Perceptual compensation for reverberation in human listeners and machines

Amy V. Beeston

Doctor of Philosophy in Computer Science

Department of Computer Science
University of Sheffield

January 2015

Abstract

This thesis explores compensation for reverberation in human listeners and machines. Late reverberation is typically understood as a distortion which degrades intelligibility. Recent research, however, shows that late reverberation is not *always* detrimental to human speech perception. At times, prolonged exposure to reverberation can provide a helpful acoustic context which improves identification of reverberant speech sounds. The physiology underpinning our robustness to reverberation has not yet been elucidated, but is speculated in this thesis to include efferent processes which have previously been shown to improve discrimination of noisy speech. These efferent pathways descend from higher auditory centres, effectively recalibrating the encoding of sound in the cochlea. Moreover, this thesis proposes that efferent-inspired computational models based on psychoacoustic principles may also improve performance for machine listening systems in reverberant environments.

A candidate model for perceptual compensation for reverberation is proposed in which efferent suppression derives from the level of reverberation detected in the simulated auditory nerve response. The model simulates human performance in a phoneme-continuum identification task under a range of reverberant conditions, where a synthetically controlled test-word and its surrounding context phrase are independently reverberated. Addressing questions which arose from the model, a series of perceptual experiments used naturally spoken speech materials to investigate aspects of the psychoacoustic mechanism underpinning compensation. These experiments demonstrate a monaural compensation mechanism that is influenced by both the preceding context (which need not be intelligible speech) and by the test-word itself, and which depends on the time-direction of reverberation. Compensation was shown to act rapidly (within a second or so), indicating a monaural mechanism that is likely to be effective in everyday listening. Finally, the implications of these findings for the future development of computational models of auditory perception are considered.

Acknowledgements

This work was supported by an Engineering and Physical Sciences Research Council project grant, EP/G009805/1.

Enormous thanks are due, first and foremost, to my supervisor Guy Brown for all his guidance and support during my time in Sheffield. In particular, his sustained encouragement has been invaluable in my struggle to manage competing demands of part-time study, employment, and family life.

Secondly, I would like to thank Tony Watkins for a great number of interesting discussions over the past six years, and for sharing the speech stimuli, room impulse responses, and human listener data without which this work could not have begun.

Further thanks are due to my collaborators in this area, including Simon Makin, Andrew Raimond, Kalle Palomäki, Ray Meddis and Robert Ferry. Thanks also to my mentors, Candice Majewski and Emiliano Cancellieri, and to Emina Kurtić and Cleo Pike for commenting on parts of this work. Additionally I would like to thank Mauro Nicolao, Sarah Al-Shareef, Robin Hofe and other present and former members of the Speech and Hearing group for sharing space, time and ideas, and to thank all the participants who took part in my listening experiments.

Finally, I thank my friends and family members for their seemingly limitless emotional and practical support through this time. In particular, I am extremely grateful to my siblings for constant conversation, and to Michael Beeston, Heather Bond, Brian Cordingley, Kate Miguda, Su and Ian Summers and Christine Ure for repeatedly helping in our home. I am completely indebted to Nicol and Mark Summers for sharing every day with me, good and bad. Special thanks to Mark for help with editing and tidying graphics, and for much love, support and delicious food.

Contents

List of figures

List of tables

Chapter 1

Introduction

Contents

1.1 Perceptual compensation for reverberation

Reverberation is a term which describes the multitude of reflections that usually accompany directly propagating sound. Affected by the surfaces in the nearby environment, each reflection arrives at the ear as a slightly delayed, attenuated and spectrally altered copy of the original sound source. It is commonly understood that 'early reflections' are beneficial to speech perception, but 'late reverberation' is an unwanted signal distortion which degrades performance in speech identification (e.g., Bradley et al., 1999). Recent research, however, has begun to show that late reverberation is not *always* detrimental to speech perception. Rather, a prolonged exposure to late reverberation can provide some context that listeners use to help them identify reverberant speech sounds. In this way, reverberation promotes a kind of perceptual constancy in audition (somewhat akin to colour or brightness constancy in vision) which ensures that the phonetic message of a speech signal is likely to remain the same whether heard in a small room or in a large auditorium.

Although a growing body of psychoacoustic data has begun to outline the reliabilities and fallibilities of the human auditory system in regard to the reverberation content of a signal, relatively little is yet known about the physiological processes which yield robustness to reverberation. Since descending auditory pathways of the efferent system are thought to contribute to a human listener's resilience in complex listening environments, this thesis proposes a candidate model for perceptual compensation for reverberation which is based on auditory efferent processing. Though speculative, the resulting computational model provides data consistent with human listener responses in a range of monaural speech identification tasks. Further, a series of perceptual experiments are undertaken to address questions arising during the modelling process, and to improve our understanding of how perceptual compensation for reverberation manifests itself with human listeners.

Reverberation still poses a serious problem for 'machine listening' systems such as automatic speech recognition (ASR). This problem becomes ever more pressing, because as the demand for flexible interfacing with technology increases, so too does our desire for distant (far-field) speech recognition. Although reverberation-robust ASR is not the main focus of the current thesis, the contribution that computer models of perceptual compensation for reverberation might make to reverberation-robust speech recognition systems is additionally discussed.

1.2 Background

Reverberation exists in most of the environments in which humans spend time. Indoors, the sound that reaches a listener (human or machine) is a combination of energy received *directly* from the sound-source and *indirectly* from a multitude of reflections from the surrounding surfaces. These reflections are collectively known as reverberation.

While reverberation allows a human listener to gain a spatial impression of the environment, and to locate the distance of a sound-source within that space, reverberation has generally been considered to be an unwanted distortion for acoustic signals because, in excess, it degrades speech intelligibility (Bolt and MacDonald, 1949; Nábělek et al., 1989). Reverberation reduces the envelope modulation depth of a signal because dips in its temporal envelope are filled with reflected sound energy (Houtgast and Steeneken, 1973). Although the identification of speech depends largely on such temporal envelope modulations, human listeners are nonetheless quite robust to the effects of reverberation. In contrast, reverberation is highly problematic for machine listening systems (e.g. distant speech recognition) and for human listeners dependent on machine-mediated hearing (e.g. using assistive listening devices such as cochlear implants).

ASR applications typically become error-prone in the presence of reverberation and, as a result, researchers have been attempting to remove reverberation from signals since the early 1970s. As this challenging problem has not yet been satisfactorily resolved, dereverberation is gaining increasing importance in research as the number of real-world uses for ASR continues to grow. One such example is the goal of making an automatic transcription of everything spoken in a meeting. This requires distant speech recognition (for talkers who may potentially move around the room while speaking), irrespective of the severity of room reverberation. For real technological advance in this field, solutions are required for both reverberation and noise (fans, electrical motors, interfering talkers). The fact that background noise can often be relatively consistent from one moment to the next has allowed a number of successful techniques to be developed to mitigate against the effect of stationary noise sources. In contrast, reverberation is highly non-stationary. Moreover, its perceptual effects are determined by the previous content of the signal itself in addition to the room acoustics in which it is experienced. Thus, despite a great deal of effort spent on engineering-based approaches to the problem of reverberation, currently there is no 'industry standard' method to adapt a machine listener to its physical environment.

A different approach is to consider why human listeners have comparatively little trouble identifying speech in reverberant environments. That is, while reverberation constitutes a large distortion acoustically, it appears to be a somewhat lesser problem when considered from a psychoacoustic perspective (cf. Table 2.2 below). There has been an increased effort within the research community to understand perceptual compensation for reverberation in the last decade, but the topic remains poorly understood at present.

Two groups in particular have led the research in this field, those of Watkins and Zahorik (detailed in § 2.4 below). Watkins' experimental paradigm makes use of a categorical perception task based around a continuum of speech sounds which are typically, though not exclusively, presented monaurally. The continuum synthetically interpolates between 'sir' and 'stir' tokens at either extreme, with the so-called 'category boundary' locating the point along the continuum at which the listener's percept switches between the two words. In Watkins' experiments, compensation for reverberation is measured by locating the category boundary in a variety of listening conditions, for instance when a test-word contains a low or high level of reverberation. Compensation for reverberation is then quantified by considering the benefit that arises in the interpretation of a reverberant test-word when the listener additionally has access to a similarly reverberated preceding context phrase. In Zahorik's lab, compensation effects have been demonstrated with a much wider range of speech material, but only seem to arise in binaural listening conditions. Listener performance is typically measured in terms of the percentage

of correct word identifications in a given experimental condition, and compensation for reverberation is quantified in terms of the performance benefit arising from a prolonged exposure to the same room condition.

The major finding arising from these experiments is that human listeners perceptually compensate for the effects of reverberation experienced in typical listening environments. Human listeners seem to 'take into account' their surroundings when judging a particular situation so that the underlying categories of speech (e.g., consonants) are still consistently recognised despite considerable acoustic distortions introduced by real-room reverberation patterns. In such cases, the inclusion of a reverberant context prior to the test signal actually *assists* identification of a reverberant speech sound, even though it increases the overall level of late reverberation in the signal. Thus, it appears that the human auditory system somehow – through a process of perceptual compensation for reverberation – recalibrates itself to maximise its speech processing abilities in its current listening environment. Psychoacoustic experiments are now beginning to elucidate the various factors contributing to this auditory recalibration, asking how quickly such compensation mechanisms act, whether they are restricted to speech perception or work more generally, and whether monaural or binaural factors dominate.

As yet, however, very little is known about the physiological mechanisms underlying perceptual compensation for reverberation. One particular area of physiology which we can speculate may prove relevant is the efferent auditory system. Although the role of efferent feedback in auditory perception is not yet fully understood, efferent activity *is* generally accepted to underpin our ability to understand speech in complex, noisy environments. Multiple efferent pathways exist in the human auditory system, each carrying descending electrical signals from a more central brain area back towards an earlier stage of the auditory system. Two such efferent pathways reach right back to the auditory periphery itself: the acoustic reflex acts on the middle ear bones as they transmit sound through the ear drum, and the medial olivocochlear reflex innervates the outer hair cells in the cochlea. It is primarily through the second of these mechanisms that the efferent auditory system is thought to facilitate listening *in noise*, by performing a series of adjustments that ultimately improve the neural representation of sound by controlling the encoding of the signal's dynamic range (Guinan, 2006, 2011). Little evidence yet demonstrates whether the efferent system can similarly enhance the neural coding of speech *in reverberation*. Nonetheless, since reverberation brings about a reduction in the signal's dynamic range (an acoustic effect which is similar in this regard to that of additive noise), it seems feasible that the efferent system may also work to improve neural representation of the reverberant speech envelope, perhaps by revealing temporal modulations of speech that had been distorted by the additional reflected energy.

To amalgamate findings from different strands of experimental research into a single conceptual framework, and to improve hypothesis design for further psychoacoustic investigation, researchers have at times built computer models which instantiate a set of principles thought to be relevant to a particular listening task. Such functional models are particularly useful in cases where the systems in question are complex in structure, and where the overall output behaviour depends on the interactions among many potentially nonlinear components (Weintraub, 1985). Along these lines, a handful of efferent-inspired computational auditory models have been developed in recent years and used to simulate — and hence, explain — listener responses for tasks in which speech is heard in noise: Brown et al. (2010) use the model of Ferry and Meddis (2007); Lee et al. (2011) use that of Messing et al. (2009); and Chintanpalli et al. (2012) build on the work of Zilany et al. (2009). Though the details of each implementation differ, these models each lay out a scheme in which the medial olivocochlear efferent pathway acts to regulate the activity of the afferent (ascending) auditory pathway. In this way, processes deriving information from the acoustic signal may be modified according to clues gained about the specific contextual environment, and the machine listener may thus achieve a degree of robustness to the background noise present in the signal. A similar modelling process might yet help understand whether the efferent system can help explain perceptual compensation for reverberation.

1.3 Problem and purpose

The key finding of research into perceptual compensation for reverberation in human audition – that listeners can recalibrate to their acoustic environment – has been consistently replicated in the experiments carried out by Watkins' and Zahorik's research teams. However, since their two approaches differ significantly in terms of the speech material and audio presentation methods employed, it is not yet known whether the *same* compensation effect is being demonstrated in each lab, or whether different processes are in fact being examined. Furthermore, the physiological processes which confer human listeners' robustness to the effects of reverberation have not yet been uncovered. Here, the twin tools of psychoacoustic experimentation and computer modelling, hand-in-hand, can help to propose and to test hypotheses regarding the form and function of perceptual mechanisms involved in compensation for the effects of reverberation.

An important question to answer is whether monaural listening is sufficient to achieve compensation for the effects of reverberation on naturally spoken speech material. The phoneme-continuum task first described in Watkins (2005a) has repeatedly been used since to demonstrate monaural compensation. However, when

a different speech database was used in Brandewie and Zahorik (2010), a monaural effect was reported for only 2 of the 14 listeners. It is noteworthy, however, that this experiment by Brandewie and Zahorik included a spatialised noise source in addition to reverberation. This does arguably increase the realism of the listening task, but it also exacerbates the difficulty of interpreting listener results since it is unclear whether the compensatory effects investigated by them relate directly to speech identification itself, or might rely instead on spatial hearing mechanisms arising in binaural auditory pathways. As a result, it has not yet been established whether monaural pathways in the auditory system provide sufficient compensatory processes to deal with the effects of reverberation on naturally spoken material.

If monaural compensation effects *do* turn out to be relevant for real-room listening with naturally spoken material, then this could prove to be a significant find. An improved understanding of the relevant mechanisms might suggest further advances in far-field ASR techniques, in speech intelligibility prediction models, or in signal-processing strategies for assistive listening devices such as hearing aids and cochlear implants which are (at present) often worn single-sided.

The auditory efferent system is believed to enhance the neural representation of speech in noise through a process of dynamic range adaptation, and thus may underpin our facility to understand degraded speech signals (Guinan, 2006, 2011). Since reverberation also has an effect on the signal's dynamic range, it is reasonable to propose that the efferent system might also be involved in improving the neural representation of reverberant speech. While efferent-inspired computational auditory models have several times been used in recent years to simulate listener responses for tasks in which speech is heard in noise, their performance in reverberant listening tasks has not been investigated.

If an efferent-inspired computational model of the auditory system *does* provide a good match to listener response data in reverberant speech tasks, then such a model might help to elucidate the physiological processes which control our ability to compensate for reverberation.

1.4 Research objectives

To further investigate perceptual compensation for reverberation, a dual approach underlies the current thesis. As shown in Figure 1.1, a computational auditory model is configured to include an efferent-inspired circuit which enables it to compensate for the effects of reverberation in its environment. In addition, a series of psychoacoustic experiments allow a deeper investigation of the phenomenon in

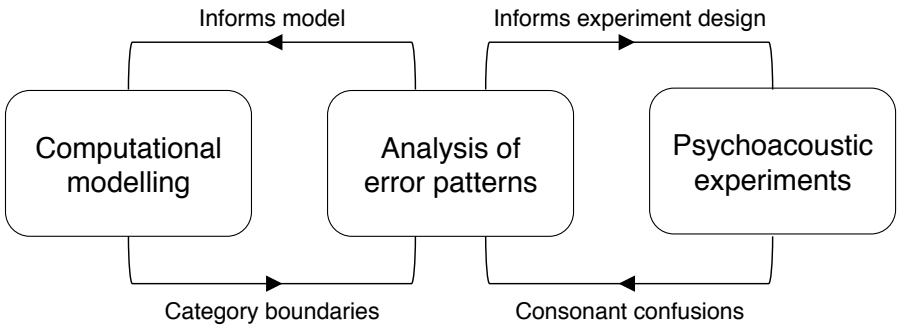


Figure 1.1: Illustration of workflow which involves simultaneous investigation of perceptual compensation for reverberation in human listeners and in machines. A computational model simulates monaural compensation effects in listener data of Watkins (2005a), measured by means of the movement of category boundaries on the 'sir-stir' continuum. Psychoacoustic experiments use naturally spoken speech material (in which the talker, test-word and context phrases change from trial to trial) to address questions arising during the computational modelling process and to further investigate the nature and time course of the monaural constancy mechanism.

human listeners, following and extending the pre-existing experimental paradigms using questions raised by the modelling process.

The computational model presented in this thesis develops the efferent model of cochlear processing first published by Ferry and Meddis (2007). Their model allows efferent suppression to be simulated in a particular frequency region by setting a manual reduction of the gain applied to the afferent processing chain. The model has recently been used to simulate human recognition of speech in noise (Brown et al., 2010; Clark et al., 2012), but has not previously been applied to the study of reverberant speech perception. The current study does just this, by developing a model comprising an afferent auditory pathway with an efferent regulation (feedback) loop, where the level of efferent activity is altered in response to the level of reverberation detected in the signal.

The central idea behind the computational model is that the main effect of late reverberation is similar to the detrimental effect of additive noise, since it increases the noise floor of the signal, and thereby decreases its dynamic range. This thesis asks whether a model of efferent regulatory control can provide additional robustness to reverberation, as has been shown for noise. That is, the thesis is concerned with the question of whether an efferent-inspired auditory model can produce results consistent with human listener data in reverberant speech identification tasks, and eventually thereby help to explain perceptual compensation for reverberation.

The computational model in the current thesis is tested against human listener responses from Watkins (2005a), where listeners monaurally identified 'sir' and 'stir' test-words in a variety of reverberation conditions. In Experiment M1, human re-

sponse data from two room-positions ('near' and 'far') are used to address questions regarding the kind of metric that should drive efferent feedback, and to set corresponding model parameters that determine how efferent suppression should be implemented. Trained thus, the model is then challenged to produce qualitatively similar category boundaries to Watkins' listeners, for stimuli in which abrupt changes in reverberation may occur mid-utterance (so that the sentence is heard in part from nearby, and in part from further afield). Experiment M2 asks whether compensation for reverberation persists in the model, as it does for human listeners, when the speech direction of the utterance is time-reversed and its linguistic content is destroyed. Finally, Experiment M3 considers whether compensation occurs in the model when the time-direction of the reverberation itself is reversed, since the absence of reverberation decay tails at signal offsets is sufficient to block the compensation process in human listeners.

The process of building an auditory model inevitably exposes gaps in our knowledge of how perceptual phenomena arise in human listeners. In order to begin to address a number of questions surrounding perceptual compensation for reverberation that arose during the modelling process, a series of four psychoacoustic experiments are presented.

The first study with human listeners (Experiment H1) fundamentally asks whether the *monaural* compensation for reverberation effect first demonstrated in the 'sir-stir' continuum experiments of Watkins holds relevance in ecologically valid listening environments. In doing so, it seeks evidence of the monaural constancy effect that is sufficient to remove doubts about the monaural mechanism which were raised by Brandewie and Zahorik (2010) and Nielsen and Dau (2010). The listener task in Experiment H1 follows that used in the speech identification paradigm of Watkins' work, but the stimuli comprise naturally produced speech from twenty different adult voices, where the talker, context speech and nonsense test-word vary trial-by-trial. This approach allows perceptual compensation for the effects of reverberation to be measured as a reduction in the number of consonant confusions that listeners make in a given experimental condition.

The remaining three perceptual experiments further probe the nature of monaural compensation for reverberation. Reversals of the time-direction in the speech and/or reverberation content of the stimuli, as in the modelling work above, are used in Experiment H2 to investigate the robustness of the compensation effect to linguistic and acoustic aspects of the signal when naturally spoken speech material is heard. Experiment H3 examines the temporal extent of the signal influencing identification of the test-word, querying whether information derived from reverberation occurring *after* the test item plays a role in addition to the effect arising from the preceding context. Finally, Experiment H4 investigates the time course

of the compensation effect, and examines the way in which a human listener's robustness to reverberation appears to build-up over time.

1.5 Thesis structure

In order that the focus of the current research may be more clearly defined, the next two chapters review existing work which leads towards an understanding of perceptual compensation for reverberation. Chapter 2 first discusses reverberation from an acoustical point of view, examining objective measures of room acoustics and engineering solutions to reverberation-robust ASR. The effects of reverberation on speech identification by human listeners are then reviewed before the body of evidence contributing to our current understanding of perceptual compensation for reverberation is introduced. Chapter 3 considers the compensation mechanism from a computational modelling perspective, and aims to establish which auditory components would be relevant for simulating perceptual compensation for reverberation. A particular focus of this chapter is the adaptation of the auditory system to its present environment, which can be simulated with the inclusion of efferent components as such mechanisms offer robustness in noisy or reverberant environments.

The following two chapters describe the original research carried out into perceptual compensation for reverberation. Chapter 4 presents a computational model based on auditory efferent processing, and describes its customisation for a categorical perception listening task. In three modelling experiments (M1 – M3), word identity decisions made by the model are tested against human listener data from Watkins' phoneme-continuum experiments. A series of four experiments with human listeners (H1 – H4) are presented in Chapter 5 in order to confirm the relevancy of the monaural constancy effect to real-room listening with naturally spoken stimuli, and to address a number of questions arising during the computational modelling process.

Finally, Chapter 6 draws together the findings of the modelling and experimental work, and outlines their mutual implications. Potential criticisms of the work are discussed in light of wider research perspectives, and some suggestions for future work are outlined.

Chapter 2

Identification of reverberant speech

Contents

Chapter overview

This thesis is concerned with understanding and modelling effects of perceptual compensation for reverberation. Overviewing the necessary research background for this topic, this chapter opens with an introduction to reverberation from an acoustical point of view. Following this, the problem reverberation poses to machine listeners is described, and some engineering-approaches that increase robustness to reverberation in speech recognition are discussed. The second half of the chapter outlines the psychoacoustic effects of reverberation on human listeners, and considers speech intelligibility in reverberation. The body of evidence demonstrating compensation for the effects of reverberation is introduced, highlighting specific research questions that are addressed later in the thesis.

2.1 Room acoustics and quantification of reverberation

In a room, there are many routes by which sound may travel from a talker's mouth to a listener's ear. This is sometimes termed 'multipath propagation', and is shown schematically in the upper row of Figure 2.1. In the left-most panel, the direct path of sound from a source to a microphone is shown. Being the shortest distance, this is the sound which arrives first at the receiver's position. Next, the centre panel depicts a number of 'early' reflections arriving after hitting nearby surfaces in the room. These reflections consist of delayed and attenuated copies of the direct signal. They are therefore highly correlated with the direct signal, but are spectrally 'coloured' due to the fact that the rooms' surfaces do not absorb energy equally at all frequencies. The final panel depicts the diffuse sound field arising some time later from numerous tightly-packed reflections. These make up the 'late' reverberation, and cause the archetypal signal degradation that is commonly associated with the presence of reverberation.

The room impulse response (RIR) shown in the lower half of Figure 2.1 provides an alternative description of the way in which sound propagates in an enclosure, now in the time-domain. Mathematically, the RIR represents the response of an acoustic environment, assumed to be a linear and time-invariant system, to an impulse produced within it. As described by the room sketches above, the initial, left-most impulse represents the arrival of the direct sound. Following this, a smattering of early reflections can be seen, their relative strength and timing depending on the geometry of the room in question. The final portion of the RIR constitutes the diffuse 'reverberation tail' containing densely packed higher-order reflections which decay exponentially to silence.

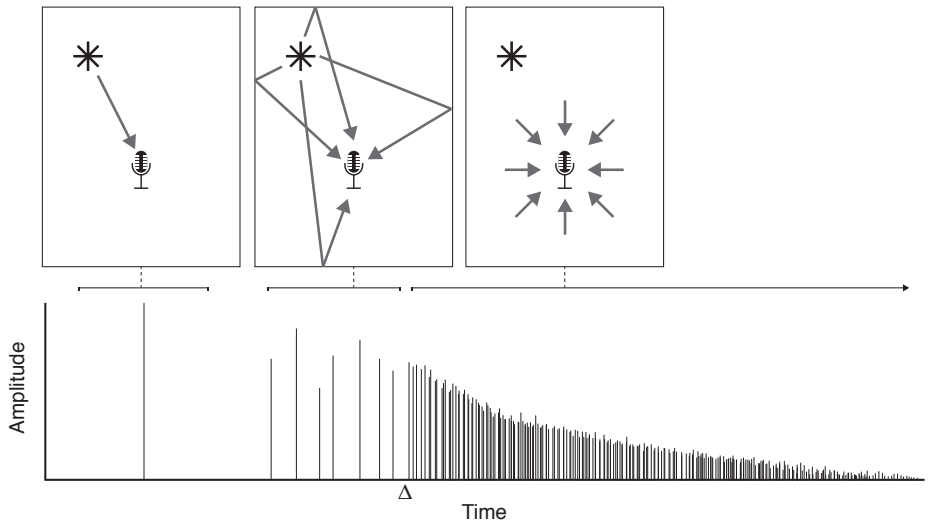


Figure 2.1: Multipath propagation and the room impulse response (RIR). *Above:* Panels show the direct sound path from the source (marked with an 8-point star) to the receiver (*left*); some early reflections (*centre*); and the diffuse late reverberation (*right*). *Below:* Schematic representation of the RIR which fully characterises the acoustic system. The direct signal from an impulsive sound is seen first (*left*). Subsequently, Δ marks the temporal boundary between the early reflections (*centre*) and exponentially decaying late reverberation tail (*right*).

RIRs can be recorded by a variety of methods including the maximum-length sequence technique of Gardner and Martin (1994) or the swept-sine technique of Farina (2000). When recorded using a mannekin with microphones in its ear canals, the resulting binaural room impulse response (BRIR) captures one such impulse response for the left-ear and another for the right-ear. The BRIR thus describes the reverberant qualities of the space as it would be experienced by a human listener, quantifying the spatial and temporal distribution of reflections between a particular pair of source and receiver positions in the room.

The capture and storage of RIRs have facilitated psychoacoustic experimentation into the perceptual effects of reverberation, because the properties of the dry (unreverberated) signal, x, and room impulse response, h, may be independently manipulated, and the reverberant signal $y(t)$ simulated afterwards using linear convolution $(*)$ as follows,

$$y(t) = h * x \tag{2.1}$$

$$= \int_{\tau=0}^{T} h(\tau)x(t-\tau)\,\mathrm{d}\tau \tag{2.2}$$

where the impulse response is of length T, and τ indexes its samples[1]. When the results of such a convolution are later played back binaurally to listeners over headphones, this re-creates the sound-pressure vibrations recorded *in situ* at the ears of the mannekin in the reverberant room, and listeners report that their aural impression is of hearing sound situated in a reverberant room (Allen, 1979; Blauert, 1997).

Although late reflections degrade the signal, early reflections appear to make a positive contribution to the intelligibility of the signal (Bradley et al., 1999). The temporal boundary between the so-called early and late regions of the RIR has therefore been a topic of considerable study (see e.g., Hidaka et al., 2007). Marked Δ on the lower panel of Figure 2.1, this boundary reflects the fact that the energy contributed by convolution with $h(t)$ is therefore initially *helpful* where $t \leq \Delta$, but subsequently *un*helpful for $t \geq \Delta$.

Objective measures of reverberation, discussed below, are largely concerned with quantifying aspects of the early or late portions individually, or else determining the relative balance of energies between these two parts. However, perceptual effects of the reverberation depend not only on the room acoustics, but also on the temporal and spectral features of the signal itself (Houtgast and Steeneken, 1985). Acoustic and psychoacoustic effects of the early reflections and late reverberation are first introduced below; objective methods that attempt to quantify such effects are examined afterwards.

2.1.1 Early reflections

Early reflections occur sufficiently close in time to the direct signal that they may be perceptually fused with it, introducing a spectral colouration to the signal that is defined according to the source-receiver configuration and the particular enclosure in question (see e.g., Arweiler and Buchholz, 2011; Bech, 1995). In other words, the timbral alteration due to the early reflections is strongly dependent on the room positions of the talker and listener (or speaker and microphone) respectively. Early reflections are understood to assist in delivery and understanding of speech in a room, since the reflected energy essentially sums with the direct sound. These

[1] Equation 2.2 reveals that at its heart, convolution is the integral of the product of two functions, one of which is reversed and shifted. This formulation, like the remainder of the thesis which follows, does not include the contribution of a static background noise source (or rather, considers only the cases in which such noise is entirely absent or of sufficiently low-power as to be considered negligible). If desired, however, additive terms could be introduced to the right-hand side of Equations 2.2 and 2.1 to represent the presence of stationary noise sources (e.g., fans, heaters, air-conditioning units) in the environment.

early reflections increase the level of the signal at the ear, and boost the effective ratio of signal to noise components in the sound (Bradley et al., 2003; Nábělek and Robinette, 1978).

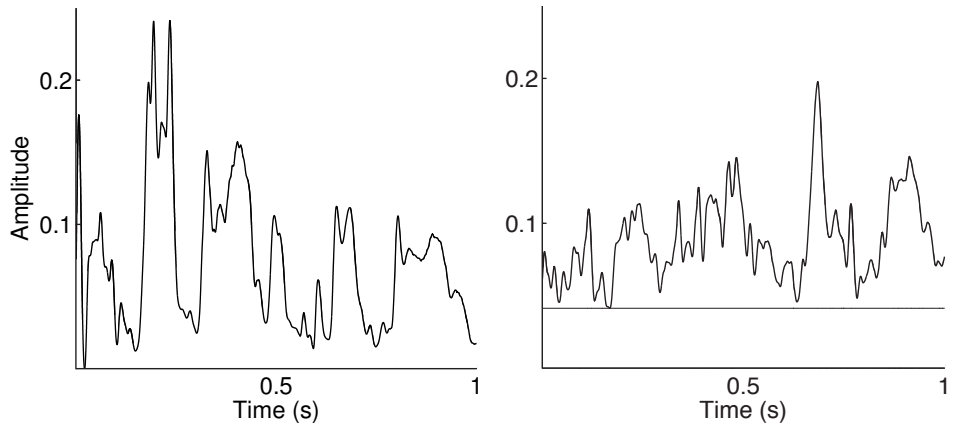
A number of researchers are working on ways to characterise the perceptual influence of early reflections, using either monaural (Meynial and Vuichard, 1999) or binaural (Georganti et al., 2014) approaches. However, much work remains to be done in order to understand the influence of the original signal content itself (particularly its distribution across frequency) on sound perception generally, and on speech intelligibility more specifically (Arweiler and Buchholz, 2011). That is, speech intelligibility depends not only on the room condition, but also on type of voice, rate of speech, and so on. Additionally, it will be important to understand the time course over which adaptation to colouration or timbral differences can occur (Pike et al., 2013). These topics are important because early reflections appear to play a dual role in speech intelligibility when considered from both the listener's and talker's points of view. For example, early reflections in school classrooms can promote a student's ability to hear their teacher (Bradley et al., 2003; Ellison and Germain, 2013), and can reduce the vocal effort required by a teacher to present their words at a given signal level (Pelegrín-García et al., 2011).

2.1.2 Late reverberation

As outlined above, early reflections are reported to be beneficial for the perception of speech in a reverberant environment. On the other hand, the succession of late-arriving reflections, collectively referred to as 'late reverberation', is usually understood to degrade listener performance in reverberant speech-based tasks.

The main effects of late reverberation on a speech signal can be viewed in Figure 2.2, which plots the amplitude envelope of the signal of a voice heard at a nearby distance (*left*) with little influence of reverberation, and when the voice is heard from a far distance (*right*) and the influence of reverberation is considerably stronger. Perhaps the most obvious effect here is that the tightly packed late reflections have combined to effect an increase in the noise floor at the far distance (Hidaka et al., 2007; Koening et al., 1977). Secondly, the reverberation can be viewed as causing an attenuation of a signal's amplitude modulation, as the reflected energy fills the momentary dips in the dry signal (Houtgast and Steeneken, 1985). The dynamic range of the far-reverberated signal appears correspondingly reduced.

As was seen above for the early reflections, the effects of late reverberation are also dependent on temporal and spectral features of the original (voice) signal. This topic is discussed further in regard to speech perception in § 2.3 below.



(a) Near distance: little reverberation **(b)** Far distance: increased reverberation

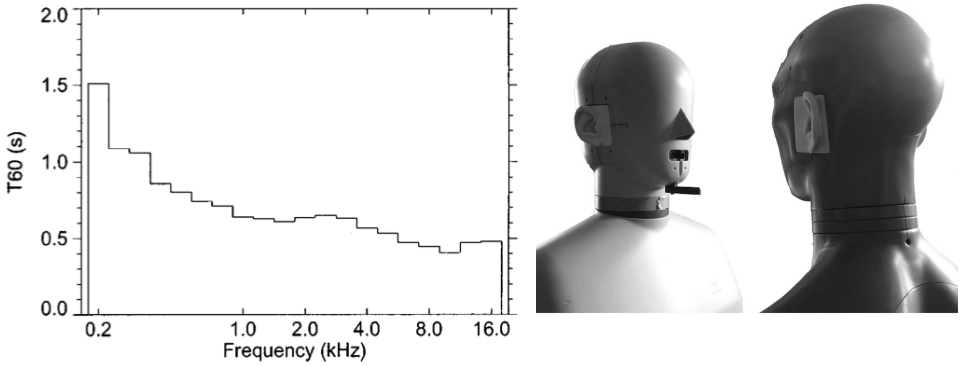
Figure 2.2: Demonstration of the effects of reverberation on a single speech utterance heard from 0.32 m distance (*left*) and from 10 m distance (*right*). The traces show the temporal characteristics of the wideband signal, plotting the amplitudes of low-pass filtered Hilbert envelopes (cutoff frequency 50 Hz). The near distance signal consists primarily of direct sound and thus has a strongly modulated amplitude envelope and very low noise floor (visually indistinguishable from the abscissa). The amount of reverberation increases when sound is heard from the far distance. Here, the amplitude envelope trace reveals an increase in the noise floor (depicted with the raised horizontal line) and an attenuation in modulation depth contributing to a lessening of the signal's dynamic range.

2.1.3 Room measures and room-position measures

Since the perceptual effects of reverberation depend on qualities of both the source signal and room properties, it is hard to wrap up all the relevant aspects of room acoustics in a single measure. Nonetheless, such a simplification has often been sought since it would be convenient to represent a room's effects in such way.

Since the early 1920s, the most commonly-used reverberation measure is 'reverberation time', or T60 as it is usually stated. This defines the time interval required for the reverberant signal to drop by 60 dB relative to the level of the direct sound, i.e. to decay to one millionth of its original intensity (Sabine, 1922). In general, short reverberation times (small T60) are preferred for listening to speech, since this minimises the persistence of reflected sound from one syllable into the next (Náběl ek et al., 1989). On the other hand, longer reverberation times (larger T60) are preferable for music listening (Hidaka et al., 2007; Lokki et al., 2012).

When expressed as a single number, T60 is likely to have been measured from a wideband signal that is averaged across frequency from c. 20 Hz to 20 kHz. However, it is clear from Figure 2.3a that the value of T60 in fact varies with frequency, primarily because the various surfaces in the room (including any people present) absorb sound non-uniformly. Moreover, since the idea of T60 is to summarise the



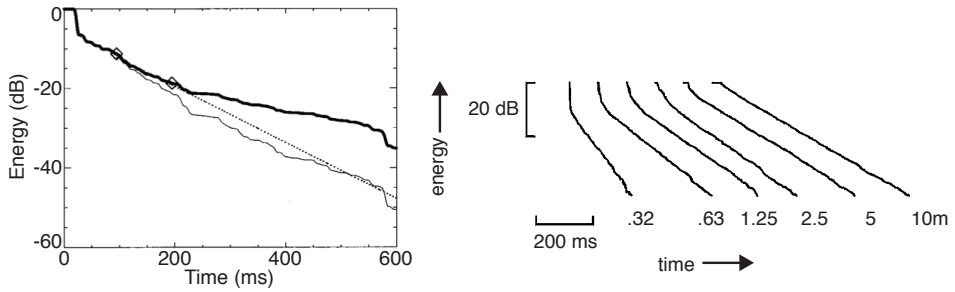
(a) Reverberation time, T60 **(b)** Binaural mannekins

Figure 2.3: Directional and non-directional measurements of reverberation. Figure 2.3a plots reverberation time, T60, the average time required for the reverberant signal to decay by 60 dB relative to the level of the direct sound (averaged across 12 positions in an auditorium). T60 varies across frequency, with longer T60s recorded for lower frequency bands (figure from Zahorik, 2002). Rather than characterise the average room conditions, the binaural mannekins shown in Figure 2.3b allow the highly directional characteristics of talkers and listeners to be captured in particular source-receiver positions within a given room (photograph by the author of mannekins used in Watkins, 2005a).

room's overall acoustic reflection characteristic, measurements are usually taken at a number of different source-receiver positions and averaged. This reduces the influence of the selected recording position within the room, and thereby tries to summarise the overall room response.

An alternative approach to studying reverberation is to try to capture exactly that variation which has just been 'averaged out' in the T60 measure. This is the approach taken, for instance, by Lokki et al. (2012) whose aim was to simulate the room acoustics that would be experienced at a given seat in a particular concert hall. The position-specific approach was also taken by Watkins (2005a), where the reverberation characteristics of individual source-receiver positions are captured by a pair of binaural mannekins shown in Figure 2.3b. This allows recreation of the highly directional characteristics of talkers, whose mouths project most of the sound energy forwards, and of listeners, whose individual ears are subject to acoustic effects from the room itself and from interactions of the sound with the body as discussed below.

An alternative way to characterise reverberation is with an energy decay curve (EDC). The EDC can be found whenever the RIR, $h(t)$, is known, and is arguably a more intuitive description of the room characteristic as a listener would experience it. Introduced by Schroeder (1965), the decay curve is computed from the RIR



(a) Zahorik (2002) **(b)** Watkins (2005a)

Figure 2.4: Energy decay curves. Figure 2.4a plots the energy decay function for a mid-frequency band (centred at 635 Hz) recorded c. 5 m distance (as shown in Zahorik, 2002). The time to arrival of the direct sound can be seen in the initial horizontal part of the curve, then a sharp drop characterises the early reflections and an approximately linear decay constitutes the late reverberation. The *dotted* line is a linear fit to the decay observed in the region marked with *diamonds* before 200 ms. The *dark* curve is contaminated by noise in the environment, which introduces an additive component with cumulative effects which raises the decay curve through time. When the noise contamination is removed, the upward-swing of the dark line is less pronounced and the resulting *light* line more closely approximates the anticipated linear decay. Figure 2.4b, redrawn from Watkins (2005a), shows energy decay slopes for six source-receiver positions in an L-shaped room. The rate of decay appears independent of the positions of talker and listener beyond a certain distance, but directional effects (e.g. head shadow) influence the rate of decay considerably when the talker and listener are nearby.

directly, as the tail integral of the squared impulse response at time t, so that

$$EDC(t) = \int_{\tau=t}^{\infty} h^2(\tau)\,\mathrm{d}\tau, \tag{2.3}$$

and $EDC(t)$ represents the energy remaining in the impulse response at time t. The smooth, linear decay of the EDC is apparent from Figure 2.4, particularly in relatively noise-free conditions. In Figure 2.4a, the upward drift of the dark curve reveals the effect of noise contamination in the enclosure which acts cumulatively through time to gradually reduce the incline of the EDC (Schroeder, 1965). In the example shown here, a process of noise reduction has subsequently been undertaken by Zahorik (2002) to observe the reverberant effects of the room alone. The resulting light curve more closely matches the theoretical linear decay that was extrapolated from the observed data marked between the two diamond shapes when the signal levels were sufficiently strong that the reverberation effect dominated that of the noise.

Figure 2.4b shows six decay curves recorded by Watkins (2005a), again after a noise-reduction technique has been applied, for multiple source-receiver distances (SRDs) within a single room. While late-reverberation decayed approximately exponentially in the RIR (cf. Figure 2.1), the EDC shows a linear decay. A number of

other features additionally become apparent. First, the fast drop (i.e., vertical portion) at the start of the EDC at near distances shows the high proportion of energy that was experienced in the direct sound and its early reflections. This proportion gradually reduces so as to be almost imperceptible at the furthest distances. Conversely, the horizontal portion at the start of the EDC at far distances represents the delay in arrival of the direct sound travelling from the source position (which is almost imperceptible at the short distances).

Since the EDC gives an intuitive representation of the room's decay, it is often used to calculate the reverberation time of an environment: a drop of 60 dB in energy may be measured (or extrapolated when noise sources are present) in the final linear portion of the curve more easily that it could in the amplitude envelopes of the RIR[1]. However, the variation in slope in Watkins' curves in Figure 2.4b highlight that caution should be taken in regard to the claim that the rate of decay of the late-reverberation is independent of the positions of talker and listener. While the assumption in this claim holds approximately true for the longer SRDs, where the linear portion of the curves for distances greater than 0.5 m are approximately parallel, the rate of decay is considerably steeper for the closest talker-listener position (0.32 m). This is likely due to the directional characteristics of sound production at the mouth of the talker (i.e., the transducer loudspeaker of the source mannekin), and of sound reception at the ear of the listener (i.e., involving sound interaction with the pinnae, head and torso of the receiver mannekin).

For sufficiently large source-receiver distances, the assumption that the magnitude of the late-reverberation is insensitive to the exact talker and microphone locations has proved useful in speech technology applications. An improvement has been recorded in far-field ASR, since algorithms based on this assumption can become robust to movements in the talker's location (Yoshioka et al., 2012).

2.1.4 Measuring reverberation from an impulse response

As discussed above, early reflections are generally considered beneficial to sound propagation and perception while late-reverberation is not. Most reverberation measures are therefore, at their root, an attempt to somehow quantify the balance between these helpful and unhelpful parts of the room's distortion, shown in the second and third panels (upper row) of Figure 2.1 respectively. When reverberation is artificially simulated, the distinction between early reflections and late reverberation, marked Δ on Figure 2.1, is easy to determine, in particular since different

[1] It is common to report the T60 value even if has been extrapolated from the corresponding T20 or T30, i.e. the time taken for a drop of 20 dB or 30 dB respectively (Campanini and Farina, 2009).

techniques are often used to create the individual early reflections (e.g. using the image model of Allen, 1979) and the subsequent diffuse decay (e.g. using independent Gaussian noise samples as in Zahorik, 2009). In real rooms, on the other hand, the distinction between these two parts of the impulse response is a matter of some debate. Two main approaches to characterise the relative importance of the early and late portions of the impulse response, *time*-based and *level*-based approaches, are discussed below.

The time-based approach divides the impulse into two parts at a certain point in time, Δ, and compares the balance of energies present in the initial (early) and subsequent (late) portions. This method underlies a number of early-to-late indices, perhaps the most often used of which is the 'clarity index', C_Δ, given by

$$C_\Delta = 10 \log_{10} \frac{\int_{t=0}^{\Delta} h^2(t)\,\mathrm{d}t}{\int_{t=\Delta}^{\infty} h^2(t)\,\mathrm{d}t} \tag{2.4}$$

which is measured in dB and compares early contributions (in the numerator) with late contributions (in the denominator). A directly related measure, the 'definition index', D_Δ, quantifies the early portion as before, but quantifies this against the *total* sound energy described in the impulse response using

$$D_\Delta = 10 \log_{10} \frac{\int_{t=0}^{\Delta} h^2(t)\,\mathrm{d}t}{\int_{t=0}^{\infty} h^2(t)\,\mathrm{d}t} \tag{2.5}$$

which is again measured in dB (note here that only the lower limit of the integral in the denominator has altered).

For speech perception, setting $\Delta = 50$ ms seems to correlate highly with speech recognition performance (Nishiura et al., 2007). The corresponding clarity index at this time constant is known as C_{50} and is seen to be a good indicator of intelligibility; other values of Δ may be appropriate for alternative listening situations, however, and C_{80} is often a more relevant measure than C_{50} in rooms in which music is to be heard (Chesnokov and SooHoo, 1998; Hidaka et al., 2007; Marshall, 1996). Here, if the clarity index $C_\Delta \simeq 0$, then the ratio implies the early and late sound energies are approximately equal. A large positive C_Δ value implies the late sound is effectively absent, while a negative C_Δ value implies that there is more late sound than early sound present (as might occur, for example, when the 'direct' sight-line is broken and the listener and talker are not facing one another, or are around a corner in an L-shaped room).

If the value of Δ is reduced sufficiently, then at a certain time-point, T_d, it serves to separate the direct sound-path from the remainder. This is the method underlying calculation of the direct-sound to reverberant-sound ratio (DRR). Here, the contents of the impulse response from zero up to T_d, are said to correspond to the anechoic direct path of the signals captured. The time value should therefore depend upon the distance between the source and receiver (Naylor et al., 2010), and is correspondingly short (e.g., a value c. 2.5 ms was used in Zahorik, 2002). The 'direct' portion is then compared with the remainder of the values in the impulse response to obtain the DRR value,

$$DRR = 10 \log_{10} \frac{\int_{t=0}^{T_d} h^2(t)\,\mathrm{d}t}{\int_{t=T_d}^{\infty} h^2(t)\,\mathrm{d}t} \tag{2.6}$$

which directly mirrors the relationship of initial and latter parts of the RIR described in Equation 2.4 above.

Rather than calculating energy ratios in a fixed time period, the second approach to determining the room's influences depends instead on measuring the time taken for a particular energy level relationship to transpire. Such an approach was seen above for the T60 reverberation metric, where the time required for a 60 dB drop in the signal level was recorded. A related measure captures the influence of early reflections, the early decay time (EDT), which records the time over which a signal makes its initial 10 dB decay. Referring back to the energy decay curves in Figure 2.4b, at nearby source-receiver distances this 10 dB drop comprises a sharp loss associated with the cessation of the direct sound and the early reflections. At long SRDs however, there may be insufficient energy received in either direct signal and early reflections to create this sharp 10 dB drop. Here, were the EDT measure to be used, then it would instead be mainly describing the linear slope of the late reverberation decay.

2.1.5 Measuring reverberation from a reverberant signal

If the impulse response of the room is not known in advance, then aspects describing the reverberation quality of the environment must be estimated from a reverberant signal instead. One approach to this problem is to estimate the impulse response first, and then proceed to quantify reverberation using the estimated RIR with the methods described in the previous section. However, a number of different approaches have arisen to quantify reverberation without access to the real or estimated impulse response. Two such methods are briefly discussed.

The first of these measures, the signal-to-reverberation ratio (SRR), relies on the availability of the clean speech signal, and is thus usually used to assess *de*reverberation of a dry speech signal that had previously been artificially reverberated (Furuya and Kataoka, 2007; Naylor et al., 2010; Tsilfidis and Mourjopoulos, 2011; Westermann et al., 2013). By considering the effect of reverberation as noise, the calculation is analogous to that of the signal-to-noise ratio (SNR) and measures the ratio of the direct sound power and reverberant sound power (expressed in dB).

A second signal-based approach, the noise-to-mask ratio (NMR), relies on prior knowledge of the human auditory system rather than on knowledge of the dry speech signal as was the case above. Again, the reverberation is treated as a noise-like distortion, and an analysis is made to determine which parts of the reverberation may produce audible noise components (Furuya and Kataoka, 2007; Tsilfidis and Mourjopoulos, 2011; Tsoukalas et al., 1997; Westermann et al., 2013). NMR increases from 0 dB when the reverberation is at the threshold of audibility, with higher values indicating more perceptible distortion.

2.1.6 The Modulation Transfer Function

Described by Houtgast and Steeneken (1973, 1985), the Modulation Transfer Function (MTF) characterises the transmission of sound in a room. The term was borrowed from optics, where it was used to assess sharpness in visual scenes. Here, the fundamental concept underlying the MTF in the audio domain is that the signal reaching a listener's ear through the room is a 'blurred' (rather than exact) copy of the original source signal.

In this method, the room is treated as a linear time-invariant system. To measure the distortion it exerts on a signal, the room is tested with a modified 'sine in/sine out' paradigm, with the modification being that the analysis is done in the modulation domain. That is, a series of test signals (typically noise bands across seven octaves of the frequency range from 125 Hz up to 8 kHz) are modulated by sine waves at a range of modulation frequencies (0.63 to 12.5 Hz in third-octave intervals), and any attenuation in magnitude or modulation depth in the intensity envelope that results from transmission through the room is recorded.

By taking as their focus the frequency ranges and amplitude modulation rates required for understanding speech, Houtgast and Steeneken proposed the MTF as a predictor of the speech intelligibility that would be recorded were it to be investigated in the room in question. Indeed, the MTF has proved useful in the decades since and has formed the basis of several subsequent methods of determining speech perception in rooms, for instance, the Speech Intelligibility Index (SII) discussed below.

It is noteworthy at this stage that the phase delay is disregarded in the MTF formulation specified by Houtgast and Steeneken (1985), with only the magnitude (amplitude) characteristic being used. Discussion returns to this point several times in the following thesis, since the exclusion of the complex component of the function renders the MTF insensitive to variation in the time-direction of reverberation. Admittedly, this limitation would not readily be exposed in real architectural acoustics, but it has nonetheless been investigated psychophysically by a number of researchers. Data from Watkins (2005a) and from Longworth-Reed et al. (2009) both suggest that human speech perception in rooms is strongly affected by the time-direction of the reverberation. In these experiments, the RIR is time-reversed before convolution with dry speech signals, and thus simulates the effects of reverberation in reverse (with slowly rising ramps being present before signal onsets rather than slowly decaying tails being present after signal offsets). Since the real part of the MTF does not capture the resulting differences in the signal, any subsequent models of speech perception which rely on it (such as the Speech Transmission Index (STI) discussed next) consequently fail to predict human performance in these listening conditions.

2.1.7 Predicting speech intelligibility

To understand the perceptual effects of acoustic distortion introduced by real rooms, it is also instructive to consider a family of methods developed primarily for objectively predicting the intelligibility of speech signals. The common thread underlying these methods relates to an assumption that the intelligibility of the whole signal may be estimated from a weighted sum of the individual spectro-temporal modulations of which the signal is comprised (Bronkhorst, 2000). The methods typically comprise a two-stage analysis: firstly, sound is transformed into an internal 'auditory' representation; secondly, a decision metric transforms the auditory representation into a judgment regarding the signal's intelligibility (Chabot-Leclerc et al., 2014).

The Articulation Index (AI) was developed at Bell Telephone Laboratories in the 1920s (published around 25 years later), and used the concept of 'articulation' to characterise the probability of correctly transcribing phonemes and syllables transmitted by telephone (Allen, 1996; Fletcher and Galt, 1950; French and Steinberg, 1947). Defined in this sense, articulation is a useful measure to judge transcriptions of nonsense words or unknown languages since transmission of meaning, required for signal intelligibility, is not relied upon. Rather, the AI method describes joint physical properties of the talker, listener and the channel between them, and describes the influence (or absence) of energetic masking on the signal. The speech signal is divided into separate frequency bands (usually 20), each of

Table 2.1: Intelligibility prediction methods reported in Chabot-Leclerc et al. (2014). The Speech Transmission Index (STI) considers the audibility of spectral regions and assesses reduction in temporal modulation of the speech envelope using the Modulation Transfer Function (MTF). Modifications to the STI model include the two-dimensional spectro-temporal modulation index (STMI) of Elhilali et al. (2003), and the speech-based envelope power spectrum model (sEPSM) which calculates the ratio of modulations in the speech and noise envelopes (Jørgensen and Dau, 2011). Further abbreviations used in table (alphabetically): E, envelope compression; N, stationary noise; P_j, phase jitter; P_s, phase shift; R, reverberation; S_s, spectral subtraction; S_d, spectral distortion.

Study	method	decision metric	can account for	cannot account for
Steeneken and Houtgast (1980)	STI	MTF(t)	R, N	E, P_j, P_s, S_s
Elhilali et al. (2003)	STMI	MTF(f,t)	N, R, P_j, P_s	S_s
Jørgensen and Dau (2011)	sEPSM	SNR$_{env}(t)$	N, R, S_s	S_d

which independently makes a contribution to the overall articulation score. SNRs are calculated in each channel, and a weighted sum across the frequency axis of values proportional to the channel's SNR then results in a single number to describe articulation. This index is scored intuitively between 0 and 1, such that an AI of less than 0.3 is considered to be poor, but an AI above 0.7 describes excellent conditions for listening to speech[1].

The AI was later altered and standardised in ANSI S3.5-1997 as the Speech Intelligibility Index (SII). The SII provides a measure of intelligibility of speech, dependent on the talker and individual listener profile. However, it can only describe intelligibility in cases in which background noise, if present, is stationary (see e.g., Chabot-Leclerc et al., 2014).

In contrast, the Speech Transmission Index (STI), based on the Modulation Transfer Function (MTF) which was described earlier in § 2.1.6, provides a measure of intelligibility that can account for the effects of reverberation as well (Steeneken and Houtgast, 1980). Since this intelligibility measure now includes a description of the modulation character of a room, the STI value is correspondingly sensitive to the modulation depth of specific low-frequency bands in a test-signal that, if reduced during transmission through a reverberant room, would be associated with loss of intelligibility (Houtgast and Steeneken, 1973).

