Neural Coding of Overlapping Speech and the Effects of Hearing Loss

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

Samuel Steven Smith, BSc. MSc.
October 2019

Abstract

How listeners identify speech that is masked by an overlapping speaker remains a key problem in auditory neuroscience. A mild hearing impairment can profoundly alter a listener's capacity to identify speech amongst interfering talkers. Existing theories propose that the auditory system uses cues (e.g. pitch, onset/offset timing) to segregate neural representations of overlapping speech. Historically, much attention has been paid to the benefit of fundamental frequency cues to the identification of concurrently presented vowels, but segregation-based models fail to adequately explain human behaviour. Proposed here is a model void of segregation but that embodies an optimal strategy of predicting the combined representations of vowels. The model was able to quantitatively replicate listeners' identification of concurrent-vowel pairs, an improvement on segregation-based models. This was the case for neural representations simulated from a filter recreation of the auditory periphery, or recorded from multi-channel electrodes in the guinea pig midbrain. The compressive nature of auditory encoding appeared to facilitate a strategy of predicting the combined representation of speech. In addition, neural coding did not support previously proposed segregation cues such as temporally encoded pitch and harmonic beats. Neural recordings were also gathered in response to overlapping vowel-consonants, at variable onset delays, from animals with normal hearing or a noise induced hearing loss. Poorer speech recognition was predicted for the hearing loss data, compounded by an inability to optimally utilise knowledge of interfering speech sounds. Overall, investigation of the auditory system's goals seems a promising approach to understanding the neural processes underlying speech identification when masked by a competing talker.

Contents

Chapter 1 General Introduction 1

1.1. Motivation 1

1.2. Statement of the (Cocktail Party) Problem 3

1.3. Auditory System 4

1.4. Auditory Scene Analysis 8

1.5. Concurrent-vowel Identification 10

1.6. Glimpsing 19

1.7. Hearing Loss 21

1.8. Summary and Aims 23

1.9. Overview of Thesis 26

Chapter 2 Modelling Auditory Perception 27

2.1. Introduction 27

2.2. Statistical Decision Theory 30

2.3. Signal Detection Theory 31

2.4. Bayesian Decision Theory 37

2.5. Proposed Modelling Framework 40

2.6. Discussion and Conclusion 47

Chapter 3 Concurrent-vowel Identification: Auditory Filter Model 48

3.1. Introduction 48

3.2. Methods 50

3.3. Results 61

3.4. Discussion 78

3.5. Conclusion 82

Chapter 4 Concurrent-vowel Identification: Neurophysiology 83

4.1. Introduction 83

4.2. Methods 86

4.3. Results 96

4.4. Discussion 121

4.5. Conclusion 126

Chapter 5 Neural Recognition of Overlapping Vowel-consonants following Hearing Loss 127

5.1. Introduction 127

5.2. Methods 130

5.3. Results 137

5.4. Discussion 155

5.5. Conclusion 158

Chapter 6 General Discussion and Future Work 159

6.1. Motivation (i): "how humans identify speech masked by a competing speaker" 159

6.2. Motivation (ii): "changes to identification with hearing loss" 165

6.3. Summary 169

Appendix 170

References 172

List of Figures and Tables

Figure 1.1 Schematic of the superposition principle 3
Figure 1.2 Diagram of the classical auditory pathway 7
Figure 1.3 Listeners' performance in CVI tasks 11
Figure 1.4 Effect of syllable asynchrony on recognition 20
Figure 2.1 Probability diagram for the G-SDT model, and corresponding curves plotted in ROC space 34
Figure 2.2 Key components of reviewed decision theories 40
Figure 2.3 Visual representation of Eq. 2.8 43
Figure 3.1 Harmonic structure of interleaved concurrent-vowel pairs 51
Figure 3.2 Schematic of a filter recreation of the auditory periphery 55
Figure 3.3 Schematic of how predictions of confusion matrix entries are calculated by the proposed back-end classifier 58
Figure 3.4 Overview of computational modelling framework 59
Figure 3.5 Fitting parameter when vowels are independently identified 61
Figure 3.6 Fitting parameter under an optimal scheme 64
Figure 3.7 Identification of concurrent-vowel pairs under an optimal scheme 65
Figure 3.8 Scatter plots comparing confusion matrix entries 66
Figure 3.9 Prediction of listeners' CVI performance during training 69
Figure 3.10 MI between spectro-temporal features and model predictions 72
Figure 3.11 Model predictions for meta-parameter variants 74
Figure 3.12 Predictions made from the model inclusive of a PACF 76
Figure 3.13 Identification of interleaved vowel pairs 77
Figure 4.1 Schematic of chapter methodology 86
Figure 4.2 Enzyme activity associated with electrode tracks 90
Figure 4.3 Example of a FRA 91
Figure 4.4 PSTH examples 94
Figure 4.5 CAPs 96
Figure 4.6 Spontaneous and driven firing rates 97

Figure 4.7 Neural BF tuning 98
Figure 4.8 CVI predictions from optimally configured model using IC data 100
Figure 4.9 Scatter plots comparing confusion matrix entries 102
Figure 4.10 Model predictions for low-pass cut-off variants 103
Figure 4.11 MI between neural representations and model predictions 106
Figure 4.12 Changes in firing rates with F0 differences 108
Figure 4.13 Filter explanation schematic 110
Figure 4.14 Comparison of beating with changes in MI 112
Figure 4.15 Comparison of VS at beating periods with changes in MI 114
Figure 4.16 Example SACs 116
Figure 4.17 SAC measurements at F0s of vowels and concurrent-vowel pairs 117
Figure 4.18 Rate responses to concurrent-vowel pair /i,u/ 119
Figure 4.19 Neural responses to concurrent-vowel pairs compared with neural responses to constituent vowels 120
Figure 5.1 Schematic of overlapping VC task 131
Figure 5.2 Overview of modelling framework 136
Figure 5.3 ABRs before and after NE 137
Figure 5.4 CAPs for NH and HL animals 138
Figure 5.5 Spontaneous and driven firing rates in NH and HL animals 139
Figure 5.6 Neural BF tuning in NH and HL animals 140
Figure 5.7 Average FRAs in NH and HL animals 141
Figure 5.8 Neural recognition of VCs with a suboptimally configured model 143
Figure 5.9 Vowel and consonant identification 144
Figure 5.10 VC identification with an optimally configured model 145
Figure 5.11 Identification of the vowels /u/and /ɝ/ 147
Figure 5.12 Identification of the consonants /k/ and /p/ 148
Figure 5.13 Identification of the vowels /a/ and /ɒ/ 149
Figure 5.14 Identification of the consonants /s/ and /ʃ/ 150
Figure 5.15 Histogram of BFs for NH and HL datasets 151
Figure 5.16 VC identification predicted from frequency constrained models 154
Figure A.1 Flat laminar cells of the IC 171

Table 2.1 Generic confusion matrix for binary decision tasks. 28
Table 3.1 Confusion matrices for human listeners and suboptimal model.....63
Table 3.2 Confusion matrices for optimal model 67
Table 4.1 Confusion matrices for optimal model with neural data 101

Publications

Peer-reviewed Journal Articles

Smith, S. S., Chintanpalli, A., Heinz, M. G., & Sumner, C. J. (2018). Revisiting Models of Concurrent Vowel Identification: The Critical Case of No Pitch Differences. *Acta Acustica united with Acustica*, *104*(5), 922-925.

Selected International Conference Proceedings

Integrating classifiers and electrophysiology to better understand hearing loss. *Computational Neuroscience Meeting. Barcelona, Spain. July 16, 2019.*

Decoding the neural representations of vowel-consonants following hearing loss. *International Hearing Loss Conference. Niagara-on-the-lake, Canada. May 7, 2019.*

Classifying the neural code of concurrently presented vowels. *Society for Neuroscience Annual Meeting. San Diego, USA. November 7, 2018.*

A model of concurrent vowel identification without segregation predicts perceptual errors. *Association for Research in Otolaryngology Midwinter Meeting. San Diego, USA. February 10, 2018.*

Abbreviations

ABR	Auditory brainstem response
ACF	Autocorrelation function
AM	Amplitude modulation
AN	Auditory nerve
ANF	Auditory nerve fiber
ASA	Auditory scene analysis
AUC	Area under curve
BF	Best frequency
CAP	Compound action potential
CN	Cochlear nucleus
CVI	Concurrent-vowel identification
ENV	Envelope
F0	Fundamental frequency
FM	Frequency modulation
FRA	Frequency response area
(G-)SDT	(Gaussian) signal detection theory
HL	Hearing loss
IC	Inferior colliculus
NE	Noise exposure
NH	Normal hearing
MI	Mutual information
ML	Machine learning
MUA	Multi-unit activity
PCA	Principal component analysis
PCC	Pearson's correlation coefficient
PSTH	Peri-stimulus time histogram

RMS(E)	Root mean square (error)
ROC	Receiver operating characteristic
SAC	Shuffled autocorrelogram
SNR	Signal to noise ratio
SOC	Superior olivary complex
SPL	Sound pressure level
TFS	Temporal fine structure
VC	Vowel-consonant
VS	Vector strength

Acknowledgements

I would like to sincerely thank my lead supervisor, Dr. Christian Sumner, for his guidance and continuous support throughout the Ph.D. process. Thanks are duly extended to the other members of my supervisory committee, Dr. Mark Wallace and Prof. Michael Akeroyd. The community of hearing researchers in Nottingham, staff and students alike, have all contributed to what has been an enriching and stimulating few years. I have particularly fond memories of my time spent on Floor 3 with members of the Neuro group. Finally, special acknowledgements are offered to my parents, brother, and partner for their resolute encouragement and willingness to help in any way they could. Thank you.

Chapter 1 General Introduction

1.1. Motivation

Humans are able to communicate in complex auditory scenes. At a party, for example, an individual must try to understand their conversation partner despite multiple other audible conversations occurring simultaneously. Speech of interest is potentially modified due to the interaction of undesirable speech sounds. To successfully infer a conversation partner's speech is to solve the aptly-named "cocktail party problem" (Cherry, 1953; McDermott, 2009). In spite of this name, aspects of the cocktail party problem apply in a wide array of environments such as a classroom or a restaurant. Although intrinsically hard, observations reveal that humans are remarkably good at identifying speech masked by competing talkers (see Assmann and Summerfield's (2004) review).

Notably, an inability to communicate in complex auditory scenes is a leading complaint coincident with hearing loss (HL; Pichora-Fuller, 1997). The shortcomings of a failure to communicate following HL are vast, having a proven negative impact on employment, mental illness, and development (Action on Hearing Loss, 2015; 2018). Despite the benefits of hearing aids in quiet environments, deficits to the identification of overlapping speech does not seem attributable to simply a lack of audibility (Moore, Peters, & Stone, 1999; Plomp, 1978). Difficulties therefore likely stem from complications in auditory processing that are not currently understood.

The closely intertwined topics of

(i) how humans identify speech masked by a competing speaker,

(ii) and changes to identification with hearing loss,

constitute the main considerations of this thesis. This chapter introduces scientific literature relevant to these topics, followed by a summary, and declaration of the thesis' aims. Finally, an indication as to the content within the remainder of the thesis is provided.

1.2. Statement of the (Cocktail Party) Problem

Physical sounds are vibrations characterised by compressions and rarefactions in pressure. A sound's waveform captures these oscillations in pressure as a function of time. Speech waveforms are aggregated from rapid oscillations, called temporal fine structure (TFS), that are slowly modulated by a temporal envelope (ENV). TFS and ENV characteristics affect the perception of the frequency and amplitude properties of sound (Moon & Hong, 2014).

When two sounds are presented together in air their waveforms obey the superposition principle (Verma, 2013, p.304). This means that the displacement of two waves when presented together is equal to the sum of the individual wave displacements. If the waveform of a target speech sound is described as a function $target(t)$, where t is time, and a waveform of an interfering speech sound $interferer(t)$, then their superposition is a waveform $w(t)$, where $w(t) = target(t) + interferer(t)$. When trying to identify the target speech, the problem that the auditory system must solve (i.e. the cocktail party problem) is to determine the identity of $target(t)$ having only observed $w(t)$. Additionally, the system likely only has limited knowledge of the interfering speech. In this thesis, the problem as stated is considered almost exclusively in the monaural domain.

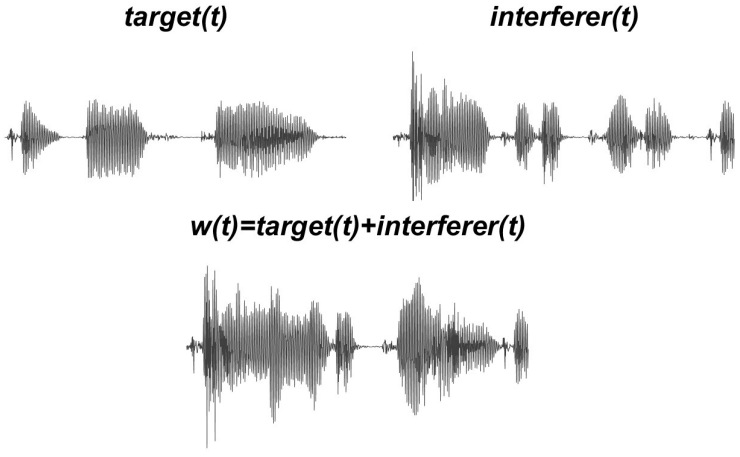


Figure 1.1 Schematic of the superposition principle for waveforms $target(t)$ and $interferer(t)$ representative of target speech and interfering speech accordingly.

1.3. Auditory System

The system that solves the problem stated in Section 1.2 is broadly segmented in to two sections, the auditory periphery and the central auditory nervous system. The auditory periphery is by far the best understood of the two sections, and concerns the conversion of sound from its mechanical form in air to electronic signals in the brain. The nervous system is a chain of nuclei that process information about sound in the form of neural activity. This neural code is used to create perceptual representations of the auditory environment. Due to ongoing references to the auditory system throughout this thesis, a brief description of it follows.

1.3.1. Periphery

Sound vibrations travel through the external pinnae, auditory meatus, and tympanic membrane, to the middle ear. The ossicles of the middle ear transmit vibrations into the liquid medium of the cochlea. The cochlea, with chambers divided by the basilar membrane, performs a frequency spectrum decomposition (Von Békésy, 1944). Often represented by a bank of overlapping bandpass filters, this decomposition splits a sound into parallel narrowband signals variable in their TFS and ENV information. This is determined mechanically by the stiffness of the basilar membrane. The membrane is narrow and thick at its basal end, only permitting movements associated with high frequencies, and monotonically reduces in stiffness along its length, isolating lower frequencies towards its apex. These signals undergo mechano-electronic transduction in the organ of corti. Outer hair cells actively amplify (Dallos, 1992; Fettiplace & Hackney, 2006) signals prior to the depolarisation of inner hair cells due to the deflection of stereocilia. Due to this mechanism being most active at lower sound levels (Cooper & Yates, 1994) signals are nonlinearly transformed. Subsequent to the deflection of stereocilia, the opening of calcium ion channels leads to a depolarisation of the membrane. This

results in the release of neurotransmitters into the synaptic cleft that bind to postsynaptic receptors of spiral ganglion cells triggering action potentials.

1.3.2. Nervous System

At this point along the auditory pathway, sounds are represented in the form of electrical signals passing along 30,000 auditory nerve fibres (ANFs), on each side of the head. ANFs are characterised as low in their spontaneous firing of action potentials (also known as spikes) if rates are less than 0.5 s^{-1}, medium if between 0.5 and 18 s^{-1}, and high if more than 18 s^{-1} (Liberman, 1978). ANF spontaneous rates are inversely correlated with their thresholds. In the auditory nerve (AN), frequency mapping (termed tonotopicity) established at the cochlea is maintained.

The AN leads to a complex grid of nuclei that, although structurally well defined (see Figure 1.2), are functionally less well understood. Initially, each nerve splits into at least two branches and innervates the dorsal and ventral divisions of the ipsilateral cochlear nucleus (CN) of the brainstem, whilst still preserving tonotopicity. Immediately a large array of distinct neuron types are found in the CN (e.g. fusiform cells, granular cells, and brush cells in the dorsal division; spherical and globular bushy cells, medium and large stellate cells in the ventral division; Rhode and Greenberg, 1992) as well as a variety of interneurons. These are mainly projection neurons which provide at least five different parallel output channels. Neural signals then pass ipsilateral, and contralateral, via a variety of pathways one of which is the nucleus of the trapezoid body to the superior olivary complex (SOC; thought to play a key role in the processing of position in space). Most ascending nerve fibres then collect in the lateral lemniscus before terminating in the inferior colliculus (IC). Still, tonotopicity is maintained, with a laminar low to high frequency distribution mediated from the dorsolateral to the ventromedial side of the central nucleus of the IC (Schreiner & Langner, 1997). The IC has been demonstrated to be sensitive to speech-like frequency modulation (FM) sweeps (Rees & Møller, 1983; Vartanian, 1974) and different amplitude modulation (AM)

frequencies (Kay, 1982). Yet, the existence of a modulation filter bank, similar to the frequency filter bank in the cochlea, is unclear (Schnupp, Garcia-Lazaro, & Lesica, 2015). Multimodal sensory information is integrated in the IC, thought to be influential in the understanding of speech via processes such as lip reading (Champoux et al., 2007).

Auditory signals then relay by way of the medial geniculate nucleus before reaching the auditory cortex (AC). Here both simple and complex processing occurs to aid the creation of sound perception. Specialised areas of the AC process high level characteristics of speech and music, for example, focusing on harmony, pitch, and timing (Deutsch, 2010). The AC sends efferent feedback projections towards the IC, thalamus, and lower structures, allowing for multiple feedback loops and parallel processing of auditory information (Huffman & Henson, 1990). Indeed, afferent and efferent processing integrate across the extent of the auditory nervous system (Burguetti & Carvallo, 2008).

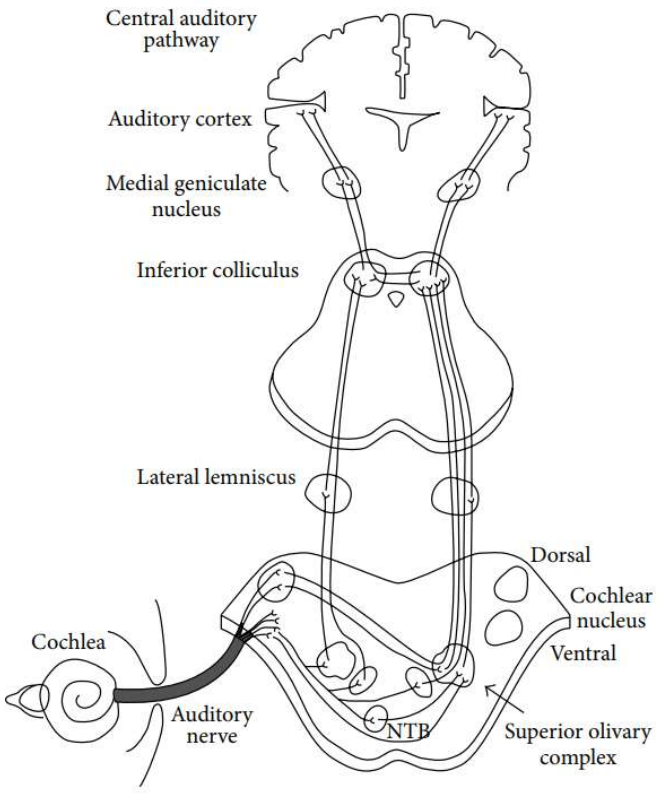


Figure 1.2 Schematic of the primary (lemniscal) auditory pathway within the nervous system (Lucker & Doman, 2015).

1.3.3. Guinea Pig Auditory System

The guinea pig is a commonly studied animal model of the human auditory system, with their cochlear mechanics first described by Von Békésy and Wever (1960) over 50 years ago. Guinea pigs have similar low frequency hearing thresholds as humans, although slightly poorer frequency selectivity (Glasberg & Moore, 1990). Additionally, guinea pig vocalizations incorporate frequencies common to those found in human speech (Green et al., 2018).

1.4. Auditory Scene Analysis

The cocktail party problem is one exemplar of auditory scene analysis (ASA; Bregman, 1990). ASA investigates how the auditory system organises sounds into perceptual units. In his book, Bregman (1990) posited that the system segregates a scene into separate auditory streams where each stream represents a single sound source. A stream is formed from the simultaneous integration of correlated auditory features (e.g. harmonicity, dynamics, onset/offset timing, spectral formants). By integrating auditory features into streams, the auditory scene is at the same time being segregated. Following auditory stream segregation, a listener can then attend to one of the segregated streams. The field of ASA has been interested in the auditory features used to group an auditory scene and how this is implemented. Grouping properties have commonly been investigated with alternating pure tone stimuli varying in presentation rate and frequency (Carlyon, Plack, Fantini, & Cusack, 2003; Scholes, Palmer, & Sumner, 2015).

In real life, auditory scenes are built from the interaction of a multitude of sources with characteristics beyond the complexity of pure tones. There are two ways in which sources can interfere with one another, energetic masking and informational masking (Bee & Micheyl, 2008). In the auditory domain, energetic masking refers to a signal that encounters direct acoustic interference by a signal of another source. At the level of the cochlea, signals overlap and obscure one another in frequency and time. Informational masking is an encompassing term for masking that is not energetic (Durlach et al., 2003). In essence, it refers to instances where auditory signals do not overlap in frequency or time yet their coincidence results in poorer intelligibility. Informational masking can be explained within the framework of ASA as a malfunctioning of the formation of auditory streams (Shinn-Cunningham, 2008). The physical instantiation of ASA and energetic/informational masking, particularly with regard to the roles of bottom-up processes and top-down attention, is an ongoing debate (Macken, Tremblay, Houghton, Nicholls, & Jones, 2003).

The desire to comprehensively explain ASA has spurred the subfield of computational auditory scene analysis (CASA; Brown and Cooke, 1994). Several prominent theories of the parsing and processing of auditory scenes have been implemented as computational models. CASA models that apply Bayesian inference have recently risen in prominence (Barniv & Nelken, 2015; Boes, de Coensel, Oldoni, & Botteldooren, 2011; Nix & Hohmann, 2007). These rather abstract models make predictions about auditory scenes by optimally calculating the conditional probability of observing auditory features given the presence of a source. This probability is moderated by the prior probability of that source being present in the scene. Bayesian CASA models have been demonstrated to track and segregate binaural mixtures of voices, and to positively correlate with a listener's perceptual phases (a phase is a time period where the perception of a stream is constant (Denham, Gyimesi, Stefanics, & Winkler, 2013; Pressnitzer & Hupé, 2006)). Other CASA models include those created to imitate neurocomputation (Pichevar & Rouat, 2007; Wang & Chang, 2008) and models that specifically focus on processing of temporal coherences between auditory features (Elhilali & Shamma, 2008). A critique of existing CASA models is that they are focussed on assessing specific algorithms for specific problems, rather than creating a comprehensive computational framework for performing ASA (Szabó, Denham, & Winkler, 2016).

1.5. Concurrent-vowel Identification

Of the multitude of auditory features available to facilitate the analysis of speech in noise, much attention has been given to fundamental frequency (F0). F0 is the lowest frequency existent within a repeating signal. Two competing speakers will produce speech at F0s that likely contrast with one another, making it a potentially useful cue for communication (Brokx & Nooteboom, 1982; Brungart, 2001). Researchers have commonly sought to probe F0 processing using synthesised vowels (harmonic steady state complexes; Klatt, 1980). These highly controlled stimuli attempt to restrict the availability of auditory features beyond F0.

A basic concurrent-vowel identification (CVI) experiment involves presenting two synthesised vowels, with identical onset and offset times and levels. Vowel identities are determined by their distinct formant structures, but vowels can also vary in their F0s. The experimental participant is asked which two vowel identities were presented from all possible combinations. CVI attempts to focus the complexity of speech identification in the presence of interfering speech into a controlled laboratory task. The main result from CVI experiments is that as the F0 difference between vowels increases, identification improves, denoted an 'F0 benefit' (Figure 1.3). This improvement occurs with F0 differences up to 1 semitone in size, after which performance plateaus (Assmann & Summerfield, 1990; Chintanpalli & Heinz, 2013; Scheffers, 1983; Zwicker, 1984).

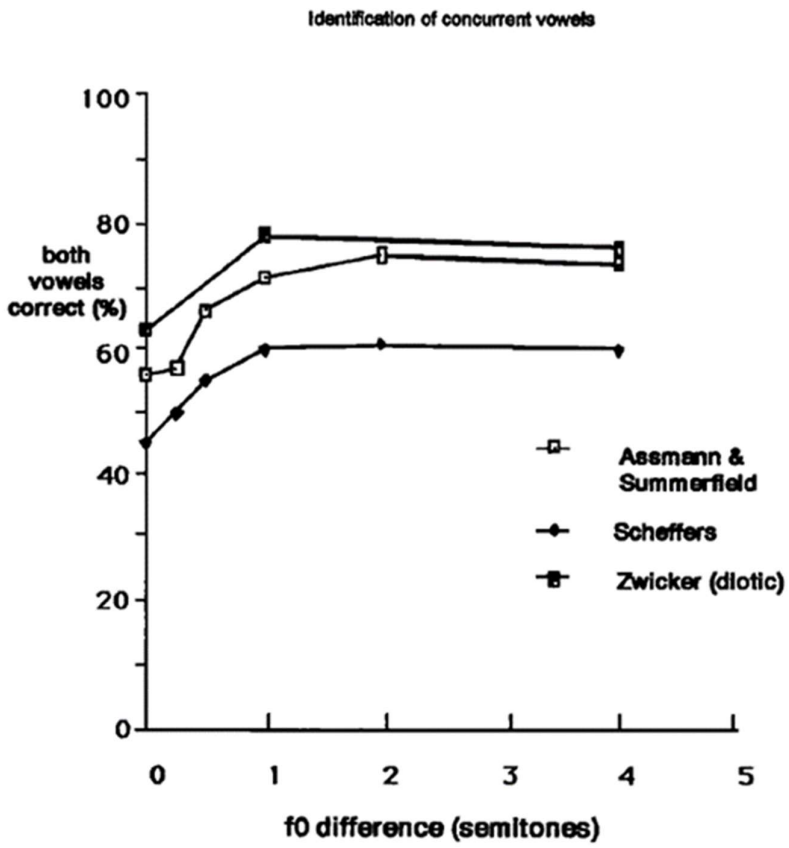


Figure 1.3 CVI results taken from Assmann and Summerfield (1990), Scheffers (1983), and Zwicker (1984). Published in Meddis and Hewitt (1992).

1.5.1. Computational Models

To understand the role of F0 differences in CVI, researchers have repeatedly turned to computational modelling. This tool allows hypotheses about processing and mechanisms underlying CVI to be tested. Much like the ASA framework, all existing models of CVI attempt to segregate representations of a concurrent-vowel pair. This stage is then followed by a comparison of the resulting segments with stored templates describing individual vowels. However, the representational form of concurrent-vowel pairs, segregation cues, mechanisms by which segregation is implemented, template comparison procedure, and predictive performances, vary between models.

1.5.1.1. Pitch Cues

Scheffers (1983) first proposed the CVI paradigm in his PhD thesis for which he attempted to build an explanatory computational model. His model was applied to a rate-place representation of concurrent-vowels that characterises firing rate as a function of tonotopic place (established at the basilar membrane). Representations of concurrent-vowels were produced with a simulation of the auditory periphery and temporal variations in firing rate were averaged over time. Scheffers proposed pitch as a segregation cue, where pitch is a perception of the F0 of a harmonic complex. The notion of pitch as a perceptual entity is crucial, exemplified in instances where, despite the exclusion of an F0 in a complex, pitch is implied by higher frequency components (Plack, Oxenham, & Fay, 2006). In his model, pitch was computed as the fundamental harmonic spectrum which best fit the rate-place representation of a sound, implemented using a harmonic sieve (Duifhuis, Willems, & Sluyter, 1982). For a given concurrent-vowel pair two pitches were identified. The accompanying rate representation was segregated into two patterns by sampling at harmonics associated with each pitch. Predictions of what the presented concurrent-vowel pair was were made by template matching the sampled segments with representations of individual vowels. Scheffers' model was largely only able to identify one vowel of a pair correctly, the model's overall average proportion of correct response was lower than listeners', and it did not predict an F0 benefit. From this he concluded that pitch was not used to perform segregation when performing CVI.

Persisting with the notion of pitch as a cue, Assmann and Summerfield (1990) implemented a model of CVI conceptually similar to Scheffers', but which analysed the TFS of the vowel waveforms (they also implemented other model variations that are not relevant within the scope of this review). Implementation involved the inclusion of an autocorrelation function (ACF; Rabiner and Schafer, 1978, p.141) which correlated a concurrent-vowel pair's filtered waveform with delayed copies of itself. By applying a peak-picking algorithm to an ACF representation of a concurrent-vowel pair, the two strongest periodicities were

selected as pitches. Representations were then segregated based on the degree to which constituent channels matched the periodicities of the two pitches identified. Finally, a template comparison was applied. Despite the different mechanism by which the model implemented pitch-based segregation, the model failed to predict the behaviour of listeners. No F0 benefit was predicted, and the Pearson's correlation coefficient (PCC; Pearson, 1956) between the model's predicted decisions (referred to as a confusion matrix) and listeners' ranged from 0.4 to 0.62. This result seemed attributable to a failure of the pitch mechanism as, when provided with the 'true' pitches of concurrent-vowel pairs, the model was better able to predict an F0 benefit and PCCs rose to between 0.61 and 0.75.

Meddis and Hewitt (1992) investigated a model similar in nature to Assmann and Summerfield's, but with a crucial distinction. The peak-picking algorithm only identified the single most salient pitch from the ACF. This is in line with an observation by Scheffers (1983) that, following the presentation of a concurrent-vowel pair, a single vowel is dominant and coloured by a less easily identifiable second vowel. Segmentation of the ACF representation was thus achieved by selecting channels that suitably matched the periodicity of this one pitch, or channels that did not. This model was autonomously able to predict an improvement in CVI as the F0 difference between the vowels increased from 0 to 1 semitones. Meddis and Hewitt attributed this improvement with F0 differences as a direct consequence of an improvement in segregation. This model has been widely cited and further variations of this model have been implemented (de Cheveigné's (1997) harmonic cancellation model and Brown & Wang's (1997) dynamical system model).

Regardless, Meddis and Hewitt's model failed to quantitatively predict listeners' specific decision probabilities, such as errors. This failure can be explained in part by the deterministic nature of the model. For a concurrent-vowel pair presentation, the model would always make the same prediction as to what that concurrent-vowel pair was (i.e. decision types were either made 100% of the time or 0% of the

time, whereas listeners are less certain about their errors and make variable decisions). Attempts to reconstruct Meddis and Hewitt's model have helped to highlight an inability to predict the specific decisions that listeners make, notably their errors (Chintanpalli & Heinz, 2013). A normalised similarity metric was calculated between a confusion matrix derived from human listeners and that predicted by a reconstruction of Meddis and Hewitt's model. For each deterministic prediction made by the model, the similarity metric equalled the average human response percentage for that decision, normalised by the highest percentage human response for the presented concurrent-vowel pair. The metric was generous in that it was maximal (100%) if the model identified the concurrent-vowel pair with the highest human response percentage. Regardless, the model's mean normalised similarity scores ranged from ~60% (at F0 differences of 0 and 0.25 semitones) to ~80% (otherwise).

In addition to the poor predictive performance of pitch-based segregation models, the mechanisms by which they operate have also been subject to criticism. Assmann and Paschall (1998) asked listeners to adjust a tone complex to match the F0s of concurrent-vowel pairs. Listeners only perceived two distinct pitches when the F0 difference between vowels was above or equal to 4 semitones. An inability to rank the pitches of concurrent-vowels at F0s less than 1 semitones has also been cited (Culling & Darwin, 1994). Given that the largest improvements in CVI occur when F0 differences are smaller than one semitone, these findings potentially compromise a model dependent upon pitch identification. Arguably, the perception of pitch could be a process separate to that which is used to take advantage of harmonicity when performing CVI. An altogether different possibility is that whilst pitch is used for segregation at large F0 differences, a different mechanism is at work for smaller F0 differences.

1.5.1.2. Beating Cues

Culling and Darwin (1994) suggested an alternative cue which could facilitate CVI at smaller F0 differences. When two harmonics interact within a single auditory

filter, it results in ENV modulations with a frequency the difference of the harmonic frequencies. Termed beating, Culling and Darwin argued that unresolved harmonics could explain the poor performance of purely rate-based models. Yet, they proposed that due to beating, relative wave phases could at points enhance one another, and at points cancel one another, temporarily making spectral features of vowels more prominent.

Darwin and Culling (1993) created a CVI experiment where stimuli only contained F0 differences at low frequencies (first formant). Listeners still exhibited an F0 benefit at small F0 differences. Culling and Darwin (1994) also ran a CVI experiment using vowels with interleaved harmonics designed to preserve spectral ENVs and beating patterns, but would break a pitch-based segregation process. Again, listeners still exhibited an F0 benefit at small F0 differences, but this was slightly reduced in magnitude relative to the original vowel stimuli.

Culling and Darwin (1994) produced a computational model to explain the results they observed using interleaved vowels. The model used a two-layer perceptron to compare rate-based representations of a concurrent-vowel pair with templates of individual vowels. Comparisons were repeatedly performed at regular time intervals, motivated by the idea that beating would make the constituent vowels more salient at different instances of their phase cycles. Vowels were compared with concurrent-vowel representations independently of one another. Their model was able to match average human performance, and predict a modest F0 benefit, for both the original and interleaved concurrent-vowel pairs. Correlating the F0 benefit (first order changes in identification) predicted by the model with listeners' resulted in PCCs ranging from 0.47 to 0.80. Correlating all the entries in a confusion matrix produced by their model with listeners', resulted in a low PCC of 0.27 for vowels with no F0 differences.

Like harmonic cues, the plausibility of beating cues has also been contested. When concurrent-vowel phase relationships were carefully altered to retain harmonic

cues but minimise beating, CVI performance was unaffected at a number of F0 differences between 0 and 1 semitones (de Cheveigné, 1999). de Cheveigné (1999) suggested that Culling and Darwin's results could be explained by harmonic features that were not completely removed from the interleaved vowel stimuli. He also noted how dependent beating cues are on the starting phase relationship of the stimuli presented (phase of Klatt synthesizer is highly conducive to modulation patterns).

1.5.1.3. Non-F0 Cues

In CVI experiments, listeners have consistently identified concurrently presented vowels greatly above chance in the absence of any F0 differences. In this situation there are no F0 cues for the auditory system to utilise for segregation. As a result, all existing models of CVI that try to segregate concurrent-vowel pair representations are unable to do so. The models compare individual vowel representations with representations of entire concurrent-vowel pairs. Consequently, these models fail to predict the ability of listeners to identify concurrent-vowel pairs in the absence of any F0 differences. Assmann and Summerfield (1989) modelled CVI without F0 cues, exploring the effects of transformations to auditory spectral patterns on identification. Predictions of presented concurrent-vowel pairs were made by an independent comparison of individual vowel patterns with a concurrent-vowel's. They found that emphasising spectral peaks (i.e. formants) best predicted listener data. Across all the transformations they attempted, PCCs ranged between 0.42 and 0.71. Largely ignored in the literature since 1989, the cues used when vowels have no F0 difference remain speculative, yet are important as they likely support identification when F0 differences are introduced.

1.5.2. Neurophysiology

Additional to computational models of CVI, researchers have investigated neuronal responses to concurrent-vowel stimuli in animal models. Palmer (1990, 1992) looked at AN encoding of overlapping vowels in guinea pigs. Recording from single neurons, he determined that spike timing patterns in the AN were capable of representing F0 and formant information of concurrently presented vowel stimuli. He showed this with multiple analytical approaches including Fourier analysis, spike interval histograms, and ACFs. Beyond showing that there was sufficient temporal information in AN encoding to perform F0-based segregation, if and how the auditory system extracts this information was not addressed. Additionally, comparison of the segmented vowel profiles with individual vowel profiles was only performed qualitatively. Experiments were only performed for two concurrent-vowel pairs, /a,i/ and /ɔ,i/ with a fixed F0 difference of 25 Hz. Consequently, the extent to which these findings generalise is speculative.

In the hindbrain, Keilson, Richards, Wyman, and Young (1997) recorded responses to concurrently presented vowels in chopper and primary-like neurons in the ventral CN of cats. They found that chopper neurons with best frequencies (BFs) above 500-700 Hz well represented F0 information in their temporal synchrony (i.e. did not represent 2nd/3rd formants) and vowel spectrums in their rate responses. This was less the case in the primary-like neurons and chopper neurons with BFs below 500-700 Hz. Temporal encoding of F0s in the CN was again discussed within the context of harmonic segregation models. How F0 information would be neurally decoded was not addressed. One candidate region could be the central nucleus of the IC. Nakamoto, Shackleton and Palmer (2010) presented concurrent harmonic series, not vowels, to guinea pigs and recorded from the IC. They discovered one of the two harmonic complexes were encoded significantly more than the other. This enhancement diminished following the deactivation of the AC. This suggests a role for temporal processing of F0, mediated by top-down auditory processing. Note these findings should be interpreted with caution due to the

methodology involving cooling of the AC, which could have altered descending connections other than to the IC.

At the level of the AC, Fishman, Micheyl and Steinschneider (2016) observed the multiunit activity of A1 neurons in alert macaque monkeys to concurrently presented vowels. Sufficiently resolved lower harmonics reliably encoded F0s, and from this determination vowel identities could be recognised using template matching procedures. Assuming that concurrent-vowel pairs are temporally encoded earlier in the auditory system, the fidelity with which concurrent-vowels are spectrally encoded in the macaque A1 would advocate some form of temporal-rate conversion. Indeed, this would be required due to the low temporal encoding limit of the AC (Langner, 1992). This experiment only investigated responses to one concurrent-vowel pair at a single F0 difference (26 Hz). Additionally, for each neural region that was recorded from, stimuli were scaled relative to the region's BF. Thus, the results are only valid under an implicit assumption of scaling invariance in the auditory system (Larsen, Cedolin, & Delgutte, 2008).

1.6. Glimpsing

Glimpses are another auditory feature thought to aid speech identification in noisy environments (Cooke, 2006). A glimpse is defined as a spectro-temporal region of target speech that is relatively unaffected by interfering signals. Glimpsing can be thought of as akin to visual occlusion, where objects are only partially visible due to the obstruction of another object (Scholl & Pylyshyn, 1999). The value of glimpsing has been demonstrated with measurements of intelligibility where average signal to noise ratios (SNR) are held constant, but maskers varied in their temporal profiles. Speech maskers or AM noise resulted in better intelligibility than speech babble which, in turn, was better than steady state noise (Festen & Plomp, 1990; Simpson & Cooke, 2005). It is by taking advantage of vowel glimpses that Culling and Darwin (1994) were suggesting that beating aided CVI.

Vestergaard, Fyson, and Patterson (2011) constructed a psychoacoustic paradigm where syllables were presented temporally overlapping one another to various degrees. They showed that reducing the extent of temporal overlap, increasing the availability of glimpses, resulted in the best identification. This was found for both overlapping vowel-consonant (VC) stimuli and consonant-vowel stimuli (Figure 1.4). When acoustic properties of the stimuli were modified to simulate syllables emanating from vocal tracts of different length, identification improved further. This was largely attributed to less energetic masking of the stimuli (i.e. better spectral glimpses).

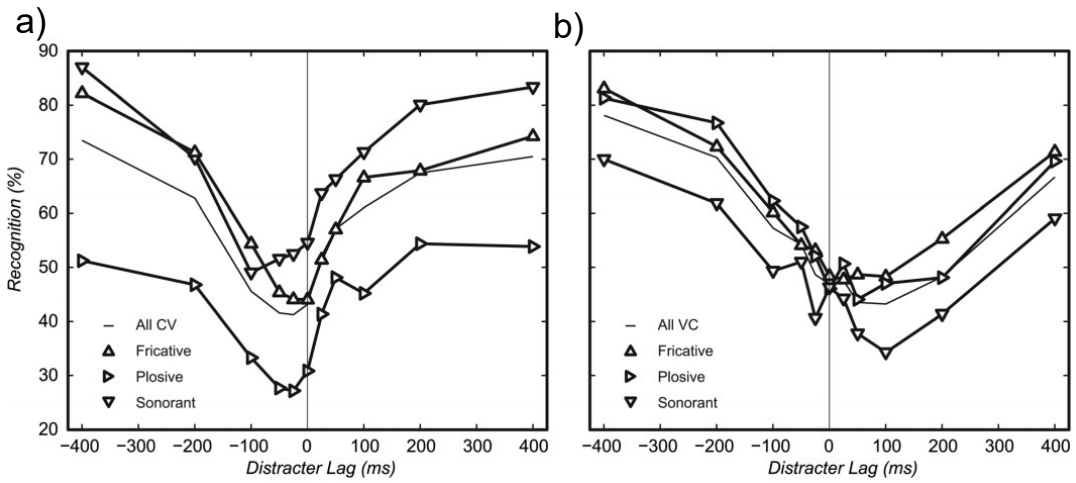


Figure 1.4 (from Vestergaard, Fyson and Patterson, 2011) Effect of syllable asynchrony on recognition. (a) shows recognition for consonant-vowels, and (b) VCs.

Computational models of glimpsing have been built (Best, Mason, Swaminathan, Roverud, & Kidd, 2017; Cooke, 2006). Cooke's (2006) model, incorporating simple heuristics regarding the determination of glimpses (e.g. analysing areas where SNR is greater than 0 dB), was able to account for human intelligibility of consonants in a selection of maskers, although the model could not so much predict listeners' confusions. The ability of this model to determine SNRs was because it was provided with complete knowledge of the signal and noise waveforms. For naïve humans, without knowledge of the interfering noise, this SNR calculation would be difficult. Engineering approaches to synthesise enhanced speech have extracted glimpses by applying an ideal binary mask (Wang, 2005). The ideal binary mask removes spectro-temporal regions where a SNR is less than a predefined threshold. Like Cooke's model this has been performed with full knowledge of both the signal and the noise, but approaches involving estimation of noise signals have also been constructed (Hu & Loizou, 2008).

1.7. Hearing Loss

As mentioned in Section 1.1, even a mild HL can significantly impair the ability of listeners to solve the cocktail party problem. Compared to individuals with normal hearing (NH), a HL has been shown to impede performance in a number of experimental tasks involving simultaneously presented sounds (Dubno, Dirks, & Morgan, 1984; Helfer & Freyman, 2008). Even when the identification of individual vowels in quiet is unaffected, adults with HL perform poorer at CVI and have reduced F0 benefits when compared to individuals with NH (Arehart, Souza, Muralimanohar, & Miller, 2011; Chintanpalli, Ahlstrom, & Dubno, 2016).

The causes of hearing impairment are largely concentrated within the cochlea. Here, outer hair cells, essential for nonlinear frequency selectivity and compression (Oxenham & Bacon, 2003), are damaged resulting in poorer functioning (Baer & Moore, 1994). The ensuing smearing of the frequency spectrum reduces the sparsity with which signals are represented. In addition, outer hair cells become depolarised following HL (Cody & Russell, 1985). This results in smaller dynamic ranges with elevated thresholds (Ruggero, Rich, & Recio-Spinoso, 1996). HL can also spur 'dead regions' where inner hair cells are missing. Hard to spot in audiograms (Moore, 2001), it is only when pure tone audiometry thresholds are markedly abnormal that inner hair cell dysfunction is attributed. Together these cochlear morbidities directly alter how stimuli are encoded entering the auditory nervous system.

The death of spiral ganglion cells was previously considered a secondary event following cochlear hair cell death (Spoendlin, 1975). Kujawa and Liberman (2009) found a prompt reduction in afferent nerve terminals following a two hour long noise over exposure at 100 dB SPL. Yet, they observed complete recovery of hair cell function and audiometric thresholds. Consequently, a number of morbidities can occur in the AN without altering audiometric thresholds. Damage to low spontaneous rate ANFs with high thresholds impairs how informatively they can

represent loud environments (Bakay, Anderson, Garcia-Lazaro, McAlpine, & Schaette, 2018). Additionally, the loss of ANFs results in stimulus undersampling resulting in less robust neural codes describing noisy conditions (Lopez-Poveda & Barrios, 2013).

Peripheral damage following HL has been shown to alter temporal coding. For example, the benefit of introducing AM to an interfering noise provided an improved threshold of 6 dB SPL within NH listeners but only 2 dB SPL for individuals with cochlea damage (Peters, Moore, & Baer, 1998). In the AN, noise exposure (NE) was found to enhance ENV information about signals in noise, whereas TFS information was degraded (Henry & Heinz, 2012; Henry, Kale, & Heinz, 2014). Swaminathan and Heinz (2012) pointed out that findings that disambiguate ENV and TFS cues should be interpreted with caution as these cues are not orthogonal to one another. In addition, neural encoding of temporal onsets also changes following HL (Scheidt, Kale, & Heinz, 2010).

Robertson and Irvine (1989) explored tonotopic disturbances in the AC as a result of unilateral HL. Neurons in cortex regions associated with cochlea damage underwent altered responses towards frequencies not previously observed in the region. Remapping could also occur in the neural instantiation of more abstract auditory features important for solving the cocktail party problem (Shinn-Cunningham, 2000). However, neuronal plasticity is enhanced in critical periods, such as during early postnatal years, and less enhanced during old age when HL is more commonly reported (Hübener & Bonhoeffer, 2014). Alternatively, deficits could be attributed to altered cognition and attention (Füllgrabe, Moore, & Stone, 2014).

1.8. Summary and Aims

A summary of the literature, as well as a declaration of the thesis' aims (in bullet points), is provided separately for each of the two topics motivated in Section 1.1.

