code between spike timing and spike rate codes with spike timing referring to millisecond precision and rates referring to the duration of a given stimulus, often of the order of 100 ms. However in this case spike counts over intermediate windows are optimal for producing discriminable representations of medial consonants. This could still constitute a rate-code as defined by Theunissen and Miller (1995) if the salient spectro-temporal features of the stimulus occur at the same rate or faster. For example, it has been demonstrated that periodicity can be represented spike rates over longer epochs in the auditory cortex indicated by bandpass rate modulation transfer functions (Joris et al., 2004). Similarly, this could constitute a temporal code if temporal features over longer time scales are more important.

If the former is true, it may be that rate codes are being used to represent idiosyncratic fine grain acoustic features of each token that can be utilised to identify repeated presentations of the same token, a task that human listeners are surprisingly good at (Agus and Pressnitzer, 2013), but that these representations become corrupted after combining across each of the 3 talkers, and are not useful for forming robust representations of phoneme classes. This causes the reduction in discriminability as more talkers are added, which could be assumed to continue decreasing as more talkers are added. If the latter, then

these representations could relate to spectro-temporal dynamics over longer epochs that do generalise across talkers, but high feature selectivity and a small number of sites precludes the classifier from performing perfect consonant discrimination, and a much higher number of recording sites forming a higher-dimensional representation of each stimulus is necessary to create phonemic clusters that can be utilised by this or other classifiers.

It has been previously noted that cortical responses are highly locally heterogeneous. It may be that this optimal response integration window relates to some compromise across the population and that individual units produce responses that are optimally discriminable using different integration windows, either because they encode acoustic features occurring at disparate timescales or because a multiplicity of neural codes are used for acoustic features happening at similar timescales. The classifier was trained and tested using responses from individual SU and MU units, such that the training set comprised PSTHs using a 1 ms bin size which were likewise parametrically smoothed by convolution with a rectangular window. The results of these analyses are shown in Figure 4.9, in which site CF is denoted by the ordinate of each bar. The units that produce the most discriminable responses are shown as darker shades. For both single- and multi-talker paradigms, there is no clear range of CFs that corresponds to better discriminability. However, those units that produce the most discriminable responses, similarly to the population analysis, show optimal integration windows of approximately 10 – 100 ms. The single-talker paradigm relies on response reliability for successful classification. Despite evidence of a high degree of temporal precision in responses in the

auditory cortex, these do not enable discrimination between a large set of speech sounds when individual units are considered. This is likely due to the combination of sparse spiking and the classifier temporal shifting mechanics whereby a single spike occurring at a certain time after stimulus onset cannot be distinguished from one happening up to 100 milliseconds later (corresponding to the maximum relative shift allowed by the classifier).

In this section, it was demonstrated that VCV phoneme sequences produce discriminable representations in the auditory cortex of the anaesthetised guinea pig. Integration windows of between 10 and 100 milliseconds are optimal for correct classification of neural response patterns. If only millisecond precise responses are considered, discriminability is very poor. Responses are highly heterogeneous and poorly correlated with stimulus envelope and the majority of units are characterised by low spike rates, although a minority of SU sites show sustained responses. The discriminability of individual unit responses shows no clear relationship with pure tone tuning, however only the portion of the cortical surface that was optimally sensitive to pure tones within the bandwidth of the stimuli was targeted. Discriminability of responses decreases with additional talkers. It is not clear whether the neural representations comprise a rate code for high rate envelope cues, a spike timing code for slower temporal stimulus dynamics or an abstract representation of higher order spectro-temporal features. If the first is true, the attenuation of high rate envelope cues should decrease discriminability of the neural representations. The relative importance of spectral and temporal cues in the stimuli for the production of discriminable representations of speech

tokens produced by a single talker and consonant classes produced by multiple talkers is examined in the following section.

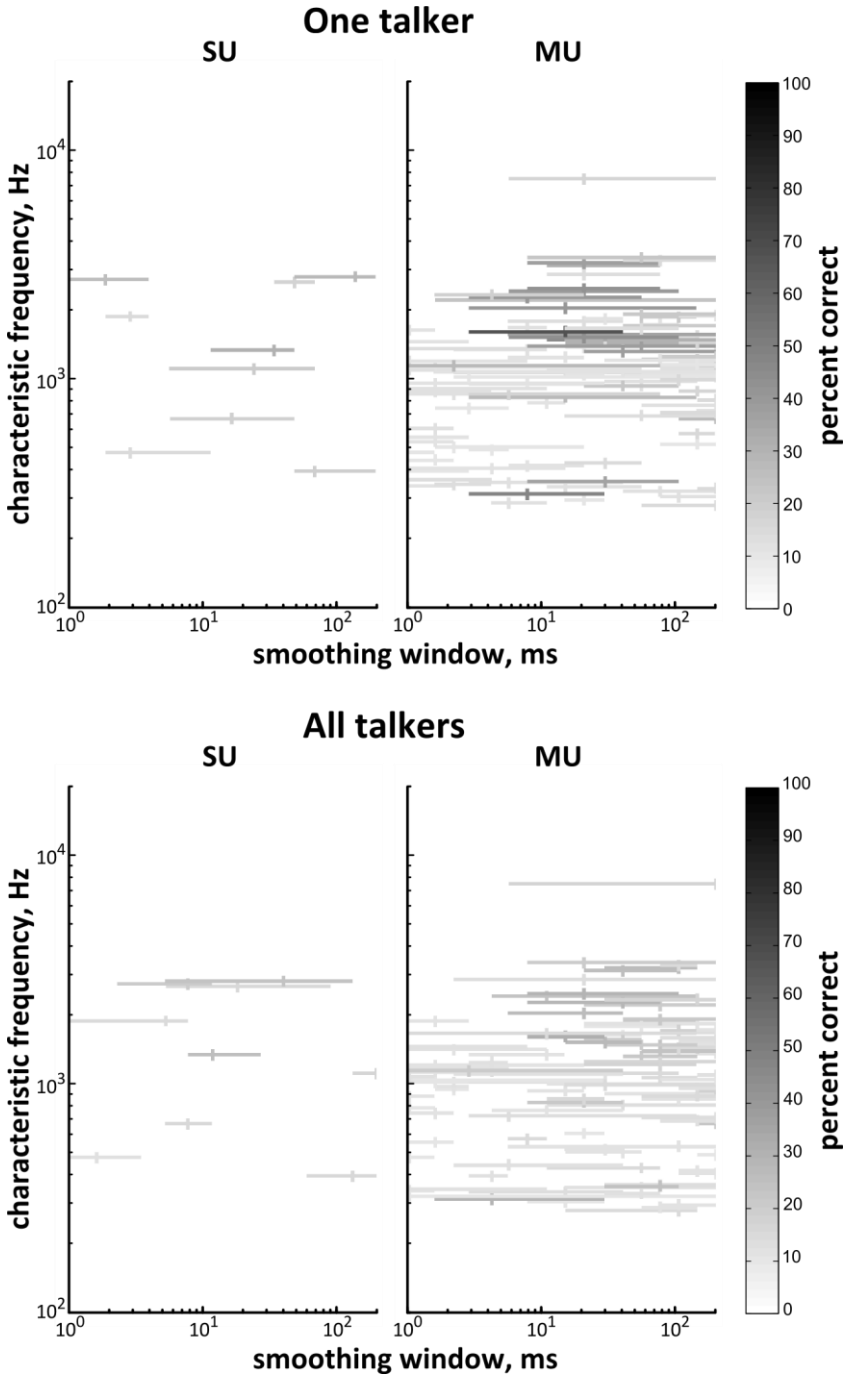

Figure 4.9: Discriminability of individual SU and MU responses as a function of CF and smoothing window for a single- (upper) and multi-talker (lower) paradigm. Vertical bars denote optimal smoothing window. Horizontal bars denote the range over which discriminability is over 90% of the maximum for each unit.

4.2.3 Vocoded speech

The speech corpus in the previous section was parametrically degraded using a noise vocoder. The number of spectral channels used was set to 1, 2, 4 and 8 for which the envelope extraction filter cutoff was set to 16 Hz. Envelope extraction filter cutoffs were also set to 50, 160 and 500 Hz for the single channel condition, resulting in a total of 8 vocoding conditions including natural speech. At SU recording sites, only data for the single channel, 16 and 500 Hz envelope conditions and 8 channel, 16 Hz condition are presented here as few sites achieved the required stability for the presentation of the extended stimulus set. Responses to at least 6 presentations of each speech token for each vocoding condition were recorded at 18 of the SU sites. Responses to 10 presentations of each token were recorded at all of the 196 MU recording sites in the previous analysis for each of the 8 vocoding conditions.

The spatio-temporal response patterns were similarly smoothed by convolution with a rectangular window of parametrically varied duration. And discriminability was quantified using the same neural classifier separately for each vocoding condition using both single- and multi-talker paradigms. The results are summarised in Figure 4.10. For the single talker paradigm, the discriminability of SU representations is significantly above chance for all conditions. Discriminability is significantly greater for the 8 channel, 16 Hz condition than for the 1 channel, 16 Hz condition ($p = 0.01$) but is not significantly greater in the 1 channel, 500 Hz condition than in the 1 channel 16 Hz condition. Where MU representations were used, mean optimal

discriminability is greater than that measured perceptually for single channel vocoded speech. There is a significant, but small increase in discriminability where high rate envelope cues are not attenuated ($p = 0.01$), but no significant increase when the envelope extraction filter cutoff is increased above 50 Hz. Discriminability is greatly improved with additional spectral channels with the classifier reaching asymptotic performance with only 2 channels. Surprisingly, when the number of vocoder channels is increased from 4 to 8 channels, discriminability decreases.

That classifier performance based on SU representations is poor may seem unsurprising, given that the number of individual neurons is small. However, it has been demonstrated previously in this thesis that the discriminability of representations formed by similarly sized populations in the auditory midbrain is much larger and can exceed behavioural discriminability for a single talker task even for the most heavily degraded vocoder conditions. The average firing rate of these cortical neurons is also significantly lower than those in the IC. This, in conjunction with the success of the classifier based on the substantially larger MU population is supportive of a distributed, sparse representation of the speech sounds in which individual neurons show a higher degree of selectivity to stimulus features.

In previous chapters, it was demonstrated that a lack of spectral information may be offset by temporal cues such that neural discriminability is restored to levels that closely match those for natural speech. Whilst there is some benefit from the presence of high rate envelope cues, it is not as pronounced, which

reflects observations in human psychophysics. Envelope extraction filter cutoff has a pronounced effect on the perceived quality of the speech sounds, and therefore must produce distinct internal representations. The extra temporal detail increases discriminability in more peripheral nuclei in a profound way, but this result is suggestive that in the auditory cortex, the acoustic features that give rise to the percept of sound quality are encoded in a different way such that representations of timbre are distinct from concrete representations of specific high rate temporal features. This can explain why the addition of high rate temporal features increases the "distinctness" of neural representations peripherally, but do not appear to benefit speech token recognition.

The pattern of results is very similar for the multi-talker paradigm. Again, the addition of multiple talkers reduces classifier performance for every vocoding condition. The mean discriminability of responses to single channel vocoded speech still increases monotonically as the envelope extraction filter cutoff is increased, although there was no significant difference between any of the groups (one-way ANOVA, $p = 0.05$). The use of 2 or 4 spectral channels increased discriminability, although there was an decrease in discriminability when the number of channel was increased to 8.