A criticism of the STI, however, is that it fails to predict intelligibility in conditions with nonlinear processing such as envelope compression, phase jitter, phase shifts or spectral subtraction (Chabot-Leclerc et al., 2014). Some recent attempts

[1] Kryter (1962) describes practical methods to adapt the AI score for certain factors, e.g. subtracting an offset in reverberant conditions (depending on T60), or adding an amount to account for the beneficial presence of visual cues. However, Kryter also reminds us that the AI is only defined for male voices, and does not predict the audibility of female voices.

to improve on the STI are outlined in Table 2.1. These methods differ in regard to whether the pre-processing and decision stages use modulation information in the spectral, temporal or spectro-temporal domains. Inspired by neural responses observed in the auditory cortex of ferrets, the spectro-temporal modulation index (STMI) of Elhilali et al. (2003) introduces two-dimensional modulation processing to consider modulation across the frequency domain in addition to the temporal modulations considered in the STI. An alternative modification to the STI was made by Jørgensen and Dau (2011), in which the ratio of speech and noise envelopes determines the predicted intelligibility score (rather than the reduction in modulation due to distortion). More recent improvements to these models are reviewed by Chabot-Leclerc et al. (2014).

This area of research is still receiving a great deal of attention: clearly there is a strong desire in the community for an accurate prediction of speech intelligibility in the face of varied signal distortion. Further, it is interesting to note the type of distortion under investigation has gradually shifted through the years from dealing with stationary background noise sources to fluctuating maskers (competing speakers) and room reverberation, and now includes the types of distortion which occur in digital transmission. Though increasingly accurate at making a global prediction of signal intelligibility in a fixed setting, such methods do not yet include a temporal adaptation component that accounts for a human listener's continual recalibration to their listening environment.

2.2 Machine listening in real-room reverberation

Perhaps the most significant benefit of overcoming reverberation in machine listening would be to allow *distant* speech recognition. This could, for instance, allow meetings involving numerous participants to be transcribed via a single microphone, or allow speeches to be subtitled automatically even across large public spaces. Understanding the effects of reverberation would also be advantageous in the development of scene-recognition hearing aids that would allow wearers to adapt quickly to changes of room condition (Büchler et al., 2005), an application which will likely grow in relevance as the prevalence of hearing disorders continues to increase. There is still much work to be done on reverberation, even after a half century of research.

Many studies report a decrease in automatic speech recognition (ASR) performance as the level of reverberation in the signal increases owing to an increase either in the source-receiver distance or in the reverberation time of the acoustic enclosure (see e.g., Couvreur and Couvreur, 2004; Eneman et al., 2003; Giuliani et al., 1996; Palomäki et al., 2004). For example, a study by Kingsbury et al.

(1997) showed a deterioration from around 15% word error rate (WER) in clean conditions to around 77% WER in reverberation. Diverse signal-processing based strategies have been developed to address the problem that reverberation poses to machine listeners. For instance, by modelling temporal envelopes of the speech signal in narrow sub-bands, Thomas et al. (2008) showed a vast improvement in recognition rates (but would nevertheless render many applications unusable as WER remained about 6% for reverberant speech compared with 1% in clean conditions). In the main, these methods are unable to achieve recognition rates for reverberated speech comparable to those for unreverberated speech, and a substantial gap remains between human and machine performance in reverberation.

Four main strategies for reverberation-robust ASR are discussed below. The first strategy attempts to improve the input signal representation provided to the recogniser by means of some pre-processing or feature enhancement. The second strategy attempts instead to improve the manner in which the recogniser itself deals with the reverberant speech. The third strategy deals with reverberation by quantifying the signal uncertainty that it has introduced. The fourth strategy draws from more auditory-like processing techniques, and hints that bio-inspired methods may yet lead to improvements in reverberant signal processing techniques. Of course, in order to achieve the best recognition results, authors often combine a variety of methods.

Initially, however, this section next outlines a standard machine listening system and shows that, among the major theoretical and technical advances in ASR, there have additionally been bio-inspired gains, particularly in cases where signals are degraded by background noise. However, the success of these methods has been limited when applied to reverberant speech, primarily because the reverberation spreads over a much longer time-window than noise-based methods were designed to deal with.

Figure 2.5 depicts the general framework (described by Yoshioka et al., 2012) which underlies the majority of speech recognisers developed in the past few decades. In the so-called 'front-end' of the system, a time-series of amplitude measurements of the sound are transformed into short-term spectral estimates every 10 ms or so. In training, these spectral frames are used as feature-vectors (i.e. observations) to create a set of acoustic models, for example, Gaussian-mixture-based hidden Markov models (HMMs), capturing the various speech sounds encountered in the training data. Similarly, language and pronunciation models aim to represent the common linguistic structures, by exploiting prior probabilities across many sequences of words and phones respectively. During recognition, the 'back-end' of the system then decodes the extracted feature-vectors to provide the most likely word sequence, given the previously trained statistical models.

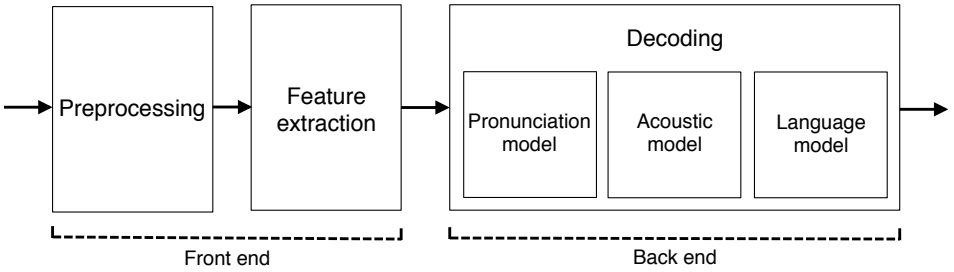


Figure 2.5: A common automatic speech recognition (ASR) framework, redrawn from Yoshioka et al. (2012). The front-end of the recogniser encodes the input audio signals in a series of feature-vectors which are then passed to the back-end to determine the most probable word-sequence given the sets of pre-stored statistical models.

Stern and Morgan (2012) list a series of gains that are sought within such a framework: improving the feature representation in the model's front-end; using a larger amount of training data; finding more representative statistics for the acoustic and language models in the back-end. Beyond this, a number of successful techniques can combat the effects of both additive and convolutional signal distortions arising from interfering sound sources and/or transmission characteristics of the channel (e.g., mismatches in frequency responses of the microphones used)[1]. Despite their effectiveness in noise, however, these methods provide little benefit in the face of real-room reverberation since they typically work on the time-scale of the frame-size of the feature-extraction method and thus cannot account for the type of distortion encountered in reverberant speech, where the reflected energy from a particular sound at a particular point in time will cover several consecutive time-frames in the feature representation. Indeed, methods to combat such longer-term distortions (discussed below) do not yet provide machine listeners with the same robustness to reverberation that a human listener enjoys.

2.2.1 Bio-inspired gains in speech recognition

At times, the state of the art in machine hearing has been increased by incorporating system components designed after observation of psychoacoustic processes (Hermansky et al., 2013; Stern, 2011). The human auditory system has evolved specialised neural circuits allowing us to 'focus' on particular sounds of interest even when they are heard among potentially louder or nearer distracting sounds. As

[1] These include, for example, the maximum a posteriori (MAP) estimation technique of Gauvain and Lee (1994), the maximum likelihood linear regression (MLLR) method described by Leggetter and Woodland (1995), the parallel model combination method of Gales et al. (1996), or the vector Taylor series (VTS) methods suggested in Acero et al. (2000) and Droppo and Acero (2008).

a result, the benefit from biologically-inspired approaches tends to become more apparent in adverse listening conditions, i.e. when signals are degraded rather than clean. Indeed, such approaches may eventually prove useful for machine listeners in reverberation, whether for reducing errors in ASR performance or for improving the representation of target sounds in complex environments for the signal processing algorithms used in hearing aids or cochlear implants.

Observation of the human auditory sensitivities to frequency and intensity have guided the representation of sound that is encoded in standard ASR feature-vectors. Three frequency scales are typically used in auditory modelling studies – Mel (Stevens et al., 1937), Bark (Zwicker, 1961) and equivalent rectangular bandwidth (ERB) (Glasberg and Moore, 1990) – the first two of which additionally form the basis of standard ASR feature-vectors. Each scale approximates the way in which the auditory bandwidth varies with frequency (low-frequency channels have approximately constant bandwidths, while at higher frequencies the bandwidth increases with constant-Q). The Mel or Bark scales are used not because they approximate human perception but rather because they have led to higher accuracy in ASR than was achievable with a linear-frequency scale. A similar story underlies the handling of intensity in both Mel-frequency cepstrum coefficient (MFCC) and perceptual linear predictive (PLP) features. Using the Mel-frequency axis, MFCCs include a log-amplitude compression factor so that equal intervals on the intensity scale represent *equal increments* of perceived loudness (Davis and Mermelstein, 1980). PLP features, based on the Bark frequency scale instead, include amplitude compression via a nonlinear power law to approximate perceptual loudness judgements (Hermansky, 1990). In this case, equal intervals on the intensity scale thus represent *equal ratios* of perceived loudness. More recently, the power-normalized cepstral coefficient (PNCC) features developed by Kim and Stern (2012) claim improved recognition achieved in part by following auditory principles.

Spectral distortions arising from the early reflections in reverberant signals can largely be compensated by the noise-based methods used to combat linear channel effects, microphone mismatch, interfering sounds and so on as mentioned above (e.g., using cepstral normalisation techniques described by Liu et al., 1993). However, to compensate for the effects of the late-arriving reflections, new methods are required which can account for the longer-term temporal evolution of the reverberant signal. Despite some bio-inspired gains in reverberant speech processing (discussed below in § 2.2.6), it remains the case at present, however, that abstract functional models provide the best speech recognition performance. The next part of the thesis briefly overviews such methods.

2.2.2 Reverberation-robust signal processing

The idea behind reverberation-robust signal processing is to achieve a type of context-sensitive machine hearing by which the detrimental effects of unknown environments can be somehow cancelled in signal recognition.

As described above, ASR systems typically use statistical algorithms to match the input acoustic signal against an acoustic model for each speech-component expected (e.g. on the phone- or word-level). Reverberation causes temporal smearing in the signal, and as a result, information from a specific speech sound is spread over several frames. The duration of the room impulse response is usually much longer than the frame length of the analysis window used for feature extraction, perhaps a half-second or so in comparison to the 10 ms time-frame used for feature extraction. To be successful in reverberation, therefore, a speech recogniser would ideally put this longer-term context (encapsulated in a number of consecutive previous frames) to good use. Capturing such long-term temporal evolution with HMMs, however, would require extremely long left contexts for each HMM state (using polyphone rather than triphone models) which would quickly become problematic due to the high number of resulting speech classes and the diminishing number of training examples remaining in each class (Yoshioka et al., 2012).

A more popular way to deal with reverberation within the ASR framework outlined in Figure 2.5 has been to train the acoustic models using reverberated speech data. This method indeed improves recognition results provided that the reverberant conditions of the test material match that of the training but, somewhat counter-intuitively, results in an increased WER when the speech signal is clean. Moreover, since the exact pattern of reflections varies considerably whenever the source or receiver moves, when trained in this way the recogniser can be particularly vulnerable to alteration in the physical position and environment of the talker and microphone (McDonough et al., 2008). Multi-condition training attempts to resolve such problems by conditioning the recogniser using speech recorded in a *range* of acoustic conditions. However, it remains an open question whether this can provide high performance in *all* environments that might be encountered.

The remainder of this section introduces four strategies which have improved the state of the art in reverberation-robust speech processing. Each strategy is discussed sequentially below, though in practise real-world systems may employ multiple strategies in combination in order to maximise recognition rates. Firstly, a set of methods are discussed which focus attention on the front-end and attempt to remove the effect of reverberation from feature vectors themselves (§ 2.2.3). Secondly, methods which alter the acoustic models or decoder to deal more appropriately with reverberant feature vectors are considered (§ 2.2.4). Thirdly, approaches which deal with reverberation through missing data and uncertainty are introduced

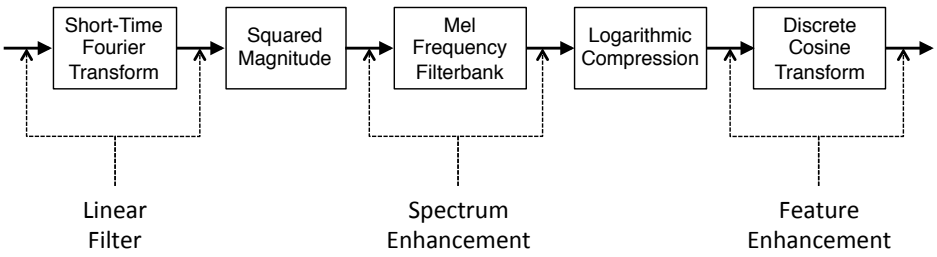


Figure 2.6: Front-end enhancements for reverberation-robust ASR. The schematic redrawn from Yoshioka et al. (2012), shows three enhancement methods (*below*) which may be employed during feature-extraction (*above*) to reduce the effects of reverberation.

(§ 2.2.5). Finally, some methods inspired in part by aspects of biological audition are discussed (§ 2.2.6).

2.2.3 Front-end approaches

The front-end of the recogniser combines optional signal pre-processing and feature-vector extraction. As an extension of this concept, front-end approaches to reverberation-robust ASR aim to remove the detrimental effect of reverberation from the feature vectors themselves. Spatial filtering techniques address this problem by exploiting the fact that speech is directional, and thereby enhancing sound from one direction whilst suppressing everything else. Perhaps the simplest example involves use of a single directional microphone, where the talker's position can be accurately predicted. More generally, an array of microphones and a *beamforming* technique could be used to target a specific (stationary) talker and enhance the direct sound while attenuating the reflected portions in the signal (van Veen and Buckley, 1988). Yoshioka et al. (2012) classify a variety of additional front-end feature-enhancement methods in relation to the stage of feature extraction at which they act, as shown in Figure 2.6.

The first front-end method, *linear filtering*, makes use of a number of consecutive time-frames and thus can account for longer-term temporal evolution of a reverberant signal. This approach addresses the very start of the signal processing chain, using an adaptive filter to estimate and remove the effects of reverberation either on the time-domain signal itself or on its short-time Fourier transform (STFT) representation. This forms the basis of many dereverberation techniques, in particular the blind deconvolution method which first estimates the room impulse response and uses this to estimate a clean version of the reverberant signal (Gillespie and Atlas, 2003; Hopgood and Rayner, 2003). In conjunction with long-term linear

prediction, the method has improved ASR results both for single-input and multi-microphone systems (Kinoshita et al., 2009; Nakatani et al., 2010).

A second front-end approach attempts to remove the effect of reverberation when the signal has been converted into a spectral representation. Given a sequence of reverberation-corrupted observations, the *spectrum enhancement* method attempts to estimate coefficients equivalent to the corresponding clean signal. Since it occurs after the squared magnitude step of feature extraction (see Figure 2.6), the method offers robustness against talker movement since the late reverberation is of comparable magnitude irrespective of the relative positions of the talker and microphone (Yoshioka et al., 2012). This also allows the reverberant distortion to be removed in a manner similar to that of additive noise distortion. Here, a time-shifting spectral subtraction method is required, one option being to make use of the anticipated exponential decay in the magnitude of the late reverberation (Lebart et al., 2001).

A third front-end approach, *feature enhancement*, attempts to infer clean features directly from the reverberant features, i.e. to dereverberate the reverberant features. Rather than exploiting characteristics of the room, feature-based techniques capitalise on structural properties of speech itself, and aim to reduce the mismatch between reverberant and clean speech representations. For instance, the speech signal can be characterised by its modulation rate and, in particular, the low-frequency temporal envelope modulation characteristics can be used to find an inverse modulation transfer function (Avendano and Hermansky, 1996; Langhans and Strube, 1982).

2.2.4 Back-end approaches

Rather than focus on the feature representation of the signal, back-end approaches instead attempt to improve the processing stage in which the reverberant features are decoded (cf. Figure 2.5).

The decoder essentially comprises statistical methods accounting for anticipated acoustic and linguistic variation of the input signal. A number of *model compensation* methods now exist whereby the standard features (e.g. MFCCs) are fed to HMM acoustic models whose parameters are altered prior to recognition (in an attempt to account for the distortion expected in the input signal), and a standard decoder then transcribes the reverberant utterances (Sehr et al., 2009). This family of techniques, including maximum a posteriori (MAP) (Gauvain and Lee, 1994) and maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995), are now widely used to increase robustness to background noise and unfamiliar talkers in addition to reverberation. Since these methods attempt to directly model

the reverberant observations, the mismatch between clean acoustic models and a reverberant input is reduced. However, the reported ASR gains have been somewhat limited in cases of reverberation since these methods do not account for the temporal evolution of this particular distortion, i.e. the dependency of the current state on the previous feature vectors is not exploited.

Methods which attempt to capture such *inter-frame relations* are now appearing in reverberation-robust ASR. Extending the model compensation techniques just mentioned, a similar approach is taken by Takiguchi et al. (2006) where the HMM adaptation parameters (i.e. the means and covariances) are updated frame-by-frame, varying in line with the recently experienced acoustic context. Incorporating time-varying reverberation estimations, the REMOS (REverberation MOdeling for robust Speech recognition) method of Sehr et al. (2010) uses a statistical reverberation model to describe a reverberant utterance as the convolution of a clean speech phrase with a room impulse response. Its main advantage therefore lies in the fact that changing acoustic conditions (e.g. a moving talker, or an unknown room) can be accommodated by updating the reverberation model without retraining the recognition engine itself.

2.2.5 Missing and uncertain data

A different class of techniques deal with reverberation by means of missing data techniques (Palomäki et al., 2004; Raj et al., 2004) and uncertainty decoding (Deng et al., 2005; Maas et al., 2013). Such techniques exploit knowledge of which parts of the signal contain reverberant energy (vs. which parts are relatively uncorrupted) and, at times, make use of the degree of confidence in the underlying estimation.

The missing data approach described by Raj et al. (2004) aims to reconstruct the original signal from a distorted one, i.e. to estimate an enhanced representation of the reverberant signal which is then passed forwards into the speech recogniser. Since the decoding process in the back-end is not affected in this method, large vocabulary continuous speech recognition can be undertaken following standard procedures without much additional computational cost. An alternative missing data approach described by Palomäki et al. (2004) uses 'bounded marginalisation' to deal with feature values that have been corrupted by reverberation. Rather than ignore the unreliable feature values, their true values are assumed instead to lie within a certain range (e.g., between zero and the observed (corrupted) value, for an auditory firing rate representation). During recognition, in a step which requires some modification to the usual back-end of the recogniser, the decoder can then compute likelihoods by integrating over the range of possible feature values.

Uncertainty decoding approaches have been implemented as both front-end and back-end methods for robust ASR (Deng et al., 2005; Liao and Gales, 2008). The principle trait of these methods is that, in addition to the feature vectors themselves being passed to the recogniser, a measure of the uncertainty of these observations (due to environmental conditions such as noise or reverberation) is also included. Following this approach, Maas et al. (2013) have recently extended the REMOS procedure (cf. § 2.2.4) with some success to additionally cope with a range of noisy environments.

2.2.6 Auditory inspired approaches

A final class of reverberation-robust signal-processing methods are those that are inspired in some way by biological audition (yet which are not overly concerned about the physiological resemblance). As mentioned above in § 2.2.1, bio-inspired approaches have brought gains to ASR, particularly in adverse, noisy situations, but such methods have proved of limited benefit in reverberation. To compensate for the effects of the late-arriving reflections, new methods are required which can account for the longer-term temporal evolution of the reverberant signal. As such, although a better understanding of how human listeners achieve perceptual compensation for reverberation might eventually impact on the encoding of reverberant sound in a number of machine listening applications, detailed nonlinear cochlear models (as discussed in the following chapter) are not expected to bring an immediate solution to the problem of reverberation-robust ASR. Nonetheless, useful applications may arise from considerations such as those examined in the model originally proposed in Lavandier and Culling (2010) and further developed in Jelfs et al. (2011), Lavandier et al. (2012) and Culling et al. (2013). Here, two binaural listening effects (better-ear listening and binaural unmasking, further discussed in § 2.3.1 below) are combined in order to generate 'intelligibility maps' of a room which indicate where to stand for optimal communication given the architectural constraints of the room.

In an attempt to account for longer-term temporal distortion that reverberation effects, Kingsbury (1998) used the modulation spectrum (which analyses frequency components of the signal's temporal envelopes) with the observation that phonetic identification in human listeners is governed largely by the components with lower modulation frequencies. Another perceptually-inspired (front-end) method, RASTA, attempts to make perceptually salient spectral transitions more obvious, while reducing sensitivity to irrelevant steady-state factors (Hermansky and Morgan, 1994). More recently, Petrick et al. (2008) used a 'temporal power envelope feature analysis' to uncover phone-level events that contribute most to intelligibility by looking at the modulation rate in the speech signal. In the same study, this

technique is compared with 'harmonicity based feature analysis' which assumes that the harmonic components of speech are *relatively* undistorted by reverberation, and which removes low-frequency energy present during unvoiced speech segments presuming it to be reverberant energy.

Though these methods bring some benefit for reverberant ASR, it is an oversimplification to assume that harmonic components of speech are undistorted by reverberation. Reverberation does distort the harmonic structure, and (for short reverberation times) a measure of the pitch strength is therefore inversely proportional to the reverberation present. This is shown by Wu and Wang (2003) who measure the pitch strength to determine $\mathrm{T60}$, based on a method that tracks the distribution of fundamental frequency ($F0$) values. The distribution of $F0$ is sharply peaked in dry conditions but spreads out with increasing reverberation time since the $F0$ of reflected sound may differ slightly from the direct $F0$. By examining the signal to reverberant component ratio, the linear prediction residual also provides a means for categorisation of dry and reverberant speech (Yegnanarayana et al., 1998). Another method, harmonic dereverberation (HERB), combines feature-based techniques with inverse filtering to design filters that remove the small deviations in pitch (and hence eliminate reverberant components) so that $F0$ can be made periodic in these local time regions (Nakatani et al., 2003).

The dependence of the perceptual effects of reverberation on the original signal content itself has additionally motivated modelling studies in a number of labs, since it validates the search for objective methods of speech perception which can work directly on reverberant signals without access to the corresponding room impulse responses. Van Dorp Schuitman et al. (2013), for instance, point out that since reverberation estimations based on RIR measures alone (cf. 2.1.4) do not depend on the signal content, they cannot adequately represent a listener's experience if, say, the room is occupied rather than empty, or if it is used for music listening rather than for speech. A related point is also exposed by a bio-inspired dereverberation study by Tsilfidis and Mourjopoulos (2011), in which an estimated clean signal may be synthesised from a reverberant signal at input[1]. The method was reported to be fairly successful for speech stimuli, but worked less well for musical signals (e.g., a continuous 'cello sound) in which the reverberation components were harder to identify and remove.

[1] Their method involved monitoring several independent cues (including an estimate of $\mathrm{T60}$ and a reconstruction of the clean signal) which together informed a mask locating portions of prominent vs. inaudible late-reverberation, and determined parameters of the audio resynthesis.

2.3 Human listening in real-room reverberation

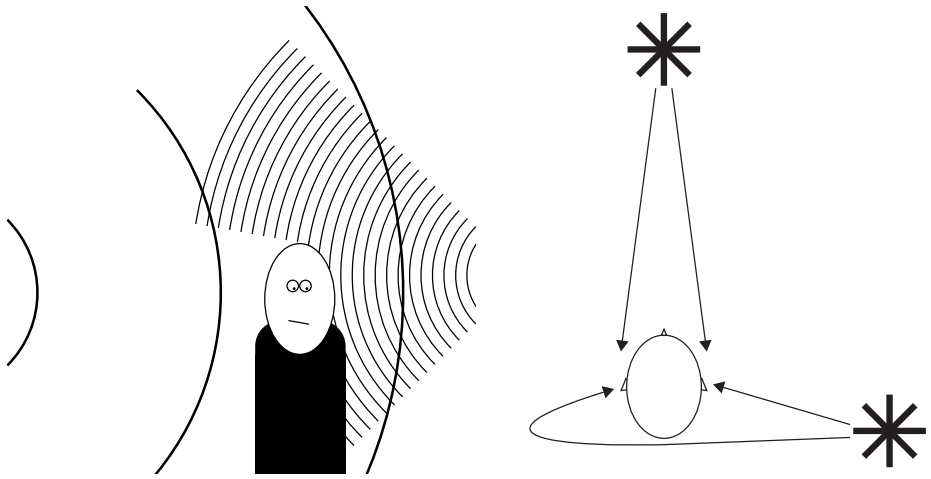
Reviewing psychoacoustic research into subjective effects of early and late room reflections, van Dorp Schuitman et al. (2013) report eight perceptually important attributes of room acoustics that appear regularly in the literature: reverberance, clarity, intimacy, apparent source width, listener envelopment, loudness, brilliance and warmth. From these descriptions of sound quality, it is clear that reverberation influences perceptual judgements both of *what* a sound is, and *where* it came from.

The majority of the thesis that follows is concerned with speech identification, however it must be noted from the outset that the 'where' and 'what' perceptual aspects of a sound source are clearly intertwined. The location of a sound (the 'where') is likely to affect its identification when heard in a reverberant room (and for a voice, its intelligibility). Conversely, the spectro-temporal content of a sound source (the 'what') also plays a role in its ability to be located, even in anechoic space. Studying composite sounds created with the envelope of one signal and the temporal fine structure of another, Smith et al. (2002) were able to pin down some of these aspects more closely. They report that properties of the fine structure are responsible for establishing the 'where' percept, while (agreeing with Shannon et al., 1995) properties of the envelope determine 'what' is heard[1].

In a reverberant room, reflected sound blurs the spectro-temporal features in the signal, making it harder for listeners to segregate and select target sounds in the presence of others (see e.g., Bronkhorst, 2000; Culling et al., 2003; Shinn-Cunningham et al., 2013). Yet reverberation is very common in our everyday listening environments, and we are often unaware of reflected sound arriving at our ears. The perceived position of a sound source is dominated by the first sound components to arrive (Wallach et al., 1949). This 'precedence effect' suggests that the direct signal path dominates perception in a reverberant room, which provides some robustness against room effects (Hermansky et al., 2013; Litovsky et al., 1999). Importantly for speech perception, the message heard in a speech signal usually remains semantically intact, regardless of whether the speaker and listener are close to each other or far apart (Watkins, 2005a).

This thesis discusses our ability to compensate for the effects of reverberation. The majority of the work undertaken below is focused on monaural speech identification tasks, however it must be noted that the binaural system clearly confers

[1] Moreover, by varying the number of filtered bands in which their stimuli were synthesised, and by placing cues in conflict with one another, Smith et al. (2002) found that as the number of bands increases above four, the envelope cues increasingly dominate perception. On the other hand, if just one or two bands are present in the signal then the fine structure cues are likely to be more influential.



(a) ILD, better ear effect **(b)** ITD, binaural unmasking

Figure 2.7: Binaural localisation cues: interaural level difference (ILD) and interaural time difference (ITD). In Figure 2.7a, a low-frequency sound, depicted with long-wavelength and travelling left-to-right, diffracts (bends) around the head and provides a similar sound level at each ear. A high frequency sound, depicted with short-wavelength and travelling right-to-left, cannot diffract around the head: a head shadow arises, and a level difference is observed between the two ears. In Figure 2.7b, a source directly ahead of the listener provides the same sound signal to both ears. When off to one side, however, the sound must travel further to reach one ear (either directly or by reflection) and a time delay is introduced between signals arriving at the ears.

significant benefit in certain listening situations. Intelligibility predictors based on monaural analyses will inevitably fail to account for the gains reported with binaural listening, especially in assisting the separation of target speech from stationary or modulated background noise (see e.g. Lavandier et al., 2012). Two significant binaural mechanisms are therefore discussed briefly below, regarding *timing* and *level* differences observed between the signals arriving at the two ears. The contribution of these binaural cues is well studied for localisation, however, their contribution to speech identification in reverberant conditions is less clear. Indeed, as discussion below reveals, a great deal of work remains to be done to understand binaural and monaural contributions to perceptual compensation for reverberation.

2.3.1 Binaural cues and monaural cues

Binaural cues are based on the comparison of the level and timing of signals received by the left and right ears, as depicted in Figure 2.7. The interaural level difference (ILD) cue described in Figure 2.7a is the first main binaural cue described. In the image, a low frequency sound can be seen travelling left-to-right across the page, with a wavelength sufficiently long as to pass around the listener

without trouble and present an approximately equal sound level to each ear. On the other hand, a high frequency sound, here travelling right-to-left, cannot diffract around the listener's head. This gives rise to an acoustic shadow at the far ear, and a large ILD between the ears. Particularly apparent for high frequency sounds, this cue has been linked to what is termed the 'better ear' effect. This is thought to be a process by which the brain can assess the two ears independently and *use* the ear providing the better SNR (Edmonds and Culling, 2006).

The binaural timing cue, the interaural time difference (ITD), arises from the fact that a sound from off to one side will take longer to reach the ear on the further side of the head (cf. Figure 2.7b). On the other hand, a sound source directly in front (or behind) of the listener will travel an identical path length to each ear, and will arrive simultaneously at each ear. This cue is linked to the concept of 'binaural unmasking' which relates to the suppression of signals arriving with a given ITD between the two ears, perhaps via the equalisation-cancellation model described by Durlach (1963). At frequencies below 1.5 kHz, ITDs are thought to arise from a comparison of the two signals themselves while at higher frequencies it was suggested that the ITD cue is conveyed by the envelope of the signal instead (Wightman and Kistler, 1992).

Taken together, ITD and ILD provide cues required to locate a sound source. The ITD cue provides a strong clue as to the azimuth of the source, i.e., its position left to right, particularly for low frequency sources such as the voice (Wightman and Kistler, 1992). Additionally, recent research suggests that processing of the ILD cue in the central auditory system may be involved with perception of distance of a source (Jones et al., 2013). However, these binaural cues are unable to provide any information in regard to the elevation of the source since both ITD and ILD will be zero for a source located ahead, irrespective of its height above or below the listener. For this, monaural cues (or head movements) are required.

In the presence of reverberation, these binaural cues become less reliable (Rakerd and Hartmann, 1985). Judgements of ITD are disrupted as the reflected energy increases the decorrelation between the two signals received at each ear. Additionally, ILD cues may alter unpredictably since they depend on the listener's exact position within the enclosure. Thus the spectral balance of the various sound components may vary dramatically depending on the location of the source and listener relative to the room modes and in proximity to particularly reflective or absorbent surfaces. Interestingly, distance perception has been reported to improve in reverberant conditions, while direction accuracy worsens (Shinn-Cunningham, 2000). The latter finding may be explained by the fact that lateralisation is cued largely by the temporal fine structure of the ITD which is decorrelated in the presence of reverberation (Smith et al., 2002). Indeed, Devore and Delgutte (2010) suggest

that ILDs may provide more reliable directional information than envelope ITDs for localising high frequency sounds in reverberant conditions.

Monaural localisation cues exist in the form of directional filtering that results from the interaction of the sound source with the various parts of the human anatomy. The pinnae in particular provides a spectral cue for elevation, as the original source sound is modified with a different frequency profile depending on 'where' it originates: level with, above or below the ear (Pickles, 1988). Secondly, the pitch percept, or its physical correlate, $F0$, might be thought of as a monaural cue assisting with 'what' is heard. Maintained throughout the various stages of auditory processing, differences of pitch assist listeners in segregating concurrent sounds, and similar pitch contours assist the grouping sounds originating from a single source (Darwin, 1984). However, Culling et al. (2003) report that reverberation additionally hinders listeners' ability to capitalise on $F0$ cues. Their experiments showed that monotonised speech (in which there is no $F0$ modulation, and thus reduced prosodic information) was harder for listeners to segregate when stimuli were reverberated rather than presented in anechoic conditions.

Integration of cues

As stated earlier, listeners are remarkably robust to the effects of real-room reverberation. In part, this is likely to be due to the fact that listeners are typically active participants in their acoustic environments. By moving, tilting and rotating the head, a great deal of ambiguity in where the signal originated can be resolved. For example, by tilting the head, an elevation cue may be transformed into an azimuth cue which is easier to resolve with a greater degree of accuracy. The ways in which such cues are swapped or integrated has not yet been well-studied, however, since the bulk of the relevant research to date has used headphone presentation which does not allow this kind of cue-swapping to occur.

Moreover, it is not yet understood how monaural and binaural aspects of hearing combine together, or take-over from one another, particularly in reverberant listening tasks. The two binaural cues discussed earlier are thought to contribute to the phenomena termed spatial release from masking (SRM), combining better-ear listening and binaural unmasking (Lavandier and Culling, 2010; Zurek, 1993). While having two ears is clearly a benefit as it allows SRM, it is still not clear whether that benefit arises from switching between the two monaural systems (left then right individually, using the better ear effect which favours the ear with the higher SNR) or from actually synthesising concurrent information from the two ears simultaneously (as occurs in binaural unmasking arising from the ITDs). Using modulated noise maskers with simulated room reverberation, ongoing research in this area aims to better understand the relative importance of these contributions

in the context of speech intelligibility (see e.g., Culling and Mansell, 2013; Weller et al., 2014).

Though questioned recently by Jones et al. (2013), the duplex theory has stood for around a hundred years, suggesting that localisation relies on time-based cues for low frequency sounds, and level-based cues for the high frequency sounds. Clearly, to achieve good speech identification a listener must also integrate information across frequency and through time. Culling et al. (2006) examined contributions of monaural and binaural cues, separately and in combination, on speech perception. In their study, the binaurally-derived information was helpful only in the low-frequency regions (below around 1.2 kHz) whereas monaurally-derived cues were beneficial throughout the entire frequency region. In other words, binaural information helped (alongside monaural cues) to uncover the pitch and formant structure of the speech in this task, whereas the high-frequency consonant identification was achieved by (only) the monaural system.

2.3.2 Categorical perception

When listening to speech, individual sounds are perceived according to categories that remain stable despite variation introduced by everyday listening environments. However, Klatt (1982) points out the trap of assuming equivalency between phonetic- and psychophysical-judgements. Here, speech perception exhibits an important flexibility, namely that these categories are adjusted continually to take changing contexts into account (Remez, 1996; Tuller et al., 1994).

The premise that a continually variable acoustic signal can be partitioned into discrete perceptual categories is of fundamental importance to speech production and recognition, and essentially allows the distinction of one word from another. There is an obvious and significant acoustical dependence on body size, speech rate, accent and so on, yet we will instinctively classify any given speech sound either as or *not* as a particular phoneme. Since speech sounds are strongly categorical in nature, category-boundary experiments have proved useful tools in speech perception research, and a number of methods have been devised to experimentally determine the 'phoneme boundary' which marks the point on the continuum where the percept changes from one speech sound to another (see e.g., Ganong and Zatorre, 1979).

The 'perceptual magnet' theory of Kuhl (1991) highlights the fact that we are more likely to notice differences *between* rather than *within* categories. This theory describes a warping of perceptual space, illustrated in Figure 2.8, where a speech sound that is close to a category 'prototype' (or unambiguous speech sound) is pulled toward the prototype itself. The perceptual magnet effect has been framed

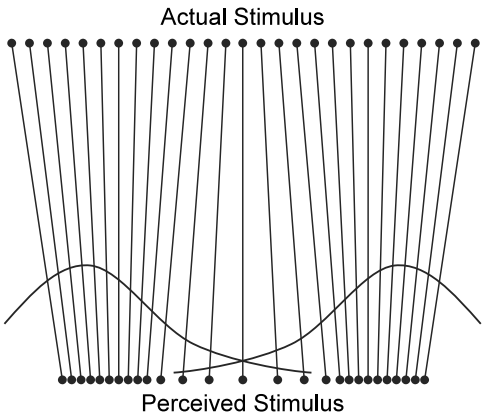


Figure 2.8: Predicted relationship between acoustic and perceptual space for two categories: perceptual space is shrunk in the neighbourhood of unambiguous speech sound categorisation, however shrinkage is weakest at category borders. Figure redrawn from Feldman and Griffiths (2007).

in Bayesian terms by Feldman and Griffiths (2007) so that an incoming speech sound is judged in relation to a phonetic category's mean and variance.

Ladefoged and Broadbent (1957) showed that identification of a test-word also depends strongly on the context in which it is observed (specifically that the perception of a test-word as *bit*, *bet*, *bat*, or *but* depends on the formant structure of the introductory sentence). The category boundary experimental paradigm has thus proved particularly useful to explore the dependencies of speech perception on a range of contextual effects. Indeed, this method underpins the experiments of Watkins (2005a) which examine perceptual compensation for reverberation.

2.3.3 Misclassification in reverberant speech

Over the past 40 or so years, studies have consistently demonstrated that reverberation degrades the intelligibility of speech. Some of these works are summarised in Table 2.2. The results are, in the main, consistent, allowing some generalisations to be drawn as follows.

Reverberation consistently degrades speech intelligibility due to the presence of real-room reflections which mask the direct sound. Even mild reverberation is sufficient to disturb the identification of test-words, such that word identification scores for both normal hearing and hearing impaired listeners are altered from the non-reverberant case (e.g., Nábělek and Robinette, 1978). A binaural advantage was reported for cases in which reverberation and noise were heard jointly (Helfer, 1994; Nábělek and Robinette, 1978), but was typically rather limited in otherwise quiet conditions (ranging from 3.8 to 6.7% relative improvement in Nábělek and

Table 2.2: Overview of reverberant speech studies, presented alphabetically. Abbreviations used in table: *Listeners* – number of participants with normal hearing (n) or some hearing loss (h). *Material* – identification of consonants (c), vowels (v), words (w) or nonsense (n) speech material. *Conditions* – quiet (q), modulation filtered (m), noisy (n), reverberant (r) or vocoded (v) material. *Presentation* – monaural (m), binaural (b) or diotic (d) stimuli.

Study	Listeners	Material	Conditions	Presentation
Cox et al. (1987)	40_n	w	n, r	m
Drullman et al. (1994a)	54_n	c, v	m	m
Drullman et al. (1994b)	60_n	c, v	m	m
Gelfand and Silman (1979)	20_n	c	q, r	m
George et al. (2008)	10_n	w	n, r	m
Harris and Swenson (1990)	$10_n, 20_h$	w	n, r	b
Helfer and Wilber (1990)	$16_n, 16_h$	n	q, n, r	b
Helfer and Huntley (1991)	$16_n, 8_h$	n	q, n, r	b
Helfer (1994)	13_n	n	q, n, r	m, b, d
Nábělek and Pickett (1974)	5_n	c	n, r	m, b
Nábělek and Robinette (1978)	$10_n, 12_h$	w	n, r	m, b
Nábělek and Robinson (1982)	60_n	w	q, r	m, b
Nábělek and Dagenais (1986)	10_h	v	n, r	m
Nábělek et al. (1989)	20_n	c	q, n, r	m
Nábělek et al. (1992)	$20_n, 20_h$	v	q, n, r	m
Nábělek et al. (1994)	$10_n, 10_h$	v	q, n, r	m
Nábělek et al. (1996)	$10_n, 7_h$	v	q, n, r	m
Poissant et al. (2006)	26_n	w	q, n, r, v	m

Robinson, 1982, where the largest scores were associated with the elderly listening group who had in any case begun from the worst baseline performance).

The position of a phoneme within a word or phrase also plays a role in its intelligibility under reverberation. This has been explained by Nábělek et al. (1989), citing earlier work by Bolt and MacDonald (1949), as the effect of two distinct types of masking. *Self-masking* refers to the blurring of energy internally within a particular phoneme, and *overlap-masking* occurs when the energy of a preceding phoneme spreads in time and extends into the following phoneme. As a result, items tested in word-initial position generally contained less reverberation than those in the word-final position, and had a correspondingly higher chance of being correctly identified (e.g., Helfer and Huntley, 1991; Nábělek et al., 1989).

Reverberation does not degrade different speech sounds equally, but misclassification among items is somewhat predictable as these tend to be within speech classes,

i.e., a stop consonant will most likely be mistaken for another stop consonant rather than for, say, a fricative. Certain sounds (e.g., sibilants) appear naturally more resistant to the effects of reverberation (Gelfand and Silman, 1979); additionally, vowels are mistaken less often than consonants (Drullman et al., 1994a, b). Reverberation typically introduces more errors involving place of articulation rather than in regard to manner or voicing (Cox et al., 1987; Drullman et al., 1994b; Gelfand and Silman, 1979; Nábělek and Pickett, 1974). Stop consonants, particularly the unvoiced plosives, are almost always reported to be the most vulnerable group of speech sounds (Drullman et al., 1994b; Gelfand and Silman, 1979; Nábělek and Pickett, 1974).

Perhaps driven by the observation that multiple reflections reach the listener in close succession and therefore act as a speech-shaped noise masker (Gelfand and Silman, 1979), the majority of these studies looked jointly at the effects of reverberation and noise (cf. the fourth column of Table 2.2). However, reverberation is a convolutional rather than additive distortion, and we would not therefore expect it to cause the same effects on speech perception. Moreover, the variety of noise types (including stationary, speech-shaped, /s/-shaped, babble and cafeteria noise) and the diversity of listening tasks (identifying consonants, vowels, words, sentences or nonsense syllables) resulted in less consistency across the noise-based studies than arose for studies examining reverberation alone.

2.3.4 Effect of reverberation on stop consonants

As mentioned above, the speech sounds most likely to be misidentified in reverberation are the stop consonants, and in particular the unvoiced plosives (Cox et al., 1987; Drullman et al., 1994b; Gelfand and Silman, 1979; Nábělek and Pickett, 1974). These sounds are produced when the lips or tongue cause a restriction of the airway, producing the bilabial [p], alveolar [t] or palatal [k] whose release bursts were reported by Allen and Li (2009) to be in the low frequency regions c. 0.7–1 kHz for [p], in the high frequency region around 4 kHz for [t], and in the mid frequency region around 1.4–2 kHz for [k].

In the temporal domain, however, the unvoiced plosives are all defined by a dip occurring in the amplitude envelope, i.e. a brief silence or period of low energy. This characteristic results in their particular susceptibility to the effects of reverberation (as was seen earlier in Figure 2.2), since the dip in the temporal envelope which helps to cue their identity may easily become obscured in the presence of reverberation.

None of the studies in Table 2.2 has yet measured the influence of reverberation on continuous speech, thus the wider relevance of the difficulty in identifying rever-

berant stops in everyday listening must be inferred from the frequency of their occurrence. In an American English conversational speech dataset studied by Mines et al. (1978), the unvoiced plosives [p], [t] and [k] together accounted for around 10.7% of all phonemes (or about 18.2% of all consonants) encountered. Thus the relative abundance of these consonants in English may contribute substantially to the difficulties anecdotally reported by many listeners in reverberant environments.

Interestingly, Nábělek et al. (1989) reported that unvoiced stop consonants were even more vulnerable to reverberation when presented after an [s] sound than when they were presented alone. Again, this finding would appear to remain relevant in more conversational settings, since the [s] occurred at a rate comprising 4.6% of phonemes encountered in Mines et al. (1978). Further, occurrences of [s] being followed by [p], [t] or [k] were examined in the dataset used in Lanchantin et al. (2013). Of all occurrences of [s] that were not in the word-final position[1], 44% of the remaining [s] occurrences were followed by an unvoiced plosive. Among these phoneme pairs, [st] accounts for 77.4% of the data observed, with [sp] and [sk] appearing on 12.1% and 10.4% of occasions, respectively. This analysis suggests that the three pairs of consonants most likely to be lost in reverberation, [sp], [st] and [sk], might account for just over 2% of all sounds heard in typical English conversation.

2.3.5 Perceptual constancy in vision and audition

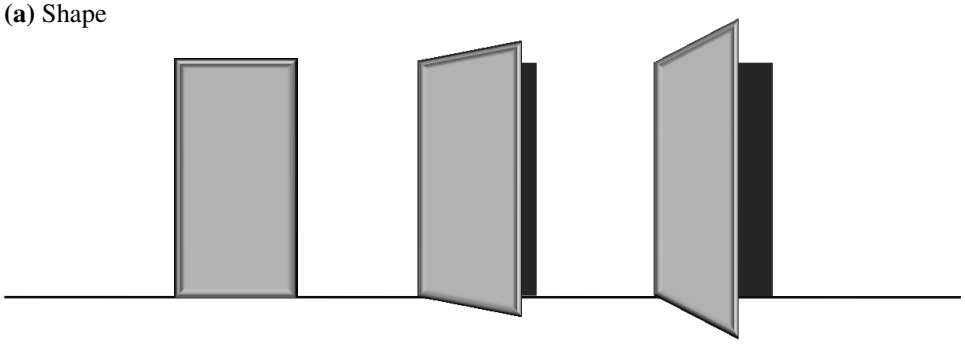
Perceptual constancy describes our ability to recognise an object or quality as fundamentally unchanged despite it appearing to be different due to the circumstances in which it is encountered. Perceptual constancy is increasingly important in auditory research owing to the growing desire for machine listening systems to work robustly in natural environments. Motivated by a wide variety of practical applications and theoretical concerns, researchers have investigated perceptual constancy for varied aspects of sound[2]. Before restricting the discussion tightly to the topic of reverberation it is helpful to briefly consider some other examples of perceptual constancy both in speech perception and more widely.

Perceptual constancy has received greater attention in vision than in hearing, examining constancy for lightness, colour, shape and size among other attributes

[1] Pronunciation probability statistics were obtained using a dictionary file rather than a transcription of spoken utterances, thus statistics crossing word-boundaries were not available. Thanks to Oscar Saz for providing these numbers.

[2] See, for example, Ardoint et al. (2008); Gockel and Colonius (1997); Johnson and Strange (1982); Kuhl (1979); Li and Pastore (1995); Stecker and Hafter (2000); Summerfield (1981); Zahorik and Wightman (2001)

(a) Shape



(b) Focal point

Figure 2.9: Visual constancy. Figure 2.9a depicts shape constancy: the opening door projects different shapes visually, but is still understood to be rectangular. In Figure 2.9b the focal point depends on the stimulus surrounds (Figure from Webster et al., 2002). The middle row is repeated left and right. On the left-hand side it is flanked by blurred bars and appears sharp in comparison. On the right-hand side it is flanked by sharpened bars and thus appears correspondingly blurred.

(Adelson, 2000). Shape constancy is depicted in Figure 2.9a and shows that the perception we have of an object may be quite different from the physical properties that have brought about that perception. Explanations of visual constancy differ among vision researchers, however Yang and Purves (2004) have recently explained a wide range of phenomena (including optical illusions) by describing brightness with a probability distribution function that records particular luminance values within local and global surroundings. In this way they suggest that the perception of a target will depend on its context.

We can also approach perceptual compensation for reverberation with reference to a visual analogy: image blur. Work by Webster et al. (2002) has examined neural adjustments to image blur, showing that humans label an image as 'blurry' or 'sharp' depending on their recent exposure to blurry or sharp scenes. If more

blurry pictures are shown, then the viewer becomes accustomed to this new blurry world so that a non-blurry image will be interpreted as 'too sharp'. Similarly in an overly sharpened world, a normal-focus image will be labelled as 'too blurry'. Webster et al. propose that visual responses compensate for spatial sensitivities by a process of continual recalibration arising from cortical adaptation, and allow constancy for image structure to be achieved by renormalizing the perceived point of best focus as shown in Figure 2.9b.

For speech perception, it is clear that vowels and consonants are perceived as constant categories across a wide range of acoustic conditions, for example with differing talkers, unknown accents, varied rates of speech (Summerfield, 1981), and a wide range of listening environments (Cox et al., 1987). A perceptual normalisation process seems to exist whereby listeners disregard the variation of specific instances and arrive at a general underlying category. For example, perceptual constancy for vowels has been studied in relation to rapid speech (Johnson and Strange, 1982) and coarticulation (Strange et al., 1983). The former paper suggests that a vocal context surrounding a test-word makes it easier to identify. The latter paper examines various speakers, speech rates and surrounding consonants using a stimulus composed of /b/-vowel-/b/. These authors report that dynamic spectral information from the transitions from and to the consonant is enough to identify the vowel, even when the centre of the vowel itself is attenuated to silence. The idea that transient attack portions of sound (i.e. the time-varying parts) are important is also supported by Ardoint et al. (2008). They find that perceptual constancy exists for at least partial (or incomplete) normalization of temporal-envelope patterns when compressed or expanded in the time domain, thereby suggesting an explanation for why speech perception is robust under variation in presentation rate.

2.4 Compensation for reverberation in human listeners

Earlier sections of this thesis showed that when speech occurs in rooms, the direct sound is amalgamated at the ear with many time-delayed and attenuated reflections from the room's surfaces. From an acoustical point of view, the effect of reverberation was shown to reduce the modulation depth of the speech envelope and to adversely affect its intelligibility (cf. Figure 2.2). Regardless of these facts, speech perception in rooms remains remarkably robust under diverse reverberation conditions. The auditory mechanisms underpinning this robustness, and permitting compensation for the effects of reverberation on the speech signal, are attracting a lot of attention in research today. A series of recent experiments (see for example: Brandewie and Zahorik, 2010, 2012, 2013; Longworth-Reed et al., 2009; Srinivasan and Zahorik, 2013; Ueno et al., 2005; Watkins, 2005a, b; Watkins et al.,

2011) suggests that the auditory system achieves *perceptual constancy* in reverberation, as was seen above for visual properties of surfaces such as their shape and focal point.

A substantial body of evidence now shows that the perception of a reverberant sound is influenced by the properties of its temporal context (e.g., Brandewie and Zahorik, 2013; Longworth-Reed et al., 2009; Watkins, 2005b). In these experiments, listener performance improves in a range of speech identification tasks when the listeners hear consistent room cues just prior to the test stimulus. Together, these studies suggest a conceptual model of compensation for reverberation in which listeners glean information from the acoustic context preceding a test-sound, and accumulate this information over a period of time in order to achieve perceptual constancy. Much work is yet needed, however, to understand both the nature of the information that is used, and the time course over which it is gathered.

As in the reverberant speech perception studies above, research into compensation for the effects of reverberation has also employed speech stimuli presented in (spatialized) noise. Zahorik and colleagues have demonstrated binaural compensation effects using the Coordinate Response Measure database in Brandewie and Zahorik (2010), the Modified Rhyme Test in Brandewie and Zahorik (2012), and from the PRESTO subset of TIMIT sentences in Srinivasan and Zahorik (2013). Additionally, Zahorik and colleagues have begun to probe the mechanisms that may account for these effects, and have reported that prior binaural exposure in a particular room condition improves listeners' ability to detect amplitude modulation in that room (Zahorik and Anderson, 2013; Zahorik et al., 2012).

In addition to this binaural compensation effect, a large body of psychophysical data gathered in a phoneme-continuum identification task supports the idea of a monaural compensation mechanism (e.g., Watkins, 2005a, b; Watkins and Makin, 2007a, b, c; Watkins et al., 2011). These experiments suggest that the monaural mechanisms are informed primarily by the temporal envelope of the signal (which may or may not be speech-like). This proposition is broadly in line with findings of a recent neural coding study by Kuwada et al. (2012) which observed neurons that have a higher modulation gain in reverberant conditions relative to anechoic conditions, and which might therefore help to counteract the deleterious effects of reverberation on modulation depth.

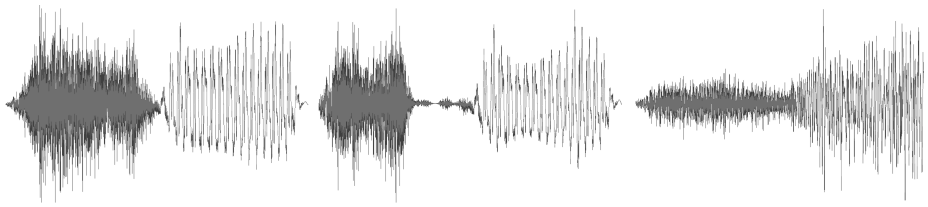
Watkins' work into monaural constancy effects has inspired much of the work for this thesis and is therefore described in greater detail below. However, it is important to note that monaural compensation has not always been apparent in the published literature. In Brandewie and Zahorik (2010), for instance, only two of fourteen participants were reported to derive an appreciable benefit from monaural room exposure. There are several reasons why this performance difference may

arise, in particular the fact that the listeners' speech identification task in their study might conflate aspects of localisation and spatial unmasking with the compensation effect being studied (since listeners were required to identify reverberant speech while a masking noise was presented off to the side in binaurally simulated room reverberation). An alternative possibility is that monaural effects may be limited to phoneme-continuum type experiments (where only a small number of response tokens are available) and thus would not occur in experiments during which the speech stimuli differ substantially from trial to trial. Further, a third possibly remains that the 'morphing' effects observed in Watkins' phoneme-continuum identification may not arise with naturalistic speech, and instead an even spread of errors across the various responses alternatives might arise (Phatak et al., 2008).