1.8.1. Motivation (i): "how humans identify speech masked by a competing speaker"

Progress in understanding how listeners recognise speech masked by a competing speaker has largely occurred within the confines of ASA. A fundamental idea of ASA is that the brain segregates auditory scenes, and this is done on the basis of auditory cues. Different descriptions of auditory segregation (e.g. qualitative explanations, computational models, psychoacoustic tests), predicated on different auditory cues, have been tested with varying degrees of success. Moreover, no comprehensive testable framework successfully unifies these ideas. Ultimately it cannot be concluded that how listeners solve the cocktail party problem is understood.

These limitations are well demonstrated in CVI, arguably the most widely modelled instance of the cocktail party problem (over 35 years old), yet one of the simplest. Foremost, testable models of CVI cannot quantitatively predict listeners' behaviour. In fact, many models do not even make probabilistic predictions. The auditory cues proposed to drive segregation (pitch or temporal beating) have significant confounds. Further, when there are no F0 differences to facilitate segregation the models fail. Unsegregated representations of concurrent-vowel pairs are independently compared with representations of individual vowels. Attempts to investigate the neural basis of CVI have confirmed that there is sufficient information in the auditory nervous system to facilitate CVI. However, this has not been quantitatively tested due to the limited number of stimuli presented, and amount of data collected, in neurophysiological studies.

Additionally, the means by which this information would be extracted is speculative.

It seems that an understanding of CVI, and possibly other instances of the cocktail party problem, may be hindered due to the application of bottom-up segregation heuristics motivated by ASA. It is not that the concept of segregation is necessarily incorrect, but it may at present be a limiting factor in converging on an understanding of the cocktail party problem. Bregman (1981) has himself promoted the concept of asking what a system's purpose is, as opposed to its finer implementation details. To this end, when solving the cocktail party problem, the purpose of the auditory system is to correctly identify a conversation partner's speech with the highest probability. This could be achieved by correctly predicting the representation of a partner's speech in combination with any interfering speech sounds. This approach does not make reference to traditional ASA concepts. Conversely, a listener could attempt to identify their partner's speech ignorant to the existence of any interference. Somewhat motivated by the contemporary success of prediction theory in wider science (Kersten & Yuille, 2003; Maloney & Zhang, 2010), these ideas represent a principled non-heuristic way to confine the theoretical space within which the identification of overlapping speech occurs.

In this thesis I aim to:

- Construct a predictive framework to investigate the identification of overlapping speech sounds.
- Apply this framework to understanding how listeners identify concurrently presented speech tokens (vowels/VCs), and the neural coding supporting this.

1.8.2. Motivation (ii): "changes to identification with hearing loss"

The effects of HL on the processing of overlapping speech are highly interrelated with auditory processing in unimpaired listeners. Hence, the literature already discussed with regards to motivation (i) also applies here.

The best understood impact of hearing impairment is in the cochlea. Changes to its nonlinear processing are thought to reduce the amount of useful information in the resulting representation of masked speech. Aside from cochlear damage, recent work has demonstrated the existence of limitations to AN coding not visible audiometrically (hidden HL; Schaette and McAlpine, 2011). Changes to neural dynamics on a population level, temporal coding, and cortical organisation, are a few ways in which HL could compromise the processing of overlapping speech.

The relationship between damage to the cochlea following HL and changes to afferent neural processes remains understudied. Yet, together, these comorbidities likely result in problems amongst complex auditory settings. How these comorbidities affect the neural coding of speech amongst non-stationary interferers (i.e. a sound whose stochastic properties vary over time, unlike white noise for example) is an especially challenging aspect. The application of a framework like that suggested in 1.8(i) that can optimally account for the complex interaction of overlapping speech sounds, may be a good approach. An initial confirmation of this approach would be to determine that speech sounds masked by interfering speech sounds are more poorly represented following cochlear damage.

In this thesis I aim to:

- Establish the extent to which the neural encoding of speech sounds masked by non-stationary speech sounds is worse for an auditory system with HL, relative to NH.

1.9. Overview of Thesis

Chapter 2 proposes and details a computational modelling framework designed to predict the perception of auditory stimuli. This is detailed subsequent to the description of some influential theories of decision-making.

In Chapter 3, the modelling framework of Chapter 2 is used to predict human behaviour in a basic CVI task. The model is configured to identify vowels independent of one another, or to predict the combined representation of concurrently presented vowels. Rate-place and temporal coding are also examined.

Chapter 4 describes neurophysiology experiments using a full set of CVI experimental stimuli, recorded from the IC of anaethetised guinea pigs. The modelling framework of Chapter 2 is applied to this data to predict the behaviour of human listeners in a CVI task. The features of the neural responses determining the model's predictions are also investigated.

Chapter 5 addresses how the neural representation of overlapping speech sounds changes with HL. To quantify this, neural data gathered from animals with either NH or a noise induced HL is integrated with the model proposed in Chapter 2.

Chapter 6 collates the ideas investigated within the thesis, offering a broad discussion of the findings and suggestions for future work.

Chapter 2 Modelling Auditory Perception

2.1. Introduction

This chapter deals with theories and models of decision-making that aid the interpretation of perceptual data produced amongst uncertainty. Psychoacoustics relates physical sounds to the behaviour of listeners, in order to understand the processes that underlie perception. Audition, like many aspects of our internal and external worlds, is fundamentally uncertain. This results in the production of perceptual errors in auditory scenes. For example, in a clinical pure tone audiogram assessment, the misidentification of a pure tone when no pure tone is presented is a common occurrence.

In the simplest of psychoacoustic experiments a participant may be presented either a stimulus classed as a signal (c_{sig}) or noise (c_{noi}), where the latter indicates the absence of the signal. The participant would then be asked to report whether they had been presented a signal (r_{sig}) or noise (r_{noi}). A grid of responses (as in Table 2.1), also known as a confusion matrix, would be obtained denoting the participant's proportion of hits, misses, false alarms, and correct rejections. This data would then be used to determine the validity of a researcher's original hypothesis regarding perceptual processes. Classically, a researcher may want to identify the lowest value of a stimulus for which it can be detected or alternatively the smallest detectable change in a stimulus.

		Response	
		Signal (r_{sig})	*Noise* (r_{noi})
Stimulus -class	*Signal* (c_{sig})	Hit $P(r_{sig}\|c_{sig})$	Miss $P(r_{noi}\|c_{sig})$
	Noise (c_{noi})	False alarm $P(r_{sig}\|c_{noi})$	Correct rejection $P(r_{noi}\|c_{noi})$

Table 2.1 A 2x2 confusion matrix. A signal or noise is presented (rows), and the presence of the signal is either confirmed or denied (columns). Each decision type is labelled along with the associated probabilistic notation in grey.

Uncertainty is inherent in data resulting from such psychoacoustic experimentation and approaches to interpreting this are widespread. Researchers often attempt to characterise the decision space, connecting stimuli to behaviour, with some sort of descriptive measure. In the simplest case this descriptive measure may be the proportion of hits made. However, this would ignore all presentations of the noise stimulus-class, an entire row of a confusion matrix. To counteract this, one could simply record the proportion of correct identifications that are made. This is normally achieved by taking the average of hits and correct rejections, assuming an equal presentation rate. Alternatively, accounting for the relative proportion of hits made in comparison to false alarms encapsulates all of the information present in a 2x2 confusion matrix because the rows in a confusion matrix sum to 1. These two decision types, hits and false alarms, have consequently been used to formulate basic measures underpinning entire theories of decision-making.

In this chapter, a model of decision-making is proposed in which several key ideas from existing theories of decision-making amongst uncertainty have been employed. In combination with computational advancements, this proposed model is able to predict the perception of non-trivial auditory stimuli making it well suited for investigating the cocktail party problem. However, rather than

embodying notions of segregation the framework supports modelling the purpose of a system (e.g. optimality) as proposed in Section 1.8.

This chapter is comprised of the following content:

- As arguably the most popular theory applied to psychoacoustic experimentation, signal detection theory (SDT) is summarised along with its origins within statistical decision theory (Sections 2.2 & 2.3).
- Following this is a related discussion of Bayesian decision theory (Section 2.4).
- This leads in to a description of the model of decision-making developed and applied throughout this thesis (Section 2.5).

2.2. Statistical Decision Theory

Prior to the publication of Green and Swets' (1966) famous book titled "Signal detection theory and psychophysics", a significant amount of work was done to establish the statistical foundations of decision-making. Termed "Statistical decision theory", the theory posits how to optimally make decisions given uncertain evidence (Wald, 1950).

In this definition, the word evidence can be thought as an abstract description for a quantity that is informative to hypotheses about the state of the world. For example, a high frequency "ding" could be evidence of a bicycle wanting to pass, or a microwave. Statistical decision theory promotes calculating the probability with which, given a hypothesis about the state of the world, evidence would be observed. This probability is calculated using a likelihood estimate. A statistically optimal decision is specified to be the selection of the hypothesis about the world that has the largest associated likelihood estimate. Note how this approach ignores how probable the hypotheses themselves are, independent of evidence.

2.3. Signal Detection Theory

SDT is similar to statistical decision theory in that it is also a model of decision-making in the context of uncertain evidence. It was originally developed to analyse the detection of an auditory signal amongst white noise, conceptualised with the goal of "unifying a considerable area of psychoacoustics within the statistical-decision framework" (Licklider, 1964, p. 119). Pertaining to domains other than audition, Macmillan & Creelman (2004, p. 25) have defined SDT to be a "theory relating choice behaviour to a psychological decision space".

2.3.1. Gaussian Signal Detection Theory

By far the variation of SDT that is most widely known, widely used, and in many textbooks directly synonymous to SDT itself, is the Gaussian equal variance model denoted Gaussian signal detection theory (G-SDT). This model assumes that there are two likelihood distributions describing the signal and noise that are both Gaussian with equal variance. Accordingly, descriptive statistics can be measured. Measures look at two aspects of observers. The first characterises the distances between likelihood distributions, determining the limits of an observer's performance (sensitivity). The second measures how certain decisions are favoured independent of an observer's sensitivity (bias). Previous methods (e.g. early statistical decision theory) only considered the optimal application of likelihood functions and hence did not adequately distinguish between these entities.

2.3.1.1. Sensitivity

An observer that makes hits 100% of the time and correct rejections 100% of the time would be described as having perfect sensitivity. On the contrary, a completely insensitive observer would make the same proportion of hits as they do false alarms. Sensitivity measures aim to formalise this intuitive notion, and the region between these two behavioural extremes.

To reiterate, G-SDT's decision space is conceptualised as two distributions describing the likelihood of evidence, given a signal or a noise. It is the mean and variance parameters of these distributions that set a limit on the maximum achievable performance, and hence determine sensitivity. G-SDT's measure of sensitivity is denoted d' (Green & Swets, 1966) and is defined as the standardized difference between the means of the signal and noise likelihood distributions (see Figure 2.1a). Due to the assumptions implicit in G-SDT, this is equivalent to the difference between the z-score (cumulative Gaussian distribution) of hits and false alarms. This is mathematically written as:

$$\begin{aligned} d' &= \frac{\mu_{sig} - \mu_{noi}}{\sigma} \\ &= z\left(P\left(r_{sig}\middle|c_{sig}\right)\right) - z\left(P\left(r_{sig}\middle|c_{noi}\right)\right) \quad , \end{aligned} \tag{2.1}$$

where μ_{sig} and μ_{noi} are the expected values of the signal and noise Gaussian likelihood distributions, σ is standard deviation, z represents a z-score, and r_{sig}, c_{sig}, and c_{noi}, are as in Table 2.1.

d' is an infinite measure because Gaussian distributions are defined over an infinite range. To avoid infinite measurements, d' is commonly set to have a maximum value of 4.65 (equivalent to $P\left(r_{sig}\middle|c_{noi}\right)$=0.01 and $P\left(r_{sig}\middle|c_{sig}\right)$=0.99).

2.3.1.2. Bias

A motivating property of the d' measure is that it is invariant to factors other than sensitivity. Subsequently, d' can have the same value for two confusion matrices with different numerical entries. The aspect that distinguishes such confusion matrices as different is bias. Bias is a favouring of one hypothesis over another. This behaviour is hypothesised to occur from the interaction of a complex set of assumptions an observer makes about the world. If an individual has no bias then their decisions are statistically optimal in that they maximise the proportion of correct decisions made.

In the G-SDT space (see Figure 2.1a), bias can be characterised with a vertical line positioned over a criterion point. When evidence accumulates beyond this point one hypothesis is selected, and when below, the alternate hypothesis is selected. Aside from when the criterion point is exactly midway between the two likelihood decisions, the criterion is biased. Termed the C-criterion, this commonly applied measure of bias is calculated as:

$$C = -\frac{\left(z\left(P\left(r_{sig} \mid c_{sig}\right)\right) + z\left(P\left(r_{sig} \mid c_{noi}\right)\right)\right)}{2} \quad . \tag{2.2}$$

If the assumptions of G-SDT hold then the C-criterion is orthogonal to d' (Ingham, 1970).

2.3.1.3. Assumptions

The assumptions made in G-SDT have a number of rationalisations:

- The central limit theorem explains that, under certain conditions, the average of a sufficiently large number of independent random variables is normally distributed (Clapham & Nicholson, 2009). Given the large number of processes linking physical stimuli to behaviour this is not an unreasonable assumption. Importantly for this thesis, the Gaussian distribution is a limit of multiple Poisson distributions, a commonly employed distribution to mimic neural dynamics (Teich & Khanna, 1985).
- The Gaussian distribution is theoretically convenient. Random variables can be defined using few parameters, and results, such as the calculation of likelihood ratios (refer to section titled "Decision Variable") and statistically optimal decisions (Fukunaga, 2013), become analytically tractable.
- Implementation of Gaussian assumptions in the context of SDT is no more, nor less, important than when considering the application of commonly employed parametric statistical tests such as student T test and chi-squared tests.

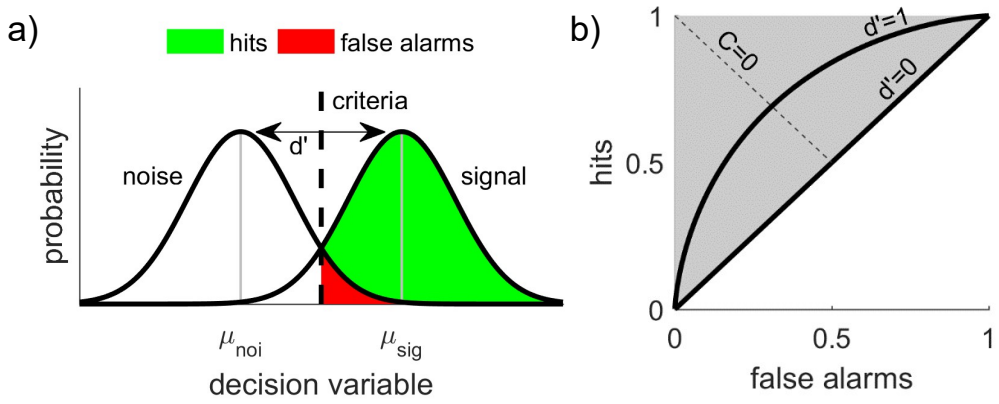


Figure 2.1 (a) A Gaussian equal-variance decision space (as in G-SDT). The horizontal dashed line represents the decision criterion point, and on the signal side of this the hits and false alarms are highlighted in green and red. (b) ROC space with two curves representing $d' = 0$ (complete insensitivity) and $d' = 1$. The diagonal dashed line represents decisions made with an unbiased criterion.

2.3.2. Parametric and Non-parametric Variants

2.3.2.1. Decision Variable

In its purest form SDT is unconstrained by Gaussianity. Instead SDT is predicated on the concept of a decision variable, a quantity used to inform a decision-maker. Stemming from ideas in statistical decision theory, when presented with a task offering a binary decision, one can calculate likelihood estimators for the two available hypothetical decisions. Formally, any monotonic function of the ratio between the two likelihood estimates, termed the likelihood ratio, is the decision variable of interest in SDT (Swets, Tanner Jr, & Birdsall, 1961). By definition, an increase in a decision variable's value favours the association of evidence with one particular hypothesis. A decrease in the decision variable's value describes evidence shifting to favouring that hypothesis less, but instead the alternative. An observer uses this variable to select a point where for values that are larger than this point one hypothesis is selected, and when less than the point the alternative is selected. There is a point along the decision variable where neither decision is favoured by the observer. This would specify a statistically optimal decision point.

Any alternative decision point is inherently biased. The independence of the location of this decision point with the structure and dynamics of the decision variable is a key aspect of SDT. Under the assumptions of G-SDT this decision point is the criterion point.

The decision variable is often confused as to merely be describing a physical or sensory dimension (Schiffman, 1990). This is possibly because SDT was designed for auditory signal detection in white noise, and the variable is mistaken to have exactly represented the stimulus statistics. Consequently it has been argued that SDT suffers to account for stimuli described by multi-dimensional stimuli (Norris, 1995). However, this is wrong. SDT's decision variable is clearly defined, and was conceptually intended to represent an abstract cognitive and decision space. SDT's decision variable, likelihood ratio, is a unidimensional measure which can reflect an evaluation of evidence in many dimensions.

2.3.2.2. Receiver Operating Characteristic

Prior to SDT, insight into decision-making was obtained by plotting an individual's proportions of false alarms along the ordinate and hits on the abscissa of a graph (Woodward, 1953). A point on a receiver operating characteristic (ROC) graph can therefore illustratively capture the entirety of the information in Table 2.1. Perfect discrimination would be represented by a point in the top left of the graph ([0,1]). The closer a point is to the top left, the more accurate an observer's response. In turn, the G-SDT conceptions of sensitivity and bias can be captured by such a graph. Highlighted in Figure 2.1b are ROC curves which differ in their d' values, calculated by solving Eq. 2.1. A ROC curve with a d' of 0 is shown where the propotion of hits matches false alarms, and an ROC curve with a d' of 1 is also shown for comparison.

Many researchers are concerned about the violation of Gaussian and equal variance assumptions when applying G-SDT measures. If the assumptions of G-SDT have been deemed as inappropriate, published articles have been content to

calculate sensitivity using the scaled mean difference of non-Gaussian distributions (Jones, 2016). Alternatively, specific formulations of SDT have been developed for non-Gaussian parametric distributions. For example, issues connected to the non-monotonicity of non-equal variance Gaussian likelihood ratios has led to the construction of Poisson based SDT variations, predicated on replicating asymmetric ROC curves (Egan, 1975; Kaernbach, 1991).

Further, nonparametric SDT statistics have been developed solely based on the ROC curve. One example of a nonparametric statistic is area under ROC curves (AUC). Green and Swets (1966) explain how the AUC makes no parametric assumptions about any underlying likelihood distributions. However, the calculation of AUC requires at least two points in ROC space. A' is an example of a nonparametric measure calculated from an individual point in ROC space (Pastore, Crawley, Berens, & Skelly, 2003). A' determines an estimate of area based on an individual point, however this necessitates stronger assumptions about shape and therefore implicit distributional properties.

2.4. Bayesian Decision Theory

The Bayesian theory of decision-making builds on many of the ideas underpinning statistical decision theory and SDT. Some have gone as far to endorse Bayesian decision theory as the only principled solution for making decisions amongst uncertainty (Vincent, 2015). Whereas statistical decision theory is predicated on calculating and comparing likelihood estimates, Bayesian decision theory also takes into account prior information about the state of the world, as well as costs associated with decisions. Like SDT in its most general form, Bayesian decision theory makes no assumptions as to the parametric properties of uncertainty.

2.4.1. Bayes Theorem

In Section 2.2, the concept of a likelihood estimate was introduced. This estimator allows one to calculate the likelihood of evidence if a particular hypothesis about the world were true. However, the question which is asked of perceptual entities is the inverse of this. Given some evidence, what is the probability a hypothesis about the state of the world is true? To answer this question in a principled manner requires the implementation of Bayes' theorem (Bayes & Price, 1763).

To gain an intuitive understanding of the logic that motivates Bayes' theorem, imagine you have tuned into an unknown radio station. You are only 50% sure that the radio host says the word "basketball". If you were then made aware that you are in fact tuned into a sports radio station you would be much more confident that the word spoken was "basketball", compared to if the radio station was of a classical music format. Bayes theorem formalises this type of inference. It has been applied to a wide variety of situations, even attempts to determine the existence of God (although not by the Reverend Thomas Bayes himself; Swinburne, 2012).

Bayes theorem advocates calculating the probability with which a hypothesis produces evidence, whilst taking into account the prior probability of that

hypothesis being true. This, scaled by the relatively unimportant measure of how frequently the set of evidence is observed, solves the inverse problem. Bayes' theorem is written:

$$P(h|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|h)P(h)}{P(\boldsymbol{x})} \quad , \tag{2.3}$$

where $\boldsymbol{x}$ is a vector that is an empirical measure of evidence, h is a hypothesis about the state of the world, and $P(h|\boldsymbol{x})$ is referred to as a posterior probability.

2.4.2. Decision Rules

To best predict reality is to select hypotheses about the state of the world in a manner that produces the least errors. This has proven to be achieved by selecting the hypothesis with the highest posterior probability for unfamiliar evidence, not dissimilar to the maximum likelihood approach of statistical decision theory. This strategy is called the maximum a posteriori (MAP) decision rule and is formally written as:

$$\hat{\theta}_{MAP}(\boldsymbol{x}) = \underset{\theta}{\operatorname{argmax}}\, P(h_\theta|\boldsymbol{x}) \quad , \tag{2.4}$$

where $\hat{\theta}_{MAP}(\boldsymbol{x})$ identifies the hypothesis with the MAP estimate.

Although the MAP decision rule optimally reduces the total proportion of errors made, it does not necessarily take into account the cost of making correct and incorrect decisions. A decision rule that incorporates a loss function that equates a decision with a corresponding value is preferable. To this end, Bayesian decision theory promotes minimising a posterior expected loss function. Denoted Bayes' rule, this is mathematically written as:

$$\hat{\theta}_B(\boldsymbol{x}) = \underset{\theta}{\operatorname{argmin}} \sum\nolimits_{\alpha} \lambda(h_\theta|h_\alpha)P(h_\alpha|\boldsymbol{x}) \quad . \tag{2.5}$$

$\lambda(h_\theta|h_\alpha)$ is the loss function describing the numerical loss that results from the selection of the hypothesis h_θ when the true state of the world was the hypothesis h_α.

As an example, consider a decision between two hypotheses where the likelihood of hypothesis 1 is assumed to be 40%. Given there is no prior knowledge favouring either hypothesis, the MAP decision rule in Eq. 2.4 would recommend selecting hypothesis 2 for which the observer has attributed a probability of 60%. Now consider that incorrectly selecting hypothesis 1 results in a loss of £5 ($\lambda(r_1|c_2) = £5$), incorrectly selecting hypothesis 2 loses £20 ($\lambda(r_2|c_1) = £20$), and a correct decision loses £0 ($\lambda(r_1|c_1) = \lambda(r_2|c_2) = £0$). Selecting hypothesis 1 would result in an expected loss of £3 ($£0 \times 0.4 + £5 \times 0.6$), as opposed to £5 ($£20 \times 0.4 + £0 \times 0.6$) when selecting hypothesis 2. Therefore, using Bayes' rule, the decision that minimises expected loss is hypothesis 1.

Statistical decision theory, SDT, and Bayesian decision theory are unified in their application of likelihood estimates. It is both the application of prior knowledge and the attribution of value to decisions that separates Bayesian decision theory from statistical decision theory. Likewise, it is the notion of bias that separates SDT from statistical decision theory. If all hypotheses are equal in the loss they incur (in which case Bayes' rule simplifies to the MAP decision rule) and have equivalent prior expectations, this is the same as a SDT model without any bias, which is equivalent to the statistically optimal calculation of statistical decision theory (see Figure 2.2 for schematic overview). It may then not come as a surprise that SDT's notion of bias implicitly embodies information about prior expectation and costs. Thus, although in SDT experiments bias measures may be deemed to be statistically suboptimal, it may be optimal within the framework of Bayesian decision theory. It seems pertinent to acknowledge the practical ability of machine learning (ML) algorithms, such as deep neural nets (LeCun, Bengio, & Hinton, 2015), to optimally make decisions. However, compared to the reviewed theories of decision-making, ML algorithms are less transparent, less able to elucidate the principles underlying their behaviours.

2.5. Proposed Modelling Framework

The theories discussed in this chapter are examples that have proven successful in efforts to understand the space that lies between stimuli and perceptual decisions. SDT has traditionally been favoured in the psychoacoustics literature due to its numerical simplicity, yet its application has limited scope in experiments utilising increasingly complex stimuli and paradigms. Following advancements in computational power, research using more progressive frameworks is achievable (Kersten & Yuille, 2003; Körding & Wolpert, 2006; Maloney & Zhang, 2010). These advancements allow predictions of perceptual behaviour to be made directly from sensory representations, or even neural representations recorded from the brain (Steadman, 2015). With this in mind, I present a computationally tractable model of auditory perception. The model amalgamates some of the simplifying assumptions of G-SDT with theoretical concepts underpinning statistical, and Bayesian, decision theory.

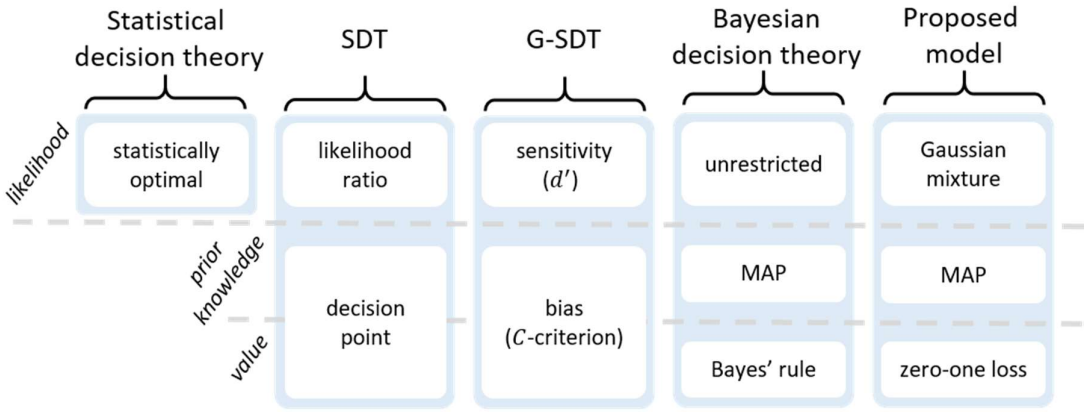


Figure 2.2 Diagram highlighting the key components of the decision theories introduced in this chapter, as well as the proposed model.

2.5.1. Creating Likelihood Estimates

The first stage of the proposed framework is to create likelihood estimators. To reiterate, a likelihood estimator is a distribution describing the probability of observing evidence given a hypothesis about the state of the world. In the proposed framework, this is constructed using a generative model describing how an auditory representation of a stimulus occurs. This can be mathematically written as:

$$P(\boldsymbol{x}|s) = f(s) \quad , \tag{2.6}$$

where $f()$ embodies the process that occurs between the presentation of the stimulus, s, and the production of some evidence, $\boldsymbol{x}$. The form of Eq. 2.6 is purposefully liberal and could, for example, be a computer simulation or a real representation recorded from a brain.

In this thesis, all generative models will yield Gaussian distributions. The assumption of Gaussianity is inspired by G-SDT, and is accompanied by a number of favourable properties (see "2.3.1.3. Assumptions" in Section 2.3.1). Additionally, likelihood estimators describing classes of stimuli (termed class-conditional likelihood estimators) will be Gaussian mixtures. This reflects the fact that stimulus-classes are often comprised of multiple different stimuli. For example, the vowel /i/ could be presented at 6 different F0s, yet for each vowel stimulus the stimulus-class of /i/ remains the same. A class-conditional likelihood estimator is calculated as:

$$P(\boldsymbol{x}|c) = \sum_{i=1}^{S} P(\boldsymbol{x}|s_i)P(s_i) \quad , \tag{2.7}$$

where s_i is one of S different stimuli belonging to the stimulus-class c, and $P(s_i)$ is the prior probability of s_i.

2.5.2. Experimental Assumptions

The application of Bayes' theorem (Eq. 2.3) converts class-conditional likelihood estimators into posterior probability distributions, or posterior estimators. This conversion requires the estimation and application of prior knowledge regarding how often a stimulus-class is presented. In a laboratory setting it is often the case that each class has an equivalent prior presentation probability, however factors such as previous exposure to a certain stimulus-class during training may require careful consideration of prior probabilities. In this thesis it will be assumed that prior to participation in a psychoacoustic experiment, participants will be unfamiliar with the experimental stimuli.

Additionally, in order to reduce the number of hyper-parameters and speculative assumptions made, this thesis will assume a zero-one loss function. This implies that there is no loss for a correct identification (zero) and all errors are equal in the loss they incur (one). Consequently, Bayes' decision rule (Eq. 2.5) simplifies to be equivalent to the MAP decision rule (Eq. 2.4).

2.5.3. Making Predictions

When attempting to model and understand perception it is insightful to consider sources of error. Errors in a binary task have already been introduced in Table 2.1 as misses and false alarms. However, when there are multiple stimulus-classes (>2), more classes are incorrect than correct. Rather than simply signal and noise stimulus -classes and response options (i.e. $\{c_{sig}, c_{noi}\}$ and $\{r_{sig}, r_{noi}\}$; see Table 2.1), experimental stimuli will be described as presented from one of the classes in the set $\{c_1, c_2, \dots\}$ and identified from one of the classes in the set $\{r_1, r_2, \dots\}$. It is common for these sets to be equivalent.

Using the MAP decision rule, the proposed framework predicts that the stimulus-class c_i is identified as r_j with the probability:

$$P(r_j|c_i) = \int_{R_j=\{\boldsymbol{x}|\widehat{\theta}_{MAP}(\boldsymbol{x})=j\}} P(\boldsymbol{x}|c_i)\, d\boldsymbol{x} \quad . \tag{2.8}$$

R_j represents evidence, $\boldsymbol{x}$, for which the stimulus-class j has the MAP estimate. The calculation of Eq. 2.8 (conceptualised in Figure 2.3), across all combinations of classes presented and identified, produces a confusion matrix directly comparable with experimental participants.

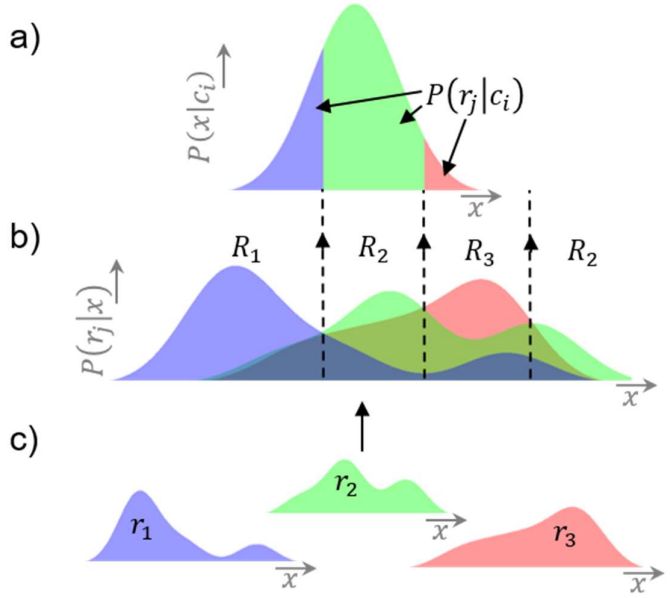


Figure 2.3 Visual representation of Eq. 2.8. (a) The class-conditional likelihood distribution of a stimulus-class, c_i, is selectively integrated over. These integrations predict the probability of c_i being identified as a stimulus-class r_j ($P(r_j|c_i)$; i.e. entries of a confusion matrix). (b) Integrations are performed over regions, R_j, where a stimulus-class, r_j, has a posterior distribution with the MAP estimate (decision rule 2.5). (c) Posterior distributions for three different response options, r_1, r_2, and r_3.

2.5.4. Computation

Although modestly stated, Eq. 2.8 is non-trivial to calculate. The complexity of the calculation heavily depends on the form of $f(\,)$ and the experimental paradigm that is being modelled. If the set of classes presented match the set of classes which can be identified, and are equal-variance Gaussian, there is a pseudo-analytical approach to solving Eq. 2.8 (Fukunaga & Olsen, 1970). However, due to the fact that stimulus -classes are constructed by averaging over a mixture of distributions (Eq. 2.7), this analytical approach is seldom achievable. In general, the calculation is best performed numerically using Monte Carlo simulations. This numerical approach produces evidence using the likelihood estimator for a given stimulus-class, e.g. c_i. The decision rule is then applied to classify this evidence. When repeated a very large number of times, the proportion of instances that evidence is identified as stimulus-class r_j tends to $P(r_j|c_i)$.

The Monte Carlo calculation of Eq. 2.8 is greatly increased in speed as all likelihood estimators are Gaussian. Additionally, it is useful to assume that all dimensions in the class-conditional likelihood estimators are independent of one another. This assumption has proven successful in the ML literature for both dependent and independent data sets (Zhang, 2004). This assumption directly benefits computational speed.

One further computational convenience is the option to apply principal component analysis (PCA) to a dataset as a pre-processing step. The assumption of independence is better supported following PCA due to the transformation producing sets of linearly uncorrelated variables. Fan and Poh (2007) have previously shown that this can benefit the performance of a naïve Bayes classifier. Additionally, if the number of stimulus -classes are less than the number of features, this can be favourable to the dimensionality of the dataset, making computations faster with minimal loss of information.

2.5.5. Optimal (and Suboptimal) Predictions

The proposed modelling framework can be understood as an ideal observer that makes the best decisions possible, necessitated by the task constraints and information available for the model to utilise (Geisler, 2003). It is by varying the information available to the model which permits interrogation of how the auditory system identifies speech in tasks where interfering sounds are present.

In Section 1.8.1, it was proposed that to best identify a conversation partner's speech an auditory system might predict the representation of a conversation partner's speech in combination with interfering sounds. Hereafter, this approach will be referred to as an optimal scheme. The word optimal is selected because such a scheme, when instigated perfectly, would result in a 100% identification performance (assuming the system is absent of noise). To implement this within the proposed modelling framework is a matter of storing class-conditional likelihood estimators that describe the combined representations of target speech and the interfering speech. Given a representation of target speech, $target$, and its representation amongst interfering speech, w, the probability with which $target$ is identified from the presentation of w under an optimal scheme is:

$$P(target|w) = \int_{\boldsymbol{x}\in R_w} P(\boldsymbol{x}|w)\, d\boldsymbol{x} \quad , \tag{2.9}$$

where R_w is a region of evidence where w is the MAP estimator. In this instance, as variance (noise) in the likelihood estimators tends to 0, $P(target|w)$ tends to 1.

Alternatively, the auditory system may perform speech identification ignorant to the possible interference of overlapping sounds. This is reminiscent of existing models of CVI which, when vowels have identical F0s, compare unsegregated representations of concurrent-vowel pairs with individual vowels. Only if speech identification is performed in a quiet setting, free of any interfering sounds, will such a strategy default to the optimal scheme. Otherwise, identification will never consistently be 100%. Thus, as this thesis is predominantly interested in the identification of overlapping speech sounds, this approach will hereafter be

referred to as a suboptimal scheme. Along these lines, the proposed framework configured suboptimally would store only class-conditional likelihood estimators of target speech in quiet. Under a suboptimal scheme, the probability of identifying w as $target$ would be calculated as:

$$P(target|w) = \int_{\boldsymbol{x} \in R_{target}} P(\boldsymbol{x}|w)\, d\boldsymbol{x} \quad , \tag{2.10}$$

where R_{target} is a region of evidence where $target$ is the MAP estimator.

2.6. Discussion and Conclusion

In this chapter, three prominent theories of decision-making have been surveyed (although it should be noted that there is a vast literature beyond that which has been presented). This culminated in a proposed model of decision-making that draws influence from these three theories. The initial aspect of the proposed model is the construction of likelihood estimators, as in statistical decision theory. The notions of prior information and loss that are inherent to SDT, but explicit in Bayesian decision theory, inspired the application of the model's decision rule. Finally, the proposed model, like G-SDT, advocates the use of Gaussian assumptions to make the calculation of predictions computationally tractable.

The proposed model has a number of positive attributes. The model makes a limited number of assumptions, yet outputs probabilistic (as opposed to deterministic) predictions directly comparable to listeners. Subsequently, the framework allows one to posit optimal and suboptimal approaches to decision-making. This approach encourages a principled approach to understanding the goal orientated aspects of perceptual processing, rather than only biological mechanisms, as recommended by Marr (1982). Further, although the proposed model does not run anywhere near the speed of the highly parallelized and efficient human brain, it is fast enough to simulate non-trivial psychoacoustic experiments. Simulating such experiments will be the first step in validating this approach to modelling auditory perception.

Chapter 3 Concurrent-vowel Identification: Auditory Filter Model

3.1. Introduction

Humans are able to understand speakers within complex auditory environments, solving the famously termed cocktail party problem (Cherry, 1953). Particular attention has been paid to the benefit that F0 differences have on the correct identification of concurrent-vowel pairs (Scheffers, 1983; Zwicker, 1984). This benefit occurs for F0 differences up to 1 semitone following which performance plateaus. In this simple instance of multi-speaker identification both overall performance and the benefit of F0 differences degrade with HL and age (Arehart et al., 2011; Chintanpalli, Ahlstrom, & Dubno, 2016; Summers & Leek, 1998).

Computational models have been developed to explain the relationship between auditory processing and human performance in CVI tasks. Despite this being an open research area for more than 35 years, models have not yet been able to successfully predict human perception. Meddis and Hewitt's (1992) widely cited model does not quantitatively match human performance. When there are no F0 cues available, this model separately compares unsegregated representations of concurrent-vowel pairs with representations of individual vowels. In fact, when there are no F0 differences to facilitate segregation, all existing models of CVI succumb to this comparison.

Additionally, existing CVI models are unable to predict the specific decisions made by listeners (Chintanpalli & Heinz, 2013). This is often a consequence of deterministic predictions (i.e. the identification of concurrent-vowel pairs 100% or 0% of the time). Models that do incorporate probabilistic predictions assume vowels are identified independently of one another (Assmann & Summerfield,

1989; Culling & Darwin, 1994), occasionally with supplementary ad-hoc criteria (Assmann & Summerfield, 1990).

At present there is no consensus as to how, or even what, cues are used to solve the CVI task. Along with conflicting evidence regarding the utilisation of harmonic cues (Assmann & Paschall, 1998; Meddis & Hewitt, 1992) there is conflicting evidence regarding the utilisation of beating cues (Culling and Darwin, 1994; de Cheveigné, 1999). Additionally, there is no encompassing model that permits the influence of multiple kinds of cue.

From this brief recap (see Section 1.5 for a more detailed review) it is clear that how humans perform CVI is still very much unknown. Rather than continuing to investigate this instance of the cocktail party problem within the conceptual framework of segregation, it seems pertinent to apply a stripped back model of ASA that more broadly questions the goals of the auditory system.

This chapter applies the modelling framework of Chapter 2 to the task of CVI:

- Quantitative predictions are made using the model proposed in Chapter 2 under a suboptimal scheme (Section 3.3.1), and an optimal scheme that predicts the combined representation of concurrently presented vowel (Section 3.3.2). The latter is demonstrated to qualitatively and quantitatively predict listeners' behaviour (Section 3.3.1 and 3.3.2).
- The proposed model is shown to account for listeners' experimental training data as well as testing data (Section 3.3.3).
- To determine the auditory features influential to the model's predictions an information-theoretic analysis is applied (Section 3.3.4). In addition, the effect of the model's meta-parameters (e.g. temporal resolution and filter shapes) are probed (Section 3.3.5).
- Lastly, concepts prominent in existing CVI literature are used to further scrutinise the proposed model (Section 3.3.6 and 3.3.7).

3.2. Methods

3.2.1. Stimuli

Five synthetic vowels (/i/, /a/, /u/, /æ/, /ɜ˞/; steady-state harmonic complexes) were created using a Klatt-synthesizer (Klatt, 1980). The formant frequencies and bandwidths matched those specified by Chintanpalli and Heinz (2013). The vowels were synthesised at 6 different F0s: 100 Hz, 101.5 Hz, 103 Hz, 112 Hz, or 126 Hz. All vowels had a duration of 400 ms (including 10ms onset/offset raised cosine ramps) and were set to 65 dB SPL. The vowel identities, F0s, duration, and sound level were chosen as to align with Chintanpalli and Heinz's (2013) study.

A concurrent-vowel pair's waveform was constructed by adding together the waveforms of two vowels. There were 25 pairwise concurrent-vowel permutations, symbolically denoted $/v_i, v_j/$ where $v_i, v_j \in \{v_1 = \text{i}, v_2 = \text{a}, v_3 = \text{u}, v_4 = \text{æ}, v_5 = \text{ɜ˞}\}$. For each permutation, the F0 of v_i was set to 100 Hz and v_j had an F0 of k Hz, where $k \in K$ and $K = \{100,101.5,103,106,112,126\}$. 25 permutations of vowels presented at 6 F0 differences (0, 0.25, 0.5, 1, 2, and 4 semitones) meant there was a total of 150 separate concurrent-vowel pairs.

Additionally, a corresponding set of interleaved vowel stimuli were prepared. Each vowel was synthesised at two F0s. Harmonics were alternatively selected and combined from each F0. This resulted in an interleaved vowel having the same spectral ENV as a normal vowel, but a harmonic structure embodying two F0s.

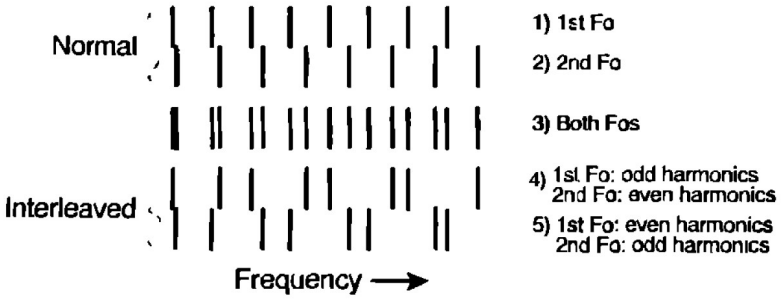


Figure 3.1 (from Culling and Darwin (1994)) The harmonic structure of a normal concurrent-vowel pair, and the equivalent interleaved concurrent-vowel pair.

3.2.2. Task

The psychoacoustic data was collected by Chintanpalli and Heinz (2013), who agreed to share their raw data (and have collaborated as far as contributing comments on posters and publications). The task had a training and a testing component.

Training: Prior to the testing stage, five NH subjects were successfully trained (with feedback) to identify individual vowels with above 90% accuracy. Immediately following, subjects were trained (with feedback) to correctly identify concurrent-vowel pairs that had F0 differences of either 0, 1, or 4 semitones (75 pairs in total) until a gradual increase in identification was exhibited as a variable of F0 difference. A correct identification was defined as correctly selecting the combination of vowels that had been presented (from a 5x5 grid). A concurrent-vowel pair permutation, at a given F0 difference, was presented a maximum of 12 times during training.

Testing: Following 2-4 hours of training, each subject was tested (with feedback) on their ability to identify all 25 concurrent-vowels pair permutations, at all 6 F0 differences, 12 times. Again, a correct identification was defined as correctly selecting the combination of vowels that had been presented.

3.2.3. Computational Model

A computational model (framework introduced in Section 2.5) was used to predict human behaviour at the CVI task (3.2.2). The model was made up of two parts:

- An auditory front-end, a generative model, that determines the likelihood of observing a given representation of a stimulus (i.e. a likelihood estimator as in Eq. 2.7).
- A classification back-end that associated unlabelled representations (constructed using the auditory front-end) with a particular stimulus-class.

3.2.3.1. Auditory Front-end

A model of peripheral auditory processing was implemented. Rather than mimicking biophysical dynamics (Sumner, Lopez-Poveda, O'Mard, & Meddis, 2002; Zilany & Bruce, 2006, 2007), the model largely consisted of linear filters and simple nonlinearities. This was done in an attempt to aid interpretation of the model's predictions and to increase computing speed. Additionally, models with similar simplifications have been shown to make predictions in agreement with speech intelligibility data (Jørgensen and Dau, 2011). The proposed auditory front-end crudely simulated filtering in the auditory meatus and on the basilar membrane, nonlinear neural transduction, and adaptation and stochasticity along the auditory nerve.

Input to the auditory front-end were waveforms of auditory stimuli, s. The frequency response of the outer/middle ear was replicated. Coefficients for a cascade of bandpass filters were fit such that pure tones between 0.1 and 4 kHz, presented at the threshold of audibility (Fletcher & Munson, 1933), all had filter outputs within 5 dB of one other.

Basilar membrane displacement was modelled by passing the resulting signal, in parallel, to 100 Gammatone filters. The linear Gammatone filter was originally developed to replicate impulse response functions recorded from ANFs (de Boer

& de Jongh, 1978). Its name stems from the filter's functional form which combines an ENV of a Gamma distribution with a carrier tone. Gammatone filter bandwidths were related to the filter's centre frequency by functions determined psychophysically (Glasberg & Moore, 1990). The centre frequencies of the 100 Gammatone filters were logarithmically spaced between 0.1 and 4 kHz, encompassing the harmonic range of vowel stimuli.

To mimic the bending of hair cells in only one direction the output of each filter was half-wave rectified. This was followed by four 2nd order low-pass Butterworth filters. Cut-off frequencies were set at either 0, 10, 50, 100 (default), 500, or 4000 Hz. The resultant signals were then logarithmically transformed to imitate gain control in inner hair cell synapses resulting in nonlinear compression . The log-transformed signals were again low-pass filtered (again four 2nd order Butterworth) prior to downsampling (the Nyquist frequency was always at least 1.25× the value of the low-pass cut-off frequency).