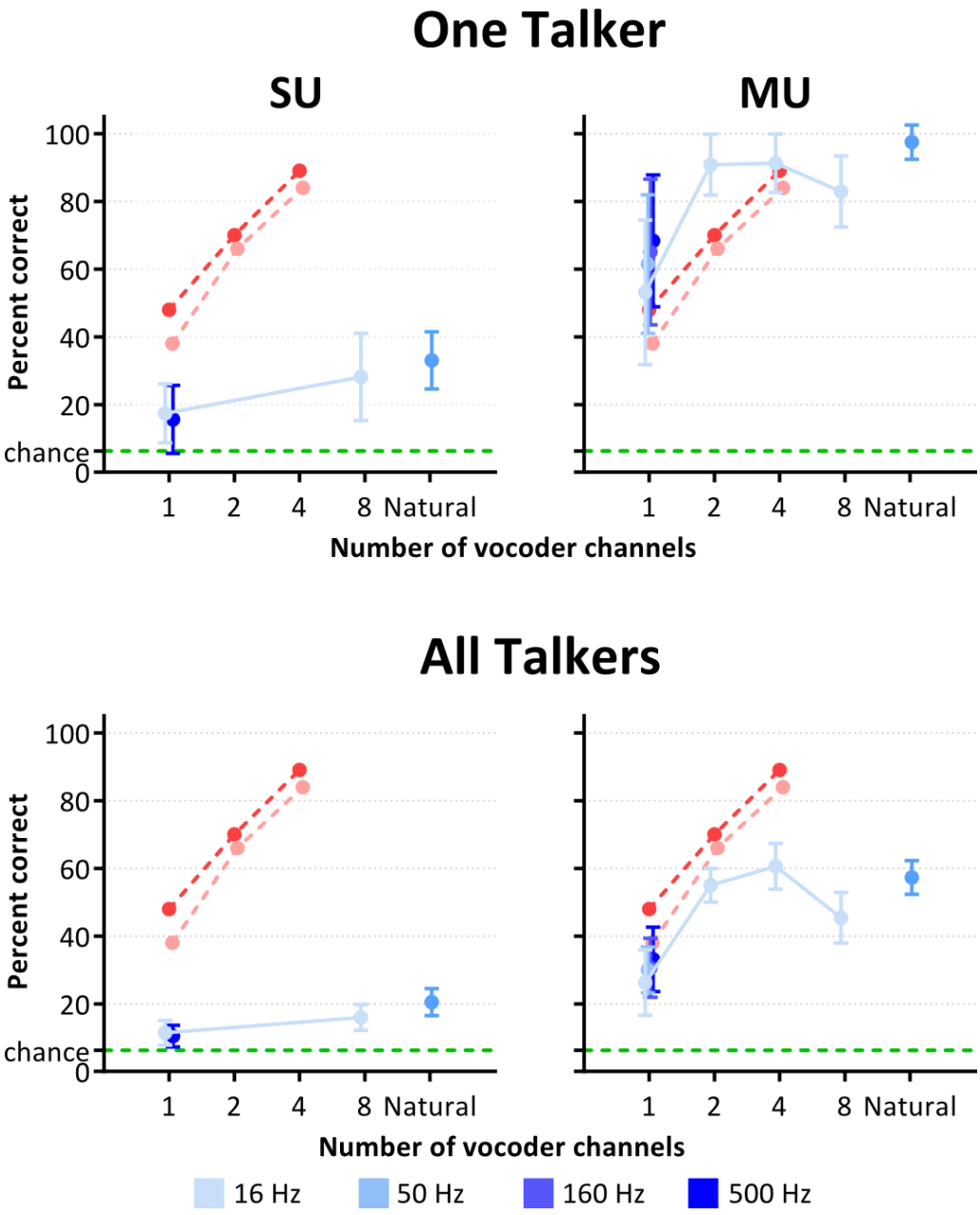

Figure 4.10: Discriminability of cortical representations of vocoded speech. The classifier was trained on representations averaged across multiple presentations of token produced by a single talker (upper) or combined across all 3 talkers (lower). All points correspond to optimal smoothing windows. Red lines indicate human behavioural discriminability of an analogous stimulus set (reproduced from Shannon et al., 1995).

The effect of smoothing window duration on classifier performance for each of the vocoding conditions for the single talker paradigm is shown in Figure 4.11. The upper 2 plots present data from all vocoding conditions where the envelope extraction filter cutoff was set to 16 Hz. Regardless of the number of spectral channels, these curves are a remarkably similar shape to that observed for natural speech, whereby classifier performance shows a non-monotonic, peaked relationship with smoothing window duration. Optimal discriminability corresponds to smoothing window durations of approximately 8 – 30 ms. This is despite the fact that stimulus modulations corresponding to these timescales have been attenuated by the envelope extraction filter. Although the smoothing window duration is of the order of 10s of milliseconds, this is indicative of a temporal code in that stimulus changes occurring over epochs corresponding to at most 16 Hz, the reciprocal of which is equal to 64 milliseconds, are optimally represented by neural responses with finer temporal precision. However, the possibility that classifier performance is driven by envelope fluctuations at higher rates cannot be ruled out, as the filter roll-off was relatively mild (18 dB/octave), commensurate with that used in the human psychophysics study that used an analogous stimulus set (Shannon et al., 1995). Whilst envelope fluctuations at frequencies much greater than an octave faster than 16 Hz are severely attenuated, this does not preclude the use of cues corresponding to integration windows of the order of 30 ms being available.

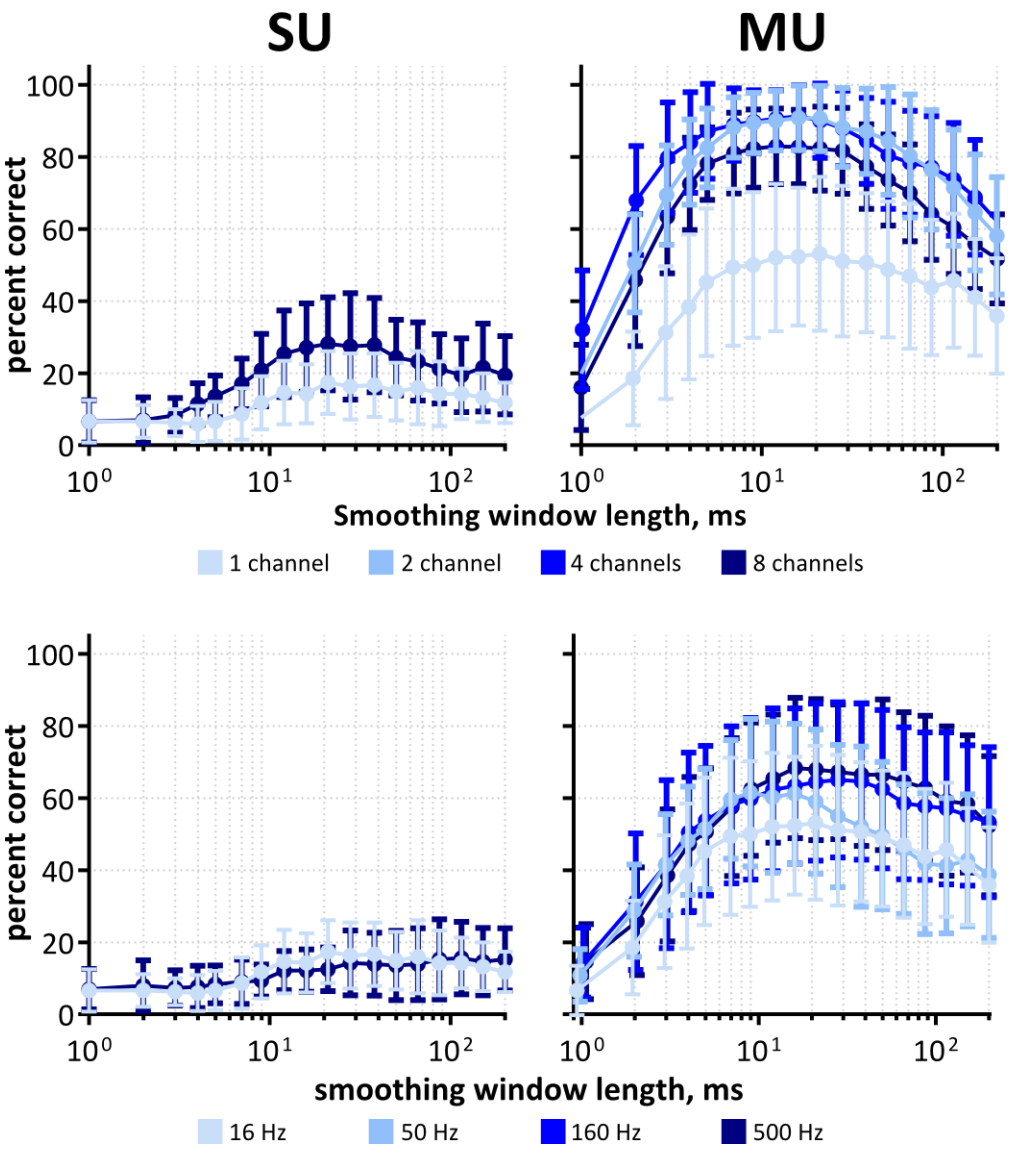

Figure 4.11: The effect of smoothing on neurogram discriminability of tokens produced by a single talker. SU representations comprised 18 sites, MU comprised 196. Upper plots correspond to 16 Hz envelope extraction filter cutoff, lower plots correspond to single channel speech.

The lower right plot of Figure 4.11 shows discriminability of MU representations of single channel vocoded speech as a function of temporal smoothing as the envelope filter cutoff is varied between 16, 50, 160 and 500 Hz. An increase in the intensity of high rate envelope cues does not cause a shift towards shorter window durations. In fact, increasing the envelope extraction filter cut-off creates neural representations that are more robust to sever temporal degradation showing that spike rates are modulated by envelope fluctuations corresponding to periodicity information improving discriminability if long integration windows are assumed.

A similar plot for the multi-talker classifier training paradigm is shown in Figure 4.12. Whilst SU performance remains poor, the discriminability of MU representations of 2, 4 and 8 channel vocoded speech as a function of temporal smoothing shows a very similar profile to that in shown in the upper right plot of Figure 4.11. For single channel speech, however, the function does not show a clear central peak and remains relatively flat as the smoothing window increases to above approximately 10 ms. The fact that these curves remain the same shape with the peaks diminished could be indicative again that similar cues are being used by the classifier to identify the correct speech token produced by the same talker as that which produced a give test pattern, but that the average representation of this token is noisy as a result of being combined with the representations of token produced by the other two talkers. The other possibility is that the central peak does relate to an intrinsic optimal timescale for discriminable neural representations of consonants, but that the subsampled population is unable to take advantage of additional spectral cues

available in the 8 channel condition, for example, due to the particularly small sample of units sensitive to low frequencies. There is no benefit to increasing the envelope extraction filter cutoff for single channel speech.