The relevance of a monaural compensation effect to natural speech perception tasks thus remains to be demonstrated: this is the subject of the first perceptual experiment reported below (Experiment H1). Moreover, the factors *contributing to* or *detracting from* the monaural compensation effect have also yet to be elucidated with natural speech.

Watkins' work suggests that compensation effects can arise even when the context speech cannot be understood, e.g. for cases in which the context speech is time-reversed (Watkins, 2005a). Indeed, provided there are appropriate temporal modulations across a sufficiently broad range of the spectrum, compensation can arise even when the context is not a speech signal at all (e.g. Watkins and Makin, 2007a, b, c, which all study compensation arising from contexts comprised of modulated noise bands). However, Srinivasan and Zahorik (2011) warn that semantic expectations arising from a previous context sentence influence perception of subsequent reverberant test words to such a degree that they can override such compensation effects. Another recent study by Longworth-Reed et al. (2009) appears to support the earlier finding of Watkins (2005a) that monaural compensation is blocked by a reversal of the time-direction of *reverberation*. Thus the time-direction of reverberation appears critical for it to be successfully integrated by the auditory system and turned into 'useful information' contributing to the beneficial compensation effect. Experiment H2 below investigates whether the previous findings of Watkins (2005a) and Longworth-Reed et al. (2009) persist when the time-reversals (of the speech direction and of the reverberation direction) are examined monaurally with naturally spoken material.

Importantly, the time course of the monaural compensation effect has yet to be investigated. However, a number of studies with varying listener tasks have recently queried the timescales on which *binaural* compensation effects are apparent. At the 'slow' end of the scale, long-term learning of a particular room condition (c. 5 hours) appears to improve localisation accuracy (Shinn-Cunningham, 2000). Contrastingly, just seconds of inconsistent reverberation on the preceding context was



(a) 'sir' **(b)** 'stir' **(c)** reverberant 'stir'

Figure 2.10: Waveform display showing the temporal smearing effected by reverberation. The plosive closure of the [t] sound in an unreverberated utterance of 'stir' (Figure 2.10b) is filled with reflected energy as reverberation increases (in Figure 2.10c). This makes it appear more similar to a 'sir' utterance (Figure 2.10a).

sufficient to disrupt listeners' ability to determine the azimuth of a test pulse (Zahorik et al., 2009). For binaurally-presented speech identification tasks, a benefit of prior room experience has typically been reported at the minimum timescale permitted by the analysis in use: measured in minutes for the sentence sets in Longworth-Reed et al. (2009); occurring within six sentences in Srinivasan and Zahorik (2013); and within a few seconds for the two-sentence carriers used in Brandewie and Zahorik (2010). Recently, Brandewie and Zahorik (2013) designed a study specifically to measure the time course of the binaural effect, and reported that around 850 ms of room exposure was sufficient to achieve considerable speech intelligibility enhancement. Interestingly, this study also reported that the compensation mechanism appeared to act more slowly at higher noise levels: that is, it may take longer for a listener to adapt to a noisy room. Experiment H3 below queries the temporal extent of the signal region contributing to monaural identification of the test-word itself, and Experiment H4 examines the influence of the extent of matching context reverberation on this judgement.

2.4.1 Watkins' sir-stir paradigm

Watkins and colleagues have demonstrated perceptual compensation for the effects of reverberation using a paradigm in which the preceding context of a speech sound influences its identity (see for example: Watkins, 2005a, b; Watkins and Makin, 2007a, b, c; Watkins and Raimond, 2013; Watkins et al., 2010b, 2011). Since these 'sir-stir' continuum experiments form the basis of the auditory modelling study in Chapter 4 and, in addition, greatly informed the design of the perceptual listening experiments in Chapter 5, their general scheme is described here in some detail.

In Watkins' experiments, listeners identify 'sir' and 'stir' test-words [TEST] embedded in a fixed context phrase ("OK, next you'll get [TEST] to click on"), while reverberation conditions of the context and test-word portions of the signal are var-

ied. The identification task exploits the vulnerability of the stop consonant [t] in the context of [s] as was discussed above (cf. § 2.3.4), and is highly sensitive to the way that reverberation tends to 'morph' one sound into another[1]. This can be understood in relation to Figure 2.10, which depicts the temporal waveform for unreverberated utterances of 'sir' and 'stir', and the effects of reverberation on the word 'stir'. Figure 2.10b shows a token test-word 'stir', evidenced by the dip in the amplitude envelope that characterises the 't' closure in the non-reverberant situation. The reverberant decay adds energy into this dip in Figure 2.10c and reduces the dynamic range of the signal which then appears more similar to a typical 'sir' utterance as shown in Figure 2.10a.

Test-words are drawn from a synthetic continuum of 11 steps that was originally created by interpolating between the temporal envelopes of naturally spoken tokens of 'sir' and 'stir' (Watkins, 2005a). Here, the amplitude modulation of the [t] was gradually introduced across the continuum items so that a sample drawn from one end of the continuum gave the percept of 'sir', and a sample from the other end of the continuum gave the percept of 'stir'. However, due to the categorical nature with which humans perceive speech (cf. § 2.3.2), samples drawn from intermediate continuum steps do not appear to listeners to be 'intermediate' or ambiguous in any way. Rather, they are immediately and unthinkingly categorised as either 'sir' or 'stir'. The category boundary – that is, the step in the continuum at which the percept on average switched from 'sir' to 'stir' – is then recorded at a given experimental condition by counting the number of 'sir' responses achieved across all items in the continuum and subtracting 0.5. Across the 11 steps of the continuum, this results in a category boundary which may range from -0.5 (no 'sir' responses) to 10.5 (all 'sir' responses).

The key features of Watkins' paradigm are summarised in Figure 2.11 where the vertical axis displays the category boundary measured across the 11 continuum steps (here labelled 0 to 10 inclusive). Listeners identify the continuum test-words when they are embedded in a fixed context phrase. Context distance and test-word distance are independently varied, simulating voices being heard from different positions in a room. In any experiment, the typical set of stimuli heard then contains two same-distance conditions (with near-near and far-far *context-test* distances) and two mixed-distance conditions (near-far and far-near).

[1] A similar 'morphing' process is described by Phatak et al. (2008) for the types of confusion made in the presence of noise in which a spectro-temporal region critical to the recognition of a particular consonant feature may be heavily masked. In such cases, rather than an even spread of responses among all possible consonants, the obfuscated test-item promotes the majority of responses in one particular category; thus its identity can be said to have 'morphed' in the presence of the noise masker.

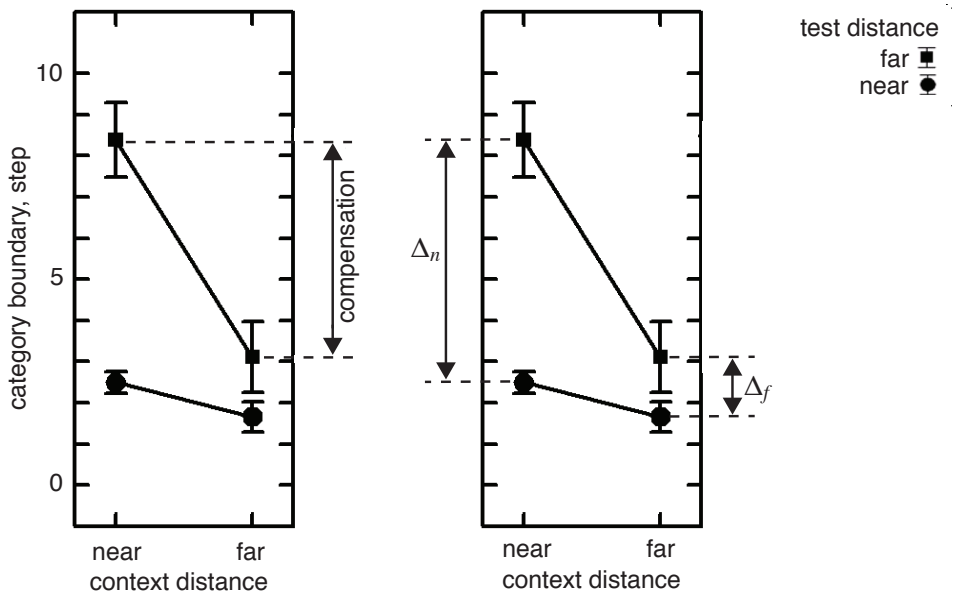


Figure 2.11: Two methods for measuring compensation for reverberation in Watkins' 'sir-stir' paradigm. Listener data (which is the same in each panel) reveals that when the test-word alone is far-reverberated, the category boundary shifts upwards due to the increase in 'sir' responses. When the context and test-word are both far-reverberated, more steps are again reported to be 'stir' and the category boundary shifts downwards again. In the left-hand panel, compensation is defined with regard to far-distance test-word alone, as the category boundary recovery that a far-distance context introduces (Watkins and Makin, 2007c). The right-hand panel presents a more widely-used strategy (e.g. Watkins and Makin, 2007a), where compensation is quantified by first measuring the difference in category boundary due to test-word reverberation (i.e. the vertical distance observed at each level of the context (marked Δ_n and Δ_f for the near and far distance contexts respectively), and then by calculating the difference (recovery) that arose due to the context condition, so that compensation $= \Delta_n - \Delta_f$.

For the near-near continuum stimuli, in which the sound is heard to come from a consistently nearby voice throughout, a relatively low category boundary is usually recorded (c. 2.5, for the bottom-left data-point in either panel in Figure 2.11).

In the near-far stimulus condition, the level of reverberation present during the test-word is now much higher than that during the preceding context and listeners typically respond 'sir' to more steps of the continuum. The reverberation has, in effect, undone the work of the amplitude modulation that was introduced to create the continuum, as though the increase in reverberant energy has concealed the dip in the temporal envelope that previously cued the [t] consonant. The category boundary shifts upwards as a result of the effect of reverberation on the test-word (to c. 8.5, for the top-left data-point in Figure 2.11).

For the far-far continuum stimuli, the level of reverberation on the context is increased to match that of the far-reverberated test-word. Here, the reverberation which had earlier (at near-far) seemed to obscure the [t] in the far-distance test-word is still present. Moreover, at far-far there is a *further* increase in the overall amount of reverberation present during the portion of the signal containing the test-word, since overlap masking additionally results in reverberant energy from the context being prolonged into the test region. Given that an increased level of reverberation has repeatedly been shown to degrade intelligibility in the signal, one might anticipate that this increase in reverberant energy would further smooth signal modulation, increase the number of 'sir' responses and raise the category boundary still higher. However, exactly the opposite pattern of results is seen in the data: at far-far, more of the continuum steps are again identified as 'stir' and the category boundary is partially restored towards its original position (to c. 3, for the bottom-right data point in the same panel). Thus the category boundary movement here displays the net-result of a compensation effect which far outweighs the detrimental effect of overlap-masking.

This pattern of listener responses has been replicated throughout Watkins' studies, leading him to conclude that listeners routinely use information about the temporal envelope of surrounding speech to compensate for the effects of reverberation on a particular word. This has usually been validated by means of a statistically significant interaction between two factors for context-distance and test-distance (each with two levels) in a repeated-measures analysis of variance (ANOVA). This analysis emphasizes that the effect of one factor alters depending on the level of the other. That is, the far-distance test-word brings about a large increase in the category boundary, but only for near-distance contexts. Alternatively, the far-distance context brings about a large reduction in the category boundary, but only for far-distance test-words. In other words, the degrading effect of test-word reverberation is greatly reduced at far-distance context conditions.

The matter of how to quantify the compensation effect numerically has not been so straightforward to settle. Figure 2.11 shows two methods which have been used. The left-most panel redraws the method used in Watkins and Makin (2007c), where the near-distance test-words are effectively neglected. Here, compensation is calculated solely from the recovery of the category boundary that the far-distance context brings about for the far-distance test-words.

An alternative method to quantify compensation, shown in the right-hand panel of Figure 2.11, was first described in Watkins and Makin (2007a) and has since been reused extensively to compare different experimental conditions including vocoded noise-band contexts in Watkins et al. (2011) and silent contexts in Watkins and Raimond (2013). Here, the effect of reverberation on the test-word itself is quantified first by measuring the vertical difference in category boundaries at each

context condition individually. In Figure 2.11 these quantities are marked Δ_n and Δ_f for the near and far distance contexts respectively. In this formulation, the magnitude of the compensation effect is subsequently computed by calculating the category boundary recovery that arises due to the change in context condition, i.e. by calculating the difference in the test-word reverberation effect between the two context conditions. Thus, compensation $= \Delta_n - \Delta_f$.

2.4.2 Alternative measures of compensation for reverberation

Watkins' category boundary paradigm provides an elegant method to observe the effects of compensation for reverberation, however his experiments are restricted in the sense that they observe listeners' responses to only two speech categories for stimuli arranged along a phoneme-continuum. In recent years there has been a move to investigate compensation for the effects of reverberation on speech identification using more diverse speech material. The major findings of these studies were discussed above. This section instead surveys the various experimental methods, summarised in Table 2.3, which have been employed to investigate compensation.

The first two studies presented in this table (*top*) were directly influenced by Watkins' work, and asked listeners to identify test-words differentiated by a single consonant. In the paper by Ueno et al. (2005), stimuli were arranged similarly to the 'sir-stir' paradigm so that the preceding speech context and subsequent test-word could be processed with independent reverberation conditions (here with two BRIRs recorded in different rooms). Listeners identified the test-word's consonant in two experiments (selecting between 16 alternatives in their first experiment, and 6 alternatives in their second), and performance was measured in terms of the percentage of correct identifications that the listener made. Secondly, using the two-alternative forced-choice (2AFC) 'sir-stir' identification task, Nielsen and Dau (2010) attempted to replicate the findings of Watkins (2005a). Methods (and hence, results) in both of these studies were somewhat unsatisfactory[1], but nevertheless did hint overall that consistent exposure to the room reverberation condition would improve listeners' consonant identification.

[1] These two studies were limited by a number of practical details. Ueno et al. (2005) assumed that the 'silent' context would act as a control condition, however this is seen below (in Experiment H3) not to be the case. The same assumption was made by Nielsen and Dau (2010), who may have had additional experimental confounds arising from the fact that the test-word's distance did not vary unpredictably from trial to trial in their second experiment but was instead held fixed at the far room distance throughout.

Table 2.3: Studies examining compensation for the effects reverberation on speech identification, excluding those of Watkins (2005a). The table is split into three portions, each of which is presented chronologically. *Top:* One-off studies inspired by Watkins' work. *Middle:* studies by Zahorik and colleagues. *Bottom:* Experiments reported in Chapter 5 of this thesis. Abbreviations: *Listeners* – number of participants with normal hearing (n) or some hearing loss (h). *Material* – identification of consonants (c), words (w) or sentences (s). *Conditions* – anechoic (a), low-pass filtered (f), background noise simultaneous with test item (n_s), noise context preceding test item (n_c), reverberation (r), silent context (s), time-reversed reverberation (t), vocoded (v). *Stimuli presentation* – monaural (m), binaural (b) or diotic (d) stimuli. *Task* – 2AFC (2), 4AFC (4), 6AFC (6), consonant (c), number/colour (n), last word in sentence (l), or as many words as possible in sentence (w).

Study	Listeners	Material	Conditions	Stimuli	Task
Ueno et al. (2005)	25_n	c	a, r, s	b	c
Nielsen and Dau (2010)	19_n	c	n_c, r, s	d	2
Longworth-Reed et al. (2009)	10_n	s	r, t	b, d	w
Brandewie and Zahorik (2010)	14_n	w	n_s, r	b, m	n
Srinivasan and Zahorik (2011)	21_n	w	n_s, r	b	l
Zahorik and Brandewie (2011)	$14_n, 12_h$	w	n_s, r	b	n
Brandewie and Zahorik (2012)	14_n	w	a, n_s, r, s	b	6
Srinivasan and Zahorik (2013)	60_n	s	n_s, r	b	w
Brandewie and Zahorik (2013)	16_n	w	a, n_s, r, s	b	n
Srinivasan and Zahorik (2014)	30_n	w	a, r, v	b	w
Experiment H1	60_n	c	f, r	m	4
Experiment H2	64_n	c	r, t	m	4
Experiment H3	60_n	c	r, s	m	4
Experiment H4	40_n	c	r	m	2

The central section of Table 2.3 presents a series of studies by Zahorik and colleagues. In the first of these studies, Longworth-Reed et al. provided evidence in support of Watkins' claim that compensation for reverberation is strongly affected by the time-direction of the reverberation (discussed earlier in § 2.1.6). This finding could not be predicted by the available models of reverberant-speech perception, however, and a great deal of further research into compensation for reverberation therefore ensued.

The bulk of the studies by Zahorik and colleagues (*the middle 6 of 8*) share a lot of similarities, although a variety of speech datasets (and listener tasks) are employed. In these six studies, speech stimuli are presented binaurally, with target speech directly ahead of the listener and a simultaneous background noise spatialised to appear from off to one side. The listener task varies a little with the speech database in use, but in each case involves identification of either *some* words

(selected keywords, alone or in combination) or *all* words (i.e. as many as possible) in a sentence. Percentage correct is calculated from the collected data, and is sometimes transformed (e.g., using the arc-sine transform (Kirk, 1968, p. 66) when the bulk of participant responses are toward the edges of the measurement range, i.e. close to 0% or 100% correct). Further, examining the influence of the background noise across different SNRs, the participant data is fitted in some studies with a logistic function approximating the psychometric curve, allowing them to derive the Speech Reception Threshold (SRT) at the 50% intelligibility point (Brandewie and Zahorik, 2010, 2012; Zahorik and Brandewie, 2011).

It was argued above that the inclusion of the spatialised masking noise in this case is likely to engage different, or perhaps additional, listening strategies than those at work in the purely monaural task where reverberant speech is heard without noise. The last study in this section, that of Srinivasan and Zahorik (2014), is therefore particularly encouraging to see since it examines perceptual compensation for reverberation in the absence of masking noise.

The work from Zahorik's lab takes a sizeable step forward from Watkins' in the search for a connected-speech measure of compensation for reverberation. Their studies ask about the effects of prolonged exposure to reverberation on speech as an entity in itself. Thus, any findings they can draw will clearly be relevant to speech in every-day listening situations. The down-side of their approach, at least to readers who do not have access to the raw data and are reliant instead on the published results, is that there is no analysis of which *types* of speech sounds are misheard (or recovered) in the differing reverberation conditions. Moreover, it is not obvious whether the effects of noise (where present) and reverberation could be satisfactorily separated in these cases of mistaken identity.

The final section of Table 2.3 looks ahead to four perceptual studies which are described fully in Chapter 5 below. These studies resemble Watkins' original work in two fundamental respects: they focus on perception of reverberant stop consonants, and they query the monaural compensation mechanism. However, these studies also echo work from Zahorik's lab, in that they use naturalistic speech materials which vary from trial-to-trial in regard to the talker, context and test-words. Here, consonant confusions are assessed with an information-theoretic measure which quantifies the *consistency* of mistakes, as well as the overall proportion of correct responses (Miller and Nicely, 1955).

2.4.3 Temporal envelope constancy

Constancies are named in the literature according to the thing that is constant, rather than the thing that does the distorting, e.g. shape constancy, colour con-

stancy, brightness constancy, and so on. In other words, it is the constancy of the percept, not of the underlying signal properties or stimulus conditions, that is named. Two recent studies suggest that compensation for the effects of reverberation in speech identification is underlain by constancy for perceptual attributes of the temporal envelope, though *which* attributes these are is not yet entirely clear. It is well-established that speech identification typically depends more strongly on temporal envelope cues than temporal fine structure (TFS) cues (see e.g., Shannon et al., 1995; Smith et al., 2002)[1]. It is perhaps not surprising therefore that compensation for the effects of reverberation *also* appears to be influenced more by envelope cues than by TFS cues. This account was proposed by Watkins et al. (2011) using 8-band vocoded speech material in monaural phoneme-identification tasks; moreover, recent evidence from Srinivasan and Zahorik (2014) directly supports this explanation with binaural stimuli as well.

Another important conclusion of Watkins et al. (2011) is that the monaural constancy effect appears to work in a band-by-band manner. That is, the level of reverberation in one auditory channel determines its contribution to the constancy effect relatively independently of the level of reverberation in other frequency regions. Conceptually, this band-by-band hypothesis implies that each auditory channel may independently adapt to the room conditions experienced, as interpreted by the temporal variation in the narrowband amplitude envelope in that limited spectral region. However, it is not yet clear *which* aspects of the narrowband envelopes influence the constancy mechanism; two different interpretations exist.

Since reverberation attenuates the modulation in temporal envelopes of individual frequency-bands of the speech signal, the first interpretation considers the amplitude modulation character of the signal. This bears some relation to the Speech Transmission Index (STI) and Modulation Transfer Function (MTF) on which it is based (cf. Sections 2.1.7 and 2.1.6 respectively), and essentially quantifies the preservation of the amplitude envelope spectrum. Alongside the speech perception studies discussed above, Zahorik and colleagues are also pursuing this line of research by investigating amplitude modulation (AM) detection thresholds (Zahorik and Anderson, 2013; Zahorik et al., 2011, 2012). Interestingly, these studies find that AM thresholds in reverberant rooms are lower than those predicted by the acoustical MTF (i.e., whether measured binaurally or monaurally, human sensitivity to AM is higher than anticipated); and in addition, AM thresholds are enhanced by prior exposure to the room. In related work using binaurally modulated signals, Reed and van de Par (2014) and Reed et al. (2014) are additionally attempting to

[1] Indeed, this is the principle by which cochlear implants bypass the acoustic stages of hearing and deliver electric impulses along the cochlear partition that are sufficient for understanding speech (Rubinstein, 2004).

pick apart cues carried in the AM content of the signal that either assist speech identification directly or enhance intelligibility indirectly (e.g. through spatial unmasking). A recent study by Kuwada et al. (2012) has also begun to examine the neural coding of the envelope signal in the midbrain, and to study the transformation of this internal representation in various reverberant conditions.

Rather than study the modulation character of the signal itself, a second approach to temporal envelope constancy considers the tails that reverberation adds to the channel offsets in a signal. This line of thought may be motivated by the finding that the constancy mechanism breaks down under conditions where the reverberation signal is time-reversed, yet the modulation character of the acoustic signal is relatively unaffected (Longworth-Reed et al., 2009; Watkins, 2005a). The proposal here is that the tail-like portions are 'used' somehow to 'deal with' reverberation, perhaps to allow a perceptual filtering out or dampening of its effects, and that the same process cannot take place when the reverberation appears in ramps prior to signal onsets rather than tails from its offsets. Indeed, such temporal asymmetries are observed throughout the auditory system, for example in loudness judgements (Stecker and Hafter, 2000) or in perception of sound timbre (Rupp et al., 2013).

Chapter summary

This chapter examined human and machine identification of reverberant speech. It began by describing room-acoustics in § 2.1 and the way in which early reflections and late reverberation brought about differing perceptual effects, depending both on signal content and the room characteristic. Reverberation was seen to pose a much larger problem for machine listening systems than it did for human listeners, and § 2.2 described the tendency for ASR to become error-prone in real-room reverberation. Robustness to reverberation was realised by a range of engineering solutions acting either at the front- or back-end of the recogniser, and bio-inspired approaches were also suggestive of potential improvements that could be made in machine hearing, particularly in adverse conditions.

The second half of this chapter considered the effects of reverberation on human listeners from a psychoacoustic point of view. The effects of reverberation in speech identification were seen in § 2.3 to be largely localised to specific groups of speech sounds, the most affected being the unvoiced stop consonants. Finally, § 2.4 outlined the growing body of work investigating perceptual compensation for reverberation with human listeners, and identified a number of questions about the monaural mechanism that have yet to be investigated with naturally spoken stimuli. To this end, Chapter 3 now asks which auditory components may be relevant to the investigation of perceptual compensation for reverberation.

Chapter 3

Biological and computational auditory systems

Contents

Chapter overview

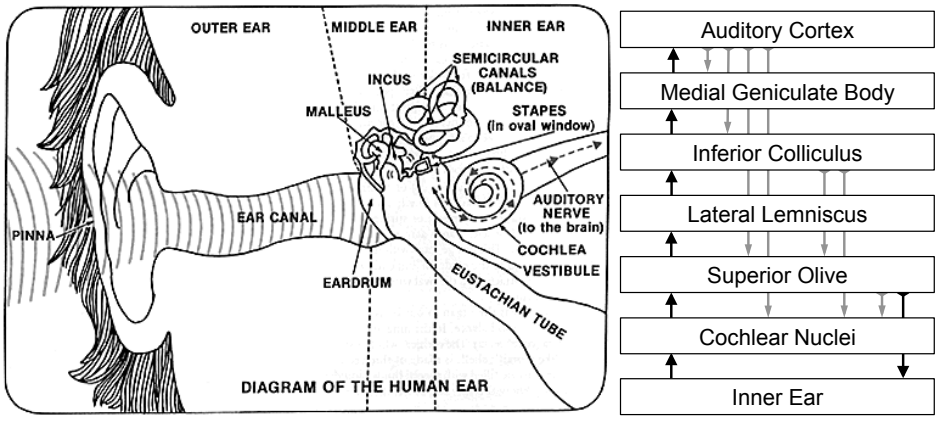
Chapter 3 gives an overview of biological and computational audition, and an insight into the benefits that an adaptive listening system might bring in reverberant environments. Firstly, the relevance of auditory modelling to the current research objectives is discussed in § 3.1. Secondly, the human auditory system is examined in detail, studying peripheral processing in § 3.2, central processing in § 3.3, and efferent feedback to the periphery in § 3.4. Finally, in § 3.5 the chapter ends with a discussion of the state of the art in efferent-inspired auditory models. Chapter 3 thus aims to uncover auditory processes relevant to perceptual compensation for reverberation, and to establish which components of the auditory system should be modelled in order to simulate these compensation effects.

3.1 Auditory modelling for reverberant listening tasks

The human auditory system turns sound into electrical signals that the brain can interpret. One stimulus can, however, lead to many different neural representations depending on its presentation level, surrounding environment, and recent historical context. Additionally, it follows that one pattern of neural activity can in fact be evoked by many diverse real-world circumstances. This chapter describes how the efferent system is thought to re-calibrate the mapping between the sound experienced and neural activity transmitted.

The external ear is only a fraction of what we hear with. Rather, our hearing is vitally dependent on a number of delicate systems buried deep within our skulls. The peripheral and central auditory systems are overviewed in Figure 3.1. The peripheral auditory system, seen in Figure 3.1a, comprises the outer, middle and inner ear. Together, these transform pressure variations in air into nerve impulses delivered via the brain stem to the auditory cortex. This processing chain forms part of the ascending pathway, described below in § 3.2. Also known as the *afferent* pathway, the primary function achieved here is the basic analysis of sound: frequency is encoded in nerves by location along the cochlear partition; intensity is encoded by the numbers of nerve fibres responding and by their firing rates.

Crucially, the peripheral auditory system is under the influence of the central auditory system (as well as the exterior world). A series of *efferent* pathways descend through the central auditory system, reaching back into the earlier stages of the auditory system as shown in Figure 3.1b. In the auditory periphery, efferent signals are believed to alter the sound-encoding behaviour of the cochlea, and ultimately



(a) Peripheral **(b)** Central

Figure 3.1: Peripheral and central auditory processing. Figure 3.1a shows the structure of the human peripheral auditory system, as depicted by the Laurent Clerc National Deaf Education Center (LC-NDEC, 2009). Figure 3.1b, adapted from Ryugo (2011), depicts a simplified schematic of central auditory system, showing afferent (*upward arrows*) and efferent (*downward*) connections. The final descending arrow (shown *dark*) depicts the stage at which the olivocochlear (OC) neurons provide efferent feedback to the peripheral system, and innervate the outer hair cells in the inner ear.

to confer robustness in adverse listening conditions. This topic is the focus of § 3.4 below.

3.1.1 Relevance of the efferent system to reverberant listening

Efferent signals appear to continually adjust the sound-encoding behaviour of the cochlea, adapting the auditory system to suit the listening conditions being experienced. Increasingly, evidence suggests that efferent processing assists compensation for environmental noise through its role in controlling the dynamic range of hearing (see e.g. Guinan, 2011; Guinan and Gifford, 1988, and further discussion in § 3.4 below). In turn, this cochlear re-adjustment affects neural transduction by the inner hair cells, and is thought to result in an enhanced signal representation in the auditory nerve, ultimately leading to improved perception of speech in noise.

The physiological process by which efferent feedback to the periphery might be involved in perceptual compensation for reverberation is entirely speculative at present, however, it is not without foundation. The main effects of reverberation, depicted earlier in Figure 2.2, appear essentially similar to the effects of additive noise when considered in certain analysis domains (i.e., the noise floor increases and the dynamic range of the signal reduces). This similarity underpins the pro-

posal in the current thesis that efferent processing may also account for our robustness to the effects of reverberation in speech identification.

3.1.2 Motivating an auditory modelling approach

Biological audition is rendered vulnerable to the effects of noise when the efferent system is compromised or lost (Guinan, 2006; Henderson et al., 2001). By analogy, a machine listener with no efferent simulation may be considered similarly disadvantaged, and it is speculated that performance may become more robust in adverse conditions by incorporating an element representing efferent processing.

The view that ASR and other machine hearing tasks may be improved by modelling the human auditory system has gained some momentum in recent years (see e.g., Elhilali and Shamma, 2008; Ellis and Weiss, 2006; Nix and Hohmann, 2007; Regnier and Allen, 2008; Seneff, 1988; Stern and Morgan, 2012). However, it is worth noting that 'increased robustness in difficult environments' is not typically expressed as a decrease in word error rate. Rather, an increased robustness to noise or reverberation typically is taken to mean a closer match to human listener error patterns on the same set of data.

A secondary benefit to auditory modelling becomes apparent when we consider that such models are useful for bringing together potentially disparate ideas, and can allow response data from several independent experiments to be combined in a single modelling task. By doing this, the mechanisms within a system may be examined in detail, and explanations offered. As a direct result of this, modelling questions often give rise to hypotheses about how a particular system may function (or may be dysfunctional). Auditory models are therefore useful in the process of making uncertainties explicit, and in designing experimental tests that may answer such points.

The auditory modeller's task is not trivial, however, as neural coding of sound in reverberation is poorly understood at present. While the afferent (ascending) auditory pathways are relatively well understood, particularly in the peripheral stages, the efferent (decending) pathways as yet leave many questions unanswered (and indeed, unasked). Crucially, it is these efferent pathways which are thought to underpin our remarkable ability to compensate for environmental obstacles.

The remainder of this chapter looks in more detail at how individual components of auditory systems are combined in nature and in simulation. Description of actual physiological systems are interleaved with computational models that attempt to capture something of their relevant form and function.

3.2 The peripheral auditory system

The peripheral auditory system is overviewed in Figure 3.1a. Regulated by efferent pathways from the central auditory system shown in Figure 3.1b, the peripheral system transforms acoustic vibrations in the environment into an electrical representation which is delivered through the auditory nerve into higher centres of the auditory system.

3.2.1 Outer and middle ear

The human outer ear comprises the *pinna* and *meatus* (or auditory canal). The geometry of the pinna leads to interference effects that attenuate certain frequencies while simultaneously boosting others. The pinna thereby filters the high-frequency components of the incoming sound, in a manner which depends on the angle of incidence to the head (Pickles, 1988). This provides a monaural cue for sound localisation which can assist with resolving front-to-back locations and can help to determine sound elevation. Ryugo (2011) suggests that descending auditory signals may be relevant even at these outermost reaches of the auditory system. The relationship between cue values and sound localisation is presumed to be learned through experience, thus as the head/body matures and grows, the values of such cues must also change over time. Here, Ryugo suggests that efferent circuits could facilitate the constant re-adjustments required to recalibrate our 3D coordinate system to preserve auditory space.

Following the pinna, sound is transmitted into the middle ear through the meatus, a cylindrical channel approximately an inch long which boosts frequencies around 2 to 3 kHz. This canal transmits sound to the eardrum and thereby induces a vibration in the smallest bones in our bodies, collectively known as the *ossicles* (and individually as the *malleus*, *incus* and *stapes*, or 'hammer', 'anvil' and 'stirrup' respectively). Accordingly, the frequencies that are particularly important for speech perception are transferred efficiently through the *oval window* to the fluids of the cochlea.

Additionally, the middle ear may also reduce the transmission of sound from differential bone movements (e.g. from chewing) through the skull directly to the cochlea (Moore, 2004). The descending pathway of the acoustic reflex (AR)[1], also known as the 'middle ear reflex', acts as a safety mechanism somewhat akin to

[1]The acoustic reflex (AR) was not pictured in the simplified auditory system overviews of Figure 3.1, but is seen in the more detailed schematic of Figure 3.7 below to carry feedback from central to peripheral regions of the auditory systems.

reducing the pupil size of the eyes in bright sunshine. When exposed to intense sounds, minute muscles attached to the ossicles contract and reduce the transmission of low frequencies (below about 1000 Hz) through the middle ear. This acoustic reflex is too slow to give protection from real-world shock sounds such as unexpected cracks and bangs, but is thought to protect against sustained high-level noise (Handel, 2006). Additionally, the AR may help to reduce the sound we hear of ourselves talking and to reduce the upward spread of masking (Moore, 2004). By attenuating low frequencies most strongly, the higher-frequency components that are critical for speech perception become relatively enhanced.

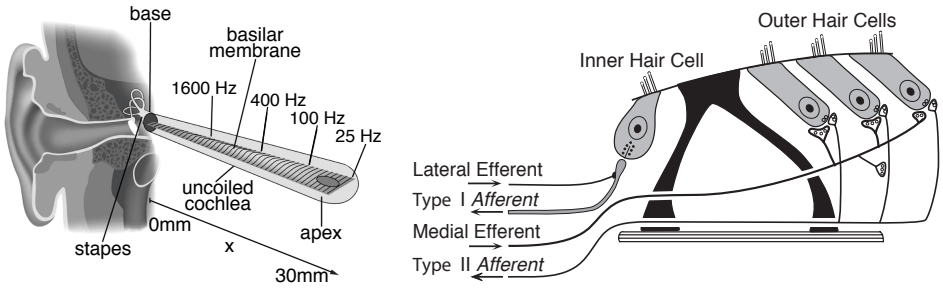
A filter representation of the outer-middle ear

Taken together, the outer and middle ear thus transfer sound pressure variation to the oval window and filter the signal's spectrum (aiding sound localisation and strengthening the speech-bands). The outer and middle ear can be modelled with a single linear filter that provides a broad resonance around 2.5 kHz that imitates the ear canal effect, with a subsidiary resonance around 5.5 kHz that corresponds to the external ear (Pickles, 1988). Alternatively, two filters may be combined in series to imitate each step in turn: the first filter characterises the outer ear, and the second, the eardrum-pressure to stapes-velocity transformation in the middle ear (Lopez-Poveda and Meddis, 2001).

3.2.2 Inner ear mechanics

The basic role of the cochlea is to convert the incoming sound waves into neural activity. The cochlea is filled with an almost incompressible fluid. Pressure exerted by the stapes on the oval window is transferred through this fluid, setting up a vibration effectively instantaneously on the basilar membrane (BM) that travels the length of the cochlea. The response on the BM is frequency-dependent, largely due to mechanical properties of the BM itself. If conceptually unwound from its spiral shape as shown in Figure 3.2a, the BM is stiff and narrow at the base (where high frequencies cause a maximal displacement) and wider and more flexible at its apex (where low frequencies show a maximal response). This has led to a view of the BM as a tonotopic frequency analyser such that each area of the BM appears to be tuned to a specific resonance frequency.

Cochlear mechanics have been studied for almost a century, yet our understanding remains incomplete. Early work focussed on afferent processing, examining the way in which signals are passed upwards through inner hair cell (IHC) transduction into the auditory nerve. However, as can be seen in Figure 3.2b, there are also clearly efferent connections within the cochlea, the majority of which terminate on



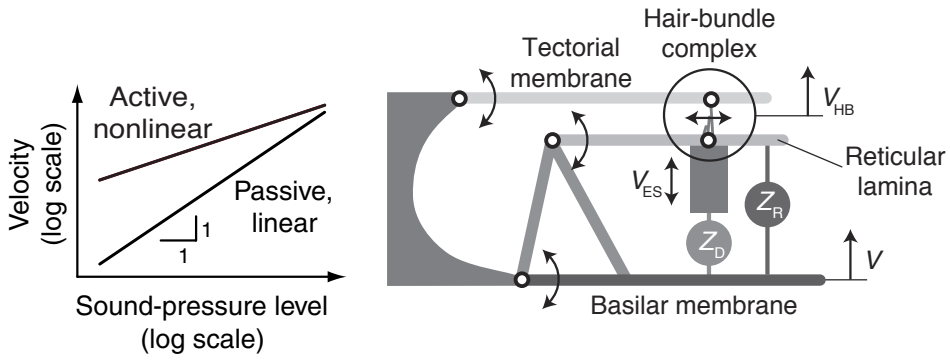
(a) Cochlear frequency analysis **(b)** Cochlear innervation

Figure 3.2: Cochlear frequency analysis and innervation. Figure 3.2a (redrawn from Kern et al., 2008) imagines the cochlea uncoiled to reveal a tonotopic frequency analyser. Figure 3.2b (from Guinan, 2011) shows the afferent and efferent connections in the *organ of Corti*. The majority (c. 95%) of afferent fibres (Type I) transmit information from the inner hair cells, while relatively few (c. 5%) afferent fibres (Type II) originate on the outer hair cells (OHCs) (Spoendlin, 1969). Lateral olivocochlear (LOC) efferent signals synapse on the afferent Type I fibres themselves, while medial olivocochlear (MOC) efferent signals innervate the OHCs.

the outer hair cells (OHCs). In this way, signals originating from higher centres of the auditory system can exert control over the sensitivity of the cochlea (Russell and Murugasu, 1997), such that the sound analysis undertaken within it is adapted to the current setting and listener task (Guinan, 2011; Reichenbach and Hudspeth, 2014). This is thought to arise because of the electromotility of the OHCs, whereby variations in the electrical stimulation cause contraction and elongation of the cell to such a degree that the surrounding tissues are additionally affected (Frolenkov et al., 1998).

As understanding of the efferent processing in the cochlear partition slowly but steadily increases, we are beginning to understand more about the resulting sound processing which occurs (and additionally, our models are gradually improving). This is illustrated in Figure 3.3a which describes the BM velocity near the resonant frequency. The passive, linear relationship (lower line) describes the relationship that was classically described by studying BM responses to sound in cadavers (von Békésy, 1947), where the resonance frequency depends on the local stiffness and mass of the membrane.

However, as shown in the upper line of Figure 3.3a, animal studies have shown that live cochlear responses are larger than postmortem cochlear responses (Rhode, 1971). This so-called 'cochlear amplification' has been studied non-invasively by means of otoacoustic emissions (OAEs) (Kemp, 1978; Manley and Köppl, 1998; van Dijk et al., 1989; Yates and Kirk, 1998). Despite the wide adoption of these techniques, however, there remains much that is not yet understand about efferent action at the level of the cochlea. In particular, Guinan (2014) warns that it is



(a) BM input-output relation **(b)** Model of the cochlear amplifier

Figure 3.3: Active cochlear modelling. The 'passive' response displayed in Figure 3.3a shows the motion recorded in early studies of the BM, where an increase in sound-pressure level corresponded linearly to an increase in BM velocity. The 'active' response recorded in a living cochlea instead becomes nonlinearly compressive at high signal levels. Figure 3.3b presents a model of the organ of Corti, which rests on top of the BM. The tallest stereocilia in the hair bundle of the OHC is firmly embedded in the tectorial membrane (TM) above. Vertical motion of the BM thus depends on both the vibration arising from the input sound signal *and* the motility of the OHCs. Both figures are redrawn from Reichenbach and Hudspeth (2014).

not yet understood how well modifications in the OAEs represent actual changes occurring in the auditory nerve[1].

Despite these open questions, the scheme in Figure 3.3a remains generally accepted in the research community at present. The active response is greater than the passive cochlear response, and boosts the amplification near the resonant frequency. However, BM vibrational response does not grow in proportion to the level of the input stimulus, but instead is nonlinearly compressive: the response is maximal for low-level sound intensities (above threshold) but is increasingly attenuated for high intensity sounds, especially for high frequencies.

Thus, for a living human, the BM response derives from the physical properties of the BM (stiffness, mass) and any active tuning mediated through the efferent system. To account for this, the explanation of cochlear mechanics proposed by Reichenbach and Hudspeth (2010) describes BM response near its resonance frequency by means of its local impedance which includes terms for the local stiffness and mass of the membrane as before (for the passive system), and now additionally

[1]Nor is it yet fully understood how the lateral olivocochlear (LOC) and medial olivocochlear (MOC) efferent signals may interact, nor how ipsilateral and contralateral effects relate, nor to what degree the sizes of measured effects would vary with listener attention and task-difficulty. Discussion returns to such queries in § 3.4 below.

depends on a viscous damping coefficient which varies with the (active) cochlear amplification. A recent micro-mechanical model of the BM response by Reichenbach and Hudspeth (2014) is shown in Figure 3.3b. In this scheme, the vertical motion of the BM depends in part on the properties of the OHC which are themselves innervated under efferent control.

At the furthest extreme, the compressive non-linearity in the inner ear could be viewed as a safety-mechanism preventing the ear from shaking itself apart with strong input signals. More generally, the nonlinear cochlear response can be viewed as a mechanism to circumvent response saturation and retain perceptual contrast (Handel, 2006). The involvement of the MOC efferent system in controlling the sensitivity of the BM is discussed in greater detail below (cf. § 3.4).

Cochlear models

Cochlear models represent the vibration at a given place on the basilar membrane in response to a stimulus. In attempting to confer the same responses on a computer simulation that were experimentally measured in cadavers (von Békésy, 1947), early cochlear models historically fell into two broad categories: *transmission line* models and *filterbank* models.

Zwislocki-Moscicki (1948) provided an early simulation of the motion of travelling waves along the BM using a transmission line framework where the mass and friction of the BM are taken to be constant while the flexibility increases with distance from the oval window. This idea was later taken up by Allen (1981) and by Lyon (1982), using a cascade of notch filters to gradually remove the high frequency components from the pressure wave in the cochlear fluid, each with a parallel resonance filter that converts the pressure on the BM into displacement.

Motivated by psychophysical studies of the frequency-resolving power of the human ear, filterbank models require channels whose bandwidth and spacing increase with frequency. To this end, early experimental work by Zwicker (1961) tabulated the critical bandwidth as a function of centre frequency, and led to the Bark frequency scale (mentioned earlier in regard to ASR feature creation, cf. § 2.2.1). The Mel frequency scale was similarly inspired by perceptual correlates of pitch-resolution (Davis and Mermelstein, 1980; Stevens et al., 1937). An alternative method that matches well to human data first describes the magnitude response of vibration at a particular point on the BM with a gammatone function (de Boer, 1979), and secondly distributes such filters across the frequency axis in proportion to the measurements of the human auditory filter bandwidth (Moore, 1986; Patterson et al., 1988). This produces a warped-frequency mapping known as the equivalent rectangular bandwidth (ERB)-rate scale which is almost logarithmic,

and may include any number of filters dependent on the resolution desired for auditory modelling.

Although practical for many purposes, the assumed linearity and symmetry of the gammatone filter has not stood up to empirical measurements of the BM which in fact shows an asymmetric and level-dependent frequency response (Glasberg and Moore, 2000). In order to model the broadening of auditory filter bandwidth that is seen with an increased stimulus level, a number of extensions to the model have been proposed including the *gammachirp* filter (Irino and Patterson, 1997) and the dual-resonance nonlinear (DRNL) filter (Lopez-Poveda and Meddis, 2001). The latter is of particular interest here as it provides a framework that can later be extended to introduce efferent auditory processing[1].

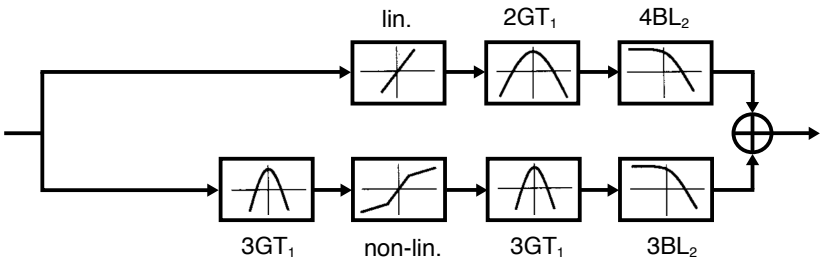
The DRNL filterbank model combines a linear and nonlinear path in parallel to transform stapes displacement to BM vibration (Lopez-Poveda and Meddis, 2001; Meddis, 2006; Meddis et al., 2001). As shown in Figure 3.4a, the linear pathway models the passive mechanical properties of the BM by cascading gammatone and low-pass filters. The nonlinear pathway simulates the influence of outer hair cells by including a 'broken stick' function whose compressive effect varies along the length of the BM. In Figure 3.4b, Ferry and Meddis (2007) introduce an attenuator at the beginning of the nonlinear pathway in order to model the cochlear compression that was shown to be a feature of active cochlear processing (cf. Figure 3.3a). This model is discussed further in § 3.5 below, where the state of the art in efferent modelling is introduced and assessed. Before this, discussion first continues along the afferent processing chain to briefly examine the higher auditory centres, and then returns to introduce efferent feedback to the periphery.

3.2.3 Hair cell transduction at the auditory-nerve synapse

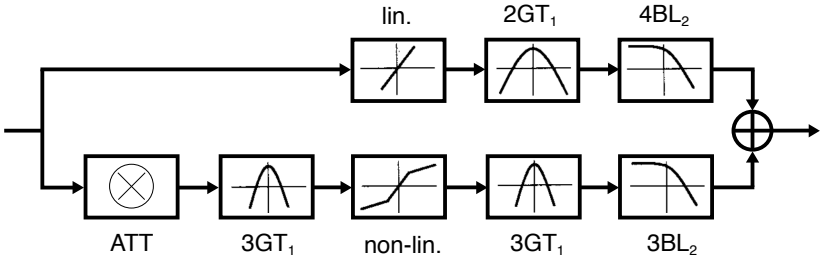
The environmental information that has been derived by the cochlea is passed upwards into the central auditory system in such a way that the tonotopic 'frequency axis' is maintained at every level. In general, the pitch of a sound tends to be coded in terms of which neurons are responding, and its loudness is determined by the rate of response and the total number of neurons activated.

Mechanical-to-neural transduction is achieved in the *organ of Corti*, positioned between the basilar membrane (BM) and the tectorial membrane (TM). While the outer hair cells (OHCs) are largely understood to be involved in efferent control of the sensitivity of the cochlea (Russell and Murugasu, 1997), it is predominantly the inner hair cells (IHCs) which transmit the detail of our local sound environment

[1] This cochlear model forms the centre-piece of the modelling studies undertaken in Chapter 4.



(a) Dual-resonance nonlinear (DRNL) filter of Lopez-Poveda and Meddis (2001).



(b) Efferent attenuation introduced by Ferry and Meddis (2007) in the nonlinear pathway.

Figure 3.4: The dual-resonance nonlinear (DRNL) filter, redrawn from Lopez-Poveda and Meddis (2001), is shown in Figure 3.4a. The input to the model is the stapes velocity (m/s). The output, representing the basilar membrane velocity (m/s), is the sum of the signal from linear (*top*) and nonlinear (*bottom*) pathways. In addition to the linear (lin.) or nonlinear (non-lin.) gain after which the branches are named, each pathway comprises a number of gamma tone (GT) or Butterworth lowpass (BL) filter cascades; the number of times the filter is applied is shown in large script, and the filter order is represented in the subscript. Figure 3.4b presents the adaptation by Ferry and Meddis (2007) which simulates the effects of efferent suppression by applying an attenuation, ATT, at the start of the nonlinear pathway.

further upwards to the auditory nerve (AN)[1]. Deflection of IHC stereocilia occurs due to shearing forces between the BM and TM, instigating a chemical exchange process where potassium ions flow into the hair cell. This leads to a release of a chemical neurotransmitter (when the hairs bend in one direction) and, when a sufficient quantity has been released, an action potential or *spike* travels up the auditory nerve.

[1]Spoendlin (1969) reports that 95% of the afferent AN fibres transmitting sound towards the brain, known as Type I afferents, originate from the IHCs (cf. Figure 3.2b). These fibres transmit the details of our local sound environment further upwards to the AN, encoding the frequency, intensity and timing of sounds present. Around 5% of the afferent fibres contact the OHCs instead, and are known as Type II afferents. Though comparatively less is known about these fibres, recent data in Weisz et al. (2014) is consistent with an earlier proposal that the Type II afferent fibres integrate acoustic information over a wider frequency range, and respond in particular to high-intensity sounds.

There is a one-to-many relationship between IHCs and cells in the AN: each hair cell may be connected to as many as 10 nerve cells (though each nerve cell is only connected to 1 hair cell), and approximately 30,000 neurons in each auditory nerve are available for carrying auditory information from the cochlea to the central nervous system (Moore, 2004). Since an incident sound wave dissipates most of its energy near the resonance frequency of the BM, different frequencies at input cause maximal displacement at different *places* along the cochlear partition. In this way, hair cells at different locations are maximally stimulated by different input frequencies. Increased intensity of stimulation causes a more rapid rate of response, but nonlinear compression assists in maintaining sensitivity across widely-varying sound levels.

Additionally, *timing* theories suggest that the auditory system may also derive information from the interval between successive spikes and thereby encode frequency from the periodicity of the spike train (Young and Sachs, 1979). This theory relies on the fact that, for frequencies below about 4 kHz, neural spikes tend to occur at a particular phase of the stimulating waveform. That is, although auditory nerve fibres have a probabilistic (i.e. random) chance of firing in any given cycle, when they *do* fire they tend to do so at the same point in the stimulus cycle. A temporal regularity therefore exists in the firing pattern of a neuron in response to a periodic stimulus, such that inter-spike intervals are very close to integer multiples of one period of the stimulating signal. Above around 4 kHz, phase locking can no longer occur as the chemical changes of the hair cell transduction process are not sufficiently fast to encode signal detail. Thus the cochlea preserves both *temporal*- and *rate*-information, allowing signal details to be conveyed through the phase locking of stimulus-synchronised discharges even when the firing-rate has become saturated.

In addition to maintaining frequency selectivity through the tonotopic arrangement of hair cells, a number of other phenomena help ensure that a large dynamic range is maintained in the AN spike train in order to accommodate the vast range of intensities that are encountered in daily life.