There was an option to apply a pooled ACF (PACF) as specified in Meddis and Hewitt's (1992) model of CVI. In each channel, the expected value was calculated for a signal multiplied with a delayed copy of itself and an exponential decay. The decay was incorporated to reduce the influence of distant periodicities. The running ACF of a signal x, at time, t, at a lag L, was calculated as:

$$h(t, L) = \sum_{i=1}^{\infty} x(t - i\delta t)x(t - i\delta t - L)e^{\frac{-i\delta t}{\tau}} \quad , \qquad (3.1)$$

where δt was the sampling interval and τ a time constant (10 ms). Running ACFs were calculated for, and averaged across, all 100 channels to create the PACF.

If the PACF was not applied, for each stimulus, the outputs across the 100 channels were concatenated and standardised (mean of 0 and standard deviation of 1). If the PACF was applied, for each stimulus, the PACF was standardised (mean of 0 and standard deviation of 1). Analogous to the standardisation performed in Meddis and Hewitt's model (1992), this was done in an attempt to crudely simulate

adaptation and habituation in the neural system (Rabinowitz, Willmore, Schnupp, & King, 2011).

The previous sequence of processes, that occurred following the input of a stimulus' waveform (s), collectively notated as the function $a(\,)$, produced a vector:

$$\boldsymbol{\mu}_s = a(s) \quad . \tag{3.2}$$

Independent, Gaussian, zero-mean noise, with identical variance, was then added to each value in the vector $\boldsymbol{\mu}_s$. This resulted in a distribution:

$$P(\boldsymbol{x}|s) = N(\boldsymbol{\mu}_s, \sigma^2\boldsymbol{I}) \quad , \tag{3.3}$$

where $\boldsymbol{x}$ represents a feature vector (previously referred to as evidence), $\boldsymbol{I}$ is the identity matrix, and σ^2 is variance. Variance of the Gaussian noise was a free fitting parameter in the auditory front-end. Eq. 3.3 can be thought of as describing the relative likelihood of observing a feature vector given the presentation of the stimulus, s. The term given for this type of distribution is a likelihood estimator.

Some aspects of this auditory front-end simulation were varied for investigative purposes. These meta-parameters included Gammatone filter bandwidths, the cut-off frequency of low-pass filters, standardisation, and the inclusion of the PACF.

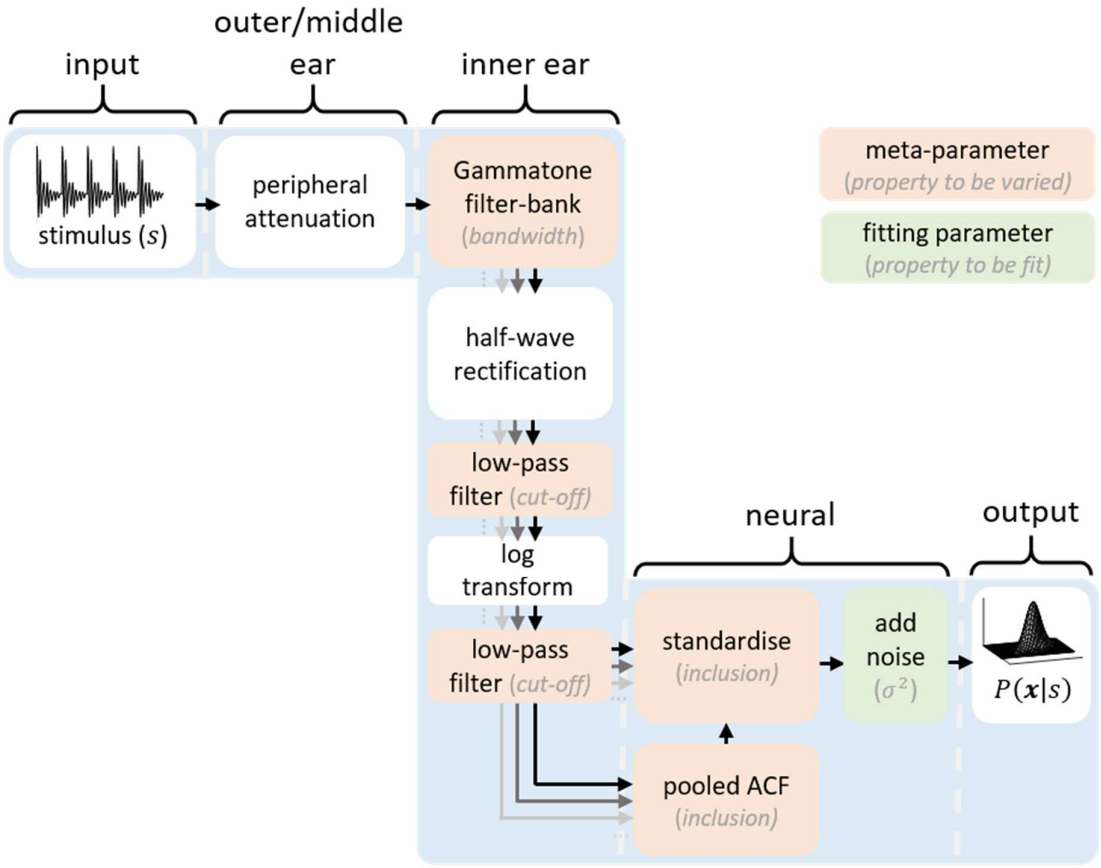


Figure 3.2 Schematic of the simulation of the auditory periphery used as the front end to the proposed classifier. Braces describe what part of the auditory system is characterised in the corresponding column of flowchart processes.

3.2.3.2. Classification Back-end

The classification back-end is where likelihood estimators, constructed by the auditory front-end, were used to predicted listener's behaviour. Varying the stimulus-classes stored in the model's back-end enabled the investigation of different schemes by which the auditory system could be performing CVI.

Suboptimal Scheme

As in existing models of CVI (Assmann & Summerfield, 1989; Culling & Darwin, 1994), this scheme identified vowels independently of one another. Essentially vowels were identified ignorant to the existence of an interfering vowel. Class-conditional likelihood estimators were constructed for individual vowels by averaging across likelihood estimates of said vowel at all F0s:

$$P(\boldsymbol{x}|/v_j/) = \sum_{k \in K} P(\boldsymbol{x}|/v_j/, k)\, P(k) \quad . \tag{3.4}$$

Accounting for the fact that concurrent-vowel pairs with an F0 difference of 0, 1, or 4, semitones had been presented twice as much as others by the end of testing (Section 3.2.2), $P(k) = 0.\dot{2}$ when $k = 100{,}106{,}126$, else $P(k) = 0.\dot{1}$. As specified in Section 2.5.2, the modelling framework otherwise assumes equal priors and a zero-one loss function. Given these assumptions, regions (R) could be determined for which each individual vowel had the MAP estimate (Eq. 2.4).

Given the presentation of a concurrent-vowel pair $/v_i, v_j/$ (where $/v_i/$ had an F0 of 100 Hz, and $/v_j/$ an F0 of k Hz), the probability of identifying $/v_i/$ and of identifying $/v_j/$ were separately calculated (as stated in Eq. 2.10). Assuming independence, the resulting probability of identifying $/v_i, v_j/$ (where $/v_j/$ had an F0 of k Hz) as $/v_I, v_J/$ was calculated as (shown schematically in Figure 3.3a):

$$\begin{aligned} P(/v_I, v_J/|/v_i, v_j/, k) = & \int_{\boldsymbol{x} \in R_{/v_I/}} P(\boldsymbol{x}|/v_i, v_j/, k)\, d\boldsymbol{x} \\ & \times \int_{\boldsymbol{x} \in R_{/v_J/}} P(\boldsymbol{x}|/v_i, v_j/, k)\, d\boldsymbol{x} \quad . \end{aligned} \tag{3.5}$$

Optimal Scheme

Alternatively, of substantial interest was whether a back-end configured to emulate an optimal strategy of identifying the combined representation of concurrent-vowel pairs predicted listeners' behaviour. Class-conditional likelihood estimators were constructed for a concurrent-vowel pair by averaging likelihood estimators of said concurrent-vowel pair, across all F0 differences:

$$P(\boldsymbol{x}|/v_i, v_j/) = \frac{1}{2}\sum\nolimits_{k\in K}\big(P(\boldsymbol{x}|/v_i, v_j/, k) + P(\boldsymbol{x}|/v_j, v_i/, k)\big)P(k) \quad . \qquad (3.6)$$

Note how the formulation reflects that, in the task described in 3.2.2, feedback was given without reference to the ordering of F0s within a vowel pairing. $P(k)$ was defined as for the suboptimal scheme, otherwise equal priors and a zero-one loss function were again assumed. Regions (R) were determined for which each concurrent-vowel pair had the MAP estimate (Eq. 2.4).

This optimal scheme attempted to identify the entire representation of a concurrent-vowel pair, calculated as stated in Eq. 2.9. Thus the probability with which the concurrent-vowel pair $/v_i, v_j/$ (where $/v_j/$ had an F0 of k Hz) was identified as $/v_I, v_J/$ was (shown schematically in Figure 3.3b):

$$P(/v_I, v_J/|/v_i, v_j/, k) = \int_{\boldsymbol{x}\in R_{/v_I, v_J/}} P(\boldsymbol{x}|/v_i, v_j/, k)\, d\boldsymbol{x} \quad . \qquad (3.7)$$

The high dimensional integrals stated in Eq. 3.5 and Eq. 3.7 were numerically solved using Monte Carlo methods. This produced confusion matrices directly comparable with human listeners' confusion matrices.

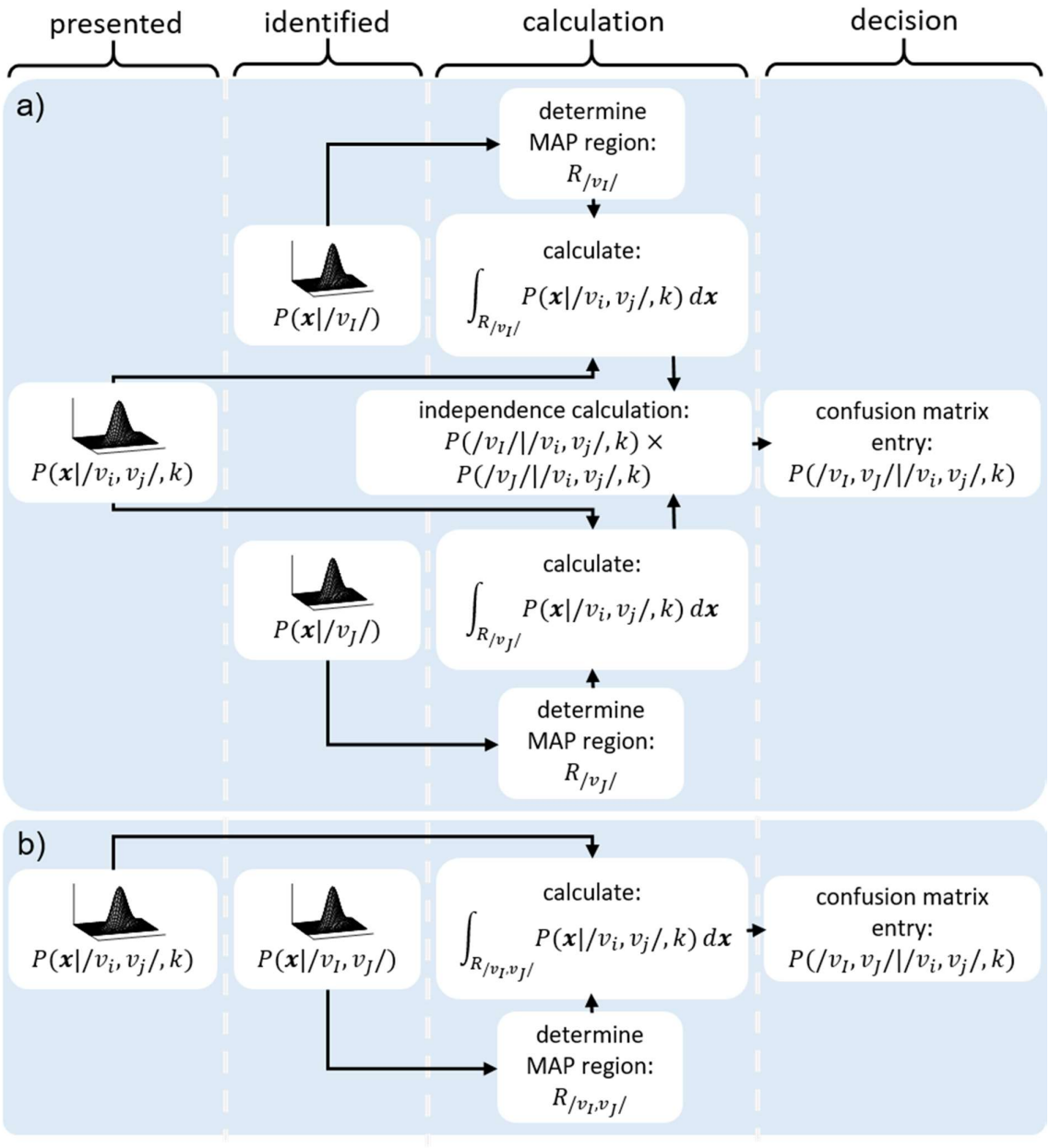


Figure 3.3 Schematic of how a confusion matrix entry predicted by both schemes of the proposed back-end classifier, specifically the identification of the concurrent-vowel pair $/v_i, v_j/$ with an F0 difference of k Hz as $/v_I, v_J/$. Diagram (a) shows the process for classification using representations describing individual vowels, whereas (b) shows the equivalent process with stored representations of concurrent-vowel pairs.

3.2.3.3. Overview

To recap, for a given concurrent-vowel pair a spectro-temporal representation was produced via a simulation of the auditory periphery. Noise was then added to each element of this representation in the form of independent equal variance Gaussian distributions. These probabilistic representations were then compared with probabilistic representations of either individual vowels (suboptimal scheme; in which case this was followed by calculation of joint probability assuming independence) or concurrent-vowel pairs (optimal scheme). A confusion matrix was built from the probability of identifying presented concurrent-vowel pairs as the same, or another, concurrent-vowel pair. All Gaussians had the same variance parameter which was fitted (by stochastic convergence) to best match the average number of correct identifications predicted by listeners (72.6%) when vowel pairs have no F0 differences.

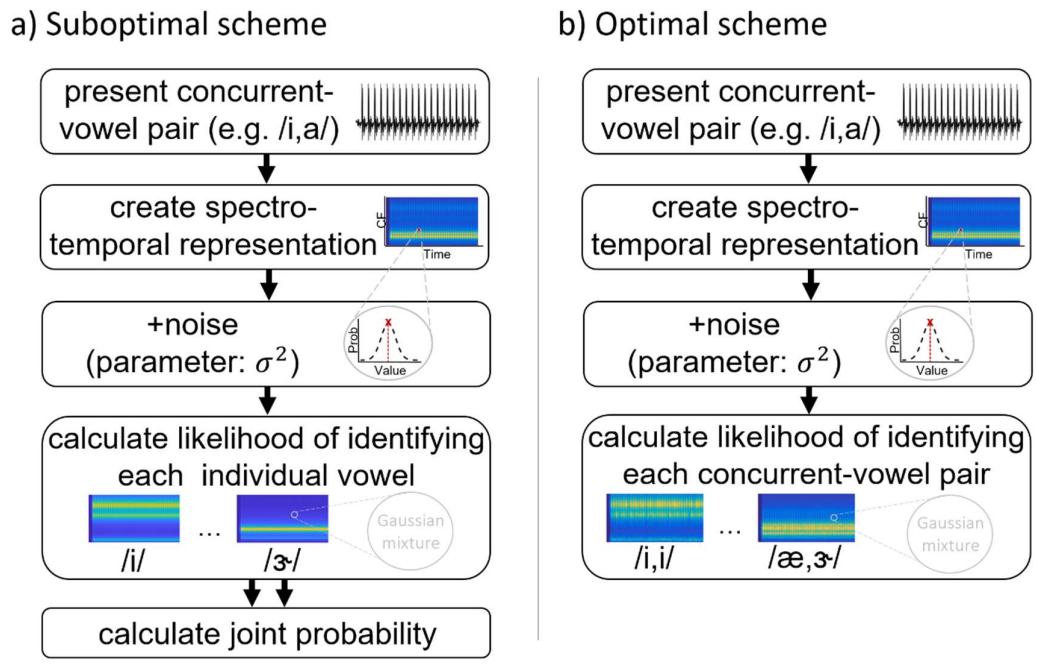


Figure 3.4 High level schematics of the proposed computational model for the (a) suboptimal and (b) optimal schemes.

3.2.4. Information-theoretic Analysis

Information theory was utilised to study the transmission of information in the proposed models. Information can be thought of as an entity that reduces uncertainty, commonly measured in bits (Shannon, 1948). One bit represents the information that is gained when the value of a binary random variable is known. A fundamental quantity in information theory is entropy, which quantifies the uncertainty of a system. The entropy (in bits) of a random variable X, which exists in a probability space $x \in \boldsymbol{X}$, is

$$H(X) = -\sum_{x \in \boldsymbol{X}} P_X(x) \log_2 P_X(x) \quad . \tag{3.8}$$

For example, the random variable X could represent the model's identification of concurrent-vowel pairs, and thus the probability space is made up of all the available concurrent-vowel pairs that could be identified.

Additionally, the relationship between the model's predictions and representational features was quantified by measuring how much information they share. Termed mutual information (MI), this measures how much information is gained about a random variable X by observing another random variable Y (X and Y exist in the probability spaces $x \in \boldsymbol{X}$ and $y \in \boldsymbol{Y}$ respectively), and vice versa. X could still denote the model's identification of concurrent-vowel pairs, and Y an element in a feature vector. Assuming X and Y take the form of discrete variables, MI is formally defined as:

$$I(X;Y) = \sum_{y \in \boldsymbol{Y}} \sum_{x \in \boldsymbol{X}} P_{X,Y}(x,y) \log_2 \left(\frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)} \right) \quad . \tag{3.9}$$

MI variants can be calculated by summating over a subset of the probability space. This can be used to determine the information shared between random variables in specific instances. For example, the relationship between a specific type of error and a representational feature could be understood (in this instance Eq. 3.9 would be calculated for a single event, x). It should be noted, that as Eq. 3.9 requires the discrete sampling of feature PDFs, MI measures are not absolute, yet adequately permit at least qualitative assessments.

3.3. Results

3.3.1. Independent Identification of Vowels

When presented with concurrent-vowel pairs without F0 differences, existing models of CVI were unable to perform segregation. This resulted in a suboptimal comparison of representations describing entire concurrent-vowel pairs with individual vowels. Presented is a computational model that was configured to implement a probabilistic equivalent of this comparison, implementing the commonly held assumption that vowels are identified independently of one another (embodied in Eq. 3.5). A fitting parameter, Gaussian noise, was set to best match the model's mean performance with that of listeners' (72.6%), when concurrent-vowel pairs were presented without F0 differences.

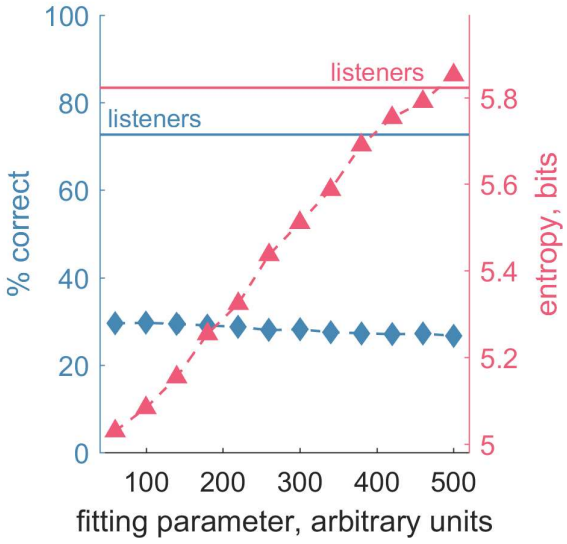


Figure 3.5 Predictions of proposed model when configured to independently identify individual vowels (no F0 difference). The blue plot represents the proportion of correct identifications predicted by the model (diamonds) and listeners (line). The salmon coloured plot shows the entropy of predicted confusion matrices predicted by the model (triangle) and listeners (line).

Akin to existing models of CVI, this approach failed to predict human performance when concurrent-vowels were presented without an F0 difference. The maximum proportion of correct identifications predicted by the model was 30.2%. As the fitting parameter increased, i.e. extra Gaussian noise was added within the model, the proportion of correct identifications made monotonically reduced (see Figure 3.5). This reduction in performance was traded off with an increase in how uncertain the predictions were, quantified with information entropy. When the fitting parameter increased in size such that the model's entropy matched that of listeners' (5.82 bits), the proportion of correct identifications made dropped to 26.7%. This demonstrates that, relative to listeners, this model is intrinsically more certain in the predictions it makes. This is only altered under conditions that are detrimental to the model's overall performance.

Similar to entropy, the PCC at each F0 differences between the confusion matrix predicted by the model and listeners' increased in tandem with the fitting parameter ($r = 0.32$ monotonically increased to $r = 0.36$, $p \ll 0.01$). This range of PCCs were all higher than what was achieved with Culling and Darwin's (1994) model ($r = 0.27$), but below Assmann and Summerfield's (1989) model ($r = 0.42$ to $r = 0.710$) which explored different transformations to representations of concurrent-vowel pairs. Both of these models (and the presented model) contain the assumption that vowels are identified independently of one another. However, this assumption alone limits the predictive performance of models. The maximum possible PCC between such models, and the obtained listener data, is $r = 0.88$ ($p \ll 0.01$). This was determined by applying a nonlinear optimisation algorithm to derive a maximally correlated confusion matrix, whilst under the constraint of independence (i.e. the probability of identifying one vowel cannot alter the probability of identifying another vowel). From this result, it seems likely that a less trivial classification scheme is at work.

		Concurrent-vowel pairs identified, %														
		/i, i/	/ɑ, ɑ/	/u, u/	/æ, æ/	/ɝ, ɝ/	/i, a/	/i, u/	/i, æ/	/i, ɝ/	/a, u/	/a, æ/	/a, ɝ/	/u, æ/	/u, ɝ/	/æ, ɝ/
Concurrent -vowels presented (listeners)	/i, i/	83						18								
	/ɑ, ɑ/		73		2		3				17	3	3			
	/u, u/			100												
	/æ, æ/				48				3			11		21		16
	/ɝ, ɝ/					51							2		47	
	/i, a/						90	1	5	2	2	2				
	/i, u/	20						79	1	1						
	/i, æ/				1		2		86	1		3		8		1
	/i, ɝ/							2	1	94					2	1
	/a, u/		15	1			13		1		70	1	1	1		
	/a, æ/				24				6		3	44		16		8
	/a, ɝ/		7			3	1				12	8	64	1	2	4
	/u, æ/				12		1		14		2	5		64		3
	/u, ɝ/					17		1					1		81	1
	/æ, ɝ/								3	2	2	14	11	7	1	63
Concurrent-vowels presented (model)	/i, i/	100														
	/ɑ, ɑ/		99													
	/u, u/			100												
	/æ, æ/				100											
	/ɝ, ɝ/					100										
	/i, a/		100													
	/i, u/	100														
	/i, æ/				100											
	/i, ɝ/					100										
	/a, u/		100													
	/a, æ/		48		9							42				
	/a, ɝ/		40		1	7						14	32			6
	/u, æ/				100											
	/u, ɝ/					100										
	/æ, ɝ/				32	15						5	4			44

Table 3.1 Confusion matrices predicted by human listeners (top) and the suboptimal model (bottom; fitting parameter = 60) when vowels had no F0 difference. Correct identification entries are shaded in light green. Confusions that listeners made more than 10% of the time are shaded in light yellow.

3.3.2. Identifying the Combined Representation of Vowels

An alternative to the widely adopted suboptimal scheme examined in Section 3.3.1, is a scheme that predicts the combined representation of concurrently presented vowels. This is an optimal scheme, because when implemented perfectly, predictive performance tends to 100% as the variability in the system is reduced. A computational model was configured to implement this strategy (embodied in Eq. 3.7). As in Section 3.3.1, the model's fitting parameter was set to best match the model's mean performance with that of listeners' when concurrent-vowel pairs were presented without F0 differences (see Figure 3.6). This parameter was held constant for the presentation of concurrent-vowel pairs with F0 differences.

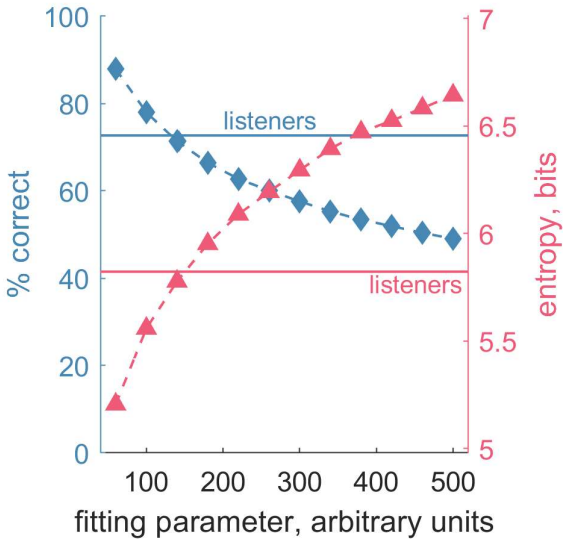


Figure 3.6 Plot is of the same format as Figure 3.5, but for predictions made with the optimally configured model (for vowel pairs presented with no F0 difference).

This model was able to match human performance when concurrent-vowel pairs without F0 differences were presented. Moreover, as the F0 difference between the vowel pairs increased, the model improved. This F0 benefit both qualitatively and quantitatively predicted listeners' behaviour (Figure 3.7). Between 0 and 4 semitones the model predicted an improvement of 15.1%, only 3.2% less than that of listeners. The model's predictions were similar to listeners' at all 6 F0 differences

(root mean square error (RMSE) of 3.6%). This is a considerable improvement on previous attempt to model this dataset using segregation models (Chintanpalli & Heinz, 2013).

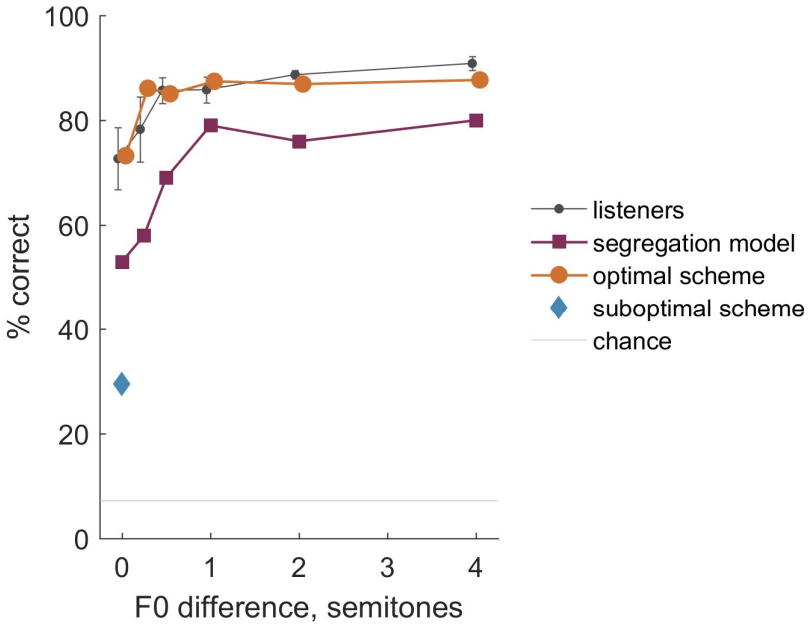


Figure 3.7 Mean percent correct, as a variable of F0 difference between concurrent-vowel pairs, predicted by: human listeners, a segregation model (Chintanpalli & Heinz, 2013), and the model configurations proposed in this chapter. Chance performance is also shown for reference. The model's predictions are calculated by Monte Carlo approximation of an integral. Error bars are not shown for the model as they would describe the convergence of the numerical approximation method used.

A criticism of existing models of CVI was that they could not predict listeners' underlying decisions. Often models were deterministic, thus not able to account for uncertainty inherent in human decision-making. The proposed model makes probabilistic decisions directly comparable to listeners' decisions. Additionally, when configured to optimally identify concurrent-vowel pairs, predictions were highly correlated with listeners'. The PCC between the model and listeners', for all decision probabilities, across all F0 differences was $r = 0.98$ ($p \ll 0.01$). The PCC analysis, performed conditionally at each F0 difference, produced a spread of high

coefficients (0 semitones: $r = 0.93, p \ll 0.01$; 0.25 semitones: $r = 0.96, p \ll 0.01$; 0.5 semitones: $r = 0.99, p \ll 0.01$; 1 semitones: $r = 0.99, p \ll 0.01$; 2 semitones: $r = 0.99, p \ll 0.01$; 4 semitones: $r = 0.99, p \ll 0.01$), displayed visually in Figure 3.8. When isolating and correlating only correct identifications, predictions still correlated well with listeners' performances ($r = 0.57$, $p \ll 0.01$). The same was true when only errors were accounted for ($r = 0.49$, $p \ll 0.01$). The largest error made by listeners was the misidentification of the identical concurrent-vowel pair /ɜ, ɜ/ for the pairing /u,ɜ/, when vowels had identical F0s. This error was made on average 47.1% of the time, and the model predicted that this error occurs 44.6% of the time, 2.5% different.

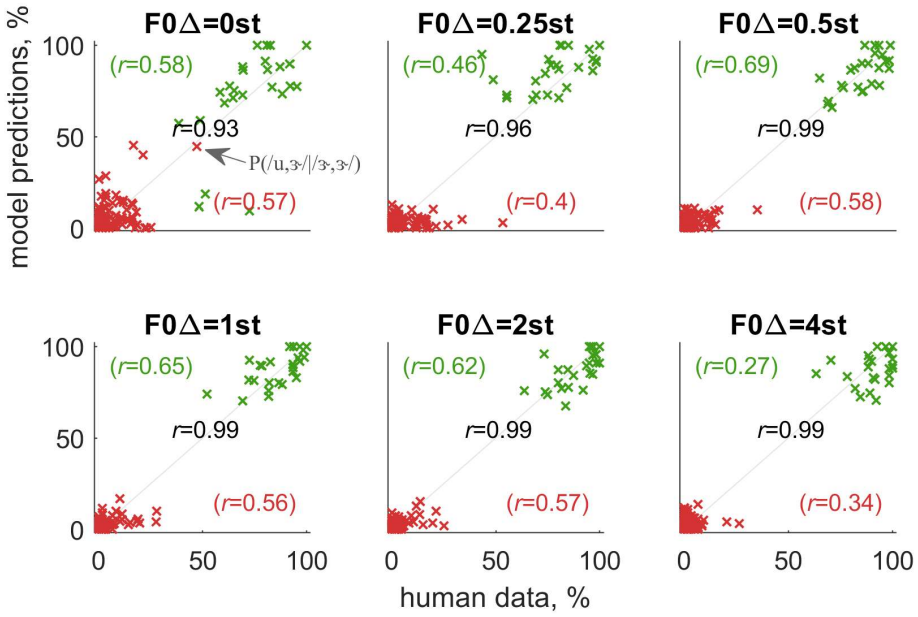


Figure 3.8 Scatter plots comparing the probability with which individual decisions were predicted by listeners (abiscissa) relative to the proposed model (ordinate), at the 6 F0 differences (written F0Δ). This is equivalent to a scatter diagram comparing entries within a listeners' confusion matrix with entries from a confusion matrix predicted by the model. Decisions that were correct identifications are marked as green, and errors are red. The corresponding PCCs are displayed (in black is the coefficient for all decision types). Marked in grey is the misidentification of the concurrent-vowel pair /ɜ, ɜ/ for the pairing /u,ɜ/, when vowels had identical F0s.

Concurrent-vowels presented (model)	Concurrent-vowel pairs identified, %														
	/i, i/	/ɑ, ɑ/	/u, u/	/æ, æ/	/ɝ, ɝ/	/i, a/	/i, u/	/i, æ/	/i, ɝ/	/a, u/	/a, æ/	/a, ɝ/	/u, æ/	/u, ɝ/	/æ, ɝ/
/i, i/	100														
/ɑ, ɑ/		9				19				45	11	15			
/u, u/			100												
/æ, æ/				12				29			6		40		14
/ɝ, ɝ/					19				27			2		45	9
/i, a/		1				88				6	2	3			
/i, u/							100								
/i, æ/				3				77			1		14		5
/i, ɝ/					2				77			2		17	2
/a, u/		2				4				88	2	4			
/a, æ/		2				7		1		12	57	12	9		1
/a, ɝ/		1				2			1	5	2	78	1	3	10
/u, æ/				3				13			4	1	74		6
/u, ɝ/					2				5			2		91	1
/æ, ɝ/				1				3	1		2	14	6	3	71

Table 3.2 Confusion matrices predicted by the optimal model when vowels were presented with no F0 differences. Correct identification entries are shaded in light green. Confusions that listeners made more than 10% of the time are shaded in light yellow.

3.3.3. Training Data

At the beginning of the CVI task (Section 3.2.2) listeners were given training. They were trained to correctly identify individual vowels, but also trained to identify concurrent-vowel pairs presented at three separate F0 differences (0, 1, and 4 semitones). This latter aspect was specifically designed to elicit an increased F0 benefit in listeners. When listeners were tested on their ability to identify concurrent-vowel pairs at F0 differences for which they had already been trained, performance had significantly improved (unpaired t-test equal mean unknown but equal variance, null hypothesis rejected at the 5% level with $p \ll 0.01$). Arguably, the optimal scheme of the proposed model may be successful as it imitates listeners learning the entire representations of concurrent-vowel pairs during the training phase of the CVI task. However, although decisions improved in accuracy following training, the qualitative pattern of decisions made during training and testing were highly correlated (PCC between confusion matrices collected during training and testing is $r = 0.98, p \ll 0.01$).

Listeners were also tested on concurrent-vowel pairs presented at three F0 differences for which they had not been trained (0.25, 0.5 and 2 semitones). Listeners still demonstrated an F0 benefit for these stimuli. Additionally, listeners exhibited an F0 benefit in the training process (black dots in Figure 3.9), during which, listeners improved by more than 20% when concurrent-vowels were presented with an F0 difference of 1 semitone rather than 0. When presented with concurrent-vowel pairs with an F0 difference of 4 semitones, in turn, performance slightly reduced. However, in this instance the data was more variable across the experimental participants.

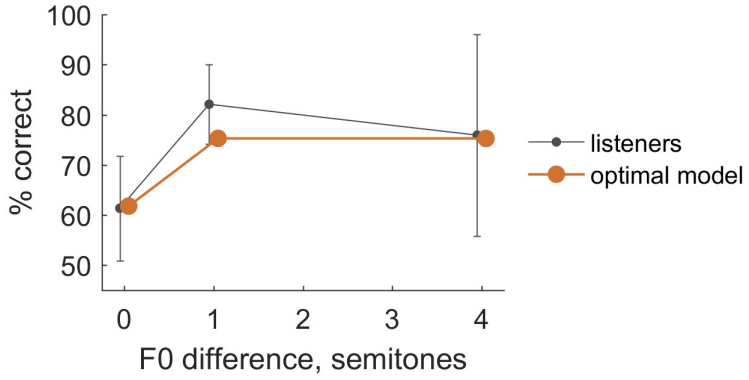


Figure 3.9 The mean number of correct identifications predicted by human listeners during training and the proposed optimal model of CVI, fitted accordingly. Only 3 F0 differences were tested during training.

The proposed model of CVI, when predicting the combined representation of vowels, was able to account for the training data (orange circles in Figure 3.9). By increasing the fitting parameter to match the model's average performance to listeners' when presented concurrent-vowels that had no F0 differences during training, the model qualitatively and quantitatively predicted an F0 benefit. The predicted decisions were, again, also highly correlated with listeners ($r = 0.94, p \ll 0.01$). Consequently, this demonstrates that the improvement between training and testing can feasibly be explained by a reduction of Gaussian noise in the decision-making process (i.e. the model's fitting parameter). This does not suggest any change in strategy for recognition in training compared to testing. As prior to training on concurrent-vowel pair stimuli listeners had only been exposed to individual vowel stimuli, learning of the entire representations of concurrent-vowel pairs does not feasibly explain the presented findings.

3.3.4. Mutual Information

By applying information-theoretic metrics (Shannon, 1948) it was possible to account for why the proposed model made the predictions it did. Figure 3.10a displays the MI between all decisions predicted by the proposed model and representational features. When a feature's probability distribution was independent of the model's decisions that feature was deemed uninformative, quantified by MI equalling zero bits (white pixels). Conversely, increasing mutual dependence between these two variables was signified by increasing MI values (proportional to darkness of pixels). This provides an intuition as to what spectro-temporal features are important, ambiguous when merely considering the model's outputs. Salient in Figure 3.10a are distinct rows for which high values of MI are attributed. The two most prominent horizontal regions correspond to channels in the auditory front-end simulation particularly responsive to frequencies between 0.5 and 1.5 kHz. This frequency region spans the first and second formants of the presented vowels.

MI calculations were also performed limited to decisions made when concurrent-vowels pairs were presented at particular F0 differences (Figure 3.10b). For vowel pairs with non-identical F0s, MI was not stationary across time. Instead, a temporal oscillation in MI was observed in features corresponding to channels in the auditory front-end responsive to low frequencies. The frequency with which MI oscillated increased as the F0 difference between concurrent-vowels increased. These frequencies closely correspond with ENV modulations of unresolved harmonic interactions attributable to the F0s of the vowels presented, i.e. beating. In the auditory front-end simulation, Gammatone filters centred at lower frequencies were dominated by a single pair of unresolved harmonics. This would culminate in deeper, more visible, beats than in channels with higher characteristic frequencies responsive to a greater number of partials.

MI analysis was also used to explain why the model made specific decisions, including errors. Highlighted in Section 3.3.2, the misidentification of /ɜ,ɜ/ for

/u,ɝ/, when vowels had identical F0s, was the most prevalent confusion made by listeners. Figure 3.10c displays the mutual dependence of the model's prediction of this specific error and representational features. Noticeable is a darker region at the top of the plot, indicative of features principally responsible for this error. When /ɝ,ɝ/ was presented, high frequency (>2.5 kHz) features regularly took values which were more similar to values observed when /u,ɝ/ was presented. Due to the similarity between the model's predictions and listeners', this MI analysis offers a plausible and data driven explanation for why listeners made this specific confusion.

It was also viable to look at how the features informed improvement in the model's predictions with increasing F0 differences. This was achieved by subtracting MI calculated for vowel pairs presented at one F0 difference from MI calculated at another F0 differences, and observing positive and negative changes. Figure 3.10d displays this calculation for concurrent-vowel pairs presented with F0 difference of 4 semitones, relative to F0 difference of 0 semitones. Changes in MI appear somewhat volatile, widely spread across spectral and temporal features. This implies that the auditory features driving the model's improvement with F0 differences are manifold.

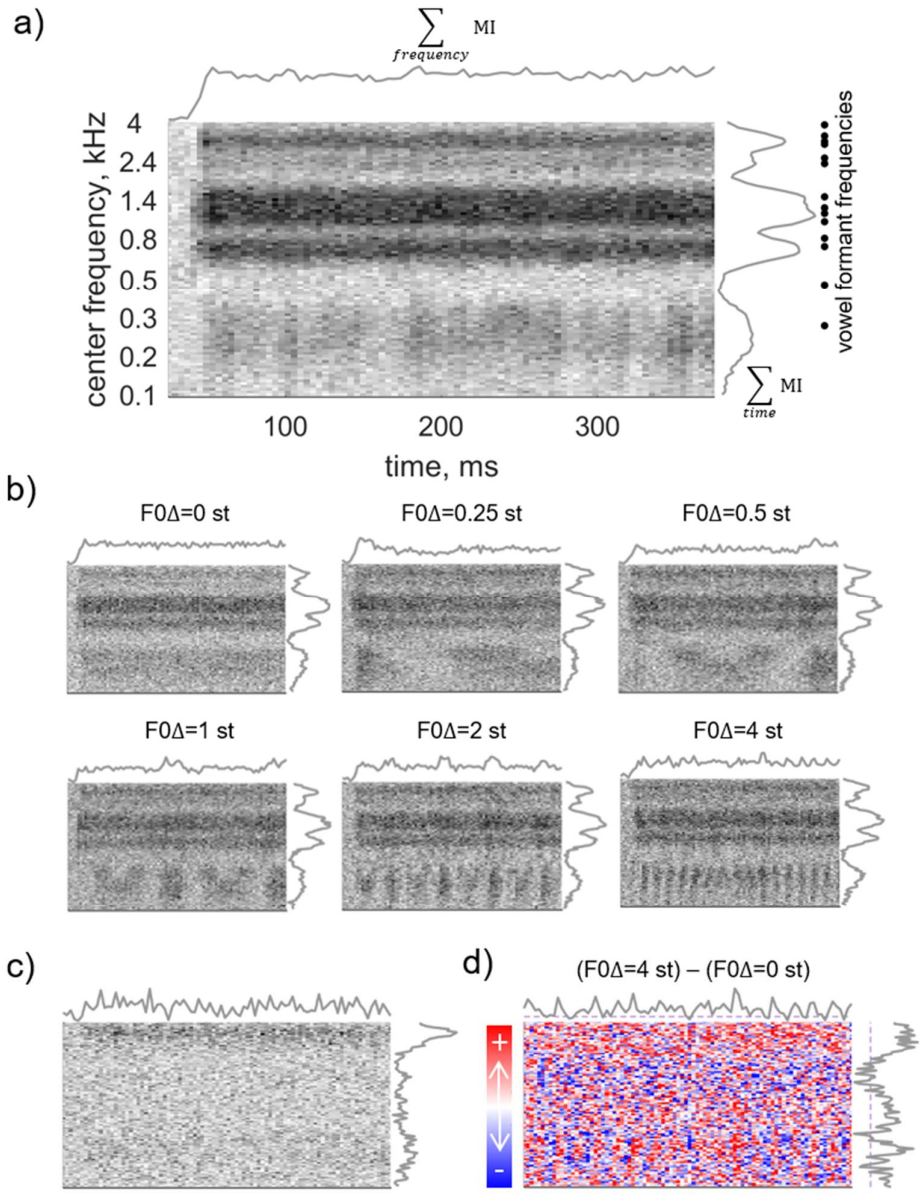


Figure 3.10 Plots displaying to what degree each representational feature informed: (a) all the model's decisions. (b) the model's decisions at 6 separate F0 differences. (c) the model's confusion of /ɜ,ɜ/ with /u,ɜ/ (identical F0s). (d) the model's improvement with an F0 difference of 4 semitones relative to 0 semitones.

3.3.5. Meta-parameters

So far, the back-end of the proposed model has been investigated using an identical auditory front-end simulation (detailed in Section 3.2.3). Varying the auditory front-end's meta-parameters allowed the significance of these meta-parameters on the model's predictions to be determined. A particularly influential meta-parameter was the cut-off frequency of low-pass filters. This directly limited the temporal information contained in the representation of vowel stimuli. As the cut-off increased in frequency the accuracy with which the proposed model predicted listeners' percent correct identification, as a function of F0 differences, increased. Congruently the size of the predicted F0 benefit increased (red dots in Figure 3.11a). However, at higher cut-off frequencies the magnitude of the F0 benefit was traded for the accuracy with which individual decisions were predicted. Figure 3.11b (red dots) shows how higher cut-off frequencies resulted in less correlated error predictions between the model and listeners. The trade-off between F0 benefit and decision accuracy culminates in an optimum cut-off frequency intermediate of extreme values. Figure 3.11c (red dots), correlating all decisions, across all F0 differences, captures this balance best and indicated an optimum cut-off frequency of 100 Hz. This permits the existence of low FM, such as auditory ENVs, but not high frequency TFS.

The effect of having Gammatone filter bandwidths determined by psychophysical data collected by Oxenham and Shera (2003) was also probed. These narrower measures of human auditory filters retain the trends predicted for a model with filter widths derived by Glasberg and Moore (1990) (red and black dots in Figure 3.11). The narrower filters did modestly improve the PCC between the model and listeners errors, but predicted a marginally smaller F0 benefit. The more significant effect of filter bandwidths was the amount of Gaussian noise (fitting parameter) in the model's front-end that was required to match the performance of listeners. As would be expected due to decreased correlation between features, the narrower Gammatone filters required more noise to match the performance of listeners.

An additional meta-parameter of interest was the exclusion of the standardisation process. Plots Figure 3.11d, Figure 3.11e, and Figure 3.11f, are the equivalents to Figure 3.11a, Figure 3.11b, and Figure 3.11c but for an auditory front-end without standardisation. The qualitative dynamics of the according predictions were similar to the front-end with standardisation, however predictions were generally less accurate (Figure 3.11e and Figure 3.11f). In order to predict a substantial non-negative F0 benefit, the model without standardisation required a higher cut-off frequency.