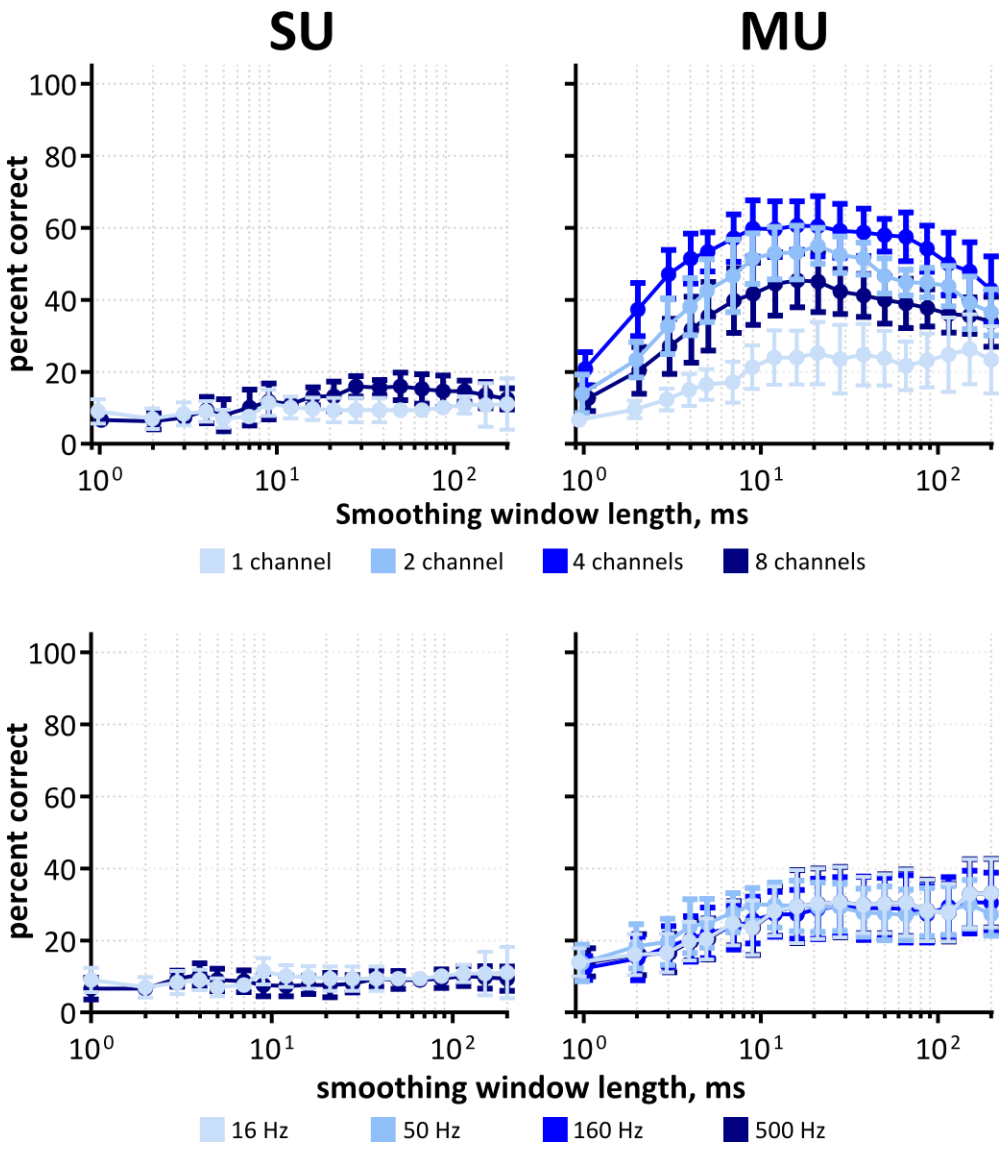

Figure 4.12: The effect of smoothing on neurogram discriminability of consonants produced by three talkers. SU representations comprised 18 sites, MU comprised 196. Upper plots correspond to 16 Hz envelope extraction filter cutoff, lower plots correspond to single channel speech.

4.3 Discussion

Electrophysiological recordings were made in the auditory cortex of anaesthetised guinea pigs. Electrode placement was done relative to anatomical landmarks and pure-tone frequency response areas were recorded to verify approximate electrode placement, which was in the rostral, low frequency portion of the primary region.

The neural classifier that was developed in Chapter 2 was applied to cortical representations of natural and vocoded speech in anaesthetised guinea pigs. Two classification paradigms were used for each of the vocoding conditions. The first was a single talker paradigm in which the classifier was trained and tested using neural activity in response to speech tokens produced by a single talker. This is analogous to several published studies, which used comparable nearest neighbour classifiers to discriminate sets of speech tokens (Engineer et al., 2008; Perez et al., 2013; Ranasinghe et al., 2012; Centanni et al., 2014). These studies differ in the number of consonants used ranging from 20 to only 4, however a common finding is that classifier performance is best correlated with pair-wise behavioural discriminability of the same speech sounds in the same animal model only when spike timing information precise to 1 ms is conserved. Correlations decrease as a result of compromised discriminability of some consonant pairs if spike timing information is discarded.

This finding does appear to be in contrast to results presented here, which demonstrate that to millisecond precise position of spikes in spatio-temporal activity patterns essentially constitutes noise, which makes token recognition

difficult. Only a degree of temporal smoothing, of the order of 10 – 100 ms enables reliable token discrimination. However, due to several methodological differences these findings are likely to be representative of different mechanisms. Firstly all of the aforementioned studies examined representations of consonant sounds occurring at sounds onset. Clearly there are profound differences between onset and sustained representations of steady-state sounds in the auditory cortex, which could feasibly affect differential neuronal encoding window durations for word-initial and medial consonant sounds.

The use of medial consonant sounds in the present study also necessitated the development of a more sophisticated classifier incorporating relative temporal shifts between test and training activity patterns. Whilst it has been shown that millisecond precise responses do exist in the auditory cortex and can be used to differentiate differences in temporal features such as voice onset time, for example, this may be contingent on an external reference to stimulus onset time, something that is not available for the brain to use whilst performing speech recognition and a task that is compounded by adaptation caused by sounds preceding the target consonant.

Investigations were extended further by combining classifier training patterns across 3 talkers. As talkers were added, neural discriminability decreased suggesting that the acoustic features encoded by the smoothed spatio-temporal response patterns were in part specific to idiosyncrasies of each individual utterance and did not generalise across variations with consonant

classes. The range of the optimal smoothing windows, however, remained much the same. This could be viewed as an indication of an intrinsic optimal encoding window that is efficacious for the representation of generic spectro-temporal dynamics of consonant sounds. However, if cannot be ruled out that the classifier is predominantly performing token recognition as other responses to the same token that elicited a given test pattern was incorporated into the stimulus set. One way to investigate to what extent the former is true would be to produce a stimulus set that included a large number of talkers and perform similar analyses, however due to the difficulties in maintaining stable electrophysiological recordings over long epochs (due to fluctuations in the effects of anaesthesia or tissue movements around the electrode), this may be at the expense of a large set of phonemes and could not easily be extended to incorporate vocoded counterparts.

The stimulus set was parametrically degraded using a noise vocoder. For the single taker paradigm, the neural discriminability of the set of 16 VCVs was quantitatively similar to human performance for single channel vocoded speech where a large population of MU recording sites were used. The classifier was unable to use a small number of SU recordings to discriminate the stimulus sets. Classifier performance over-predicted human performance where only 2 spectral channels of information were available and reached near perfect performance for 4 spectral channels. This over-prediction has also been demonstrated at more peripheral nuclei (see Chapters 1 and 2) and this chapter provides further evidence that behavioural discrimination of speech tokens is not limited by primary sensory representations. Surprisingly, where the

number of spectral channels is doubled to 8, there is a slight decrease in classifier performance, which is more pronounced in the multi-talker paradigm. This is a phenomenon that is not observed perceptually, however the stimulus set used in human psychophysical study that was emulated in the present investigation did not include this 8 channel condition. Studies using noise vocoders, however, show quantitative differences attributable in part to the selection of vocoder parameters (Whitmal III et al., 2007). It may be that the comparatively restricted overall bandwidth used by Shannon et al. (1995) results in detrimental random fluctuations in the noise carriers when the spectral channels are further bisected.

Neural representations of the vocoded speech sounds were similarly parametrically temporally degraded. Despite envelope fluctuations above 16 Hz having been attenuated by the envelope extraction filter, discriminability reaches optimal values across the vocoding conditions when smoothing corresponding to integration windows 10 ms is used, which is also the case for natural speech. This is a robust finding regardless of the number of spectral channels, or whether a single or multi-talker paradigm is used. Response patterns that incorporated spike timing on finer time scales do not become increasingly discriminable when high rate envelope cues are not attenuated. At more peripheral nuclei, the inclusion of high rate envelope cues makes the neural representations of consonants more dissimilar (see Chapters 2 and 3), which does not appear to be the case for cortical representations. This suggests that, whilst an internal representation of high rate temporal features must exist

in order to elicit percepts of timbre, specific features that could be used to differentiate speech tokens are not directly represented.

Chapter 5
General discussion

5.1 Novel aspects of the experimental paradigm and neural decoding

One of the fundamental aims of auditory neuroscience is to explore the link between spatio-temporal patterns of neural activity and behavioural observations and thus to elucidate a 'neural code' for speech; the aspects of the spatio-temporal pattern of neural activity that underlie perceptual phenomena. Early studies focussed on peripheral representations of physical attributes of speech stimuli. For example, Sachs and Young (1979) demonstrated that the positions of formant frequencies of the vowel sounds /ɪ, ɛ and *ɑ:*/ were discernible when the average spike rate elicited by each stimulus was plotted against the pure tone CF of a large population of auditory nerve fibres, at least for low stimulus intensities. Other studies considered the temporal pattern of spikes and found that peripheral neurons tend to phase-lock to formant frequencies close to their CF (Young and Sachs, 1979; Palmer et al., 1986) constituting and alternative 'code' for vowel sounds.

A number problems preclude using such approaches to understand how auditory neural processing underlies speech recognition in general, however. One of the foremost of these is that assumptions have to be made regarding the salient physical properties of the stimulus. For vowel sounds, perceptual studies show that the relative location of formant frequencies is the most salient cue for vowel discrimination (Klatt, 1982). Consonants, however, present a much more diverse array sounds, many of which are inherently dynamic and thus characterisation of salient acoustic parameters presents a more challenging problem. Further, such physical attributes of the stimulus are far from stable across multiple talkers, particularly between those of different genders, for example. Temporal attributes could also be heavily affected by reverberation. Despite this variability, human listeners are still able to correctly classify phonemes or syllables even where top-down linguistic cues are not available.

In perceptual studies relating to human perception of speech, it is common practice to manipulate acoustic stimuli, reducing spectral and temporal content in order to investigate which aspects of the speech signal are required for speech understanding (e.g. Drullman et al., 1994; Shannon et al., 1995; Friesen et al., 2001; Xu et al., 2005; Stone et al., 2008). Such studies have made significant advances towards elucidating the relative importance of spectral and temporal cues for speech recognition and, along with investigations into the efficacy of cochlear implants (e.g.Fishman et al., 1997; Nie et al., 2006), have revealed the remarkable robustness of the auditory system to severe degradation of the input signal.

The tools developed in such psychophysical studies provide a means to parametrically degrade speech signals by removing various spectral and temporal cues. A vocoder in particular is often used as the nature of these degradations is considered to be analogous to the nature of degradations inherent in contemporary auditory neural prosthetics, the most common of which is the cochlear implant (although the processing strategies used in less established central neural prostheses, such as auditory brainstem or midbrain implants, are currently very similar).

This body of literature, then, constitutes a powerful tool for examining which aspects of the spatio-temporal patterns of neural activity in various brain regions underlie the perception of speech; The fact that the salient physical attributes of a diverse set of natural speech sounds is unknown can be overcome by firstly parametrically simplifying them using a vocoder and secondly by focussing on the discriminability of their representations rather than on explicit representations of acoustic features.

This approach presented a number of challenges, however. Firstly, it was desirable to emulate the stimulus set from an existing study such that the perceptual effects of the various signal degradations were already known. Secondly, the discriminability of the neural representations needed to be quantified in such a way that it could be directly related to the psychophysical data. The chosen study utilised medial consonants produced by multiple talkers. This has the advantage of having a greater degree of ecological validity compared to those using speech tokens produced by a single talker; if there

was only a single exemplar of each token, it is plausible that a much greater array of acoustic cues could be used to discriminate the speech tokens than those that are useful for phoneme classification. Further, phonemes do not generally occur in isolation in connected speech and are often preceded by other sounds. It is feasible that target sounds that are preceded by a period of silence that is long compared to those between words could be represented in a different way to those that are not. For example, they may elicit responses by populations of neurons in central brain regions that are particularly sensitive to sounds onset, which would respond in a different way if the target sounds was preceded by others. It has also been demonstrated that spike timing is more precisely locked to the first of repetitive acoustic events such as acoustic pulse trains (Wang et al., 2008), suggesting that the importance of spike timing changes depending on context.