Firstly, the nerve cells themselves are differentiated such that individual AN fibre groups encode different ranges of intensities present. Nerve cells fire even in the absence of any stimulus: the level of this background activity is known as the spontaneous rate (SR). A continuum exists such that individual nerve fibres have different thresholds and thereby encode different ranges of intensities present (Manley, 2000). Splitting the intensity range across different types of nerve fibres increases the chance that a set of nerve fibres exist that are not yet saturated. The hearing sensitivities of different species may depend partially on the distribution of such fibres: whereas humans are often described as having low and high SR fibres,

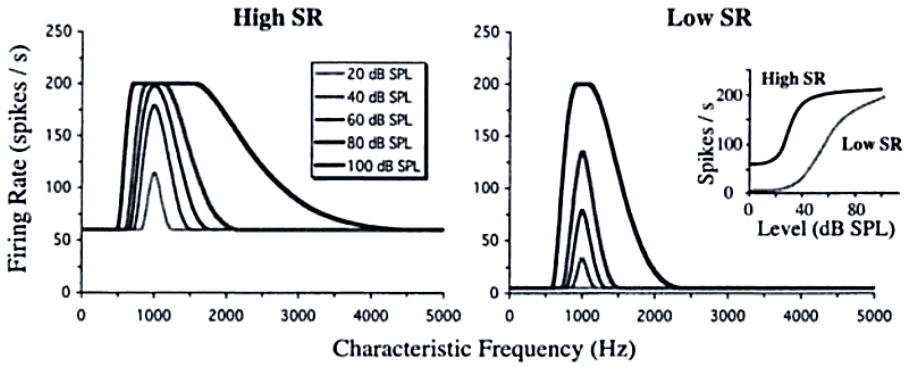


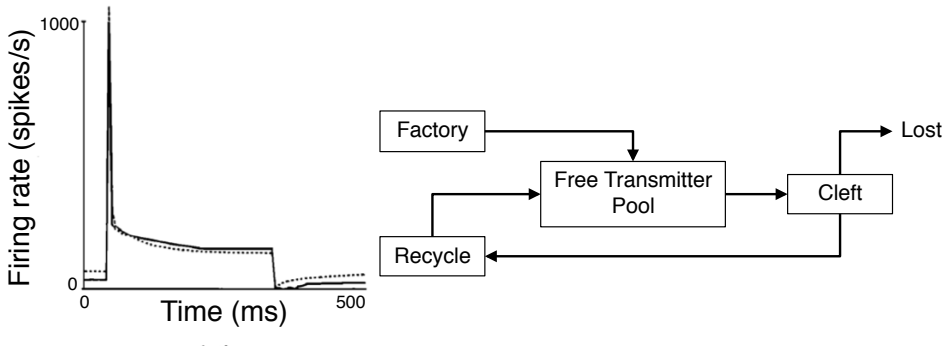
Figure 3.5: High and low spontaneous rate (SR) auditory nerve fibres. Firing rate is shown as a function of characteristic frequency for a high SR fibre (*left*) and a low SR fibre (*right*) in response to a 1000 Hz pure tone presented at different levels. Both auditory nerve (AN) fibres can be seen to respond to a narrow range of frequencies at low intensities, and a broader range at higher intensities. The rate-level functions of the two fibres to this tone, inset on the right hand side of the figure, and show comparatively larger encoding of dynamic range that can be achieved by the low SR nerve fibre. Image from Plack (2005, p. 124).

cats, for example, are thought to have medium SR fibres in addition (Liberman, 1978).

Reviewing this topic, Handel (2006) reports that for humans, approximately 85% of the neurons are high-SR fibres. The high SR fibres fire spontaneously at a rapid rate (tens of times per second) and saturate at low signal levels. We have a far smaller number of low SR fibres (that spike fewer than c. 15 times per second). These fibres saturate at much higher signal levels, and thus encode signal strength over a much wider dynamic range. Figure 3.5 illustrates the differences between these fibre types. The inset panel (Figure 3.5, right) shows that high SR fibres are already saturated by around 40 dB, and suggests that at normal conversation levels of around 60 dB (here, without involvement of the efferent system) it is the low SR fibres that are particularly relevant for speech perception[1].

However, each of the individual IHCs that provides a signal to the AN fibres additionally benefits from a gain-control mechanism that allows the response to alleviate saturation and adapt to the intensity levels they detect (Handel, 2006). Thought to be under efferent control, this adjusts the dynamic range between spontaneous and saturation levels, and assists in maintaining an optimum response pattern to perceive our ever-changing world. This type of effect, brought about by efferent feedback to the periphery, is described in detail in § 3.4 below.

[1] As such, it is the low SR fibres that are modelled in Chapter 4, cf. § 4.2.1 below.



(a) Post-stimulus time histogram **(b)** Components of the Meddis hair cell model

Figure 3.6: Figure 3.6a is a post-stimulus time histogram showing hair cell adaptation to a steady tone stimulus. The solid line is the function derived by Westerman (1985) and the dotted line results from the Meddis hair cell model. Figure taken from Meddis (1988). Figure 3.6b describes the hair cell transduction process implemented by the Meddis model (redrawn from Meddis et al., 1990). The onset of a tone causes an immediate rise in firing rate proportional to the (half-wave rectified) motion of the BM, represented by the arrow pointing from the *free transmitter pool* to the *cleft*. This peak onset-response is immediately followed by an initially rapid decline in firing-rate and then a slower adaptive decline, resulting in a firing-rate representation that emphasises the leading edges of stimulus inputs. In the model, the free transmitter pool is replenished over time, partially by the *factory* (a source representing new creation of the necessary chemicals) and partially through the *recycling* or re-uptake of previously ejected neurotransmitter. A relatively small amount is *lost* but this loss is eventually balanced by the slow manufacture from the factory which gives rise to the steady-state condition of the adapted auditory nerve (whose level depends on the amplitude of the stimulating tone).

Irrespective of its particular threshold or associated dynamic range, all auditory nerve fibres are thought to respond with a similar temporal discharge pattern, shown in the post-stimulus time histogram (PSTH) of Figure 3.6a. Stepping along the time-axis, this PSTH shows the non-zero spontaneous firing rate in quiet conditions and a sharp onset spike when the stimulus tone begins. Following this, a two-stage adaptation is seen: first a rapid decline, then a slower decline (with time constants around 10 and 70 ms respectively). Eventually the hair cell becomes well-adapted to the stimulus and a steady-state is reached. When the stimulus stops, the spike activity momentarily drops below the spontaneous rate before recovering to it. In the review discussed above, Handel reports that for a typical high SR fibre with a characteristic frequency of 6900 Hz, a 60 dB (i.e. $1,000$-fold) difference in intensity results in only a factor of three increase in the onset rate, and a factor of two increase in the steady-state response.

Modelling mechanical-to-neural transduction

Auditory models represent hair cell transduction using the velocity or displacement of the basilar membrane as an input, resulting in a measure of nerve activity across an array of inner hair cell models. Implementations vary, and the output of a single IHC model may take various forms. Information about neural spikes may be output either as a firing rate map, as a spike train (with which timing information is encoded) or as a probabilistic measure i.e., spike-likelihood. In the computational model described by Ghitza (1988), for instance, the phase-locking of the auditory nerve is modelled with an 'ensemble interval histrogram' that preserves fine spectral details in low-frequencies only. In this way, useful aspects of auditory function are abstracted and incorporated into the front-end of a speech recogniser.

The review by Hewitt and Meddis (1991) examines eight IHC models by examining their ability to reproduce mammalian responses to specific tone-burst stimuli. While no single model could replicate all the desired features, Hewitt and Meddis recommended Meddis (1988) in terms of both the match to physiological data and computational efficiency. The scheme underpinning this model is described in Figure 3.6b. Viewed as a functional model, the Meddis IHC model is capable of exhibiting the major features of the mechanical to neural transduction process (as were seen previously in Figure 3.6a): spontaneous firing, saturation, a maximal sensitivity to a particular frequency and, after an onset peak, a rapid decline to the adaptation level whose rate is independent of tone intensity and background level (Meddis, 1986, 1988; Meddis et al., 1990).

However, the biological plausibility of the Meddis IHC model has recently been questioned. The model by McEwan and van Schaik (2004), for instance, proposed a simplification that achieves a similar result, and Sumner et al. (2002) presented a revision which aims to be more faithful to known physiology. AN activity was also modelled by Zhang et al. (2001), and updated by Zilany et al. (2009) to simulate the dynamical response observed over several timescales at the synapse between IHCs and AN fibres. In their model, a slower adaptation accounts for the recovery after stimulus onset, and a faster rate adaptation simulates the response to alterations in signal level for a continuous stimulus.

3.3 The central auditory system

A simplified schematic of the central auditory system was presented at the start of this chapter (cf. Figure 3.1b). It shows the major nuclei involved in the ascending (afferent) auditory processing chain, and labels a number of descending (efferent) pathways in addition. Importantly, it can be seen here that a large number of the

organs involved in the afferent pathway receive efferent signals, and are thus under control of the higher levels of the auditory system. In other words, the behaviour of the afferent sound-encoding process can be modulated at almost every stage in its upward journey.

This section briefly introduces a few parts of the central auditory system that are particularly relevant to reverberant speech processing in the brainstem, midbrain and cortex.

3.3.1 Brainstem

The key idea of audio processing in the brainstem is that the previously single-track processing of the periphery now becomes massively parallel. Progressing upwards through the auditory system, the acoustic signal information which was originally spread over multiple fibres in the AN is gradually encoded in a more robust manner at the level of the single cell (Joris and Smith, 2008).

Ascending information from the auditory nerve fibres is delivered to the brainstem via the cochlear nucleus (CN). The CN is split into separate regions, each of which has its own tonotopic map. Its various specialised cells and synapses provide for different routes through the brainstem, each of which passes a differently conditioned set of data onwards to the midbrain. Cao and Oertel (2010) discuss three of the major cell types found in the CN. Firstly, *bushy* cells provide 'primary responses', encoding timing information by phase locking to individual cycles at low-frequency, and following the temporal envelope for high-frequency sounds. Secondly, *octopus* cells integrate across a population of AN fibres and thus encode 'onset responses' in the presence of transient broadband pulses. Thirdly, *T stellate* cells signal energy transients over a small range of frequencies, effectively detecting spectral alterations in the input.

Interestingly, Bürck and van Hemmen (2007) propose the CN as a site for monaural echo suppression (such that a lead-click may suppress a lag-click which follows at a delay of around 2–3 ms). However, the advantage from such processing is clearly not sufficient to fully remove the effects of reverberation. Sayles et al. (2013) discuss *chopper* cells in the CN which are likely to be of particular relevance in reverberant listening situations since they encode amplitude modulation with a high degree of sensitivity (indeed, the most sensitive of these units matched the thresholds of human psychophysical performance). Analogous to the 'best frequency' of an AN fibre, chopper units can be characterised by their 'best modulation frequency' (independently of the modulation depth). Since reverberation alters the modulation content apparent in a signal, these cells might be implicated in the degradation

of the neural representation of pitch (extracted from timing information) that was reported for the anaesthetised guinea pigs studied by Sayles and Winter (2008).

Following the CN, the superior olivary complex (SOC) is the next level in the afferent pathway. The SOC is the first level of the auditory system at which binaural information is available, that is, that inputs from the right side and left side converge in the superior olive (SO) on each side of our head. Concerned primarily with sound localisation (particularly azimuth detection), the medial SO performs a low-frequency analysis to consider interaural time differences (ITDs), and the lateral SO computes the interaural level difference (ILD) cue for high-frequency signals (Darrow et al., 2006).

The SOC is also of particular relevance to the current study since the descending olivocochlear feedback derives from here. Discussed extensively in § 3.4.1 below, these efferent signals are thought to confer robustness in complex and noisy listening situations (Henderson et al., 2001). Indeed, it has also been suggested that these descending circuits might act to mitigate asymmetries in our physiology, and assist in the calibration of left- and right-side derived information in our binaural auditory system (Cullen and Minor, 2002; Darrow et al., 2006). Further, it appears that effects of attention and experience are also be apparent at the level of the brainstem response. For instance, Bidelman and Krishnan (2010) have recently shown that reverberation effects in the brainstem were less detrimental for musicians than those without such auditory training.

3.3.2 Midbrain

The inferior colliculi (IC) are large nuclei in the midbrain which receive monaural input from the CN and binaural input from the SO. Moreover, the ICs are involved with multi-sensory integration, responsible for the startle response, attention reflexes, and learned reflexes (Gruters and Groh, 2012).

A recent report by Devore et al. (2009) suggested that the IC continues the binaural localisation processing described above. A single neuron was shown to provide a similar pattern of data to that of psychophysical performance measured with human listeners, in that the presence of reverberation degraded the sensitivity to direction overall. Here, Devore et al. reported an 'onset dominance' cue which effectively highlights the more reliable spatial cues present in the signal (typically those which arrive early in the stimulus history before reflections have time to build up). Additionally, Devore and Delgutte (2010) suggest that the ILD pathway in the IC provided more accurate cues than envelope-based ITDs for localisation of high frequency sounds in reverberation.

Kuwada et al. (2012) caution, however, that although studies often report IC responses collected from anesthetised animals, this anesthetisation is itself known to alter binaural processing in the IC. Using unanesthetised rabbits, Kuwada et al. demonstrated that single neurons in the IC were able to encode both amplitude and azimuth in such a way that the effects of reverberation were largely ameliorated. Here, the example neuron encoded source distance in response to binaural input: the input signal level was held constant, but the firing rate decreased systematically with the distance of a reverberant source. Additionally, the example neuron adjusted the modulation gain of the signal in response to monaural input, and thus compensated for the level of reverberation present in the signal.

3.3.3 Cortex

The varied subcortical processes occurring in the brainstem and midbrain suggest that the auditory cortex is highly specialised for certain types of pre-conditioned inputs. The auditory cortex is interconnected with other cerebral areas, and also transmits descending signals to several lower parts of the auditory system (cf. Figure 3.1b). The role of cortical processing remains somewhat ambiguous, however, activity in the cortex can itself be studied with non-invasive techniques such as magnetoencephalography (as used by Ding and Simon, 2011) or high-density electroencephalography (as used by Horton et al., 2013).

Lomber and Malhotra (2008) have recently reported that different regions of the cortex are involved with recognition and with localisation (reminiscent of the division of 'what' and 'where' processing suggested by the psychophysical studies of Smith et al., 2002, discussed earlier in § 2.3). Focusing on the former of these pathways, recent studies have revealed that cortical delta band (1–4 Hz) and theta band (4–8 Hz) oscillations entrain to the slow temporal modulations of the envelope of a speech signal at syllable and word rates (Ding and Simon, 2013a, b, 2014). Moreover, an *attended* signal is represented more accurately at the cortex than is a distractor signal containing either noise or a competing speaker (Horton et al., 2013). Taken together, these studies suggest that by the level of the auditory cortex, the brain is capable of encoding a representation of speech that is largely invariant to the sonic background in which it occurred.

3.4 Efferent feedback to the periphery

Two main feedback systems regulate the *peripheral* auditory system: the acoustic reflex to the middle ear, and efferent innervation of the inner ear (Giguère and Woodland, 1994). These are depicted in Figure 3.7, which overviews the bridge

between the peripheral and central hearing processes. The first of these descending pathways, the acoustic reflex (AR), is a slow-acting reflex which is thought to be triggered mainly when high sound levels are experienced[1] (see discussion above in § 3.2.1). The second descending pathway, the olivocochlear (OC) system, provides a number of routes by which efferent neuron spikes can travel from the central nervous system back towards the cochlea.

Guinan (2006) reports two separate types of nerve travelling down the olivocochlear bundle (OCB), named due to their origin in the SOC and termination in the cochlea, as lateral olivocochlear (LOC) or medial olivocochlear (MOC) efferents. The *anatomy* of the OC system has now been fairly well mapped (see e.g. Brown, 2011), such that it is known in what manner the inner and outer hair cells are separately innervated by LOC and MOC neurons (as was shown earlier, cf. Figure 3.2b). On the other hand, the *function* of the OC is still incompletely understood.

The role of the MOC neurons has been studied in a number of species, and is now understood to act on the OHCs to reduce cochlear sensitivity. As was described above, research in this area relies on the cochlea being *active*, and thus requires a non-destructive form of measurement. The clinical description of otoacoustic emissions (OAEs) by Kemp (1978), has led to a family of OAE-based methods being developed over the past 35 years. This is because it became apparent that the faint sounds emitted by the OHCs could be modulated, and thus recorded (non-invasively) in the audio domain, by activation of the MOC efferent pathway (Mountain, 1980; Siegel and Kim, 1982).

However, little is yet known about the LOC system since there is no sound-evoked test of its function (Guinan, 2014). Thus the contribution of LOC efferents, either separately or in concert with MOC signals, remains a subject awaiting future investigation (Brown, 2011; Guinan, 2006, 2014).

3.4.1 The medial olivocochlear system

Substantial experimental evidence now supports the claim that electrical stimulation of the MOC fibres results in a reduced effect of acoustic stimulation (e.g. Robles and Ruggero, 2001; Russell and Murugasu, 1997). As was shown schematically in Figure 3.7, fibres of the efferent MOC system descend from the brainstem and terminate on the OHCs. The length and stiffness of the OHCs are modulated

[1] Indeed, the AR is omitted from the modelling study presented in the following chapter, since the sounds heard are presented at a level well below that thought to give rise to contraction in the middle ear muscles.

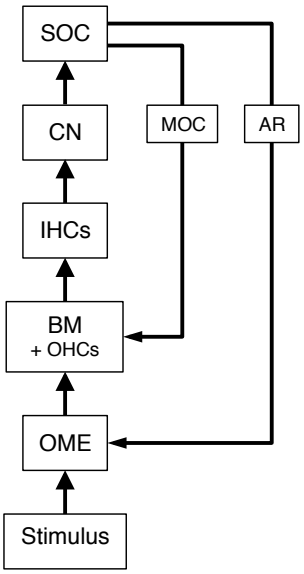


Figure 3.7: Efferent feedback to the periphery. A stimulus is processed through the afferent chain (depicted with upward arrows), passing first through the outer and middle ear (OME), where it may be attenuated by the acoustic reflex (AR) (outer downward arrow) if the level is sufficiently loud. The transmission of sound through the basilar membrane (BM) may also be adjusted by the medial olivocochlear (MOC) reflex (inner downward arrow), effected through changes to the motility of the outer hair cells (OHCs). Following this, transduction at the inner hair cells (IHCs) transforms the vibrational energy to a neural firing pattern which is transmitted up the auditory nerve (AN) to higher centres of the auditory system (of which the cochlear nucleus (CN) and superior olivary complex (SOC) alone are shown).

by the descending efferent signals, and since the OHC tips are embedded into the TM, the motion of the IHCs with BM-TM shearing is altered in turn (Kim, 1986). That is, activation of the efferent fibres inhibits the cochlear response, and thereby alters the signal properties encoded at the level of the auditory nerve.

Two major subgroups of MOC neurons exist, ipsilateral and contralateral, defined according to which ear (monaurally) excites their response (Guinan, 2006). In addition, a third (though more minor) subgroup appears to respond to sound in either ear (Brown, 2011). In natural listening conditions, however, the medial olivocochlear reflex (MOCR) evoked in one ear would typically arise from sound coming into both ears (i.e., ipsilaterally and contralaterally). In this way, the resulting cochlear suppression would depend on the combination of efferent activity in *both* these descending pathways.

However, whether examined monaurally or binaurally, the effects of MOC activation on the auditory nerve response are very difficult to measure and quantify at present. Chintanpalli et al. (2012) note, for instance, that while the ipsilateral effect may be measured by shocking the MOC bundle, it is extremely difficult to measure this effect using sound itself to elicit the response without producing ancillary effects such as AN excitation and subsequent suppression. Moreover, Guinan (2006) notes that studies may not draw pertinent conclusions from animal physiologically in this regard: it appears that for cats the ipsilateral reflex may be two or three times stronger than the contralateral, whereas for humans they appear approximately equal. Approaches based on OAEs separate the elicitor sound in

either frequency or time from the expected response (measuring distortion-product or transient-evoked OAEs respectively), and thus check for adequate presence and function of the efferent effects. Yet such tests still do not directly estimate the psychopysical effect size of OHC suppression (Guinan, 2014). Any MOC activity would clearly alter the magnitude and phase of the afferent cochlear processing; however, it is extremely hard at present to investigate the impact of such sound-evoked efferent suppression on interaural cues such as ILD and ITD. Further, recent studies are beginning to reveal that MOC effects are also subject to attention and experience; thus their involvement may differ between active and passive listening occasions (Guinan, 2014).

Cooper and Guinan (2003, 2006) report that, under efferent control, the OHCs influence the sound-processing function of the BM by two separate mechanisms. Firstly, a *fast* effect, thought to aid signal discrimination in noisy environments, has been attributed to an overall decrease in OHC motility. Secondly, a *slow* effect, thought to perhaps help protect the auditory system from excessive stimulation, has been linked with a change in the axial stiffness of the OHCs. Signals were reported to occur in the region of 10-100 milliseconds for the fast effect, and 10-100 seconds for the slow effect (Sridhar et al., 1997; Zhao and Dhar, 2011). On the other hand, Backus and Guinan (2006) reported effects acting on three timescales ('fast' $\simeq$ 70 ms, 'medium' $\simeq$ 330 ms and 'slow' $\simeq$ 25 s). Though the details of the various analyses differ, the studies all concur that the efferent pathway acts over a much slower time-scale compared with the high speeds at which the afferent system carries information to the brain. Thus the efferent system could be thought to provide a series of continual adjustments, recalibrating our auditory system as the environment around us gradually shifts.

3.4.2 MOC unmasking in noise

In this section, Figure 3.8 is used to explain the presumed role of the MOC efferent system in controlling the dynamic range of hearing (Guinan, 2011; Guinan and Gifford, 1988).

Figure 3.8a presents AN fibre responses to short tone bursts heard in an otherwise quiet environment. Without involvement of the efferent system, the firing rate here increases with stimulus level until the fibre's saturation level is reached at around 40 dB. In Figure 3.8b, MOC activation effects a shift of the rate-level curve to higher sound pressure levels (to the right). A tone presented at 40 dB now lies on the steep part of the response curve, and the fibre can increase its firing rate further still for sounds presented at stronger levels.

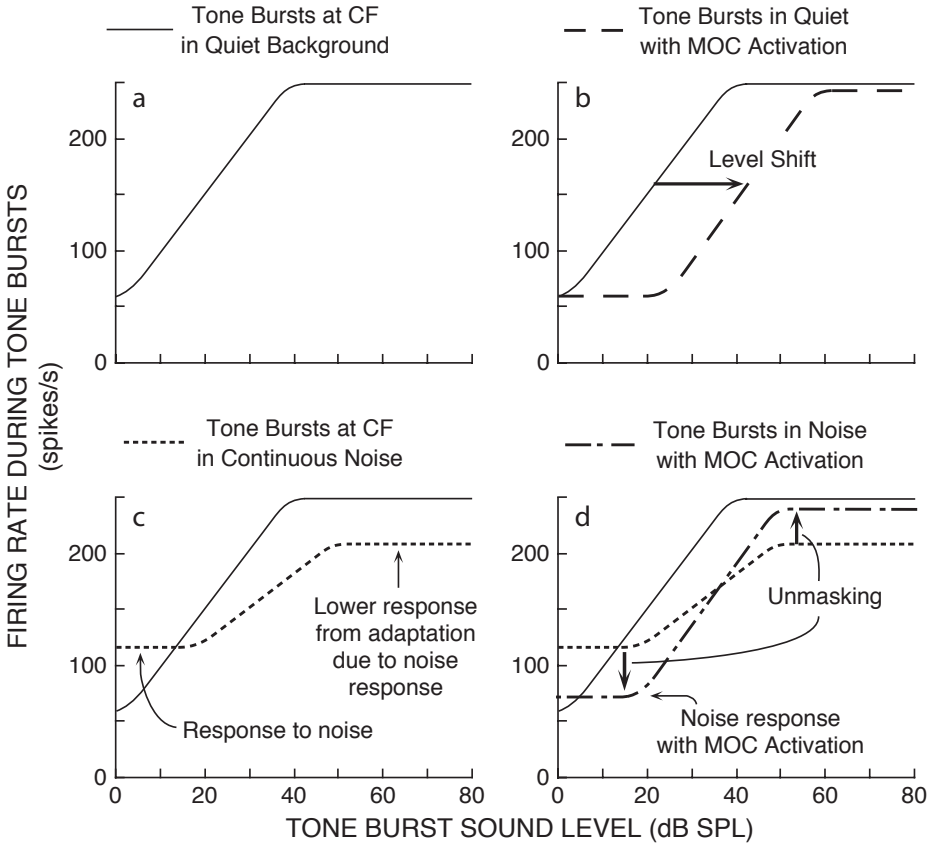


Figure 3.8: Figure 3.8 is from Guinan (2011) and shows medial olivocochlear (MOC) unmasking in the presence of low-level background noise. The top panels depict auditory nerve (AN) fibre responses to short tone bursts in quiet backgrounds (a) without– and (b) with– MOC activation, which effects a shift of the rate-level curve to higher sound pressure levels. The lower panels show equivalent rate-level curves in the presence of a continuous background noise. This has the effect of reducing the fibre's dynamic range (c): at low test-signal levels the AN responds to the presence of the noise (raising its overall response curve), while at high sound-levels the continuous noise causes additional adaptation which lowers the response to the test-tone bursts. The MOC activation in (d), however, increases the dynamic range of the fibre once again: the reduced cochlear gain causes the response to the low-level sounds to be suppressed; the background noise therefore causes less adaptation and the high-level tone bursts are responded to more strongly once more.

The lower panels of Figure 3.8 show equivalent rate-level curves in the presence of a continuous background noise. This has the effect of reducing the fibre's dynamic range in Figure 3.8c. Here, at low test-signal levels the AN responds more strongly to the presence of the noise than the test-tone, which raises its overall response. At high sound-levels the continuous noise causes additional adaptation which lowers the response to the tone bursts. The MOC activation in Figure 3.8d, however, increases the dynamic range of the fibre once again. Now, the reduced cochlear gain causes the response to the low-level sounds to be suppressed. As a direct result, the background noise causes less adaptation in the AN, and the high-level tone bursts are responded to more strongly once more.

Understood in this way, release from adaptation has the effect of increasing the dynamic range available. If these results can be carried over from test tones to speech signals, then the implication is that when speech begins in a noisy environment, a new surge may be seen in the AN response despite the concurrent presence of background noise. Indeed, mirroring the situation examined in Figure 3.8, Henderson et al. (2001) and Strickland and Krishnan (2005) report that in cases where the efferent system was impaired or missing it proved very difficult to perceive test stimuli in noise. Given that the AN response had already become well-adapted, it could not respond further to a new sound, even at an increased level. When the efferent system is functioning, however, the situation is quite different. Here, efferent activity reduces the BM displacement, the response curves of the AN fibres shift to higher sound pressure levels, and the dB region of normal conversation would once more lie on the sloping section of the AN response curve so that variation in the speech-signal may again cause a variation of the output response. Thus, the background noise can be tuned out to some extent because efferent suppression reduces the previously-adapted firing rate that the noise would ordinarily have produced.

3.4.3 Proposed relevance to reverberant listening

The proposal that the MOC efferent system might be involved in perceptual compensation for reverberation can now be re-examined. The effect of reverberation on a speech signal was demonstrated early in this thesis, where the overall dynamic range of the reverberant signal decreased as the late-reverberation caused a prolongation of energy which raised the noise floor of the signal (cf. Figure 2.2). There is clearly a high degree of similarly in this respect between the effects of reverberation on a speech signal, and the effects of additive noise. In a direct parallel to that outlined above in § 3.4.2, where the MOC system has been shown to assist perception of speech in noise through a process of release from adaptation, the current thesis proposes that an essentially similar process may underpin our robustness to the effects of reverberation in speech identification.

It should be noted that no particular claim is made at this stage about the *origin* of any signal controlling the MOC efferents, only that the resultant efferent activation might be involved in the process of perceptual compensation for reverberation. Indeed, if it can be assumed that the conversational sound levels studied in the majority of the reverberant speech studies are insufficient to trigger the acoustic reflex, then, as Figure 3.7 shows, the MOC efferents provide the main pathway through which the auditory system might be re-calibrated to adjust the peripheral encoding of sound. As was seen in Figure 3.1b at the start of this chapter, however, descending inputs to the olivocochlear pathway arise from both the inferior colliculus and the auditory cortex (Brown, 2011). It still remains a possibility, therefore, that the effects seen in perceptual compensation for reverberation might be explained only with reference to these more central layers of the auditory system.

3.5 State of the art in efferent auditory models

Early computational models of the auditory system (e.g., Ghitza, 1986; Lyon, 1982; Seneff, 1988) benefitted from plentiful physiological evidence on afferent processing in the peripheral auditory system. Such models were of interest to the community since they provided a representation of sound that was thought to be more faithful to auditory processing than MFCCs or PLPs, and at times could improve recognition in difficult listening conditions (Stern, 2011; Stern and Morgan, 2012). However, these models were computationally expensive and produced output features in a format that required further treatment before use in an HMM-based recogniser[1].

Later auditory models benefitted from computing efficiencies, and from methods designed to address the mismatch with statistical assumptions of the recognition back-end (as discussed in Stern, 2011). Computational models began to examine effects of longer-term temporal evolution (e.g., the modulation spectrum of Kingsbury et al., 1998) and to predict speech intelligibility (cf. § 2.1.7 where different spectral regions were viewed as containing complimentary information). Other researchers were inspired by afferent effects observed in more central auditory areas (e.g., Chi et al., 2005, modelled the spectro-temporal response fields of neurons in the auditory cortex). When used as a front-end for a speech recogniser, these representations again generally resulted in improved performance for conditions in

[1] Auditory features are typically computed at a high data rate, and multiply the data by a factor equal to the number of frequency channels. Moreover, neighbouring frequency areas are highly correlated in auditory representations, requiring use of a discrete cosine transform or similar to decorrelate the channels.

which the signal was degraded by noise (e.g., Kim et al., 1999; Kleinschmidt et al., 2001; Tchorz and Kollmeier, 1999).

Compared to the mass of physiological evidence on afferent pathways in the auditory system, relatively little physiological data exists regarding efferent processing. Perhaps as a result, descending pathways have often been completely absent from the auditory models developed. Four notable exceptions which *include* efferent processing, based on work by Giguère and Woodland (1994), Goldstein (1990), Zilany and Bruce (2006) and Ferry and Meddis (2007), are introduced in the remainder of this chapter.

While the afferent pathways of auditory models are tuned to quickly emphasise change in the environment, the efferent system is comparatively sluggish: so-called 'fast' effects occur over tens of milliseconds, but 'slow' effects unfold over tens of seconds (Sridhar et al., 1997). In other words, the time-scales over which afferent models of the auditory periphery may be regarded as adaptive are relatively short-term (in the region of 50 ms), offering little ability to compensate for slowly varying changes in the environment or to deal with the longer-term effects of reverberation. On the other hand, the efferent-inspired models described below appear to be able to make use of the longer-term contextual information in the signal, for instance by monitoring the AN response internally. Although configuration details differ between research groups, each model implements efferent suppression by means of a gain-control mechanism so that MOC feedback (directly or indirectly) decreases the effect that an input signal would ordinarily have on BM motion. In turn, the reduction in BM motion results in a change in activity recorded at the AN stage of the model. This may allow a degree of disambiguation of AN responses, effectively mimicking the human ability to suppress irrelevant information in speech signals (Hermansky, 1998).

Though they have not been tested with reverberant signals, the efferent models discussed below claim to improve machine listening in *noisy* environments (Brown et al., 2010; Chintanpalli et al., 2012; Lee et al., 2011), much as was observed in physiological data (cf. § 3.4.2). Rather than modelling physiological or psychoacoustic data directly, two recent studies follow a physics-based approach to understanding cochlear function. The description of non-linear amplification by Reichenbach and Hudspeth (2014) was introduced above (cf. Figure 3.3). Additionally, Gomez et al. (2014) are researching clinical applications of active cochlear modelling, and aim to inspire the next-generation of cochlear implants.

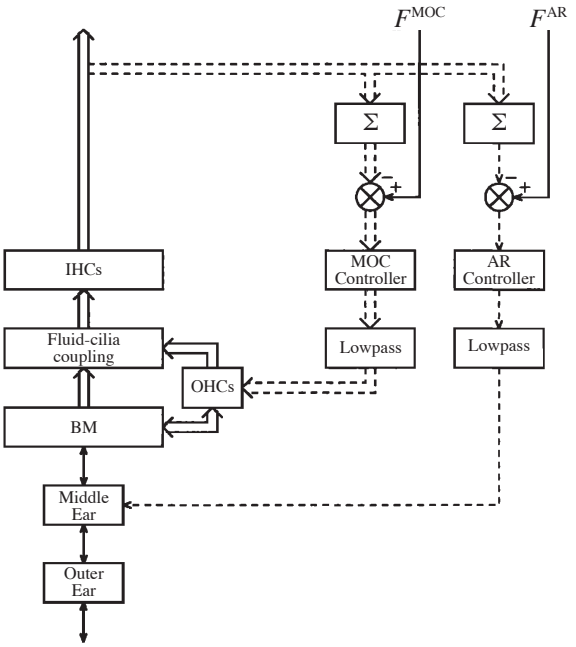


Figure 3.9: Auditory model comprising afferent pathways (*solid* lines) and two efferent pathways descending to the periphery (*dashed*), redrawn from Giguère and Woodland (1994). The afferent pathway comprises modules for the outer ear, middle ear, basilar membrane (BM) and its coupling with the inner hair cells (IHCs). The inner efferent pathway models the effect of medial olivocochlear (MOC) efferents on the outer hair cells (OHCs), and the outer efferent pathway represents the acoustic reflex (AR). Fixed model parameters F^{MOC} and F^{AR} represent target firing rates of narrowband MOC and wideband AR efferent signals respectively.

3.5.1 Giguère and Woodland

Giguère and Woodland (1994) model both of the main feedback systems operating in the auditory periphery (cf. Figure 3.7 above): the acoustic reflex of the middle-ear and the efferent innervation of the inner ear. To simulate the effects of MOC feedback, their model inserts a component representing the fluid-cilia coupling between the BM and IHC, and controls this with an OHC model as shown in Figure 3.9. The stated objective of the efferent feedback circuits is to regulate the average (afferent) AN firing rate. This is achieved in the model by monitoring the *average* firing rate (across all channels for the AR circuit, and within particular frequency bands for the MOC circuit), however, Giguère and Woodland suggest that at higher frequencies a measure of *synchronous* firing activity may be a more appropriate control structure.

This model later formed the basis of ASR experiments carried out by Giguère et al. (1997). Simulations of normal and impaired hearing were realised by altering the model parameters, and the resultant AN representation was used as a front-end for a neural-network recogniser. Modelling peripheral hearing impairment lowered the speech recognition accuracy in a manner that was broadly in line with human data.

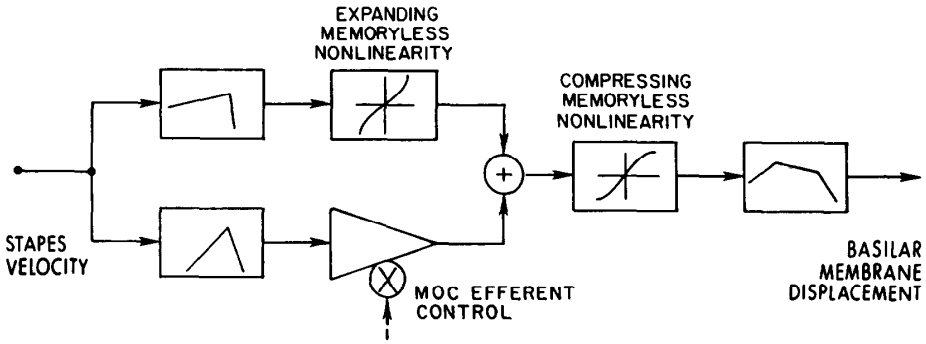


Figure 3.10: Multiple band-pass non-linearity (MBPNL) model, redrawn from Goldstein (1990). The cochlear model is developed further in Ghitza (2007), Messing (2007) and Messing et al. (2009), and is used by Lee et al. (2011) for noise-robust ASR experiments.

3.5.2 Goldstein

Figure 3.10 pictures the multiple band-pass non-linearity (MBPNL) model of cochlear mechanics by Goldstein (1990), in which stapes velocity is converted into BM displacement. Like the DRNL filterbank, this model also comprises two pathways which are summed to represent the tuning curves of the BM. The upper path represents the overall broadband response of the BM tuning curves by using an expanding function followed by its inverse compressive function, thereby generating a linear filter response. The lower path represents the sensitive compressive non-linear tip of the BM tuning curves, comprising a compressive nonlinear filter whose gain may be altered to mimic the efferent-induced suppressive effects on the BM that occur in the presence of noise.

The MBPNL model by Goldstein (1990) lies at the heart of the efferent-inspired studies undertaken by Ghitza (2007), Messing (2007) and Messing et al. (2009). Aiming to predict human performance in diphone discrimination tasks in background noise, the MBPNL is embedded in a closed-loop model of the auditory periphery in such a way that the parameters controlling the efferent response are determined, per channel, based on the amount of (sustained) noise present in the signal. In this simulation, the MOC adjustment affects the gain in the MBPNL channel: low levels of noise in the input signal have little suppressive effect, while high levels of noise in the input signal bring about a larger efferent suppression and reduce the nonlinear amplification of low-amplitude sounds as a result. In this way the efferent processing effectively normalises the input sound level, ensuring that the output of the model always lies within an appropriate dynamic range window.

While earlier works focused on modelling psychoacoustic responses in particular listening tasks (e.g. Messing et al., 2009, examined confusions among word-

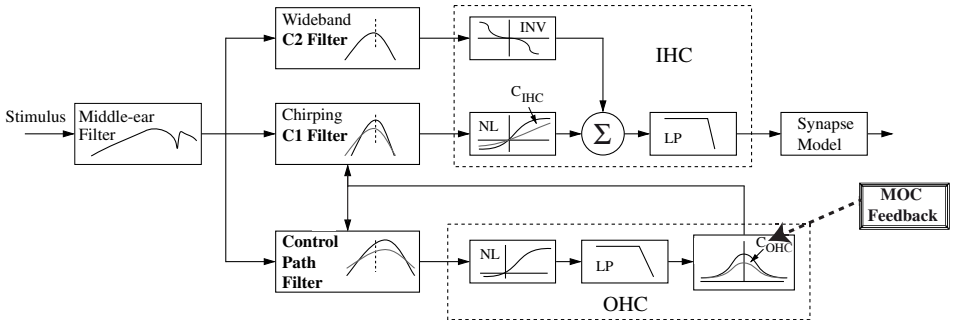


Figure 3.11: A model of peripheral processing by Zilany and Bruce (2006). MOC feedback affects the OHC behaviour in the lower pathway, which in turn adjusts IHC transduction at the AN synapse. The model has also been used to match psychophysical data in studies by Jennings et al. (2011) and Chintanpalli et al. (2012), and has received further parameter tuning in Zilany et al. (2009, 2014).

initial consonants when corrupted by speech-shaped additive Gaussian noise), recent studies have also applied this model in ASR experiments. Lee et al. (2011) used the model of Messing et al. (2009) to show that efferent processing could improve word accuracy when speech was presented in diverse noise backgrounds. Five types of noise were examined, three of which were stationary (white, pink, speech-shaped), and the remainder (train, subway) represented real-world listening environments.

3.5.3 Zilany and Bruce

Figure 3.11 shows an alternative model of peripheral processing by Zilany and Bruce (2006), which again allows the main MOC efferent effect to be modelled with a reduction of the nonlinear cochlear response. Previous versions of the model were tuned in order to match (feline) physiological data captured in response to tones, tone-complexes, broadband noise and vowels (Bruce et al., 2003; Carney, 1993; Zhang et al., 2001). The model has more recently been used to match a number of other psychophysical datasets. Zilany et al. (2009) simulated low, medium and high spontaneous rate fibres, and showed firing rate adaptation effects as discussed in § 3.2.3 above. More recently, parameters describing mechanical-to-neural transduction at the IHC–AN synapse have undergone fine-tuning in Zilany et al. (2014) in order to better represent the data published by Liberman (1978).

The Zilany et al. (2009) model formed the basis of a study undertaken by Jennings et al. (2011) in which model parameters were tuned to match perceptual data showing a longer-term temporal effect sometimes referred to as ‘overshoot’. Here, under certain conditions, a preceding sound can improve the detection of a masked tone (Fletcher et al., 2013). Chintanpalli et al. (2012) also extended the model of

Zilany et al. (2009) to query the involvement of the MOC reflex in noise-based listening. In agreement with the MOC unmasking effect outlined in § 3.4.2 above, their study reported that manual reduction of the OHC gain parameter (simulating efferent suppression) re-mapped the dynamic range of simulated AN fibres, and allowed discrimination of tones-in-noise over a wide range of background noise levels.

3.5.4 Ferry and Meddis

The dual-resonance nonlinear (DRNL) model of cochlear filtering (introduced above in Figure 3.4) underpins another family of efferent modelling studies. The original DRNL filter consists of two parallel paths, one linear and one nonlinear, which together describe the ascending processing pathway which transforms stapes displacement to BM vibration (Lopez-Poveda and Meddis, 2001; Meddis, 2006; Meddis et al., 2001). To simulate the effects of efferent processing, Ferry and Meddis (2007) incorporated an attenuator at the start of the nonlinear pathway of the DRNL, as shown in Figure 3.4b. Their paper simulates empirical studies carried out at different laboratories in order to confirm that the model fits physiological data recorded on the BM and in the AN.

Ferry and Meddis (2007) state that the aim of the model, in the longer term, is to study the role of the efferent system in processing complex stimuli such as music and speech, with and without noise. Some of these conditions, and others besides, have since been examined by embedding the efferent-DRNL in computational models of auditory function which variously simulate human response data or investigate environmental robustness in ASR (e.g., Beeston and Brown, 2010; Brown et al., 2010; Clark et al., 2012; Meddis et al., 2013).

In Brown et al. (2010), the model of Ferry and Meddis (2007) was used as a front-end for speech-in-noise recognition experiments. The MOC attenuation parameter was varied manually (in an open-loop configuration), and was fixed across all frequency channels. Similar to the study of Messing et al. (2009) discussed above, optimal recognition scores were achieved when the attenuation parameter was proportional to the amount of noise present in the signal. Clark et al. (2012) reported further ASR improvements when MOC feedback was derived from a measure of firing activity in the simulated AN, and was adjusted separately in each channel. In the latest revision, Meddis et al. (2013) describe a closed-loop model with frequency- and time-dependent circuits implementing both the MOC and AR pathways. In this version of the model, the MOC signal is derived from the high SR fibres of the AN response after processing in the cochlear nucleus. The AR, on the other hand, is derived after a second-level brainstem response which receives input from the AN's low SR fibres.

Beeston and Brown (2010) also used the efferent-inspired model of Ferry and Meddis (2007), but diverged from the approach taken in Brown et al. (2010) above. The efferent-DRNL filterbank was embedded in a closed-loop model of auditory processing, in which the value of efferent attenuation was derived on the fly in response to the dynamic range of the input signal's simulated AN response. Further investigation into candidate metrics to control the efferent feedback mechanism in reverberant speech-based tasks was also presented in Beeston and Brown (2013).

3.5.5 Efferent model choice

The four state-of-the-art efferent auditory models discussed above are broadly equivalent in that they all aim to model the role of efferent suppression on the response of the basilar membrane to an incoming sound signal. Since its publication in 2007, the Ferry and Meddis model has become well-established in auditory modelling studies, in part because the model has been validated against a number of physiological data sets, and in part because a Matlab implementation of the model has been made available for use by other researchers. As a result, the Ferry and Meddis model was adopted to represent cochlear function in the computational work presented in Chapter 4 below, where the efferent-inspired auditory model is tested to see if it may replicate and help explain the effects of perceptual compensation for reverberation. It is noteworthy, however, that since the three models deriving from work by Giguère and Woodland (1994), Goldstein (1990) and Zilany and Bruce (2006) perform essentially the same task, they have potential to act as substitutes for the efferent-DRNL component in this regard.

Chapter summary

To allow an examination of perceptual compensation for reverberation from a computational modelling perspective in the chapter which follows, Chapter 3 first reviewed existing works which either shed light on auditory processing relevant to reverberant speech identification, or described computational models capable of simulating such auditory processes. This chapter therefore discussed biological and computational auditory systems (§ 3.1) and reviewed the process by which the peripheral auditory system transforms acoustic vibrations into electrical messages which the brain interprets (§ 3.2, § 3.3). Crucially, this encoding of sound is regulated by a series of efferent pathways which descend from higher auditory areas and ultimately confer robustness in adverse conditions (§ 3.4).

Two efferent pathways reach directly back into the auditory periphery: the acoustic reflex to the middle ear, and the medial olivocochlear (MOC) innervation to

the inner ear. Focussing on the latter of these mechanisms, both physiological and computational studies suggested that MOC neurons reduce cochlear sensitivity by attenuating the nonlinear amplification of the basilar membrane. Activation of this pathway appears to adjust the process of neural transduction by the inner hair cells and, by a process known as MOC unmasking, improves the effective dynamic range of signals encoded in the auditory nerve. Further, a number of studies using state-of-the-art efferent-inspired auditory models as front-ends for speech recognisers suggest that this enhancement in signal representation can improve recognition accuracy for speech in noise (§ 3.5).

By considering the acoustical effects of late reverberation in terms of an increased noise-floor and reduced dynamic range, this chapter additionally set out the relevance of MOC unmasking to perceptual compensation for reverberation. The physiological mechanisms conferring human robustness to the effects of reverberation on speech identification have not yet been confirmed, and it is as yet too early to know whether the evidence gathered so far regarding reverberation-robust processing in the central auditory system will turn out to explain this phenomenon. Nonetheless, it appears that MOC efferents are likely to be implicated in the process of recalibrating the sound-encoding at the cochlea in reverberant listening environments, much as they are for situations in which background noise is a factor.

Chapter 4

Modelling perceptual compensation for the effects of reverberation[1]

Contents

[1] An early version of this model was first published in Beeston and Brown (2010). The reverberation metric controlling the efferent feedback loop was investigated further in Beeston and Brown (2013) resulting in the model described here and in Beeston and Brown (2014). Additionally, the reverberation estimation technique described in § 4.3.2 was used in reverberant speech recognition experiments by Kallasjoki et al. (2014).

4.1 Introduction

Motivated by the main areas of research just described, this chapter presents a model in which auditory efferent suppression is used as a candidate theory for explaining the effects of perceptual compensation for reverberation. Human auditory processing seems to be underlain by a set of constancy mechanisms which result in robust speech perception even when listening conditions are degraded by noise or reverberation. As a result, state-of-the-art techniques in automatic speech recognition (ASR) have sometimes incorporated components that were either loosely or more directly inspired by aspects of biological audition. This strategy produced substantial gains in noisy speech recognition, but to date, has not yet solved the challenge that reverberation poses to machine listeners.

The computer model presented below is an extension of that proposed by Ferry and Meddis (2007), in which efferent suppression regulates activity in the afferent pathway (§ 4.2). Previously used to simulate data that displayed increased noise-robustness with increased efferent activity (Brown et al., 2010; Clark et al., 2012), the model is examined in the current chapter to see if efferent processing may help explain the relative robustness to the effects of reverberation that is observed with human listeners. Palomäki et al. (2004) previously reported an approach to speech recognition in which reverberation may be treated similarly to noise; their insight underpins the current work, where the Ferry and Meddis model is applied to the task of simulating perceptual compensation for the effects of reverberation.

Here, the effect of the late reverberation is considered to be similar to the effect of additive noise, in that it reduces the dynamic range of the signal and increases its noise floor (as was previously depicted in Figure 2.2). Since the efferent system is known to be involved with the regulation of dynamic range in such noisy listening conditions (see e.g., Guinan, 2006, and § 3.4 earlier), it follows that the efferent system might also be involved in suppressing the auditory nerve response to late-arriving reflections. This is the key proposal which is implemented in the model.

Further, this chapter describes how the model of Ferry and Meddis may be embedded within a feedback mechanism which monitors the level of reverberation in the environment and adjusts the sound-encoding behaviour of the simulated auditory nerve response accordingly (§ 4.3). To see whether the model can indeed simulate and help explain the effects of perceptual compensation for reverberation, the model is evaluated against human listener data published by Watkins (2005a). In addition to modelling his general 'sir-stir' category boundary paradigm (§ 4.4), the model is tested in further experiments with time-reversed speech (§ 4.5) and time-reversed reverberation (§ 4.6). The final section of the chapter then discusses the proposed involvement of efferent processing in compensation for reverberation more widely, and highlights a number of areas where psychophysical research may help to answer questions arising during the modelling process.

4.1.1 Research questions

This chapter presents a computational model of auditory function in which the simulated auditory nerve response in the afferent pathway is monitored, and at times attenuated, by an efferent feedback control mechanism. The manner in which this efferent control is implemented is therefore of primary concern. Since human listeners make use of the temporal context immediately prior to a test sound, the auditory model follows a similar principle: a time-windowed segment of the simulated auditory nerve signal is observed, and used *somehow* to determine a value for efferent attenuation which is then applied in the model. Opinion appears divided in the literature over exactly *what* aspects of a signal may determine a listeners' ability to compensate for the reverberation present in a given test signal. Two putative measures of the perceptual influence of reverberation are examined in § 4.3, and are tested as controllers for the efferent feedback circuit.

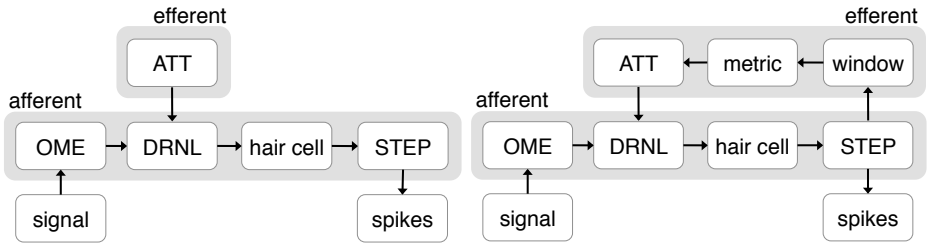
The first measure of reverberation concerns the dynamic range of the simulated auditory nerve signal, quantified by examining the signal's mean-to-peak ratio (MPR). Reverberation is typically understood to reduce dynamic range, as dips in a signal's temporal envelope are filled with reflected energy. One way to view this process is to note that the 'noise floor' of a signal appears to increase when it is reverberated. Since the efferent system is known to be active in noisy conditions,

it seems feasible therefore that reverberation may also give rise to efferent activity. In the model described below, an increase in reverberation is detected as a raised MPR value, and used in turn to trigger the efferent suppression.

A second measure that has been proposed to inform the compensation mechanism concentrates instead on the 'tails' that reverberation adds at offsets in a signal's temporal envelope. Here, a low-pass mask (LPM) signal measure is used whereby signal envelopes are smoothed in each auditory channel, and the signal's energetic content is examined during the envelope's downward-going slopes (i.e. during the reverberation tails). In the model, an increase in the LPM value is detected when the level of reverberation increases; this again corresponds to an increase in efferent suppression.

Employing these two measures of reverberation, Experiment M1 is a calibration exercise which asks whether measures of either the dynamic range or the reverberation tails of a signal could be appropriate candidates for controlling efferent suppression in a model of perceptual compensation for reverberation. A computer model is built to simulate the experimental findings of Watkins (2005a): it 'listens to' (or processes) a sound file and 'responds with' (or outputs) a 'sir' or 'stir' result. If the model is capable of simulating perceptual compensation for reverberation, then results should follow the pattern of human listener responses in Watkins' work. The influence of increased reverberation on the test-word is expected to reduce the number of 'stir' responses since the dip that cued the [t] becomes obscured by reflected energy protruding from the preceding context into the test-word region. However, when the level of reverberation on the context is also increased, then in human listeners the factors obscuring the [t] appear to be somewhat overcome, so that the number of 'stir' responses increases again.

Further aspects of the compensation mechanism are investigated in two following experiments. Human listener data in Watkins (2005a) indicates that compensation does not rely on phonetic perception, so the compensation effect should be maintained even when the time-direction of the speech signal in the preceding context is reversed. This is investigated in Experiment M2. Conversely, when the time-direction of the *reverberation* is reversed in the preceding context, it appears that compensation no longer occurs (Longworth-Reed et al., 2009; Watkins, 2005a). This finding is of particular interest since it is not consistent with the predictions of objective measures of reverberant speech perception. Time-reversal of reverberation is therefore investigated in Experiment M3.



(a) open-loop implementation **(b)** closed-loop implementation

Figure 4.1: Schematic depictions of the open-loop (fig. 4.1a) and closed-loop (fig. 4.1b) implementations of the computational auditory model which receives an audio signal at input and provides a simulation of firing activity in the auditory nerve (AN) at output. Afferent components comprise an outer and middle ear (OME) filter, a bank of dual-resonance nonlinear (DRNL) filters representing basilar membrane vibration, inner hair cell transduction and temporal integration as a spectro-temporal excitation pattern (STEP). Efferent attenuation, ATT, is controlled manually in the open-loop implementation, but is derived via a metric from a windowed portion of the STEP in the closed-loop model configuration.

4.2 Auditory model overview

Efferent feedback from the olivocochlear system, discussed earlier in Sections 3.4 and 3.5, has recently been introduced into the Meddis auditory model[1], the Goldstein model[2], and the model of Zilany and Bruce[3]. This approach has proved fruitful in modelling a number of tasks involving listening to speech in noise (e.g., Brown et al., 2010; Clark et al., 2012; Lee et al., 2011). The following experiments ask whether the efferent system may also be implicated in reverberant listening. The insight underlying this work is that efferent suppression might similarly reduce the auditory response when the noise floor is raised by reverberation in the signal. If so, then periods of low amplitude in the signal's temporal envelope which had been previously filled in by reflected energy might now be revealed once more by the application of efferent attenuation.