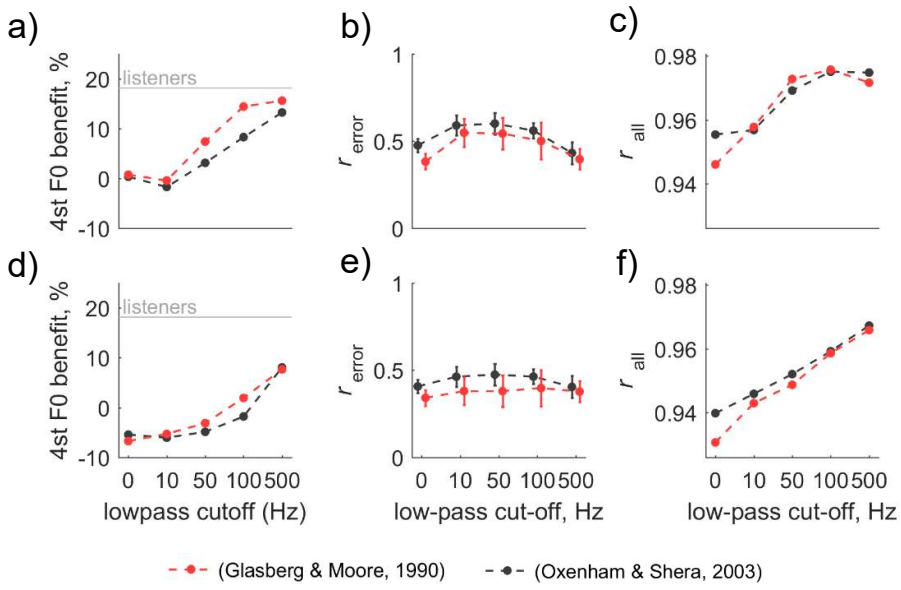


Figure 3.11 (a) Improvement in identification with an F0 difference of 4 semitones. (b) PCCs between errors made by the model and listeners, calculated at all 6 F0 differences. (c) PCCs between all of the model's predictions and listeners. (d), (e), and (f) display the same measures as (a), (b), and (c), but for an auditory front-end without standardisation. All measures were calculated for an optimal model configuration ran at 5 different low-pass cut-offs (0, 10, 50, 100, and 500 Hz) and for two Gammatone filter shapes (Glasberg & Moore, 1990; Oxenham & Shera, 2003).

3.3.6. Autocorrelation

A contentious process found in Meddis and Hewitt's (1992) harmonic segregation model was a PACF. The process removes phase information from auditory signals, before combining signals across channels. The proposed optimal model was run with an auditory front-end configured to include a PACF. One version of the PACF took into account temporal delays up to 0.0045 ms, denoted the 'timbre region'. A second version of the PACF accounted for the timbre region as well as temporal delays longer than 0.0045 ms, the 'pitch region'. Aside from the inclusion of a PACF, the rest of the model was unchanged.

When the PACF only included the timbre region, the model was able to predict an F0 benefit of 8.4% from 0 to 4 semitones (model was fitted to match listeners' performance when vowels were presented with identical F0s). When the PACF also included the pitch region, the predicted F0 benefit increased to 18.5%. Therefore, information contained in periodicities < 222 Hz (lags ≥ 0.0045 ms) do seem to influence the size of the model's predicted F0 benefit. However, the PACF inclusive of both the timbre and pitch regions substantially over performed listeners by 12.2% when concurrent-vowel pairs were presented with an F0 difference of 0.25 semitones. In both PACF conditions the specific decisions predicted were again well correlated with listeners' ($r = 0.93, p \ll 0.01$ and $r = 0.92, p \ll 0.01$ respectively). However, despite the notable predictive power of the proposed model inclusive of a PACF, its inclusion seemed to have no obvious advantage over the same model without a PACF. Arguably, an autocorrelation process could be justified as an efficient processing strategy where the removal of phase information is a form of data compression. The removal of phase information could also make a model insensitive to stimulus phase differences across channels.

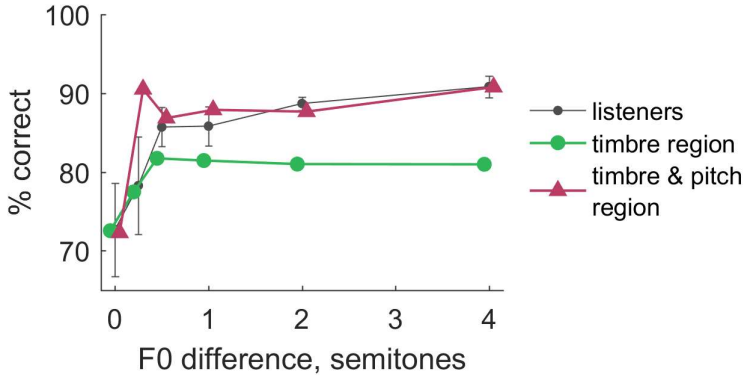


Figure 3.12 The mean number of correct identifications predicted by the proposed model of CVI when the auditory front-end included a PACF.

3.3.7. Vowels with Interleaved Harmonics

The information-theoretic analysis in Section 3.3.4 revealed a relationship between the model's predictions and low frequency temporal modulations. Such beating has been previously proposed as a segregation cue for CVI (Culling & Darwin, 1994). They presented concurrent-vowel pairs with interleaved harmonics (see methods 3.2.1), and listeners were found to exhibit a small F0 benefit.

Interleaved vowel stimuli were presented to the proposed optimal model of CVI. The model predicted an insignificant F0 benefit from 0 to 0.25 semitones (Figure 3.13; yellow squares). However, beyond this, identification of vowels pairs with non-identical F0s was worse than for vowels pairs with no F0 differences. When the model was instead trained to store representations of the interleaved concurrent-vowel pairs an F0 benefit was predicted (brown diamonds). This improvement of 12.7% over 4 semitones was similar in scope to listeners tested on normal concurrent-vowel pairs. In fact, in this condition the predicted confusions were highly correlated with listeners that had been presented the normal, non-interleaved, vowel pairs ($r = 0.97, p \ll 0.01$).

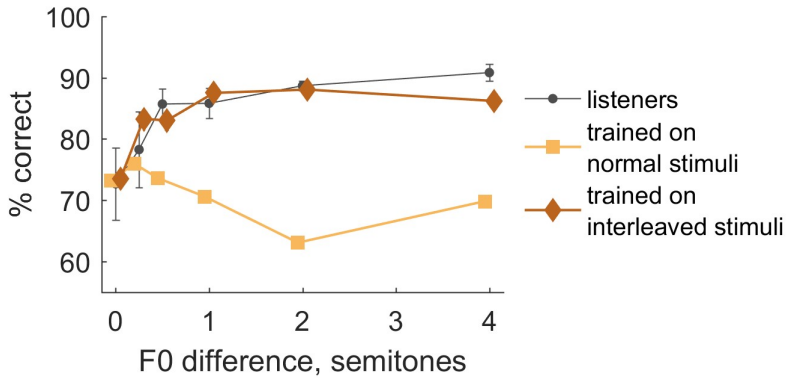


Figure 3.13 The mean number of correct identifications predicted by the proposed optimal model, when presented concurrent-vowel pairs with interleaved harmonics. The model was run without (yellow squares) and with (brown diamonds) prior knowledge of the interleaved vowels. Data is shown for listeners tested on normal concurrent-vowel pairs.

3.4. Discussion

This chapter described a computational model constructed to predict the behaviour of listeners identifying concurrently presented vowels with varying F0 differences. The model's architecture had two key components, a front-end simulation of the auditory periphery, and a back-end configured to make probabilistic predictions based upon front-end representations. The model did not perform any segregation, but different classification schemes encoded in the model's back-end were examined.

Initially a scheme that identified individual vowels independently of one another, termed the suboptimal scheme, was interrogated. Operating under this scheme, the proposed model replicated the failure of alike existing models (Assmann & Summerfield, 1990; Assmann & Summerfield, 1989; Chintanpalli & Heinz, 2013; Culling & Darwin, 1994; Meddis & Hewitt, 1992; Scheffers, 1983) to predict listeners' behaviour. The proposed model equally weighted all spectro-temporal features, but previous studies have shown the predictive value of enhancing features such as spectral peaks and shoulders (Assmann & Summerfield, 1989). A model that appropriately weighted features to benefit performance, such as a locally weighted naïve Bayes classifier (Frank, Hall, & Pfahringer, 2002), may provide a more tolerant impression of what is achievable with an independent identification scheme. Irrespective, the assertion of independent identification was alone found to severely limit a model's capacity to match listeners' confusion matrices.

Alternatively, a scheme was proposed that optimally predicted the combined representation of concurrent-vowel pairs. This approach resulted in predictions that qualitatively and quantitatively matched listeners', a vast improvement on previous attempts to model this data (Chintanpalli & Heinz, 2013). The proposed optimal scheme was designed to address what goals the auditory system may be solving when performing CVI (Marr & Poggio, 1977). However, a trivial

interpretation of the scheme is that participants had learnt the representations of concurrent-vowel pairs during training. Training has been previously found to improve the ability of listeners to perform CVI, and this has been associated with changes in auditory cortices within one hour of training (Alain, Snyder, He, & Reinke, 2007; Reinke, He, Wang, & Alain, 2003). However, although training effects were observed, the model could also account for listeners' training data which was not qualitatively different from their testing data.

Explaining how the proposed optimal scheme would be implemented in the brain is difficult given its highly abstract definition. Predicting the combined representation of vowels could be conceptualised as a predictive feedforward/feedback system (Bastos et al., 2012; Rao & Ballard, 1999) operating at cerebral levels of the auditory system. However, experiments feeding listeners precursory or retroactive information about one vowel of a concurrent pair were not found to advantage the identification of the remaining vowel, aside from for an ipsilateral acoustic precursor (Brandewie & Oxenham, 2015; Summerfield & Assmann, 1989). Seemingly this is evidence against predictive coding. Yet, such experiments do not rule out the possibility that if listeners were able to predict the combined representations of vowels this would render any additional information about constituent vowels useless. Beyond such speculation, disentangling bottom-up implementations of CVI from top-down is beyond what is achievable with this solely computational model.

The model described in this chapter did not have any segregation cues hard coded, yet improved in predictive performance when presented vowel pairs with increasing F0 differences. This begs the question of what features in the representations of concurrent-vowel pairs drove this behaviour. MI analysis did not highlight any spectral harmonicity cues. Additionally, the model's inclusion of a PACF, used to simulate temporal encoding of harmonicity (Meddis and Hewitt's, 1992), did not better a model without its inclusion. Regardless, a PACF could be valuable in offering a more efficient encoding of stimuli that is insensitive to phase

information (Meddis & Hewitt, 1991). Together with previously raised discrepancies as to the role of pitch cues, especially at small F0 differences (de Cheveigné, 1999; Culling & Darwin, 1994), these findings throw into doubt the oft-proposed segregation cue of pitch (Micheyl & Oxenham, 2010). Indeed Scheffers (1983), who initially proposed pitch as a cue to facilitate CVI, came to the conclusion that pitch may serve to merely indicate the presence of multiple sounds in an environment. The presence of two vowels is assumed by default in the proposed model.

In line with Culling and Darwin's (1994) beating hypothesis, oscillations in MI implied that there were peaks in time that were more informative to the model and dips that were less informative. Yet, to what extent this is due to the spectral cancellation of individual vowels, as originally proposed by Culling and Darwin (1994), was unclear. The model was unable to predict an F0 benefit for interleaved vowel pairs (designed to retain beating cues but break a harmonic segregation mechanism), possibly due to its phase sensitivity. However, when trained on representations of interleaved vowels the model did predict an F0 benefit. Listeners have been proven to quickly adapt to atypical sounds, such as frozen noise tokens (Felty, Buchwald, & Pisoni, 2009; Langhans & Kohlrausch, 1992). When storing representations of interleaved vowels, the model may have imitated listeners' rapid familiarisation with the atypical interleaved vowel stimuli. Overall, however, it is difficult to say whether the relationship between beating and the model's predictions is causal. A MI analysis instead indicated widespread spectro-temporal changes with F0 differences that could be suggestive of auditory cues not previously considered in the CVI literature.

This discussion of F0 benefit has largely neglected the model's accurate prediction of individual decisions, such as errors. Spectral bands, in the frequency regions of the vowels' first and second formants, were found to be highly dependent on the model's predictions. This finding coincides well with psychoacoustic experiments that have highlighted the importance of these two formants for performing CVI

(Chintanpalli & Heinz, 2013; Darwin & Culling, 1993). In turn, purely spectral representations most accurately predicted listeners' specific decision probabilities. This could be why models operating on temporally enhanced representations of concurrent-vowel pairs fail to predict listeners' confusions (Chintanpalli & Heinz, 2013). A model with more narrowly tuned Gammatone filters (Oxenham & Shera, 2003) predicted listeners' data to a similar degree of accuracy as the model with more broadly tuned filters (Glasberg & Moore, 1990). However, the model with narrower filters required a larger amount of Gaussian noise to approach human like performance. As recent evidence supports narrower auditory filters in the human auditory system (Sumner et al., 2018), this would suggest that CVI in humans is more robust to neural noise than animals with larger auditory filters such as guinea pigs (Evans, 1972).

3.5. Conclusion

A computational model of auditory perception was implemented to question how listeners solve a simple instance of the cocktail party problem, CVI. A proposed model, configured to optimally identify the combined representation of concurrently presented vowels, predicted listener's behaviour both quantitatively and qualitatively. This is a vast improvement on previous efforts. The model's optimal strategy could be interpreted as directly representative of listener's top-down processing, or a high-level abstraction of an optimal bottom-up process. Despite attempts made to relate the model's predictions to low-level auditory features, it is difficult to say what is driving the model's improvement with F0 differences. The model does not seem to be utilising pitch cues, and the role of beating information is unclear. Additionally, analysis suggests that the model may be influenced by auditory cues that are currently difficult to characterise. The model's combination of a simple simulation of the auditory periphery with a predictive back-end clearly has considerable predictive power. Yet a more realistic front-end may help to understand the means by which individuals process concurrently presented vowels.

Chapter 4 Concurrent-vowel Identification: Neurophysiology

4.1. Introduction

In Chapter 3, a model of CVI that optimally predicted the combined representation of vowel pairs was able to predict the behaviour of human listeners. This model, which was fronted by a filter simulation of the auditory periphery, was a successful proof-of-concept. However, it seems wanting to investigate the proposed model with a front-end that encompasses more realistic, nonlinear, auditory processing. To date, the best understood models of auditory processing are based on animals (e.g. the guinea pig). Animals permit investigators to probe processing on a neuronal level. In fact, most existing models of CVI used a front-end computer simulation based on animal physiology.

Attempts have been made to directly investigate the neurophysiological basis of CVI in the AN, CN, and AC of animals (Fishman, Micheyl & Steinschneider, 2016; Keilson, Richards, Wyman, & Young, 1997; Palmer, 1990). Whereas, neural data in the IC has previously only been collected in response to concurrent harmonic series (Nakamoto, Palmer, & Li, 2010; Sinex and Li, 2007), not concurrently presented vowels. Efforts have demonstrated that proposed CVI mechanisms could feasibly operate using the information present in neural representations of concurrently presented vowels. However, the existence of the proposed mechanisms is largely speculation. Additionally, efforts have not been made to make predictions directly comparable with listeners when performing CVI, based on neurophysiology data. This is in part due to the limited set of stimuli that experimenters have applied, as well as the limited spread of neural responses collected.

Running a version of the proposed CVI model, using neural data recorded in response to a complete set of CVI experimental stimuli, would provide a test of the model's robustness and generality. If successful, analysis of the dynamics driving the model's predictions could provide insights into the neural processing underlying CVI. The proposed model makes no explicit assumptions regarding the cues or mechanisms used to perform CVI. Consequently, any improvements in its predictions are directly attributable to an increase in the information describing concurrent-vowel pairs, coded in their neural responses. It would be valuable to understand the form of this information, particularly as a variable of F0 differences between concurrent-vowel pairs. Is it a previously proposed cue, such as harmonicity or beating interactions, that explains the model's behaviour?

The IC is a promising brain nucleus to investigate the previously outlined ideas. As a principal relay station along the auditory pathway, the IC processes a large proportion of auditory related inputs. This, in combination with the ease with which large probes can be inserted in to the IC, makes it a good candidate for recording large quantities of neural data. Additionally, the IC is midway along the auditory pathway making it suitably placed to investigate bottom-up vs top-down processing. The proposed model is predicated on having idealised knowledge of representations describing concurrent-vowel pairs. Understanding how this potentially abstract concept could be substantiated in biology is important. The brain may be performing a high-level prediction process very similar to that proposed in the model. Otherwise, the auditory system may have evolved a bottom-up segregation process that mimics the optimal behaviour of the proposed model, where both are similar in performance due to similar external constraints. Finally, as a brain nucleus which has not yet been probed in response to concurrent-vowel stimuli, the IC potentially offers understanding beyond what it is possible to speculate here (see Appendix A.1 for a more in-depth discussion of IC neuroanatomy).

Finally, confirming whether physiological recordings combined with a computational decision model predicts human behaviour has importance beyond the scope of CVI. An animal model of HL combined with a largely unrestrictive back-end classifier would be an extremely useful tool to help understand the effects of HL on neural coding. Showing the proposed decision model can predict human data based on animal neurophysiology is an important proof-of-concept for future work (Chapter 5).

Within the following chapter I attempt to demonstrate the following:

- Combining neural recordings, in response to a full set of CVI experimental stimuli, with a computational decision model that predicts the combined representation of concurrently presented vowel, predicts human behaviour (Sections 4.3.3 and 4.3.4).
- At larger F0 differences the spectra of concurrent-vowel pairs are less concentrated, resulting in milder cochlear compression, and leading to more distinct neural encoding within the auditory midbrain (Section 4.3.5 and 4.3.6).
- Previously proposed segregation cues, harmonicity and beating, are not well supported (Sections 4.3.7 and 4.3.8).
- Compressive auditory encoding facilitated constructing representations of concurrent-vowel pairs by providing an interleaved binary representation of the interfering sounds (Section 4.3.9).

4.2. Methods

4.2.1. Stimuli, Task, and Model

The stimuli and task are identical to the stimuli and task methods described in Sections 3.2.1 and 3.2.2, aside from the exclusion of interleaved vowel stimuli. The computational decision model was based on the model proposed in Section 3.2.3, aside from the use of a different auditory front-end (the back-end classifier was configured identically as detailed in Section 3.2.3). The alternative front-end consisted of neurophysiology recordings (Section 4.2.2) made in the IC of guinea pigs.

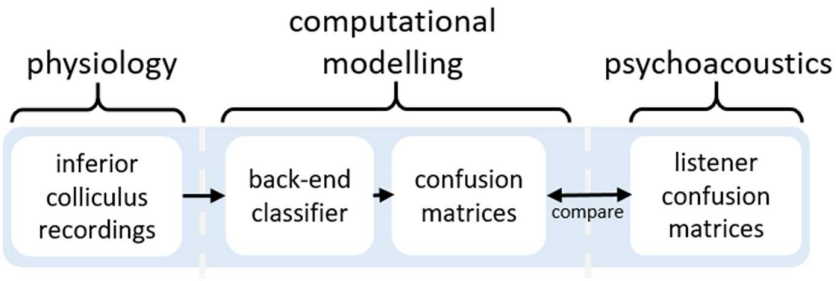


Figure 4.1 Schematic of methodology

4.2.2. Neurophysiology Recordings

4.2.2.1. Subjects

Experimental subjects were six tricolour guinea pigs (*Cavia porcellus*). 4 were male and 2 were female, all bred in house. Animals weighed between 392 and 616 g. Experiments were carried out in accordance with UK Home Office regulations.

4.2.2.2. Surgical Preparation

Anaesthetic induction was induced with bilateral intraperitoneal injections of urethane (4.5 ml kg^{-1}; 20% solution). The forepaw withdrawal reflex was continuously monitored and supressed with intramuscular Hypnorm (0.2 ml; fentanyl citrate 0.315 mg/ml; fluanisone 10 mg/ml) injections. Atropine sulphate (0.2 ml; 6 µg/kg) was subcutaneously injected to supress bronchial secretions. 5 ml saline injections were intermittently delivered to maintain hydration. Animals were laid on a heating blanket and sustained at a core temperature of 38°C monitored by a rectal probe.

Once anaesthetised, hypodermic needles could be placed through skin over the left and right mastoids, and a reference vertex point. Acting as electrodes, needles were used to record auditory brainstem responses (ABRs; see 4.2.2.5). Excess wax in the auditory meatus and obscuring tragus was removed. A tracheal cannula provided 100% oxygenation which was supplied by an artificial ventilation unit (3.5ml oxygen, 100rr). In a sound-proofed and electrically earthed booth, the head was secured within a stereotaxic holder, a bite bar, and hollowed speculae allowing direct aural access to the tympanic membranes.

A midline incision was made in the scalp and the dorsal skull was exposed. Cerebrospinal fluid pressure was reduced draining off excess fluid with a small incision in the posterior fossa. This reduced brain pulsations caused by systolic blood pressure. Holes were made bilaterally in the ethmoid bone of the skull to

provide access to the bulla. This access permitted compound action potential (CAP) recordings made using an insulated wire placed on the round window (see 4.2.2.3). Pressure equalization was maintained by placement of a polythene tube to the bulla and the opening was sealed with vaseline. The skull was levelled between 5 and 13 mm anterior to lambda and a craniotomy, 5 mm^2 in size over the right IC, was made (Rees & Palmer, 1988). The craniotomy was positioned so its centre was 2.5 mm lateral of midline and the rostral boundary aligned with Bregma. Dura was removed taking care to minimise damage to the surface of the cortex. The exposed brain was kept moist using warm saline (0.9%). A hole was drilled in the skull and a thin wire was attached as a ground. Multi-electrode arrays were then inserted in the centre of the craniotomy to record neural responses to auditory stimuli (see 4.2.2.5). Following the experiment, animals were terminated with a 3ml intraperitoneal injection of pentobarbital.

Surgery was performed in sterile conditions: the surgeon wore a sterile gown, gloves, mask, and hat; any exposed animal skin was sterilised with iodine prior to the scalpel incision; an assistant was always present during surgery to help with making injections and monitoring anaesthesia level. They helped to prevent nonsterile contact between the surgeon, instruments, and the animal.

4.2.2.3. Sound Delivery and Calibration

The presentation of auditory stimuli was controlled from either Brainware (developed by J. Schnupp, University of Oxford, UK) or custom built Matlab (MathWorks, Natick, MA, USA) software. Stimuli were passed to an RX8 processor (Tucker Davies Technologies, System 3), following which stimuli were amplified and delivered to Radioshack 40-1377 speakers, inserted into the aural speculae.

A microphone (Brüel & Kjær 4143 condenser microphone) connected to a 26AC pre-amplifier and 12AR power supply (both G.R.A.S, Holte, Denmark), and fitted with a 1mm calibrated probe tube, was placed a few mm from the tympanic membrane. The sound delivery system was calibrated using averaged recordings

to 50 ms unattenuated white noise, repeated 20 times. A frequency decomposition was applied and inverted to produce a flat frequency response.

4.2.2.4. Auditory Brainstem Responses and Compound Action Potential Audiograms

Needles acting as electrodes were connected through a RA4LI, RA16PA, and RX7 (all Tucker Davies Technologies). 5 ms pure tones bursts (ramped 1 ms) were presented bilaterally at a frequency of 1, 4, or 9 kHz, and at a level of 20, 30, 40, 60 or 80 dB SPL, using custom software written in Matlab. ABR signals were recorded from the left mastoid and filtered between 0.1 and 5 kHz. For each frequency-level combination, ABRs were averaged across a maximum of 500 presentations of which 50% of trials were phase inverted. Peak-to-peak amplitudes between 2 and 6 ms were used as a measure of audibility.

Insulated wire placed on the round window of the left cochlea was used to record CAPs with the same hardware/software setup as ABR recordings. 30ms pure tones were presented in 0.5 octave steps between 0.25 and 32 kHz and in 5 dB SPL steps between 30 and 100 dB SPL, with a repeat period of 100 ms. CAP signals were filtered between 0.3 and 12 kHz. The root mean square (RMS) of the CAP response to each frequency-level combination was averaged across 10 stimulus repeats.

4.2.2.5. Neural Recordings

64 channel multi-electrodes arrays in an 8×8 (200 μm × 200 μm spacing) configuration (Neuronexus; Technologies, Ann Arbor, MI, United States) were used to record extracellular potentials from neurons in the central nucleus of the IC. An array was lowered vertically from the brain surface to the midbrain using a Burleigh Inchworm, with a Burleigh 6000 Controller and 6005 handset (Burleigh Instruments, Fishers, NY). The electrode was advanced ~3-5 mm until units responded to repeated 50 ms bursts of ~70 dB SPL broadband noise.

On occasion the anatomical position of electrode penetrations was validated histologically using cytochrome oxidase staining (Figure 4.2; Liu, Palmer and Wallace, 2006). This enzyme is a mitochondrial marker and sensitive to neural damage. In the brain studies very little damage was observed but confirmation was obtained that the electrode tracks were in the central nucleus of the IC.

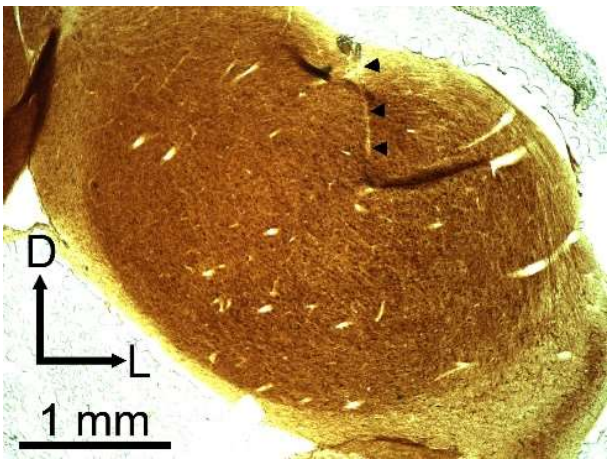


Figure 4.2 Enzyme activity associated with electrode tracks in the centre of the IC, marked with black arrowheads. Lateral (L) and dorsal (D) directions are indicated.

Once an electrode array was at a suitable depth, desired auditory stimuli were delivered by the system described in 4.2.2.3. In response, 64 voltage traces were recorded at a sample rate of 24.1 kHz into Brainware. Gain was applied and traces were bandpass filtered between 0.3 and 6 kHz for offline analysis following the experiment. Offline, a custom spike collection algorithm in Matlab was applied to each channel of the electrode array. Signal statistics were used to determine a spike detection threshold, T, defined by the formula:

$$T = -4 \times median\left(\frac{|x|}{0.6745}\right) \quad , \tag{4.1}$$

where x is a voltage trace (Quiroga, Nadasdy, & Ben-Shaul, 2004). When this threshold was surpassed the presence of a spike was recorded. An empty window of 1 ms prior to, and following, a spike accounted for neural refraction. No additional spike sorting was performed, and hence all neural data was multi-unit activity (MUA).

4.2.2.6. Frequency Response Area

A frequency response area (FRA) for each unit was constructed by presenting 50 ms pure tones (ramped 10 ms), generated using RX8 (Tucker Davies Technologies) hardware. Frequencies ranged from 0.2 to 25.6 kHz in 0.25 octaves steps (across 7 octaves), and levels ranged from 5 dB SPL to 100 dB SPL in 5 dB steps. Frequency-level combinations were presented twice, and all presentations were uniformly randomised. The repeat period was 200 ms.

For each channel, its FRA was constructed from the average spike count in response to each pure tone. To determine the which channels were most likely driven by auditory stimuli a two-sample t-test was performed on firing rates recorded in response to the lowest level tones (spontaneous rates) and the highest level tones (driven rates). For all channels, frequencies at which a channel responded maximally, on average, between 30 and 50 dB ($BF_{30\text{-}50}$) and 50 and 70 dB ($BF_{50\text{-}70}$) were calculated. Although characteristic frequencies were also derived (via curve fitting) it was decided that BFs provided a more reliable and adequate description of MUA frequency response for these investigations, despite having less precision than CFs.

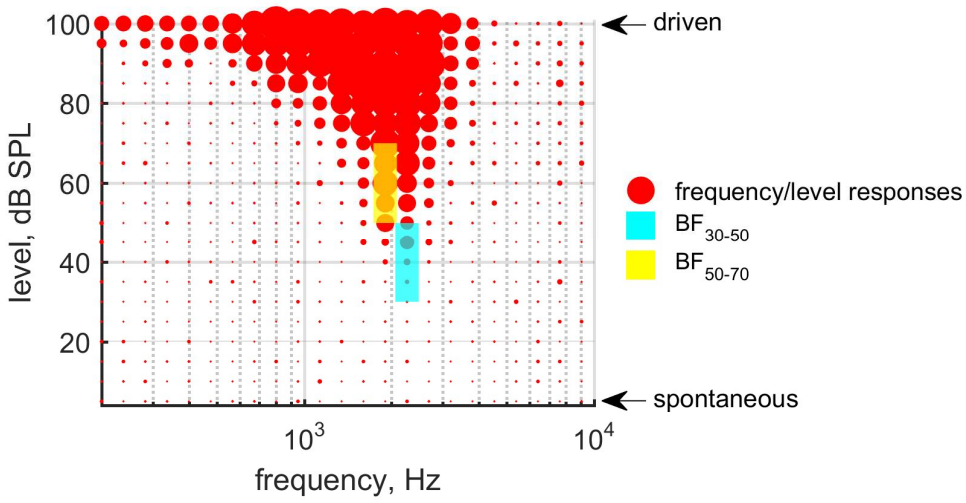


Figure 4.3 Example of a FRA. Size of the red circles are directly proportional to the firing rate in response to frequency/level of pure tone presented. $BF_{30\text{-}50}$ and $BF_{50\text{-}70}$ regions, and levels associated with spontaneous and driven firing rates, are also depicted.

4.2.2.7. Presentation of Concurrent-vowel Identification Stimuli

Stimuli were identical to the CVI experiment stimuli described in Section 3.2.1. To recap, five synthetic vowels (/i/, /a/, /u/, /æ/, /ɝ/), with F0s of 100, 101.5, 103, 106, 112, and 126 Hz were prepared. Each vowel was 400 ms long (ramped 10 ms) and set to 65 dB SPL. All individual vowels, as well as twenty five permutations of concurrent-vowels at F0 differences of 0, 0.25, 0.5, 1, 2, 4 semitones, were presented monotonically (left ear). Each vowel or concurrent-vowel pair stimulus was presented 30 times, randomised, and with a repeat period of 800 ms.

4.2.3. Computational Model

Neural recordings in response to the CVI experimental stimuli were used in combination with the computational modelling framework introduced in Chapter 2 to predict listener behaviour. As applied in Chapter 3, the model consisted of a front-end and a classification back-end. However, as used here, the front-end was an animal model of hearing as opposed to a filter simulation of the periphery.

4.2.3.1. Front-end

For a presentation of a CVI experimental stimulus, the spike count in 0.125 ms (Nyquist limit of 4000 Hz) bins over the proceeding 400 ms was calculated. This was cosine ramped (onset/offset of 25 ms) and passed through a cascade of four 2nd order Butterworth low-pass filters. The cut-off frequency was selected as either 0, 10, 50 (default), 100, or 500 Hz. This signal was then down-sampled to have a Nyquist limit at least 1.25 × the low pass cut-off frequency and windowed between 25 and 375 ms. Low-pass spike counts were averaged across multiple presentations of a concurrent-vowel pair, and combined across the 300 electrode channels least likely to have measured responses to auditory stimuli by chance (using two-sample t-tests), to create a peri-stimulus time histogram (PSTH). 300 channels represented a compromise between retaining a majority of the driven channels, and computational running time (models were also run with 100 or 200 channels and were found to make similar predictions). The variance in spike counts across repeat presentations was also calculated.

For each CVI experimental stimulus, the associated PSTH operated as the mean of an independent multivariate Gaussian distribution. The associated variance in spike counts operated as the variance of this Gaussian distribution. Independent, Gaussian, zero-mean noise, with identical variance, was then added in each dimension. The variance of the additional noise was a fitting parameter in the model and was kept identical in value across concurrent-vowel pairs, electrode channels, and dimensions.

The Gaussian distributions built from the neural responses to a CVI experimental stimulus acted as a likelihood estimator in the model's classification back-end (e.g. $P(\boldsymbol{x}|/v_i, v_j/, k)$ from Section 3.2.3). Eight randomised repeats of 2-fold cross validation were used to avoid overfitting. Consequently, neural responses to 15 presentations of a stimulus were used to build a likelihood distribution.

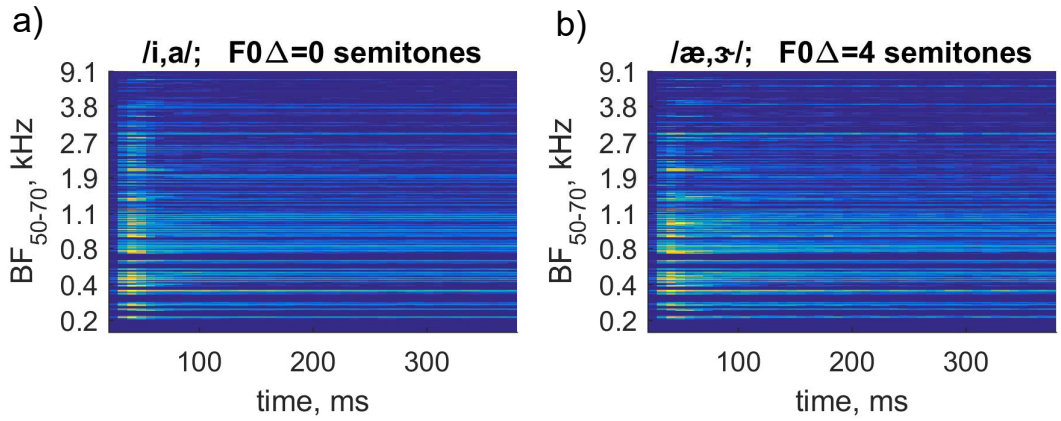


Figure 4.4 PSTHs in response to the concurrent-vowel pairs (a) /i,a/ and (b) /æ,ɝ/. The PSTHs compile MUA responses from the 300 electrode channels least likely to have measured responses to auditory stimuli by chance. Firing rates were averaged across 30 repeats of each stimuli and low-pass filtered (50 Hz cut-off). Spike rates are encoded by the colour map ranging from low (dark blue) to high (bright yellow).

4.2.3.2. Classification Back-end

The classification back-end was as stated in Section 3.2.3.

4.2.4. Analysis

4.2.3.1. Vector Strength

For a channel, the synchronization of neural responses to a temporal modulation was described by vector strength (VS; Goldberg and Brown, 1968):

$$\mathrm{VS} = \frac{1}{n}\sqrt{\left|\sum \sin(a_i)\right|^2 + \left|\sum \cos(a_i)\right|^2} \quad , \tag{4.2}$$

where n is the total number of spikes in a channel's analysis window, i represents an individual spike, and a_i is the phase angle of spike i relative to the modulation frequency of interest. A VS of 1 represents perfect synchronisation whereas 0 indicates no synchronisation.

4.2.3.3. Shuffled-autocorrelograms

Statistically robust, within channel, shuffled autocorrelograms (SAC; Joris, 2003; Joris et al., 2006) were computed. In response to individual and concurrent-vowel stimuli, SACs collated spike-train inter-spike intervals into histograms. Only across spike-train inter-spike intervals were compiled in to the SAC histograms, in order to remove the effects of neural refraction. The SAC had a normalisation factor (Louage, van der Heijden, & Joris, 2004) of the dimensionless unit 'normalised number of coincidences', described by:

$$N(N-1)r^2\tau\omega \quad , \tag{4.3}$$

where N denotes the number of spike trains that intervals are compared between (30), r is the average number of spikes per second, τ is the width of the histrogram bins ($\frac{1}{45}$ of the modulation frequency of interest), and ω is the total length of the analysis window (350 ms).

4.3. Results

4.3.1. Auditory Brainstem Responses and Compound Action Potential Audiograms

ABRs were made in four, out of six, animals. ABRs were recorded in response to short pure tones presented at 4 and 9 kHz. For all four animals, increasing peak-to-peak amplitude responses were observed with increasing pure tone levels. In addition to the ABR measurements, CAPs were recorded on the round window of two animals (Figure 4.5). In isolation these CAP measurements offer more detailed auditory profiles, measured across 15 frequencies (2 per octave between 0.25 and 32 kHz) and 15 levels (5 dB SPL steps between 30 and 100 dB SPL). As pure tones increased in level, the magnitude of the CAPs also increased. Tips of the CAP tuning curve were located at approximately 9 kHz, between 40 and 50 dB SPL. The CAP profiles correspond well with normal guinea pig auditory profiles.

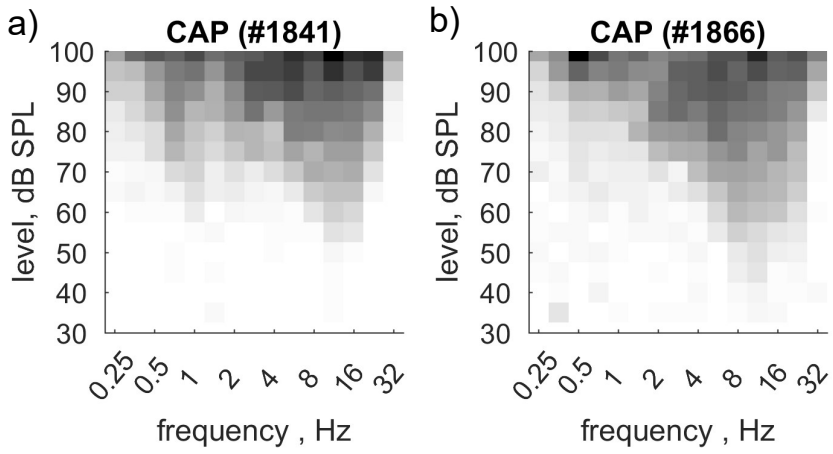


Figure 4.5 CAPs were recorded on the round window of two guinea pigs (#1841 (a) and #1866 (b)). Darker pixels represent increased magnitude of CAPs in response to pure tone frequency/level combinations.

4.3.2. Pure Tone Responses

Pure tone FRAs were derived for all electrode channels (e.g. see Figure 4.3). Looking at the relationship between spontaneous and driven firing in response to pure tones (bottom and top rows of FRA respectively) allowed the determination of whether a neural channel was likely to have measured responses to auditory stimuli through chance, and hence be unsuitable for further analysis. Figure 4.6 depicts the mean spontaneous firing rates (grey dots) and mean driven firing rates (green) for all probe channels. The horizontal location of these dots represents how different spontaneous and driven firing rates are within a channel, taking into account variability in firing rates (two-sample t-test). The left of the plot contains channels that were unlikely to be recording auditory driven responses by chance. It is apparent that the majority of channels were recorded from brain regions responsive to auditory stimuli.

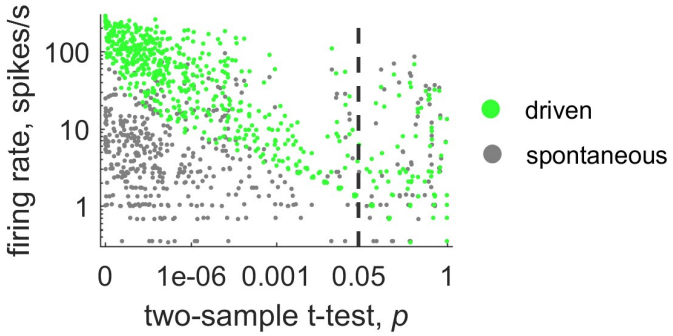


Figure 4.6. Spontaneous and driven rates in response to pure tones. The abscissa captures how significantly different driven and spontaneous responses were in a given electrode channel.

From the FRAs, two further metrics were calculated: the frequency to which a channel best responded when presented tones at a level between 30 and 50 dB SPL ($BF_{30\text{-}50}$), and similarly for between 50 and 70 dB SPL ($BF_{50\text{-}70}$). $BF_{30\text{-}50}$ and $BF_{50\text{-}70}$ had a resolution equal to that of the FRA, a quarter of an octave. Figure 4.7a displays the distribution of $BF_{30\text{-}50}$ in logarithmically spaced frequency bins (indicated by colour). The left side of the plot shows the $BF_{30\text{-}50}$ of a few channels, building up to the $BF_{30\text{-}50}$ distribution for the 300 channels least likely to have measured responses

to auditory stimuli by chance. Frequency bins are fairly equal in their width, with the exception of the highest frequency bin which is larger in size. This is because the tips of tuning curves tend towards lower thresholds for higher frequencies. Figure 4.7b is equivalent to Figure 4.7a but for $BF_{50\text{-}70}$. In this case, the width of the highest frequency bin reduces in size to be equivalent to the others. Overall, the observed distributions of $BF_{30\text{-}50}$ and $BF_{50\text{-}70}$ are consistent across the abscissa of Figure 4.7a and Figure 4.7b, and depict frequency tuning that would be expected in the central nucleus of the IC.

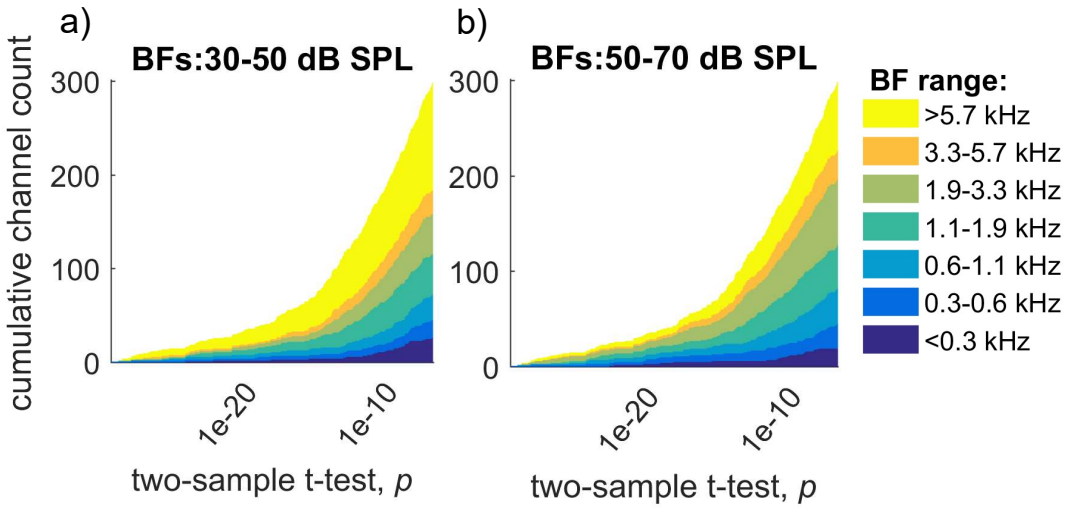


Figure 4.7 Channels that are likely driven by auditory stimuli are cumulatively displayed. BFs, measured as the frequency to which a channel responded most between 30-50 (left) or 50-70 dB SPL (right), are colour coded and displayed proportionally.

4.3.3. Identification of Concurrently Presented Vowels

Neural data was recorded in response to a complete set of CVI experimental stimuli. These responses (MUA in 300 electrode channels, selected from across 6 GPs, 50 Hz low-pass filter cut-off) were used to train and test two variations of a computational decision model. Initially, the model was suboptimally configured to identify vowels independently of one another (Eq. 3.5). The model's fitting parameter (additional Gaussian noise) was adjusted such that the model's average performance, when vowels were presented with identical F0s, best matched human listeners' (72.6%). However, the model was unable to exceed 18.0% correct (blue diamond in Figure 4.8) and in this instance predictions were only loosely correlated with listeners' predictions (PCC of $r = 0.31, p \ll 0.01$).

Alternatively, the model employed an optimal strategy of predicting the combined representations of concurrently presented vowels (Eq. 3.7). As was the case for the suboptimal scheme, the model's fitting parameter was set to best match the model's mean performance with that of listeners', when vowel pairs had identical F0s. This was done separately for listeners' training and testing data. In both instances, the model predicted an improvement in performance as the F0 difference between concurrently presented vowels increased. The F0 benefits predicted were qualitatively and quantitatively similar to listeners' (see Figure 4.8). When fitted to match listeners' performance during training, between F0 differences of 0 and 4 semitones, the model predicted an improvement of 15.5%, and when fitted to listeners' performance during testing the predicted improvement was 16.4%. In fact, at all 6 F0 differences the model's predictions were similar to listeners (RMSE of 3.9% and 2.3% for training and testing respectively).

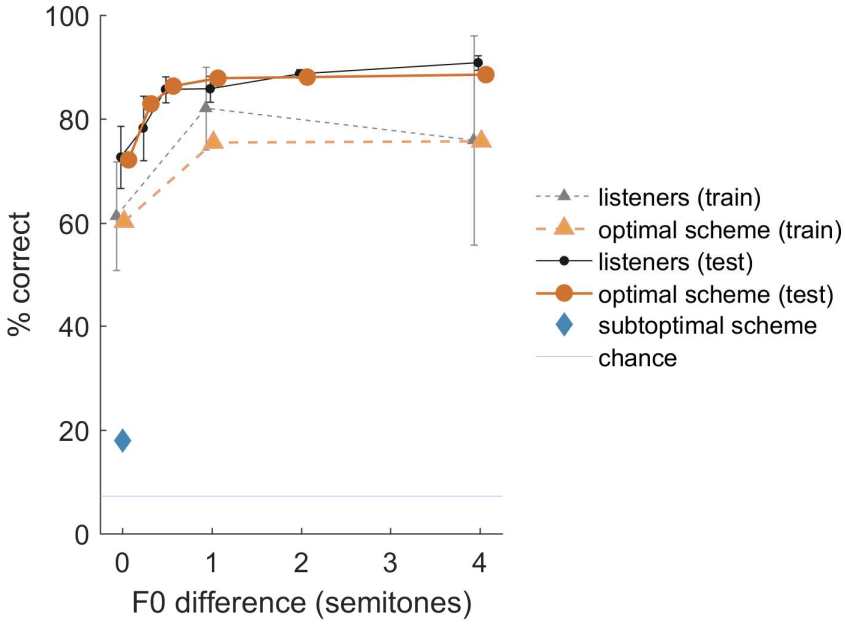


Figure 4.8 Mean percent correct, as a variable of the F0 difference between concurrent-vowels, predicted by: human listeners and the neural classifier (low-pass cut-off 50 Hz). The classifier was configured to implement an optimal scheme as well as a suboptimal scheme. Data is shown for the task's training and testing sessions.