The increased ecological validity of utilising multiple exemplars of each speech token as well as medial target sounds presented a departure from existing studies that utilised comparable nearest neighbour decoding approaches. A key issue is one of how representations could be aligned for comparison. Initial investigations referenced spike times to a computer clock, however this constitutes an external reference frame that is not available to the brain. One idea is to reference spike times from some population response or LFP (Chase and Young, 2007; Perez et al., 2013; Panzeri et al., 2014), however this approach clearly has limitations where the target sound is preceded by another of variable length, as it is in this study.

Such issues could potentially be overcome implicitly by more sophisticated methods of classification, such as artificial neural networks, for example. However due to the high dimensionality of the data set, the training set would need to be prohibitively large for current electrophysiological techniques in order to avoid over-fitting (Quiroga and Panzeri, 2009). A compromise, then, was to introduce a relative temporal shifting mechanic into the classifier, which produced templates averaged across each of the talkers where the responses to each token were shifted such that the distances between them were minimised. Likewise, test patterns were allowed to shift temporally relative to the training set. This mechanic is clearly not representative of any biophysical mechanism, but is a pragmatic way to deal with the problem of defining appropriate response windows for target stimuli that vary in duration and position relative to absolute stimulus onset. The auditory system has to deal with a much greater variability in everyday speech such as that in speech rate, but how this might be implemented is yet to be explored.

5.2 The relative importance of spectral and temporal cues for neural discriminability

The relative importance of spectral and temporal information in speech was explored by measuring neural responses to vocoded speech whereby the envelope bandwidth and numbers of spectral channels were parametrically varied. The discriminability of responses to 16 VCVs produced by 3 male talkers was quantified using the nearest neighbour classifier described in detail in 2.1.3 and discussed above. The data corresponding to all vocoding conditions in

which the envelope extraction filter was set to 16 or 500 Hz for each brain region are reproduced in Figure 5.1. The data presented here correspond to the multi-talker paradigm, analogous to the emulated human behavioural task (Shannon et al., 1995), the results of which are reproduced on the same axes.

The first thing to note is that, for auditory nerve and IC representations, there is a significant increase in discriminability where the envelope extraction filter was set to 500 Hz compared to when it is only 16 Hz, regardless of the number of spectral channels. The same cannot be said for representations in the auditory cortex, where non-attenuated high rate envelope cues do not significantly increase the discriminability of either SU or MU representations. This observation could be related to the well-established observation that the frequency of peaks in tMTFs decrease throughout the ascending auditory system, with best modulation frequencies typically of the order of tens of Hertz in the auditory cortex. In more peripheral regions such as the IC and auditory nerve these extend to hundreds or even thousands (see 1.1; Joris et al., 2004). However, there is evidence that high rate temporal information is transformed by the level of the cortex and represented instead in the firing rates of neurons over longer timescales (Wang et al., 2008). The presence of sustained firing rates required to implement such codes, however, appears to be contingent on the absence of anaesthesia (e.g. Wang et al., 2005), which suggests that high rate envelopes may not be represented effectively in this experiment.

For 16 Hz envelope speech (represented by the lighter coloured lines), increasing the number of spectral channels from 1 to 2 and then to 4 increases

the discriminability of the neural representations in each of the brain regions, which is in general agreement with other studies that have investigated the effects of reduced spectral cues on neural discriminability of conspecific vocalisations in rodents (Ter-Mikaelian et al., 2013) as well as in human speech (Ranasinghe et al., 2012). In peripheral regions, the frequency sensitivity of the auditory system underlies this phenomenon, enabling separate populations to encode envelope modulations in each of the spectral regions independently. The increase is most pronounced in the auditory cortex where the number of spectral channels is increased from 1 to 2. Surprisingly, in each of the brain regions, there is no corresponding increase in discriminability when the number of vocoder channels is increased from 4 to 8; the discriminability of cortical representations actually decreases in this case. This was an unexpected result, the putative causes of which are discussed below.

In the auditory nerve, the lack of increase in discriminability between 4 and 8 channels may be a result of undersampling; as the number of channels increases, there is a corresponding decrease in the number of nerve fibres predominantly responding to energy within each spectral channel. Due to the stochastic nature of spike generation the auditory nerve, this may lead to stochastic undersampling of envelope features within each channel which is exacerbated as the number of channels increases, thus offsetting the potential benefit of greater spectral acuity. Such undersampling has been proposed to underlie speech processing deficits that are not accompanied by corresponding auditory threshold shifts (Lopez-Poveda and Barrios, 2013) although this phenomenon has been linked in particular to a degraded representation of

high-rate or low intensity envelope cues as well as difficulties recognising speech in the presence of maskers.

A second factor in the auditory nerve is the attenuating effect of the outer-middle ear filters of the model on the lower channels of the vocoder. The frequency response of these filters was shown previously in Figure 2.7, and it can be seen that the lower channels are heavily attenuated by these filters. The neural classifier has no normalisation mechanism, so temporal shifting and distance measures are driven by the sites that generate the most spikes, so these lower channels are dominated by the higher ones and cannot contribute effectively to classification. It may be that some central gain mechanism is able to account for this, or perhaps fibres with different physiological properties resulting in different dynamic ranges assume differential importance in various spectral regions. For example, for conversational speech perhaps high spontaneous rate fibres exhibiting lower thresholds could be utilised in these lower frequency, attenuated channels.

A similar argument could be made for the IC. Pure tone tuning characteristics are shown in Figure 3.6, which demonstrates an under-representation of neural sites optimally tuned to frequencies corresponding to the lower channels. A similar plot is shown Figure 4.3 for the cortical recordings. However, fewer than half of the cortical sites exhibited pure tone tuning, and so many of the responses could not be characterised in this way. There is evidence that cortical neurons are selective to higher order stimulus features such as frequency sweeps or repetition rates (Mendelson and Cynader, 1985; Wang et al., 2005)

or even complex sequences of spectral and temporal modulations such as those in birdsong in the avian auditory forebrain (Woolley et al., 2005). Such selectivity leads to a 'sparse' representation of complex sounds in which only small populations of neurons are active at any one time (Hromádka et al., 2008). This again could underlie an undersampling issue, but one of a different nature to that in the auditory nerve; it could be that the stimulus set utilised in this experiment contained spectro-temporal modulations that were optimal for driving populations of cortical neurons that were not recorded in this study, perhaps because a particular region of the primary cortex was targeted based solely on pure-tone response characteristics.

In summary, the relationship between spectral and temporal cues and neural discriminability appears to depend on the brain region. In the auditory nerve and the inferior colliculus, the representation of high rate envelope cues makes average representations of phonemes more distinct at the timescale associated with optimal discriminability. Spectral cues also serve to increase the discriminability of neural representations, but to a lesser extent. Put another way, reducing the spectral complexity of the signal has less effect on the discriminability of peripheral representations of speech than reducing temporal complexity, consistent with previous studies (Loebach and Wickesberg, 2006). This remains the case even where average responses are combined across multiple talkers. In the auditory cortex, high rate envelope cues do not increase the discriminability of the representations and instead a greater weight is put on the spectral complexity of the signal. This may indicate that stimulus features associated with high rate envelope cues such as timbre are re-encoded

in the auditory cortex perhaps on separate timescales such that they are separate from features that occur over the longer salient timescales for communication vocalisations. This could explain why the classifier does not appear to benefit from such cues in the single channel case as, despite having the potential to indicate the presence of voicing, for example, this feature is represented using a neural code over a different timescale to that which is optimal for consonant discrimination in general.

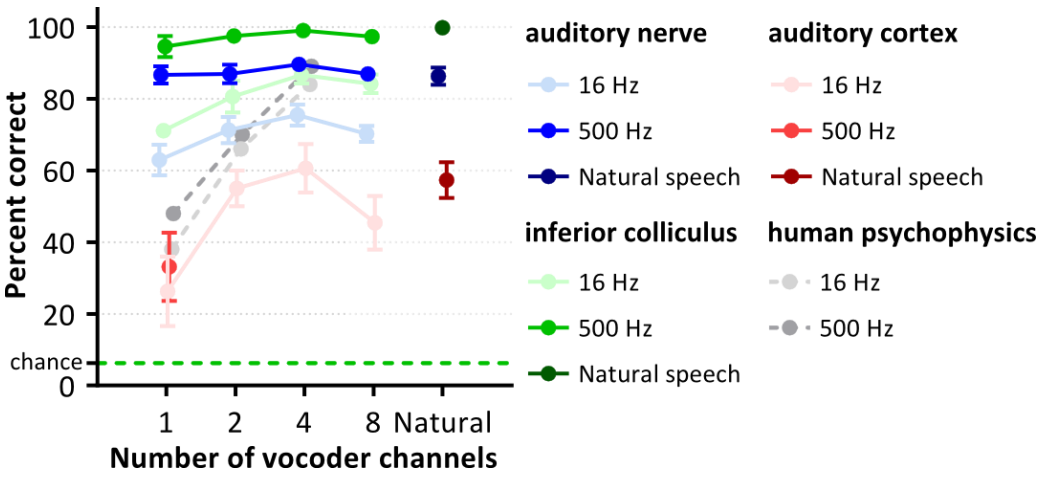

Figure 5.1: The discriminability of neural representations of VCVs produced by 3 male talkers in the auditory nerve, inferior colliculus and primary auditory cortex of anaesthetised guinea pigs. Human data reproduced from Shannon et al. (1995). Auditory nerve representations comprised 100 simulated fibres. IC and cortical representations formed of 106 and 198 MU recording sites respectively.

5.3 The effect of carrier on neural discriminability

One of the potential reasons for the inability of the classifier to benefit from spectral cues in auditory nerve representations alluded to previously was the increase in intensity of intrinsic random envelope fluctuations in the noise carrier as the channel bandwidths decrease. In order to test the effect of these envelope fluctuations, the noise carriers were replaced with tones, which have intrinsically flat envelopes. The frequency of each tone bisected each analysis band on a logarithmic scale. Other than this, the simulation of auditory nerve responses and neural classification were carried out exactly as described in 2.1. The result of this change in carrier is shown in Figure 5.2.

A very similar pattern of results can be seen; there is a clear separation between performances for the vocoding conditions where the envelope extraction filter cutoff was 500 Hz and where it was 16 Hz. Classifier performance as a function of the number of vocoder channels is relatively flat compared to human performance. Clearly, the inability of the classifier to make use of the additional spectral cues is not due to the deleterious effects of intrinsic fluctuations of the noise carrier. The main cause of this effect, then, may be undersampling within each channel. However, since the responses from individual nerve fibres are treated independently by the classifier it is unlikely that simply adding more fibres could alleviate this and that integration across multiple fibres with the same CF is required. Further experiments could investigate the effects of within or across channel integration by forming the rows of the input neurograms by producing average PSTHs from responses to single presentations combined across multiple fibres.