The linchpin of the current simulation[4] is the efferent DRNL cochlear model of Ferry and Meddis (2007), which extends the earlier model of basilar membrane

[1]Efferent processing was introduced by Ferry and Meddis (2007), and subsequently used in modelling studies by Beeston and Brown (2010); Brown et al. (2010); Clark et al. (2012).

[2]See for example Ghitza (2007); Lee et al. (2011); Messing et al. (2009).

[3]See Chintanpalli et al. (2012); Jennings et al. (2011).

[4]The starting code base for the efferent dual-resonance nonlinear (DRNL) component is the MAP (Matlab auditory periphery), version 1.6, as described in Meddis (2006). Described below in § 4.2.2, efferent suppression is simulated using the cochlear attenuation parameter in the non-linear path of the DRNL as identified by Ferry and Meddis (2007).

(BM) response proposed by Meddis et al. (2001) by adding an attenuator to the non-linear path of the DRNL. When embedded in an auditory model with further components representing hair cell transduction and firing activity in the auditory nerve as shown in Figure 4.1a, the effect of an increase in efferent attenuation is that the response in the auditory nerve is suppressed (i.e., fewer spikes per second at output).

Figure 4.1b presents a schematic overview of the computational auditory model described in this chapter. The central block of the model consists of two major parts, and is an extension of the computer model proposed by Ferry and Meddis (2007), in which an efferent feedback loop monitors and regulates activity in the afferent pathway. The afferent pathway is described fully in § 4.2.1. Here, a simulation of auditory nerve (AN) firing rate is obtained by passing the acoustic signal through an outer and middle ear (OME) filter, a bank of DRNL cochlear filters and a model of inner hair cell function. The simulated AN activity in each channel of the model is temporally integrated to give a spectro-temporal excitation pattern (STEP). In the efferent pathway, discussed below in § 4.2.2, an estimate of the amount of reverberation present in the signal is calculated over a windowed portion of the STEP, using a metric based either on dynamic range or on reverberation tails. This estimate is finally used to control the amount of efferent attenuation applied, ATT.

4.2.1 Afferent pathway

To understand the progression of an audio stimulus through the afferent auditory pathway, Figure 4.2 presents a visualisation of a male voice speaking the word "stir" as it would appear after each stage of the peripheral auditory processing. Each element was previously introduced in § 3.2.1, and is discussed in turn below with the implementation details of how it is used in the computational model for investigating perceptual compensation for reverberation.

i. Outer and middle ear (OME)

The outer and middle ear (OME) transfer sound pressure variation to the oval window while filtering the signal's spectrum to aid sound localisation and strengthen the frequency-regions likely to contain speech. The OME can be modelled with a single linear filter that provides a broad resonance around 2.5 kHz that imitates the ear canal effect, with a subsidiary resonance around 5.5 kHz that corresponds to the external ear (cf. Pickles, 1988, fig. 2.3 and fig. 2.6). Here, two filters were combined in series to imitate each step in turn: the first filter characterises the outer

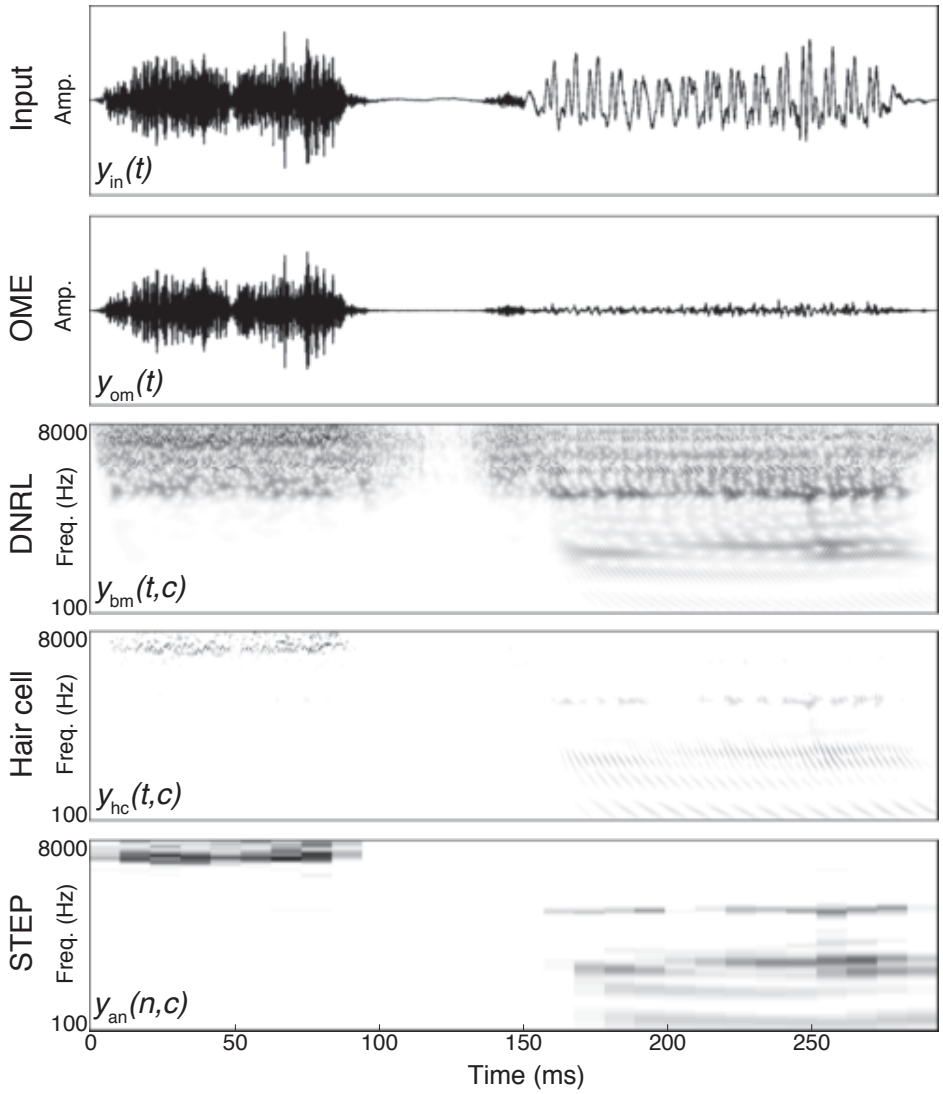


Figure 4.2: Afferent processing pathway (with efferent attenuation fixed at 0 dB). A single word ('stir' spoken by a male voice) is shown progressing through the model *from top to bottom* as: input signal $y_{\mathrm{in}}(t)$ indexed by the time sample t; stapes displacement after the outer and middle ear (OME) filter, $y_{\mathrm{om}}(t)$; as a basilar membrane (BM) representation $y_{\mathrm{bm}}(t, c)$ after the dual-resonance nonlinear (DRNL) filterbank, additionally indexed by the channel number c with centre frequencies log-spaced in the range between $c_1 = 100$ Hz and $c_{80} = 8$ kHz; as a hair cell response, $y_{\mathrm{hc}}(t, c)$; and as a spectro-temporal excitation pattern (STEP) response, now indexed by the time frame n, simulating firing activity in the auditory nerve (AN), $y_{\mathrm{an}}(n, c)$.

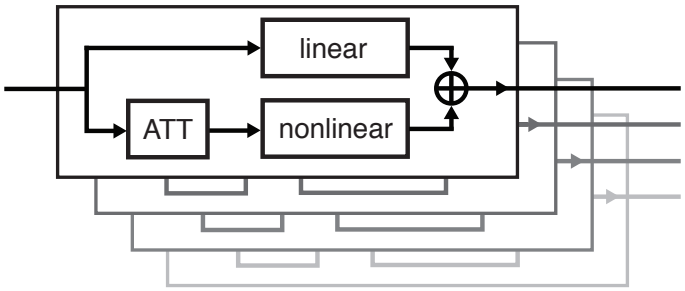


Figure 4.3: Bank of efferent DRNL filter components (cf. Figure 3.4b). The current model uses 80 filters (only four of these are shown), with centre frequencies log-spaced in the range between 100 Hz and 8 kHz. Efferent attenuation (ATT) is applied in each channel at the start of the nonlinear pathway as described by Ferry and Meddis (2007).

ear, and the second, the eardrum-pressure to stapes-velocity transformation in the middle ear (Lopez-Poveda and Meddis, 2001).

The stapes displacement, $y_{\mathrm{om}}(t)$, resulting after OME filtering of input waveform, $y_{\mathrm{in}}(t)$, may be examined by comparing the first and second panels of Figure 4.2. Here, the initial [s] and [t] of the spoken word 'stir' remain strong while the subsequent voiced part, the vowel [ɜ˞], appears attenuated in comparison since it is comprised predominantly of energy in lower frequency regions than the two filters' peak frequencies.

ii. Inner ear mechanics

As was shown previously in Figure 3.4a, a single dual-resonance nonlinear (DRNL) filter models the frequency analysis performed at a particular location in the cochlea. The two pathways – one of which is strictly linear, and the other, nonlinear owing to the inclusion of a static 'broken-stick' nonlinearity – are added together at output, and together introduce level-dependent filtering behaviour. Further, Figure 3.4b depicted the innovation introduced by Ferry and Meddis (2007), where an attenuator at the start of the nonlinear pathway models the effect of efferent suppression.

Since a single DRNL filter can represent the vibration resulting at a particular place along the basilar membrane (BM) in response to an input stimulus (Lopez-Poveda and Meddis, 2001; Meddis et al., 2001), a *bank* of efferent-input DRNL filters thus acts on the OME-filtered signal, $y_{\mathrm{om}}(t)$, to represent the whole BM in the current model. Configured to represent responses of a human listener (Meddis, 2006), the filterbank in the current model uses $C = 80$ filters, with centre frequencies, c, log-spaced in the range between $c_1 = 100$ Hz and $c_{80} = 8$ kHz. The resulting

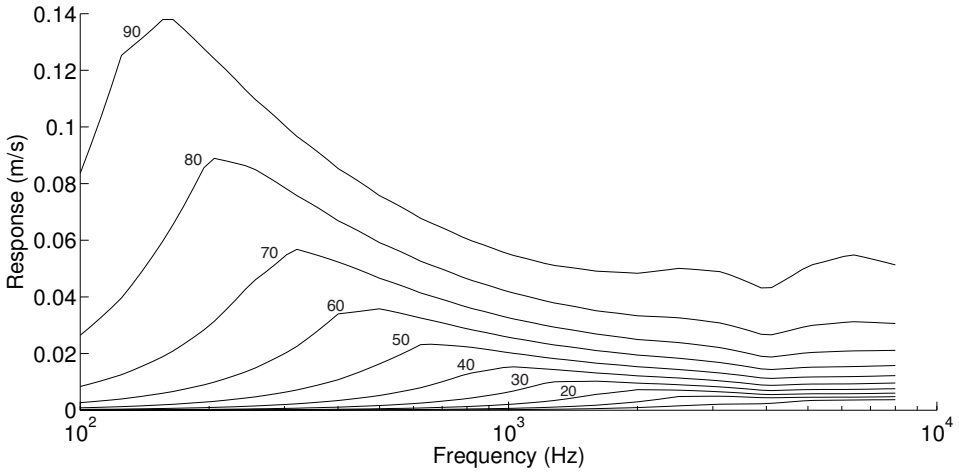


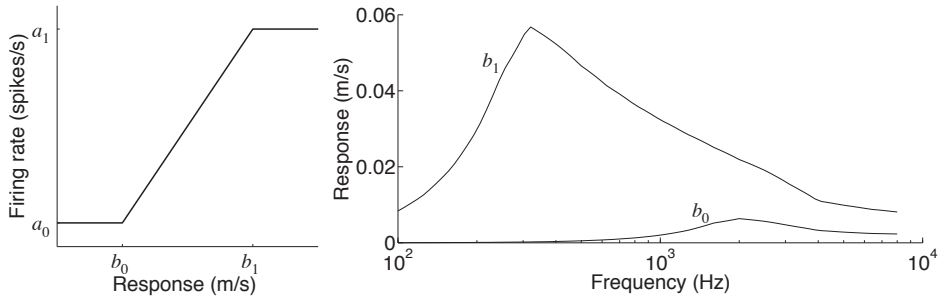
Figure 4.4: Iso-intensity contours for the basilar membrane velocity response, $y_{\mathrm{bm}}(t, c)$, to input signal levels of 0 to 90 dB SPL (in steps of 10 dB).

representation, $y_{\mathrm{bm}}(t, c)$, models the frequency-dependent and level-dependent response of the BM, and is depicted for the word 'stir' in the third panel of Figure 4.2.

The frequency-dependent and level-dependent response of the BM is examined more closely in Figure 4.4. Here, iso-intensity contours are plotted for input signals of 0 to 90 dB SPL (in steps of 10 dB). It is evident that auditory regions corresponding to speech are again favoured, and that the frequency producing the maximal BM response clearly varies with sound level. Responses were first calculated for pure tone test signals in single channels whose frequencies matched values defined in the international standard for equal loudness contours (ISO226, 2003). From there, an interpolation function derives intermediate response values (at each signal level) for each of the $C = 80$ channels used in the full auditory model. After this interpolation, channels are now log-spaced in the region $c_1 = 100$ Hz to $C_{80} = 8$ kHz as required.

iii. Hair cell transduction

The current study models transduction by inner hair cells with a simple scheme similar to that described by Messing (2007). The resulting hair cell representation is displayed in the fourth panel of Figure 4.2. Representing the fact that hair cells fire when bending in one direction only, the output of the DRNL was first half-wave rectified, giving $y\prime_{bm}(t, c)$. Secondly, the break down of phase-locking at high frequencies was simulated by lowpass filtering the half-wave rectified out-



(a) Rate-limiting function **(b)** $y\prime\prime_{bm}(t, c)$ response at threshold b_0 and saturation b_1

Figure 4.5: (a) Input-output function for the simple rate-limiter hair cell model described in Equation 4.1, redrawn from Messing (2007). The dynamic range of the signal is compressed by clipping the inner hair cell module's output to lie within a specified set of values between a lower limit (spontaneous rate, a_0) and upper limit (saturation rate, a_1). (b) Response levels of $y\prime\prime_{bm}(t, c)$ selected as threshold and saturation values at $b_0 = 20$ dB SPL and $b_1 = 70$ dB SPL pure tones respectively.

put to a cutoff frequency of 3 kHz by a second order Butterworth filter[1], giving $y\prime\prime_{bm}(t, c)$. Finally, the resulting response was mapped to a representation of hair cell activity, $y_{\mathrm{hc}}(t, c)$, by applying a channel-dependent rate-limiting function as shown in Figure 4.5a, where

$$y_{\mathrm{hc}}(t,c) = \begin{cases} a_0 & \text{if } y\prime\prime_{bm}(t,c) < b_0 \\ a_1 & \text{if } y\prime\prime_{bm}(t,c) > b_1 \\ a_0 + (a_1 - a_0)\frac{y\prime\prime_{bm}(t,c) - b_0}{b_1 - b_0} & \text{otherwise.} \end{cases} \tag{4.1}$$

The spontaneous and maximum firing rates were fixed throughout all channels of the model at $a_0 = 0.5$ spikes/s and $a_1 = 250$ spikes/s, respectively. The BM velocity parameters, b, varied channel-by-channel[2]. To approximate the response of a typical *low*-spontaneous-rate auditory nerve (AN) fibre (cf. Moore, 2004, fig. 1.17), the threshold, and saturation BM velocities were mapped to occur at around $b_0 = 20$ and $b_1 = 70$ dB SPL respectively, as shown in Figure 4.5b. This results in the fibre having a fairly large dynamic range of around 50 dB[3].

The resulting level-dependent response is depicted across a population of such hair cell fibres in Figure 4.6, whose centre frequencies span the log-spaced region of

[1] Messing's model uses a Johnson filter whose bandwidth increases with centre frequency.

[2] This is a significant departure from the model presented in Beeston and Brown (2010), where BM parameters were selected based on the response of a single channel in the 1 kHz region. This point is discussed further in § 4.6.4 below.

[3] This contrasts the model implementation of Meddis et al. (2013) in which the response of *high*-spontaneous-rate fibres are implicated in MOC efferent feedback, and low-spontaneous-rate fibres are instead used to control the middle ear muscle and the trigger the acoustic reflex.

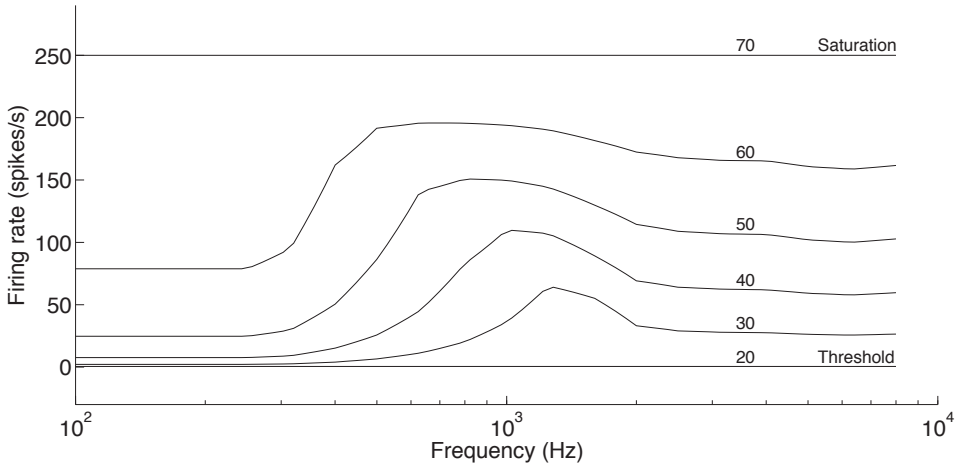


Figure 4.6: Iso-intensity contours are depicted for the hair cell population response, $y_{\mathrm{hc}}(t, c)$, to 10 dB stepped input signals between threshold level (defined as $a_0 = 0.5$ spikes/s when $b_0 = 20$ dB SPL) and saturation level (defined as $a_1 = 250$ spikes/s when $b_1 = 70$ dB SPL). Here, a population of nerve fibres is simulated covering the hearing range of the model from 100 Hz to 8 kHz, and the response is shown for a test tone occurring in the centre-frequency of each channel. The frequency region of maximum response can be seen to vary with the level of the input signal.

100 Hz to 8 kHz used in the auditory model. In each channel, the response to a test tone matching the channel's centre frequency is seen to be $a_0 = 0.5$ spikes per second for the $b_0 = 20$ dB SPL contour. The number of spikes per second gradually increases as the signal level increases until saturation ($a_1 = 250$ spikes per second) is reached for the $b_1 = 70$ dB SPL contour. For the 50 dB range in which the simulated low-spontaneous rate fibres respond, the frequency region for maximum response varies in line with the BM response characteristics (cf. Figure 4.4).

The iso-intensity contours in Figure 4.7 describe the characteristic of a single fibre at levels above the threshold. Here, the response in a single auditory channel centred at 2.1 kHz is shown for test-tones at equal sound levels which rove in frequency above and below the channel centre frequency. The response is clearly strongest when the test-tone is aligned with the channel centre-frequency. However, a fairly strong response can still be achieved for frequencies several hundred Hz above or below this value provided that the signal level is great enough.

iv. Temporal integration in the auditory nerve (AN)

The final stage of the afferent processing chain simulates the auditory nerve (AN) reponse, $y_{\mathrm{an}}(n, c)$, from the hair cell response, $y_{\mathrm{hc}}(t, c)$ by a process of tempo-

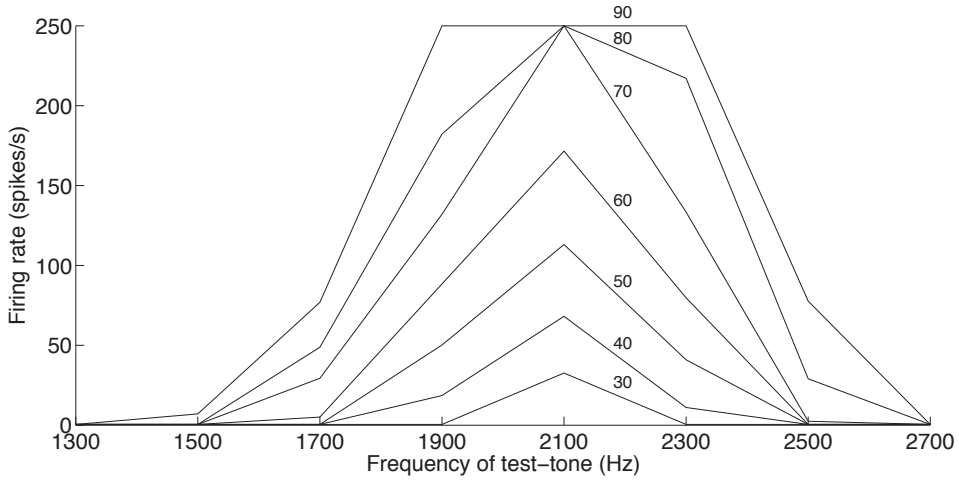


Figure 4.7: Response in a single-channel centred at 2.1 kHz when a test-tone varies (in steps of 200 Hz) below and above this frequency. Following data from Rose et al. (1971), additionally reported in Pickles (1988, Figure 4.7) and Moore (2004, Figure 1.16), the frequency axis is here plotted linearly. Again, iso-intensity contours are plotted for 10 dB steps, here from 30 to 90 dB SPL. The frequency-selectivity is again level dependent: at low levels, the channel responds most strongly to tones at its own centre-frequency. As the level of the test tone increases, the channel tuning broadens and provides a strong response even for signals of several hundred Hz difference.

ral integration. The spectro-temporal excitation pattern (STEP) which results is depicted in the lowest panel in Figure 4.2 for the spoken test-word 'stir'.

Here, a number of neighbouring time-steps are summarised together to provide a feature representation of the input signal for speech recognition. The temporal resolution is reduced from the original 48 kHz sample rate used in early stages of the model, to a much lower output-rate which is set here[1] at 100 Hz, i.e. a new output frame is calculated every 10 ms.

Temporal integration is achieved in the current model using a raised cosine window, defined as

$$w_{\mathrm{rc}}(k) = \begin{cases} 1 + cos(\frac{2\pi k}{K} + \pi) & \text{for } 1 \leq k \leq K \\ 0 & \text{otherwise,} \end{cases} \tag{4.2}$$

[1] This differs from the temporal integration stage used in Beeston and Brown (2010): (i) a raised cosine window now replaces the Hann window of the previous work in order that neighbouring frames sum to one; (ii) the output frame rate is halved (and window length correspondingly doubled) for compatibility with standard speech recognition software such as the Hidden Markov Model Toolkit (HTK, Young et al., 2009).

where k counts the sample index within a window, up to the maximum $K = 960$ samples, which corresponds to the 20 ms window size used in the current model. Windows had a hop-size of $K/2$ samples (i.e. overlap at 50%) which provided a new output frame every 10 ms (i.e. at 100 Hz), and ensured that neighbouring windows summed to one. These raised cosine windows were applied in a channel-by-channel manner, $w_{\mathrm{rc}}(k, c)$, to the hair cell response in order to simulate temporal integration and obtain the firing rate in the auditory nerve

$$y_{\mathrm{an}}(n, c) = \frac{2}{K} \sum_{k=1}^{K} w_{\mathrm{rc}}(k, c)\; y_{hc}(\frac{nK}{2} + k, c) \tag{4.3}$$

where n indexes the time frame in the STEP which results. By this process, the temporally integrated signal was proportional to the mean of the input hair cell signal in each channel, but was additionally scaled such that the maximum value of the output $y_{\mathrm{an}}(n, c)$ was equal to the maximum value of the input $y_{\mathrm{hc}}(t, c)$.

4.2.2 Efferent pathway

The human auditory system, as yet incompletely understood, was overviewed in Chapter 3. The output of the afferent pathway is a representation of instantaneous firing rate in the auditory nerve, $y_{\mathrm{an}}(n, c)$, where n indexes the time frame and c the frequency channel. The current chapter introduces efferent feedback to the periphery by simulating activity of the olivocochlear system (see § 3.4.1) which regulates the behaviour of the afferent pathway by attenuating the input to the nonlinear pathway of the DRNL (see Figure 4.1a).

The role of the efferent pathway in the model can be illustrated by considering the effect of efferent attenuation on the rate-level response curve shown in Figure 4.8. Here, the auditory response to a spoken stimulus[1] presented at varying input signal levels has been computed in a 7-channel model, with 1 channel per octave across the whole hearing range. For a given presentation level and attenuation value, the firing rate here reports the maximum response achieved in any of the 7-channels, when efferent attenuation was applied throughout the spoken phrase at a fixed level of either 0 dB or 10 dB. Efferent suppression turns down the gain on the nonlinear path of the DRNL, reducing its overall output and causing the rate-level curve to shift toward higher sound pressure levels. This causes the response to stimuli with low sound levels (which lie on the toe of the curve) to fall below threshold.

[1] The spoken stimulus is the same 'sir-stir' reference file later used for calibration of the input signal levels for the modelling task in § 4.4.2.

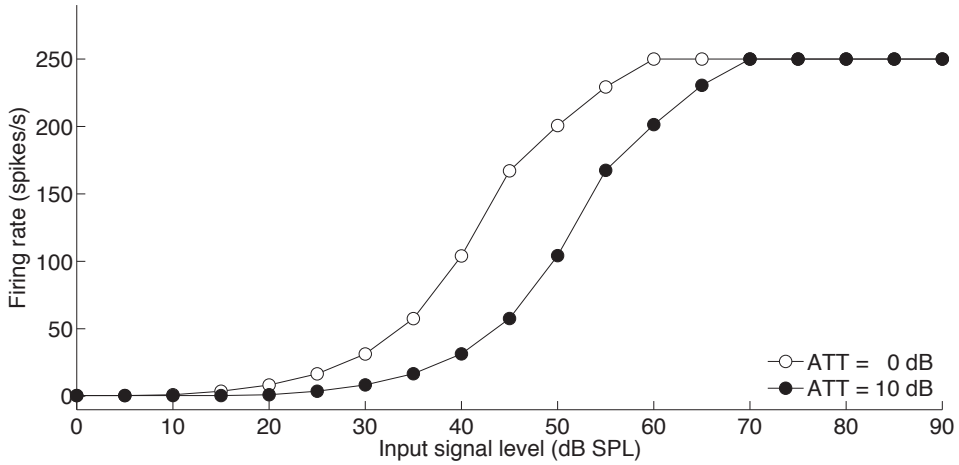


Figure 4.8: Multi-channel rate-response curve. For demonstration, the simulated auditory nerve firing response was computed in a 7-channel model (1 channel per octave) for a spoken reference stimulus presented at varying input signal levels, with efferent attenuation fixed at either 0 dB (white markers) or 10 dB (black markers). For a given presentation level and attenuation value, the firing rate here reports the maximum response achieved in any of the 7-channels. The response curve shifts to higher sound pressure levels as efferent attenuation is applied (here, 10 dB). For low signal presentation levels, this has the effect of reducing the firing response back to its threshold value.

The central hypothesis underlying the following modelling work is that the shift in the rate-level curve observed in Figure 4.8 will subdue the response to low-level reflections (from the late reverberation) sufficiently that spectro-temporal dips will re-appear in the signal. Since such periods of low amplitude are cues to the identification of stop consonants, an increase in efferent attenuation might thus bring about an improvement in the representation of reverberant stop consonants.

4.3 Control of efferent suppression

In the previous section, the consequence of manually applying efferent attenuation was demonstrated using a model based around the efferent-DRNL implementation of Ferry and Meddis (cf. § 4.2.2, and Figure 4.8 in particular). As shown in Figure 4.1b, the current model extends the work of Ferry and Meddis by introducing an ecosystemic control mechanism (Di Scipio, 2003) that determines the effectiveness of the efferent feedback loop. By varying the level of attenuation applied in the model in proportion to the amount of reverberation detected in the recent history of the acoustic surroundings, auditory efferent suppression is thus used as a candidate theory for explaining the effects of perceptual compensation for reverberation.

This section describes the conditions that such a control mechanism must satisfy if it is to allow simulation of perceptual compensation for the effects of reverberation that is observed in human speech perception. Two prospective metrics for reverberation estimation are first described below, each of which makes an assessment of the amount of reverberation present in the signal, and then (automatically) uses this value to update the value of the efferent attenuation applied in the model. The selected reverberation measures are subsequently used to model human listener data collected by Watkins (2005a) in a series of experiments below (Experiments M1, M2 and M3).

In an ideal world, the metric deriving control for efferent attenuation would be based on human physiology and function. While behavioural data gathered in psychoacoustic studies is beginning to elucidate various mechanisms underlying perceptual compensation for reverberation, relatively little is yet known about the physiological factors influencing these processes. Thus compensation for reverberation is not yet well-enough understood for a model to be biologically accurate. Rather, the current model makes a crude summarisation of known auditory processing in order to create a functional simulation of the compensation effect. Here, all higher levels of the auditory system are lumped together and treated as a 'black-box' which outputs a single efferent signal. This efferent signal acts on the peripheral auditory system and thereby regulates the afferent processing behaviour. This simplification is easily seen in Figure 3.1b; only the final descending pathway is modelled.

Functionally, then, the job of the feedback circuit is to make an assessment of the amount of reverberation present in the signal, and to use this value to automatically update the attenuation parameter in the DRNL filter bank. Two questions thus immediately arise. The first asks how reverberation should be quantified; the second queries the time-period over which this quantification should be done. These two areas are discussed in turn below.

It is generally accepted that the preceding context appears to inform a listener's decision about a subsequent test-word. It is not yet fully understood, however, what the nature of this influential information may be. One theory, termed 'modulation masking', was put forward by Nielsen and Dau (2010). This explanation suggests that human listeners may adapt to the degree of modulation present in the preceding context signal; thus it would appear that a measure of dynamic range might prove useful in modelling compensation for the effects of reverberation. One such measure, the mean-to-peak ratio (MPR) is described below. On the other hand, Watkins (2005a) argues that listeners are informed by mechanisms that detect and compensate for the 'reverberation tails' present in a signal. More recently, Watkins et al. (2011) have suggested further that this compensation mechanism may be informed by an assessment of temporal envelopes within individual auditory chan-

nels. The second measure described below, the low-pass mask (LPM) reverberation estimator is based on these principles. Many other approaches could of course be investigated to quantify reverberation in the auditory modelling task but are not investigated further in the current work. One that has proved successful in ASR involves modelling the excitation signal for voiced speech with the linear prediction residual (Ananthapadmanabha and Yegnanarayana, 1979), and examining the higher order statistics (i.e. the kurtosis) to quantify the 'peakiness' of the resulting probability distribution (e.g., Gillespie et al., 2001). Another approach that looks interesting from a biological point of view would be to examine the auditory *onset*- rather than offset- response (see e.g., Heil, 2003; Longworth-Reed et al., 2009).

The time course of the monaural compensation effect has yet to be studied in detail[1]. Since Watkins (2005a) has repeatedly demonstrated compensation effects using single utterances, however, it seems that we are interested in a fairly rapid mechanism. However, in the earlier discussion of compensation for reverberation it became apparent that various constancies might be active on different time-scales (cf. § 2.4). Indeed, looking across recent studies that examined the timescales on which binaural compensation effects operate, it appears that the relevant timescale for contextual information may depend critically on the listener task. Long-term learning (over around 5 hours of exposure to a particular room condition) can improve performance in listeners' localisation accuracy (Shinn-Cunningham, 2000). On the other hand, listeners' ability to determine the azimuth of a test pulse appears to be impeded by just seconds of inconsistent reverberation on the preceding context (Zahorik et al., 2009). For binaural speech-perception tasks, experiments have shown compensation occurring at the minimum temporal resolution of the analysed data: in minutes for sentence sets in Longworth-Reed et al. (2009); within six sentences in Srinivasan and Zahorik (2013); and in just a few seconds in Brandewie and Zahorik (2010). Brandewie and Zahorik (2013) recently designed a study specifically to measure the time course of the binaural effect, and reported that 850 ms[2] of room exposure was sufficient to achieve considerable speech intelligibility enhancement.

Having thus derived a measure of reverberation over a particular time period, whether it be based on dynamic-range or on reverberation tails, the measure is then used to linearly control the efferent attenuation applied in the model as described below, i.e. attenuation increases as the level of reverberation increases. This is similar to the noise-based modelling strategies employed elsewhere, where

[1] Experiment H4 below directly addresses this point in Chapter 5.

[2] Interestingly, they found that the compensation mechanism appeared to slow down when the listener task involved dealing with an additional noise component in the stimuli.

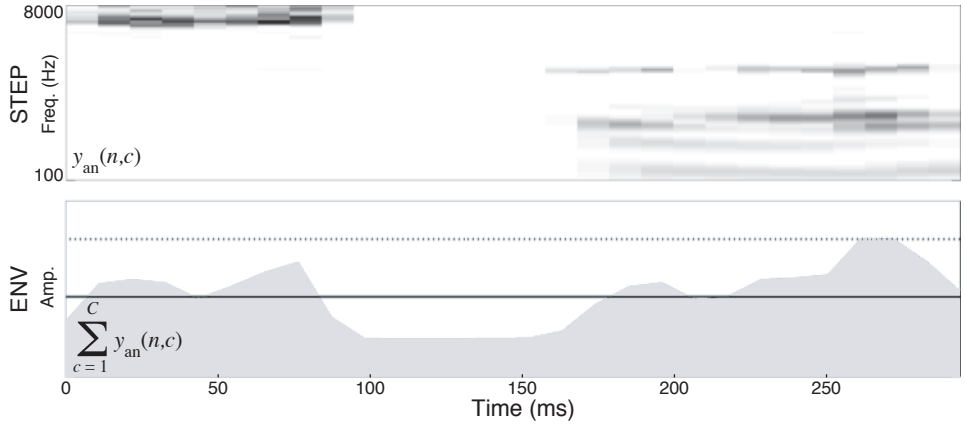


Figure 4.9: Demonstration of the mean-to-peak ratio (MPR) reverberation estimator. *Above:* STEP simulated auditory nerve response, $y_{\mathrm{an}}(n, c)$. *Below:* Envelope (ENV) of the across-channel summed auditory nerve response (as described by Equation 4.4). In the lower panel, the mean level of the ENV signal is shown with a solid line; the peak is shown dotted. An increase in reverberation would raise the noise floor and reduce the dynamic range. Thus the peak would be little changed, while the mean value would correspondingly increase.

attenuation increases as the level of noise increases (Brown et al., 2010; Lee et al., 2011; Messing et al., 2009).

4.3.1 Dynamic range estimation: mean-to-peak ratio (MPR)

The mean-to-peak ratio (MPR), related to the 'blurredness' metric of Palomäki et al. (2004), was proposed in Beeston and Brown (2010) as a method to monitor the dynamic range of the simulated auditory nerve signal (or more specifically, it's temporal envelope), and thereby arrive at an estimate of the amount of reverberation present in the signal. The method relies on the assumption that late-arriving reflections add additional energy to a signal which reduces its dynamic range as the noise floor rises. While the peak value of the signal remains more-or-less unchanged, the mean value rises with additional reflected energy. Thus, with MPR defined simply as the ratio of the mean and peak values, an increase in the level of reverberation (raising the mean value) will bring about a corresponding increase in MPR value recorded.

The upper panel of Figure 4.9 shows the STEP simulated in the auditory nerve (AN) in the afferent pathway of the model in response to the spoken word 'stir' (this STEP was previously shown in the final panel of Figure 4.2). The response in all C frequency channels is summed at each time step n, giving a pooled estimate

of auditory nerve activity,

$$\mathrm{ENV}_{an}(n) = \sum_{c=1}^{C} y_{\mathrm{an}}(n, c). \tag{4.4}$$

The estimated level of reverberation, $R_{\mathrm{mp}}(n)$, at time step n, is then quantified by the ratio of the mean and peak values of the previous AN temporal envelope computed over a windowed portion of duration Z time frames,

$$R_{\mathrm{mp}}(n) = \frac{\frac{1}{Z} \sum_{z=n-Z}^{n} \mathrm{ENV}_{an}(z)}{\max_{z=n-Z}^{n} \mathrm{ENV}_{an}(z)} \tag{4.5}$$

where z indexes time frames within this temporal window.

Posed in this way, later-arriving reflections can be regarded as contributing additional energy to a signal, filling dips in its temporal envelope with reflected energy, raising the noise floor and thereby reducing its dynamic range. The previous chapter reviewed work suggesting that the efferent auditory system is involved in gain control, and appears to bring about a suppression of the auditory nerve response in situations of additive background noise (see § 3.4.1). Thus it seems that a close relationship may exist between reverberation, noise suppression and dynamic range control. A model that adjusts efferent suppression by detecting the effects of reverberation on the signal's dynamic range thus explores the idea that low-level mechanisms controlling dynamic range in the auditory nerve might be involved in compensation for reverberation.

4.3.2 Reverberation tails estimation: low-pass mask (LPM)

The low-pass mask (LPM) metric is not concerned with the dynamic range of the signal, but instead attempts to capture information regarding offsets in the signal since these are frequently prolonged by the presence of a reverberant tail (Beeston and Brown, 2013; Kallasjoki et al., 2014). Inspired by missing data approaches to robust speech processing, LPM considers individual areas in the spectro-temporal representation to be either informative (reliable) or not (unreliable) about the offsets in individual frequency channels of the simulated auditory nerve response. Subsequent processing is based on the informative parts of the signal only; other parts are discarded in the reverberation estimation calculation. A single-channel demonstration, from which it can be inferred that LPM attempts to judge the amount of reverberation present based on the proportion of energy present during tails (offsets) in the signal, is presented in Figure 4.10.

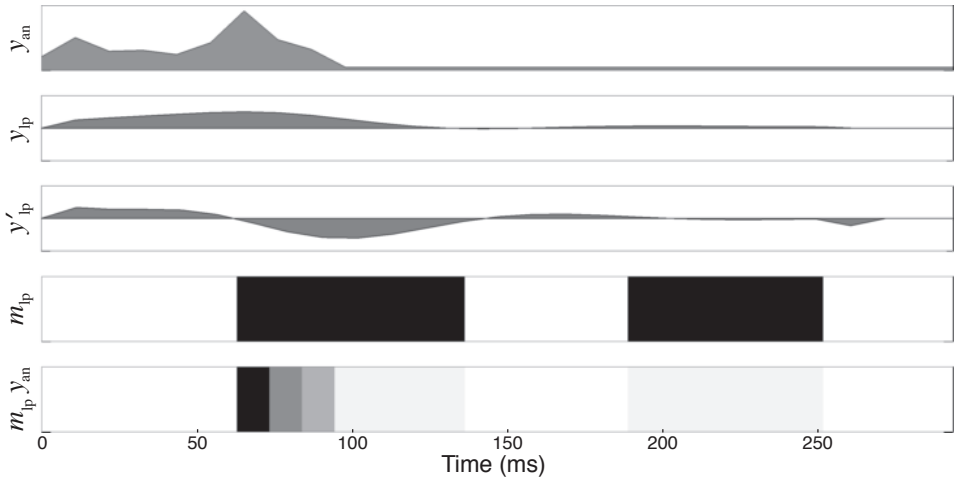


Figure 4.10: Demonstration of the low-pass mask (LPM) offset capture technique in a single-channel of the simulated AN response. *Top to bottom:* simulated AN response, $y_{\mathrm{an}}(n, c)$, in a single high-frequency channel (where $c = 76$); smoothed temporal envelope in this channel, $y_{\mathrm{lp}}(n - \tau, c)$; derivative of the envelope, $y'_{\mathrm{lp}}(n - \tau, c)$; binary mask, $m_{\mathrm{lp}}(n, c)$, which locates offsets via the negative portions of the derivative; the masked signal resulting in that channel, $m_{\mathrm{lp}}(n, c)\ y_{\mathrm{an}}(n, c)$.

Figure 4.10 describes how the LPM method locates 'tail-like' regions in $y_{\mathrm{an}}(n, c)$, the simulated STEP resulting in the auditory nerve simulation. First, the smoothed temporal envelope in each channel, $y_{\mathrm{lp}}(n - \tau, c)$, is estimated using a second-order low-pass Butterworth filter with cutoff frequency at 10 Hz, and is temporally corrected to remove the filter delay, τ. The derivative of the envelope, $y'_{\mathrm{lp}}(n-\tau, c)$, is then calculated and the binary mask, $m_{\mathrm{lp}}(n, c)$, subsequently locates its negative portions such that

$$m_{\mathrm{lp}}(n, c) = \begin{cases} 1 & \text{if } y'_{\mathrm{lp}}(n - \tau, c) < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

Finally, the amount of reverberation present in the signal at time frame n is estimated with a single number, $R_{\mathrm{lp}}(n)$, by computing the mean masked signal strength present in each of the C channels over a preceding context window of Z frames duration (so that z indexes frames within the window), and taking the across-channel mean to summarise these values:

$$R_{\mathrm{lp}}(n) = \frac{1}{C}\frac{1}{Z}\sum_{c=1}^{C}\sum_{z=n-Z}^{n} m_{\mathrm{lp}}(z, c)\ y_{\mathrm{an}}(z, c). \tag{4.7}$$

An increase in the level of reverberation would typically cause a longer reverberation tail, thereby increasing the proportion of signal contributing toward the mea-

sure, i.e. with value 1 in the binary mask $m_{\mathrm{lp}}(n, c)$. An increase in reverberation would thus likely give rise to a corresponding increase in the value of $R_{\mathrm{lp}}(n)$.

There is some support in the both psychoacoustic and speech-technology literature for a 'reverberation tail' based approach to dealing with reverberation. For example, Watkins and colleagues propose that listeners are informed by a reverberation tail-based perceptual mechanism (Watkins, 2005a; Watkins et al., 2011). Additionally, Javed and Naylor (2014) have recently suggested a metric based on the detection of such tails which correlates with objective measures of room reverberation and aims to predict the perceived impact of reverberation on a given speech signal. Since reverberation tail metrics are asymmetric in time, they hold the potential to further examine and possibly explain findings regarding speech perception in time-reversed rooms.

4.4 Experiment M1: Application of the efferent model to sir-stir continuum experiments

In this section, the auditory model is applied to the reverberant 'sir-stir' listener task of Watkins' continuum experiments.

The main findings of Watkins' paradigm are detailed first (§ 4.4.1), including listener results for conditions investigating time-forward and time-reversed speech, and time-forward and time-reverse reverberation. Input and output stages of the modelling study are also described (cf. the first and final boxes in Figure 4.1b). At input, the signal level must be scaled appropriately to provide input to the auditory model that is equivalent to the level heard by human listeners (§ 4.4.2). The proposed reverberation estimators are then aligned to assess the part of the preceding context immediately prior to the test-word (§ 4.4.3), and at output, the simulation of auditory nerve firing is converted into a 'sir' or 'stir' decision (§ 4.4.4). The effect of efferent attenuation on 'sir-stir' continuum tokens is examined manually at first (§ 4.4.5). Finally, the efferent-feedback-monitoring parameters are trained on data points for human listener category boundary points for two naturalistic listening conditions (§ 4.4.6).

After these calibration stages, the remainder of the chapter evaluates the model against Watkins' human listener data to investigate time-reversal of the preceding context speech in Experiment M2, and time-reversal of the preceding context reverberation in Experiment M3.

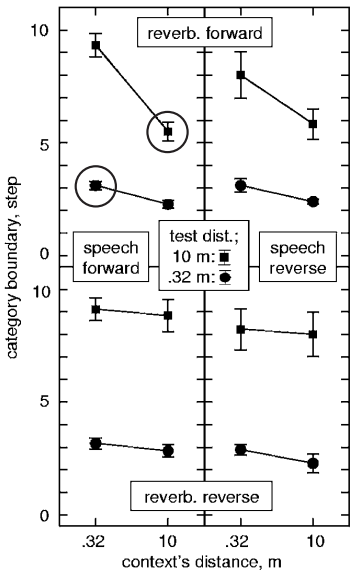


Figure 4.11: Human listener response data in Watkins (2005a, Experiment 5), showing mean and standard error for six listeners' category boundary positions. In forward-reverberation conditions (*above*), the category boundary shifts upward when the test-word alone is far-reverberated. Compensation for reverberation is indicated by the partial recovery of the boundary's earlier position when the context is also far-reverberated. Compensation is disrupted in reverse-reverberation conditions (*below*). Reversal of the context speech had little effect on the category boundaries recorded in either reverberation condition. Additional markers (*circles*) indicate the two same-distance conditions for the forward-speech and forward-reverberation stimuli, with near-near (*left*) and far-far (*right*) *context-test* distances respectively. These two data-points are used to train the efferent attenuation mappings in the following computational study (cf. § 4.4.6 below).

4.4.1 Modelling task: Watkins' 'sir-stir' continuum data

Watkins' continuum experiments were outlined previously in § 2.4.1. In these experiments, the listener task is underlain by a process of categorical perception which appears to disregard the gradual amplitude modulation imposed across continuum steps that creates the impression of the [t] at one end (vs. its absence at the other). Categorical perception ensures that even intermediate continuum steps are identified by listeners as either a spoken 'sir' or 'stir' test-word.

Using such a continuum of test-words, compensation for reverberation has been repeatedly demonstrated[1]. Experiment M1 is a calibration stage aimed to prepare the auditory model to simulate this listener task. The main findings were previously depicted in Figure 2.11, and are visible again in the top left panel of Figure 4.11 which shows human listener data from Watkins (2005a). Here, the influence of far-distance reverberation on the test-word increases the number of 'sir' responses (i.e., shifts the category boundary upwards), as though the dip in the temporal envelope that cued the [t] consonant had been concealed by reverberant energy. However, increasing the level of reverberation in the context to far-distance restores a number of 'stir' responses, even though the factors that had seemed to obscure the [t] are still present.

[1] See for example: Watkins (2005a, b); Watkins and Makin (2007b, c); Watkins and Raimond (2013); Watkins et al. (2010a, b, 2011)

Two further findings of Watkins are of particular interest and are selected for the subsequent modelling experiments. Listener data for these conditions are shown in Figure 4.11.

Firstly, it is apparent from a comparison of the two upper panels in Figure 4.11, that reversing the direction of the context speech has little effect on the listener data: compensation for reverberation is qualitatively similar in both cases. Since the context speech was not intelligible in the reversed speech case, this result suggests that the compensation mechanism does not rely on phonetic perception. Experiment M2 investigates whether the MPR- and LPM-driven efferent models are able to replicate this finding.

Secondly, comparing upper and lower panels in Figure 4.11 reveals that the compensation mechanism is disrupted when the time-direction of reverberation is reversed on the preceding context. As was discussed earlier (cf. § 2.1.6), this result is of interest in the current auditory modelling study since it cannot be predicted by objective measures of reverberant speech perception based on the room's modulation transfer function. Experiment M3 then tests the suitability of the MPR- and LPM-based efferent suppression controllers on this task.

4.4.2 Input signal level calibration

In order for the computational auditory model to undertake the 'sir-stir' listening task, a process of input signal level calibration was required. The aim of this step was to ensure that the experimental stimuli were presented to the model in an equivalent manner to that in which they had been presented to listeners in Watkins (2005a), i.e. with single-channel sound files sampled at 48 kHz, at a maximum root mean square (RMS) level of 48 dB SPL (measured with a 1-second time constant).

A particular audio file was used as a reference signal[1]. This file was played through the equipment that delivered stimuli to listeners in Watkins' laboratory, and the voltage arising in the headphone wire was measured with an analogue RMS voltmeter (B&K 2425) set to the 'slow' setting. This resulted in a continuously varying voltage signal, measured internally with a 1 second time constant, whose maximum value was read off by eye at 60 mV. Since this equipment had previously been calibrated by Watkins, this voltage was known to result in a signal in the left-channel of the headphones at a level of 48 dB SPL.

The analogue voltmeter RMS signal observation was simulated in Sheffield in order that an appropriate scaling factor could be calculated for the reference file

[1] Additionally, this audio reference file was re-used to calibrate the presentation level for the human listener experiments as described in the following chapter (cf. § 5.2.3).

(based in the same way on its maximum RMS value). Thereafter, all 'sir-stir' signals could be multiplied by this scaling factor prior to their presentation to the model, resulting in signals being delivered at 48 dB SPL as was done in Watkins' experiments.

RMS values were therefore calculated for the 'sir-stir' reference audio file, using time-windowed blocks of $K = 48,000$ samples (of duration 1 second at the 48 kHz sample rate in use), and whose starting index t was incremented by 1 sample on each iteration of the calculation. In this way, a series (length T) of local RMS values $\mathrm{RMS}(t)$ were calculated using

$$\mathrm{RMS}(t) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(y_{\mathrm{in}}(k) - \bar{y}\right)^2} \tag{4.8}$$

where k indexes the sample within the RMS time-window, and $\bar{y}$ is the mean value within the signal portion being considered (which has the effect of removing any DC bias in the signal). The peak RMS level of the reference signal, $\mathrm{RMS}_{\mathrm{pk}}$, was then found by simply taking the maximum value in this series such that

$$\mathrm{RMS}_{\mathrm{pk}} = \max_{t=1}^{T}\left(\mathrm{RMS}(t)\right). \tag{4.9}$$

When the audio presentation level is defined as L, a sound pressure level (SPL) value measured in decibels (dB), then the scale-factor adjustment, l, that presents the reference file to the model at the selected level is given by

$$l = \sqrt{2}\left(\frac{p_0}{\mathrm{RMS}_{\mathrm{pk}}}\right)10^{L/20} \tag{4.10}$$

where the standard reference sound pressure value is used, $p_0 = 10^5$ Pa (20 micro pascals).

Calculated once for the 'sir-stir' reference signal with the desired presentation level of $L = 48$ dB SPL, the corresponding value of l was stored and subsequently used to scale every other 48 dB SPL presentation of the monaural 'sir-stir' stimuli to the auditory model.

4.4.3 Monitoring reverberation in the preceding context

Two signal-based approaches to reverberation quantification were proposed in § 4.3, and important aspects of their function were described with reference to an unreverberated signal. In the current section, these two techniques are discussed

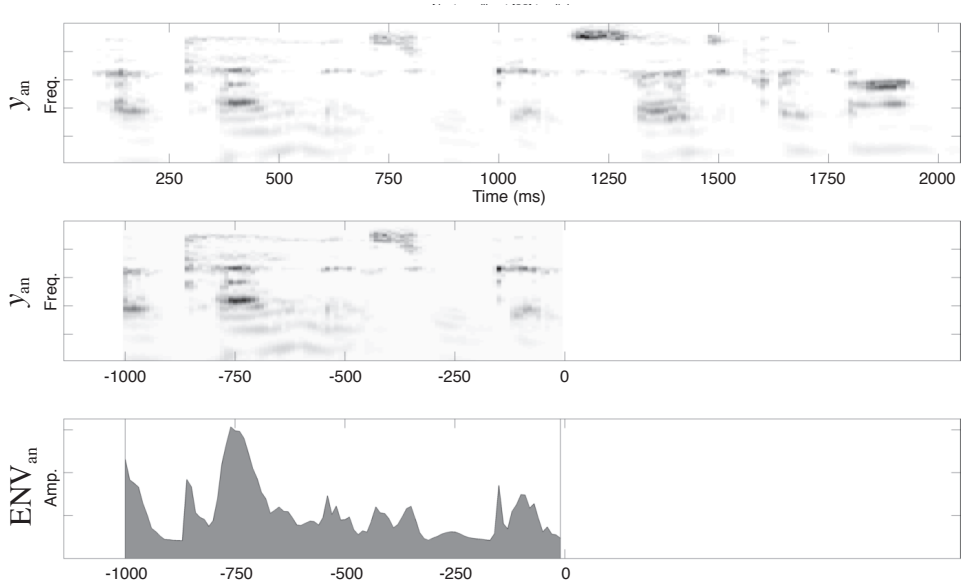
in turn to examine whether they do indeed capture something useful about the reverberation content of the signal in the acoustic context immediately prior to the test-word itself.

The mean-to-peak ratio (MPR) metric works on a specific windowed portion of the context (the recently experienced time), and measures the mean and the peak values experienced, thus keeping track of the signal's dynamic range. This is displayed in Figure 4.12. In Figure 4.12a, the STEP representing AN firing rate is shown in the top panel for a test utterance reverberated at the near distance. The context portion just prior to the test-word, here $Z = 100$ frames, corresponding to context window of 1 second duration at the model's output frame rate, is displayed in the second panel. Finally the third panel shows the across-channel sum, defined earlier in Equation 4.4. This reveals a signal with a large dynamic range, strong modulation content, and relatively sharp onsets and offsets.