Additionally, this optimal configuration of the model predicted individual decision probabilities that were highly correlated with listeners'. The PCC between the model fit to training data, and listeners', for all decision probabilities, across all F0 differences was $r = 0.94$ ($p \ll 0.01$). For the model fit to testing data, the according PCC was $r = 0.97$ ($p \ll 0.01$). Moreover, high PCCs were calculated conditionally at each F0 difference (see PCCs in black text in Figure 4.9). For reference, the confusion matrix predicted by the model fit to testing data, when vowels were presented with identical F0s, can be found in Table 4.1. For the testing data and the accordingly fit model, when isolating and correlating the decision probabilities that were correct identifications, the lowest PCC was $r = 0.27$ ($p \ll 0.01$) for concurrent-vowel pairs with an F0 difference of 0.25 semitones and the highest PCC was $r = 0.52$ ($p \ll 0.01$) for an F0 difference of 2 semitones. When isolating decisions that were errors, the lowest PCC was $r = 0.28$ ($p \ll 0.01$) for concurrent-vowel pairs with an F0 difference of 1 semitones and the highest was

$r = 0.45$ ($p \ll 0.01$) for an F0 difference of 0.25 semitones. There was no systematic relationship between PCCs calculated between correct identifications or errors, and the magnitude of F0 differences at which they were calculated.

		Concurrent-vowel pairs identified, %														
		/i, i/	/ɑ, ɑ/	/u, u/	/æ, æ/	/ɝ, ɝ/	/i, a/	/i, u/	/i, æ/	/i, ɝ/	/a, u/	/a, æ/	/a, ɝ/	/u, æ/	/u, ɝ/	/æ, ɝ/
Concurrent-vowels presented (model)	/i, i/	97						3								
	/ɑ, ɑ/		60		3		1				1	21	11	2		1
	/u, u/			100												
	/æ, æ/		4		55				6			29		6		
	/ɝ, ɝ/		1			60				1	1	1	8	8	8	12
	/i, a/		1				73		8	10	1		2	5		
	/i, u/							100								
	/i, æ/				2		5		49	5		1		37		1
	/i, ɝ/						9		5	76				9		1
	/a, u/						1				91		2	1	4	1
	/a, æ/		13		17		1		3			49	4	8		6
	/a, ɝ/		6			5	1		1		17	3	48	15	1	3
	/u, æ/				1		1		5	1	1	1	2	88		1
	/u, ɝ/					1					3		1	1	94	
	/æ, ɝ/				1	3	1		3	7	1	6	4	30	1	45

Table 4.1 Confusion matrices predicted by the proposed model with neural data when vowels had no F0 difference. Correct identification entries are shaded in light green. Confusions that listeners made more than 10% of the time are shaded in light yellow.

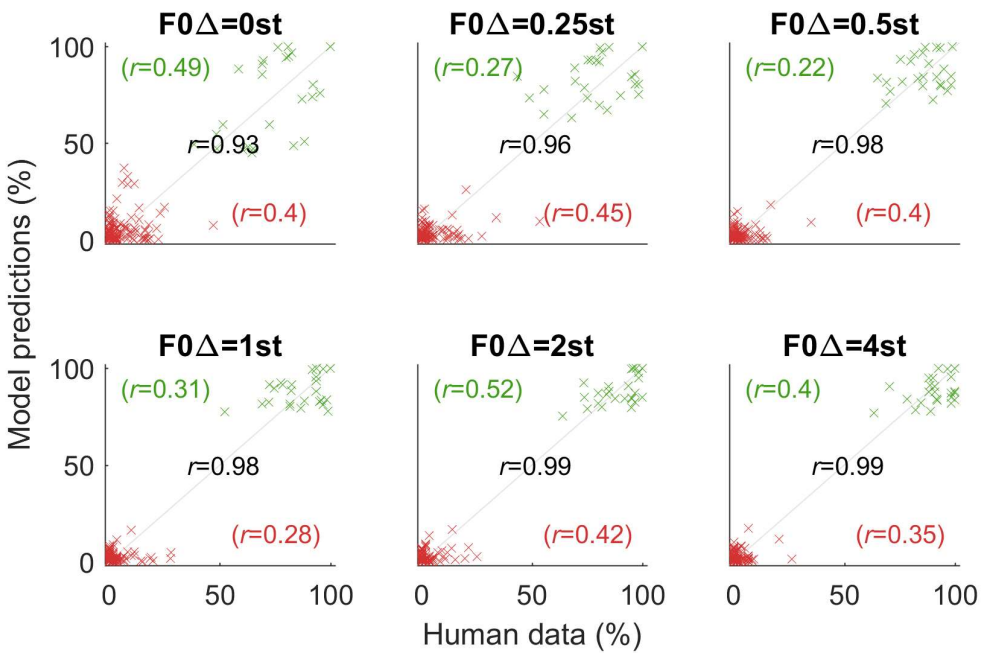


Figure 4.9 Scatter plots as in Figure 3.8, but for the proposed model of neural classification (300 channels, 50 Hz low-pass cut-off frequency, fit to testing data).

4.3.4. Low-pass Meta-parameter

The model presented in Section 4.3.3 used neural data that had been low-pass filtered to preserve temporal variations below 50 Hz. This cut-off frequency was selected from five options (0, 10, 50, 100, 500 Hz) as it gave the best quantitative prediction of listeners' F0 benefits. Regardless, the optimal scheme configuration of the model (fit to listeners' testing data), was also run at the four alternative low-pass cut-off frequencies. All iterations of the model predicted improvements in identification as the F0 difference between vowel pairs increased from 0 semitones to 0.25 semitones, and these improvements were larger in magnitude for F0 differences of 4 semitones (see Figure 4.10a). As the cut-off incrementally increased from 0 Hz to 100 Hz the size of the model's 4 semitone F0 benefit increased, over predicting listeners' F0 benefit by 1.5% for a cut-off at 100 Hz. For a higher low-pass cut-off frequency of 500 Hz, the model predicted a smaller F0 benefit than for 100 Hz. In the GP, neural synchronisation has only been observed for temporal modulations below 150 Hz (Joris, Schreiner, & Rees, 2004). Hence, a low-pass cut-off frequency of 500 Hz is likely only adding noise to the representations of the stimuli.

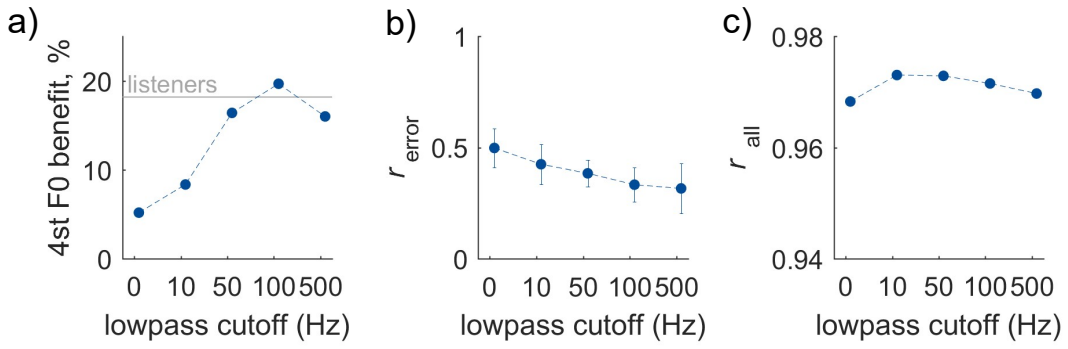


Figure 4.10 (a) Improvement in identification with an F0 difference of 4 semitones. PCC of errors (b) and all decisions (c) between the model and listeners.

At all low-pass cut-offs, the probabilistic confusion matrices predicted by the model correlated well with listeners' confusion matrices (Figure 4.10b and 4.9c).

However, as the model's low-pass cut-off increased in frequency, the PCCs between the model's prediction of errors and listeners' errors decreased. The model best predicted listeners' errors when using purely spectral representations of concurrent-vowel pairs (PCCs ranged between $r = 0.42$ and $= 0.64$, $p \ll 0.01$ for both). Correlating all of the model's decisions with listeners' shows how, overall, low cut-off frequencies (10 and 50 Hz) best predicted listeners' decisions. At these cut-offs, the model was able to predict listeners' F0 benefits, whilst partially accounting for the individual decisions, such as errors, that listeners made. Worth noting is that, all of the confusion matrices predicted by the optimal scheme of the model effectively saturate Chintanpalli and Heinz's (2013) normalised similarity metric.

4.3.5. Features Informing the Model

The proposed computational model, configured to classify neural responses, was able to predict listeners' behaviour when performing CVI. However, given the relatively few assumptions made beyond the model's optimal strategy, the auditory features influencing the model remain to be characterised. As in Section 3.3.4, dependence between representational features and the model's decisions can be interpreted using MI.

Figure 4.11a shows the MI between features and all of the model's decisions (darker pixels represent higher MI). MI is maximal immediately after the onset of a concurrent-vowel pair, prior to neural adaptation. This can clearly be seen above the main 2D plot in Figure 4.11, where MI is summed across channels. To the right of Figure 4.11's 2D plot is the MI for each channel (indexed by their $BF_{50\text{-}70}$) summed over time. It is apparent that a peak in MI occurs in channels best tuned to low frequencies between 0.5 and 1.5 kHz, a region found to be crucial for predicting an F0 benefit (Darwin & Culling, 1993). This frequency range corresponds to the first and second formant of the experimental vowel stimuli. For channels best responsive to frequencies above 4 kHz, there is a drop in MI corresponding to a lack of energy above 4 kHz in the vowel stimuli.

It was viable to observe how MI changed in relation to changes in the F0 differences between concurrently presented vowels (as in Figure 3.10d). This was done in an attempt to characterise representational features driving improvements in the model's predictions with increasing F0 differences. Figure 4.11b displays negative (blue) and positive (red) changes in MI as concurrent-vowel F0 differences increase beyond 0 semitones (to 0.25, 0.5, 1, 2, and 4 semitones). Changes in MI, summed across time (right of the plots), reveal seemingly erratic fluctuations, both positive and negative. Changes in MI were also summed across channels (above the plots). This provided an intuition as to temporal characteristics influencing improvements in the model's predictions. For all increases in F0 difference (particularly noticeable for 2 and 4 semitones), there were features periodically

spaced in time that had larger increases in MI. The relevance of these temporal periodicities will be further examined in Section 4.3.7.

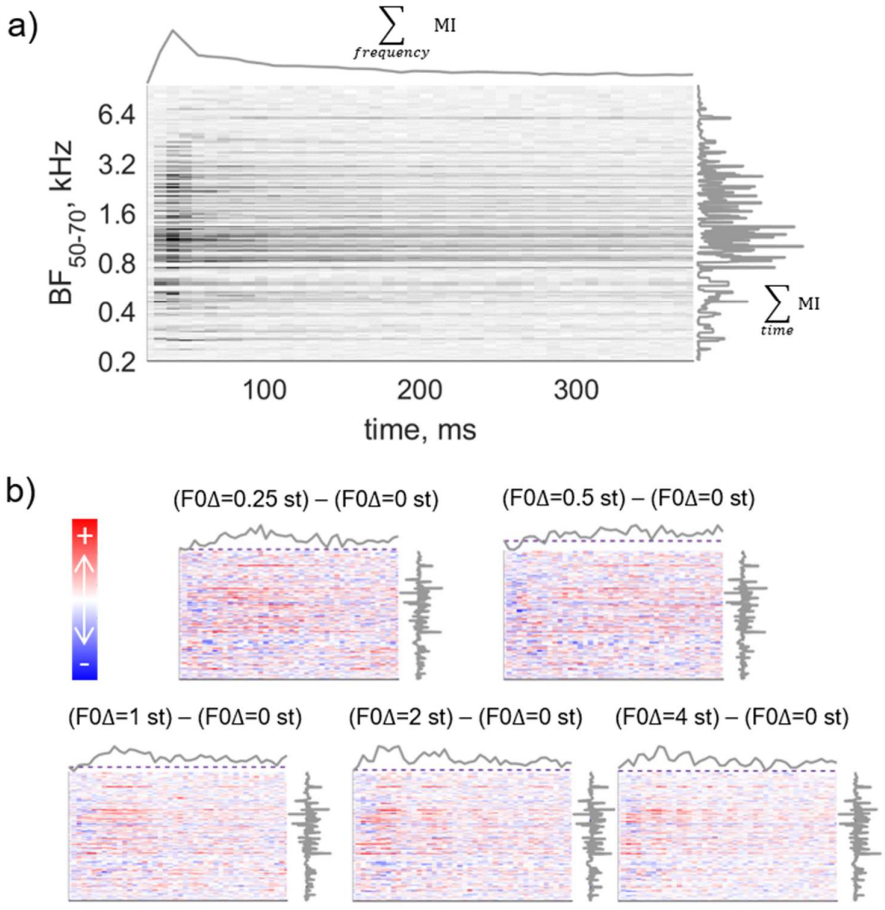


Figure 4.11 Plots displaying to what degree each representational feature informed: (a) all the model's decisions (b) changes in the model's predictions when presented vowel pairs with F0 differences larger than 0 semitones (0.25, 0.5, 1, 2, and 4 semitones).

4.3.6. Increased Firing Rates with Fundamental Frequency Differences

Average firing rates, in response to F0 differences between concurrent-vowel pairs increasing from 0 to 4 semitones, increased from 30.4 to 31.8 spikes per second. Absolute firing rates were found to positively correlate with absolute MI ($r = 0.67, p \ll 0.01$; calculated within time-channel bins). This findings suggests a relationship between increased firing rates and improved identification of concurrent-vowels at larger F0 differences. Figure 4.12a displays changes in firing rates compared with changes in MI for 5 increases in F0 difference, with a linear regression fit to each. As the magnitude of the change in F0 difference increased, so did the positive gradient of the linear fit (β coefficients (with 95% confidence bounds) for 0.25 semitone: 278 (262.5, 294.9), 0.5 semitone: 333 (316.9, 349.4), 1 semitone: 390 (373.1, 406.7), 2 semitone: 410 (392.9, 427.4), 4 semitone: 504 (486, 522.9)). This implies that for larger increases in the F0 difference between vowels, the same increase in firing rate corresponds to a larger increase in MI. Please note that these results should be interpreted with an understanding that the measures of MI are numerical approximations, and this in itself will produce a small decorrelation of the data.

There are a number of possible explanations for the increase in firing rates with F0 differences. For example, increased firing rates could be explained by the introduction of higher frequency spectral energy in vowel pairs with larger F0 differences. If neurons tuned to higher frequencies emit more spikes relative to neurons tuned to lower frequencies, an F0 difference could produce increased firing rates. Figure 4.12b shows mean changes in firing rates averaged across channels with the same $BF_{50\text{-}70}$. However, rather than a positive trend between $BF_{50\text{-}70}$ and increased firing rates, mainly positive changes alongside fewer negative changes in firing rates are inconsistently distributed across the length of the abscissa. Channels best tuned to 0.48 kHz saw the largest mean increase in firing rates, whereas channels best tuned to 1.13 kHz saw the largest mean decrease.

Additionally, if the proposed increase in firing was a product of some form of linear gain at higher frequencies, increased deviation between the firing rates encoding the concurrent-vowel pairs would be counteracted by an increased variance with which each individual concurrent-vowel pair is encoded. This would not result in more MI with F0 differences.

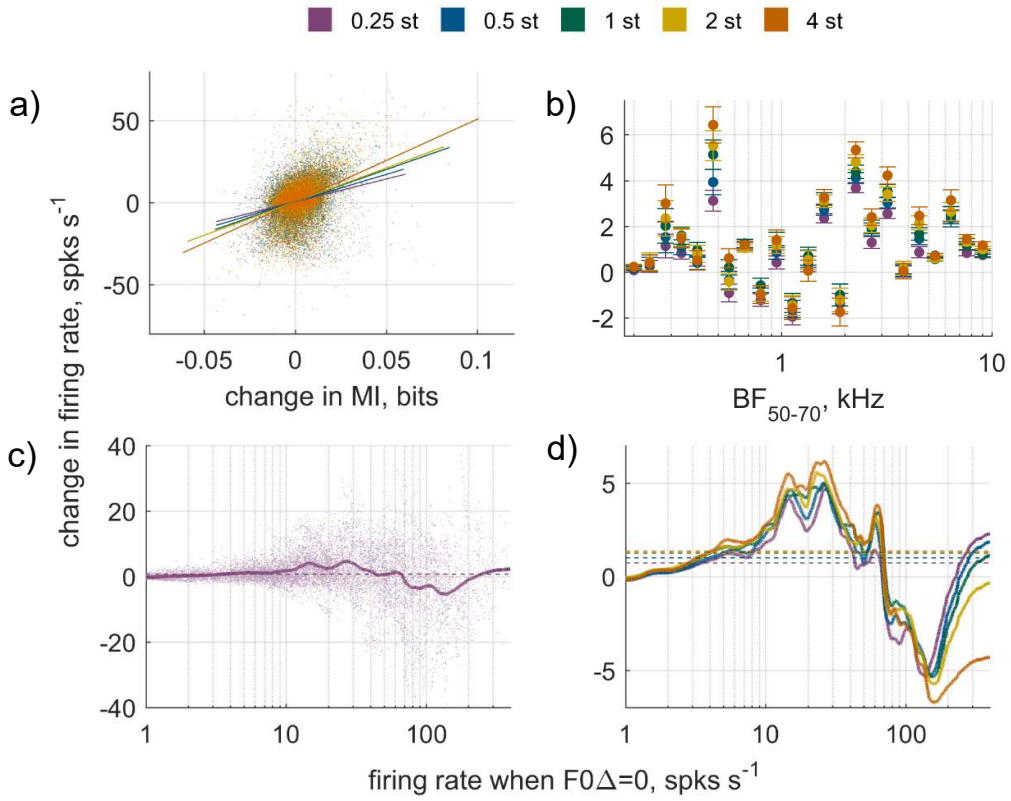


Figure 4.12 Shows changes to firing rates (0.75 ms time windows) in response to vowel pairs increasing in their F0 differences relative to no F0 differences (see legend; semitone is abbreviated as st). (a) Changes in MI are compared to changes in firing rates. Each dot represents a representation feature, i.e. a time/channel bin, and linear fits to data are overlaid. (b) Changes in firing rates are plotted as a function of channel $BF_{50\text{-}70}$. (c) Firing rates in response to vowel pairs with no F0 differences are plotted against changes in firing rates with the introduction of a 0.25 semitone F0 difference. A locally weighted linear regression is overlaid (span: 0.05). (d) Locally weighted linear regressions, like in plot (c), are displayed for all F0 differences. Mean change in firing rates are represented by dashed lines.

Figure 4.12c and Figure 4.12d show how, relative to firing rates in response to vowels with identical F0s, firing rates change as F0 differences are introduced between vowel pairs. Instances where vowel pairs with identical F0s were encoded with low firing rates (10-50 spikes per second) saw increases in their firing rates with F0 differences. This was the case for the majority of time-channel bins. Alternatively, in the minority of instances where vowel pairs were encoded with large firing rates, firing rates dramatically dropped with the introduction of F0 differences. Averaging across this majority of small positive changes, and the minority of large negative change, saw an overall positive increase in firing rates with F0 differences (horizontal dashed lines Figure 4.12d).

The mechanics of cochlear filtering can go a way to explain the observed changes in firing rates (and hence increase in MI with F0 differences). When vowel pairs have no F0 differences, harmonics of the constituent vowels are concentrated within few cochlear filters. Neurons connected to these regions of the cochlea will have relatively high firing rates (right side of abscissa in Figure 4.12c and Figure 4.12d). Neurons connected to other regions will have relatively lower firing rates (left hand side of abscissa in Figure 4.12c and Figure 4.12d). When an F0 difference is introduced between concurrent-vowel pairs, the harmonics of the constituent vowels will spread across the spectrum. Consequently, neurons previously innervated by harmonics of both vowels will reduce in their firing rates as one of the harmonics exits the corresponding cochlea filter. In turn, this widespread distribution of harmonic spectral information will increase the firing rate in a majority of neurons.

This explanation clarifies the dynamics observed in Figure 4.12c and Figure 4.12d. However, it does not explain the increase in MI with F0 differences, attributable to increasingly distinct firing rates describing different vowel pairs. In a purely linear system, the changes in firing rates would cancel each other out. However, the auditory system is not linear but known to be compressive in nature. A difference in energy at high levels is less well represented by the same difference at low levels.

Correspondingly, as vowel pairs increase in their F0 differences, energy levels are lower but spread across more neurons. Thus, concurrent-vowel pairs become more distinctly represented due to being described by energy at the lower levels of a compressive function. The dynamic range of the system is better used. These intuitions are partially verified in a toy model of CVI (where vowels are represented by a single harmonic at their F0, only two Gammatone filters) depicted in Figure 4.13. In the toy model example, the difference between concurrent harmonic pairs is 1.9× larger when they have an F0 difference of 4, relative to 0, semitones. Arguably these effects could be exaggerated in an auditory system with narrower filters and different, sharper, compression functions.

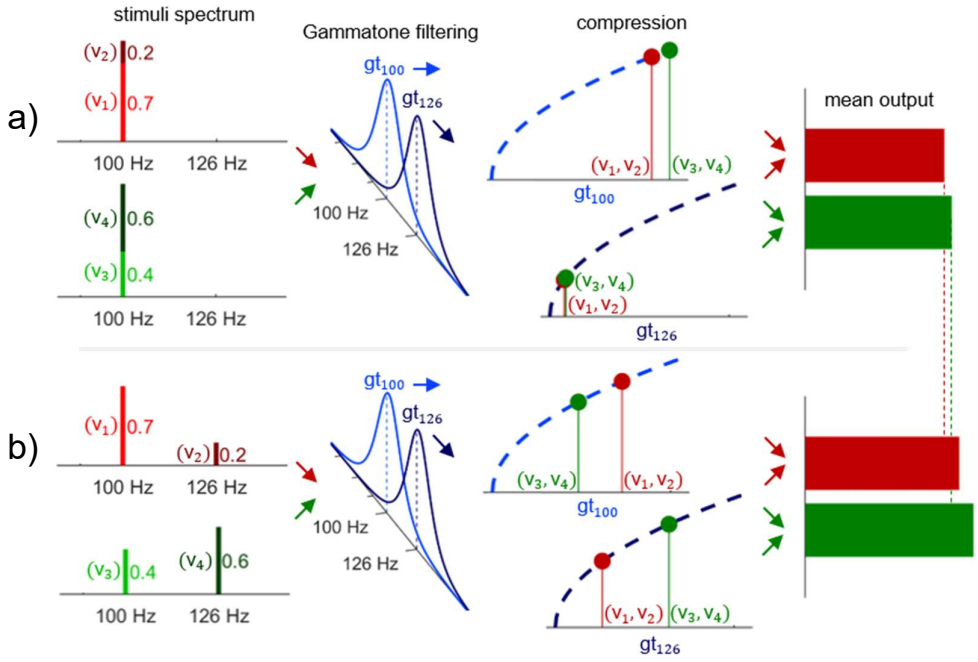


Figure 4.13 Filter theory explanation. Two concurrent-vowel pairs, v_1,v_2 (reds) and v_3,v_4 (greens) are illustrated. Each vowel is defined as a single harmonic of a particular numerical magnitude. (a) is representative of a vowel pair with no F0 difference, hence why the vowels are stacked in the stimuli spectrum, and (b) represents a vowel pair with an F0 difference of 4 semitones. These stimuli are passed through two Gammatone filters, centred on 100 Hz (gt_{100}) and 126 Hz (gt_{126}). If we look at (a), the filter outputs in response to the two concurrent-vowel pairs are depicted by lines emanating out of abscissas associated with each Gammatone filter. These filter outputs

are separately compressed for each concurrent-vowel pair, and then averaged. These averages are shown as horizontal bar graphs allowing comparison between concurrent-vowel pairs, with and without F0 differences.

The idea that CVI is driven by changes in firing rates, explained by rudimentary cochlear mechanics, is a new proposal. The proposal also helps to explain F0 benefits observed at small F0 differences, where previously proposed temporal segregation mechanisms would fail and/or are unrealistic. However, this filter-based explanation is a spectral argument, yet the proposed model improved in its predictive power when retaining low frequency temporal information. Why is this?

4.3.7. The Role of Beating

MI changes indicate representational features that may be improving predictions as F0 differences are introduced between vowels. In Section 4.3.5 this idea motivated investigation in to the role of changes in firing rates with F0 differences. Salient in Figure 4.11b are changes in MI with F0 differences that are characterised by regular temporal fluctuations. These oscillations seem to have frequencies corresponding to the unresolved interactions of unresolved harmonics, i.e. beating. Although noticeable for all changes in F0 differences, this is especially noticeable for larger changes in F0 differences where beating frequencies are higher. Figure 4.14 plots temporal changes in MI, averaged across electrode channels, for all changes in F0 differences relative to no F0 differences (in black). Overlaid are sinusoids with frequencies matching the beating frequencies of the F0s of the vowel pairs analysed (in violet). The two plots correspond very closely, for all of the 5 displayed increases in F0 differences. However, it is unclear whether beating is the cause of improvements with F0 differences, or an epiphenomenon.

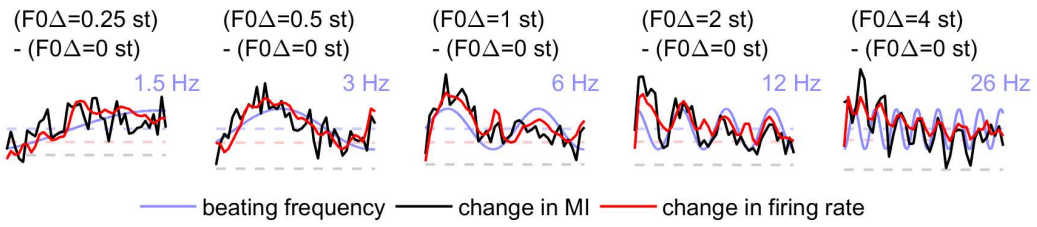


Figure 4.14 Normalised changes in MI compared to normalised changes in firing rates. Changes were summed across electrode channels and calculated for changes in F0 differences of 0.25, 0.5, 1, 2, and 4 semitones (relative to no F0 differences). Beating frequencies of concurrent-vowel pair F0s are superimposed. Faintly coloured dashed lines represent adjusted zero values.

A previous proposal for how beating may facilitate improvements in CVI is that beating creates glimpses of individual vowels due to periodic suppression of harmonics associated with the competing vowel (Culling and Darwin, 1994). If the case, then these glimpses would occur in the trough of beating oscillations and an

according increase in MI would co-occur. In Figure 4.14, temporal changes in MI are displayed alongside changes in firing rate (in red). However, increases in MI are temporally aligned with peaks in firing rates, not the troughs.

Additionally, for each electrode channel, the amount of beating in the neural firing responses, within separate channels, was quantified using VS. VS determined the degree to which spikes clustered around a given beating period. Figure 4.15 shows the relationship between a channel's VS and the same channel's mean change in MI, for an increase in F0 difference of 2 semitones. In this instance, and for an F0 difference of 4 semitones, VS was not positively correlated with changes in MI ($r = -0.32, p \ll 0.01$ and $r = -0.03, p = 0.56$ respectively). In fact, for an F0 difference of 2 semitones, a negative trend was observed. This could be because channels that saw a large drop in firing rate (and hence MI) observed a spreading of harmonics with F0 differences, making them vulnerable to unresolved harmonic beating (hence high VS). A negative trend was not found for an F0 difference of 4 semitones, but this may be due to the high frequency of beating relative to the resolution of the PSTHs (hence the high p value). Analysis was not performed at F0 difference with a beating period less than one quarter the length of the windowed stimulus (>5.7 Hz). Beating if anything is a hindrance. Yet, this characterisation could explain the need for temporal resolution in the proposed model. Without this resolution, information about concurrent-vowel pairs that is well represented in the peaks of beating oscillations would be averaged with inferior information in the troughs. This would result in the prediction of a reduced F0 benefit, much like that predicted by the '0 Hz' iteration of the model.

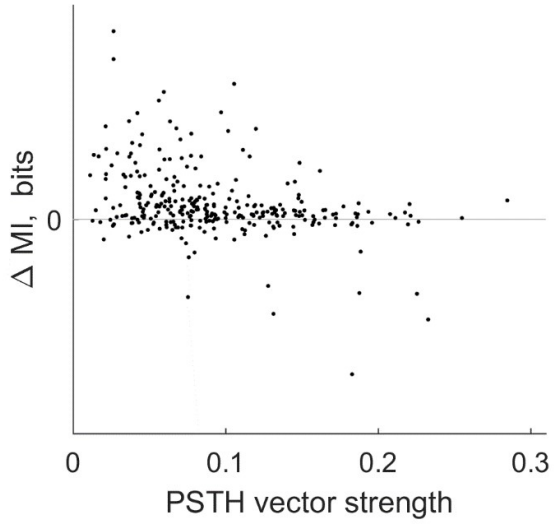


Figure 4.15 Each point represents an electrode channel. The location of the data point represents VS (12 Hz) calculated in response to vowel pairs with an F0 difference of 2 semitones (abscissa), and changes in MI as vowel pairs have an F0 difference of 2 semitones introduced (ordinate).

4.3.8. Periodicity Coding of Fundamental Frequency

In Section 4.3.4, a model was presented that was trained/tested on neural responses that had temporal variations up to 500 Hz retained. However, this model, that had the highest temporal resolution of all the model variations presented, had arguably the worst predictive performance. Never-the-less, high frequency temporal oscillations have been instrumental in previously proposed models of CVI that are predicated on extracting F0 information (see Section 1.5.1). Meddis and Hewitt (1992) went as far to attribute the improvement in their model's performance directly to an improvement in the model's ability to perform F0 segregation, which was based on high fidelity temporal analysis. Additionally, neurophysiology studies have shown that the auditory midbrain, in response to concurrent harmonic series, can temporally encode F0 information (see Section 1.5.2). Further, brainstem neurons lock to ENVs in frequency regions corresponding to F0 information (Liu, Palmer, and Wallace, 2006; Sayles, Stasiak, and Winter, 2014). Given this context, it is pertinent to discover whether temporal coding of F0 cues in the IC increase in prominence with F0 differences.

For each electrode channel, in response to individual vowels and concurrent-vowel pairs, normalised SACs were calculated in order to quantify the degree to which F0 information was encoded in high frequency periodicities. SACs reveal the distribution of normalised spike coincidences at a given time delay (at a high resolution $\leq$ 0.22 ms, not subject to low-pass smoothing as in model), statistically strengthened by comparing intervals ac**r**oss stimulus repetitions within any given channel. Figure 4.16 provides an example of an SAC for the concurrent-vowel pair /i,æ/ with an F0 difference of 4 semitones. Clear peaks can be seen above 1 on the ordinate axis (where 1 represents the absence of any temporal coding) corresponding to increased temporal coding of the F0 periodicities 100 and 126 Hz (lags of 0.01 s and 0.0079 s respectively).

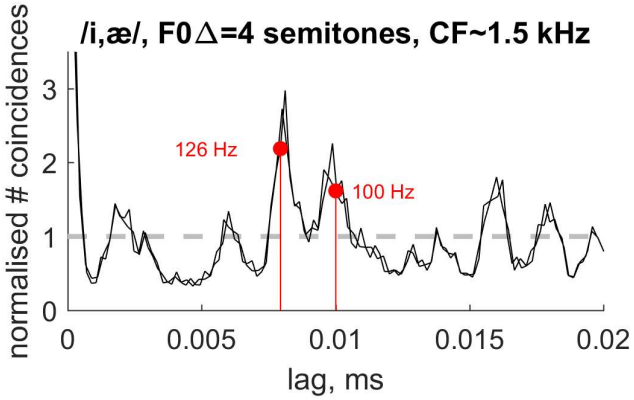


Figure 4.16 Example SACs calculated using MUA from a single electrode channel. There are quite clearly two peaks for 1 cycle in the 100 Hz scaled, and 126 Hz scaled, SAC. The SACs are calculated uniquely for each F0, hence the overlay of two SACs in the same plot.

Channels were found to temporally encode the F0s of individual vowels (green and pink points in Figure 4.17). Channels were also found to encode the F0s of concurrently presented vowels. However, on average, F0s were more poorly encoded when vowels were presented concurrently, compared with vowels presented individually. Notably, the normalised number of coincidences dropped as a function of F0 difference, and this decrease was larger for concurrent-vowel pairs than it was for the constituent individual vowels. When vowels had no F0 differences, concurrent-vowel pairs encoded constituent F0s with 1.95 normalised number of coincidences, compared to 1.2 at an F0 difference of 4 semitones.

The salience of periodicity coding varied across different vowel identities when presented individually and in combination with other vowels. The saliency of F0 coding seems to qualitatively be related to spectral energy, notably second formant frequencies. For example, /i/ had the highest frequency second formant (2250 Hz, ≥ 800 Hz higher than the other 4 vowels) and had the highest averaged normalised number of coincidences. On the other hand, /u/ had the lowest frequency second formant (850 Hz) and had the lowest average normalised number of coincidences. Additionally, /u/ was the only vowel that showed more salient periodic coding of F0 information for concurrent-vowel pairs than individual vowels (at F0

differences < 1 semitone). Here, however, no further attempt is made to explain this relationship between vowel identity and temporal encoding of F0. This is mainly because the SAC analysis indicates that temporal encoding of F0 information is not more salient at larger F0 differences. This is further evidence against models of temporally guided F0-based segregation.

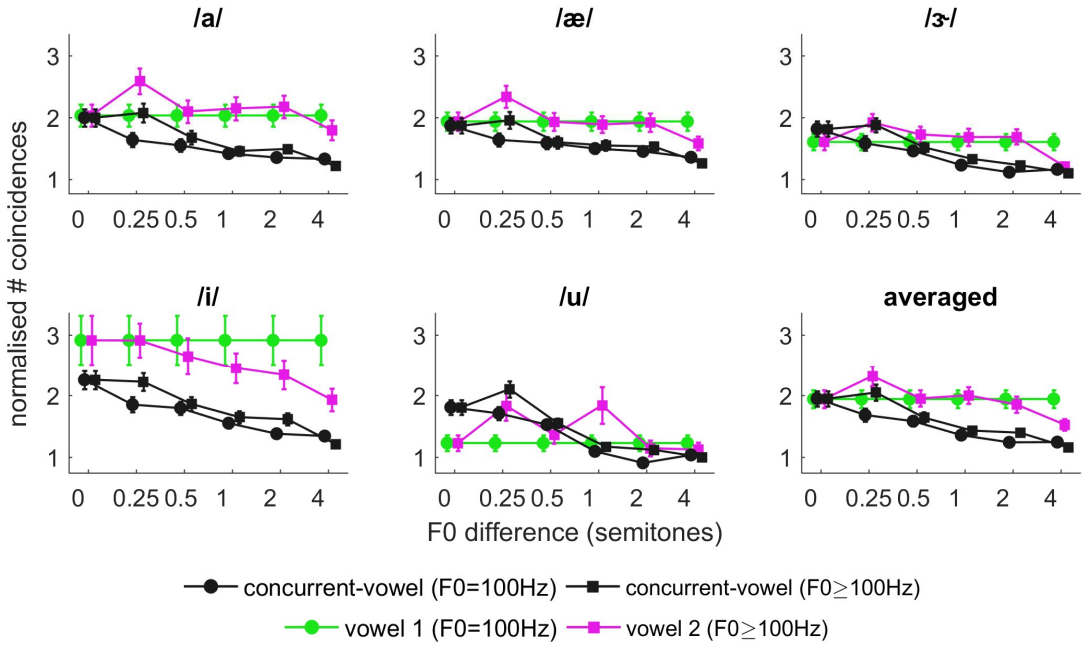


Figure 4.17 The normalised number of coincidences, averaged across all channels. Number of coincidences was calculated for concurrent-vowel pairs at the periods of the component F0s. Number of coincidences were also calculated individually for a pair's constituent vowels at the period of their F0 (a green circle is vowel 1: the constituent vowel with F0 of 100 Hz; a pink square is vowel 2: the constituent vowel at an F0 equal to or higher than 100 Hz). Results are displayed for 6 F0 differences, for each vowel identity as well as averaged across all vowel identities.

4.3.9. Approximating the Representation of Concurrently Presented Vowels

The proposed optimal model of CVI was trained on the combined representation of concurrently presented vowels. However, the means by which the auditory system might implement this strategy has so far been unconstrained. This may be achieved through high-level predictive coding, or a purely bottom-up process. How accurately can the combined neural representation of concurrent-vowel pairs be predicted using representations of individual vowels? By the IC, has any segregation occurred? To unravel these two ideas, I compared the neural representation of concurrent-vowel pairs with representations of the constituent individual vowels.

Neural firing rates were collated and compared. Figure 4.18 shows the isolated (crosses) and averaged (lines) firing rates of 300 channels in response to the concurrent-vowel pair /i,u/ (black) at 3 F0 differences. As F0 difference increases, average firing rates increase in the manner described in Section 4.3.6 (although here difficult to observe visually). Overlaid, at the 3 F0 differences, are firing rates in response to the constituent individual vowels /i/ (green) and /u/(magenta). Responses to concurrent-vowel pairs conform closely to those of the individual vowels. This observation is in contrast to sounds presented in air which superpose on one another.

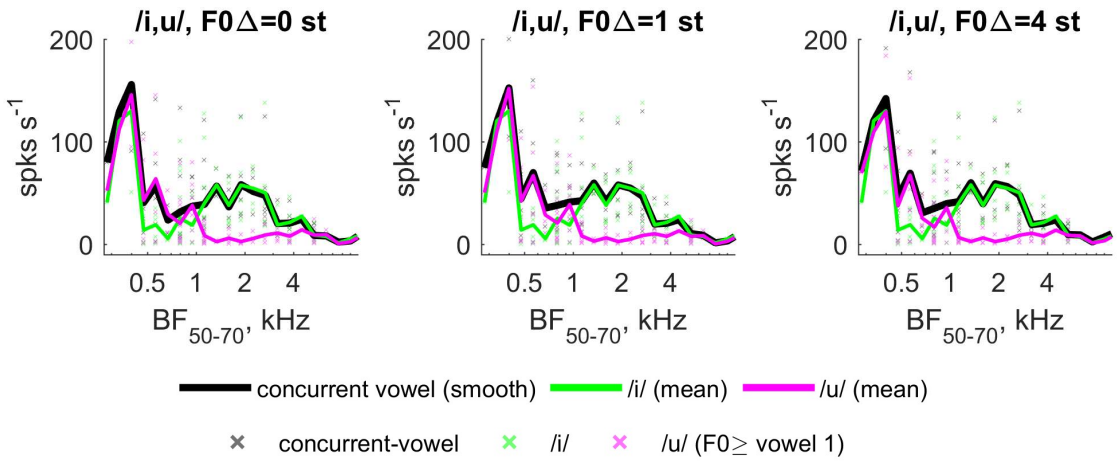


Figure 4.18 Firing rates of channels in response to concurrent-vowel pair /i,u/, and its constituent vowels, at 3 F0 differences (x markers). Firing rates are also shown averaged across $BF_{50\text{-}70}$ (sold lines).

Generalising the observations made about Figure 4.18 to all of the presented concurrent-vowel pairs, it was found that the linear summation of individual vowel representations (firing rates) was a poor predictor of the associated concurrent-vowel pair representations (RMSE of 35.9 spikes per second). This disparity was less when fitting a scaled linear function (scaling of 0.61 results in RMSE of 13.7 spikes per second). In Figure 4.18, it qualitatively appears that the representation of an individual vowel dominates any one region in the concurrent-vowel representation. Correspondingly, concurrent-vowel firing rates were better predicted by selecting the larger of the constituent individual vowel firing rates within a channel (*maximum* function results in RMSE of 12.3 spikes per second). The ideal approach to estimating the combination of stimuli in a logarithmic system would be to reverse the logarithm (i.e. apply an exponential), summate, and then reapply the logarithm. This is more formally known as a l*ogsumexp* function and is a commonly applied smooth approximation to the *maximum* function in ML literature. The *logsumexp* function predicted the combined neural representation of concurrently presented vowels even better than the *maximum* function (RMSE of 11.29 spikes per second for a natural logarithm scaled by 106.4).

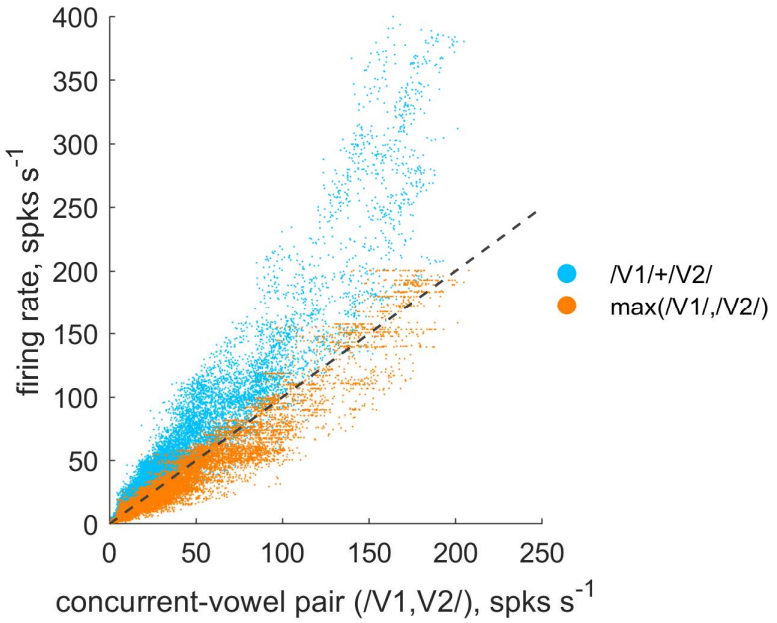


Figure 4.19 Comparison of firing rates in response to concurrent-vowel pairs (/V1,V2/), with different formulations of the constituent individual vowel firing rates (/V1/,/V2/). Blue dots represent a linear combination, and orange the output of a *maximum* function. The dashed black line represents unity.

4.4. Discussion

Of the limited number of studies that have investigated neural responses to concurrent-vowel pairs, all have used small stimulus sets (Fishman, Micheyl & Steinschneider, 2016; Keilson, Richards, Wyman, & Young, 1997; Palmer, 1990). Likely due to this, neural responses to concurrent-vowels have not previously been used to make empirical predictions directly comparable with listeners' psychoacoustic data. In this chapter, an optimally configured computational model integrated with neural recordings made in response to a full set of CVI experimental stimuli was able to quantitatively predict listeners' F0 benefits, as well as their specific decision probabilities.

This predictive performance was similar to the predictive performance of the model's back-end classifier, but fronted with a simple filter simulation of the auditory periphery (Chapter 3). In itself this attests to the robustness of the proposed framework. As was also the case for the model investigated in Chapter 3, information-theoretic analysis demonstrated that neuronal responses best tuned to frequency ranges associated with the first and second formants of the presented vowel stimuli were most informative to the model. Chintanpalli and Heinz (2013) found that the spectral distance between the first and second formants of vowels could explain 52% of variance in F0 benefit (larger distance $\propto$ smaller F0 benefit). Additionally, listeners have demonstrated an F0 benefit for vowel pairs modified such that differences in F0 are constrained to the first formants of vowel, but not when restricted to higher formants (Darwin & Culling, 1993). These findings corroborate the presented MI analysis indicating the importance of lower formants for CVI.

Previous studies have found that at the level of the AN, although individual vowels in quiet can be sufficiently encoded in rate-place representations, vowels amongst a wideband noise or another vowel were most robustly represented as temporal codes (Palmer, 1988; Sachs, Voigt, & Young, 1983; Sachs & Young, 1979).

Yet, by the level of the macaque AC, concurrent-vowel pairs are represented by a rate-place code (Fishman, Micheyl & Steinschneider, 2016). For CVI, this implies a temporal-rate coding conversion as information ascends the auditory system. There has however been evidence in support of spike timing with a precision of 1-10 ms for consonant encoding in the AC (Engineer et al., 2008). For both variations of the proposed decision model (Chapter 3 and Chapter 4), the predictions most similar to listeners' were made when an intermediate amount of temporal information was preserved in the representation of concurrent-vowels (representations low-pass filtered to attenuate frequencies above 100 Hz and 50 Hz respectively for the filter simulation of the auditory periphery and the IC data). It could be interpreted that both types of representations characterise a midpoint along a temporal-rate conversion. As predictions made with the IC data can largely be accounted for by a simple simulation of the auditory periphery, rather than a conversion involving temporally based harmonic extraction mechanisms (like those proposed in Meddis and Hewitt (1992)) this conversion could merely be a loss of high fidelity oscillations.

An SAC analysis of the unfiltered IC neural responses to vowel stimuli provided evidence against an F0-guided segregation mechanism predicated on TFS. Although F0 information was temporally encoded in neural responses to concurrent-vowel pairs, the saliency of encoding decreased with increasing F0 differences. Similar SAC analysis of neural responses to reverberant concurrent-vowel pairs, recorded in the ventral CN of anaesthetised guinea pigs, have shown periodic coding to be further impaired (Sayles, Stasiak, & Winter, 2014). It seems that the exactitude of TFS-based segregation would suffer as F0 differences were introduced between vowels. Correspondingly, CVI models utilising ACFs struggle to identify both F0s of a vowel pair (Assmann & Summerfield, 1990). However, models that instead aim to identify only the most salient F0 of a pair seem to be less impacted (Meddis & Hewitt, 1992). Altogether the findings of the computational modelling and the SAC analysis seem to indicate that high resolution TFS is less important than previously thought for CVI. If the case,

however, this leaves open the question of why a loss of TFS causes severe deficits identifying speech amongst background talkers when tested experimentally using vocoders, or in individuals with HL and/or cochlear implants (Hopkins, Moore, & Stone, 2008; Qin & Oxenham, 2005).