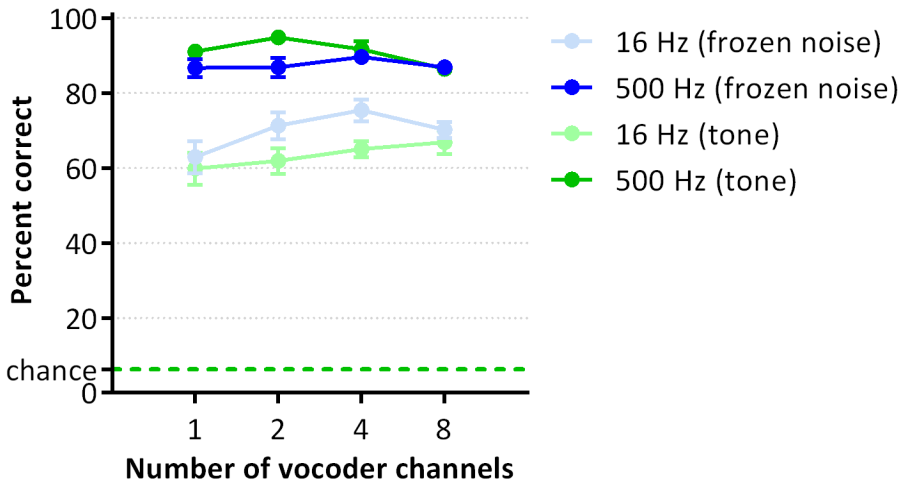

Figure 5.2: The effect of carrier type on the discriminability of neural representations of vocoded speech in the simulated auditory nerve fibres. Representations formed of 100 nerve fibres logarithmically spaced across frequency.

Results from perceptual experiments using tone and noise vocoders are inconsistent, with some investigators reporting very little difference between the two vocoder types across a range of speech discrimination tasks (Dorman et al., 1997) and others reporting a marked improvement in speech recognition in quiet by human listeners for a tone over a noise vocoder (Whitmal III et al., 2007). To what extent these discrepancies can be attributed to differences in other vocoder parameters, such as envelope bandwidth, training or various cognitive factors remains unclear.

A similar process was carried out for a stimulus set in which separate noise carriers were used for each token. In this case, classifier performance reached ceiling for every vocoding condition provided that temporal smoothing was of the order of a few milliseconds. This is presumably indicative that in this

condition the high rate envelope fluctuations and temporal fine structure of the noise carriers themselves provide a viable cue for discriminating between the tokens, and highlights the importance of choice of carrier type and methods used for stimulus generation in these kinds of experiments.

5.4 The relationship between neural discriminability and perception

The human psychophysical data reproduced from Shannon et al. (1995) in Figure 5.1 utilised an analogous stimulus set in a 16 alternative forced choice paradigm. For single channel vocoded speech, the discriminability of representations in the auditory nerve and the inferior colliculus over-predict the ability of human listeners to correctly identify the correct phoneme regardless of the envelope bandwidth. This is the case even where the envelope extraction filter is set to 16 Hz. It is often assumed in perceptual studies that the limiting factor in the ability of human listeners to correctly identify degraded speech sounds is the information content in the input signal. However, this shows that the peripheral representations of even the most heavily degraded speech sounds are highly discriminable by a simple nearest neighbour classifier even when average representations are combined across multiple talkers. For a single talker task, these representations are almost perfectly discriminable by such a classifier in both the auditory nerve (see Figure 2.11) and the inferior colliculus (see Figure 3.16) given an appropriate choice of encoding window.

Where the envelope extraction filter cutoff is increased to 500 Hz, the discriminability of neural representations in the auditory nerve and the inferior colliculus is increased significantly by approximately 10-15%. While high rate envelope cues invariably improve performance in human listeners in a consonant discrimination task (Van Tasell et al., 1992; Shannon et al., 1995; Stone et al., 2008; Souza and Rosen, 2009), the effect is invariably comparatively small. In light of the present results, the inability of human listeners to use such cues seems surprising, particularly when such cues carry information about consonant voicing; the presence or absence of quasi-periodic fluctuations that could aid the distinction between voiced and unvoiced phonemes with otherwise similar envelope characteristics such as */z/* and */s/*, particularly where rich spectral cues are not available. However, stimulus features over such timescales are unlikely to generalise well across more extensive environmental variability such as talker identity or reverberation, for example. With a small number of talkers of one gender, the classifier is able to explicitly use peripheral representations of these cues, particularly as other presentations of an identical token form part of the training set, whereas human listeners appear to systematically disregard them for the purpose of phoneme identification.

Discrepancies between the neural discriminability and human performance may be in part due to the different nature of the task performed by the neural classifier and human listeners. Often in human psychophysical experiments, listeners are provided with a degree of training for a given vocoding condition. This training process invariably involves a process of mapping the new,

degraded speech tokens to a previously learned phonology of natural speech. During testing, the listeners are required to respond by selecting which phoneme a given degraded token corresponds to. Conceptually, this process is different to the operation of the classifier, which does not attempt to map the degraded token to their non-degraded counterparts but rather produces new internal classes of each phoneme for each of the vocoding conditions. This would be more analogous to presenting human listeners with the vocoded tokens without providing the information that they correspond to particular speech sounds and instead asking listeners to designate classes using abstract labels, for example. This type of process may well be more informative if the question relates to which cues enable the formation of robust internal representations of phonemic classes, which is particularly relevant for pre-lingual cochlear implantation where no learned phonology exists. This mapping process may indeed be one of the major contributors to the high degree of inter-subject variability reported in vocoder studies and potentially conceptually one of the major "cognitive factors" referred to by Van Tasell et al. (1992) underlying such phenomena.

Another potential source of discrepancy between results in the studies presented here and the human psychophysics literature could be inter-species differences. Clearly the discriminability of neural representations does not benefit to the same extent as human listeners from increasing the number of vocoder channels. Arguably for auditory nerve and IC datasets, discriminability as measured by the neural classifier was substantially higher than that measured in humans even for single channel vocoded speech, so there was less

room for improvement. However a similar effect could be expected to be observed if the model species had substantially broader cochlear filters such that the individual frequency channels in the vocoder were not sufficiently resolved. Whilst estimates of the frequency resolution of the peripheral auditory system in the guinea pig indicate broader auditory filters than in the human, this is unlikely to be to such an extent that it could cause such an effect as even the narrowest vocoder frequency bands used in this study should be resolved (see 1.7). There could also be differences in temporal processing across the two species. For example, phase-locking to fine structure components extends to 2-3 kHz in the guinea pig, but this limit is unknown in humans. It is often assumed to be higher or at least a similar value, however (see Joris and Verschooten, 2013).

5.5 Evolution of neural representations of speech in the auditory system

The approach taken in this series of experiments was to systematically manipulate which aspects of the neural representation were used as inputs to a neural classifier. Two manipulations were used; the number of sites of neural activity, corresponding to a manipulation of spatial sampling, and the temporal resolution of the neural responses. This second manipulation assumes that a downstream process that may form categorical representations of speech sounds is able to perform such operations on these sensory representations.

The optimal timescales for discriminability of the 16 VCVs used throughout this thesis in each of the 3 brain regions are shown in the upper plots of Figure 5.3.

Figure 5.3 A shows the effect of temporal smoothing on representations of each of these VCVs where the training set comprised average responses to each token produced by a single talker. In the auditory nerve and the inferior colliculus, representations on timescales ranging from a few milliseconds to hundreds of milliseconds form robust response classes corresponding to each individual token. In fact, for auditory nerve representations the ensemble PSTH is sufficient for perfect discrimination (see Figure 2.5). This indicates a highly redundant representation in these peripheral brain regions for performing discriminability of a closed set of speech sounds. By the auditory cortex, though, longer smoothing windows are required to perform optimal discrimination indicative that neural activity at finer timescales constitutes noise in the sense that it does not represent features of the stimulus specific to any token. As discussed previously, this result may seem contrary to published studies examining the neural code for consonant sounds (Engineer et al., 2008; Ranasinghe et al., 2012; Perez et al., 2013), which suggest that millisecond precise timing plays a key role in cortical representations of consonants. Indeed, it seems that in some auditory structures, information about speech stimuli is carried in the sub-millisecond precise timing of spikes (Garcia-Lazaro et al., 2013). However, it is important to draw a distinction between neural codes that carry most information about physical attributes of individual stimuli, those that result in correlations with behavioural performance either in the confusions generated (Engineer et al., 2008) or as a function of signal degradation (Ranasinghe et al., 2012) and those that are optimal for discriminability.

For the multi-talker paradigm, temporal smoothing has a dramatically different effect on representations in the auditory nerve and the inferior colliculus. Temporal degradation of the spatio-temporal activity pattern in the auditory nerve with increasing smoothing window lengths causes a corresponding monotonic decrease in discriminability. The spike rate profile, which was sufficient to perfectly discriminate speech tokens cannot be used by the classifier as effectively to form robust classes for each of the phonemes. This implies that whilst the number of spikes elicited by each token is relatively robust across multiple presentations, the differences between responses to exemplars of each phoneme produced by each of the talker are of comparable magnitude to those between the different phonemes. Millisecond-precise representations in the IC do not form intrinsic response classes for this set of speech sounds and discriminability peaks where smoothing windows of approximately 10 ms are used.

Interestingly, the discriminability of representations in the IC is far more robust to severe temporal degradation of the spike trains that those in the auditory nerve. This could imply a re-encoding of cues that were previously represented by the precise timing of spikes in the auditory nerve into spike rates. This could correspond to the emergence of rate-codes for amplitude modulation (Langner and Schreiner, 1988; Lorenzi et al., 1995; Joris et al., 2004). However, it is difficult to make a direct comparison since the number of neurons differs between the two datasets (106 MU sites vs 100 individual fibres).

In the auditory cortex, discriminability as a function of smoothing window duration has a very similar shape for the multi-talker paradigm as it does for the single talker paradigm, albeit with a lower peak. This is tentatively indicative of an intrinsic optimal timescale of the neural code for consonant-like dynamic stimuli being of the order of tens of milliseconds. However, since prior representations of the test stimulus formed part of the training set it cannot be ruled out that recognition is driven by representations of acoustic features corresponding to an individual token, which is essentially masked by combination across the multiple talkers. A much larger set of talkers would be needed to successfully train a classifier to recognise novel tokens successfully, although there is evidence that representations of these timescales do form response classes that are robust across large numbers of talkers (Mesgarani et al., 2008).

One of the limitations of the classifier is that it assumes the same encoding windows are optimal for every recording site. Analyses of individual recording sites reveals that optimal smoothing windows differ even between units with similar CFs in the IC and in the cortex (see Figure 3.12; Figure 4.6). Further, there is evidence that various stimulus attributes are represented optimally over different timescales (Panzeri et al., 2010). For example, an important cue for distinguishing between voiced and unvoiced consonants is the presence or absence of quasi-periodic amplitude modulations; a feature which could be extracted and represented explicitly but perhaps over timescales other than those optimal for consonant discrimination in general. In perceptual studies on speech recognition, it is common practice to perform information transmission

analysis (see 2.1.4) on confusion matrices by classifying the consonant sounds by particular phonetic features such as the place or manner of articulation. The same classifier used in this study could be trained and tested to find optimal encoding windows for various acoustic features.