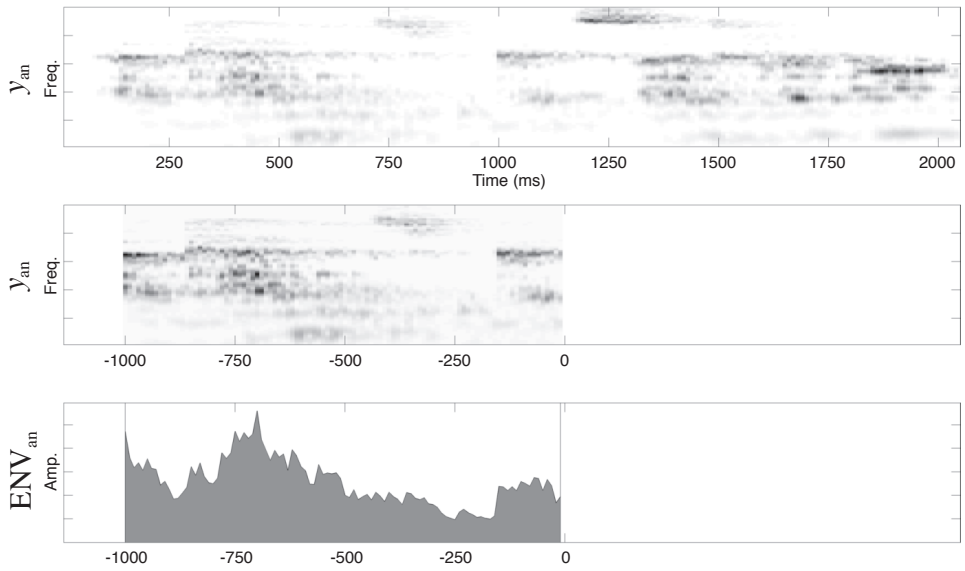
The far-distance context is similarly shown in Figure 4.12b. The peak value is of the same order of magnitude as in the near condition, with the most significant difference being that the dips in the temporal envelope have largely been filled with reverberant energy (effectively increasing the noise floor). This increase in energy shows as a rise in the signal's mean value, which thereby causes a corresponding raise in the value estimating the reverberation content of the signal R_{mp} (cf. Equation 4.5). If an increase in R_{mp} value were to be mapped in the efferent circuit to a process of stronger attenuation, it is anticipated that this would decrease the simulated AN response to the late-arriving, low-level reflected energy, and thereby uncover some of the dips in the temporal envelope.

A similar demonstration is given in Figure 4.13 for the low-pass mask (LPM) reverberation estimator. In this case, only the $Z = 100$ frames of the context window are shown, again for the response in a single high-frequency auditory channel. For the major region of activity in this channel (at roughly 350 ms before the end of the context portion), a comparison of the binary masks in the near distance (Figure 4.13a) and far distance (Figure 4.13b) reveals that the increased reverberation adds longer 'tails' to the window in the far condition. The negative-going part of the smoothed temporal envelope suggests a masked tail-contribution in this channel over approximately 12 frames in the near condition, and over double this duration in the far context condition.

Again, a higher value of R_{lp} results for the far-reverberated speech context than for the near-reverberated context. Following the same logic as described for the MPR estimator above, if an increased R_{lp} were again mapped to an increased attenuation value, then the increase in reverberation could be thought of as activating the efferent suppression mechanism and may possibly simulate effects of MOC unmasking.

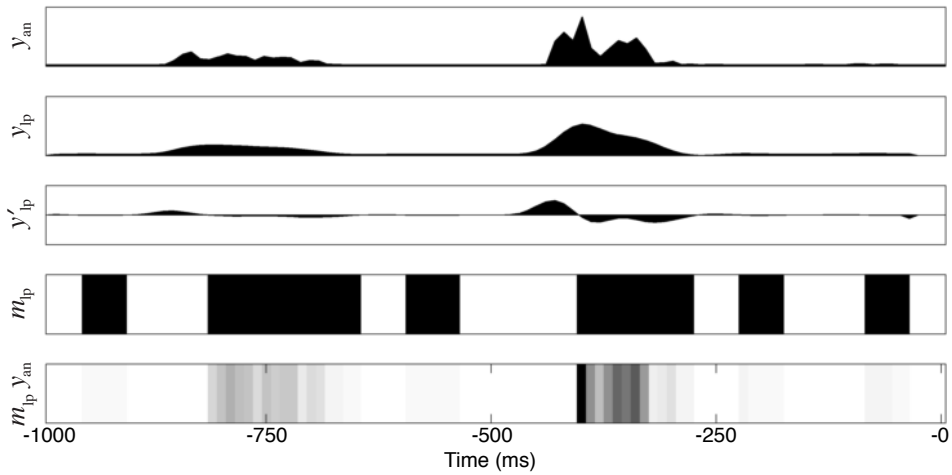


(a) near distance context: MPR measure, $R_{\mathrm{mp}} = 0.34$

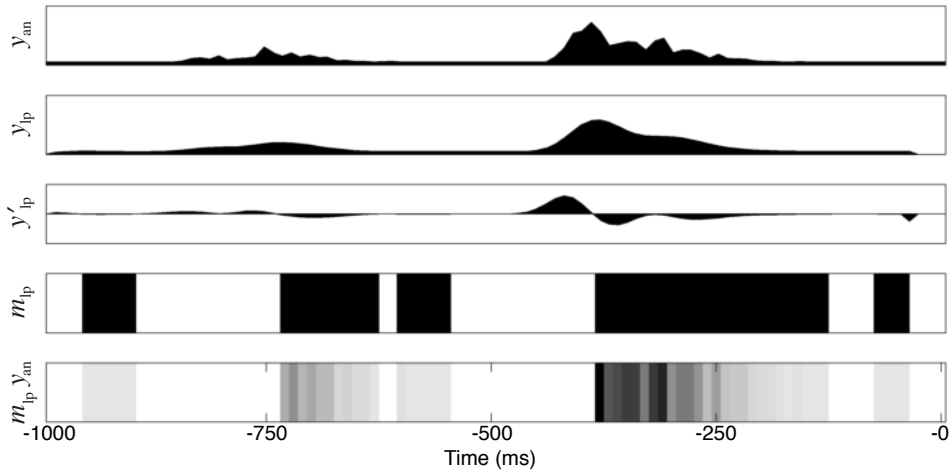


(b) far distance context: MPR measure, $R_{\mathrm{mp}} = 0.47$

Figure 4.12: Mean-to-peak ratio (MPR) measures of the preceding context. STEPs resulting from the simulation of auditory nerve activity, $y_{\mathrm{an}}(n, c)$, for the forward-speech, forward-reverberation continuum stimulus (step 00) are shown in the upper panels of Figure 4.12a for the near-near *context-test* condition, and in Figure 4.12b for the far-far condition. The second and third panels for each stimulus reveal the MPR assessment of the level of reverberation in the portion of the context of 1 second duration ($Z = 100$ frames) immediately prior to the occurrence of the test-word. The across-channel envelope, $\mathrm{ENV}_{an}(n)$, as defined by Equation 4.4, shows sharp offsets and a high dynamic range for the near-distance context and resulted in the value $R_{\mathrm{mp}} = 0.34$. The far-distance context appears less strongly modulated, with reflected energy partially filling some of the dips in the temporal envelope. The increase in reverberation in this condition raised the MPR value to $R_{\mathrm{mp}} = 0.47$. Other details are as described in Figure 4.9.

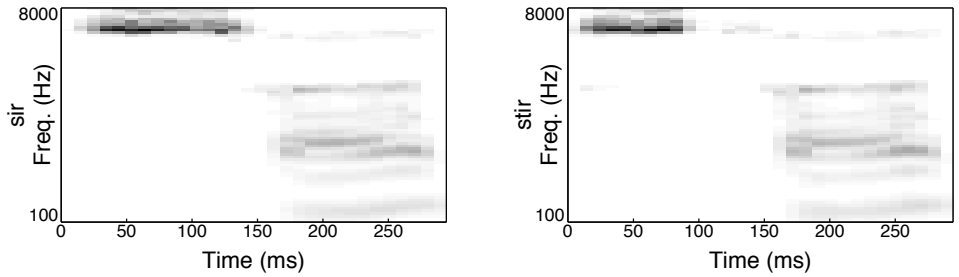


(a) near distance context: LPM measure, $R_{\mathrm{lp}} = 0.63$

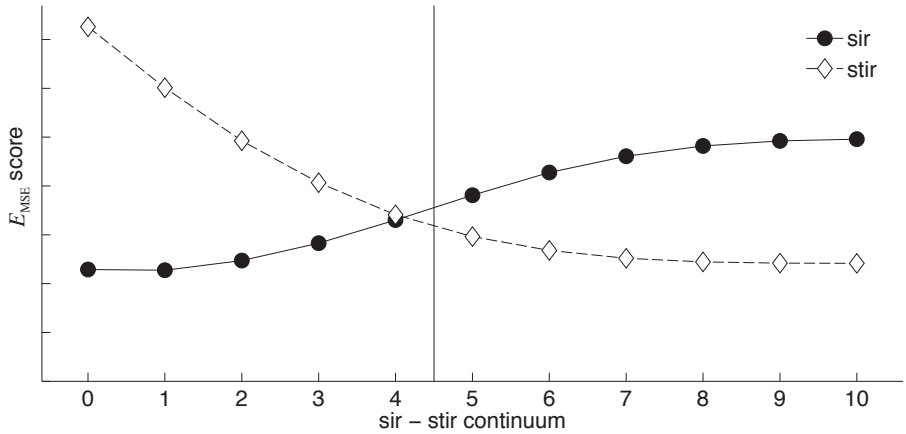


(b) far distance context: LPM measure, $R_{\mathrm{lp}} = 0.80$

Figure 4.13: Single-channel demonstration of the low-pass mask (LPM) estimation technique. A single high-frequency channel ($c = 72$) of the STEP is shown, highlighting the simulated auditory nerve response for the portion of context of 1 second duration ($Z = 100$ frames) immediately prior to the temporal location of the test-word, first in conditions of near distance reverberation (4.13a), and subsequently with far distance reverberation (4.13b). For either stimulus, the smoothed temporal envelope in the selected channel, $y_{\mathrm{lp}}(n-\tau, c)$, and its derivative, $y'_{\mathrm{lp}}(n-\tau, c)$, are calculated. Negative portions in the derivative signal create a binary mask, $m_{\mathrm{lp}}(n, c)$, which locates the portions of the original signal most likely to contain the reverberation tails, $m_{\mathrm{lp}}(n, c)\; y_{\mathrm{an}}(n, c)$ (cf. 4.7). The value of R_{lp} rises from 0.63 in the near distance condition to 0.80 in the far distance condition.



(a) STEP representations for canonical 'sir' and 'stir' templates



(b) 'Sir-stir' identification using mean squared error (MSE) distance.

Figure 4.14: (Fig. 4.14a) STEP representations of the word 'sir' (left) and 'stir' (right) from the extreme ends of the unreverberated 'sir-stir' continuum. The vowel is included here for demonstration, though it is discounted in the experiments reported below. (Fig. 4.14b) Example of the E_{MSE} score (cf. Equation 4.11) altering across the stimuli in the continuum, quantifying the distance from 'sir' and 'stir' templates for each continuum step. Here, the first five steps of the continuum were selected as 'sir', whilst the remainder of the steps bring about 'stir' responses. The category boundary B (cf. Equation 4.12), results in a boundary quantised to the value of 4.5.

4.4.4 'Sir-stir' speech identification

A simple template-matching approach was employed for speech identification: for each sound file, the model responds with a 'sir' or 'stir' decision in much the same fashion that a human listener does. To simulate this 2AFC task, STEP templates were derived from the 'sir' and 'stir' words at either end of Watkins' unreverber-

ated 'sir-stir' continuum (with the efferent attenuation parameter fixed at 0 dB[1]). Resulting templates for canonical 'sir' and 'stir' test-words are shown respectively in Figure 4.14a.

During simulation, utterances for each step of the continuum were presented to the model and the corresponding STEP token, $y_{\mathrm{an}}^{\mathrm{tok}}$, was computed from the simulated auditory nerve response. The time frames corresponding to the test sound were compared in turn with the 'sir' and 'stir' templates, $y_{\mathrm{an}}^{\mathrm{tem}}$, using a standard MSE metric[2], given by

$$E_{\mathrm{MSE}}(y_{\mathrm{an}}^{\mathrm{tok}}, y_{\mathrm{an}}^{\mathrm{tem}}) = \frac{1}{C}\frac{1}{N}\sum_{c=1}^{C}\sum_{n=1}^{N}\left[y_{\mathrm{an}}^{\mathrm{tok}}(n,c) - y_{\mathrm{an}}^{\mathrm{tem}}(n,c)\right]^2 \tag{4.11}$$

where $y_{\mathrm{an}}^{\mathrm{tok}}$ and $y_{\mathrm{an}}^{\mathrm{tem}}$ are STEPs of dimension C frequency channels and N time frames. Since listeners rely on a specific phonetic cue (the presence or absence of a [t]) in order to distinguish between 'sir' and 'stir', the template matching process was similarly restricted to the part of the test-word that contains the initial sibilant and stop[3]. For each utterance, the template with the smallest value of E_{MSE} was then chosen as the test sound identity ('sir' or 'stir'). This process is visualised in Figure 4.14b.

Finally, the category boundary reported the point along the 11-step continuum at which the 'percept' switched from 'sir' to 'stir', as shown by vertical line in Figure 4.14b. By analogy to the numerical method outlined in Watkins (2005a), the category boundary B was calculated essentially from a count of the number of sir responses, I_{sir}, with

$$B = I_{sir} - 0.5 \tag{4.12}$$

so that its values lie in the range -0.5 to 10.5 as used by Watkins. Since Watkins reported results across a number of human listeners, each of whom received a number of presentations of each stimulus, his category boundary results varied smoothly across the entire range. Contrasting this, the current model is quantised: it can only output category boundaries directly at $\{-0.5, 0.5, 1.5, ..., 10.5\}$. Since the model is fully deterministic, repetitious presentation of stimuli to the model would bring about exactly the same result each time. A number of suggestions are discussed in § 4.7.3 below to work around this issue.

[1] In Beeston and Brown (2010), efferent attenuation was not fixed at 0 dB at this stage. Rather, templates were created using the linear-fit-derived attenuation values at each time-step.

[2] A similar approach, comparing unknown tokens to frozen speech templates using the MSE distance, was also used in Messing et al. (2009).

[3] With the benefit of hindsight, this appears overly restrictive (cf. discussion in § 4.7.3).

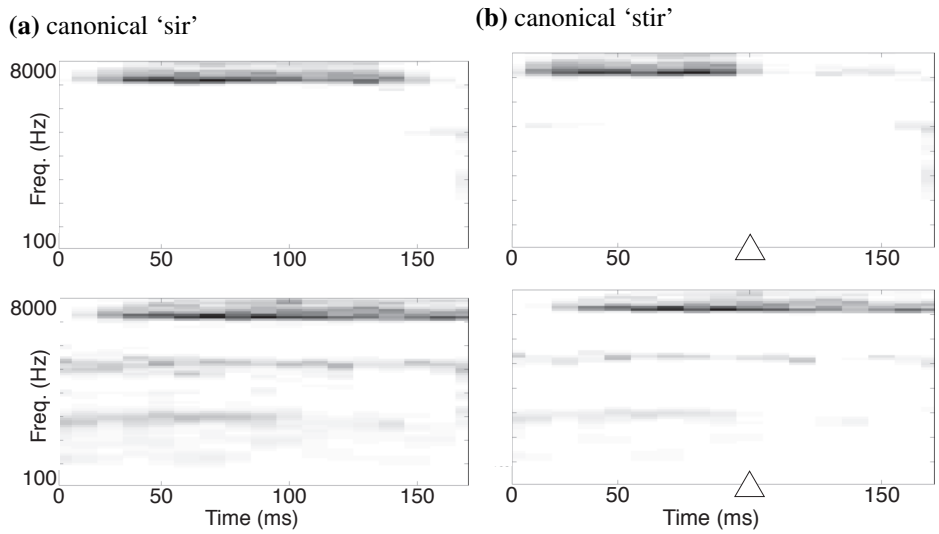


Figure 4.15: STEP representations of the consonant portions of 'sir' (4.15a) and 'stir' (4.15b) from the extreme ends of the unreverberated 'sir-stir' continuum. A test-word from the middle of the continuum is initially reported as 'sir' in the presence of reverberation (4.15c). When efferent attenuation is applied, the /t/ closure (*triangle*) is partially revealed and the word is reported as 'stir' (4.15d).

4.4.5 Efferent attenuation applied to the continuum

Section 4.2.2 previously explained the central hypothesis underlying the efferent model, namely that the shift in the rate-level curve (cf. Figure 4.8) will suppress the response to low-level reverberation in signal areas that were previously characterised by spectro-temporal gaps. In this way, the influence of reflected energy protruding into the signal regions of low amplitude will be reduced. Since stop consonant perception is influenced by dips in the temporal envelope, and since such regions are easily obscured by reflected energy, it follows that an increase in efferent attenuation might allow improved recognition of reverberant stop consonants.

To investigate the effects of efferent attenuation on the 'sir-stir' continuum stimuli, a model was configured in the 'open-loop' fashion as displayed in Figure 4.1a. Here, the efferent attenuation parameter, ATT (measured in dB), could be manually specified, and held fixed for the entire duration of the sound file.

Figure 4.15 illustrates this process with 'sir-stir' continuum files. Panels 4.15a and 4.15b show the initial consonant portions [s] and [st] of the unreverberated continuum extremes for 'sir' and 'stir' respectively. In 4.15c, the STEP resulting from a reverberated signal from the middle of the continuum is shown. Here, reflected

energy is smeared across the signal, and the model responds by labelling the token 'sir'. In 4.15d, a fixed amount of efferent attenuation has been applied to the same reverberant sound file viewed in 4.15c. The attenuation applied has reduced the simulated auditory nerve response overall, and the model this time selects the label 'stir' as the closest match.

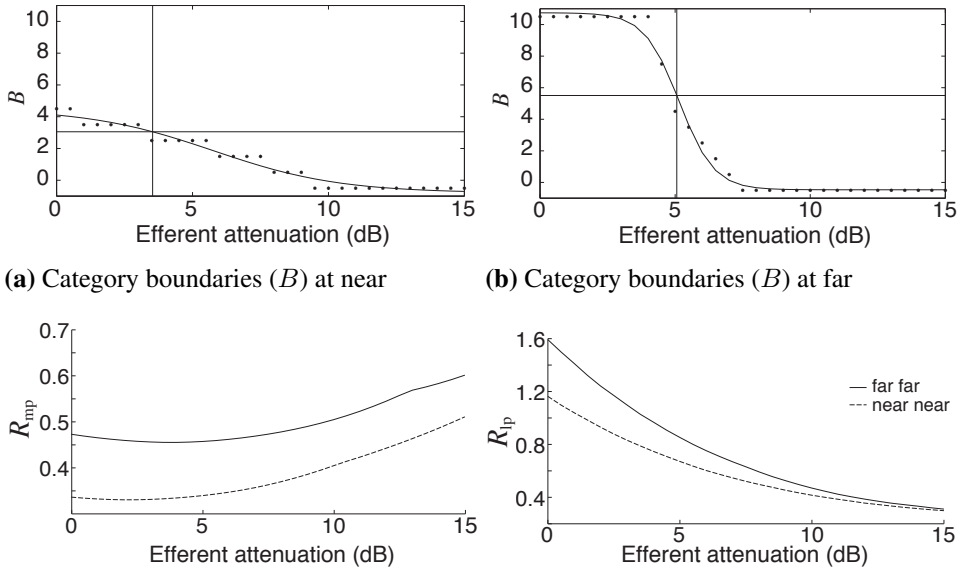
This result confirms that manipulation of the efferent attenuation parameter provided to the DRNL filter bank (Ferry and Meddis, 2007) can indeed influence the subsequent recognition of a reverberant stimulus from the middle of the Watkins' 'sir-stir' continuum. Here, a token was first recognised as 'sir' when reverberation obscured spectro-temporal dips in the signal, and secondly recognised as 'stir' when the simulated auditory nerve response to low-level reflected energy was suppressed by the efferent signal.

4.4.6 Tuning the efferent feedback circuit

This section describes how the approximate operating point of the efferent feedback circuit was found, in order that the model may be run in the 'closed-loop' implementation shown in Figure 4.1b where the efferent attenuation parameter, ATT, is derived on the fly from the measured metric value (MPR or LPM). The process deriving the metric-to-attenuation mappings is as follows.

Two experimental conditions from Watkins (2005a, Experiment 5) were selected for use as training data. These conditions are marked on Figure 4.11, and constitute the two naturalistic conditions of everyday listening: near- and far-distance reverberation, while both the speech and reverberation remain in the usual time-forwards direction. For each of the stimuli at these training conditions, sound files were scaled and input to the model as described in § 4.4.2; the STEP for each continuum step was computed in the auditory model, and matched against 'sir' and 'stir' templates to obtain the category boundary as described in § 4.4.4. Concurrently, the value of efferent attenuation applied in the model, ATT, was varied systematically from 0 to 15 dB in steps of 0.5 dB.

The results of this process are shown for the near-distance reverberation condition in Figure 4.16a, and for the far-distance condition in Figure 4.16b. The quantisation of the model, discussed in § 4.4.4, is apparent in the descending staircase-like pattern of data points. The near-distance condition appears to be relatively robust to changes in attenuation: the staircase has broad, flat steps. Furthermore, the mean category boundary for the human listeners, at $B \simeq 3$, lies directly between two possible model outputs ($B = 2.5$ or $B = 3.5$). On the other hand, the far-distance condition has a steep slope more typical of psychometric functions: moving outwith a fairly small range of efferent attenuation values (c. 4 to 8 dB) is enough to



(a) Category boundaries (B) at near

(b) Category boundaries (B) at far

(c) Variation of R_{mp} with ATT

(d) Variation of R_{lp} with ATT

Figure 4.16: Derivation of the efferent attenuation mapping function in Experiment M1. *Above:* Determining the value of efferent attenuation that approximates category boundaries, B, in the human data for near-near *context-test* distance stimuli (Figure 4.16a) and for far-far stimuli (Figure 4.16b). Quantisation in the boundary decisions is overcome using an interpolated sigmoid fitting process. *Below*: The corresponding reverberation estimation values are shown as a function of the efferent attenuation applied. The MPR measure (R_{mp} in Figure 4.16c), and LPM measure (R_{lp} in Figure 4.16d) consistently produce larger values for the far reverberation conditions (plotted with *continuous* lines) than for the near distance conditions (*dotted*).

flip the response of the computational model from 'sir' to 'stir' for *all* steps of the continuum.

To obtain the value of efferent attenuation that would approximate the category boundaries recorded in Watkins' human listener data, the data-points arising from the model simulations are fitted with a sigmoid curve (shown in the solid line in Figures 4.16a and 4.16b) at each of the near- and far-distance conditions. The human category boundary is marked with a horizontal line, and the resulting ATT value with a vertical line. Thus for the open-loop model configuration, where the attenuation is applied at a fixed level (throughout the duration of the sound file), the near-near category boundary would be well matched (were the 'sir-stir' identification quantisation not an issue) at around $ATT \simeq 3.5$ dB. A correspondingly higher value of $ATT \simeq 5$ dB would be needed to achieve the far-far category boundary. Here, the principle underlying the model becomes apparent: when there is more reverberant energy, more efferent attenuation is required.

Section 4.3 previously proposed two reverberation estimation methods to drive the efferent feedback circuit when run in the closed loop model configuration. There is little evidence with which to determine exact time-scales over which such measures might be best implemented, however the efferent system is generally thought to act on rather slower timescales than the afferent system (Backus and Guinan, 2006; Cooper and Guinan, 2003). In the present study, the context-assessing measures were each calculated over a signal period of duration 1 second (i.e. Z=100 frames) prior to the test-word's position in the stimulus. The resulting estimator values, again varying with the fixed level of attenuation applied in the open loop configuration, are shown for R_{mp} estimated by the MPR in Figure 4.16c, and for R_{lp} estimated by the LPM in Figure 4.16c. Here, the reported metric values were obtained by averaging across all steps in the 'sir-stir' continuum.

The behaviour of these metrics differs as efferent attenuation is increased. In the case of R_{mp}, an ever-increasing value of attenuation eventually pushes the majority of the signal back under the threshold for firing. The result therefore is a significant reduction in the peak value. The mean value however is slightly boosted relative to this by the fact that the spontaneous firing rate is always positive and non-zero in this implementation. Thus the ratio of mean and peak tends to increase (cf. Equation 4.5) at higher values of efferent attenuation. In the case of R_{lp}, the ever-increasing attenuation reduces the signal strength so severely that there is little energy left to capture in the reverberant tail mask (cf. Equation 4.7) and the estimation of reverberant energy gradually dwindles. Importantly, however, the lower row of Figure 4.16 shows that both reverberation estimation techniques consistently measure smaller values for the near-distance reverberation, and larger values for the far distance condition containing a correspondingly higher degree of reflected energy.

For each reverberation estimator, the values that arose from the 0.5 dB steps of Figure 4.16 were additionally interpolated to obtain values corresponding to the attenuation level that approximates the human near- and far-distance category boundaries. A linear mapping is then assumed between the measured metric value and the required attenuation, so that the level of attenuation applied in the model, ATT, increases monotonically with increasing reverberation. This allowed the two measures to be tested as controllers of the efferent feedback circuit, where the amount of efferent attenuation to apply at each time-step is derived from the following linear fits[1]. For MPR, the attenuation applied (in dB) was subsequently derived on the fly using

$$ATT_{\mathrm{mp}} = \max\left[\left(11.18 \times R_{\mathrm{mp}} - 0.24\right), 0\right], \tag{4.13}$$

[1] Numerical values differ from those in Beeston and Brown (2010, 2014) due to knock-on effects from alterations in the input signal level scaling method and hair cell mapping parameters.

and for LPM, using

$$ATT_{\mathrm{lp}} = \max\left[\left(3.57 \times R_{\mathrm{lp}} - 0.39\right), 0\right]. \tag{4.14}$$

Inclusion of the $\max[..., 0]$ term in these equations ensures that any potential negative values for efferent attenuation are excluded. In practise, however, values of R_{mp} and R_{lp} were never sufficiently small for this situation to arise. Moreover, the simulated attenuation values appear to be in the range predicted by physiological experiments reported in the literature (see e.g., Cooper and Guinan, 2003; Meddis et al., 2013; Murugasu and Russell, 1996).

Section 4.4 has described the process by which behaviour of the efferent circuit was tuned on the conditions of Watkins' experiment where both context and test-word were heard at the consistently near or consistently far distance reverberation. Once the corresponding linear approximations had been found (equations 4.13 and 4.14 respectively for the MPR and LPM estimation techniques), no further tuning of the model was performed; the same parameters were used for all following simulations.

4.5 Experiment M2: Compensation for reverberation with time-forward and time-reversed speech

Experiment M2 investigates whether an efferent-inspired auditory model is able to simulate compensation for the effects of reverberation in the time-forward reverberation conditions reported by Watkins (2005a) as was previously depicted in the upper panels of Figure 4.11. Two different methods (MPR and LPM) are investigated as estimators of the reverberation. Each of these is used to control the efferent feedback signal via a mechanism which regulates the afferent stage of audio processing.

The MPR- and LPM-driven models are tested in conditions where the acoustic context preceding the test-word may contain either time-forward or time-reversed speech signals. Firstly, the time-forward speech conditions allow an examination of the basic compensation for reverberation paradigm first introduced in § 2.4.1. Here, the addition of reverberation to a test-word influences its identity; addition of similar reverberation to the surrounding context appears to reduce the perceptual degradation of the initial test-word reverberation. Secondly, by including conditions where the context speech signal was reversed, Watkins asks whether human listeners' need to understand the speech signal in order to compensate for the effects of reverberation. This question is of interest to the current study since the auditory model has no language model component as presently described. This

is contrary to most machine listening techniques, where statistical language models are usually included alongside acoustic models representing the various speech sounds. Examining human responses in this dataset therefore allows us to query whether a language model would be required in order to simulate compensation effects. If this component were necessary, then the compensation effect would be expected to break down in conditions where the speech in the context is time-reversed.

4.5.1 Watkins' stimuli and human response data

The test stimuli in this experiment follow the pattern of the 'sir-stir' paradigm outlined earlier (cf. § 2.4.1) for the baseline time-forward reverberation, time-forward speech condition. By manipulating the temporal envelope of a test-word, Watkins constructed an 11-step test-word continuum, where tokens varied between 'sir' at one end and 'stir' at the other. The test-word was embedded in a context phrase ("OK, next you'll get [...] to click on") and the reverberation conditions of the context portion and test-word portion of the signal were varied independently by convolving them with room impulse responses recorded at either near or far source-receiver distances.

Listeners were asked to identify the test-word as either 'sir' or 'stir'. The category boundary position (the step in the continuum where listeners' percept shifted on average from 'stir' to 'sir') was recorded in a mid-continuum position for the near-reverberated test-word embedded in the near-reverberated context. This is shown by Watkins' data-points marked at '0.32 m' distance in the upper left panel of Figure 4.11. When the test-word alone was reverberated at the far-distance (marked '10 m'), the category boundary shifted upwards as listeners responded 'sir' to more steps of the continuum. However, compensation for reverberation was evident in the downward slope of the upper line in both upper panels (cf. Figure 4.11). Here, an increase in reverberation distance of the preceding context speech lowered the category boundary position, and more stimuli were again reported as 'stir' by the listeners.

While the test-word was always presented in the conventional time-forward direction, in some experimental conditions the context speech surrounding the test-word was temporally reversed. That is, the sequence of samples in the unreverberated speech signal has been inverted before convolution with the impulse responses for near or far source-receiver distance. This process destroyed the semantics of the speech context but maintained a similar overall spectral profile. The pattern of listener data in Watkins' results indicates that phonetic understanding is not necessary for compensation to occur: compensation for reverberation is observed even when the time-direction of the context speech was reversed.

4.5.2 Methods

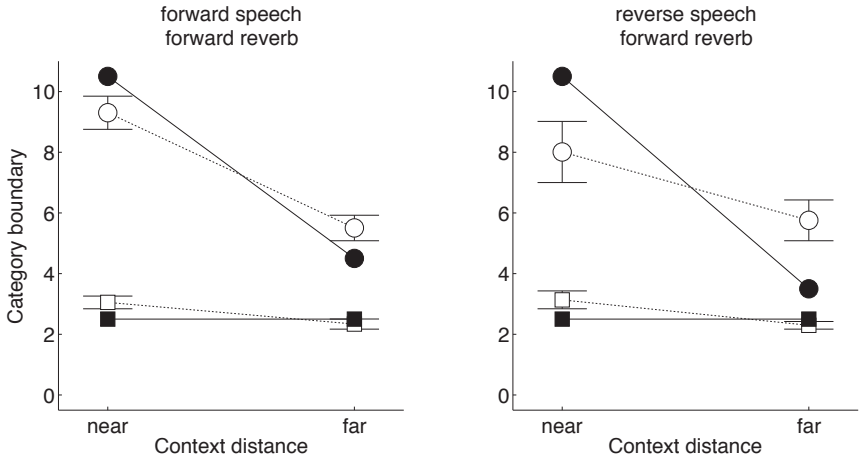
The model was applied to the 'sir-stir' continuum files as described in Experiment M1. All experimental stimuli were sampled at 48 kHz and scaled in level to the 48 dB presentation level required. The resulting single-channel files were processed by the afferent pathway of the model, with attenuation initially set at 0 dB. This stage comprised the model components representing the filtering stages of the outer and middle ear, the afferent cochlear processing, inner hair cell transduction and temporal integration at the auditory nerve as described in § 4.2.1.

The STEP resulting from the afferent processing was then monitored over a windowed portion of the context preceding the test-word (examining Z = 100 frames, equivalent to 1000 ms duration at the model's output sample rate) as shown in § 4.4.3. In separate simulations, the level of reverberation in this portion of the signal was assessed by the MPR or LPM estimators, and the value of the efferent attenuation parameter subsequently provided to the DRNL filterbank was found using the linear mappings defined in equations 4.13 and 4.14. In this way, the efferent feedback circuit both monitored and adjusted the behaviour of the auditory model's afferent processing chain.

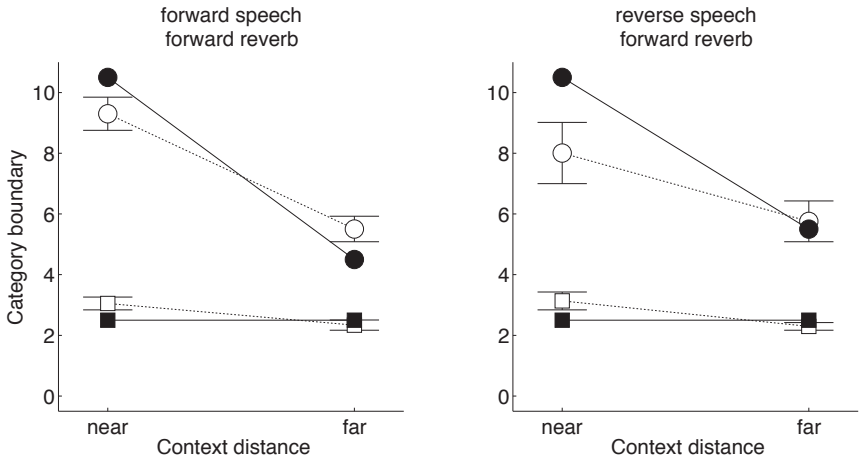
Finally, the STEP arising in the test-word portion of the signal was compared to canonical 'sir' and 'stir' representations using the MSE comparator as described in § 4.4.4, and the closest match was selected as the test-word identity. Across the entire set of stimuli for the continuum (at a particular experimental condition), the category boundary result was then derived using Equation 4.12 as previously described.

4.5.3 Results

Figure 4.17 displays the category boundaries resulting in the auditory model simulation for time-forward and time-reversed speech conditions (in time-forward reverberation). The same overall pattern of results is observed for both model simulations, as can be seen by comparing Figure 4.17a for the efferent control based on dynamic range (R_{mp}, from MPR), and Figure 4.17b for the mechanism based on a measure of reverberant tails (R_{lp}, from LPM). For the conditions where speech and reverberation were presented in their usual time-forward senses, left-hand panels replicate the major hallmarks of compensation for the effects of reverberation. Here, despite the fact that the training data for the efferent feedback circuit included only the same-distance phrases (at near and far reverberation distances), the model provides category boundary results that are qualitatively similar to Watkins' human data for the mixed-distance conditions.



(a) Time-forward and time-reverse speech results derived using MPR.



(b) Results derived using LPM.

Figure 4.17: Experiment M2: Category boundaries for time-forward (*left*) and time-reversed (*right*) speech conditions (in time-forward reverberation). Modelling results (*black markers*) arise using the MPR metric in fig. 4.17a, and the LPM metric in fig. 4.17b to monitor a windowed-portion of the preceding context (length $Z = 100$ frames, equivalent to 1000 ms) and derive ATT (dB) during the simulation. For both reverberation estimation measures, the category boundary results are qualitatively similar to human listener results (*white markers*) in Watkins (2005a, Experiment 5) and demonstrate a pattern of compensation for reverberation in both time-forward and time-reversed speech conditions.

Since the reverberation estimation metrics assess only the context portion of the signal in the current model implementation, they are not influenced by the reverberation distance of the test-word itself[1]. Rather, their value depends on both the signal content itself and the reverberation arising from the particular source-receiver distance in use.

For the forward-speech case, the near distance context reverberation gives rise to an efferent attenuation value in the region of 3.5 dB ($ATT_{\mathrm{mp}} = 3.53$ dB derived via the MPR measure in Equation 4.13, and $ATT_{\mathrm{lp}} = 3.59$ dB derived via the LPM method presented in Equation 4.14). When the test-word distance matched the near context distance, the category boundary predicted by the model was slightly lower than the human boundary result. This result can be understood with reference to Figure 4.16a: the degree of quantisation inherent in the model's 'sir-stir' identification procedure is such that a boundary of $B \simeq 3$ cannot be achieved, the closest available boundary being ± 0.5 (cf. Equation 4.12). Since both reverberation estimators have resulted in a little *over* 3.5 dB attenuation, the category boundary drops into the region where a boundary of $B = 2.5$ results. In fact, as can be read from this figure, any value of efferent attenuation from 3.5 up to around 5.5 dB will result in a category boundary of 2.5 for the near-near condition, as the category boundary is rather insensitive to variation in attenuation in this region.

When the test-word is far-reverberated but the context remains at the near distance, the number of continuum steps reported as 'sir' is maximal. In this condition, the [t] closure is obscured by overlap reverberation from preceding speech sounds and as a result the MSE matching process (cf. § 4.4.4) reports all continuum steps as being more similar to the canonical 'sir' template. The increased reverberation measured for the far-distance context conditions leads to a higher value of ATT being applied in the model. This suppresses low-level activity in the simulated AN response; a lower category boundary therefore results, as more steps of the continuum are now reported to be more similar to the canonical 'stir' template.

The far-distance context gives rise to a stronger attenuation signal, equivalent to 5.06 dB for both reverberation estimators. While the resulting boundary for the near test-word is recorded around the same position as for human listener responses, the boundary resulting for the far-distance test-word is one step lower than the human result for either reverberation mapping. Again, this can be predicted by the tuning curve shown earlier (in Figure 4.16b). The predicted value of $ATT = 5.06$ dB lies on the steep part of the curve where a small change in efferent attenuation brings about a large change in category boundary. Here, the

[1] Experiment H3 below indicates that the model should be updated to include sources of intrinsic information deriving from within the test-word in addition. This is discussed further in § 6.1.3.

boundary is marked at $B = 4.5$ for the model results rather than at around step 5.5, i.e. one step of the continuum has been labelled differently by the computational model compared with the human listener data.

In the reverse-speech cases, the computational model predicts a similar pattern of results: compensation occurs despite the temporal reversal of the context speech signal. For both efferent control metrics, the model over-estimates the influence of far-distance reverberation on the test-word when the near-context is present. The category boundary that results is correspondingly higher than was observed with human listeners.

For a far-distance test-word, the model driven by the MPR metric additionally over-estimates the recovery due to a consistently far-reverberated context; the predicted boundary is lower than the human boundary as a result. Here, the attenuation predicted by R_{mp} for the far-distance context is $ATT = 5.41$ dB. The far-far boundary predicted by the LPM model better fits the human perceptual data. In this case a lesser attenuation was predicted by R_{lp} for the far-distance context, with $ATT = 4.68$ dB. Compared with the MPR-driven model, correspondingly less of the simulated AN response is suppressed by the LPM-driven feedback loop.

4.5.4 Interim discussion

Both the MPR and LPM efferent control mechanisms achieved a simulation of the overall pattern of data for perceptual compensation for the effects of reverberation in forward-reverberation conditions. Here, increasing reverberation in the context surrounding a test-word improved recognition of a test-word's [t] even though this processes added further reflected energy into the signal. Thus, the MPR, based on a measure of dynamic range, and LPM, measuring signal energy in reverberant tails, appear to both capture some relevant properties of reverberation that can be used to drive the efferent feedback loop such that efferent suppression is proportional to the level of reverberation experienced. This suggests it may not be unreasonable to treat the noise-like aspect of reverberation (the raised noise floor from late-arriving reflections) with the kind of efferent activity (suppression) that is thought to be active in situations of noisy listening.

By including conditions where the context speech signal was reversed, Watkins asked whether human listeners' used semantic processing to compensate for the effects of reverberation. In general, humans can still understand speech well provided that the time-reversed segments are less than 100 ms in length; at longer reversal lengths, intelligibility decreases (Saberi and Perrott, 1999). The duration of the context phrase portions in Watkins work was substantially longer (at 1.15 s), thus these portions of the speech signal were completely unintelligible when heard

backwards. Nonethess, Watkins showed that perceptual compensation persisted in conditions with time-reversed speech contexts, when the speech sounds were not heard as a series of words. This (and subsequent experiments) allowed Watkins et al. (2011) to conclude that compensation for reverberation is not due to phonetic perception, but arises from a rather more general perceptual-constancy mechanism.

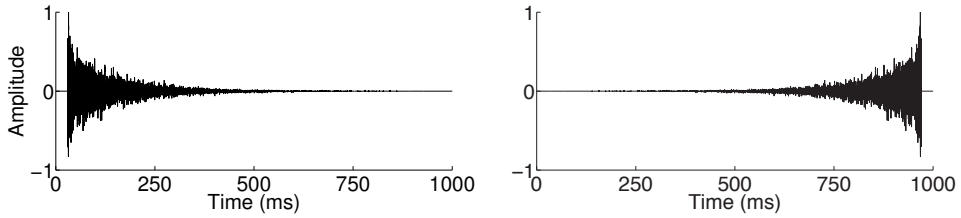
Without further tuning, the auditory model similarly displayed compensation for reverberation in conditions with time-reversed speech contexts. Both versions of the model followed the qualitative data trend overall, but the model in which efferent suppression was driven by LPM gave a slightly closer match to human category boundary results than did the model based on MPR. It appears then that a language model component is *not* required for the current auditory modelling task. However, it should be noted that this finding is likely due to the unpredictable nature of the 'sir-stir' stimuli, since Srinivasan and Zahorik (2013) have recently reported that the ability to make semantic predictions about a test-word are sufficient to overcome the influence of reverberation in the phrase.

4.6 Experiment M3: Abolition of compensation with time-reversed reverberation

Experiment M3 investigates whether the efferent-inspired auditory model simulates the *disruption* of compensation for the effects of reverberation that is observed in human listener data reported by Watkins (2005a), for conditions in which the time-direction of reverberation in the preceding context is reversed. As in the earlier work, two methods are used to estimate reverberation in the portion of the signal immediately preceding the test-word; one is based on dynamic range (MPR, described in § 4.3.1) and the other on a measure of reverberation tails (LPM, described in § 4.3.2). The resulting measures of reverberation, R_{mp} and R_{lp} respectively, are used to automatically control efferent attenuation regulating the afferent processing in the auditory model as described in § 4.3.

It is anticipated that the model based on dynamic range will not capture this effect since the dynamic range is little affected by the time-direction of reverberation[1]. On the other hand, the reverberation tail measure is sensitive to the time-direction

[1] Nonetheless, due to the nonlinear processing in an auditory model (adaptation, refractory periods etc.) the representation for time-reversed stimuli will not be simply a time-reversed version of the response to time-forward stimuli. In some implementations the MPR measure may therefore be sensitive to the time-direction of reverberation, as was seen in the model configuration of Beeston and Brown (2010). This topic is discussed fully in § 4.6.4 below.



(a) Time-forward impulse response **(b)** Time-reversed impulse response

Figure 4.18: Experiment M3 investigates conditions in which the reverberation processing applied to the context has itself been time-reversed. The sequence of time-samples in the room impulse response (*left*) is time-reversed (*right*) prior to convolution with the 'sir-stir' continuum test-words and context phrases.

of the signal and may therefore capture this effect because the time-reversal of reverberation removes tails at offsets, and adds ramps at onsets instead.

4.6.1 Watkins' stimuli and human response data

The modelling task in this experiment investigates conditions where the 'sir-stir' continuum test-words are embedded in context phrases in which the reverberation processing applied to the context has itself been time-reversed. As shown in Figure 4.18, the sequence of samples in each room impulse response was optionally time-reversed before convolution with the unreverberated speech signals. As before, the test-word was always presented with both speech and reverberation in the conventional time-forwards direction. Other aspects of stimuli presentation were the same as described in § 4.5.1, so that the same- and mixed-distance phrases were now investigated, in the presence of time-reversed context reverberation, for both time-forward and time-reversed context speech.

Human listener responses for these conditions were obtained by Watkins (2005a) as shown in the lower panels of Figure 4.11 (discussed earlier in § 4.4.1). As was the case in time-forward reverberation, the direction of speech in the context did not substantially affect the results (i.e. the left panel follows the same trend as that on the right). Increasing the test-word reverberation from near to far while the context remains at the near distance brings about a large shift in the listeners' mean category boundary position, as had been observed in the time-forward reverberation conditions. However, in conditions where the reverberation applied to the context was time-reversed, the addition of a consistently far-reverberated context no longer aids perception of the far-reverberated test-word. Thus, compensation for reverberation did not occur when the RIR was time-reversed.

4.6.2 Methods

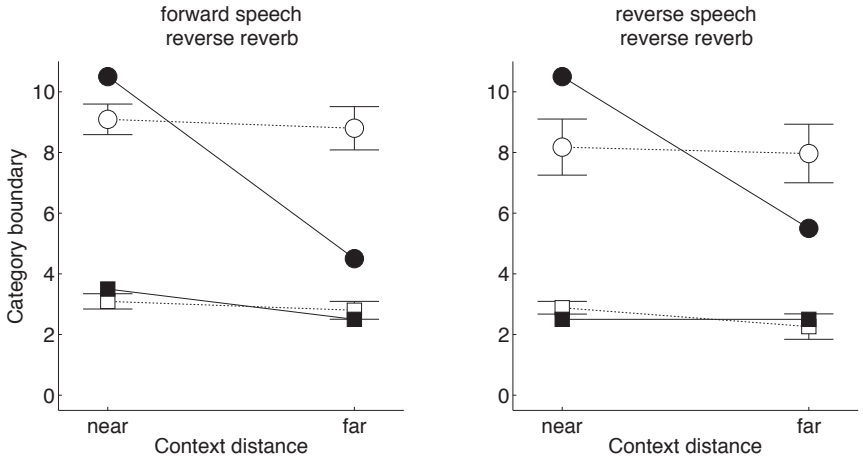
No further tuning of model parameters was undertaken for Experiment M3. Therefore, as was described in the previous experiment, the two reverberation estimators (MPR and LPM) were each used to control the efferent feedback signal sent to the auditory periphery via the existing linear mappings (cf. equations 4.13 and 4.14 respectively). Time positions locating the test-word portion of the signal were updated to account for the later starting index of the speech content in the file (which arose since the time-reversed reverberation processing introduced ramps before onsets in the temporal envelope, rather than tails after signal offsets). All other aspects of the model configuration were the same as described in § 4.5.2.

4.6.3 Results

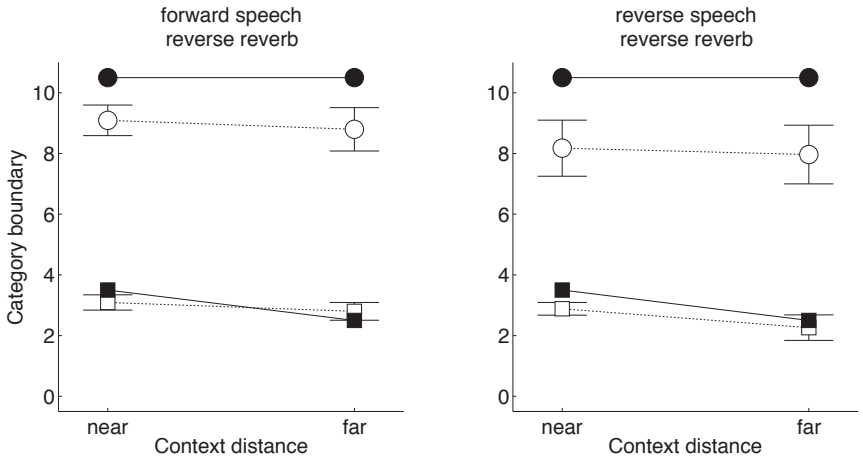
Category boundaries arising for time-forward and time-reversed context speech conditions are shown in Figure 4.19 for conditions in which the context reverberation was always applied in a time-reversed fashion. In this experiment a different pattern of results emerges from each of the efferent control mechanisms. For the LPM-driven model in Figure 4.19b, category boundaries follow the general trend underlying the perceptual data, however, the MPR-driven model in Figure 4.19a fails to simulate the abolition of compensation for reverberation that is observed for human listeners in time-reversed reverberation conditions (cf. Watkins, 2005a, Experiment 5).

The difference in these results essentially arises in the category boundary placement for the condition where the far-distance test-word is preceded by the far-distance context. In reverse-reverberation, the MPR-driven model results in an ATT_{mp} value of 5.24 dB when the context speech is presented in the time-forward direction, and 4.86 dB when the speech signal is reversed. These values are close to those that were derived via the R_{mp} in the time-forward reverberation case (cf. § 4.5.3), and hence cause a similar degree of suppression in the AN response. The knock-on effect of this is that several steps of the continuum are recognised by the MSE template-matching process (cf. eq 4.11) as 'stir' rather than 'sir' (6 steps in the case of the forward-speech, and 5 for the reverse speech condition). Thus, for the MPR-driven model, compensation for reverberation is erroneously simulated for reverse-reverberation conditions.

On the other hand, the attenuation values arising in the LPM-driven model simulation are somewhat lower in reverse-reverberation conditions compared to the forward-reverberation cases reported earlier (cf. § 4.5.3). Here, the R_{lp} value lead to an ATT_{lp} value of 3.98 dB for the forward-speech case, and 3.86 dB for time-reversed speech. Since the far-distance test-word is not sensitive to this variation

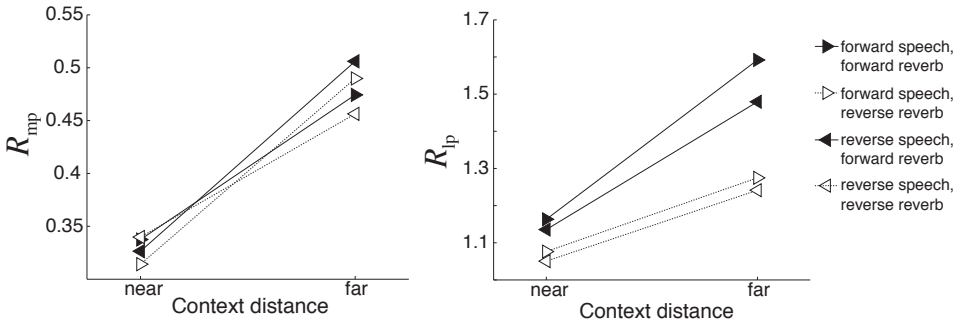


(a) Time-reversed reverberation simulations using MPR.



(b) Results derived using LPM.

Figure 4.19: Experiment M3: Category boundaries for time-reversed reverberation simulations. Time-forward (*left*) and time-reversed (*right*) speech conditions are shown for the MPR metric in fig. 4.19a, and the LPM metric in fig. 4.19b. Model results (*black markers*) follow the general trend underlying data in human listener results in Watkins (2005a, Experiment 5) (*white markers*) for the LPM-driven model. However, the MPR-driven model fails to simulate the abolition of compensation for reverberation in time-reversed reverberation conditions.



(a) MPR context measures **(b)** LPM context measures

Figure 4.20: Dependency of reverberation estimators on context distance, MPR value, R_{mp}, in Figure 4.20a and LPM value, R_{lp}, in Figure 4.20b. Both measures are relatively unaffected by time-reversing the speech; however only R_{lp} is substantially altered by time-reversing the reverberation.

in attenuation (cf. Figure 4.16b, where an attenuation value greater than 4 dB was required to achieve recognition of continuum tokens as anything other than 'sir'), the recognition of the [t] does not improve (i.e., the AN simulation is suppressed, but the suppression is not yet sufficient to make the continuum steps resemble the canonical 'stir' than the canonical 'sir'). As a result, the far-distance context does not lower the category boundary in the reverse-reverberation condition, and compensation for the effects of reverberation is not apparent in the simulation data.

4.6.4 Interim discussion

While the model whose efferent circuit was driven by the reverberation tail measure (LPM) provides a good simulation of the perceptual data in this task, listening conditions in Experiment M3 indicate that the measure based on dynamic range (MPR) is not a suitable metric with which to monitor reverberation and control efferent suppression. Thus, in the current task, just *one* of the reverberation estimators is consistent with independent reports by Watkins (2005a) and Longworth-Reed et al. (2009) that listeners can no longer perceptually compensate for reverberation when the direction of the reverberation was reversed.

Data in the current study is consistent with the explanation of compensation given by Watkins (2005a), that the perceptual compensation effect might arise from a listener's use of reverberation tails. Reverberation tails are effectively abolished by time-reversing the room impulse response prior to convolution with context speech. As such, information about reverberation is no longer contained in the (per-channel) offsets of the surrounding context speech, and thus cannot inform the compensation process. This can be seen in Figure 4.20 (right), where the val-

ues measured by the LPM metric, R_{lp}, are plotted at each context distance for each of the reversal conditions. The time-reversal of speech affects R_{lp} only slightly, and thus does not alter the overall pattern of data resulting from the ATT_{lp} value provided in the subsequent linear mapping in Equation 4.14. On the other hand, time-reversal of the reverberation direction substantially lowers the value of R_{lp} since offsets are sharper and contain proportionally less signal energy (cf. Equation 4.7). This leads to a substantially reduced level of efferent attenuation being applied in the model in the reverse reverberation cases (via Equation 4.14). With reduced attenuation, the low-level reverberant signal energy is no longer removed from the signal, thus the [t] is not revealed and the category boundary position is not recovered.