Sinex and Li (2007) found that there is poor harmonic resolution in the IC. They determined that the IC instead was low resolution, encoding ENVs of local interactions amongst adjacent harmonic components (Sinex & Li, 2005; Sinex, Li, & Velenovsky, 2005). This description facilitates Culling and Darwin's (1994) proposal of beating as a segregation cue. Although, here, beating was discovered in IC responses to concurrent-vowel pairs, it was determined that beating was an epiphenomenon rather than being responsible for improvements in CVI. Beating periodically reduced the saliency with which concurrent-vowels were represented. Hence, concurrent-vowel representations that preserved low frequency temporal information retained the beat peaks where vowel pairs were most saliently represented. This argument against a beat hypothesis agrees with studies showing that F0 benefits are observed at extremely small F0 differences (0.4%) where beating periods would negligibly alter ENVs (8× length of the stimulus), and that the effects of starting phase spectra on vowel identification is small (de Cheveigné, 1999). Running such modified concurrent-vowel stimuli would provide a further test of the proposed model.

Rather than F0-guided or beating segregation cues, in this chapter an alternative explanation for the benefit of F0 differences was provided. Improved predictions with F0 differences were found to be closely related to increased average firing rates, resulting in increased distinction in neural responses describing separate concurrent-vowel pairs. This improved distinction with F0 differences was attributed to the compressive nature of auditory filters. With F0 differences, harmonic information is less concentrated but more widely spread across probe channels, and hence at lower regions of compressive functions. This results in an increased precision of encoding. The importance of compression is evident in its

beneficial application in hearing aid algorithms (Lyon, 1982; Simpson & Cooke, 2005). However, the generality of compression assisting speech in noise identification is less clear when considering other suppressive cochlear nonlinearities (Ruggero, Robles, & Rich, 1992). Additionally, the effects of compression are likely different in human cochlea filters relative to guinea pigs (Sumner et al., 2018) and when amidst efferent activation (Yasin, Drga, & Plack, 2014).

This filter inspired explanation for listeners' improvement with F0 differences is also contingent on an assumption that listeners can perfectly predict the combined representations of sounds. Fishman, Micheyl, & Steinschneider (2016) showed that linearly combining representations (as in air) did not translate well to the neural domain. Instead, it was shown in Section 4.3.9 that individual vowels dominated any one region of their combined neural representation. Correspondingly, a *maximum* or a *logsumexp* (numerical approximation of a *maximum*) function could be used to accurately predict the combined representation of vowels. The compressive nature of the auditory system was again used to explain this. Aside from the already raised issues for compression as an explanatory mechanism, the viability of this finding can only truly be supported if a *maximum* or *logsumexp* model predicted an F0 benefit. This computational model remains to be implemented. Still, that the combined representation of vowel pairs could be trivial to estimate offers alternatives to F0-guided or beating based segregation. Additionally, an approach where top-down processing would be ambivalent to the F0s of the vowels presented may nominally explain failures to observe reliable differences in fMRI BOLD signals with F0 differences (Alain et al., 2005). A broader discussion of possible implementations and interpretations of the proposed modelling scheme is offered in the thesis' general discussion (Section 6.1.1).

This investigation was predicated on obtaining a large volume of neural data in response to a complete set of CVI experimental stimuli. Doing so resulted in a key limitation of this investigation, MUA. This resulted in a relatively poor resolution

of frequency tuning. Despite best efforts to apply state-of-the-art spike sorting methods (Kilosort; Pachitariu, Steinmetz, Kadir, Carandini, & Harris, 2016), these algorithms could not adequately compensate for electrode drift over the long experimental sessions, as well as other artefacts. Recently, spike sorting algorithms notable in their drift correction properties have emerged (Pachitariu, 2019). The use of single unit activity would be advantageous for better depiction of neural frequency response properties. However, to obtain such resolution was beyond the scope of this thesis. Additionally, MUA has been found to suitably represent neural dynamics without spike sorting (Trautmann et al., 2019). A further limitation of this work is that neural recordings were made in anaesthetised preparations. Urethane has been found to weakly affect multiple key receptors, thought to influence the precision of temporal spiking in the IC (Astl, Popelár, Kvasnák, & Syka, 1996; Wu, Ma and Kelly, 2004). Differences in auditory event-related potentials have also been observed between human listeners attending, and not attending, to concurrent-vowel stimuli (Alain, 2007). It is ambiguous how to reconcile this with the fact that data from anaesthetised animals predicted the behaviour of attentive humans. However, Schumacher, Schneider and Woolley (2011) suggest that the main difference between unanesthetized and urethane anesthetized animals is a depression of neural excitability, but otherwise auditory representations are not significantly altered.

4.5. Conclusion

I have presented a computational model integrated with neurophysiological responses to concurrent-vowel pairs that qualitatively and quantitatively predicts the behaviour of human listeners. Previously proposed segregation cues were found to be inconsequential. Although present in the neural responses, beating seemed to have little relation to the model's predictive dynamics and F0 periodicities were encoded more weakly in concurrent-vowel pairs with larger F0 differences. Instead, increasingly distinct neural encoding of concurrent-vowel pairs occurred naturally with increasing F0 differences, attributed to innate changes in the stimulus spectrum and simple cochlear mechanics. This trend was observed at small F0 differences as well as for large. The proposed model's strategy was to optimally predict the entire representation of the presented concurrent-vowel pairs. This function was well approximated numerically, and conceptually justified, by a *maximum* function. Overall, this work has provided a further proof-of-concept for the proposed model of overlapping speech identification, whilst providing new insights into processing supporting the identification of concurrently presented vowels.

Chapter 5 Neural Recognition of Overlapping Vowel-consonants following Hearing Loss

5.1. Introduction

Following a mild/moderate HL, many features of speech remain audible. Consequently, vowel identification in quiet is similar across NH and HL listeners. However, the identification of vowels amongst interfering vowels (i.e. CVI) suffers with ageing and HL (Chintanpalli, Ahlstrom, & Dubno, 2016). Following on from the conclusions of Chapters 3 and 4, this could be due to poorer frequency selectivity, reduced compression, or additional internal noise. Moreover, despite sufficient discrimination of consonants in silence, the confusion of consonants presented amongst interfering sounds is prevalent in listeners with HL (Dubno, Dirks, & Langhofer, 1982; Gordon-Salant, 1985). The pattern of consonant confusions systematically differs between NH and HL listeners (Phatak, Yoon, Gooler, & Allen, 2009). Beyond atomic speech units, a mild/moderate HL poses major challenges for the identification of syllables, words, and sentences, when in environments amongst competing talkers (Helfer & Freyman, 2008; Lee & Humes, 2012; Miller & Licklider, 1950).

Hypothesised to aid the identification of speech amongst interferers are glimpses, spectro-temporal regions of target speech that are relatively unaffected by interfering signals (Cooke, 2006). Increasing amounts of temporal glimpses have been shown to significantly improve syllable recognition in NH listeners (Vestergaard, Fyson, & Patterson, 2011). Consonant recognition was roughly 20% worse than vowel recognition when targets were completely covered by distractors. In turn, consonant recognition benefitted most from the introduction of glimpses. The positive value of glimpses has been demonstrated, to an extent, for HL listeners as well as NH listeners (Kidd & Humes, 2012; Wang & Humes,

2010). Though, overall, listeners with ageing and/or HL struggled relative to NH listeners when identifying interrupted speech.

By using a dichotic presentation paradigm, Ozmeral, Buss, and Hall (2016) explained deficits in identification amongst asynchronously modulated noise (i.e. where glimpses are present), for individuals with HL, as a result of reduced audibility and poorer frequency selectivity. In agreement, the degradation of audibility at high frequencies has been highlighted as an instigator of consonant confusions, notably for stop plosives (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Régnier & Allen, 2008). Changes to cochlear nonlinearities such as compression and suppression are also speculated to alter firing rate cues (Liberman & Dodds, 1984). Advancing along the auditory pathway, HL reduces the number of healthy ANFs (Kujawa and Liberman, 2009) as well as increasing ANF thresholds. Other neural changes attributable to a mild/moderate HL include increasingly correlated, and less robust, neural populations (Lopez-Poveda & Barrios, 2013; Willmore & Tolhurst, 2001), and altered representations of ENV and TFS information (Kale & Heinz, 2010). Surveyed are just a few, from a wide selection of, proposals as to how HL could affect auditory processing of overlapping speech. However, the relative contribution of each proposal to altered auditory perception is unclear. There are alternative factors, such as cognitive decline with age (Dryden, Allen, Henshaw, & Heinrich, 2017) which could instead be accountable for deficits identifying overlapping speech. Consequently, the extent to which neural coding of overlapping speech is degraded in an individual with HL, relative to NH, is unclear.

Here neural responses, collected in animal models of NH and HL, in response to syllables in quiet and amongst interferers, were combined with the modelling framework introduced in Chapter 2. This was done in an effort to observe how changes to neural coding, with HL, relate to speech identification. The specific paradigm allowed the benefit of temporal glimpses to be probed for both vowel and consonant identification, independent of linguistic context.

In this chapter I will present the following:

- NH and HL animals have distinct auditory profiles, measured using ABRs, CAPs, and pure tone neural responses (Sections 5.3.1 and 5.3.2).
- With a computational model, trained on neural representations of syllables in quiet, the identification of syllables in quiet does not greatly suffer following HL. However, the identification of overlapping syllables does (Section 5.3.3).
- A model configured to utilise prior knowledge of interfering syllables improves the identification of overlapping syllables for both NH and the HL dataset (Section 5.3.4).
- The vowel and consonant confusions underlying the model's overall results (Section 5.3.5 and 5.3.6).
- The superior performance of the NH dataset does not seem to be explained solely by sampling of higher spectral frequencies relative to the HL dataset (Section 5.3.7).

5.2. Methods

5.2.1. Stimuli

Two groups of four VCs were created using a Klatt-synthesiser (Klatt, 1980) with parameters derived from the IPOX phonetic dictionary (Dirksen & Coleman, 1997). Listeners have been shown to improve in the identification of VCs and CVs by utilising temporal glimpses (Vestergaard, Fyson, & Patterson, 2011). Here, VCs were selected as experimental stimuli as they had a more realistic sound quality than CVs when produced using a Klatt-synthesiser. Two pairs of vowels ({/a/,/ɒ/} and {/u/,/ɜ˞/}) were selected based on their within pair confusability in NH listeners (Chintanpalli & Heinz, 2013) and their potential to be further confused with HL (Chintanpalli, Ahlstrom, & Dubno, 2016). Vowels that were investigated in Chapter 3 and Chapter 4 (/a/,/u/,/ɜ˞/) had formant frequency and bandwidths altered to match that in Chintanpalli and Heinz (2013). All vowels had F0s set to 100 Hz. A pair of fricative consonants (/ʃ/, /s/) and a pair of plosive consonants (/p/,/k/) were also selected based on their within pair confusability in NH listeners (Cooke, 2006) and their potential to be further confused with HL (Phatak, Yoon, Gooler, & Allen, 2009). Combining vowels and consonants resulted in the creation of the two groups of four VCs ({/aʃ/, /as/, /ɒʃ/, /ɒs/} and {/uk/, /up/, /ɜ˞k/, /ɜ˞p/}). All VCs were parametrically modified to be 300 ms in length (including 10ms onset/offset raised cosine ramps). Within each group of VCs, vowel onsets and consonant onsets were time matched. Each VC was set to an average sound level of 80 dB SPL, intended to offset supra-threshold audibility limitations of HL. The relative level of consonants compared to vowels were set to sound the most natural (-10 dB SPL for /p/ and -18 dB SPL for /k/,/s/, and /ʃ/). Vowels were synthesised at a sample rate of 11025 Hz but upsampled to 97656 Hz prior to presentation, warranted by the laboratory stimulus delivery system.

VCs were monaurally presented in quiet, denoted as $/vc_i/_q$ where $vc_i, \in \{vc_1 = \text{aʃ}, vc_2 = \text{as}, vc_3 = \text{ɒʃ}, vc_4 = \text{ɒs}, vc_5 = \text{uk}, vc_6 = \text{up}, vc_7 = \text{ɜ˞k}, vc_8 = \text{ɜ˞p}\}$. VCs were

also monaurally presented in overlapping pairs. A target VC could be paired and overlapped with any within group VC, denoted the distractor VC. Target VCs were presented amongst distractor VCs at one of 14 distractor lags. Lags ranged from -262.5 ms to -37.5 ms in steps of 37.5 ms, and +37.5ms to +262.5 ms in steps of 37.5 ms (see Figure 5.1). The lag step of 37.5 ms was close to the minimal temporal glimpse observed in humans (Vestergaard et al, 2011) yet, for pragmatism, it efficiently scaled the length of the stimuli. Within a group, there were 10 pairwise combinations of VCs, $/vc_i, vc_j/$ where $vc_i, vc_j \in \{vc_1 = \text{aʃ}, vc_2 = \text{as}, vc_3 = \text{ɒʃ}, vc_4 = \text{ɒs}\}$ or $vc_i, vc_j \in \{vc_1 = \text{uk}, vc_2 = \text{up}, vc_3 = \text{ɜ˞k}, vc_4 = \text{ɜ˞p}\}$. Each combination could be presented with a distractor lag of a ms, where $a \in A$ and $A = \{-262.5, -225, -187.5, -150, -112.5, -75, -37.5, 37.5, 75, 112.5, 150, 187.5, 225, 262.5\}$. In total, there were 280 overlapping VC stimuli.

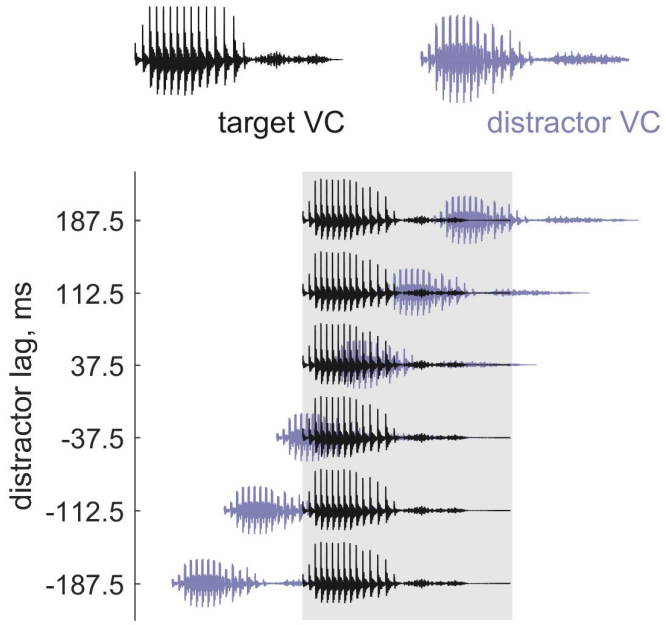


Figure 5.1 A target VC (black) is shown amongst a distractor VC (purple), at 6 different distractor lags (-187.5, -112.5, -37.5, 37.5, 112.5, and 187.5 ms). The grey rectangle depicts the temporal window within which classification was performed.

5.2.2. Neurophysiology Recordings

5.2.2.1. Subjects

Experiments were performed on 13 tricolour guinea pigs (*Cavia porcellus*), all bred in house. 7 guinea pigs were assigned to the NH control group, where 5 were male and 2 were female, weighing between 348g and 723g. 6 guinea pigs underwent a NE protocol, and hence were assigned to the HL group, where 4 were male and 2 were female, weighing between 429g and 687g. Experiments were carried out in accordance with UK Home Office regulations defined by the Animals (Scientific Procedures) Act 1986 and the European Communities Council Directive 1986 (86/609/EEC) and with the approval of the Animal Welfare and Ethical Review Body at the University of Nottingham, UK.

5.2.2.2. Noise Exposure

Subjects assigned to the HL group underwent NE. Anaesthesia was induced with bilateral intraperitoneal injections of ketamine (40 mg kg^{-1}) and xylazine (8 mg kg^{-1}). The forepaw withdrawal reflex was periodically tested and kept supressed with intramuscular ketamine/xylazine mix (0.2 ml; 3:2) injections to ensure adequate analgesia. Animals were laid on a heating blanket and sustained at a core temperature of 38°C monitored by a rectal probe and a feedback system to alter the blanket heat. Animals were placed in a sound-proofed and electrically earthed booth and monitored throughout with an infrared camera.

Once anaesthetised, hypodermic needles were placed through skin over the left and right mastoids, and a reference vertex point close to bregma. Pre-NE ABRs were recorded, as described in Section 4.2.2.4, using the presentation of auditory stimuli through sealed polyethylene tubes with diameters of 20 mm connected to the speaker system. Following pre-NE ABRs, the left ear was exposed to bandpass noise (8-10 kHz at 115 dB SPL) for 1 hour. The right pinna was folded and positioned beneath the right polyethylene tube filled with cotton wool.

Immediately following NE, post-NE ABRs were recorded in exactly the same way as the pre-NE ABRs. In order to allow some recovery of thresholds, and to avoid the possible onset of tinnitus, which normally takes about 8 weeks to develop (Coomber et al., 2014), NEs were performed 4-7 weeks prior to the non-recovery surgery and neural recordings.

5.2.2.3. Surgical Preparation, Sound Delivery and Calibration, Auditory Profiles, Spike Collection, Frequency Response Area

Aside from NE and a different stimulus set, the neurophysiology protocol remained identical to that observed in Section 4.2.2 (specifically 4.2.2.2, 4.2.2.3, 4.2.2.4, 4.2.2.5, and 4.2.2.6 remained identical). To summarise, neural responses were recorded from the IC of anaesthetised guinea pigs using 64 (8×8) channel electrodes (NeuroNexus). A threshold criterion was then used to collect within channel MUA spike times. FRAs were mapped for all channels and BFs were determined.

5.2.2.4. Presentation of VC Stimuli

Experimental stimuli were 8 VCs in quiet and 280 overlapping VC pairs (as described in 5.2.1; VCs were grouped into, and only confusable within, 4s). All stimuli were presented monaurally (left [contralateral] ear). Each stimulus was presented 30 times, in a randomised order, with a 200 ms delay between the offset of the previous stimulus and the onset of the next. To present all FRA and VC stimuli in any one electrode position took between 2 and 3 hours.

5.2.3. Computational Model

A computational model (as introduced in Chapter 2) was used to identify VCs based on neural representations. The model was composed of neural representations describing VC stimuli (front-end) and a classification back-end.

5.2.3.1. Front-end

Using similar pre-processing as described in Section 4.2.3, neural responses to each VC in quiet and for each overlapping VC pair were converted into likelihood estimators. Briefly, for each stimulus, spike counts in 0.125 ms time bins (Nyquist limit of 4000 Hz) were cosine ramped (onset/offset of 10 ms) passed through a cascade of four 2nd order Butterworth low-pass filters with a cut-off frequency of 100 Hz. The low-pass spike counts were averaged across multiple presentations, down-sampled to 4 ms bins, and collated across the 300 electrode channels least likely to be recording responses to auditory stimuli by chance, to create a PSTH (variance within an electrode channel was also retained). The PSTH was calculated within a 300 ms time window aligned with the target VC. Each PSTH bin parametrically represented the mean in an independent dimension of a multi-variate Gaussian distribution. The associated variance in spike counts operated as the variance of this Gaussian distribution. Independent, Gaussian, zero-mean noise, with identical variance, was then added to each dimension. The variance of this noise was a fitting parameter within the model, used to scale overall performance. The composite Gaussian distribution, reflecting variability in the IC and putative more central noise, acted as a likelihood estimator for a stimulus (e.g. $P(\boldsymbol{x}|/vc_i/_q)$ denotes a likelihood estimator describing the probability of observing a vector of evidence $\boldsymbol{x}$ given the presentation of a VC, $/vc_i/$, in quiet, $_q$). Eight randomised repeats of 2-fold cross validation were used to avoid overfitting to training stimuli.

5.2.3.2. Classification Back-end

Likelihood estimators built in the auditory front-end were used in the classification back-end to predict the identification of VCs in quiet and when overlapping one another. The back-end was configured suboptimally or optimally, with regards to containing knowledge about distractor VCs.

Suboptimal Scheme

The back-end classifier was configured to only store knowledge of VCs in quiet. This approach is referred to as being suboptimal because the model is ignorant to the presence of distracting VCs. The likelihood estimator for each VC, when presented in quiet, acted as the class-conditional likelihood estimator for that VC:

$$P(\boldsymbol{x}|/vc_i/) = P(\boldsymbol{x}|/vc_i/_q) \quad . \tag{5.1}$$

Assuming equal priors and a zero-one loss function, regions (R) could be determined for which each VC had the MAP estimate (Eq. 2.4).

If an overlapping pair of VCs, $/vc_i, vc_j/$ ($/vc_i/$ is the target), was presented with a distractor lag of a, was presented, the probability with which $/vc_i/$ was identified as the VC $/vc_I/$ was:

$$P(/vc_I/|/vc_i, vc_j/, a) = \int_{\boldsymbol{x}\in R_{/vc_I/}} P(\boldsymbol{x}|/vc_i, vc_j/, a)\, d\boldsymbol{x} \quad . \tag{5.2}$$

These high dimensional integrals were solved numerically using Monte Carlo methods. The model's fitting parameter, independent identically distributed Gaussian noise (in auditory front-end), was set such that the suboptimal model, trained on NH data, predicted VCs in quiet with a 95% accuracy. This parameter was kept constant for the HL data and for the optimal scheme configuration of the back-end.

Optimal Scheme

Alternatively, a model was configured that optimally utilised knowledge of overlapping VCs. For each VC, the model aggregated likelihood estimators of it amongst all other within groups VCs, at all distractor lags. The resultant class-conditional likelihood estimators were mixtures of multi-variate Gaussian distributions:

$$P(\boldsymbol{x}|/vc_i/) = \frac{1}{56}\sum_{a\in A}\sum_{j=1}^{4} P(\boldsymbol{x}|/vc_i, vc_j/, a) \quad . \tag{5.3}$$

By again assuming equal priors, and a zero-one loss function, regions (R) were determined for which each VC had the MAP estimate (Eq. 2.4).

If an overlapping pair of VCs, $/vc_i, vc_j/$ ($/vc_i/$ is the target), was presented with a distractor lag of a, the probability with which $/vc_i/$ was identified as the VC $/vc_I/$ was as stated in Eq. 5.2, but with the implementation of the newly defined class-conditional likelihood estimators in Eq. 5.3.

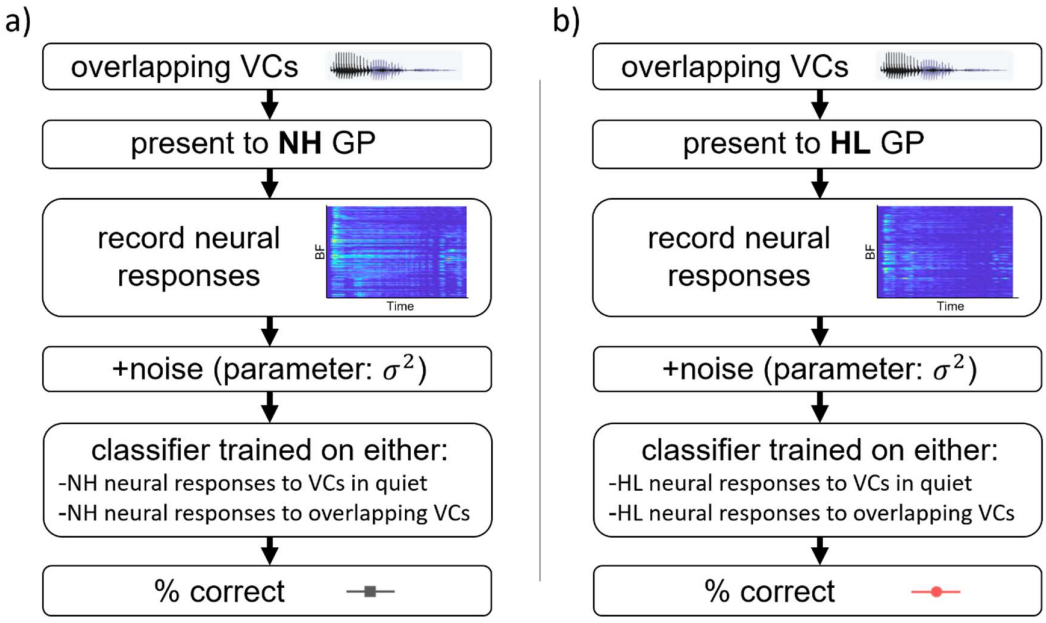


Figure 5.2 High level schematics of the proposed model for the NH (a) and HL (b) data.

5.3. Results

5.3.1. Auditory Profiles

Animals were split into two groups, a NH control group and a HL group that was exposed to high-intensity noise as detailed in Section 5.2.2.2. ABRs were recorded for all subjects in the HL group, immediately prior to NE and immediately following NE. Figure 5.3 shows the average peak-to-peak amplitudes measured in response to pure tones presented at 3 frequencies (2, 4, and 9 kHz) and at 5 levels (20, 30, 40, 60, and 80 dB SPL). Prior to NE, mean peak-to-peak ABRs increased as a function of level, and were maximal for pure tones presented at 9 kHz. This is representative of normal auditory processing. Immediately following the acoustic trauma of NE, peak-to-peak amplitudes showed an almost complete loss of ABRs, with above negligible responses only observed for tones presented at 80 dB SPL. ABRs recorded on the day of non-recovery experimentation are not presented here due to being recorded with a different system/setup, making comparison with pre- and post-NE ABRs difficult.

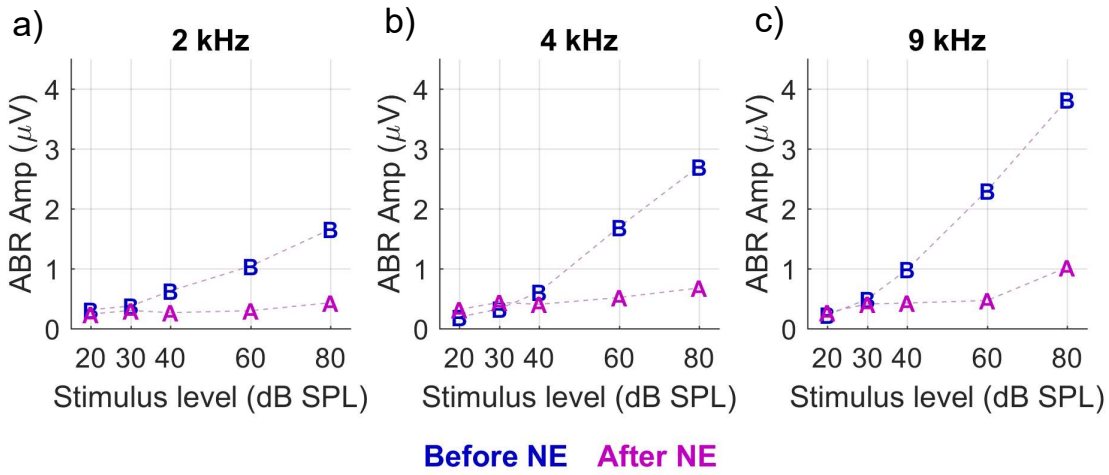


Figure 5.3 ABR peak-to-peak amplitudes averaged across HL guinea pigs at 3 frequencies (a: 2 kHz, b: 4 kHz, c: 9 kHz), and 5 sound levels. ABRs were measured before NE (blue B), and immediately after NE (magenta A).

CAP measurements were recorded in NH and HL animals immediately prior to non-recovery experimentation, which occurred 4-7 weeks after NE for the HL animals (displayed in Figure 5.4). In NH animals, CAPs revealed healthy auditory profiles, with the lowest thresholds occurring around 9 kHz. HL animals that had been acoustically over exposed (4-7 weeks prior) demonstrated a recovery in threshold responses at low/mid frequencies. However, HL animals had a distinct loss of audibility at high frequencies beyond 4 kHz, characteristic of a high frequency HL. In the HL CAP recordings, the threshold of response to frequencies below 4 kHz was slightly lower in the HL animals than the NH animals. This could be related to Liberman and Kiang's (1978) findings that cochlea tuning curves with degraded frequency tips are often accompanied by increased sensitivity in their tails. Overall, the ABRs and CAP measurements serve as evidence that animals that underwent NE did show signs of a moderate high frequency HL, whilst NH animals demonstrated healthy auditory profiles.

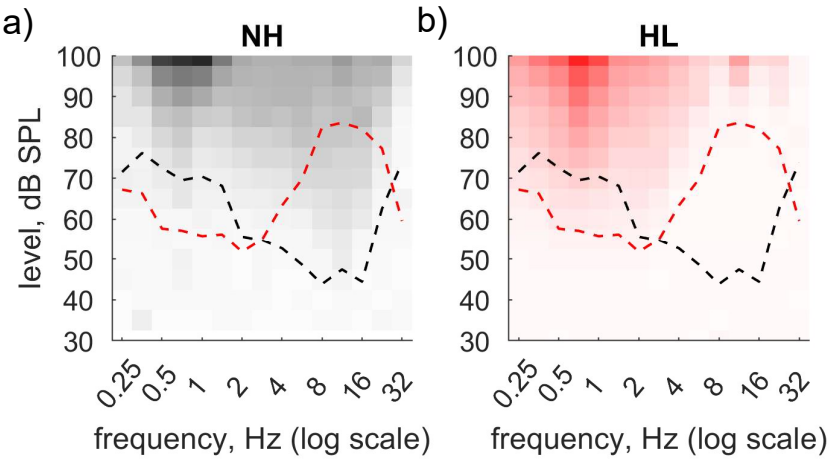


Figure 5.4 CAPs were recorded on the round window, and averaged across, NH (a, black) and HL (b, red) guinea pigs. Darker pixels represent increased magnitude of pure tone CAP response. Average NH thresholds (black dash line) and HL thresholds (red dash line), inferred via broken stick regression, are overlaid on both plots.

5.3.2. Pure Tone Responses

Neural responses to pure tones were measured and represented in FRAs for all electrode channels. From this a determination of spontaneous firing, driven firing, and how significantly different the two were (calculated by a two-sample t-test) could be made. Recordings made in both NH and HL groups reveal a large number of channels recording neural activity likely driven by auditory stimuli (Figure 5.5).

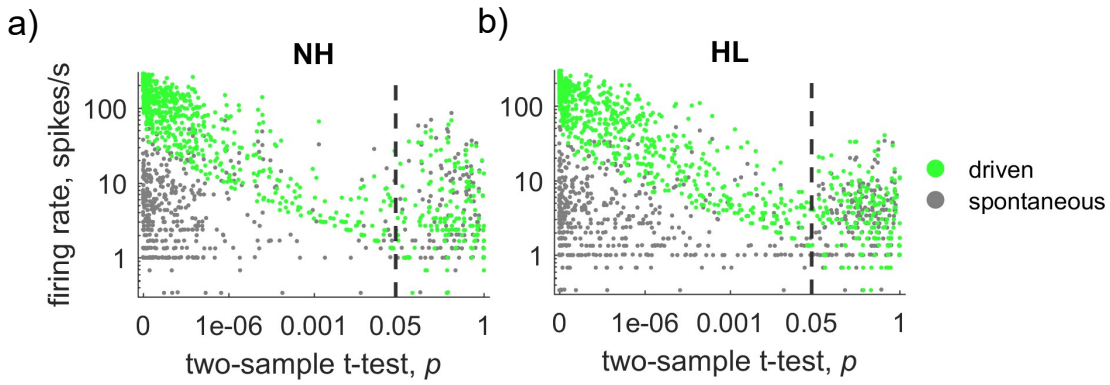


Figure 5.5 Spontaneous and driven rates in response to pure tones, for NH (a) and HL (b) animals. The abscissa captures how significantly different driven and spontaneous responses were in a given electrode channel.

In addition, a representation of a channel's frequency response characteristics was captured using the $BF_{50\text{-}70}$ metric (frequency of pure tone which elicited the maximum average firing rate, when presented between 50 and 70 dB SPL). Figure 5.6 displays the distribution of $BF_{50\text{-}70}$s for channels measured in NH and HL animals, plotted cumulatively based upon how likely the channels were driven by auditory stimuli. For NH channels, the distribution of $BF_{50\text{-}70}$ is fairly uniform across logarithmic frequency bins. This distribution is consistent across the cumulative increase of channels p-values. In comparison, HL channels, although similar in their p-values when comparing spontaneous firing rates with driven, demonstrate a clearly non-uniform distribution of $BF_{50\text{-}70}$s. This is consistent with the loss of audibility at higher frequencies measured in the ABRs and CAPs. There is a reduction in channels that are maximally responsive to frequencies above 3.3

kHz. Instead, there is a dramatic relative increase in the number of channels maximally responsive to frequencies between 1.1 and 3.3 kHz. This $BF_{50\text{-}70}$ distribution was again consistent across the number of channels analysed.

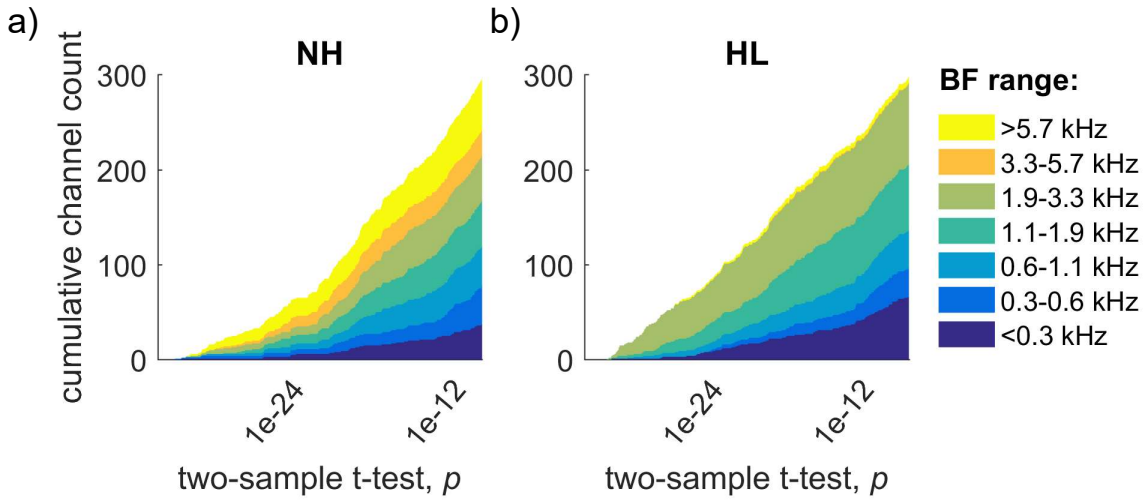


Figure 5.6 Neural channels that are likely driven by auditory stimuli are cumulatively displayed for channels in NH (a) and HL (b) animals. $BF_{50\text{-}70}$ are colour coded and displayed proportionally.

To capture the full range of frequency responses in NH and HL animals, FRAs were averaged across the 300 channels least likely to have measured responses to auditory stimuli by chance in each dataset (Figure 5.7). In tandem with the ABRs, CAPs, and $BF_{50\text{-}70}$s, the HL mean FRA had increased thresholds at higher frequency pure tones, indicative of a moderate high frequency HL. At lower frequencies, the rate profiles of neural channels remained largely consistent across NH and HL datasets. This analysis serves to show that the effects of NE to frequency/level responses in the midbrain were similar to those observed in the auditory periphery. Note, the strong representation of low frequency responses in both the NH and HL datasets may reflect electrode sampling bias when recording the neurophysiology data. The electrode was not inserted too far into the IC to try and minimise damage but this may have produced a bias towards low-frequency units as these are located in the dorsal part of the IC.

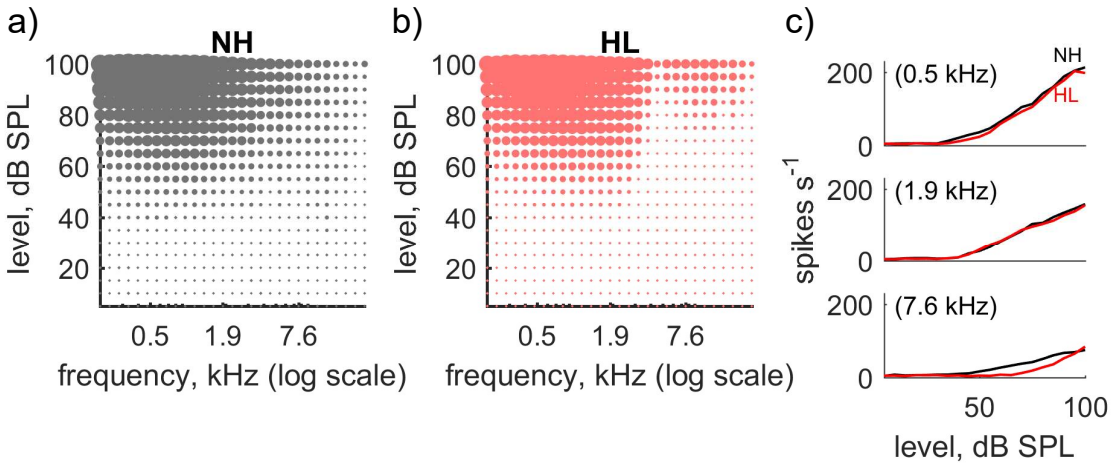


Figure 5.7 FRAs averaged separately across NH (a) and HL (b) animals (300 channels). Size of the circles are directly proportional to the number of spikes (MUA threshold crossings) emitted in response to the according frequency-level pure tone. (c) Rate level profiles for pure tones at 0.5, 1.9, and 7.6 kHz.

5.3.3. Suboptimal Identification of Overlapping Vowel-consonants

A computational model (introduced in Chapter 2) was implemented to recognise VCs based solely on neural recordings. Neural responses were either recorded from the IC of urethane anaesthetised guinea pigs that had NH or a group that underwent NE mimicking the effects of a moderate high frequency HL. Initially, the model was trained/tested on neural responses to VCs in quiet. The model's fitting parameter, Gaussian noise, was set to identify VCs in quiet with a 95% accuracy when trained/tested on NH responses. With the fitting parameter fixed, the model trained/tested on HL responses to VCs in quiet was less than 2% worse (Figure 5.8a).

The same model, trained on VCs in quiet, and with the fitting parameter remaining fixed, was presented neural responses to overlapping pairs of VCs (Figure 5.8b). When fed NH data, identification reduced to as low as 58.9% at a distractor lag of 150 ms (when the masking of consonants was maximal). Performance generally dropped as temporal delays reduced in magnitude, aside from at distractor lags of 37.5, 75, and 112.5 ms. When fed HL data, the model's identification of overlapping VCs reduced by up to a further 19.8% (distractor lag of 112.5 ms). This disparity between NH and HL data was significantly larger when overlapping VCs were identified (reject null hypothesis of 8× 2-fold cross validation unpaired t-test at 5% level, $p = 2.79 \times 10^{-12}$), relative to VCs in quiet (accept null hypothesis of 8× 2-fold cross validation unpaired t-test at 5% level, $p = 0.08$). For the NH datasets, the model's identification of overlapping VCs was qualitatively similar to that observed in NH human listeners (Vestergaard, Fyson, & Patterson, 2011). This is despite the human behavioural experiment using different, naturally recorded, VCs, overlapping at distractor lags alternative to those tested here.

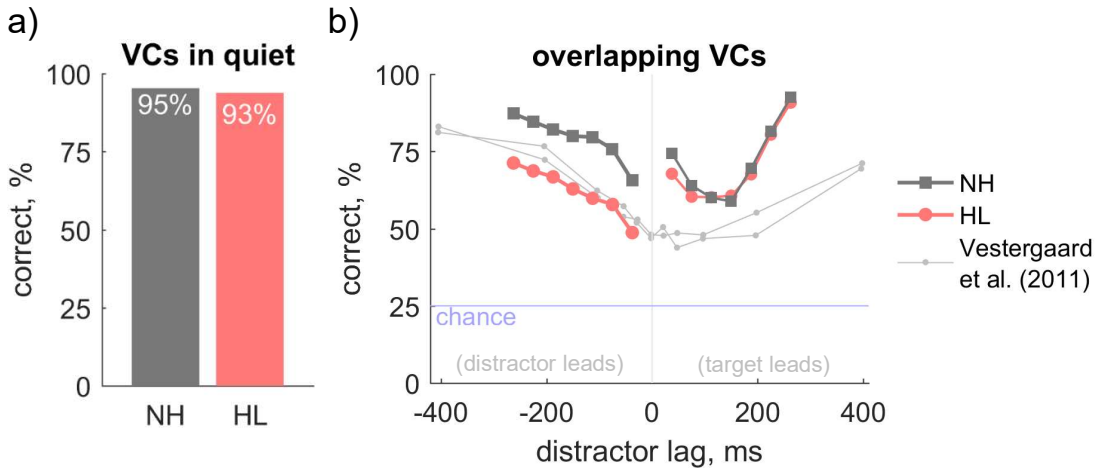


Figure 5.8 Correct identification of VCs in quiet (a), or overlapping one another as a variable of distractor lag (b). Results are averaged across VCs, and presented for NH (black squares) and HL (red circles) data. Data from Vestergaard, Fyson, and Patterson (2011) is superimposed (grey dots). Positive distractor lags represent instances when the target VC precedes the distractor VC. At intermediate positive lags (75 ms – 150 ms), the consonant of the target VC was masked by the distractor VC's vowel. Negative distractor lags represent instances when the target VC proceeded the distractor VC. At small negative distractor lags, the onset of the target VC's vowel was masked by the distractor VC's vowel.

The disparity in the identification of overlapping VCs between NH and HL was observed consistently, but only, when the distractor leads the target. This effect was found to mainly be due to the misidentification of vowel portions of VCs (Figure 5.9a). Vowel identification was up to 19.6% worse for the HL data than the NH data at a distractor lag of -112.5 ms. At positive distractor lags, where the target VC onset occurred prior to that of the distractor, vowel identification was almost 100% for the NH data. This was also the case for the HL dataset.

The model's misidentification of the consonant portion of overlapping VCs was, on average, larger than that of vowels. This was notably the case at positive distractor lags when a target's consonant is covered by the distractor VC's vowel. At increasingly large temporal delays, consonant identification slowly trended

towards performance in quiet. Notably, despite identification of consonants being poorer than vowels, there is not so much of a difference in identification between NH and HL datasets. The disparity in consonant identification between NH and HL datasets was a maximum of 8.2% (distractor lag of -37.5 ms). Notably, this difference is still larger than the difference in the correct identification of the consonant portion of VCs in quiet, between NH and HL datasets.

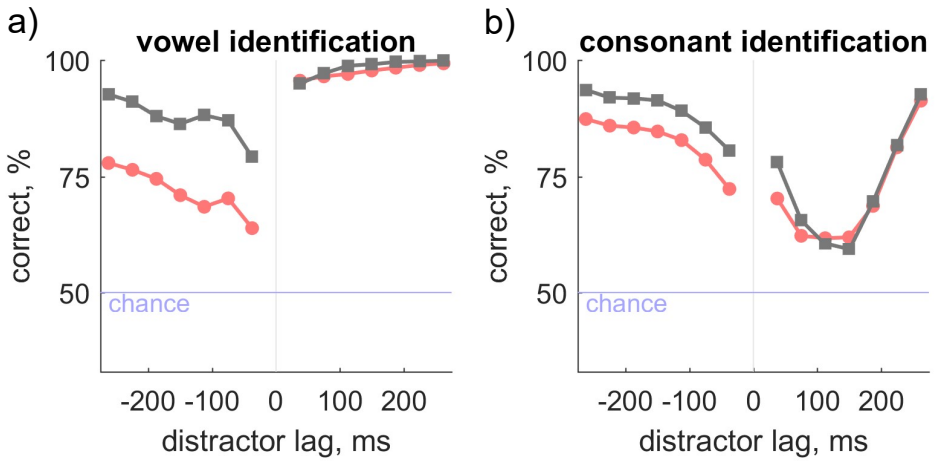


Figure 5.9 Vowel identification (a), and consonant identification (b), as a function of distractor lag. Legend same as Figure 5.8.

5.3.4. Utilising Knowledge of Interfering Vowel-consonants

The proposed computational model was alternatively trained on neural responses to overlapping VCs (fitting parameter fixed from Section 5.3.3). This imitated an optimal strategy of ideally predicting the combined representation of a target VC overlapping a distractor VC. When configured as such, the model's predictions improved for the NH dataset (12.1% on average) and the HL dataset (17.3% on average). Consequently, the difference between the performance of NH and HL data was a maximum of only 7.8% at a distractor lag of -37.5 ms. To recap, without the utilisation of this prior knowledge, the maximum difference between NH and HL datasets was 19.8%.

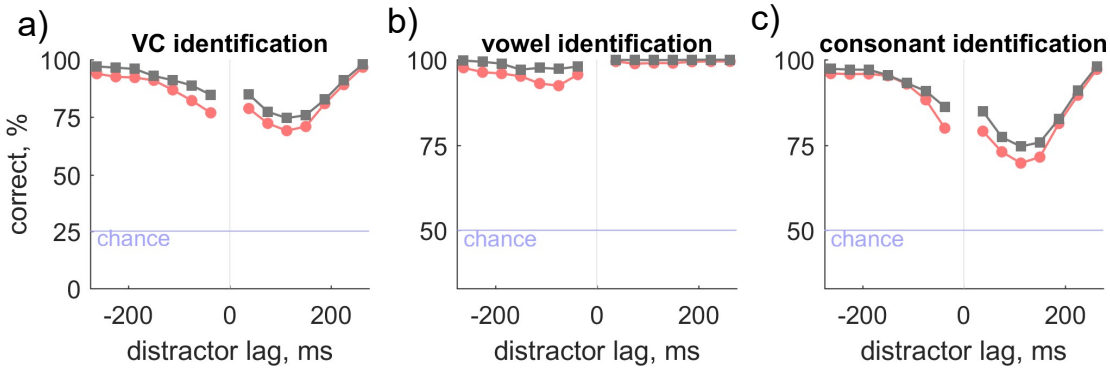


Figure 5.10 Identical to Figure 5.8(b), Figure 5.9(a) and Figure 5.9(b), but for a model with knowledge of the neural representations of overlapping VCs.