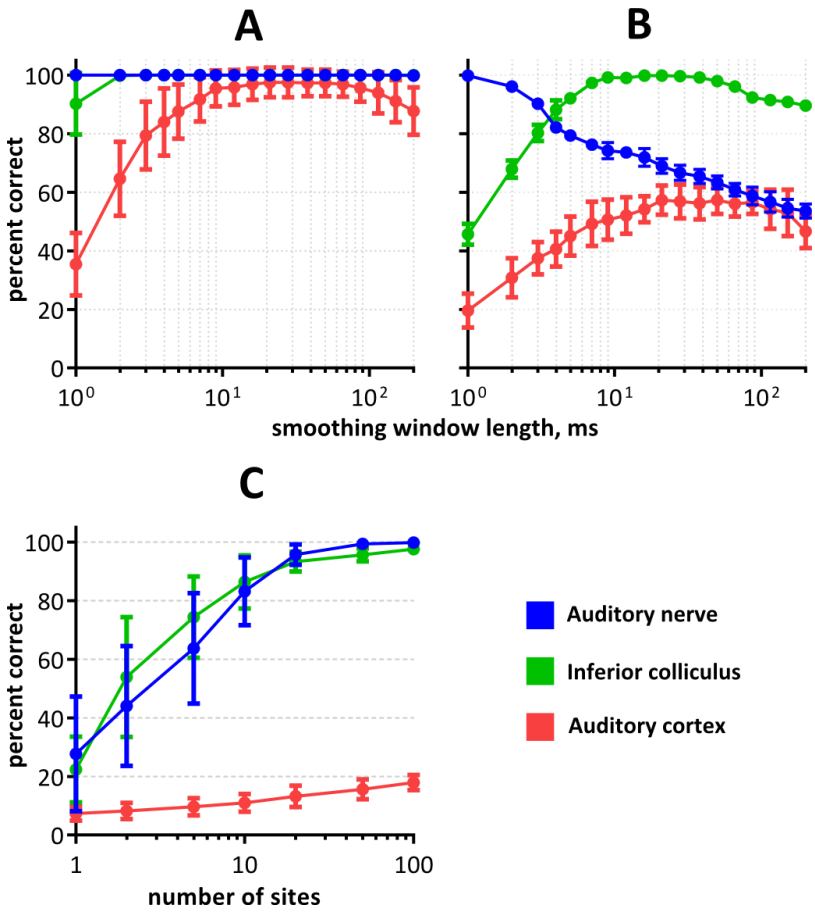

Figure 5.3: The effect of temporal smoothing on the discriminability of representations of natural speech in AN, IC and A1 for single- (**A**) and multi-talker (**B**) classifier training paradigms. Representations comprise 100 AN fibres, 106 IC MU sites and 196 A1 MU sites. The effect of number of sites on the discriminability of neural representations using the optimal smoothing windows for each brain region (**C**).

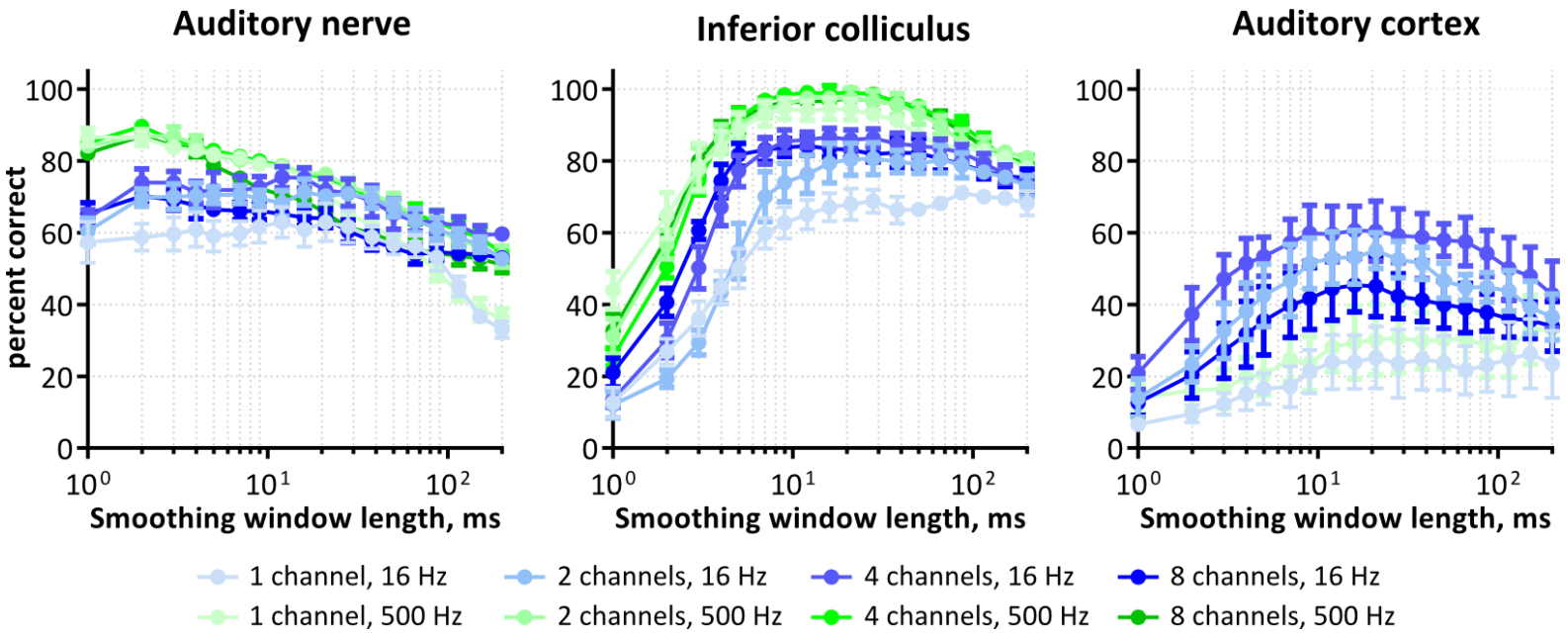

Figure 5.4: The effect of smoothing of discriminability of neural representations of 16 VCVs in the auditory nerve (100 simulated fibres), inferior colliculus (106 MU site) and auditory cortex (196 MU sites).

The effect of vocoding on the optimal timescale of the neural code for discriminable responses is shown in Figure 5.4. In the auditory nerve, the benefit of high rate envelope cues clearly corresponds to very short smoothing windows lengths. Optimal discriminability occurs where 2 ms smoothing is used, corresponding to the upper cutoff of the envelope extraction filter suggesting that these high rate envelope cues are encoded by spike timing at timescales shorter than or similar to the inverse of the modulation frequencies. This benefit disappears for temporal integration windows longer than approximately 10 ms after which discriminability is very similar for the envelope extraction filter cutoffs. Discriminability in all conditions is highly robust to temporal smoothing. This is likely to indicate the use of overall level intensity differences between the stimuli being useful for voiced vs un-voiced consonants, for example. This is supported by the observation that this robustness is decreased for the single channel condition.

The picture is very different in the IC as these high rate envelope cues increase the discriminability of the neural representation but temporal degradation of the input stimulus does not correspond to longer optimal smoothing windows. This suggests that modulation information at high rates has been re-encoded into some other form. An example could be a rate-code in which acoustic phenomena over timescales of the order of a few milliseconds are uniquely represented by the intensity of neural activity over equivalent timescales or longer, analogous to the emergence of bandpass rMTFs in the IC (Joris et al., 2004). Further, where high rate envelope modulations are attenuated, IC representations nevertheless achieve optimal discriminability where

smoothing of the order of a few milliseconds is used. Likewise, in the auditory cortex, where high rate envelope cues have been attenuated, discriminability is optimised where encoding windows shorter than the timescales associated with the predominant dynamic changes in the stimulus are used constituting a temporal representation in the formal sense (Theunissen and Miller, 1995), although not at a millisecond precise timescale.

The effect of spatial sampling on the discriminability of neural representations is shown in Figure 5.3 C. This clearly demonstrates that, for natural speech, representations in the auditory nerve and the IC are highly redundant with only a few sites required to achieve perfect discriminability of consonants. The result is very different in the auditory cortex, in which individual or small groups of sites perform very poorly. Even where almost twice the number of recording sites is available in the auditory cortex compared to the inferior colliculus (196 vs. 106), performance is still much lower. In isolation this might suggest that the cortical neural population does not effectively represent stimulus features, however perfect discrimination is possible in the single talker paradigm (see Figure 4.8). This is supportive of a cortical representation which is highly selective in which each individual neuron represents a particular stimulus feature. Such a high dimensional representation could prove advantageous for classification purposes, but only in the relevant feature space is covered by the recorded neural population. It may be that in order to reliably discriminate between consonant classes, a much larger neural population is required than that to discriminate between tokens, for which a much reduced subset of

acoustic features may be sufficient. It cannot be ruled out, however, that anaesthesia has a significant effect on the discriminability of cortical responses.

5.6 Summary

This thesis constitutes an exploratory approach to investigating some of the neural mechanisms underlying the robust nature of speech recognition by human listeners, and to what extent speech recognition is limited by information available in the stimulus or by neural processing mechanisms. It is only very recently that similar approaches have been used to investigate the neural codes used to represent complex sounds and this thesis has highlighted that any conclusions drawn are highly dependent on the specific paradigm used. For example, are the target sounds at the stimulus onset, preceded by a comparatively long period of silence? It is not unreasonable to expect a different representation to ones that are not, due in part to the recruitment of populations of neurons that specifically respond to stimulus onset but only rarely or highly selectively to on-going stimuli.

Another factor that must be considered in future investigations of a similar nature is the distinction between token recognition and consonant recognition. Often these two are interpreted in much the same way either in psychophysical or physiological studies, and yet the task posed to the auditory system is conceptually very different. This difference could mitigated in future behavioural studies by the way the task is posed to human listeners, which is often to generate a mapping between degraded versions of speech tokens and internal representations of phonemic classes as opposed to learning to utilise

reduced available cues to discriminate between classes of the altered sounds. The limitations of this mapping process could in part underlie the inability of human listeners to capitalise on the availability of high rate envelope cues in a speech recognition task involving a close set of consonants, when they demonstrably increase the distinctiveness of sensory representations.

The central findings in this thesis are:

- A set of 16 medial consonants elicit discriminable patterns of neural activity in simulated auditory nerve fibres and the inferior colliculus and auditory cortex of anaesthetised guinea pigs. The discriminability of representations in the auditory nerve and the IC remains for a multiplicity of putative neural codes, and is optimal in the auditory cortex where integration windows of 10 – 100 ms are used.
- The timescale of the neural code optimal for producing discriminable response classes in the auditory nerve depends on the rates of salient features in the stimulus. This is not so apparent in the auditory midbrain in which high rate envelope cues increase the discriminability of responses over longer timescales. High rate envelope cues do not appear to increase the discriminability of cortical responses, although further work is required to disambiguate if this is due to them not being represented at all or if they are re-encoded in some other way.
- The timescale of the neural code for optimal discrimination of dynamic speech sounds appears to depend on the context, with much longer integration windows required to discriminate cortical representations

of medial consonants than those reported for word-initial consonants. A future experiment could utilise consonant-vowel-consonant-vowel phoneme sequences to efficiently examine the effects of vowel context.

References

Agus TR, Pressnitzer D (2013) The detection of repetitions in noise before and after perceptual learning. The Journal of the Acoustical Society of America 134:464-473.

Ahrens MB, Linden JF, Sahani M (2008) Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. The Journal of Neuroscience 28:1929-1942.

Aitkin LM, Anderson DJ, Brugge JF (1970) Tonotopic organization and discharge characteristics of single neurons in nuclei of the lateral lemniscus of the cat. j Neurophysiol 33:l-440.

Aitkin LM, Phillips SC (1984) Is the inferior colliculus and obligatory relay in the cat auditory system? Neuroscience letters 44:259-264.

Anderson L, Wallace M, Palmer A (2007) Identification of subdivisions in the medial geniculate body of the guinea pig. Hearing research 228:156-167.

Augustine G, Charlton M, Smith S (1985) Calcium entry and transmitter release at voltage-clamped nerve terminals of squid. The Journal of physiology 367:163-181.

Békésy GV (1947) The Variation of Phase Along the Basilar Membrane with Sinusoidal Vibrations. Acoustical Society of America Journal 19:452.

Bladon A, Fant G (1978) A two-formant model and the cardinal vowels. Speech Transmission Laboratory Quarterly Progress and Status Report 19:1-8.

Bradbury JW, Vehrencamp SL (1998) Principles of animal communication.

Bullock D, Palmer A, Rees A (1988) Compact and easy-to-use tungsten-in-glass microelectrode manufacturing workstation. Medical and Biological Engineering and Computing 26:669-672.