In this study, the MPR-driven model did not achieve a good match to the human listener data for time-reversed reverberation conditions (cf. upper panels in Figure 4.19): the *abolition* of compensation was not simulated. Figure 4.20 (left) reveals that the context measures, R_{mp}, of the time-reversed reverberation conditions were very similar to those of the time-forward reverberation cases; thus the ATT_{mp} value that resulted (via Equation 4.13) acted similarly in both reverberation directions to suppress reverberant energy in the AN response and increase likelihood of 'stir' results.

Monitoring of dynamic range in reverse reverberation

Here, MPR displays the same problem as objective speech intelligibility room measures discussed earlier in § 2.1.6. Based on the Modulation Transfer Function (MTF), such measures are insensitive to the time-direction of the signal since they disregard the phase component and use only the magnitude. Like MPR (in the temporal envelope domain), the magnitude of the MTF is not much altered by a change in the direction of reverberation, so it cannot accurately reflect subjective experience in these conditions (Longworth-Reed et al., 2009). Nonetheless, discussion of dynamic range and signal modulation frequently occurs in relation to reverberation. One such theory relates to modulation masking[1] (Nielsen and Dau, 2010), where it is proposed that a listener may become accustomed to a particular type of modulation in a signal and thus become less sensitive to it over time. This line of thought seems similarly to underpin the auditory modelling work of Lee et al. (2011), where the channel-by-channel dynamic range of the model's response is adjusted through time to reflect the dynamic range of the input signal. Indeed, Beeston and Brown (2010) also previously reported model performance in

[1] The modulation masking theory is analysed below in § 5.5.5 in light of data presented in Experiment H3.

a reverse-reverberation task that was qualitatively compatible with Watkins' findings using an MPR-based efferent feedback circuit similar in principle to that described above. Use of the MPR metric in modelling listener tasks involving reverse reverberation therefore requires some further discussion.

Since, on the surface, the MPR metric appears to be insensitive to the time-direction of the reverberation, the question therefore arises over why the metric was sufficient to replicate human performance in a reverse-reverberation task in the study of Beeston and Brown (2010). In that study, the time-reversed reverberation conditions were measured as containing a lower degree of reverberation prior to the test-word, and this brought about less compensation than occurred in Figure 4.19a above[1]. The likely explanation for this difference rests in the non-linear processing introduced by the (differing) modelling structures: the auditory nerve representation whose MPR is monitored, $y_{\mathrm{an}}(n, c)$, did not represent the time-reversed reverberation in the same way in the current study that it was observed in Beeston and Brown (2010).

While the OME filter and underlying DRNL model was the same in each study, almost all other stages differed. Input signals were presented to the current model at the level matching listener presentation in Watkins (2005a); in the earlier model they were presented at a considerably higher level. In both models the attenuation was applied simultaneously to every channel at the same level; however the earlier model alone allowed a continual adaptation of this value every 1 ms as the utterance progressed. The 'sir-stir' identification process also differed between models: the output frame-rate is halved in the current version, and the STEP templates and tokens are thus considerably less detailed.

The most significant difference between the models, however, arises in the simulation of transduction at the inner hair cells. Beeston and Brown (2010) used the rate-limiting function (cf. Figure 4.5a) to derive the hair cell response from a fixed BM response in every channel (with minimum BM response $b_0 = 6.417 \times 10^{-4}$ m/s corresponding to the hair cell spontaneous rate of $a_0 = 10$ spikes/s, and saturation $a_1 = 500$ spikes/s arising at a maximum BM response of $b_1 = 5 \times 10^{-3}$ m/s). These values were selected by observing the 1kHz channel response, and were intended (as in the current study) to elicit a response in the AN with high dynamic range, similar to that of a low spontaneous rate fibre. Observed at the level of the STEP response, this process represented a behaviour similar to that seen in Fig 4.8, where the application of attenuation in the 1 kHz channel showed as a shift to higher sound pressure levels in the rate-intensity curve. However, on closer ex-

[1] In Beeston and Brown (2010, Figure 8), the MPR context values displayed a separation between time-forward and time-reverse conditions, similar to that seen above for the LPM measures in Figure 4.20 (right).

amination it was apparent that other channels in the model were not so smoothly responsive across the dynamic range. Rather, when observed at the level of the hair cell, each channel in the Beeston and Brown (2010) model was usually either *on* (i.e. at saturation firing rate) or *off* (i.e. at threshold firing rate). Application of attenuation was enough in some moments to effectively eliminate the response (i.e. go below threshold), and the shift in the rate-level curve observed in the 1 kHz STEP channel occurred due to the averaging process which summarised all time-steps occurring within a 5 ms time period.

The current model improves on the earlier work by implementing efferent attenuation in such a way that the MOC unmasking effect is observable in the firing rate of individual auditory nerve fibres, not only at the overall level of the STEP representation that results across the total nerve population. That is, the rate-level shift in Figure 4.8 is extracted from the model stage *before* temporal integration occurs.

Conflation of effects from the speech signal and from reverberation

The discussion above exposes a further question in regard to Watkins original study: it is not possible to say whether the lack of compensation in reverse reverberation in Watkins (2005a, Experiment 5) is (a) due to a lack of reverberation from the preceding context spilling forwards into the test-word region; (b) due to time-reversed reverberation from the following context speech signal spilling backwards into the test-word region (appearing now as onset ramps instead of offset tails); or (c) both[1]. Moreover, there is a (probably unplanned, but nevertheless interesting) symmetry in the context speech material, "OK, next you'll get {sir, stir} to click on", which broadly mirrored the phonetic classes in the local context (a sequence of 5 phonemes) on either side of the test-word. Thus, for the specific phrase used in Watkins stimuli, the spectral regions of overlap-masking from the context might be similarly distributed when reverberation is applied from either direction:

… [uɹ gɛt] … [tu klɪ] …
… V L S V S … S V S L V …

where vowels, liquids and stops are marked by the letters V, L and S respectively.

In Watkins' stimuli, the test-word and its reverberation always occur in the conventional time-forward fashion. For time-reversed reverberation conditions, this results in a reduced amount of reverberation in the signal region immediately preceding the test-word (compared with the forward-reverberation case, the reverber-

[1] Experiment H2 in the following chapter looks into this question in a further detail, schematically depicting Watkins stimuli in Figure 5.8a and an alternative condition in Figure 5.8b.

ant overlap-masking contribution is reduced). Moreover, the level of reverberation in the test-word varies with the context (reversed) reverberation because there is an overspill of reflected energy (which could perhaps be termed backwards overlap masking) from the final portion of the context ("[...] to click on") backwards into the test-word region. The model described in Beeston and Brown (2010) detected this slight reduction in reverberation. That is, the MPR measure in the context immediately prior to the test-word was reported as being lower, and a low value of efferent attenuation arose as a result, removing the compensation effect which had been simulated in time-forward conditions. Contrastingly, the MPR-driven model reported in this chapter does *not* detect the slight reduction in backwards overlap-masking, thus there is no reduction in R_{mp} for time-reverse conditions, no decrease in the ATT_{mp} applied to the non-linear path of the DRNL, and compensation mistakenly arises as a result.

It is clear from the above discussion that some caution is required when models are tuned to a particular, small set of experimental data. In order to generalise findings to an unseen data set, the key idea is to ensure that the selection of values for any free-parameters is justified by the data available. The model underlying the current simulation has been validated against a range of physiological data, allowing its various outputs to be expressed in international units such as Pascals, meters and seconds (Meddis, 2006; Meddis et al., 2013). Little is yet known about the physiological processes involved in compensation for reverberation, however, so modelling choices in the current study have, by necessity, been 'best guesses' based on what can be inferred from the available psychopysical data. Implications of this are discussed further in the following section.

4.7 General discussion

This chapter has described the development of a computer model that qualitatively matches perceptual compensation for the effects of reverberation as seen in behavioural data collected by Watkins (2005a). The key elements of the model are an estimation of reverberation in the simulated auditory nerve response, and an efferent feedback circuit which selectively attenuates the afferent auditory response based on the reverberation estimate. The goal in an auditory modelling study such as this is to match data in regard to human responses to sound stimuli. As such, research in this area of machine listening is not concerned with reducing the word error rate in recognition, but aims instead to simulate the error patterns in recognition that humans are prone to. The resulting model can be used to make predictions about the compensation mechanism, and these in turn can be tested in psychoacoustic studies.

In the model based on the low-pass mask (LPM) reverberation estimation technique, the efferent control circuit suppresses auditory nerve response in proportion to the signal energy measured in reverberation tails. This model achieves a three-fold replication of Watkins' findings. Firstly, a baseline simulation of perceptual compensation for reverberation is seen: addition of a far-reverberated context assisted the recognition of a reverberant [t] closure, even though this context actually adds more reverberant energy into the temporal region of the signal where the test-word consonant is located. Secondly, the model replicated results whereby compensation persists despite a time-reversal of the context speech. This confirms that a linguistic component is not necessary in the model, since the compensation effect is present even when the semantic content of the signal is destroyed. Thirdly, the model based on reverberation-tails additionally replicates the disruption of perceptual compensation for reverberation in conditions where the surrounding context is time-reversed. This result was of particular interest to study since it is not predicted by objective measures of speech perception in rooms (cf. § 2.1.6), but has been replicated several times for human listeners (Beeston and Brown, 2014; Longworth-Reed et al., 2009; Watkins, 2005a).

4.7.1 Proposed involvement of efferent processing

It should be noted that while functional, this model is entirely speculative: the proposal that the efferent system is involved in compensation for reverberation has yet to be investigated in a physiological sense. The efferent auditory system is a topic of much research at present since it seems to confer a number of benefits in compromised listening environments. In particular, the efferent system may assist with taking stimulus history into account, such that the spectro-temporal context dynamically adjusts the function of the afferent auditory processes. Efferent signals remain incompletely understood at present, both in regard to structure and function, but mounting evidence links the activity of MOC efferents in adjusting the dynamic range of auditory components for noisy listening tasks. The model presented in this chapter relies on the signal processing insight that some of the long-term effect of reverberation causes artefacts that are similar to the presence of additive noise components, where the added reflected energy in the reverberation decay tail raises the noise floor and lessens the dynamic range of a signal (cf. Figure 2.2), and that olivocochlear feedback might thus be additionally relevant to reverberant listening.

Moreover, the model implements a highly simplified auditory system in which a single parameter is used to model the effect of central auditory activity on the dynamics of the inner-ear. Two main types of efferent feedback exist to the auditory periphery. Firstly, at high signal input levels the acoustic reflex (also known as the middle ear muscle reflex) affects the stapes; the signal presentation levels in this

study are insufficiently powerful to activate such a system. Secondly, MOC efferents act at the level of the BM displacement; these are modelled with an attenuation of the efferent pathway of the DRNL filter bank following Ferry and Meddis (2007). Beyond this, there are many different pathways through the more central sites of the auditory system. Each set of auditory pathways has a specialised function (discussed earlier in § 3.3), some of which are now beginning to enter the realm of auditory modelling studies[1] as is discussed further in the following section below. Thus the simplification necessary to represent all central auditory activity with a single feedback parameter, as depicted in the final descending arrow of Figure 3.1b, will clearly be unable to encapsulate the full variety of the multiple auditory processing mechanisms.

While there is little physiological evidence at present to suggest that peripheral efferent effects are important in perceptual compensation for the effects of reverberation, there is evidence in *higher* centres of the auditory system that different processes offer robustness for the effects of reverberation (see for example the work of Kuwada et al., 2012, discussed earlier in § 3.3, which examines monaural responses to reverberant signals in the inferior colliculus). Since many descending pathways exist through the central auditory system, each one might have a knock-on effect on the other processes beneath it, assisting in a continual re-calibration that allows human listeners to improve their chances of perceiving important signal details in varied acoustic environments. Thus, while the processing relevant for perceptual compensation for reverberation may not actually arise directly in the processes governing MOC efferents themselves, effects stemming from higher up in the auditory system are in the end at least partially effected through the MOC efferents since this is the location at which the frequency- and level-dependent behaviour of the cochlea is determined.

Indeed, while the LPM reverberation estimator allowed a simulation of human responses to be achieved with the current model configuration, this should not be taken as a suggestion that the brain is actually calculating the value of R_{lp} in order to attenuate the auditory nerve output. Rather, this chapter has presented a functional model which explores the application of efferent suppression in reverberant listening tasks. Moreover, other model configurations are possible that may yet fit the limited human data equally well and be based on alternative reverberation metrics[2]. Despite the different individual details of the potential feedback regulation

[1] The most recent Meddis model now includes two levels of neuronal response *above* the AN, such that the MOC-efferent signal is derived after the cochlear nucleus chopper response and second-level brainstem response (Meddis et al., 2013).

[2] Discussion returns in § 4.7.4 below to the criteria – examined within a particular spectro-temporal region in the simulated auditory nerve response – that should be met in order for a candidate

metrics, the general result here is consistent with the possibility that the mechanism responsible for controlling dynamic range in the auditory nerve in noisy listening tasks might also be involved in the process of perceptual compensation for reverberation, acting to reduce the cochlear gain when higher levels of reverberation are present. If the model is a suitable representation of reality, however, it seems that a sensitivity to the time-direction of reverberation must be maintained in the monitoring process that drives this efferent feedback signal in this task[1]. Right or wrong, when a prediction arises from a modelling study in this way it can inform the design of future experiments (whether psychoacoustical, physiological or computational), and the process of modelling can thus contribute to furthering our understanding of compensation mechanisms.

The auditory model described is monaural. In everyday listening, however, different groups of MOC efferents are activated by sound through either the ipsilateral or the contralateral pathways and, for a small group of neurons, by sound into either ear (Brown, 2011; Guinan, 2006). The model described in this chapter can thus only simulate *one* of these effects. Though binaural hearing is clearly involved in many reverberant listening tasks, Watkins' single-ear demonstrations of perceptual compensation for reverberation indicate that monaural processing is sufficient for some reverberant speech identification tasks. Simulation of these monaural compensation effects in single-channel systems may eventually benefit a wide range of applications, not only for machine listening applications requiring distant ASR, but also for human listeners who are reliant on speech processing devices such as hearing aids and cochlear implants which are often worn single-sided.

Although spotting a single stop consonant [t] may appear to be a straight-forward task, it has been selected for study by Watkins since it is one of the speech sounds that is most vulnerable to the effects of reverberation (Drullman et al., 1994b; Gelfand and Silman, 1979; Náběl ek et al., 1989) and is therefore likely to be frequently misheard in everyday listening. Moreover, this consonant alone accounts for 5.78% of all phonemes spoken in American English conversational speech

reverberation estimator to be successful in ecosystemic efferent control. In overview, this would necessitate a high degree of consistency when measured across speech stimuli databases and different rooms; and predictable differences in measures to arise due to alternative source-receiver configurations within rooms. Moreover, a sensitivity to the time-direction of the reverberation processing appears beneficial, and finally such a reverberation estimator should ideally not conflate the effects of reverberation with the recent history of the signal content itself.

[1] It appears unlikely therefore the dynamic-range assessments driving efferent feedback suggested in noise-based listening task simulations (e.g., Clark et al., 2012; Lee et al., 2011) could be immediately re-used in reverberant listening tasks unless the nonlinear properties of those auditory models (e.g., adaptation in the auditory nerve) were sufficient to implicitly distinguish the time-forward and time-reversed reverberation conditions.

(Mines et al., 1978), therefore any benefit that could be brought to its interpretation in a machine listening system might have an appreciable impact on overall recognition rates. Generalising Watkins' demonstration of perceptual compensation for the effects of reverberation to a wider set of speech tokens is a central aspect of the work presented below in Chapter 5.

4.7.2 Relation to other efferent processing models

Section 3.5 discussed the state of the art in efferent auditory models, presenting, in the main, ongoing work from three teams of researchers. Each of these has been applied to the task of replicating physiological or psychological measurements of speech or tones in noise; the model described in the current chapter instead attempts to explain measurements of reverberant speech perception by means of auditory efferent involvement.

The starting code base for the current work is MAP (Matlab auditory periphery), version 1.6, as described in Meddis (2006). Described above in § 4.2.2, efferent suppression is then simulated using the cochlear attenuation parameter in the non-linear path of the DRNL as identified by Ferry and Meddis (2007). Guided wherever possible by measured physiological data, the DRNL thus underpins the current work up to the level of cochlear processing. The subsequent model stages take further inspiration from the modelling work by Messing et al. (2009). Hair cell transduction is represented by rectification, filtering and rate-limiting which attempts to marry the vibration resulting on the BM with physiological data reported for a high dynamic-range (low spontaneous rate) nerve fibre as described in Figure 4.5. Here, literature from the perception of equal loudness (ISO226, 2003) suggested an approach whereby a set of pure tones at fixed sound pressure levels be input to the model at different frequencies, and the resulting BM isointensity response contours observed. This provided a detailed map of the afferent DRNL response which could be used to set the spontaneous and saturation firing rate parameters, channel-by-channel, which subsequently scale the inner hair cell response to an unknown input. In this way, a model providing a 50 dB region of sensitivity across the entire hearing range (defined here as 100 Hz to 8 kHz) was built without yet having to select any parameter values by hand.

As a first approximation, efferent attenuation was applied equally in all frequency regions, as done in Ferry and Meddis (2007) and Brown et al. (2010), but now with values based on a measure of the pooled auditory nerve response. Recent perceptual studies however suggest that perceptual compensation is primarily a within-frequency-channel effect (Watkins et al., 2010b, 2011). In principle, the model may be extended so that the reverberation estimation and efferent feedback loop work independently within each frequency channel. Thus, as was seen in re-

cent modelling tasks, further improvements might be expected were the attenuation allowed to vary across channels (as implemented by Clark et al., 2012; Lee et al., 2011; Messing et al., 2009). For instance in Clark et al. (2012), the model version with differing attenuation values in each channel gave better ASR results than the version of the model where attenuation was fixed at the optimum wideband level. However, these models do not yet account for recently described frequency offsets, where efferent signals have recently been reported to be strongest in response to a probe tone presented a half-octave lower than the channel centre frequency (Lilaonitkul and Guinan, 2009), or (for the contralateral reflex) to be responsive only to the 500–1000 Hz spectral region in Zhao and Dhar (2012).

While it is clear that efferent signals can occur on multiple timescales, auditory models currently suffer from a lack of *consistent* data regarding the time courses of these mechanisms. So-called 'fast' effects, acting over 10-100 ms, were modelled by Ferry and Meddis (2007) (reported by Cooper and Guinan, 2003, in addition to 'slow' 10-100 s effects). A more recent experimental report, however, has split these effects into three groups, acting on 'fast' $\simeq$ 70 ms, 'medium' $\simeq$ 330 ms and 'slow' $\simeq$ 25 s timescales, irrespective of whether an ipsilateral, contralateral, or bilaterally activated unit was being considered (Backus and Guinan, 2006). Clark et al. (2012) report making a compromise between the different timescales, modelling just a single effect with an intermediate timescale. A similar approach is taken in the current model, where a window of 1 second duration is observed in order to derive a measure of reverberation that drives the efferent feedback control system.

4.7.3 Further implications of task-based modelling decisions

As described above, it is largely possible to base modelling decisions for the computational model on pre-existing data from a range of physiological or psychophysical experiments. However, as soon as the model is applied to the task of modelling a particular set of data, it is necessary to begin making such decisions since many things are left unknown about how the model should be integrated with the specific listener task that it is attempting to simulate. Here, since physiological data is not available, psychoacoustic data has instead been used to select values for model parameters: attempts were made to bring the attenuation values into reasonable ranges by tuning the parameters governing the linear mapping on the actual human response data in naturalistic conditions (cf. § 4.4.5 and § 4.4.6). The resulting values in unknown data conditions are then computed automatically, and arise in a similar range to those expected from the literature (where a maximum attenuation of around 20 dB is suggested by Cooper and Guinan, 2003; Murugasu and Russell, 1996). Moreover, allowing the efferent attenuation to vary in proportion

to the reverberation level of the signal is similar in principle to the suggestions by other researchers that efferent signal should vary in proportion with the noise level (Brown et al., 2010; Clark et al., 2012; Lee et al., 2011; Messing et al., 2009).

The time course of contextual awareness represents two difficulties in the current experiments. Firstly, there is not yet a consensus in the data about the timescales on which efferent effects are manifest. Secondly, at the time of modelling there was no data available on the specific time course of perceptual compensation for reverberation[1]. Nonetheless, the time-scale of the binaural effect could be seen to be at the lower end of the analysis window timescales (cf. Brandewie and Zahorik, 2010; Longworth-Reed et al., 2009), and was known to occur monaurally within the span of a single utterance (e.g. in Watkins, 2005a, b; Watkins and Makin, 2007b, c). The window duration was therefore set at a somewhat arbitrary 1 second time-frame. The model was then run in different configurations as shown in Figure 4.1 at the start of this chapter. First, an 'open-loop' tuning stage is undertaken in order that the efferent attenuation value providing the best match to human data could be found by an exhaustive search. In subsequent simulations the model is then able to assesses stimuli independently to derive ATT automatically in a 'closed-loop' setting. However, the model presented in this chapter is simplified in that it *only* assesses the context area prior to the text word in order to determine the subsequent attenuation value to apply (more like a 'semi' closed loop, perhaps). The assessment of the preceding speech would be better updated online in a continual fashion by means of a gradually shifting time window, following methods presented elsewhere (for example, in Beeston and Brown, 2010; Clark et al., 2012; Messing et al., 2009). Additionally such a model might include a 'forgetting' function (e.g., an exponential decay with a time-constant that varies inversely with the centre frequency of the channel so that low frequency channels contain longer histories) so that the immediate prior context of the signal contributes more strongly to the quantification of reverberation. Recent work by Watkins et al. (2011) also suggests that a frequency weighting to rate context areas signalling the [t] more strongly would benefit the high-frequency consonant distinction around which the 'sir-stir' identification task revolves, as is discussed below[2].

Additionally, the time window of test-word awareness is not straight-forward either. A keyword spotting technique was considered for the current study, and might have been closer to the manner in which a human listener approaches the 'sir-stir' task. However, this method was eventually abandoned in favour of the simpler

[1] Two studies in particular have recently provided data related to this question: Brandewie and Zahorik (2013), which investigates binaural compensation effects, and Experiment H4 (below), which investigates monaural effects.

[2] Additional support for this is found and discussed in § 5.3.4 in the following chapter as well.

template-based MSE approach (also favoured by Messing et al., 2009). Influenced by the technical considerations behind ASR segmentation techniques, the test area was at first assumed to be defined by the time frames corresponding to the unreverberated test-word position, and later, to be the frames corresponding to just the initial consonant cluster (cf. § 4.4.4). With hindsight, this appears particularly restrictive: the reverberation tails corresponding to the test item itself protrude outside of this region, yet are not assessed by the template mechanism. Moreover, recent work by Watkins and Raimond (2013)[1] suggests that these areas beyond the bounds of the test-word do indeed influence human listeners in this task. Future modellers would thus be advised to increase the duration of the word identification portion considerably to include areas outwith the test-word location itself. Moreover, alternative templates themselves would also be worthy of investigation. Here, unreverberated speech items were used in order that the exact stimuli presented in the near or far distance cases did not recur in testing. Alternatives to this could include, for instance, using acoustically averaged templates (combining near and far distances together), or using perceptually averaged templates (combining all steps of the continuum that were reported by listeners to be 'sir' or vice versa).

A further aspect of the template-matching process might yet be improved in addition, that of the relative contribution of the different spectral regions to the distance metric. Since it is not specialised to the high frequency consonant distinction inherent in the 'sir-stir' listening task, the current model retains an element of generalisability as it stands. The MSE metric described above weights all channels equally; however, the human listeners' perception of whether a word is 'sir' or 'stir' is dominated by energy in the region around 4 kHz (Allen and Li, 2009; Watkins et al., 2011). Moreover, Watkins et al. (2011) suggest that the frequency weighting of both context and test portions are such that the high frequency channels should count for more when a 'sir-stir' consonant identification task is under way (i.e., the metric should be (or becomes) specialised for the task at hand). Earlier work therefore trialled a frequency-weighted version of the MSE which favours high frequencies over low. However, the low-frequency channels in the templates (corresponding to 's' and 'st' in the current model configuration, as seen in Figures 4.15a and 4.15b) essentially contain only the spontaneous firing rate response, and therefore do not contribute much to the wideband distance calculation in any case. In practise then, this weighted metric made little difference to the overall results: numerical values for sir-scores and stir-scores differ slightly, but not appreciably; the simulated category boundary positions do not move as a result.

[1] Findings of Watkins and Raimond (2013) are indeed supported in a connected-speech task in Experiment H3 below.

This exposes a further possibility in regard to the manner in which stimuli are presented to the model. The present work follows that of countless other modelling studies: the stimulus is presented; the response is recorded; the process repeats for each item in the test suite. Whereas the human listeners in Watkins' trials received these items one after another (in a randomised order), the current model shared the simulations out between many different cores in a high performance computing grid. This was deemed a sensible given that the computer model is completely deterministic and outputs a single response for a given sound file, irrespective of what other jobs it has recently been working on. That is, since each trial is a completely separate event, there is no need to randomise the stimuli presentation order, and no knock-on effects can persist from one trial to the next. In the online closed-loop configuration of the model (where the attenuation estimate is continually updated), however, the time course is such that the previous trial may indeed have an effect on the current word identification likelihood if the inter-stimulus interval is short[1]. Therefore, it would be possible to present a randomised sequence of trials to a single instance of the auditory model (much as is done for a human listener) and record category boundaries along the same manner as that used in Watkins (2005a). Further randomisations could be investigated much as multiple people are tested; and since the model is no longer fully deterministic, category boundaries might now vary slightly in the different runs of the experiment. If so, this would allow the current limit on the quantisation to be overcome (Equation 4.12 would now be averaged across several model listeners), perhaps allowing a better match to human data since positions *in between* the labelled continuum steps could now be identified as the point at which responses flip from 'sir' to 'stir'. Since one of the pair of data points used in tuning was itself unattainable in the current model configuration (discussed in § 4.4.6), it can be inferred that this smoothing process might improve results additionally through re-tuning of the attenuation mapping equations themselves.

4.7.4 Reverberation estimation

The previous section has described the fundamental interdependence of the reverberation estimation method on the modeller's decision about the spectro-temporal 'awareness' of both context and test-word, neither of which could be determined empirically since such physiological (or psychophysical) data is not yet available. Such factors ought to be investigated systematically before deciding whether one

[1] Indeed, Watkins and Raimond (2013) criticised the study by Nielsen and Dau (2010) for failing to consider such longer-term effects when they removed the 'near' condition test-words in their 'sir-stir' replication attempt. As a result, listeners could predict the level of test-word reverberation from one trial to the next.

particular reverberation measure is better than another: this process is begun in Chapter 5 which follows. Nonetheless, for a given spectro-temporal region in the simulated auditory nerve response – here defined as 1 second context awareness in all channels, and with all channels contributing equally to the interpretation of both efferent suppression values and test-word consonant identification – we *can* assess whether one such reverberation measure at least meets a few basic requirements that would make it a likely candidate for contributing to the perceptual compensation for reverberation process. The best fit to human data was found in the current model using an estimate of reverberation based on the energy present during decaying tails (using LPM), hence this measure is tested further in an attempt to begin validation of the present model, or to highlight areas in which further work (both modelling and psychophysical) should be directed.

There are a number of criteria that must be met in order for a candidate reverberation estimator metric (such as LPM) to be assessed as a successful in an ecosystemic efferent control mechanism capable of simulating the effects of perceptual compensation for reverberation. The following list is certainly not exhaustive: there is much still to be learned about the compensation mechanism. From the preceding modelling study, however, the following list may be drawn.

1. Consistency across speech stimuli. This study has investigated a single recorded utterance as context phrase into which has been embedded a test-item from the 'sir-stir' continuum (itself created by a process of amplitude modulation). It would be important to assess potential reverberation estimation metric across a wider range of naturally spoken speech stimuli, where vocabulary and talker vary. To begin this process, an alternative speech database is therefore considered below.

2. Consistency across source-receiver distance (SRD). This study implements a linear mapping between the estimate of the level of reverberation and the subsequent value of efferent attenuation applied in the model (see equations 4.13 and 4.14 for example), however these were derived using just two room distances, 'near' (at 0.32 m) and 'far' (at 10 m). It therefore remains to be demonstrated whether different SRDs (i.e., alternative talker-listener configurations) would fit such the same assumption. To begin to address this question, four intermediary positions are examined below in addition to the original near and far distances.

3. Consistency across rooms. Again, a single room has been studied in this chapter so far, but to be considered useful, the modelling results should of course be transferrable to other reverberant environments. A second enclosure is therefore studied below.

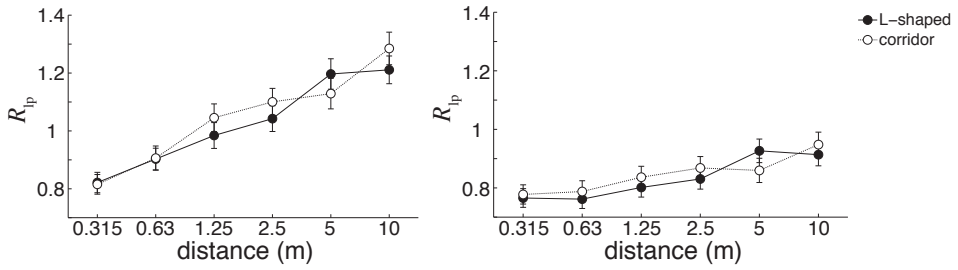
4. Sensitivity to the time-direction of reverberation. As discussed above, several studies have now shown that compensation for reverberation is disrupted when the RIR is reversed prior to convolution so that the reverberation occurs backwards (producing ramps at onsets rather than tails at offsets). To match psychophysical responses the reverberation metric must be sensitive to this difference as well.

Meeting such a set of criteria could be considered a multi-objective optimisation problem, and could be solved systematically – and updated continually – as more criteria are added based on the expanding knowledge of compensatory mechanisms in human hearing. This may lead to alternative metrics being derived, combining the 'useful' parts of each analysis in order to improve the overall match between human and machine data (see e.g., Nikulin et al., 2010). Such an approach is beyond the scope of this thesis, however, where for simplicity the tail-based LPM measure[1] alone has been selected by eye for further study in the remainder of this section since it provided qualitatively similar results to the human listener data in all experimental conditions studied so far.

Figure 4.21 presents the results of an initial investigation which asks whether the LPM measure can meet the reverberation estimation criteria just discussed. To assess whether metric results derived from the 'sir-stir' stimuli hold for more naturalistic speech tokens, 100 utterances were drawn at random from a subset of the Articulation Index Corpus (Wright, 2005)[2]. To investigate whether the LPM measure can generalise beyond the 'near' and 'far' conditions already used, to alternative talker–listener positions in a room, RIRs recorded by Watkins for a series of six logarithmically spaced source-receiver distances (SRDs) were convolved with these utterances. After convolution, the speech stimuli thus sounded as if they had been spoken from 0.32, 0.63, 1.25, 2.5, 5 and 10 m distance in the

[1]The LPM method examines the auditory nerve response simulation and assesses the energy present during the offset regions in each channel. The basic scheme of the method was previously outlined in § 4.3.2, and depicted in Figure 4.10. As the amount of reverberation increases, a reverberation decay tail is increasingly prominent at the channel-based offsets; as tails lengthen more energy is detected within these regions and the measured value, R_{lp}, rises correspondingly. This was seen to occur for the 'sir-stir' stimuli of Watkins (2005a, Experiment 5), as was shown in Figure 4.13. For instance, the offset at around time-frame 60 which was stretched in duration by around 20 frames in length in the far-distance condition (seen by comparing upper and lower figures).

[2]This corpus is described fully in Chapter 5 which follows (cf. § 5.2.1). The data subset comprises the 480 Articulation Index Corpus (AIC) utterances which formed the original speech material selected in Experiment H3 below. In total, 24 test-words (TEST) were each spoken by 20 talkers (12 male, 8 female). Each instance of the test-word has an independently selected set of context words (CW) into which it is embedded, thus forming short utterances defined by the pattern [CW1][CW2][TEST][CW3].



(a) Time-forward reverberation **(b)** Time-reversed reverberation

Figure 4.21: Estimation of the amount of reverberation using the LPM measure, R_{lp}, for 100 speech stimuli drawn at random from the Articulation Index Corpus (AIC). Error bars mark the 95% confidence interval. Conventional time-forward reverberation conditions are presented in Figure 4.21a, where the estimated reverberation increases monotonically with source-receiver distance across the six distances selected in two different rooms. In Figure 4.21b, the increase of R_{lp} with distance remains consistent across the two rooms but its rate is much reduced.

room. Additionally, RIRs recorded at these distances in a corridor were treated similarly to assess whether the metric might generalise from the L-shaped room used in the modelling task to other reverberant enclosures. Energy decay curves for all 12 impulse responses (6 distances $\times$ 2 rooms) may be viewed in Watkins (2005a, Figure 1). After convolution, stimuli were presented to the auditory model following the methods described earlier, and the amount of reverberation present in the signal was estimated using the LPM measure (R_{lp}) from the simulated auditory nerve response over a 1-second time windowed portion of the signal whose location varied with the time-direction of reverberation to ensure that the extent of overlap masking would be comparable in both cases.

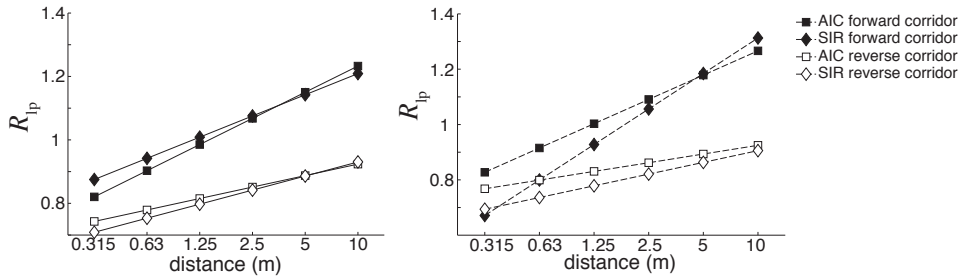
The initial investigation into the generalisability of the tail-based LPM reverberation measure reveals in Figure 4.21 that LPM measures for a new set of speech material behave in a manner that looks qualitatively similar to results obtained earlier with the 'sir-stir' stimuli (cf. Figure 4.20b). Here, R_{lp} – which stands as a proxy for the level of reverberation in the signal – appears to increase monotonically with SRD across the six distances in two different rooms. As desired, measured values of R_{lp} are always at their lowest for the shortest SRD distance, irrespective of the time-direction of room reverberation. As the talker-listener distance increases, the amount of reverberation increases and a large increase in R_{lp} is measured in the conventional time-forward reverberation condition. Values of R_{lp} also rise in the case of time-reversed reverberation, but the overall growth is considerably less pronounced. Overall, these results suggests that an efferent feedback system deriving attenuation proportional to the tail-based reverberation estimation measure may also be capable of simulating the effects of perceptual compensation for reverberation in experiments where using stimuli from a different speech corpus.

However, it is not yet clear whether the mappings for attenuation given by Equation 4.14 could be derived from one dataset (comprising arbitrary speech stimuli and room conditions) and then used directly to model listener tasks on another dataset (comprising a different arbitrary speech database and room conditions). This is considered in Figure 4.22 where a line of best fit is plotted with square markers through the AIC data just shown in the previous figure. Additionally, equivalent values for 'sir-stir' stimuli[1] are shown in each condition using diamond-shaped markers.

In each of two rooms, the R_{lp} measure behaves broadly similarly, increasing more quickly with SRD for time-forward than for time-reversed reverberation conditions as anticipated. In the L-shaped room there is a high degree of consistency between the two speech databases, and the linear predictions (the best-fit lines) lie correspondingly close together. However, results in the corridor reveal some dependency of the LPM reverberation estimation method on the signal content itself. When the reverberation content is very low (i.e., at the nearest distances) the 'sir-stir' stimuli resulted in smaller measures than the AIC analyses, indicating that the LPM measure is rather more strongly affected by channel-offsets due to the speech content and less so by the overall reverberation pattern itself. At longer SRDs the binary mask locating the regions that contribute to the LPM measure contains considerably longer 'tail' segments in each channel (since offsets in the envelopes have been somewhat smeared by the stronger reverberation conditions). In these cases, masked segments contributing to the LPM measure now contain reverberated speech signals whose temporal envelopes have been smoothed, resulting in a loss of detail. Under these conditions it is likely that the two speech databases may thus appear more similar to each other, resulting in closer values of R_{lp} being attained. Thus for the furthest SRD distances examined[2], it appears that the contribution of the reverberation content in the signal is more significant to the measure overall than is the contribution due solely to the speech content itself.

[1] For the 'sir-stir' stimuli, the line of best fit summarises the measured R_{lp} of the simulated auditory nerve response in each of the 11 continuum steps at each SRD in each room condition.

[2] Indeed, reverberant ASR results in Kallasjoki et al. (2014, Table 1) also suggest that this measure is better suited to higher levels of reverberation. In low-level reverberation cases (simulated rooms 1 and 2) the LPM-informed recogniser performed less well than the baseline system. When the level of reverberation increased, however, the LPM-informed recogniser showed some improvement over the baseline results. In the simulated data (room 3), relative improvements of 3.85% for near and 18.62% for far room distances were achieved over the baseline system's word error rate (WER) of 51.95% and 88.9% respectively. For the real-room reverberation condition the relative improvement was more consistent, at 16.95% for near and 17.96% for far, this time above baseline results of 88.71% and 88.31% WER respectively.



(a) L-shaped room **(b)** Corridor

Figure 4.22: Prediction of the amount of reverberation using the best-fit LPM measures (R_{lp}) for 100 speech stimuli drawn at random from the AIC (*square* markers) and for 11 continuum steps for the 'sir-stir' stimuli in Watkins' work (*diamonds*). In each of two rooms, the R_{lp} measure behaves broadly similarly, increasing more quickly with SRD for time-forward than for time-reversed reverberation conditions. In the 'L-shaped' room there is a high degree of consistency between the two speech databases. However, results in the corridor reveal the dependency of the reverberation estimation method on the signal content itself. When the reverberation content is very low (i.e., at the nearest distances) and the measure is more heavily affected by channel-offsets due to the speech content of the signal, and the 'sir-stir' stimuli resulted in smaller measures than the AIC analyses. At longer SRDs the speech envelopes are smoothed and appear more similar between the datasets. Here, it appears that the contribution of the reverberation in the signal content is more significant than the contribution due solely to the speech offsets.

The fact that LPM is affected by the speech source (particularly at near distances) should come as no surprise here since the estimation technique is reliant solely on temporal envelopes: it therefore inevitably conflates signal content and reverberation content because the effects of both of these factors play out in the same measurement domain. At one extreme, transient signals (e.g., a drum kit) will facilitate the perception of channel offsets, particularly at high frequencies. On the other hand, signals which are essentially continuous (e.g., a solo trumpet playing a continuous note of fixed-pitch) would provide the LPM measure with few opportunities at which to estimate the reverberation condition. Here, the number of offsets itself could be used in some way as a normalisation factor to account for different signal types encountered. Alternative computational approaches might attempt to unravel the separate energetic contributions due to the original signal and to the reverberation, either by estimating a clean version of the source signal (e.g. using spectral subtraction) or by estimating a description of the reverberation characteristic itself (by obtaining the room impulse response). On the other hand, the model described in this chapter accepts reverberant audio signals at input, and has no prior knowledge about either the room acoustic or the 'clean' speech signal. Nonetheless, since perceptual effects of reverberation *do* depend strongly on the original sonic material of the (pre-reverberated) source, a system such as this

which directly processes reverberant signals may yet help to explain behavioural data collected in reverberant speech perception tasks.

Biological auditory processing in reverberation is an area presently under intense study[1], but the underlying processes are not yet well understood. What *does* seem to be becoming apparent, however, is that our listening processes are likely to involve a weighted sum of multiple cues that arise, whether monaurally or binaurally, at different levels of the auditory system (e.g., Kuwada et al., 2012; Zahorik, 2002). Auditory models including a combination of concurrent multiple analyses have recently been proposed in order to deal with the simultaneous effects of speech content and reverberation processing, for example combining a number of different monaural cues (Tsilfidis and Mourjopoulos, 2011), or invoking parallel pathways representing binaural unmasking, better-ear listening or cross-ear glimpsing effects (e.g. Jelfs et al., 2011; Lavandier and Culling, 2010; Weller et al., 2014). This combined approach could be introduced within the model structure described in this chapter, for instance by allowing the reverberation mask (which locates the decaying tail regions in the signal) to be derived from the aggregated results of multiple reverberation estimation measures[2].

Fundamentally, however, more sophisticated cognitive models will be needed to describe the task-dependent nature which affects the combination and weighting of cues in different listening exercises. Moreover, experience, attention, active participation and other modalities of sensory information (particularly vision) can also play a role in some kinds of listening task. To understand these process more closely it may become increasingly beneficial in future to model an individual's responses to a given sequence of stimuli, rather than to model the data which is averaged across the whole listener population (and which effectively smoothes a large degree of the natural variability inherent in our responses to sound). These wider-perspectives return to the discussion in Chapter 6. More immediately, the modelling study has highlighted some specific 'unknowns' which require investigation in order to improve the similarity between model responses and human listener data in tasks demonstrating perceptual compensation for reverberation. Chapter 5 begins to address these points in a series of four psychophysical experiments. Initial experiments seek to replicate and extend the major findings of monaural compensation previously discussed, using new talkers and more varied

[1] See, e.g. Bidelman and Krishnan (2010); Bürck and van Hemmen (2007); Devore et al. (2009); Kuwada et al. (2012); Sayles and Winter (2008); Sayles et al. (2013).

[2] Alternatively, a single measure could be evaluated over a number of different timescales, thereby keeping track of its mean and variance as time progresses. In this manner, the attenuation value could be made to depend on averaged measures over the recent past and potentially thereby overcome due solely to moment-by-moment variations in the actual signal content.

speech sounds. Later experiments query which signal portions in particular should 'count' towards the processes of test-word identification and context adaptation.

Chapter summary

This chapter has investigated the proposal that auditory efferent suppression might be implicated in perceptual compensation for reverberation, and has given rise to a number of questions that may be investigated psychophysically.

A computational model of perceptual compensation for reverberation was presented which was able to qualitatively simulate the trends observed in human listener data in a categorical perception task, where the identity of a test item depended on its recent acoustic context. Since mounting physiological data has implicated medial olivocochlear efferents in dynamic range regulation, the reverberation condition of the preceding context was estimated in terms of the ratio of its peak and mean values. When the efferent feedback circuit in the auditory model was controlled with this measure of dynamic range, compensation effects for time-forward reverberation conditions were simulated. However, perceptual data in time-reversed reverberation conditions could not be accounted for. Instead, the successful model in these conditions observed the energetic content of the signal during within-channel offsets, and increased efferent suppression proportionally to the level of reverberation recently experienced. When so-driven by the measure based on reverberation tails the computational model was, like human listeners, broadly insensitive to time-reversals in the speech direction, but heavily influenced by time-reversals in the direction of reverberation. Subsequent analysis of the offset-based metric using a new speech database, a new room and further talker–listener positions suggested that the reverberation estimation technique may conflate the effects of reverberation and speech at the lowest reverberation levels, but was more robust when higher levels of reverberation were present.

Chapter 5

Perceptual compensation for the effects of reverberation on consonant identification with human listeners[1]

Contents

[1] Experiments H1, H3 and H4 were published in Beeston et al. (2014), and listener data from Experiment H2 was presented alongside a modelling study in Beeston and Brown (2014).

5.1 Introduction

Chapter 2 reviewed a body of work showing firstly the detrimental effects that reverberation typically brings about on test-word recognition, and secondly the improvement that seems to arise from a period of 'prior exposure' to the room reverberation condition. In particular, these sections overviewed a mounting body of evidence[1] which suggests that human hearing is remarkably robust to real-room reverberation since it is underpinned by auditory mechanisms that compensate for the effects of reverberation on the speech signal.

Taken together, the behavioural data in this field suggests that the perception of a reverberant sound is influenced by the properties of its context. That is, in order to achieve perceptual constancy, listeners exploit information gleaned from the acoustic context surrounding a test sound, and accumulate this information over

[1] See for example: Brandewie and Zahorik (2010, 2012, 2013); Longworth-Reed et al. (2009); Srinivasan and Zahorik (2013, 2014); Ueno et al. (2005); Watkins (2005a, b); Watkins and Raimond (2013); Watkins et al. (2011); Zahorik and Brandewie (2011).

a period of time. Chapter 4 presented a computational implementation of such a model, achieving a degree of robustness to reverberation similar to human listeners for a single reverberant consonant identification task. Chapter 5 now focuses on aspects of this model that require clarification: the nature of the information that contributes to perceptual compensation for reverberation, and the time period over which it is gathered.

The current chapter presents a series of experiments, each of which was designed with a dual purpose in mind. As will be described below, each experiment seeks to replicate and extend the work of other researchers in order to extend our understanding of the human auditory system: both its inherent ability to cope with realistic reverberant environments, and its tendencies to break down in particular ways under *unrealistic* room conditions. On the other hand, each experiment is additionally underpinned by a particular question arising during the auditory modelling study described in the previous chapter. Thus each perceptual result may potentially lead to an improvement in future auditory models or speech-processing devices. For clarity in the current chapter, however, discussion regarding the implications of this human listener data to machine listening techniques is reserved until the point at which all of the behavioural data gathered in Chapter 5 may be surveyed *en masse*. The current discussion therefore resurfaces in Chapter 6 (cf. § 6.1.3).

5.1.1 Research questions

This chapter presents a series of four perceptual experiments. The first experiment reported below asks whether perceptual compensation occurs with naturally spoken speech utterances when the talker, test and context words may vary unpredictably. In other aspects, the listener task follows the 'sir-stir' paradigm of Watkins, and asks whether perceptual compensation is apparent when only monaural cues are available. In Experiment H1, the ecological relevance of the compensation effect is assessed in terms of the consonant confusions arising in experiments using naturally produced speech from twenty different adult voices. Here, perceptual constancy for the effects of reverberation can be measured as a reduction in consonant confusions that listeners make in a given experimental condition.

Experiment H2 seeks to estimate the ecological relevance of recent reports that compensation for reverberation breaks down when the time-direction of reverberation is reversed, a finding which is of particular interest since it is not predicted by intelligibility standards based on the room's modulation transfer characteristic (Longworth-Reed et al., 2009; Watkins, 2005a). The previous chapter modelled the reverse-reverberation experiment of Watkins (2005a), using a monaural 'sir-stir' continuum category boundary task. Experiment H2 now asks whether Watkins'

findings hold using a task in which the speech is more similar to everyday listening situations in that the talker, test-word and context speech varies unpredictably from one moment to the next.

A third question relates to the respective contributions of context and test-word reverberation to the perceptual compensation effect. Watkins and Raimond (2013) noted that a compensation effect can arise due to information originating from within the test-word itself, but their experiment only examined this for cases when test-words were presented in isolation (i.e., in silence). Experiment H3 asks whether such an effect plays a role in connected speech. Further knowledge regarding this question would benefit future auditory modelling studies by helping to determine the area over which the test-word identity decision should be made.

The final experiment below clarifies the time course of monaural compensation for the effects of reverberation, and provides data which is comparable to a recent study by Brandewie and Zahorik (2013) which investigated the timescale of a binaural compensation mechanism. By applying reverberation to the context over different temporal window durations in the area immediately prior to the test-word, Experiment H4 asks how much of the context phrase must be reverberated in order to compensate for the effects of reverberation in the test-word. Data in this regard would guide future modelling studies in determining the temporal area over which the context metrics estimating the reverberation content of the signal should best be applied.

5.2 Methods

This section describes the selection and processing of the speech material in preparation for the current series of experiments, and the techniques with which perceptual compensation for reverberation is observed and examined.

5.2.1 Speech material

Human listening experiments in this thesis all use read speech material drawn from the Articulation Index Corpus (AIC), LDC2005S22 (Wright, 2005). The database contains around 2000 nonsense test syllables, among them the words ‘sir’ and ‘stir’, each spoken by 20 different talkers. Individually, corpus utterances consist of a single test syllable (TEST) embedded in a sequence of context words (CW),

[CW1][CW2][TEST][CW3]

which are similar in form to that of Watkins' 'sir-stir' continuum utterances, and well suited to the independent reverberation processing that context and test portions of the signal receive in that paradigm (e.g., Watkins, 2005a).

Context words in the AIC are drawn from a limited set, however (8 CW1 pronouns, 51 CW2 verbs, 43 CW3 codas), resulting in a quasi-predictable temporal location[1] for the test-word within the utterance e.g., "people note sir typically" or "I evoke stir precisely". This ensured that the context speech did not contain semantic cues that could be used to predict test identity or override compensation effects (cf. Srinivasan and Zahorik, 2011).

Experiments H1–H2 reported below widen Watkins' 'sir-stir' distinction to examine unvoiced plosive consonants differentiated by horizontal place of articulation: bilabial [p], alveolar [t] and velar [k]. These consonants include a period of brief silence (or low amplitude) that occurs when the airway is restricted by the articulators (the tongue, teeth, lips, and so on), resulting in a temporal dip which may easily become obscured in the presence of reflected sound energy. These consonants are thus particularly susceptible to the effects of reverberation[2]. Heightening this effect further, the initial [s] of Watkins' test-words was maintained in all experiments below, since the stop consonants were found to be even more vulnerable to reverberation when presented after an [s] sound than when they were presented alone (Nábělek et al., 1989).

To allow a direct comparison with Watkins' results, Experiment H1 begins this work using only the [ɜ˞] vowel that features in the 'sir-stir' test-words. The number of test-word vowels was increased in Experiment H3 in order to widen the test material drawn from the AIC and thereby increase the scope of the data obtained from each participant.

5.2.2 Convolution with room impulse responses

The experiments that follow all present listeners with monaural stimuli obtained by convolving speech signals with the left-channel of a binaural room impulse response (RIR) recorded by Watkins (2005a) with a pair of acoustic manikins as shown in Figure 2.3b. RIRs were recorded at two source-receiver distances in an L-shaped office (volume 183.6 m^3), with the two heads directly facing each other

[1] Participants are aware of the generic structure of the phrases, but not the exact selection of words in each trial. Moreover, the duration of the context portion is highly variable from one phrase to the next (further details are discussed below).

[2] Section 2.3.4 reviewed literature showing that reverberation tends to introduce more errors involving place of articulation than manner or voicing (Drullman et al., 1994b; Gelfand and Silman, 1979; Helfer, 1994).

in each of the positions. The two talker-listener configurations were denoted 'near' (0.32 m) and 'far' (10 m) respectively, and resulted in different levels of reflected sound at each distance. The resulting RIRs were analysed using Aurora modules in Audacity (Campanini and Farina, 2009), and are briefly characterised here by measures discussed more fully in § 2.1.4. The initial part of the energy decay curve is characterised in the ratio of early (first 50 ms) to late energy in the impulse response of 18 dB at the near distance. This reduced to 2 dB at the far distance. The later decay portion of both energy decay curves was practically linear (this was shown previously in Figure 2.4b). The slope of this decay determined an energy decay rate of 60 dB per 281 ms at the near distance, and 60 dB per 969 ms at the far distance.

Each one of the experiments that follow implements a different set of experimental conditions in relation to the specific questions it asks. The general scheme underpinning all of these experiments, however, follows Watkins 'sir-stir' paradigm as closely as possible. Test-word and context portions of the speech utterances from the AIC were independently convolved with the near or far distance RIRs, and then recombined to give the same- and mixed-distance reverberation conditions required to independently examine the effects of reverberation on the test-word and on the context[1]. Accordingly, when the stimuli were presented monaurally over headphones to listeners seated in a sound-isolating booth, the sounds at their ear were the same as those for speech arriving from sources nearby or further away in the room.

5.2.3 Headphone calibration

A particular 'sir-stir' reference audio file was used to calibrate the headphone listening level in a sound attenuating booth (IAC single-walled) to ensure that stimuli were presented to listeners in Sheffield at the same level as in the Reading Auditory Laboratory where data collection was carried out by Watkins and colleagues[2]. Stimulus presentation levels in Watkins' lab had previously been set using factory-calibrated binaural heads. Subsequently, an analogue RMS voltmeter (B&K 2425) was set to the slow (1 second) RMS averaging time, and was used to measure the

[1] An example of the resulting stimuli can be seen below in Figure 5.1.