The boost in performance observed for a model with prior knowledge of VCs was attributed to an almost complete reduction of vowel confusions at negative distractor lags. The biggest difference between the NH and HL identification of vowels (4.9%) occurred at a lag of -75 ms. Yet, at that lag, vowel identification was still at 92.4% for the HL data. Instead, consonant confusions were responsible for nearly all the classifier's mistakes. Consonant identification fell to a minimum of 74.7% for the NH data at a distractor lag of 112.5 ms, and 69.9% for the HL data, a 4.8% difference. The largest disparity in consonant identification between the NH and HL datasets was 6.2%. This occurred at a distractor lag of -37.5 ms.

5.3.5. /uk/, /up/, /ɝk/, /ɝp/

Results presented in 5.3.3 and 5.3.4 describe identification performance averaged across VC identities. However, two groups of four VCs were investigated. Each group was made up of all VC combinations of two vowel identities and two consonant identities. Vowel and consonants were chosen based on their pairwise confusability. For each presentation of a VC, there was only one possible vowel confusion and one possible consonant confusion. The model was run separately for each group, and the predictions shown in 5.3.3 and 5.3.4 were averaged across the groups.

Figure 5.11 displays the propotion of true positives (hits) and false positives (false alarms) predicted for the vowels /u/and /ɝ/. For the suboptimal configuration of the model, at positive lags, the percentage of true positives for /u/ was nearly 100% and false positives were around 0% for both the NH and HL dataset. This is close to perfect identification. However, at negative lags this was not the case. Notably, at a lag of -37.5 ms the identification of /u/ dipped to 67.4% for the NH data, and below chance level (47.1%) for the HL data. At this lag, false alarms of /u/ were not predicted for either dataset. At all other negative lags, false alarms of /u/ were, on average, 15.8% more prevelant for the HL data relative to the NH. As the vowel portion of a VC could only be identified as /u/ or /ɝ/, the proportion of false alarms for /u/ is equivalent to one minus the proporion of hits for /ɝ/. Thus, the hit proportion of /ɝ/ at negative lags (excluding -37.5 ms) was similarly, on average, 15.8% higher for the NH data relative to the HL data. The HL model's preference for /u/ may be due to it comprising of formants in the 1.5-3 kHz frequency range whereas /ɝ/ does not. As the HL data over represented this frequency range compared to the HL data it could explain the model's corresponding preference for identifying /u/. When instead configured optimally, the model's hits and false alarms for /u/ and /ɝ/ were almost perfect (i.e. ~100% and 0% respecitvely).

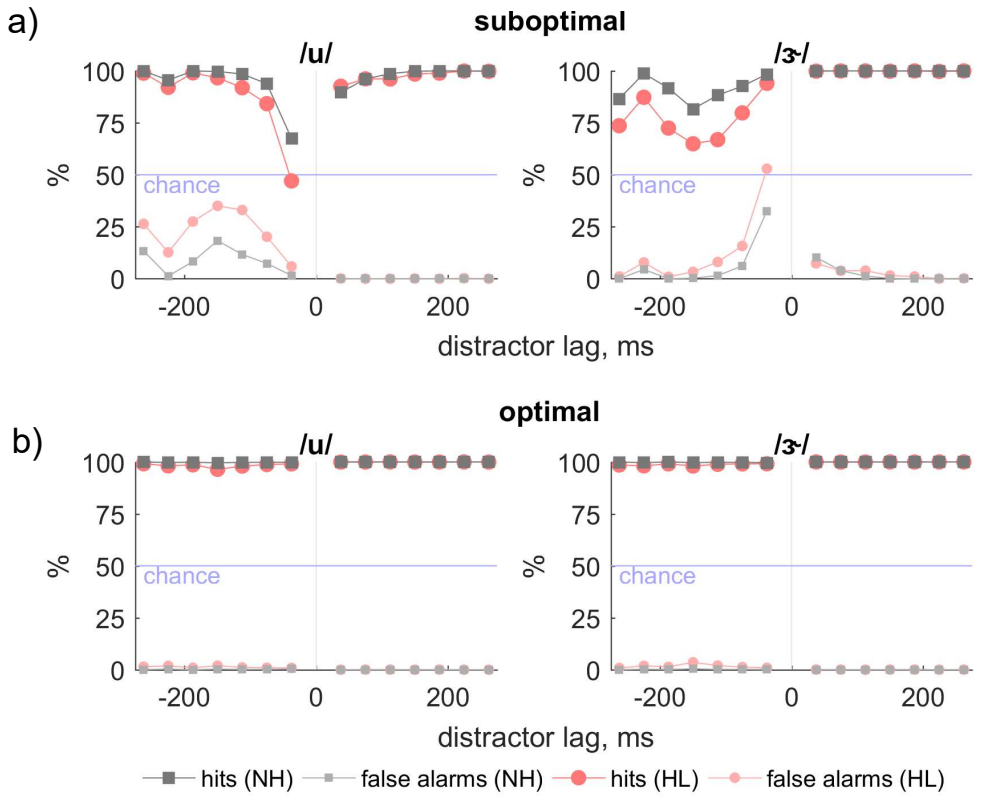


Figure 5.11 Identification of the vowels /u/ and /ɜ-/, for suboptimal (a) and optimal (b) configurations of the proposed model. Hit percentages are shown for NH (black square) and HL (red circles) data, as well as corresponding false alarm percentages (lighter shade and smaller markers).

The suboptimal configuration of the model, performed almost perfect identification of the consonants /k/ and /p/ at negative lags, for both the NH and HL data. However, as the consonant portion of VCs were increasingly covered by the vowel of the distractor VC (distractor lag between 75 and 150 ms), predictive performance dropped towards chance. At positive distractor lags, the bilabial plosive /k/ had a higher proportion of true positives for a model trained/tested on HL data than NH data. However, this was more than compensated for by a higher proportion of /k/ false alarms identified using the HL data. Essentially, the model with HL data predicted /k/ more of the time than the NH model. In turn, the number of true positives identified for the dorsal stop plosive /p/was higher for the NH data than the HL data, and this was not compensated by a comparable

increase in false alarms. Assuming the value of a hit is equivalent to a false alarm, the HL data was on average 4.6% worse when identifying the consonants /k/ and /p/ than the NH data. The model, configured to store knowledge of overlapping VCs, did not show any preference towards /k/ or /p/ for either the NH or HL datasets. However, the HL dataset again performed poorer than the NH dataset, but only by 2.6% on average.

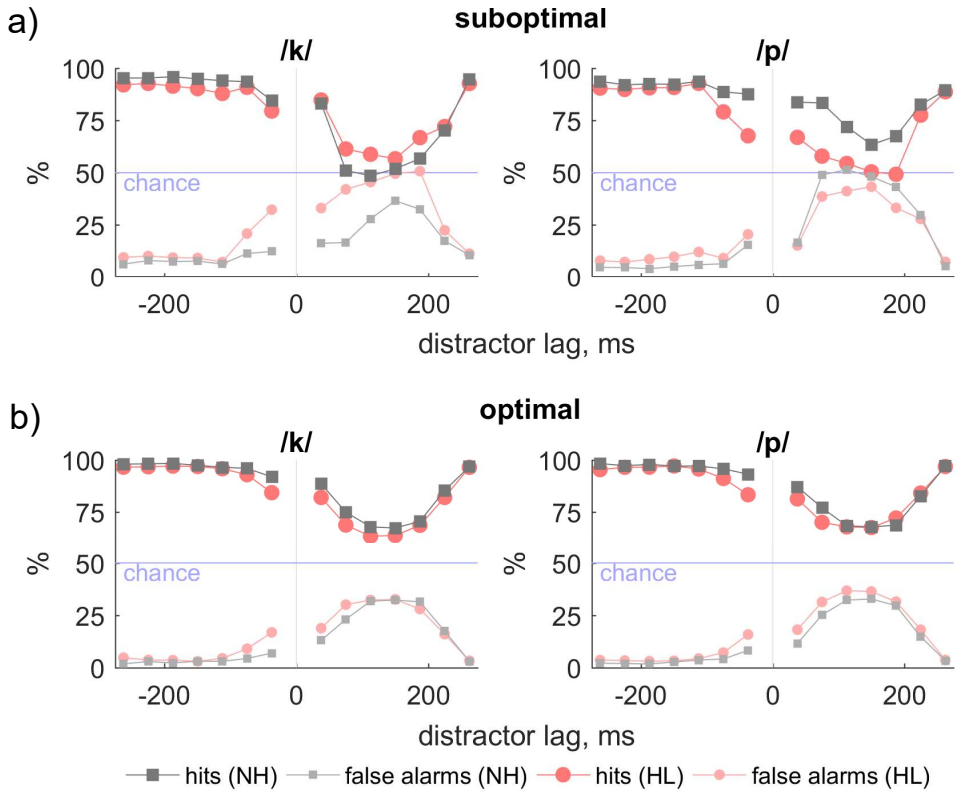


Figure 5.12 Same as for Figure 5.11, but for the consonants /k/ and /p/.

5.3.6. /aʃ/, /as/, /ɒʃ/, /ɒs/

The vowels /a/ and /ɒ/ had almost perfect true positive and false alarms statistics at positive distractor lags, for both NH and HL datasets, for both the suboptimal and optimal configuration of the model. However, at negative distractor lags, and for the suboptimal configuration of the model, /a/ had an average true positive rate of 68.6% for the NH data, which was in turn 43.2% better than for the HL data. Additionaly, for both datasets, /a/ was rarely identified when /a/ was not presented. Better identification of /a/ when presented, or equivalently reduced misidentification of /ɒ/when /a/ was presented, drove a lot of the overall differences between the NH and HL data observed in Figure 5.8b and Figure 5.9a. Alternatively, the optimal configuration of the model identified /a/ and /ɒ/ almost perfectly for both NH and HL datasets. The maximal difference in the prediction of the vowel portion of VCs, between the NH and HL datasets, was at a lag of -75 ms.

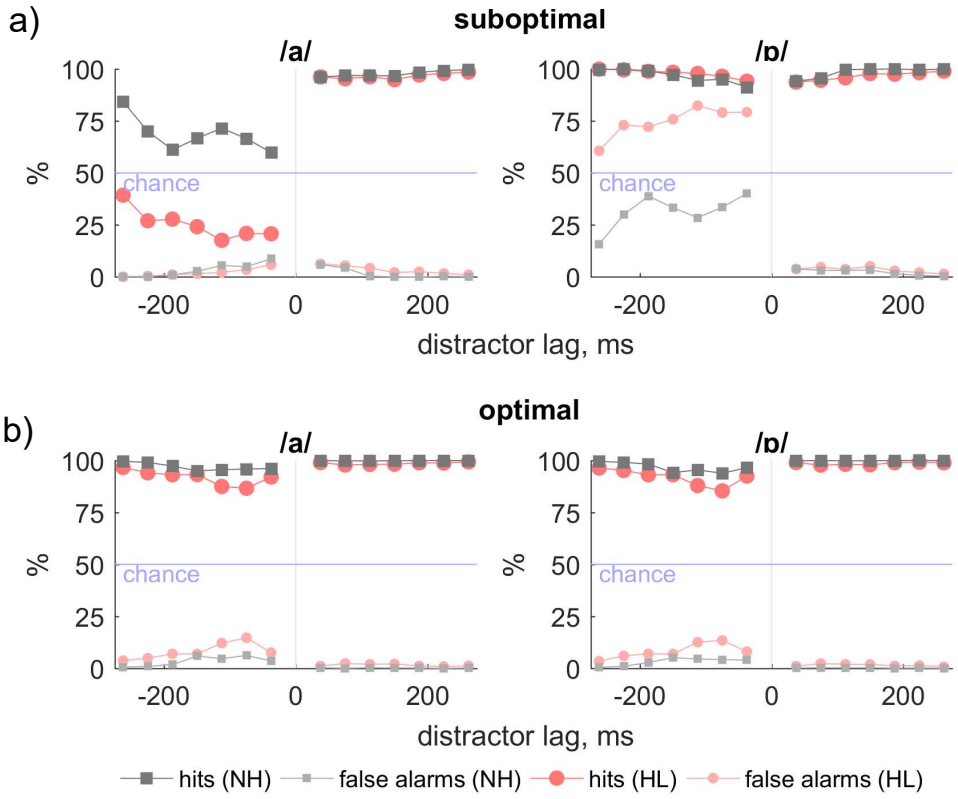


Figure 5.13 Same as for Figure 5.11, but for the vowels /a/ and /ɒ/.

For the suboptimal model configuration, the proportion of true positives predicted for /s/ was, on average, better for the HL data than the NH data. This was biggest (difference of 40.7%) at a distractor lag of 150 ms. However, the proportion of /s/ false positives was larger for the HL dataset than for the NH, at all distractor lags. Regardless, lags including and between 75 ms and 225 ms represent instances where the HL dataset better identified consonants than the NH datasets, when equally accounting for false positives and true negatives. The spectrogram of /s/ and /ʃ/, reveal that the consonant spectrums are most distinct from each other at midrange frequencies rather than high frequencies. Consequently, as the HL data overrepesents information at midrange frequencies, this may benefit recognition. As was the case for all other vowel/consonant pairs, the model's optimal scheme improved overall performance and resulted in predictions that were more similar between NH and HL data.

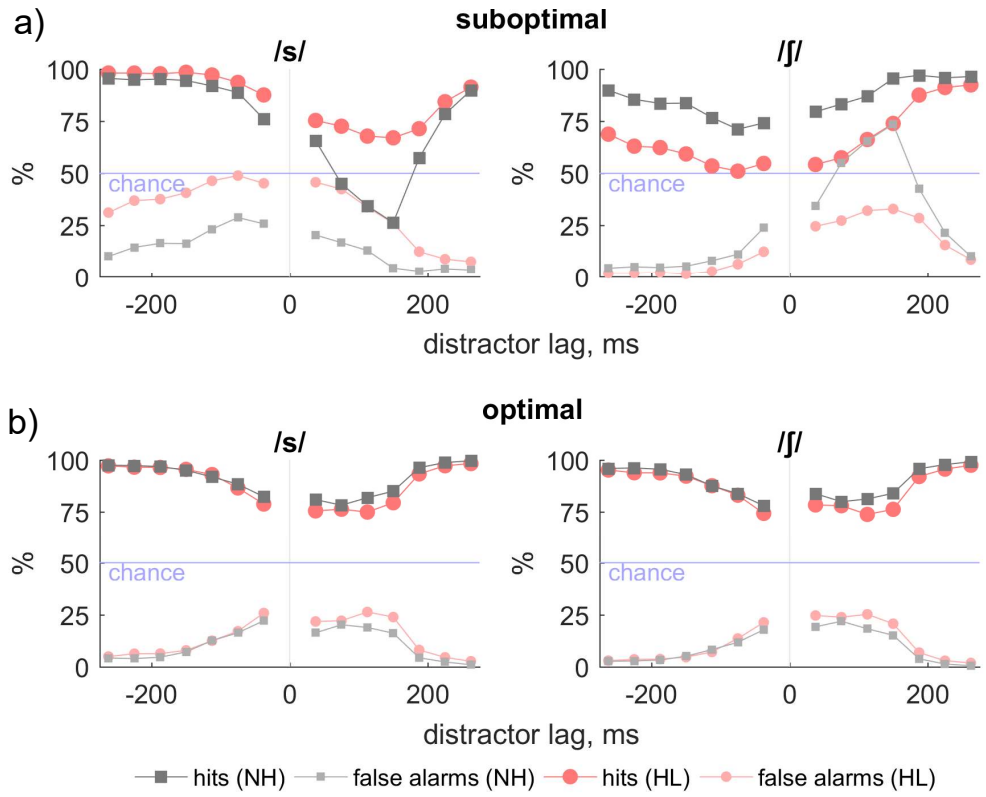


Figure 5.14 Same as for Figure 5.11, but for the consonants /s/ and /ʃ/.

5.3.7. Controlling for Frequency Sampling

Although the neural representation of vowels, overall, seem degraded following HL (notably so for the model trained on VCs in quiet), this was not so much the case for consonants (/s/). In some instances, identification of the consonant portion of VCs was worse for the NH data than for the HL data. Consequently, HL datasets may have been better representing information beneficial to the identification of particular phonemes. The NH and HL data has so far been agnostic to the frequency tuning of the neural responses from which they were constructed. Aforementioned neural representations have been simply constrained to include the 300 channels least likely to have measured responses to auditory stimuli by chance from either NH or HL animals. Channels in the NH dataset were more uniformly spread across a logarithmic axis based upon their $BF_{50\text{-}70}$s, than the HL dataset (Figure 5.15). The HL dataset, instead, oversampled channels best tuned to frequencies between 1 and 4 kHz.

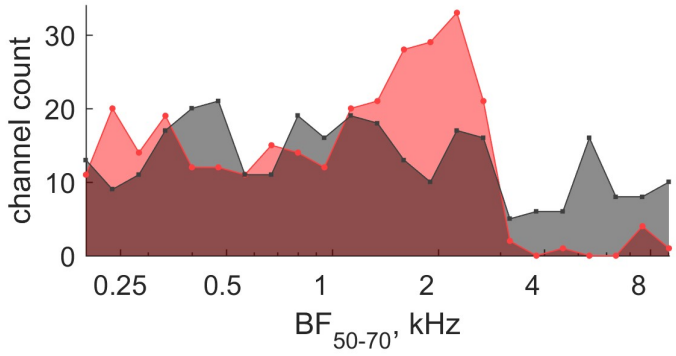


Figure 5.15 Histogram of $BF_{50\text{-}70}$ for 300 channels recorded from NH (black circles) or HL (red square) animals. Overlapping regions beneath the histograms describe the number of channels with the corresponding $BF_{50\text{-}70}$s in the frequency constrained versions of the NH and HL datasets.

The oversampling of midrange frequency channels may be beneficial to some predictions made with the HL data. However, the auditory system may not be able to favourably rewire following HL to take advantage of this information, and thus

the predictions made so far using the HL dataset may have been augmented. Conversely, the, on average, superior performance of the NH dataset may be due to its ability to represent high frequencies. As a result, the previously investigated suboptimal and optimal model configurations were re-run but for frequency constrained variations of the NH and HL datasets. Each dataset was constrained to contain no more channels responsive to a particular $BF_{50\text{-}70}$ than the other. If a dataset had n more channels of a particular $BF_{50\text{-}70}$, the n channels most likely to have measured responses to auditory stimuli by chance would be removed from the original set of 300 channels (subsequently, the constrained NH and HL datasets were compiled of 216 channels).

Predictions made with the frequency constrained HL dataset were on average 5.4% worse than the unconstrained dataset when the model was suboptimally configured, and on average 6.3% worse when the model was configured optimally (comparison of light red circles with dark red left arrow in Figure 5.16a and Figure 5.16d respectively). Although this may not appear much, in this instance suggestive of an auditory system incompetent to rewire following HL, the HL dataset is now on average 14.9% worse than the NH dataset for the suboptimal configuration, and 10.5% for the optimal configuration (compare light black squares with dark red left arrows in Figure 5.16a and Figure 5.16d respectively). Notably, for the frequency constrained HL data relative to the unconstrained NH data, the identification of consonants under the model's optimal scheme dropped to be on average 7.5% poorer (Figure 5.16f).

Conversely, the frequency constrained NH dataset was on average 5.7% worse than the unconstrained NH dataset when the model was suboptimally configured, and 4.5% worse when the model was configured optimally (comparison of light black squares with dark black right arrows in Figure 5.16a and Figure 5.16d respectively). Under the suboptimal scheme, the frequency constrained NH dataset was on average 3.6% better than predictions made with the unconstrained HL dataset. This reduced to -0.4% for the optimal scheme. Under the optimal

scheme, this finding could imply that previously observed discrepancies between NH and HL predictions can be explained by a loss of information contained in high BF channels. Alternatively, the findings may be because the HL data retains more information (216 channels) relative to the HL data (300 channels), notably for channels with BFs between 1 and 4 kHz.

When comparing the predictions made from the frequency constrained NH and HL datasets, the difference is similar to the difference between the unconstrained datasets. This similarity was also observed at the level of specific vowel and consonant confusions. These results, which control for oversampling, suggest that the superior predictions made using the NH dataset over the HL dataset (as previously described in the chapter) cannot simply be attributed to an ability to sample high frequency spectral information in neural recordings made in GPs with NH. There must be some other form of degradation to the neural representation of overlapping VCs occurring following HL.

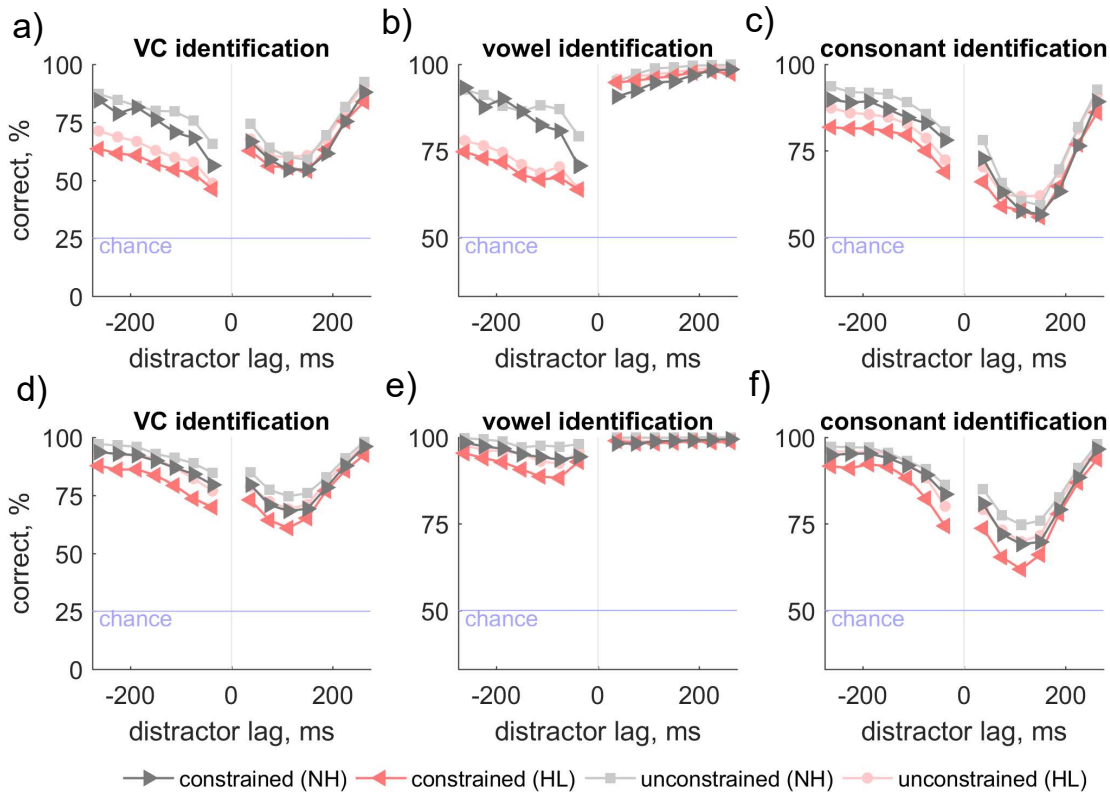


Figure 5.16 VC, vowel, and consonant identifications, predicted from frequency constrained models (left/right arrows) without (top row; plots (a), (b), and (c)) and with (bottom row; plots (d), (e), and (f)) prior knowledge of interfering VCs. Previously described frequency unconstrained models are superimposed in a lighter shade for comparative purposes.

5.4. Discussion

This chapter aimed to establish whether the neural encoding of overlapping speech sounds is degraded following HL. Given that the identification of speech in background noises is reportedly worse in individuals with a HL, and that speech signals are represented by a neural code in the brain, the proposal seems inevitable. However, this has yet to be clearly established in the auditory neuroscience literature. In an attempt to tackle this issue, I applied the computational decision model proposed in Chapter 2 to predict the identification of VCs (in quiet and overlapping one another) solely using neural data recorded from GPs. GPs either had NH or were exposed to high frequency narrowband noise. A corresponding high frequency HL was confirmed with ABRs, CAPs, and pure tone neural responses. Other measures of hearing function, such as otoacoustic emissions (Probst, Lonsbury-Martin, & Martin, 1991), would have provided a more detailed description of auditory dysfunction. If ongoing analysis implies a specific form of cochlear damage is influential, a repeat experiment with targeted characterisation of auditory function would be merited.

Generally, the model's predictions were in line with what has previously been found in humans. Neural recognition of syllables in quiet remained similar and almost optimal for both NH and HL datasets (less than 2% different). However, neural recognition of overlapping syllables degraded as they increasingly masked one another. Predicted identification was worst when the target VC was presented prior to a distracting VC, such that the target consonant was energetically masked by the distractor VC's vowel (Vestergaard, Fyson, & Patterson, 2011). Crucially, as found experimentally in humans (Dubno, Dirks and Morgan, 1984; Helfer and Freyman, 2008), the model's identification of overlapping syllables was worse for the recordings from animals with HL in comparison to ones with NH.

However, the specific differences between the NH and HL dataset were contingent on whether the model was trained to identify only syllables in quiet, or was

configured to optimally utilise knowledge of interfering speech sounds. In the former situation, errors were mostly due to the misidentification of vowel identities using the HL data. For the optimal configuration, vowel confusions largely reduced, resulting in differences between NH and HL predictions being due to discrepancies in consonant identification. Perez et al. (2013) have found that vowels, presented in quiet, are best encoded in average firing rates, whereas timing information is required for consonant encoding. Consequently, predicting speech amongst interferers may improve the neural representation of spectral information (at least in HL animals), but not temporal.

The optimal configuration improved the model's predictions for the NH data and, especially, the HL data. This scheme has proven to successfully imitate human behaviour, when applied to the neural recognition of concurrently presented vowels in Chapter 3 and Chapter 4. Additionally, predictions made under the optimal configuration maintain continuity as distractor lags go from negative to positive, as observed in human listeners (Vestergaard, Fyson, & Patterson, 2011). This transition appears discontinuous for predictions made under the model's suboptimal configuration. However, it is feasible that neither configuration is 'correct', but bound a space within which listeners can be characterised. Potentially, individuals with HL have much to gain by utilising prior information about noisy acoustic scenes. However, this may be offset by reports of increasing cognitive effort needed to facilitate such an improvement (Hicks and Tharpe, 2002; Sarampalis, Kalluri, Edwards, & Hafter, 2009).

Overall, regardless of how the model was configured, recognition using the HL neural recordings was worse than for the NH data. This in itself serves as evidence that a moderate high frequency HL does degrade the neural code of overlapping speech sounds. However, how the neural code changed remains to be disentangled. One, trivial, proposal is a simple loss of information in the frequencies most affected by HL. To challenge this, the model was re-run such that the NH and HL datasets were constrained to have equal numbers of channels

responsive to a given frequency. Although overall predictive performance dropped, it dropped similarly for both the NH and HL datasets. The previously described deficits attributed to the HL dataset still applied. Arguably, the deficits would be greater if it was assumed that HL animals could not re-wire as to optimally use the available neural information (i.e. a comparison of frequency constrained HL predictions with frequency unconstrained NH predictions). Tonotopic reorganisation (let alone more complex functional reorganisation) is still a matter of debate in, and at higher levels than, the IC (Ress and Chandrasekaran, 2013; Wang et al., 2013). Irrespective, the frequency constrained models suggest that the predicted changes to identification with HL were not solely attributable to a simple loss of high frequency information.

A drawback of the work presented is the limited set of experimental stimuli. A wider set of phonemes would allow a wider exploration of how HL can affect specific confusions and act as a confirmation that the presented findings are generalisable. Additionally, VCs were only overlapped by other VCs at a 0 dB SNR. The identification of fricative consonants has been shown to suffer more amongst white noise than in speech-shaped noise (Miller & Nicely, 1955; Wang & Bilger, 1973). Perhaps negative SNRs would have a more detrimental effect on consonant identification than observed. Additionally, less spectro-temporally generous maskers may accentuate any negative effects of HL.

5.5. Conclusion

Overall, this work offers evidence for a degraded neural representation of speech sounds masked by interfering speech sounds following HL. The extent of this degradation is dependent on whether information about interfering sounds is utilised. Potentially, prior information is more essential following a HL. The qualitative changes to speech identification that were predicted seem not to be solely attributable to a simple loss of auditory information in the frequencies most affected by HL. The principle of applying a model of decision-making to the neural coding of speech appears to be a promising method for understanding the real-world problems associated with HL. However, it would be valuable to test these ideas amongst interfering sounds with a wider variety of spectro-temporal properties.

Chapter 6 General Discussion and Future Work

For each of the two topic motivated at this thesis' opening, general interpretation of relevant findings is provided. Alongside, future avenues of research are suggested.

6.1. Motivation (i): "how humans identify speech masked by a competing speaker"

6.1.1. An Optimal Strategy

In order to determine how humans identify speech sounds amongst interfering speech sounds, I first postulated what the system was trying to achieve, i.e. its computational goals (Marr, 1982). Simple, yet unexplained, experimental instantiations of the cocktail party problem (CVI and overlapping VCs) were used to test these ideas. Models that perfectly predicted the combined representation of concurrently presented speech sounds were able to convincingly predict the behaviour of NH human listeners to identify the same sounds. One interpretation is that, on some level, the brain may be conceptualising what speech would sound like amongst an interfering talker. For example, if a waiter asked your order in a busy restaurant, they may be anticipating what your voice sounds like in combination with a conversation being had on an adjacent table.

This key result is predicated on the implementation of a scheme that optimally predicts representations of combined sounds. However, optimal performance is contingent on the consumption of neural resources (Friston, Kilner, & Harrison, 2006; Mery & Kawecki, 2005). This indicates that the identification of speech in a noisy environment is a crucial brain function, worth devoting energy towards. Complex speech vocalisations are supported by fossil records dated back to the

common ancestor of Neanderthals and modern humans (lived ~400000 years ago; de Boer, 2017). Additionally, a large proportion of neural resources are designated to facilitate sensory processing (Gaulin & McBurney, 2003). However, although the model's optimality matches the behaviour of listeners, the specific scheme modelled may be just a phenomenological demonstration of optimality. As if taken literally, the ability to predict the representations of overlapping sounds undergoes a combinatorial explosion. If we assume that humans can distinguish 330000 distinct sounds (Jacobson, 1950), to predict representations of all pairwise combinations the brain would need to perform $\sim 5.4 \times 10^{10}$ computations, which is close to half the number of neurons in the entire human brain (highest purported estimate is $\sim 12 \times 10^{10}$; Herculano-Houzel, 2009). However, this back-of-the-envelope calculation likely overestimates the combinatorial scale of the problem given the utilisation of contextual information (Benoît, Mohamadi, & Sonia, 1994; Chambers et al., 2017).

So, what computations underlie essentially optimal identification of concurrently presented sounds, but are also energy efficient? One possibility is that the top-down prediction of the combined representation of sounds is not computationally expensive. A logarithm is an analog computation that naturally arises from differential equations describing cellular biophysics (Sarpeshkar, 2014). Logarithmic encoding permits the precision of sensitivity at different intensities to remain constant, giving rise to fundamental perceptual traits in audition (decibels) as well as within other systems (Ferrell, 2009; Green & Mason, 1985). This thesis provided evidence that the compressive/logarithmic nature of the auditory system also facilitates the prediction of combined representations of sounds. The nonlinear transform means that a single sound tends to dominate any one region of the spectrum. Consequently, a *maximum* function applied to two neural representations of sounds well approximates their combination. Arguably, this function has computationally cheaper implementations (Nielsen & Sun, 2016) than predicting the superposition of sounds, as would be required in air.

Alternatively, top-down processing could be feasibly traded for more efficient bottom-up processing, without greatly reducing speech identification performance. When considering probability theory, the joint probability that two sounds have been presented is equal to the presentation probability of one sound, multiplied by the conditional probability of the other sound co-occurring. Thus, the auditory system could identify the most salient sound in a scene without any top-down influence, and in turn predict the combined representations of a limited set of ancillary sounds. This would correspond well with reports that listeners perceive a dominant sound in a scene, followed later in time by secondary sounds (de Cheveigné, 1997; McKeown & Patterson, 1995). Such an approach would also provide a means to infer and consider different numbers of sounds in a scene (Kashino & Hirahara, 1996). Determining the presence of ancillary sounds need not be a constructive process of combining representations of sounds. Instead a destructive process could be envisaged that removes an identified sound from the representation of the auditory scene. This might be based upon lateral inhibition processes observed in the IC and at higher levels of the auditory system, but currently not functionally well defined (Kral & Majernik, 1996; Shamma, 1985). In the extreme, it is arguable that exclusively bottom-up processes could result in behaviour similar to that predicted by the model. A justification for this is provided in Section 3.4. However, even a perfect segregation model would still be subject to the theoretical limits of assuming sounds are identified independently of one another.

This thesis attests to the general behaviour of the auditory system seeming optimal in a highly simplified and constrained cocktail party problem. This is an advancement on findings restricted by assumptions of segregation. Questions regarding the finer implementation of how humans identify speech in noisy environments require more consideration to clearly ascertain bottom-up from top-down processes. This could involve applying similar modelling techniques as done in this thesis, but at multiple levels of the auditory system, and in imaging modalities which will better reflect conscious processing in humans.

Improvements in computing could further facilitate improvements in the proposed model, for example by lifting assumptions of independence and Gaussianity. Applying altogether different computational models may also provide another way to tackle some of the points raised in this discussion. Models considerate of efficiency and biological evolution, such as energy-constrained neural networks (Yang, Zhu, & Liu, 2018) and genetic algorithms (Back, 1996), may help to elucidate the constraints imposed by the architecture of the auditory system. Beyond computational models, psychoacoustics relating the perception of individual sounds to their combinations would help assess to what degree the brain is able to behave optimally when identifying overlapping sounds.

6.1.2. Spectro-temporal Features

Crucial to Bregman's (1990) theory of ASA are grouping cues, such as harmonicity (Micheyl & Oxenham, 2010) and onset timing (Darwin, 1981). The proposed model was not hard coded to exploit any such auditory cues. However, the model still predicted an improvement in performance in line with an increase in the availability of such cues. Notably, there was found to be no use of classical pitch cues (Cohen, Grossberg, & Wyse, 1995; Weintraub, 1985). Spectro-temporal transformations such as an ACF, that enhances pitch related information (Meddis and Hewitt, 1991), did not improve on predictions made without its inclusion. These findings were in line with Scheffers' (1983) determination that pitch was not used to separate vowels. In spite of exhaustive efforts, no physiological basis of pitch has been conclusively found (Bendor and Wang, 2005; Walker, Bizley, King, & Schnupp, 2011). Recently, proposals of pitch as a form of probabilistic inference have been developed (Hehrmann, 2018), suggestive of pitch as an outcome of auditory processing rather than a source of it. The findings in this thesis would support this.

To determine the spectro-temporal features driving the model's decisions an information-theoretic analysis was applied (Shannon, 1948). This revealed that

with increasing F0 cues, there was simply more information distinctly representing the sounds presented. Woods and McDermott (2018) have shown how listeners are able to use abstract structures attributed to sounds, termed schemas, to aid identification. Additionally, many sounds are mediated by their statistical properties described early in the auditory system (McDermott & Simoncelli, 2011). Rather than adopting auditory cues implied by a given psychoacoustics task (e.g. in CVI experiments that vary pitch differences researchers have frequently assumed a pitch-based explanation), listeners may be learning to best exploit the general statistics of the sounds presented. Such statistical learning has found to be an implicit ability of listeners (Rohrmeier & Rebuschat, 2012; Tillmann, Bharucha, & Bigand, 2000). A statistical description of a system helps avoid issues of over characterisation of neural properties (Quiroga, Kreiman, Koch, & Fried, 2008), although can be harder to intuit.

Pertinent to a discussion of spectro-temporal features and the cocktail party problem, is how representations change along the auditory system. Emphasised in Chapter 4 was transformation of high fidelity temporal oscillations at the AN to essentially a rate code in the AC. The upper limit of phase locking (neural firing synchronised to a sound's periodic phase; Wever & Bray, 1937) in humans is of ongoing debate. A phase locking limit of 1500 Hz has been observed for binaural processing (Füllgrabe, Harland, Sęk, & Moore, 2017). For monaural processing proposals range from 1500 Hz to 10000 Hz (Verschooten, Desloovere, & Joris, 2018; Verschooten et al., 2019). However, potential central constraints and disjoints between physiological and psychoacoustic observations have led to uncertainty about the monaural function of neural phase locking in the auditory system (Lau, Mehta, & Oxenham, 2017). This view coincides with the findings of this thesis that, at least for steady state sounds, temporal encoding anywhere near the proposed phase locking cut-off frequencies (1500-10000 Hz) is not required. Corroborating this, another study that applied a model of decision-making (SDT) to rate-place coding of psychoacoustic stimuli predicted a close correspondence with listeners'

behaviour when predicting just-noticeable-differences for pure tone frequencies up to 2 kHz (Heinz, Colburn, & Carney, 2001).

6.1.3. Selective Auditory Attention

When Cherry (1953) introduced the cocktail party problem he examined whether listeners could selectively attend to overlapping speech signals. Largely neglected in this examination of the problem, auditory attention undoubtedly plays a significant role in speech perception amongst noise and is closely intertwined with any processing that occurs prior. Neural recordings from the surface of the human AC have revealed selective responses dependent on which speaker a listener is attending to (Mesgarani & Chang, 2012). Gestalt principles, the inspiration for much of Bregman's (1990) theory of ASA, argue that the brain perceives coherent objects separated from one another (Goldstein, 2009). Similar observations have been described in CVI when individuals have described perceiving a dominant vowel coloured by another (Assmann & Paschall, 1998; Scheffers, 1983). How the model proposed in this thesis can explain attention consumed by a single sound is unclear, but viable in some of the implementations speculated about in Section 6.1.1. The way a system solves a task is not necessarily equivalent to the higher level perception of the output of that system. Applying stimulus paradigms that encourage listeners to identify speech in a more naturalistic manner may help to understand this distinction (Ruggles & Shinn-Cunningham, 2011).

6.2. Motivation (ii): "changes to identification with hearing loss"

6.2.1. Amongst Non-stationary Sounds, a Neural Classifier Performs Worse following Hearing Loss

As pointed out at the start of the thesis, the two stated motivations are closely related to one another. Understanding how the neural coding of speech masked by overlapping speech changes with HL is contingent on processing without a HL. Although models of neural processing in the healthy auditory system are capable of reproducing an array of observations (Mao & Carney, 2015; Schaette & McAlpine, 2011; Sumner, Lopez-Poveda, O'Mard, & Meddis, 2002; Zilany & Bruce, 2006, 2007), over parameterisation can lead to tentative predictions when models are extended to account for auditory dysfunction (Heinz, 2010; Verhulst, Altoè, & Vasilkov, 2018). Many such models are built to account for invasive measurements made in animals, compromised by accounting for observations made in human psychoacoustics. Here, neural recordings were made in guinea pigs with NH or a moderate high frequency HL. Although animals display species specific cochlear mechanics and auditory encoding (Suga, Yan, & Zhang, 1998; Yost & Shofner, 2009), neural responses directly recorded from animals are more reliable imitations of processing than models fit to such data. Neural responses were integrated with a model of decision-making with a single fitting parameter (Chapter 2) that predicted speech identification performance. This approach, avoiding highly parametrised models, permitted less nuanced interpretation of whether neural coding is worse with HL.

Combining neural responses with ML algorithms is not an original concept, as it has previously been used to classify speech (Blakely, Miller, Rao, Holmes, & Ojemann, 2008; Mugler et al., 2014) and reconstruct it (Akbari, Khalighinejad, Herrero, Mehta, & Mesgarani, 2019). Efforts at classifying neural representations of speech in quiet have most commonly settled on distance based algorithms

(nearest-neighbour, linear support vector machine), trained on representations of speech in quiet (Centanni et al., 2014; Engineer et al., 2008; Hedrick, Moon, Woo, & Won, 2016; Mesgarani, David, Fritz, & Shamma, 2008; Perez et al., 2013; Steadman, 2015). Of the fewer studies that performed neural classification on speech amongst noise they have settled on similar distance based algorithms, and maskers have been continuous or sinusoidal modulated white noise (Christison-Lagay, Bennur and Cohen, 2017; Shetake et al., 2011). However, such synthetic maskers are not naturalistic background noises and have been demonstrated to illicit unalike identification in listeners (Meyer, Dentel, & Meunier, 2013). Although the previously referenced classifiers have successfully shown a close relationship between neural spiking and speech identification, this thesis presented a novel application of this methodology to investigate HL. Overall initial findings are promising, verifying that speech in quiet identification does not much worsen following HL, but speech identification amongst interfering sounds does. However, the extent of the differences in neural classification following HL was highly dependent on the stimuli presented. Additionally, comparison of predictions with listener data was qualitative and limited to a non-identical stimulus paradigm (Vestergaard, Fyson, & Patterson, 2011). It would be very interesting, of course, to collect human data in response to the stimuli investigated here. Additionally, due to the heterogeneity observed for speech identification amongst different noise maskers (Festen & Plomp, 1990), an extension of the investigated stimulus set amongst a wider array of spectro-temporal interferers is wanting.

Establishing that speech in noise encoding is degraded following HL naturally warrants a more detailed explanation. It is through such an explanation that interventions can be practically implemented. To appropriately do this was beyond the scope of this thesis, however preliminary insights were made. Crucially, poorer identification with HL remained when the high frequency response characteristics of the NH neural data was controlled for. Psychoacoustic studies have similarly demonstrated how high frequency HL can affect low

frequency speech understanding (Li, Guo, Yang, Feng, & Yin, 2017). In fact, the proposed computational model's default was to best utilise the information available in neural representations (incorporate neural responses least likely to have encoded auditory stimuli by chance). Although, central remapping mechanisms are more constrained than this (Eggermont, 2017), the classifier still performed poorer for the HL data than the NH data. Application of information theoretic analysis, such as MI, could help illuminate properties of the neural representations functionally altered with HL.

6.2.2. Employing Prior Knowledge

The proposed decision model had two configurations (optimal and suboptimal schemes) that were both applied to the identification of overlapping speech tokens using NH and HL neural data. Although both schemes were, overall, worse for the HL data, the predictions made did conditionally differ. Pilot human psychoacoustic data (not presented) indicates that the optimal scheme may best mimic the behaviour of NH listeners, as does visual comparison with existing human data (Vestergaard, Fyson, & Patterson, 2011). However, whether listeners with HL behave optimally is still largely a matter of speculation. As raised in Section 5.4, for individuals with HL, there is a notable potential benefit to speech in noise identification by utilising prior knowledge. This is not as much a comment on the quality of the neural code in individuals with HL, but on the ability to maximise the utility of this code. Such an ability may be best investigated with imaging of auditory cortical regions in attentive listeners (Eckert, Cute, Vaden, Kuchinsky, & Dubno, 2012; Yamamoto, Tabei, Katsuyama, Taira, & Kitamura, 2017), or by more work correlating cognition with speech identification in noise (Dryden, Allen, Henshaw, & Heinrich, 2017).

Alternatively, a multi-modal approach may help establish the importance of prior knowledge in individuals with HL. Early in this thesis, glimpsing speech amongst a noisy auditory scene was likened to visual occlusion. However, an auditory scene

is composed of sources that sum linearly, rather than opaque objects as in a visual scene. Instead a better analogy of the cocktail part problem is if visual objects were transparent (McDermott, 2009). In this way, identification within visual mixtures may invoke similar mechanisms as to those used by the auditory system. Interactions between hearing and visual processes have long been known (Blackburn, Kitterick, Jones, Sumner, & Stacey, 2019; Mcgurk & Macdonald, 1976). If HL degrades top-down processes, or makes their application more effortful (Peelle, 2018), a detrimental effect may also be observed to the identification of overlapping objects in an undamaged sensory domain (e.g. vision and the identification of overlapping transparent images).

6.3. Summary

The neural computations underlying successful recognition of speech amongst a competing talker, although often effortlessly experienced, are not trivial. A summary of the findings presented in this thesis follows:

- NH listeners appear to approach the performance of a system that optimally predicts the neural representation of speech sounds in combination with interfering speech sounds.
- Naturally available spectro-temporal information can explain listeners' behaviour, void of a need to adopt supplementary mechanisms exploiting cues such as pitch.
- Following NE, neural coding is degraded to such an extent that it produces deficits in speech sound recognition when masked by a competing speech sound, but not in quiet. This deficit is not merely attributable to a loss of audibility.
- This deficit is diminished with optimal application of knowledge regarding interfering speech sounds.

It is intended that progress on the presented findings will help direct the production of technological solutions and the improvement of hearing prostheses.

Appendix

A.1. Inferior Colliculus Neuroanatomy

The IC is the largest auditory nucleus in the brainstem and it has similar central auditory connections across all mammalian species studied. Both ipsilateral and contralateral fibres converge on the IC, as well as commissural connections between the left and right IC. More than 10 lower level brainstem nuclei (almost all) have ascending inputs to the IC, notably from the dorsal and ventral cochlear nuclei, the SOC, and the nuclei of the lateral lemniscus. Descending projections from the AC and medial geniculate nucleus are also thought to modulate neural activity within the IC. Generally inputs to the IC have reciprocal parallel projections in the opposite direction, with the exclusion of geniculocollicular fibers (Andersen, Roth, Aitkin, & Merzenich, 1980). Ascending projections from the IC enter the ventral division of the medial geniculate body of the thalamus, and thus to the cerebral cortex (Winer, 1992). Additionally, the IC projects to areas associated with motor output, notably the pons and superior colliculus. Descending projections have also been observed to the SOC and CN (Huffman & Henson, 1990).