Calford MB (1983) The parcellation of the medial geniculate body of the cat defined by the auditory response properties of single units. The Journal of neuroscience 3:2350-2364.

Calford MB, Aitkin LM (1983) Ascending projections to the medial geniculate body of the cat: evidence for multiple, parallel auditory pathways through thalamus. The Journal of neuroscience 3:2365-2380.

Centanni T, Sloan A, Reed A, Engineer C, Rennaker II R, Kilgard M (2014) Detection and identification of speech sounds using cortical activity patterns. Neuroscience 258:292-306.

Centanni TM, Engineer CT, Kilgard MP (2013) Cortical speech-evoked response patterns in multiple auditory fields are correlated with behavioral discrimination ability. Journal of neurophysiology 110:177-189.

Chase SM, Young ED (2007) First-spike latency information in single neurons increases when referenced to population onset. Proceedings of the National Academy of Sciences 104:5175-5180.

Christianson GB, Sahani M, Linden JF (2008) The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. The Journal of Neuroscience 28:446-455.

Chung DY, Colavita FB (1976) Periodicity pitch perception and its upper frequency limit in cats. Perception & Psychophysics 20:433-437.

De Boer E, De Jongh H (1978) On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. The Journal of the Acoustical Society of America 63:115-135.

Delgutte B (1980) Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. The Journal of the Acoustical Society of America 68:843-857.

Delgutte B, Kiang NY (1984) Speech coding in the auditory nerve: I. Vowel-like sounds. The Journal of the Acoustical Society of America 75:866-878.

Djourno A, Eyriès C (1957) Prothese auditive par excitation electrique a distance du nerf sensoriel a laide dun bobinage inclus a demeure. Presse médicale 65:1417-1417.

Dorman MF, Loizou PC, Rainey D (1997) Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. The Journal of the Acoustical Society of America 102:2403.

Dorman MF, Spahr AJ (2006) Speech perception by adults with multichannel cochlear implants. Cochlear implants 2nd ed New York (NY): Thieme 193-204.

Drullman R, Festen JM, Plomp R (1994) Effect of temporal envelope smearing on speech reception. The Journal of the Acoustical Society of America 95:1053.

Egorova M, Ehret G, Vartanian I, Esser K-H (2001) Frequency response areas of neurons in the mouse inferior colliculus. I. Threshold and tuning characteristics. Experimental brain research 140:145-161.

Engineer CT, Perez CA, Chen YH, Carraway RS, Reed AC, Shetake JA, Jakkamsetti V, Chang KQ, Kilgard MP (2008) Cortical activity patterns predict speech discrimination ability. Nature neuroscience 11:603-608.

Escabí MA, Schreiner CE (2002) Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. The Journal of neuroscience 22:4114-4131.

Evans E (1972) The frequency response and other properties of single fibres in the guinea-pig cochlear nerve. The Journal of physiology 226:263-287.

Evans E, Pratt S, Spenner H, Cooper N (1992) Comparisons of physiological and behavioural properties: Auditory frequency selectivity. Auditory physiology and perception 83:159-169.

Evans EF (1992) Auditory processing of complex sounds: an overview. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 336:295-306.

Fettiplace R, Hackney CM (2006) The sensory and motor roles of auditory hair cells. Nature Reviews Neuroscience 7:19-29.

Fishman KE, Shannon RV, Slattery WH (1997) Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. Journal of Speech, Language and Hearing Research 40:1201.

Friesen LM, Shannon RV, Baskent D, Wang X (2001) Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. The Journal of the Acoustical Society of America 110:1150.

Friesen LM, Shannon RV, Cruz RJ (2005) Effects of stimulation rate on speech recognition with cochlear implants. Audiology and Neurotology 10:169-184.

Garcia-Lazaro JA, Belliveau LA, Lesica NA (2013) Independent Population Coding of Speech with Sub-Millisecond Precision. The Journal of Neuroscience 33:19362-19372.

Grothe B (1994) Interaction of excitation and inhibition in processing of pure tone and amplitude-modulated stimuli in the medial superior olive of the mustached bat. Journal of neurophysiology 71:706-706.

Heffner HE, Heffner RS, Contos C, Ott T (1994) Audiogram of the hooded Norway rat. Hearing research 73:244-247.

Heffner R, Heffner H, Masterton B (1971) Behavioral Measurements of Absolute and Frequency-Difference Thresholds in Guinea Pig. The Journal of the Acoustical Society of America 49:1888-1895.

Heffner R, Koay G, Heffner H (2001) Audiograms of five species of rodents: implications for the evolution of hearing and the perception of pitch. Hearing research 157:138-152.

Henry KS, Heinz MG (2012) Diminished temporal coding with sensorineural hearing loss emerges in background noise. Nature neuroscience 15:1362-1364.

Holden LK, Finley CC, Firszt JB, Holden TA, Brenner C, Potts LG, Gotter BD, Vanderhoof SS, Mispagel K, Heydebrand G (2013) Factors affecting open-set word recognition in adults with cochlear implants. Ear and Hearing 34:342-360.

Holmes SD, Sumner CJ, O'Mard LP, Meddis R (2004) The temporal representation of speech in a nonlinear model of the guinea pig cochlea. The Journal of the Acoustical Society of America 116:3534-3545.

Hromádka T, DeWeese MR, Zador AM (2008) Sparse representation of sounds in the unanesthetized auditory cortex. PLoS biology 6:e16.

Jackson LL, Heffner RS, Heffner HE (1999) Free-field audiogram of the Japanese macaque (Macaca fuscata). The Journal of the Acoustical Society of America 106:3017-3023.

Johnson DH (1980) The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. The Journal of the Acoustical Society of America 68:1115-1122.

Jones E (2007) The thalamus. Cambridge University Press. Cambridge, UK.

Jones EG, Steriade M, McCormick D (1985) The thalamus: Plenum Press New York.

Joris P, Schreiner C, Rees A (2004) Neural processing of amplitude-modulated sounds. Physiological Reviews 84:541-577.

Joris PX, Carney LH, Smith PH, Yin T (1994) Enhancement of neural synchronization in the anteroventral cochlear nucleus I. Responses to tones at the characteristic frequency. Journal of neurophysiology 71:1022-1022.

Joris PX, Verschooten E (2013) On the Limit of Neural Phase Locking to Fine Structure in Humans. In: Basic Aspects of Hearing, pp 101-108: Springer.

Joris PX, Yin TC (1992) Responses to amplitude-modulated tones in the auditory nerve of the cat. The Journal of the Acoustical Society of America 91:215-232.

Kanold PO, Nelken I, Polley DB (2014) Local versus global scales of organization in auditory cortex. Trends in neurosciences 37:502-510.

Kettner RE, Feng J, Brugge JF (1985) Postnatal development of the phase-locked response to low frequency tones of auditory nerve fibers in the cat. The Journal of neuroscience 5:275-283.

Kiang N, Moxon E (1974) Tails of tuning curves of auditory-nerve fibers. The Journal of the Acoustical Society of America 55:620-630.

Kiang NY-S (1965) Discharge Patterns of Single Fibers in the Cat's Auditory Nerve. DTIC Document.

Kidd RC, Weiss TF (1990) Mechanisms that degrade timing information in the cochlea. Hearing research 49:181-207.

Klatt D (1982) Prediction of perceived phonetic distance from critical-band spectra: A first step. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82, vol. 7, pp 1278-1281: IEEE.

Kuwada S, Batra R (1999) Coding of sound envelopes by inhibitory rebound in neurons of the superior olivary complex in the unanesthetized rabbit. The Journal of neuroscience 19:2273-2287.

Langner G, Dinse HR, Godde B (2009) A map of periodicity orthogonal to frequency representation in the cat auditory cortex. Frontiers in integrative neuroscience 3.

Langner G, Schreiner CE (1988) Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. J Neurophysiol 60:1799-1822.

Laver J (1994) Principles of phonetics: Cambridge University Press.

Liberman MC (1978) Auditory-nerve response from cats raised in a low-noise chamber. The Journal of the Acoustical Society of America 63:442-455.

Liberman MC, Kujawa SG (2014) Hot Topics—Hidden hearing loss: Permanent cochlear-nerve degeneration after temporary noise-induced threshold shift. The Journal of the Acoustical Society of America 135:2311-2311.

Liu L-F, Palmer AR, Wallace MN (2006) Phase-locked responses to pure tones in the inferior colliculus. Journal of neurophysiology 95:1926-1935.

Loebach JL, Wickesberg RE (2006) The representation of noise vocoded speech in the auditory nerve of the chinchilla: physiological correlates of the perception of spectrally reduced speech. Hearing research 213:130-144.

Lopez-Poveda EA, Barrios P (2013) Perception of stochastically undersampled sound waveforms: a model of auditory deafferentation. Frontiers in neuroscience 7.

Lorenzi C, Micheyl C, Berthommier F (1995) Neuronal correlates of perceptual amplitude-modulation detection. Hearing research 90:219-227.

Machens CK, Wehr MS, Zador AM (2004) Linearity of cortical receptive fields measured with natural sounds. The Journal of neuroscience 24:1089-1100.

Malmierca MS, Hackett TA (2010) Structural organization of the ascending auditory pathway. The Auditory Brain 9-41.

Meddis R, O'Mard LP, Lopez-Poveda EA (2001) A computational algorithm for computing nonlinear auditory frequency selectivity. The Journal of the Acoustical Society of America 109:2852-2861.

Mendelson J, Cynader M (1985) Sensitivity of cat primary auditory cortex (Al) neurons to the direction and rate of frequency modulation. Brain research 327:331-335.

Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. The Journal of the Acoustical Society of America 123:899-909.

Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some English consonants. The Journal of the Acoustical Society of America 27:338-352.

Moore BC (2003a) Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. Otology & neurotology 24:243-254.

Moore BC (2003b) An introduction to the psychology of hearing: Academic press San Diego.

Nie K, Barco A, Zeng F-G (2006) Spectral and temporal cues in cochlear implant speech perception. Ear and hearing 27:208-217.

Nuttall AL, Dolan DF (1996) Steady-state sinusoidal velocity responses of the basilar membrane in guinea pig. The Journal of the Acoustical Society of America 99:1556-1565.

Oliver DL (2005) Neuronal organization in the inferior colliculus. In: The inferior colliculus, pp 69-114: Springer.

Oxenham AJ (2012) Pitch perception. The Journal of Neuroscience 32:13335-13338.

Oxenham AJ, Shera CA (2003) Estimates of human cochlear tuning at low levels using forward and simultaneous masking. Journal of the Association for Research in Otolaryngology 4:541-554.

Palmer A (1987) Physiology of the cochlear nerve and cochlear nucleus. British medical bulletin 43:838-855.

Palmer A (1990) The representation of the spectra and fundamental frequencies of steady-state single-and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibers. The Journal of the Acoustical Society of America 88:1412-1426.

Palmer A, Russell I (1986) Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. Hearing research 24:1-15.

Palmer A, Shamma S (2004) Physiological representations of speech. In: Speech processing in the auditory system, pp 163-230: Springer.

Palmer A, Winter I, Darwin C (1986) The representation of steady-state vowel sounds in the temporal discharge patterns of the guinea pig cochlear nerve and primarylike cochlear nucleus neurons. The Journal of the Acoustical Society of America 79:100-113.

Palmer AR, Shackleton TM, Sumner CJ, Zobay O, Rees A (2013) Classification of frequency response areas in the inferior colliculus reveals continua not discrete classes. The Journal of physiology 591:4003-4025.

Panzeri S, Brunel N, Logothetis NK, Kayser C (2010) Sensory neural codes using multiplexed temporal scales. Trends in neurosciences 33:111-120.

Panzeri S, Ince RA, Diamond ME, Kayser C (2014) Reading spike timing without a clock: intrinsic decoding of spike trains. Philosophical Transactions of the Royal Society B: Biological Sciences 369:20120467.