The central nucleus of the IC is a densely packed region, implicated in the processing of auditory information (Winer and Schreiner, 2005). As the principal relay point of ascending fibers, it has a laminar organisation provided by its fibrodentritic composition. Ascending fibers enter the central nucleus ventrally and branch out dorsomedially. Neurons in the central nucleus are densely packed and comprise laminar disc-shaped cells whose dendrites run across the ventrolateral-dorsal medial plane (Oliver & Morest, 1984; Wallace, Shackleton, & Palmer, 2012), intermixed with stellate cells. Parallel to the laminar organisation of the central nucleus of the IC exists a tonotopic frequency response gradient, initially

determined in the inner ear. The neural frequency response properties are well described by their FRA. Ramachandran, Davis and May (1999) advocated 3 distinct FRA shapes in the IC. They determined that these shapes were due to separate anatomical connections between the IC and the medial superior olive, lateral superior olive, and dorsal CN. However, more recent analysis has revealed a wider array of FRA shapes (e.g. V-shaped, nonmonotonic Vs, narrow, closed, tilt down, tilt up, double peaked) existing within a continuum, mediated by pervasive synaptic integration and inhibitory/excitatory processes (Palmer, Shackleton, Sumner, Zobay, & Rees, 2013).

In the lateral portion of the IC exists the external nucleus, spanning the ventral side of the IC, rostral and ventromedial to the central nucleus. The external nucleus receives projections mainly from the AC but also directly from the central nucleus. The third main division of the IC is known as the dorsal nucleus, and it lies medial/dorsal to the central nucleus. Each of these IC division has unique descending/ ascending connections and cell morphologies, and are known to receive non-auditory inputs. Additionally, there are other more minor divisions such as the rostral division which has a significant somatosensory input and other adjacent collections of auditory neurones such as the sagulum and the nucleus of the brachium of the IC.

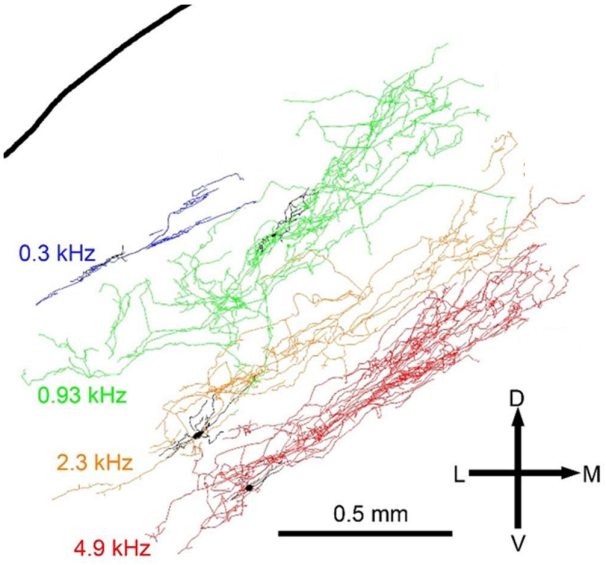


Figure A.1 Flat laminar cells (black somata, coloured axons) confined to a single lamina in tonotopic order (Wallace, Shackleton, & Palmer, 2012).

References

Action on Hearing Loss. (2015). *Hearing Matters report*. Retrieved from https://www.actiononhearingloss.org.uk/-/media/ahl/documents/research-and-policy/reports/hearing-matters-report.pdf

Action on Hearing Loss. (2018). *Working for Change Survey*. Retrieved from https://www.actiononhearingloss.org.uk/-/media/ahl/documents/campaigns/working-for-change/working-for-change-survey-results-report.pdf

Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, *9*(1), 1–12.

Alain, C. (2007). Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research*, *229*(1), 225–236.

Alain, C., Reinke, K., McDonald, K. L., Chau, W., Tam, F., Pacurar, A., & Graham, S. (2005). Left thalamo-cortical network implicated in successful speech separation and identification. *NeuroImage*, *26*(2), 592–599.

Alain, C., Snyder, J. S., He, Y., & Reinke, K. S. (2007). Changes in Auditory Cortex Parallel Rapid Perceptual Learning. *Cerebral Cortex*, *17*(5), 1074–1084.

Andersen, R. A., Roth, G. L., Aitkin, L. M., & Merzenich, M. M. (1980). The efferent projections of the central nucleus and the pericentral nucleus of the inferior colliculus in the cat. *The Journal of Comparative Neurology*, *194*(3), 649–662.

Arehart, K. H., Souza, P. E., Muralimanohar, R. K., & Miller, C. W. (2011). Effects of Age on Concurrent Vowel Perception in Acoustic and Simulated Electroacoustic Hearing. *Journal of Speech, Language, and Hearing Research*, *54*(1), 190–210.

Assmann, P. F., & Paschall, D. D. (1998). Pitches of concurrent vowels. *The Journal of the Acoustical Society of America*, *103*(2), 1150–1160.

Assmann, P. F., & Summerfield, Q. (1989). Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency. *The Journal of the Acoustical Society of America*, *85*(1), 327–338.

Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, *88*(2), 680–697.

Assmann, P. F., & Summerfield, Q. (2004). The Perception of Speech Under Adverse Conditions. In S. Greenberg, W. A. Ainsworth, A. N. Popper, & R. R. Fay (Eds.), *Speech Processing in the Auditory System* (pp. 231–308). New York: Springer.

Astl, J., Popelář, J., Kvašňák, E., & Syka, J. (1996). Comparison of response properties of neurons in the inferior colliculus of guinea pigs under different anesthetics. *Audiology*, *35*(6), 335-345.

Back, T. (1996). *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press.

Baer, T., & Moore, B. C. (1994). Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. *The Journal of the Acoustical Society of America*, *95*(4), 2277–2280.

Bakay, W. M. H., Anderson, L. A., Garcia-Lazaro, J. A., McAlpine, D., & Schaette, R. (2018). Hidden hearing loss selectively impairs neural adaptation to loud sound environments. *Nature communications*, *9*(1), 4298.

Barniv, D., & Nelken, I. (2015). Auditory streaming as an online classification process with evidence accumulation. *PloS One*, *10*(12), e0144788.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Bayes, T., & Price, M. (1763). *An Essay Towards Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London*. 370–418.

Bee, M. A., & Micheyl, C. (2008). The cocktail party problem: What is it? How can it be solved? And why should animal behaviorists study it? *Journal of Comparative Psychology*, *122*(3), 235–251.

Bendor, D., & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature*, *436*(7054), 1161-1165.

Benoît, Mohamadi, & Kandel. (1994). Effects of Phonetic Context on Audio-Visual Intelligibility of French. *Journal of Speech, Language, and Hearing Research*, *37*(5), 1195–1203.

Best, V., Mason, C. R., Swaminathan, J., Roverud, E., & Kidd, G. (2017). Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. *The Journal of the Acoustical Society of America*, *141*(1), 81–91.

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual Speech Benefit in Clear and Degraded Speech Depends on the Auditory Intelligibility of the Talker and the Number of Background Talkers. *Trends in Hearing*, *23,* 2331216519837866.

Blakely, T., Miller, K. J., Rao, R. P. N., Holmes, M. D., & Ojemann, J. G. (2008). Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids. *IEEE Engineering in Medicine and Biology Society*, *2008*, 4964–4967.

de Boer, B. (2017). Evolution of speech and evolution of language. *Psychonomic Bulletin & Review*, *24*(1), 158–162.

de Boer, E., & de Jongh, H. R. (1978). On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. *The Journal of the Acoustical Society of America*, *63*(1), 115–135.

Boes, M., de Coensel, B., Oldoni, D., & Botteldooren, D. (2011). A biologically inspired model adding binaural aspects to soundscape analysis. *40th International Congress and Exposition on Noise Control Engineering (Inter-Noise-2011)*, *5*, 3872–3878. Institute of Noise Control Engineering Japan.

Brandewie, E. J., & Oxenham, A. J. (2015). Retroactive Streaming Fails to Improve Concurrent Vowel Identification. *PLOS ONE*, *10*(10), e0140466.

Bregman, A. S. (1981). Asking the "What For" Question in Auditory Perception. In *Perceptual Organization*. Hillsdale, New Jersey: Lawrence Erlbaum.

Bregman, A. S. (1990). *Auditory Scene Analysis*. MIT Press: Cambridge, Mass.

Brokx, J. P., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*(1), 23–36.

Brown, G. J., & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language*, *8*(4), 297–336.

Brown, G. J., & Wang, D. (1997). Modelling the perceptual segregation of double vowels with a network of neural oscillators. *Neural Networks*, *10*(9), 1547–1558.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109.

Burguetti, F. A. R., & Carvallo, R. M. M. (2008). Efferent auditory system: Its effect on auditory processing. *Brazilian Journal of Otorhinolaryngology*, *74*(5), 737–745.

Carlyon, R. P., Plack, C. J., Fantini, D. A., & Cusack, R. (2003). Cross-modal and non-sensory influences on auditory streaming. *Perception*, *32*(11), 1393–1402.

Centanni, T. M., Sloan, A. M., Reed, A. C., Engineer, C. T., Rennaker, R. L., & Kilgard, M. P. (2014). Detection and identification of speech sounds using cortical activity patterns. *Neuroscience*, *258*, 292–306.

Chambers, C., Akram, S., Adam, V., Pelofi, C., Sahani, M., Shamma, S., & Pressnitzer, D. (2017). Prior context in audition informs binding and shapes simple features. *Nature Communications*, *8*, 15027.

Champoux, F., Paiement, P., Mercier, C., Lepore, F., Lassonde, M., & Gagné, J. P. (2007). Auditory processing in a patient with a unilateral lesion of the inferior colliculus. *European Journal of Neuroscience*, *25*(1), 291–297.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

de Cheveigné, A. (1997). Concurrent vowel identification. III. A neural model of harmonic interference cancellation. *The Journal of the Acoustical Society of America*, *101*(5), 2857–2865.

de Cheveigné, A. (1999). Waveform interactions and the segregation of concurrent vowels. *The Journal of the Acoustical Society of America*, *106*(5), 2959–2972.

Chintanpalli, A., Ahlstrom, J. B., & Dubno, J. R. (2016). Effects of age and hearing loss on concurrent vowel identification. *The Journal of the Acoustical Society of America*, *140*(6), 4142–4153.

Chintanpalli, A., & Heinz, M. G. (2013). The use of confusion patterns to evaluate the neural basis for concurrent vowel identification. *The Journal of the Acoustical Society of America*, *134*(4), 2988–3000.

Christison-Lagay, K. L., Bennur, S., & Cohen, Y. E. (2017). Contribution of spiking activity in the primary auditory cortex to detection in noise. *Journal of Neurophysiology*, *118*(6), 3118–3131.

Clapham, C., & Nicholson, J. (2009). *The concise Oxford dictionary of mathematics*. Oxford: Oxford University Press.

Cody, A. R., & Russell, I. J. (1985). Outer hair cells in the mammalian cochlea and noise-induced hearing loss. *Nature*, *315*(6021), 662–665.

Cohen, M. A., Grossberg, S., & Wyse, L. L. (1995). A spectral network model of pitch perception. *The Journal of the Acoustical Society of America*, *98*(2), 862–879.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562.

Coomber, B., Berger, J. I., Kowalkowski, V. L., Shackleton, T. M., Palmer, A. R., & Wallace, M. N. (2014). Neural changes accompanying tinnitus following unilateral acoustic trauma in the guinea pig. *European Journal of Neuroscience*, *40*(2), 2427–2441.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some Experiments on the Perception of Synthetic Speech Sounds. *The Journal of the Acoustical Society of America*, *24*(6), 597–606.

Cooper, N. P., & Yates, G. K. (1994). Nonlinear input-output functions derived from the responses of guinea-pig cochlear nerve fibres: Variations with characteristic frequency. *Hearing Research*, *78*(2), 221–234.

Culling, J. F., & Darwin, C. J. (1994). Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating. *The Journal of the Acoustical Society of America*, *95*(3), 1559–1569.

Dallos, P. (1992). The active cochlea. *Journal of Neuroscience*, *12*(12), 4575–4585.

Darwin, C. J. (1981). Perceptual Grouping of Speech Components Differing in Fundamental Frequency and Onset-Time. *The Quarterly Journal of Experimental Psychology Section A*, *33*(2), 185–207.

Darwin, C. J., & Culling, J. F. (1993). Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *The Journal of the Acoustical Society of America*, *93*(6), 3454–3467.

Denham, S. L., Gyimesi, K., Stefanics, G., & Winkler, I. (2013). Perceptual bistability in auditory streaming: How much do stimulus features matter? *Learning & Perception*, *5*(2), 73–100.

Deutsch, D. (2010). Hearing music in ensembles. *Physics Today*, *63*(2), 40–45.

Dirksen, A., & Coleman, J. S. (1997). All-Prosodic Speech Synthesis. In J. P. H. van Santen, J. P. Olive, R. W. Sproat, & J. Hirschberg (Eds.), *Progress in Speech Synthesis* (pp. 91–108). New York, NY: Springer.

Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017). The Association Between Cognitive Performance and Speech-in-Noise Perception for Adult Listeners: A Systematic Literature Review and Meta-Analysis. *Trends in Hearing*, *21,* 2331216517744675.

Dubno, J. R., Dirks, D. D., & Langhofer, L. R. (1982). Evaluation of hearing-impaired listeners using a Nonsense-syllable Test. II. Syllable recognition and consonant confusion patterns. *Journal of Speech and Hearing Research*, *25*(1), 141–148.

Dubno, J. R., Dirks, D. D., & Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, *76*(1), 87–96.

Duifhuis, H., Willems, L. F., & Sluyter, R. J. (1982). Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *The Journal of the Acoustical Society of America*, *71*(6), 1568–1580.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd Jr., G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *Journal of the Acoustical Society of America*, *114*(1), 368–379.

Eckert, M. A., Cute, S. L., Vaden, K. I., Kuchinsky, S. E., & Dubno, J. R. (2012). Auditory Cortex Signs of Age-Related Hearing Loss. *Journal of the Association for Research in Otolaryngology*, *13*(5), 703–713.

Egan, J. P. (1975). *Signal detection theory and ROC-analysis*. New York, NY: Academic press.

Eggermont, J. J. (2017). Acquired hearing loss and brain plasticity. *Hearing Research*, *343*, 176–190.

Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, *124*(6), 3751–3771.

Engineer, C. T., Perez, C. A., Chen, Y. H., Carraway, R. S., Reed, A. C., Shetake, J. A., … Kilgard, M. P. (2008). Cortical activity patterns predict speech discrimination ability. *Nature Neuroscience*, *11*(5), 603–608.

Evans, E. F. (1972). The frequency response and other properties of single fibres in the guinea-pig cochlear nerve. *The Journal of Physiology*, *226*(1), 263–287.

Fan, L., & Poh, K. L. (2007). A Comparative Study of PCA, ICA and Class-Conditional ICA for Naïve Bayes Classifier. In F. Sandoval, A. Prieto, J. Cabestany, & M. Graña (Eds.), *Computational and Ambient Intelligence* (pp. 16–22). Berlin: Springer.

Felty, R. A., Buchwald, A., & Pisoni, D. B. (2009). Adaptation to frozen babble in spoken word recognition. *The Journal of the Acoustical Society of America*, *125*(3), 93–97.

Ferrell, J. E. (2009). Signaling motifs and Weber's law. *Molecular Cell*, *36*(5), 724–727.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*(4), 1725–1736.

Fettiplace, R., & Hackney, C. M. (2006). The sensory and motor roles of auditory hair cells. *Nature Reviews. Neuroscience*, *7*(1), 19–29.

Fishman, Y. I., Micheyl, C., & Steinschneider, M. (2016). Neural Representation of Concurrent Vowels in Macaque Primary Auditory Cortex. *ENeuro*, *3*(3), ENEURO.0071-16.2016.

Fletcher, H., & Munson, W. A. (1933). Loudness, Its Definition, Measurement and Calculation. *The Journal of the Acoustical Society of America*, *5*(2), 82–108.

Frank, E., Hall, M., & Pfahringer, B. (2002). Locally weighted naive bayes. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence* (pp. 249–256). Morgan Kaufmann Publishers Inc.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1–3), 70–87.

Fukunaga, K. (2013). *Introduction to Statistical Pattern Recognition*. San Diego, CA: Academic Press.

Fukunaga, K., & Olsen, D. (1970). Piecewise linear discriminant functions and classification errors for multiclass problems (Corresp.). *IEEE Transactions on Information Theory*, *16*(1), 99–100.

Füllgrabe, C., Harland, A. J., Sęk, A. P., & Moore, B. C. J. (2017). Development of a method for determining binaural sensitivity to temporal fine structure. *International Journal of Audiology*, *56*(12), 926–935.

Füllgrabe, C., Moore, B. C. J., & Stone, M. A. (2014). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience*, *6*, 347.

Gaulin, S. J. C., & McBurney, D. (2003). *Evolutionary Psychology*. Pearson.

Geisler, W. S. (2003). Ideal observer analysis. *The Visual Neurosciences*, *10*(7), 12–12.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138.

Goldberg, J. M., & Brown, P. B. (1968). Functional organization of the dog superior olivary complex: An anatomical and electrophysiological study. *Journal of Neurophysiology*, *31*(4), 639–656.

Goldstein, E. B. (2009). Perceiving objects and scenes: The gestalt approach to object perception. In *Sensation and perception*. Belmont, CA: Wadsworth Cengage Learning.

Gordon-Salant, S. (1985). Phoneme feature perception in noise by normal-hearing and hearing-impaired subjects. *Journal of Speech, Language, and Hearing Research*, *28*(1), 87–95.

Green, D. B., Shackleton, T. M., Grimsley, J. M. S., Zobay, O., Palmer, A. R., & Wallace, M. N. (2018). Communication calls produced by electrical stimulation of four structures in the guinea pig brain. *PLOS ONE*, *13*(3), e0194091.

Green, D. M., & Mason, C. R. (1985). Auditory profile analysis: Frequency, phase, and Weber's law. *The Journal of the Acoustical Society of America*, *77*(3), 1155–1161.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Hedrick, M. S., Moon, I. J., Woo, J., & Won, J. H. (2016). Effects of Physiological Internal Noise on Model Predictions of Concurrent Vowel Identification for Normal-Hearing Listeners. *PLOS ONE*, *11*(2), e0149128.

Hehrmann, P. (2018). *Pitch perception as probabilistic inference* (PhD Thesis). University College London, London, UK.

Heinz, M. G. (2010). Computational Modeling of Sensorineural Hearing Loss. In R. Meddis, E. A. Lopez-Poveda, R. R. Fay, & A. N. Popper (Eds.), *Computational Models of the Auditory System* (pp. 177–202). Boston, MA: Springer US.

Heinz, M. G., Colburn, H. S., & Carney, L. H. (2001). Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. *Neural computation*, *13*(10), 2273-2316.

Helfer, K. S., & Freyman, R. L. (2008). Aging and speech-on-speech masking. *Ear and Hearing*, *29*(1), 87–98.

Henry, K. S., & Heinz, M. G. (2012). Diminished temporal coding with sensorineural hearing loss emerges in background noise. *Nature Neuroscience*, *15*(10), 1362–1364.

Henry, K. S., Kale, S., & Heinz, M. G. (2014). Noise-induced hearing loss increases the temporal precision of complex envelope coding by auditory-nerve fibers. *Frontiers in Systems Neuroscience*, *8*.

Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, *3*.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*, *45*(3), 573-84.

Hopkins, K., Moore, B. C. J., & Stone, M. A. (2008). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *The Journal of the Acoustical Society of America*, *123*(2), 1140–1153.

Hu, Y., & Loizou, P. C. (2008). Techniques for estimating the ideal binary mask. *Proc. 11th Int. Workshop Acoust. Echo Noise Control*, 154–157.

Hübener, M., & Bonhoeffer, T. (2014). Neuronal plasticity: beyond the critical period. Cell, 159(4), 727-737.

Huffman, R. F., & Henson, O. W. (1990). The descending auditory pathway and acousticomotor systems: Connections with the inferior colliculus. *Brain Research. Brain Research Reviews*, *15*(3), 295–323.

Ingham, J. G. (1970). Individual differences in signal detection. *Acta Psychologica*, *34*, 39–50.

Jacobson, H. (1950). The Informational Capacity of the Human Ear. *Science*, *112*(2901), 143–144.

Jones, P. R. (2016). A tutorial on cue combination and Signal Detection Theory: Using changes in sensitivity to evaluate how observers integrate sensory information. *Journal of Mathematical Psychology*, *73*, 117–139.

Jørgensen, S. and Dau, T., 2011. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. The Journal of the Acoustical Society of America, 130(3), pp.1475-1487.

Joris, P. X. (2003). Interaural time sensitivity dominated by cochlea-induced envelope patterns. *Journal of Neuroscience*, *23*(15), 6345-6350.

Joris, P. X., Louage, D. H., Cardoen, L., & van der Heijden, M. (2006). Correlation index: a new metric to quantify temporal coding. *Hearing research*, *216*, 19-30.

Joris, P. X., Schreiner, C. E., & Rees, A. (2004). Neural Processing of Amplitude-Modulated Sounds. *Physiological Reviews*, *84*(2), 541–577.

Kaernbach, C. (1991). Poisson signal-detection theory: Link between threshold models and the Gaussian assumption. *Perception & Psychophysics*, *50*(5), 498–506.

Kale, S., & Heinz, M. G. (2010). Envelope Coding in Auditory Nerve Fibers Following Noise-Induced Hearing Loss. *Journal of the Association for Research in Otolaryngology*, *11*(4), 657–673.

Kashino, M., & Hirahara, T. (1996). One, two, many—Judging the number of concurrent talkers. *The Journal of the Acoustical Society of America*, *99*(4), 2596–2603.

Kay, R. H. (1982). Hearing of modulation in sounds. *Physiological Reviews*, *62*(3), 894–975.

Keilson, S. E., Richards, V. M., Wyman, B. T., & Young, E. D. (1997). The representation of concurrent vowels in the cat anesthetized ventral cochlear nucleus: Evidence for a periodicity-tagged spectral representation. *The Journal of the Acoustical Society of America*, *102*(2), 1056–1071.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158.

Kidd, G. R., & Humes, L. E. (2012). Effects of age and hearing loss on the recognition of interrupted words in isolation and in sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1434–1448.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*(3), 971–995.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Kral, A., & Majernik, V. (1996). On lateral inhibition in the auditory system. *General Physiology and Biophysics*, *15*(2), 109–127.

Kujawa, S. G., & Liberman, M. C. (2009). Adding Insult to Injury: Cochlear Nerve Degeneration after 'Temporary' Noise-Induced Hearing Loss. *Journal of Neuroscience*, *29*(45), 14077–14085.

Langhans, A., & Kohlrausch, A. (1992). Differences in auditory performance between monaural and diotic conditions. I: Masked thresholds in frozen noise. *The Journal of the Acoustical Society of America*, *91*(6), 3456–3470.

Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research*, *60*(2), 115–142.

Larsen, E., Cedolin, L., & Delgutte, B. (2008). Pitch Representations in the Auditory Nerve: Two Concurrent Complex Tones. *Journal of Neurophysiology*, *100*(3), 1301–1319.

Lau, B. K., Mehta, A. H., & Oxenham, A. J. (2017). Superoptimal Perceptual Integration Suggests a Place-Based Representation of Pitch at High Frequencies. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(37), 9013–9021.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, J. H., & Humes, L. E. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *The Journal of the Acoustical Society of America*, *132*(3), 1700–1717.

Li, B., Guo, Y., Yang, G., Feng, Y., & Yin, S. (2017). Effects of Various Extents of High-Frequency Hearing Loss on Speech Recognition and Gap Detection at Low Frequencies in Patients with Sensorineural Hearing Loss. *Neural Plasticity*.

Liberman, M. C. (1978). Auditory-nerve response from cats raised in a low-noise chamber. *The Journal of the Acoustical Society of America*, *63*(2), 442–455.

Liberman, M. C., & Dodds, L. W. (1984). Single-neuron labeling and chronic cochlear pathology. III. Stereocilia damage and alterations of threshold tuning curves. *Hearing Research*, *16*(1), 55–74.

Liberman, M. C., & Kiang, N. Y. (1978). Acoustic trauma in cats: Cochlear pathology and auditory-nerve activity. *Acta Oto-Laryngologica. 358*, 1-63.

Licklider, J. C. (1964). Theory of signal detection. In J. Swets (Ed.), *Signal detection and recognition by human observers* (pp. 92-121). New York: Wiley.

Liu, L. F., Palmer, A. R., & Wallace, M. N. (2006). Phase-locked responses to pure tones in the inferior colliculus. *Journal of Neurophysiology*, *95*(3), 1926–1935.

Lopez-Poveda, E. A., & Barrios, P. (2013). Perception of stochastically undersampled sound waveforms: A model of auditory deafferentation. *Frontiers in Neuroscience*, *7*, 124.

Louage, D. H. G., van der Heijden, M., & Joris, P. X. (2004). Temporal Properties of Responses to Broadband Noise in the Auditory Nerve. *Journal of Neurophysiology*, *91*(5), 2051–2065.

Lucker, J. R., & Doman, A. (2015). Neural Mechanisms Involved in Hypersensitive Hearing: Helping Children with ASD Who Are Overly Sensitive to Sounds. *Autism Research and Treatment*, 2015.

Lyon, R. F. (1982). A computational model of filtering, detection and compression in the cochlea. In *Proceedings of the IEEE ICASSP* (pp. 1282–1285). Paris, France.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology. Human Perception and Performance*, *29*(1), 43–51.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research*, *50*(23), 2362–2374.

Mao, J., & Carney, L. H. (2015). Tone-in-Noise Detection Using Envelope Cues: Comparison of Signal-Processing-Based and Physiological Models. *Journal of the Association for Research in Otolaryngology*, *16*(1), 121–133.

Marr, D. (1982). *A Computational Investigation into the Human Representation and Processing of Visual Information*. San Franscisco, CA: W. H. Freeman and Company.

Marr, D., & Poggio, T. (1977). From Understanding Computation to Understanding Neural Circuitry. *Neurosciences Research Program Bulletin*, *14*, 470.

McDermott, J. H. (2009). The cocktail party problem. *Current Biology: CB*, *19*(22), 1024–1027.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron*, *71*(5), 926–940.

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

McKeown, J. D., & Patterson, R. D. (1995). The time course of auditory segregation: Concurrent vowels that vary in duration. *The Journal of the Acoustical Society of America*, *98*(4), 1866–1877.

Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *The Journal of the Acoustical Society of America*, *89*(6), 2866–2882.

Meddis, R., & Hewitt, M. J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, *91*(1), 233–245.

Mery, F., & Kawecki, T. J. (2005). A cost of long-term memory in Drosophila. *Science (New York, N.Y.)*, *308*(5725), 1148.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, *123*(2), 899.

Meyer, J., Dentel, L., & Meunier, F. (2013). Speech Recognition in Natural Background Noise. *PLoS ONE*, *8*(11), e79279.

Micheyl, C., & Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing Research*, *266*(1–2), 36–51.

Miller, G. A., & Licklider, J. C. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, *22*(2), 167–173.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338-352.

Moon, I. J., & Hong, S. H. (2014). What Is Temporal Fine Structure and Why Is It Important? *Korean Journal of Audiology*, *18*(1), 1–7.

Moore, B. C. (2001). Dead Regions in the Cochlea: Diagnosis, Perceptual Consequences, and Implications for the Fitting of Hearing Aids. *Trends in Amplification*, *5*(1), 1–34.

Moore, B. C., Peters, R. W., & Stone, M. A. (1999). Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips. *The Journal of the Acoustical Society of America*, *105*(1), 400–411.

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., … Slutzky, M. W. (2014). Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of Neural Engineering*, *11*(3), 035015.

Nakamoto, K. T., Shackleton, T. M., & Palmer, A. R. (2010). Responses in the Inferior Colliculus of the Guinea Pig to Concurrent Harmonic Series and the Effect of Inactivation of Descending Controls. *Journal of Neurophysiology*, *103*(4), 2050–2061.

Nielsen, F., & Sun, K. (2016). Guaranteed bounds on the Kullback–Leibler divergence of univariate mixtures. *IEEE Signal Processing Letters*, *23*(11), 1543–1546.

Nix, J., & Hohmann, V. (2007). Combined Estimation of Spectral Envelopes and Sound Source Direction of Concurrent Voices by Multidimensional Statistical Filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(3), 995–1008.

Norris, D. (1995). Signal detection theory and modularity: On being sensitive to the power of bias models of semantic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 935.

Oliver, D. L., & Morest, D. K. (1984). The central nucleus of the inferior colliculus in the cat. *The Journal of Comparative Neurology*, *222*(2), 237–264.

Oxenham, A. J., & Bacon, S. P. (2003). Cochlear Compression: Perceptual Measures and Implications for Normal and Impaired Hearing: *Ear and Hearing*, *24*(5), 352–366.

Oxenham, A. J., & Shera, C. A. (2003). Estimates of Human Cochlear Tuning at Low Levels Using Forward and Simultaneous Masking. *Journal of the Association for Research in Otolaryngology*, *4*(4), 541–554.

Ozmeral, E. J., Buss, E., & Hall, J. W. (2016). The Effects of Sensorineural Hearing Impairment on Asynchronous Glimpsing of Speech. *PLoS ONE*, *11*(5), e0154920..

Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M., & Harris, K. D. (2016). Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. *BioRxiv*, 061481.

Pachitariu, M. (2019). Kilosort2. Retrieved September 18, 2019, from https://github.com/MouseLand/Kilosort2

Palmer, A. R. (1988). The representation of concurrent vowels in the temporal discharge patterns of auditory nerve fibers. *Basic Issues in Hearing*, 244–251.

Palmer, A. R. (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*(3), 1412–1426.

Palmer, A. R. (1992). Segregation of the responses to paired vowels in the auditory nerve of the guinea pig using autocorrelation. *The Auditory Processing of Speech: From Sounds to Words*, 115-124.

Palmer, A. R., Shackleton, T. M., Sumner, C. J., Zobay, O., & Rees, A. (2013). Classification of frequency response areas in the inferior colliculus reveals continua not discrete classes. *The Journal of Physiology*, *591*(16), 4003–4025.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A'and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*(3), 556-569.

Pearson, K. (1956). *Karl Pearson's early statistical papers*. Cambridge: Cambridge University Press.

Peelle, J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear and Hearing*, *39*(2), 204–214.

Perez, C. A., Engineer, C. T., Jakkamsetti, V., Carraway, R. S., Perry, M. S., & Kilgard, M. P. (2013). Different timescales for the neural coding of consonant and vowel sounds. *Cerebral Cortex*, *23*(3), 670–683.

Peters, R. W., Moore, B. C. J., & Baer, T. (1998). Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *The Journal of the Acoustical Society of America*, *103*(1), 577–587.

Phatak, S. A., Yoon, Y., Gooler, D. M., & Allen, J. B. (2009). Consonant recognition loss in hearing impaired listeners. *The Journal of the Acoustical Society of America*, *126*(5), 2683–2694.

Pichevar, R., & Rouat, J. (2007). Monophonic sound source separation with an unsupervised network of spiking neurones. *Neurocomputing*, *71*(1), 109–120.

Pichora-Fuller, M. K. (1997). Language comprehension in older listeners. *Journal of Speech-Language Pathology and Audiology*, *21*(2), 125–142.

Plack, C. J., Oxenham, A. J., & Fay, R. R. (2006). *Pitch: Neural coding and perception* (Vol. 24). Springer Science & Business Media.

Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *The Journal of the Acoustical Society of America*, *63*(2), 533–549.

Pressnitzer, D., & Hupé, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology*, *16*(13), 1351–1357.

Probst, R., Lonsbury-Martin, B. L., & Martin, G. K. (1991). A review of otoacoustic emissions. *The Journal of the Acoustical Society of America*, *89*(5), 2027–2067.

Qin, M. K., & Oxenham, A. J. (2005). Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification. *Ear and Hearing*, *26*(5), 451–460.

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not 'Grandmother-cell' coding in the medial temporal lobe. *Trends in Cognitive Sciences*, *12*(3), 87–91.

Quiroga, R. Q., Nadasdy, Z., & Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation*, *16*(8), 1661–1687.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall.

Rabinowitz, N. C., Willmore, B. D., Schnupp, J. W., & King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron*, *70*(6), 1178–1191.

Ramachandran, R., Davis, K. A., & May, B. J. (1999). Single-unit responses in the inferior colliculus of decerebrate cats. I. Classification based on frequency response maps. *Journal of Neurophysiology*, *82*(1), 152–163.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rees, A., & Møller, A. R. (1983). Responses of neurons in the inferior colliculus of the rat to AM and FM tones. *Hearing Research*, *10*(3), 301–330.

Rees, A., & Palmer, A. R. (1988). Rate-intensity functions and their modification by broadband noise for neurons in the guinea pig inferior colliculus. *The Journal of the Acoustical Society of America*, *83*(4), 1488–1498.

Régnier, M. S., & Allen, J. B. (2008). A method to identify noise-robust perceptual features: Application for consonant /t/. *The Journal of the Acoustical Society of America*, *123*(5), 2801–2814.

Reinke, K. S., He, Y., Wang, C., & Alain, C. (2003). Perceptual learning modulates sensory evoked response during vowel segregation. *Brain Research. Cognitive Brain Research*, *17*(3), 781–791.

Ress, D., & Chandrasekaran, B. (2013). Tonotopic organization in the depth of human inferior colliculus. *Frontiers in human neuroscience*, *7*, 586.

Rhode, W. S., & Greenberg, S. (1992). Physiology of the Cochlear Nuclei. In A. N. Popper & R. R. Fay (Eds.), *The Mammalian Auditory Pathway: Neurophysiology* (pp. 94–152). New York, NY: Springer.

Robertson, D., & Irvine, D. R. (1989). Plasticity of frequency organization in auditory cortex of guinea pigs with partial unilateral deafness. *Journal of Comparative Neurology*, *282*(3), 456–471.

Rohrmeier, M., & Rebuschat, P. (2012). Implicit learning and acquisition of music. *Topics in Cognitive Science*, *4*(4), 525–553.

Ruggero, M., Rich, N. C., & Recio-Spinoso, A. (1996). The effect of intense acoustic stimulation on basilar-membrane vibrations. *Auditory Neuroscience*, *2*, 329–345.

Ruggero, M., Robles, L., & Rich, N. C. (1992). Two-tone suppression in the basilar membrane of the cochlea: Mechanical basis of auditory-nerve rate suppression. *Journal of Neurophysiology*, *68*(4), 1087–1099.

Ruggles, D., & Shinn-Cunningham, B. G. (2011). Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners. *Journal of the Association for Research in Otolaryngology*, *12*(3), 395–405.

Sachs, M. B., Voigt, H. F., & Young, E. D. (1983). Auditory nerve representation of vowels in background noise. *Journal of Neurophysiology*, *50*(1), 27–45.

Sachs, M. B., & Young, E. D. (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*(2), 470–479.

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*, *52*(5), 1230–1240.

Sarpeshkar, R. (2014). Analog synthetic biology. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *372*(2012), 20130110.

Sayles, M., Stasiak, A., & Winter, I. M. (2014). Reverberation impairs brainstem temporal representations of voiced vowel sounds: Challenging 'periodicity-tagged' segregation of competing speech in rooms. *Frontiers in Systems Neuroscience*, *8*, 248.

Schaette, R., & McAlpine, D. (2011). Tinnitus with a Normal Audiogram: Physiological Evidence for Hidden Hearing Loss and Computational Model. *Journal of Neuroscience*, *31*(38), 13452–13457.

Scheffers, M. T. M. (1983). *Sifting vowels: Auditory pitch analysis and sound segregation (PhD Thesis)*. University of Groningen, Netherlands.

Scheidt, R. E., Kale, S., & Heinz, M. G. (2010). Noise-induced hearing loss alters the temporal dynamics of auditory-nerve responses. *Hearing Research*, *269*(1–2), 23–33.

Schiffman, H. R. (1990). *Sensation and perception: An integrated approach*. John Wiley & Sons.

Schnupp, J. W. H., Garcia-Lazaro, J. A., & Lesica, N. A. (2015). Periodotopy in the gerbil inferior colliculus: Local clustering rather than a gradient map. *Frontiers in Neural Circuits*, *9*, 37.

Scholes, C., Palmer, A. R., & Sumner, C. J. (2015). Stream segregation in the anesthetized auditory cortex. *Hearing Research*, *328*, 48–58.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking Multiple Items Through Occlusion: Clues to Visual Objecthood. *Cognitive Psychology*, *38*(2), 259–290.

Schreiner, C. E., & Langner, G. (1997). Laminar fine structure of frequency organization in auditory midbrain. *Nature*, *388*(6640), 383–386.

Schumacher, J. W., Schneider, D. M., & Woolley, S. M. N. (2011). Anesthetic state modulates excitability but not spectral tuning or neural discrimination in single auditory midbrain neurons. *Journal of Neurophysiology*, *106*(2), 500–514.

Shamma, S. A. (1985). Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *The Journal of the Acoustical Society of America*, *78*(5), 1622–1632.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal,* 27(3), 379-423.

Shetake, J. A., Wolf, J. T., Cheung, R. J., Engineer, C. T., Ram, S. K., & Kilgard, M. P. (2011). Cortical activity patterns predict robust speech discrimination ability in noise: Robust speech discrimination in noise. *European Journal of Neuroscience*, *34*(11), 1823–1838.

Shinn-Cunningham, B. G. (2000). Adapting to remapped auditory localization cues: A decision-theory model. *Perception & Psychophysics*, *62*(1), 33–47.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186.

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America*, *118*(5), 2775–2778.

Sinex, D. G., & Li, H. (2005). Responses of inferior colliculus neurons to simultaneous complex tones with different fundamental frequencies. *Journal of the Association for Research in Otolaryngology*, *28*, 340.

Sinex, D G., & Li, H. (2007). Responses of inferior colliculus neurons to double harmonic tones. *Journal of Neurophysiology*, *98*(6), 3171–3184.

Sinex, D G., Li, H., & Velenovsky, D. S. (2005). Prevalence of stereotypical responses to mistuned complex tones in the inferior colliculus. *Journal of Neurophysiology*, *94*(5), 3523–3537.

Spoendlin, H. (1975). Retrograde degeneration of the cochlear nerve. *Acta Oto-Laryngologica*, *79*(3–4), 266–275.

Steadman, M. (2015). *Investigating the neural code for dynamic speech and the effect of signal degradation*. *(PhD Thesis)*. University of Nottingham, UK.

Suga, N., Yan, J., & Zhang, Y. (1998). The Processing of Species-Specific Complex Sounds by the Ascending and Descending Auditory Systems. In P. W. F. Poon & J. F. Brugge (Eds.), *Central Auditory Processing and Neural Modeling* (pp. 55–70). Boston, MA: Springer US.

Summerfield, Q., & Assmann, P. F. (1989). Auditory enhancement and the perception of concurrent vowels. *Perception & Psychophysics*, *45*(6), 529–536.

Summers, V., & Leek, M. R. (1998). F0 Processing and the Seperation of Competing Speech Signals by Listeners With Normal Hearing and With Hearing Loss. *Journal of Speech, Language, and Hearing Research*, *41*(6), 1294–1306.

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P., & Meddis, R. (2002). A revised model of the inner-hair cell and auditory-nerve complex. *The Journal of the Acoustical Society of America*, *111*(5), 2178–2188.

Sumner, C. J., Wells, T. T., Bergevin, C., Sollini, J., Kreft, H. A., Palmer, A. R., … Shera, C. A. (2018). Mammalian behavior and physiology converge to confirm sharper cochlear tuning in humans. *Proceedings of the National Academy of Sciences*, *115*(44), 11322–11326.

Swaminathan, J., & Heinz, M. G. (2012). Psychophysiological Analyses Demonstrate the Importance of Neural Envelope Coding for Speech Perception in Noise. *Journal of Neuroscience*, *32*(5), 1747–1756.

Swets, J. A., Tanner Jr, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*(5), 301.

Swinburne, R. (2012). Bayes, God, and the multiverse. *Probability in the Philosophy of Religion*, 103–122.

Szabó, B. T., Denham, S. L., & Winkler, I. (2016). Computational Models of Auditory Scene Analysis: A Review. *Frontiers in Neuroscience*, *10*, 524.

Teich, M. C., & Khanna, S. M. (1985). Pulse-number distribution for the neural spike train in the cat's auditory nerve. *The Journal of the Acoustical Society of America*, *77*(3), 1110–1128.

Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, *107*(4), 885.

Trautmann, E. M., Stavisky, S. D., Lahiri, S., Ames, K. C., Kaufman, M. T., O'Shea, D. J., … Shenoy, K. V. (2019). Accurate Estimation of Neural Population Dynamics without Spike Sorting. *Neuron*, *103*(2), 292-308.e4.

Vartanian, I. A. (1974). On mechanisms of specialized reactions of central auditory neurons to frequency-modulated sounds. *Acta Acustica united with Acustica*, *31*(6), 305-310.

Verhulst, S., Altoè, A., & Vasilkov, V. (2018). Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss. *Hearing Research*, *360*, 55–75.

Verma, N. K. (2013). *Physics For Engineers*. PHI Learning Pvt. Ltd.

Verschooten, E., Desloovere, C., & Joris, P. X. (2018). High-resolution frequency tuning but not temporal coding in the human cochlea. *PLOS Biology*, *16*(10), e2005164.

Verschooten, E., Shamma, S., Oxenham, A. J., Moore, B. C. J., Joris, P. X., Heinz, M. G., & Plack, C. J. (2019). The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints. *Hearing Research*, *377*, 109–121.

Vestergaard, M. D., Fyson, N. R. C., & Patterson, R. D. (2011). The mutual roles of temporal glimpsing and vocal characteristics in cocktail-party listening. *The Journal of the Acoustical Society of America*, *130*(1), 429–439.

Vincent, B. T. (2015). A tutorial on Bayesian models of perception. *Journal of Mathematical Psychology*, *66*, 103–114.

Von Békésy, G. (1944). Über die mechanische Frequenzanalyse in der Schnecke verschiedener Tiere. *Akust Zeits*, *9*, 3–11.

Von Békésy, G., & Wever, E. G. (1960). *Experiments in hearing* (Vol. 8). McGraw-Hill New York.

Wald, A. (1950). *Statistical decision functions*. New York: John Wiley & Sons.

Walker, K. M. M., Bizley, J. K., King, A. J., & Schnupp, J. W. H. (2011). Cortical encoding of pitch: Recent results and open questions. *Hearing Research*, *271*(1), 74–87.

Wallace, M. N., Shackleton, T. M., & Palmer, A. R. (2012). Morphological and physiological characteristics of laminar cells in the central nucleus of the inferior colliculus. *Frontiers in Neural Circuits*, *6*, 55.

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines* (pp. 181–197). Springer.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, *54*(5), 1248-1266.

Wang, D., & Chang, P. (2008). An oscillatory correlation model of auditory streaming. *Cognitive Neurodynamics*, *2*(1), 7–19.

Wang, X., & Humes, L. E. (2010). Factors influencing recognition of interrupted speech. *The Journal of the Acoustical Society of America*, *128*(4), 2100–2111.

Wang, F., Zuo, L., Hong, B., Han, D., Range, E. M., Zhao, L., ... & Liu, L. (2013). Tonotopic reorganization and spontaneous firing in inferior colliculus during both short and long recovery periods after noise overexposure. *Journal of biomedical science*, *20*(1), 91.

Weintraub, M. (1985). *A theory and computational model of auditory monaural sound separation (PhD Thesis*. Stanford University, US.

Wever, E. G., & Bray, C. W. (1937). The perception of low tones and the resonance-volley theory. *The Journal of Psychology*, *3*(1), 101–114.

Willmore, B., & Tolhurst, D. J. (2001). Characterizing the sparseness of neural codes. *Network (Bristol, England)*, *12*(3), 255–270.

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

Woods, K. J. P., & McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences*, *115*(14), 3313–3322.

Woodward, P. M. (1953). *Probability and Information Theory: With Applications to Radar*. Elsevier Science & Technology.

Wu, S. H., Ma, C. L., & Kelly, J. B. (2004). Contribution of AMPA, NMDA, and GABAA receptors to temporal pattern of postsynaptic responses in the inferior colliculus of the rat. *Journal of Neuroscience*, *24*(19), 4625-4634.

Yamamoto, K., Tabei, K., Katsuyama, N., Taira, M., & Kitamura, K. (2017). Brain activity in patients with unilateral sensorineural hearing loss during auditory perception in noisy environments. *Journal of Medical and Dental Sciences*, *64*(1), 19–26.

Yang, H., Zhu, Y., & Liu, J. (2018). Energy-Constrained Compression for Deep Neural Networks via Weighted Sparse Projection and Layer Input Masking. *ArXiv Preprint*, 1806.04321.

Yasin, I., Drga, V., & Plack, C. J. (2014). Effect of human auditory efferent feedback on cochlear gain and compression. Journal of Neuroscience, 34(46), 15319-15326.

Yost, W. A., & Shofner, W. P. (2009). Critical bands and critical ratios in animal psychoacoustics: An example using chinchilla data. *The Journal of the Acoustical Society of America*, *125*(1), 315–323.

Zhang, H. (2004). The optimality of naive bayes. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference* (pp. 562-527). AAAI Press.

Zilany, M. S. A., & Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *The Journal of the Acoustical Society of America*, *120*(3), 1446–1466.

Zilany, M. S. A., & Bruce, I. C. (2007). Representation of the vowel /ɛ/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats. *The Journal of the Acoustical Society of America*, *122*(1), 402–417.

Zwicker, U. T. (1984). Auditory recognition of diotic and dichotic vowel pairs. *Speech Communication*, *3*(4), 265–277.