Perez CA, Engineer CT, Jakkamsetti V, Carraway RS, Perry MS, Kilgard MP (2013) Different timescales for the neural coding of consonant and vowel sounds. Cerebral Cortex 23:670-683.

Peterson GE, Barney HL (1952) Control methods used in a study of the vowels. The Journal of the Acoustical Society of America 24:175-184.

Pickles JO (1988) An introduction to the physiology of hearing: Academic press London.

Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural computation 16:1661-1687.

Quiroga RQ, Panzeri S (2009) Extracting information from neuronal populations: information theory and decoding approaches. Nature Reviews Neuroscience 10:173-185.

Ramachandran R, Davis KA, May BJ (1999) Single-unit responses in the inferior colliculus of decerebrate cats I. Classification based on frequency response maps. Journal of neurophysiology 82:152-163.

Ranasinghe K, Vrana W, Matney C, Kilgard M (2013) Increasing diversity of neural responses to speech sounds across the central auditory pathway. Neuroscience 252:80-97.

Ranasinghe KG, Vrana WA, Matney CJ, Kilgard MP (2012) Neural Mechanisms Supporting Robust Discrimination of Spectrally and Temporally Degraded Speech. Journal of the Association for Research in Otolaryngology 13:527-542.

Redies H, Sieben U, Creutzfeldt O (1989) Functional subdivisions in the auditory cortex of the guinea pig. Journal of Comparative Neurology 282:473-488.

Rhode WS, Greenberg S (1992) Physiology of the cochlear nuclei. In: The mammalian auditory pathway: Neurophysiology, pp 94-152: Springer.

Rhode WS, Greenberg S (1994) Lateral suppression and inhibition in the cochlear nucleus of the cat. Journal of neurophysiology 71.

Robles L, Ruggero MA (2001) Mechanics of the mammalian cochlea. Physiological reviews 81:1305.

Rose JE, Brugge JF, Anderson DJ, Hind JE (1967) Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. Journal of neurophysiology 30:769-793.

Rosen S, Walliker J, Brimacombe JA, Edgerton BJ (1989) Prosodic and segmental aspects of speech perception with the House/3M single-channel implant. Journal of Speech, Language, and Hearing Research 32:93-111.

Rouiller E, De Ribaupierre Y, De Ribaupierre F (1979) Phase-locked responses to low frequency tones in the medial geniculate body. Hearing Research 1:213-226.

Sachs MB, Young ED (1979) Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. The Journal of the Acoustical Society of America 66:470-479.

Sadagopan S, Wang X (2009) Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. The Journal of neuroscience 29:11192-11202.

Sally SL, Kelly JB (1988) Organization of auditory cortex in the albino rat: sound frequency. Journal of neurophysiology 59:1627-1638.

Sellick P, Russell I (1980) The responses of inner hair cells to basilar membrane velocity during low frequency auditory stimulation in the guinea pig cochlea. Hearing research 2:439-445.

Shamma SA (1985) Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. The Journal of the Acoustical Society of America 78:1612-1621.

Shamma SA, Chadwick RS, Wilbur WJ, Morrish KA, Rinzel J (1986) A biophysical model of cochlear processing: Intensity dependence of pure tone responses. The Journal of the Acoustical Society of America 80:133-145.

Shannon CE (1948) The bell technical journal. A mathematical theory of communication 27:379-423.

Shannon RV, Fu Q-J, Galvin 3rd J (2004) The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. Acta oto-laryngologica Supplementum 50-54.

Shannon RV, Jensvold A, Padilla M, Robert ME, Wang X (1999) Consonant recordings for speech testing. The Journal of the Acoustical Society of America 106:L71.

Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303-304.

Shannon RV, Zeng F-G, Wygonski J (1998) Speech recognition with altered spectral distribution of envelope cues. The Journal of the Acoustical Society of America 104:2467-2476.

Shetake JA, Wolf JT, Cheung RJ, Engineer CT, Ram SK, Kilgard MP (2011) Cortical activity patterns predict robust speech discrimination ability in noise. European Journal of Neuroscience 34:1823-1838.

Sinex DG, Geisler CD (1983) Responses of auditory-nerve fibers to consonant–vowel syllables. The Journal of the Acoustical Society of America 73:602-615.

Sivian L, White S (1933) On minimum audible sound fields. The Journal of the Acoustical Society of America 4:288-321.

Souza P, Rosen S (2009) Effects of envelope bandwidth on the intelligibility of sine-and noise-vocoded speech. The Journal of the Acoustical Society of America 126:792.

Spoendlin H, Schrott A (1989) Analysis of the human auditory nerve. Hearing research 43:25-38.

Stiebler I (1986) Tone-threshold mapping in the inferior colliculus of the house mouse. Neuroscience letters 65:336-340.

Stone MA, Füllgrabe C, Moore BC (2008) Benefit of high-rate envelope cues in vocoder processing: Effect of number of channels and spectral region. The Journal of the Acoustical Society of America 124:2272.

Sumner CJ, Lopez-Poveda EA, O'Mard LP, Meddis R (2002) A revised model of the inner-hair cell and auditory-nerve complex. The Journal of the Acoustical Society of America 111:2178-2188.

Sumner CJ, O'Mard LP, Lopez-Poveda EA, Meddis R (2003) A nonlinear filter-bank model of the guinea-pig cochlear nerve: rate responses. The Journal of the Acoustical Society of America 113:3264.

Sutter ML (2000) Shapes and level tolerances of frequency tuning curves in primary auditory cortex: quantitative measures and population codes. Journal of neurophysiology 84:1012-1025.

Syka J, Popelář J, Kvašňák E, Astl J (2000) Response properties of neurons in the central nucleus and external and dorsal cortices of the inferior colliculus in guinea pig. Experimental Brain Research 133:254-266.

Tasaki I (1954) Nerve impulses in individual auditory nerve fibers of guinea pig. J Neurophysiol 17:122.

Ter-Mikaelian M, Semple MN, Sanes DH (2013) Effects of spectral and temporal disruption on cortical encoding of gerbil vocalizations. Journal of neurophysiology 110:1190-1204.

Theunissen F, Miller JP (1995) Temporal encoding in nervous systems: a rigorous definition. Journal of computational neuroscience 2:149-162.

Van Tasell DJ, Greenfield DG, Logemann JJ, Nelson DA (1992) Temporal cues for consonant recognition: Training, talker generalization, and use in evaluation of cochlear implants. The Journal of the Acoustical Society of America 92:1247.

Van Tasell DJ, Soli SD, Kirby VM, Widin GP (1987) Speech waveform envelope cues for consonant recognition. The Journal of the Acoustical Society of America 82:1152-1161.

Volta A (1800) On the electricity excited by the mere contact of conducting substances of different kinds. In a letter from Mr. Alexander Volta, FRS Professor of Natural Philosophy in the University of Pavia, to the Rt. Hon. Sir Joseph Banks, Bart. KBPRS. Philosophical Transactions of the Royal Society of London 403-431.

Wallace MN, Rutkowski RG, Palmer AR (2000) Identification and localisation of auditory areas in guinea pig cortex. Experimental brain research 132:445-456.

Wang X, Lu T, Bendor D, Bartlett E (2008) Neural coding of temporal information in auditory thalamus and cortex. Neuroscience 154:294-303.

Wang X, Lu T, Snider RK, Liang L (2005) Sustained firing in auditory cortex evoked by preferred stimuli. Nature 435:341-346.

Warren III EH, Liberman MC (1989) Effects of contralateral sound on auditory-nerve responses. I. Contributions of cochlear efferents. Hearing research 37:89-104.

Wells TT (2014) Frequency selectivity measured both psychophysically and physiologically. PhD thesis, University of Nottingham.

Whitmal III NA, Poissant SF, Freyman RL, Helfer KS (2007) Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. The Journal of the Acoustical Society of America 122:2376-2388.

Wightman FL, Kistler DJ (1993) Sound localization. In: Human psychophysics, pp 155-192: Springer.

Wilson BS, Dorman MF (2007) The surprising performance of present-day cochlear implants. Biomedical Engineering, IEEE Transactions on 54:969-972.

Winer JA (1992) The functional architecture of the medial geniculate body and the primary auditory cortex. In: The mammalian auditory pathway: Neuroanatomy, pp 222-409: Springer.

Winer JA, Schreiner CE (2005) The central auditory system: a functional analysis. In: The inferior colliculus, pp 1-68: Springer.

Winter IM, Robertson D, Yates GK (1990) Diversity of characteristic frequency rate-intensity functions in guinea pig auditory nerve fibres. Hearing research 45:191-202.

Woolley SM, Fremouw TE, Hsu A, Theunissen FE (2005) Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nature neuroscience 8:1371-1379.

Woolley SM, Gill PR, Theunissen FE (2006) Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. The Journal of Neuroscience 26:2499-2512.

Xu L, Thompson CS, Pfingst BE (2005) Relative contributions of spectral and temporal cues for phoneme recognition. The Journal of the Acoustical Society of America 117:3255.

Young ED, Sachs MB (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. The Journal of the Acoustical Society of America 66:1381-1403.

Appendix

The parameters used for the DRNL filter, inner hair cell and auditory nerve synapse used in the auditory model described in chapter 2. Parameters were taken from Sumner et al. (2003), but are reproduced here for completeness.

DRNL filter parameters that are fixed across BFs

Compression exponents, *v* (dB/dB)	0.1
Gammatone cascade of nonlinear path	3
Low-pass filter cascade of nonlinear path	4
Centre frequency of nonlinear path, CF_{NL}	equal to BF
Low-pass cutoff of nonlinear path, LP_{NL}	equal to BF
Gammatone cascade of linear path	3
Low-pass filter cascade of linear path	4
Low-pass cutoff of linear path, LP_{lin}	Set equal to CF_{lin}

DRNL filter parameters that vary with BF $p(BF) = 10^{p_0 + m\log_{10}(BF)}$	p_0	m
Bandwidth of nonlinear path, BW_{NL} (Hz)	0.8	0.58
Compression parameter, a	1.87	0.45
Compression parameter, b	-5.65	0.875
Centre frequency of linear path, CF_{lin} (Hz)	0.339	0.895
Bandwidth of linear path, BW_{lin} (Hz)	1.3	0.53
Linear path gain, G_{lin}	5.68	-0.97

IHC parameters

Endocochlear potential, E_t	0.1
Potassium reversal potential, E_k	$-70.45e^{-3}$
Resting conductance, G_0	$1.974e^{-9}$
Potassium conductance, G_k	$18e^{-9}$

E_k correction	0.04
Mechanical conductance, G_{cilia}^{max}	$8e^{-9}$
Displacement sensitivity, s_0	$85e^{-9}$
Displacement offset, u_0	$7e^{-9}$
Displacement sensitivity, s_1	$5e^{-9}$
Displacement offset, u_1	$7e^{-9}$
Total capacitance, C_m	$6e^{-12}$
Cilia/BM time constant, τ_c	$2.13e^{-3}$
Cilia/BM coupling gain, C_{cilia}	16

IHC/AN synapse parameters

Scalar, z	$2e^{-32}$
Reversal potential, E_{Ca}	0.066
β_{Ca}	400
γ_{Ca}	130
Calcium current time constant, τ_m	$0.75e^{-4}$
Calcium diffusion time constant τ_{Ca}	$0.75e^{-4}$
Replenishment rate, y	10
Loss rate, l	2580
Reprocessing rate, x	66.3
Recovery rate, r	6580
Max. Ca^{2+} conductance, G_{Ca}^{max}	$2.4e^{-9}$
Threshold Ca^{2+} concentration, ${Ca^{2+}}_{thr}$	$3.35e^{-14}$
Max free transmitter quanta, M	10