[2] Since the experiments are carried out in a quiet environment and at a fairly moderate conversational speech level, it is not anticipated that small variations in level would alter measurement of compensation effects. Nonetheless, some perceptual attributes of reverberation *have* been shown to depend on the sound presentation level (see e.g., Dubno et al., 2012; Lee et al., 2012; van Dorp Schuitman et al., 2013). Care was therefore taken to calibrate the sound delivery equipment in order to replicate the presentation level used in earlier work demonstrating the compensation effects of interest.

voltage in the wire leading to the headphones while the reference file was played. This provided a peak measure of 60 mV for the reference 'sir-stir' audio. The slow time-constant of the RMS calculation ensured that most of the trial was present during the measurement, and also that peak and average dB levels differed little across the stimulus in question.

In order to present sounds to listeners at the same level in the Sheffield lab, this level calibration stage was replicated as closely as possible with the available equipment. The audio reference file was played (through the iMac computer, M-Audio Firewire Audiophile sound interface and listening booth connectors) and the voltage was again measured in the wire leading to an identical set of headphones[1], mirroring the method described above. This time, the voltage was measured using a digital meter (UNI-T UT70D) set to its slow RMS averaging time (again, 1 second) with peak hold set. The output level of the sound interface was manually reduced until the voltage in the wire to the headphones again matched a peak level of 60 mV, averaged over multiple presentations. A preference file was saved for the sound interface to facilitate future experiment set-up.

5.2.4 Measuring the constancy effect

Various methods have been employed to measure perceptual compensation for the effects of reverberation. As was discussed earlier in § 2.4, the choice of which method to use appears to be principally influenced by the type of data collected. All measures are somehow capturing the relief obtained by the presence of a facilitative context setting in which the reverberated test trials are presented. For instance, Longworth-Reed et al. (2009) tasked listeners with repeating as many words in the sentence as possible, and reported mean word recognition scores under various experimental conditions.

For the 'sir-stir' paradigm, Watkins recorded the benefit of a consistently reverberated prior context as a recovery of the test-word's original category boundary position which was measured by determining the proportion of 'sir' (vs. 'stir') responses (Watkins, 2005a)[2]. A related approach may be seen below in Experiment H4, confusions between words beginning with 's' and 'st' were investigated and visualised in terms of the proportion of [s] responses reported by listeners.

[1] The impulse response that inverted the frequency characteristic of the Sennheiser HD480 headphones used in stimuli presentation was additionally provided by Watkins. Without this, stimuli may have been subject to colouration effects from any imbalance present in the frequency response of the headphones.

[2] The category boundary approach was introduced in § 2.4.1, and was also used in the modelling study presented in Chapter 4.

However, this approach is unsuitable for Experiments H1, H2 and H3, where confusions among *several* consonants were investigated. In these experiments, participant responses were captured and stored in consonant confusion matrices so that the pattern of misidentifications in the data may be taken into account in addition to the correct identifications.

Relative information transmitted (RIT)

Confusion matrices storing each participant's responses were subsequently analysed in terms of relative information transmitted (RIT) (Miller and Nicely, 1955; Smith, 1990). With this information theoretic approach, participants are regarded as information channels receiving input X and responding with output Y. Their information transfer characteristic is then given by

$$RIT = \frac{H(X;Y)}{H(X)} \tag{5.1}$$

where $H(X;Y)$ is the mutual information of X and Y, and $H(X)$ is the self-information (entropy) of X. Probabilities are estimated directly from the finite sample of observations contained in the confusion matrices (Miller and Nicely, 1955; Smith, 1990), where p_x is the probability of occurrence of stimulus x, p_y is the probability of occurrence of response y, and p_{xy} is the probability of the joint occurrence of x and y. In this framework, mutual information

$$H(X;Y) = \sum_{x,y} p_{xy} \log \left(\frac{p_{xy}}{p_x p_y} \right) \tag{5.2}$$

can then be interpreted as the information about the input of the system (i.e. the stimuli) that is provided by the output (i.e. the responses). To calculate the measure of information *transfer*, this value is therefore normalised by the entropy of the input to the system,

$$H(X) = -\sum_{x} p_x \log(p_x). \tag{5.3}$$

The RIT score thus summarises the consonant identification pattern of the confusion matrix with values ranging from 0 for essentially random responses, to 1 for fully consistent responses (perfect transmission).

The RIT metric offers three significant benefits over a simpler measure such as percentage correct. These are outlined below. Firstly, RIT is influenced by the patterning of *all* data in the confusion matrix, whereas percentage correct only considers whether responses are on the main diagonal. It therefore is influenced by the *kind* of mistake as well as by the number (i.e., by quality as well as by quantity).

Secondly, the RIT metric accounts for the difficulty of the listener task so that it is not influenced by chance performance level. This allows confusion matrices of different sizes to be compared in a straightforward way (Smith, 1990). Thirdly, the RIT metric is a normalised measure of stimulus-response covariation that is free from listener response bias (Miller and Nicely, 1955)[1].

Task-adapted compensation metrics

It has previously been shown how different authors – and indeed a single author at different times – have adopted a variety of methods to measure compensation effects, depending on the speech material and listener task selected in the study (cf. § 2.4). The same requirement to adapt the compensation measure to the listener task in use is true for the perceptual experiments that make up this chapter.

To present results in a visually comparable fashion to Watkins 'sir-stir' continuum experiments, consonant identification performance in Experiment H1 was inverted to provide an error (misidentification) measure instead. This was defined as

$$E_{\mathrm{RIT}} = 1 - RIT \tag{5.4}$$

so that an error value of $E_{\mathrm{RIT}} = 0$ indicated complete consistency in the participant's responses, whereas an error value of $E_{\mathrm{RIT}} = 1$ indicated an essentially random set of responses.

Experiments H2 and H3, on the other hand, adapt existing linear contrast methods (Howell, 1982) from subsequent 'sir-stir' studies to examine the compensation effect at each of a number of (non-ordinal) context conditions (e.g., Watkins and Makin, 2007b; Watkins and Raimond, 2013). Here, in each of i different experimental conditions, the participants' RIT scores for the two levels of test-word reverberation were computed – i.e. $RIT_{(n,i)}$ for a near distance test-word, and $RIT_{(f,i)}$ for a far distance test-word – and their difference was found. This quantity was labelled Δ_{RIT}, using

$$\Delta_{\mathrm{RIT}(i)} = RIT_{(n,i)} - RIT_{(f,i)} \tag{5.5}$$

where each of the participant's RIT scores were calculated as in Equation 5.1 above. This resulted in a value of Δ_{RIT} that describes the influence of reverberation

[1] The implication of this may at first seem rather surprising: performance scores on the listener response data would not have been numerically affected even if the visible category labels had been shuffled on screen prior to the experiment beginning. In fact, provided that a listener was entirely consistent about their response behaviour then it would still be possible to achieve a 'perfect' score ($RIT = 1$, so $E_{\mathrm{RIT}} = 0$) without actually responding to a single item in a 'correct' fashion.

on the test-word at each experimental condition. If Δ_{RIT} is large, then the increase in test-word reverberation has brought about a strong perceptual degradation of the signal resulting in many more consonant confusions. If Δ_{RIT} is small, the increased level of reverberation on the test-word has made little difference to the pattern of consonant confusions that were recorded. Thus, constancy is greatest for small Δ_{RIT} values.

The final experiment presented below investigates just two word initial consonant cluster conditions, 's' and 'st'. Analogous to the straight-forward counting of the number of 'sir' (vs. 'stir') identifications in Watkins continuum study (and in Equation 4.12 in the modelling study above), compensation was characterised in Experiment H4 with a simple measure of the proportion of 's' responses given by listeners.

5.3 Experiment H1: Compensation for the effects of reverberation in consonant identification

In Watkins' 'sir-stir' listening experiments, some of which were previously modelled in § 4.4, the [t] consonant distinction was cued largely by a dip in the temporal envelope of the test-word. Attempting to generalise findings from this paradigm, Experiment H1 now asks whether perceptual compensation for the effects of reverberation is apparent in a consonant identification task using natural speech produced by a range of different talkers and with varying speech contexts.

It was previously argued in § 2.3.4 that cues reliant on temporal envelope dips were particularly susceptible to reverberation, since the reflected energy may persist beyond the offsets and fill such gaps in the signal. It seems reasonable to suppose, however, that the sounds found in natural speech may contain cues which are more robust in the presence of reverberation than the amplitude modulation cue that differentiated the 'sir-stir' stimuli in Watkins' continuum experiments.

Miller and Nicely (1955) have shown that cues to *place* of articulation are severely degraded in low-pass filtered speech, causing listeners to make more confusions. Additionally, Watkins et al. (2011) found that listeners gave more perceptual weight to high-frequency bands in their 'sir-stir' experiments, presumably because the temporal envelopes of the two test-words differ the most at high frequencies. Hence, prior to being convolved with room impulse responses, the speech stimuli

used in Experiment H1 were low-pass filtered in order to reduce the likelihood of ceiling effects in listener performance[1].

In order to locate a suitable operating point at which compensation for reverberation may be observed in natural speech stimuli, a range of low-pass cutoff frequencies were selected in Experiment H1. Conversely, the expectation is that perceptual compensation will not be apparent in the more severe (lowest cutoff) filtering conditions because (i) consonant identification is likely to be poor overall and (ii) the filtering step removes temporal envelope information at the higher auditory frequencies that have been reported to be more effective in inducing compensation for reverberation on a test-word.

If perceptual constancy *does* occur in the consonant identification task, then it should become apparent in the following way. Listeners will tend to make few errors in test-word identification when both context and test sounds are heard at the 'near' distance, however the number of consonant confusions will increase when the test-word is presented at the 'far' distance yet the context remains reverberated at the 'near' distance. Subsequently, the number of confusions caused by 'far' reverberation of the test-word should be reduced (i.e., compensation will have occurred) in conditions where the context is also reverberated at the 'far' distance.

5.3.1 Stimuli

Eighty utterances were selected from the Articulation Index Corpus (AIC), with 20 talkers (12 male, 8 female) speaking each of four test-words ('sir', 'skur', 'spur' and 'stir'). Each utterance was segmented by hand using Praat software and a TextGrid file was stored locating the word-boundaries (Boersma and Weenink, 2010). Five versions of each utterance were then created by low-pass filtering with an 8^{th} order Butterworth filter at cutoff frequencies of 1, 1.5, 2, 3 and 4 kHz. The particular range of cutoff frequencies was suggested by Fig. 3 of Miller and Nicely (1955), since RIT varied for place of articulation between the two extreme values.

The method outlined by Watkins (2005a, b) was then followed to create matched and mismatched reverberation-distance conditions for each filtered utterance. Matlab v. 7.5 (R2007b) was used to read word-boundaries from the Praat TextGrid files, and thereby to identify the context and test-word portions of the phrase. Context and test portions of the signal were isolated from each utterance and zero-padded to retain the correct temporal alignment, as illustrated in Figure 5.1). This allowed

[1] Appendix A below asks whether compensation for the effects of reverberation can also be demonstrated without this low-pass filtering step.

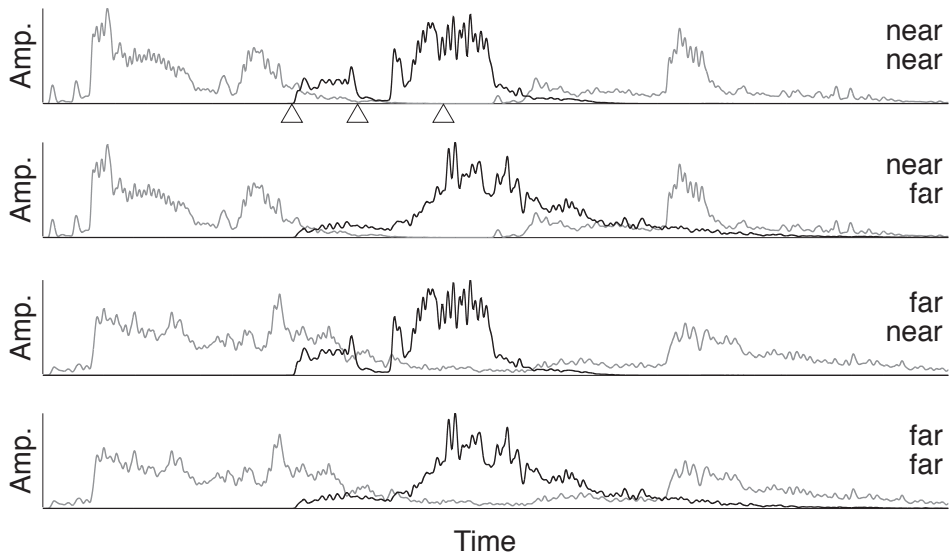


Figure 5.1: Illustration of same- and mixed-distance reverberation conditions for one representative example of the 80 utterances selected for Experiment H1. The traces are amplitudes (Amp.) of low-pass filtered (cutoff frequency 80 Hz) Hilbert envelopes derived from the temporally aligned context *(light line, upper label)* and test-word *(dark line, lower label)* before these two sounds were added, point-wise, to form the experimental stimuli. Before the addition, the context and test-word were independently reverberated at 'near' or 'far' room distances to give, from top to bottom: near-near, near-far, far-near and far-far *context-test* distance conditions. In the top panel, the test-word is annotated with pointers to show, from left to right, the start of frication, closure and voicing.

them to be independently convolved with either the 'near' or 'far' impulse response as required. The resulting waveforms were scaled and summed to give same or mixed-distance phrases, again as indicated in Figure 5.1. The near-near *context-test* condition and far-far condition were calculated first, and their root mean square (RMS) levels were equalised. Amplitude scaling factors were then derived for the context and test portions and these were applied to the mixed distance phrases, resulting in stimuli for the near-far and far-near conditions that had equal RMS levels to the same-distance stimuli.

Finally, as in Watkins (2005a), each signal was convolved with the impulse response that inverted the frequency characteristic of the Sennheiser HD480 headphones used in stimuli presentation, and the signals were scaled *en masse* to be saved as WAV files without clipping. The set of sound files for Experiment H1 thus comprised 1600 stimuli (20 talkers $\times$ 4 test-words $\times$ 5 filter cutoff frequencies $\times$ 2 context distances $\times$ 2 test distances).

5.3.2 Participants

Listening experiments reported in this study were approved by the local ethics committee, and informed consent was obtained from each participant. Sixty listeners without obvious or reported hearing deficiencies participated in Experiment H1. The group comprised fluent native or non-native speakers of English from across the student and staff population of the university. A sixth of the participants were recruited informally from the University of Sheffield's Department of Computer Science, and were not paid. The remainder responded to a university-wide email requesting volunteers, and were compensated for their time. In addition, a further 8 people completed the listening test but were discounted from further analysis since they did not meet the inclusion criterion, set at 90% correct (as in Nábělek and Robinson, 1982) for responses in the 4 kHz filter cutoff condition when both context and test-word were reverberated at the 'near' distance.

5.3.3 Procedures

Stimuli were partitioned evenly among participants so that each person heard every AI corpus utterance just once. This avoided association of the test-word with its context words by ensuring that any given phrase (comprising one of 20 talkers speaking a certain test-word in a particular context sentence) was heard in only a single experimental condition by a single listener. Each participant heard every test-word (defined according to the four initial consonant cluster conditions) at every reverberation distance and at every filter cutoff frequency combination (4 test-words $\times$ 4 distances $\times$ 5 filters = 80 trials). The stimulus partition was gathered and its order randomised immediately prior to presentation to the participant.

Matching the monaural presentation level used for the 'sir-stir' continuum experiments in Watkins (2005a), stimuli were presented to the left ear of listeners at a peak RMS presentation level of 48 dB SPL (measured with a 1-second time averaging constant). Before the experiment began there was a familiarisation session which allowed the participant to become comfortable with the computer interface and the task required of them.

Stimuli were presented with an iMac computer running Matlab Version 7.5 (R2007b) software through an M-Audio Firewire Audiophile sound interface, in a randomised order in a single session lasting approximately 6 minutes. Each experimental trial consisted of a speech context with an embedded test-word, as described above. Listeners identified the test-word with a click of the computer's mouse, positioned while looking through the booth's window at 'sir', 'skur', 'spur' or 'stir' alternatives displayed on the computer's screen. This click also initiated the subsequent trial.

Table 5.1: Confusion matrices summarising 60 participants' responses at three of the 4 kHz lowpass filter cutoff conditions in Experiment H1. Rows correspond to the stimuli presented; columns record the responses. Reverberation conditions are labelled as *context-test* distance. In the near-near condition, no confusions were recorded. In the near-far condition, listeners frequently misreported 'skur', 'spur' and 'stir' as 'sir'. These confusions were largely resolved in the far-far condition.

		near-near				near-far					far-far				
		SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR
4 kHz	SIR	60	0	0	0	SIR	56	1	0	3	SIR	52	1	0	7
	SKUR	0	60	0	0	SKUR	9	46	3	2	SKUR	2	52	0	6
	SPUR	0	0	60	0	SPUR	27	3	27	3	SPUR	4	3	47	6
	STIR	0	0	0	60	STIR	23	2	1	34	STIR	2	0	0	58

5.3.4 Results

Summarising all participants responses for the 4 kHz lowpass filter condition, the confusion matrices in Table 5.1 clearly illustrate the perceptual compensation effect. Consonant identification was not seriously disrupted in the low levels of reverberation present in the near-near *context-test* condition[1], however, confusions were frequent when more reverberation was added to the test-word alone (the near-far condition). Here, the three most numerous confusions were stimuli for 'skur', 'spur' and 'stir' each being reported as 'sir'. However, when the preceding context was also reverberated at the 'far' distance (the far-far condition), the majority of these confusions were resolved.

Perceptual compensation for reverberation

For numerical analysis, participants' responses were recorded in individual confusion matrices, and analysed in terms of their information transfer characteristics as described in section 5.2.4. Figure 5.2 shows the mean and standard error of the E_{RIT} scores at each reverberation distance and each filter condition. A three-way repeated measures analysis of variance (ANOVA) was performed on participants' arcsine-transformed RIT scores (Kirk, 1968) using IBM SPSS Statistics 20 software. All factors were within-subject; two factors had two levels each (context distance and test-word distance) and the third had five levels (filter cutoff frequency). Mauchley's test showed no cases of violation of sphericity.

The right-hand side of Figure 5.2 reveals a monaural perceptual compensation effect in the 3 and 4 kHz filter cutoff conditions. In these two conditions, a far-distance test-word was less often confused when it was preceded by a far-distance context than when it was preceded by a near-distance context (i.e. the upper line

[1] The inclusion criterion removed eight participants who misidentified an item in this condition.

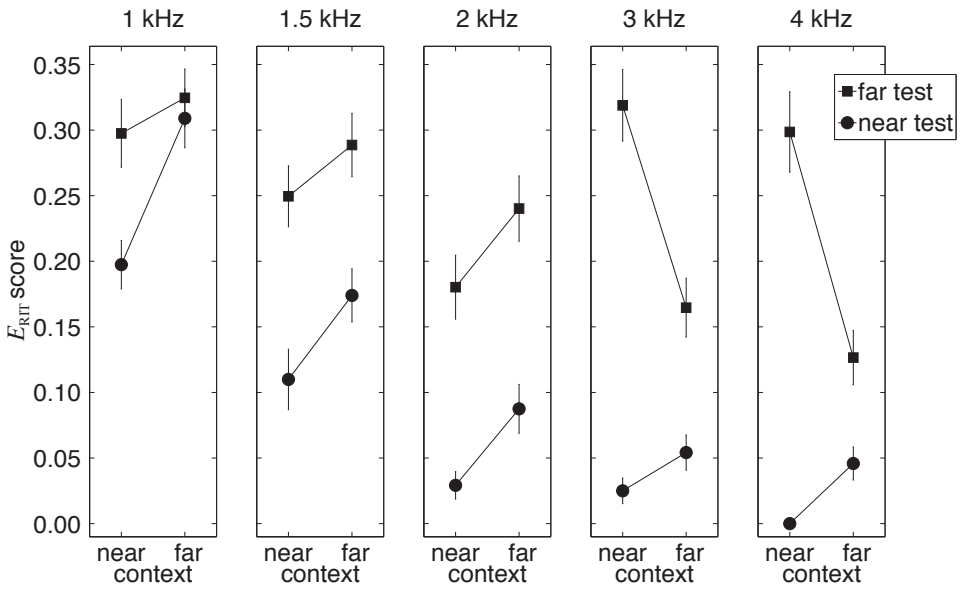


Figure 5.2: Mean and standard error of 60 participants' E_{RIT} scores (cf. Equation 5.4) at the five low-pass filter conditions of Experiment H1. Compensation for reverberation is evident in the downward-sloping upper line of the 3 and 4 kHz filter conditions. In these two conditions, an increased level of reverberation in the context, resulting from an increase in context distance, brought about an improvement in the identification of the far-distance test-words.

slopes downward to the right). For filter cutoff frequencies of 2 kHz and lower, however, a far-reverberated context did not aid identification of the far-reverberated test-word. This pattern of results was indicated in the data by a three-way interaction among the factors for filter condition, test distance and context distance, where $F_{(4,236)} = 5.94$, and $p < 0.001$. Two main effects and three two-way interactions in the analysis, described below, largely arose from this higher-order interaction.

As one would expect, consonant identification was best (i.e. had lowest values of the error metric) in the near-near reverberation condition for each filter cutoff frequency. This can be clearly seen in Figure 5.3, in which the data from Figure 5.2 has been redrawn as a conventional line plot. Additionally, and again as expected, increasing the distance of the test-word from 'near' to 'far' consistently increased the consonant identification error. This gave a main effect of the test-word's distance with $F_{(1,59)} = 306.62$, and $p < 0.001$. Further, and yet again as expected, consonant confusions became more prevalent as the cutoff frequency of the lowpass filter was reduced and the filtering became more severe. This showed in the data as a main effect of the filter cutoff frequency with $F_{(4,236)} = 53.99$, and $p < 0.001$. An interaction of these factors was also found, with $F_{(4,236)} = 9.16$, and $p < 0.001$, indicating that consonant confusions resulting from an increased

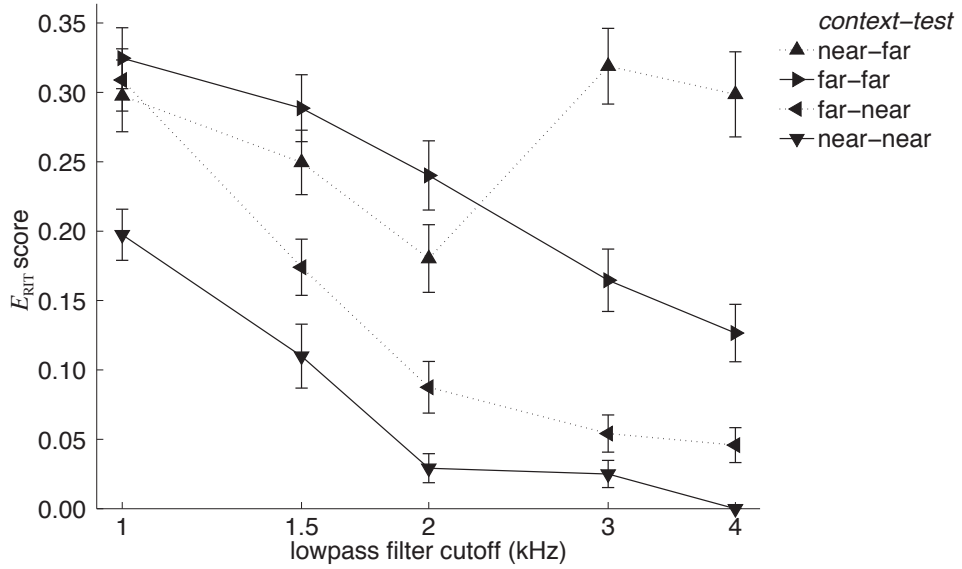


Figure 5.3: Data of Figure 5.2 replotted to show the effect of lowpass filtering on each *context-test* condition. Consonant identification error decreases monotonically with increasing lowpass cutoff frequency, except when the context is 'near' reverberated and the test-word is 'far' reverberated.

level of test-word reverberation were more prominent when higher-frequency information was retained in the signal.

A two-way interaction between the factors for context distance and test-word distance, with $F_{(1,59)} = 28.32$, and $p < 0.001$, indicated that when the far-reverberated context did cause an improvement in consonant identification, this was confined to the far-reverberated test-words. As described above, however, the effect of context reverberation varied across the filter conditions, which showed as a significant interaction of context distance and filter cutoff frequency, with $F_{(4,236)} = 9.78$, $p < 0.001$. There were no other significant F ratios.

Effect of low pass filtering on the near-far condition

As would be expected from the prior literature (e.g. Miller and Nicely, 1955), and as is apparent from Figure 5.3, consonant identification error generally decreased as the lowpass cutoff frequency increased. However, this trend was not observed in the near-far *context-test* condition. Rather, in this condition, consonant confusions increased when more high frequency information above 2 kHz was retained.

A plausible explanation for this finding might stem from the within-channel processing that is suggested to underlie the monaural perceptual compensation for

reverberation effect demonstrated in Watkins et al. (2011). If acoustic features that cue the identity of stop consonants are more strongly apparent at high frequencies, then a higher frequency cutoff condition will present the listener with more conflicting cues than would a lower frequency cutoff condition in the near-far condition. The specific acoustic-phonetics of the consonants used here, all of which are generally characterised by high-frequency cues, may provide some support for this interpretation. For example, Allen and Li (2009) report that [t], [k] and [p] can be defined primarily by their burst frequencies: at 4 kHz for [t]; at 1.4–2 kHz for [k]; and at 0.7–1 kHz for [p].

Listeners' responses for each individual consonant were therefore analysed further in the near-far condition, to seek further support for this explanation. This analysis questions whether the form of the near-far curve in Fig. 5.3 was only apparent because the data was pooled across all consonants, or whether there was a consistent pattern of behaviour for each of the consonants involved.

Participant responses at the near-far reverberation condition were therefore analysed as follows. At each filter cutoff condition, the overall 4×4 confusion matrix was refigured into four 2×2 matrices quantifying, for each consonant stimulus-response pairing, the number of hits, misses, correct rejections and false alarms that each participant reported (Gelfand, 1990, Chapter 8). As before, participants' misidentification scores were then quantified from these smaller matrices in terms of information transfer (i.e. by re-applying Equation 5.4 to calculate E_{RIT} for the new 2×2 matrices).

These results are shown in Figure 5.4, where it is apparent that a similar pattern is repeated across all test-words. A 'pivot point' is found in performance at 1.5 kHz for 'spur' and at 2 kHz for the remainder of the test items. This finding is consistent with the fact that the burst frequency for [p] is the lowest of the consonants considered here (Allen and Li, 2009). In conclusion, therefore, it would appear that the increase in error rate in the near-far condition apparent in Fig. 5.3 is not an artefact in the data caused by pooling across all consonants tested. Rather, it is most likely caused by conflicting high-frequency cues in the context and test-word of the utterance, which reduce the efficacy of within-channel compensation mechanisms.

5.3.5 Interim Discussion

By using more variation among stimuli than Watkins, we have demonstrated that a monaural perceptual compensation for the effects of reverberation is also likely to arise at the higher levels of stimulus uncertainty that tend to be present in everyday listening. In the 'sir-stir' continuum experiments described in § 2.4.1, Watkins attributed the increased number of 'sir' responses in the near-far condition to the fact

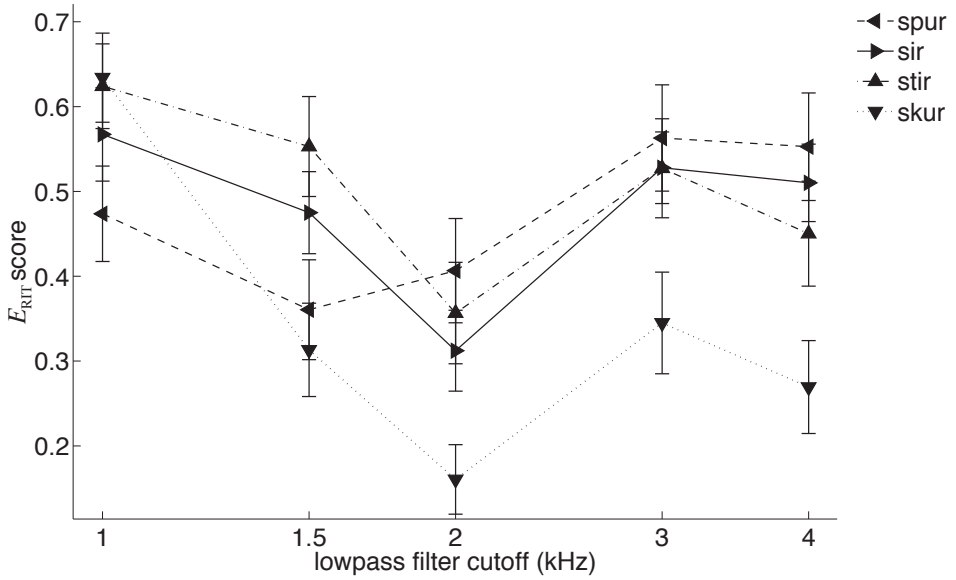


Figure 5.4: Data of the near-far condition in Figure 5.3 re-examined to show the cutoff filter effect on each test-word's responses. All test-words show a similar pattern of performance, with a 'pivot point' at 1.5 kHz for 'spur' and at 2 kHz for the remainder.

that the amplitude modulation which imposed the dip in the temporal envelope was filled with reverberant energy and no-longer cued the [t] of 'stir'. Experiment H1, on the other hand, used recorded speech utterances in which the acoustic-phonetic cues varied naturally (20 male/female talkers speaking 80 test-words, each with a different speech context). Perceptual compensation has recently been reported by other investigators using test signals with a comparable degree of variability (e.g., Brandewie and Zahorik, 2010, 2012, 2013; Longworth-Reed et al., 2009; Srinivasan and Zahorik, 2013), however, these studies were concerned with binaural listener tasks rather than the monaural task employed here.

Watkins' task repeatedly used a single recording of the speech context, which resulted in a highly predictable position for the test-word. Compared with this, there was increased uncertainty in the temporal location of the test-word in the current experiment since context durations varied from trial to trial, ranging from a minimum of 0.31 s to maximum 0.97 s, with a mean duration of 0.61 s. Results in other listener tasks, for instance in investigations of the effects of temporal uncertainty on signal detection (Egan et al., 1961) and gap detection (Green and Forrest, 1989), suggest that this temporal uncertainty is likely to have reduced listeners sensitivity in the current experiment. Despite this, perceptual compensation for the effects

of reverberation was observed, and was qualitatively similar to that reported by Watkins (2005a).

In Experiment H1, perceptual compensation did not occur when high-frequency components were removed from the speech signal. This data is consistent with Watkins et al. (2011) proposal that perceptual compensation occurs in a band-by-band manner, and, in the 'sir' vs. 'stir' distinction, that the high-frequency bands are weighted more perceptually importantly than low-frequency bands. It therefore seems likely that similar mechanisms of band-by-band processing underlie the effect of lowpass filter cutoff frequency seen with the 'sir-skur-spur-stir' distinction investigated in Experiment H1. Nonetheless, at this stage there remains the possibility that listeners were unable to compensate for the effects of reverberation at the lowest cutoff frequency filter conditions because the phonetic content of the context speech signal suffered severe degradation as a result of the filtering operation (cf. Miller and Nicely, 1955). This point is addressed in the next experiment.

5.4 Experiment H2: Compensation for reverberation with time-reversed speech and time-reversed rooms

Two main questions motivate Experiment H2. As just discussed, Experiment H1 could not rule out the possibility that compensation for reverberation might be reliant on phonetic processing. This result would be unexpected from the listener data reported by Watkins (2005a, Experiment 5) which was modelled in Chapter 4, nonetheless, it remains to be tested on more naturalistic speech material. If compensation for reverberation *were* reliant on the linguistic content of the previous speech, then it would be blocked by a process which time-reverses the context speech. On the other hand, if compensation for reverberation did not rely on the linguistic content, then one would expect the time-reversal of the context speech to leave the compensation effect largely intact.

Secondly, Experiment H2 investigates claims by Watkins (2005a) and Longworth-Reed et al. (2009) that time-reversed rooms remove the benefit otherwise gained from prior room exposure in speech-based reverberant listening tasks. As was discussed in § 2.1.6, this result is of particular interest since it is incongruent with predictions of reverberant speech perception given by objective speech intelligibility measures. Watkins' main findings were previously simulated in Experiments M2 and M3: the fast-acting monaural constancy effect occurred within the time-scale of a single utterance for time-forward rooms, but was disrupted in the time-reversed reverberation condition. In those studies, the listener task involved the binary identification of speech tokens that were differentiated by a synthetic amplitude modulation cue (which gave the listener the impression of 'sir' at one

end of the continuum and 'stir' at the other). Relatedly, Longworth-Reed et al. (2009) reported an improvement gained from a consistent room acoustic over a block of trials (sentences 21-40 had better results than sentences 1-20), but only for the time-forward reverberation direction. Their study used naturally spoken speech material, but demonstrated a longer-term effect which was measured with a binaural listener task. The monaural effect has yet to be investigated with natural speech containing real articulatory-phonetic cues. Therefore, the second hypothesis in Experiment H2 is that a monaural influence of compensation for the effects of reverberation will be observed in time-forward but not time-reversed rooms.

5.4.1 Stimuli

This experiment used exactly the same utterances from the Articulation Index Corpus (AIC) that were previously selected for Experiment H1. This comprised a set of 80 utterances, consisting of twelve male and eight female voices each uttering the test-words 'sir', 'skur', 'spur' and 'stir' in phrases with contexts of a quasi predictable nature:

[CW1][CW2][TEST][CW3].

The 4 kHz lowpass filtering process was again replicated here (and in all further experiments below) since it had proved a suitable operating point at which a clear perceptual compensation effect could be observed in Experiment H1. A two-fold reasoning motivates the choice of the 4 kHz condition over the 3 kHz condition (which had additionally permitted compensation to be observed). Firstly, the 4 kHz condition is the least severely filtered condition and is thus closest to 'normality' of those conditions tested. Secondly, this condition crops up repeatedly as a benchmarked standard in speech processing, in part since voice on the telephone network is typically represented with frequencies up to a maximum of 4 kHz.

Four preceding context ([CW1][CW2]) conditions were created by independently varying the speech direction and the reverberation direction, using either time-forward or time-reversed states for each. In all trials, however, the test-word and following context ([TEST][CW3]) were presented with the time-forward speech direction and the time-forward room reverberation condition[1].

[1] In this experiment, treatment of CW3 deviates from the reverberation conditions implemented in Watkins (2005a, Experiment 5), where the context following the test-word was treated in the same manner as the context preceding the test-word. The reasons for altering the implementation of the time-reversals on the following context were driven by primarily by auditory modelling questions. The discussion is therefore reserved until § 5.4.5 below.

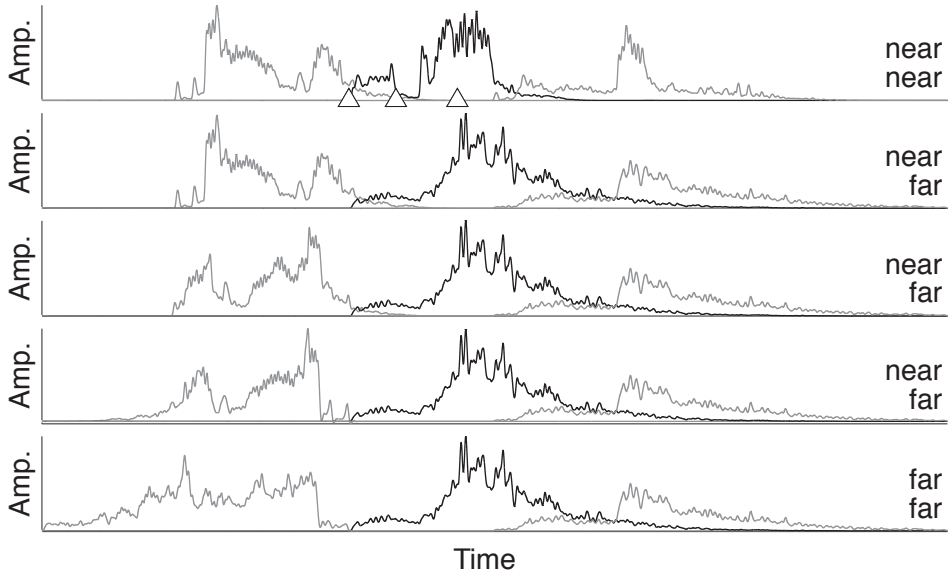


Figure 5.5: Illustration of selected stimulus conditions for one of the 80 utterances used in Experiment H2. The context *(light line, upper label)* and test-word *(dark line, lower label)* are temporally aligned and independently reverberated at 'near' or 'far' distances as before *(upper two panels)*. Here, the following context word [CW3] is reverberated at the same distance as the test-word (cf. fig 5.1, second panel). In other conditions, the time-direction of the speech signal in the preceding context is reversed *(third panel)*. Additionally, the time-direction of the reverberation processing is reversed and applied at either the near *(fourth panel)* and far *(fifth panel)* distance conditions. Other details remain as described for Fig. 5.1.

As in Experiment H1, the current study used the left-channel of the binaural RIRs recorded at 'near' (0.32 m) and 'far' (10 m) distances in an L-shaped office of volume 183.6 m^3(Watkins, 2005a, b). Context and test portions were isolated and independently convolved with either the near or far distance impulse response to create matched and mismatched conditions, as was previously described in § 5.3.1. The consistently near and consistently far reverberated conditions were calculated first, and their RMS levels equalised to obtain multiplying factors to balance the mixed-distance phrases. This resulted in the mixed-distance conditions labelled near-far (near distance [CW1][CW2] with far distance [TEST][CW3]) and far-near (far distance [CW1][CW2] with near distance [TEST][CW3]) which had equal RMS to the same-distance phrases. The resulting stimuli are depicted in fig. 5.5.

Finally, as described by Watkins (2005a), a headphone correction impulse response inverted the frequency characteristic of the Sennheiser HD480 headphones through which the stimuli were to be presented, and the utterances were scaled *en masse* to be saved as .wav files without clipping. The total set of soundfiles for this experi-

ment comprised 1280 stimuli (20 talkers × 4 test-words × 2 speech directions × 2 reverberation directions × 2 context distances × 2 test-word distances).

5.4.2 Participants

Results are presented below for 64 listeners without obvious or reported hearing deficiencies who took part in Experiment H2. This group consisted of University students and staff who were all fluent native or non-native speakers of English. Approximately one third of the participants were recruited informally from within the Department of Computer Science, and were not paid. The remainder responded to a university-wide volunteers email inviting participation in a listening study, and were compensated for their time. In addition, a further 15 people completed the listening test but were discounted from subsequent analysis since they did not meet the inclusion criterion (above 90% correct responses in the forward-speech, forward-reverberation condition with both context and test portions heard at the 'near' distance).

5.4.3 Procedures

Stimuli were again partitioned evenly among participants to ensure that participants heard a given AIC utterance in only one experimental condition. This avoided association of the test-word with its context sentence. Participants heard each test-word at each reverberation distance and time reversal combination, and in addition heard one of the test-words for a second time (4 distances × 4 time reversals × 5 test items = 80 trials).

As in the earlier experiment, listeners were seated individually in a sound-attenuating booth, and sounds were presented monaurally to the left ear over Sennheiser HD480 headphones at a peak RMS-level of 48 dB SPL (measured with a 1-second time averaging window). A familiarisation session took place before the experiment began. This allowed the participant to become comfortable with the computer interface and the task required of them.

Presentation conditions exactly replicated those of the earlier experiment, with each trial consisting of a speech context with an embedded test-word as described above. Listeners again identified the test-word with a click of the computer's mouse, positioned while looking through the booth's window at 'sir', 'skur', 'spur' or 'stir' alternatives displayed on the computer's screen. This click also initiated the subsequent trial. Stimuli were presented in a randomised order in a single session lasting approximately 7 minutes.

Table 5.2: Confusion matrices summarising 64 participants' responses at three forward-speech, forward-reverberation conditions. Rows correspond to the stimuli presented; columns record the responses. Reverberation conditions are labelled as *context-test* distance. In the near-near condition, no confusions were recorded. In the near-far condition, listeners frequently misreported 'skur', 'spur' and 'stir' as 'sir'. These confusions were largely resolved in the far-far condition.

near-near					near-far					far-far				
	SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR
SIR	80	0	0	0	SIR	71	1	2	6	SIR	65	1	1	13
SKUR	0	80	0	0	SKUR	**15**	63	2	0	SKUR	**2**	71	1	6
SPUR	0	0	80	0	SPUR	**17**	8	52	3	SPUR	**2**	8	64	6
STIR	0	0	0	80	STIR	**23**	1	0	56	STIR	**4**	5	0	71

5.4.4 Results

The time-forward reverberation condition of this experiment clearly replicates the main demonstration of compensation for reverberation that was previously observed in Experiment H1. Participant responses are tabulated in summary confusion matrices in Table 5.2, and the corresponding RIT error scores, E_{RIT}, are presented graphically in Figure 5.6. As was seen in the previous experiment, the result of the high accuracy inclusion criterion (above 90% correct) implied that no consonant confusions were recorded in the near-context, near-test condition (cf. Table 5.2, left column). When the context remained at the near distance, but the test-word was heard at the far distance, listeners frequently misreported 'skur', 'spur' and 'stir' as 'sir' (centre column). However, when the level of reverberation in the context was increased to match the far-distance test-word, these confusions were largely resolved (right column)[1].

In order to assess the consistency of results across listeners, a linear contrast was introduced which following the method of Watkins and Raimond (2013) and simplified the data sufficiently for analysis with a 3-factor ANOVA. Difference scores, here named Δ_{RIT}, were calculated for each participant as their RIT at the near-distance test-word condition minus their RIT at the far-distance test-word condition, as was previously described in Equation 5.5. Means and standard errors of this measure are shown in Figure 5.7. A repeated measures ANOVA (all within-subject factors) was performed on these difference scores. Three two-level factors described the preceding context condition: speech direction (forward and reverse), reverberation direction (forward and reverse) and reverberation distance (near distance and far distance).

[1] In the far-far condition, a considerable number confusions have arisen whereby [s] stimuli gave rise to 'st' responses. This point is discussed further in § 5.7.1 below.

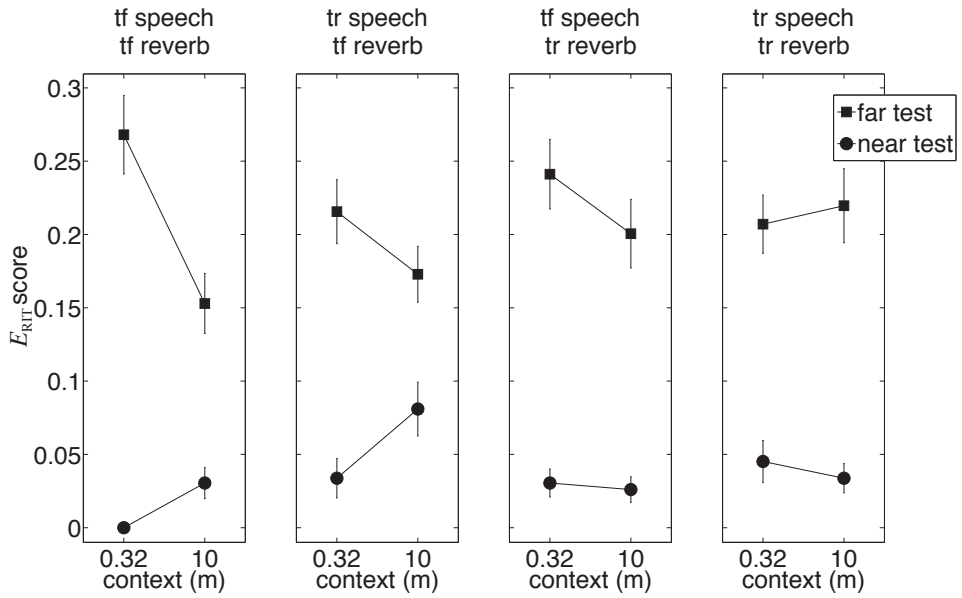


Figure 5.6: Mean and standard error are shown for 64 participants E_{RIT} scores in each of the experimental conditions, with time-forward *(tf)* or time-reversed *(tr)* speech and reverberation processing implemented independently. Here, 320 responses contribute to each data point (5 trials per participant in each condition).

The left-most line in Figure 5.7 shows the forward-speech, forward-reverberation condition that was presented in Table 5.2. For the near distance context condition, a large difference was observed on average between the near-distance and far-distance test-words for each participant, and a high Δ_{RIT} value resulted. On the other hand, the test-word distance was less influential at the far distance context condition, and a low Δ_{RIT} value resulted, which illustrated the benefit that a speech precursor with matching reverberation can bring about in the identification of a reverberant test-word. This showed in the data as a significant main effect of context distance, with $F_{(1,63)} = 13.59, p < 0.001$. A tendency for lower recognition accuracy was observed for reverse speech contexts, with $F_{(1,63)} = 7.38, p = 0.009$. Furthermore, it was evident that compensation only occurred for conditions with forward-reverberation contexts, and not for those with time-reversed reverberation. This appeared in the data as a significant interaction of context distance and reverberation direction, with $F_{(1,63)} = 8.72, p = 0.004$, and can be seen, by comparison, in the two right-most panels of Figure 5.7. In these two conditions, applying far-distance time-reversed reverberation to the speech context caused no statistical difference to the near-distance context. There were no other significant F ratios.

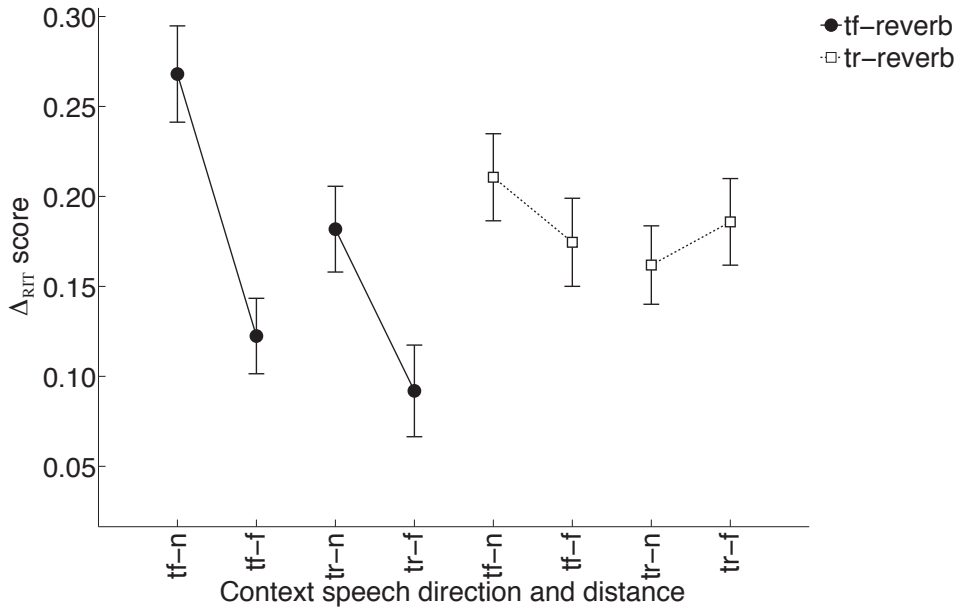


Figure 5.7: Mean and standard error are shown for 64 participants Δ_{RIT} scores at each experimental condition (cf. Equation 5.5). Each data point now characterises 640 responses (i.e., 10 trials per participant). Compensation for reverberation was observed in the steeply sloping line for time-forward reverberation conditions (black markers), with both time-forward *(tf)* and time-reversed *(tr)* speech contexts. Here, addition of a far reverberated context *(-f)* increased constancy in comparison with the near *(-n)* context. Conversely, in conditions with time-reversed reverberation (white markers), the compensation effect was not observed.

5.4.5 Interim discussion

Listener data in Experiment H2 confirms the main finding of Experiment H1 and other researchers (e.g., Brandewie and Zahorik, 2010, 2012, 2013; Longworth-Reed et al., 2009; Srinivasan and Zahorik, 2013; Watkins, 2005a; Watkins et al., 2011) that listeners perceptually compensate for the effects of reverberation. In time-forward reverberation and monaural listening conditions, a large proportion of consonant confusions again occurred when the level of reverberation applied to the test-word exceeded that of the preceding speech context. However, when far-distance reverberation was applied to both, most of these confusions were resolved.

The use of nonsense test syllables here again ensured that the context speech did not contain any semantic cues that listeners could use to predict test identity or override compensation effects (Srinivasan and Zahorik, 2011). Since Experiment H2 showed, however, that perceptual compensation for reverberation persisted despite the time-reversal of the context speech, it now seems extremely unlikely that this constancy mechanism could be reliant on phonetic perception. This finding sug-

gests that the degradation of test signal's phonetic content at the more severe low-pass filter conditions in the previous experiment (H1) would not, in itself, have caused the absence of compensation in these conditions. Rather, it appears that the outright lack of energetic components in the upper frequency channels was significant in that case (cf. § 5.3.5).

On the other hand, the pattern of consonant confusions in this listener data supported claims by Watkins (2005a) and Longworth-Reed et al. (2009) that time-reversal of the room impulse response applied to the speech context *does* prevent perceptual compensation from taking place. This result generalises the category boundary paradigm of Watkins (where a single talker and single consonant distinction were investigated), to consonant confusions among several unvoiced stop plosives spoken by 20 different talkers. Additionally, it hints at the possibility that the binaural effect reported by Longworth-Reed et al. might actually be underlain by monaural processes.

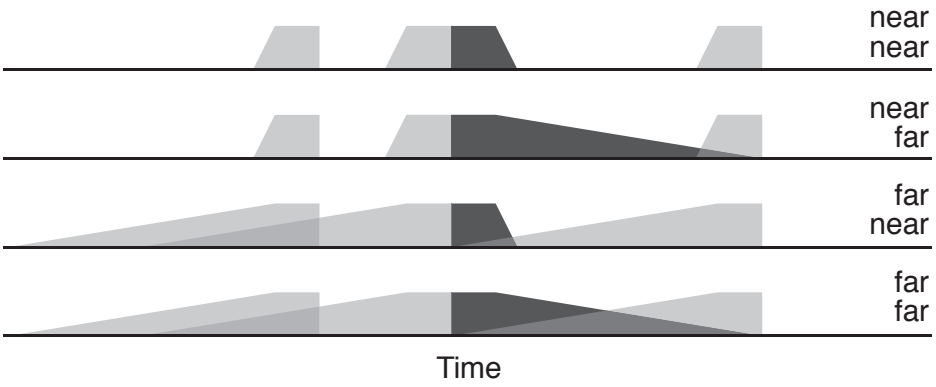
Reverberation processing on the following context

It was not clear whether the absence of compensation for time-reversed reverberation in Watkins (2005a, Experiment 5) was mainly due to effects (or lack thereof) arising from the *preceding* context treatment, or whether it largely resulted from the backward protrusion of reverberation from the *following* context into the test-word region. To help clarify this point, Experiment H2 implemented time-reversed reverberation somewhat differently from Watkins (2005a). Time-reversed reverberation conditions in Watkins' experiment are depicted in Figure 5.8a, above those of the current experiment in Figure 5.8b.

Since time-reversed reverberation from the context following the test-word could protrude backwards in Watkins' experiment, the level of reverberation present during the test-word portion of the signal varied with the reverberation distance of the final context portion in his stimuli. By restricting the time-reversal processes to the *preceding* context only, Experiment H2 ensured that the level of reverberation present in the test-word was not affected by the context distance in time-reversed reverberation conditions. This may be seen by comparison in Figure 5.8b.

Despite the altered reverberation processing on the following context portion, however, Experiment H2 replicated the disruption of the compensation effect in time-reversed room conditions. It would appear, therefore, that the monaural compensation effect was dominated in the current experiment by information arising from the context *prior* to the test-word. The contribution to compensation of effects due to reverberation in the test-word region and context region located *after* the test portion remains to be investigated.

(a) Time-reversed reverberation conditions in Watkins (2005a, Experiment 5).



(b) Time-reversed reverberation conditions in Experiment H2.

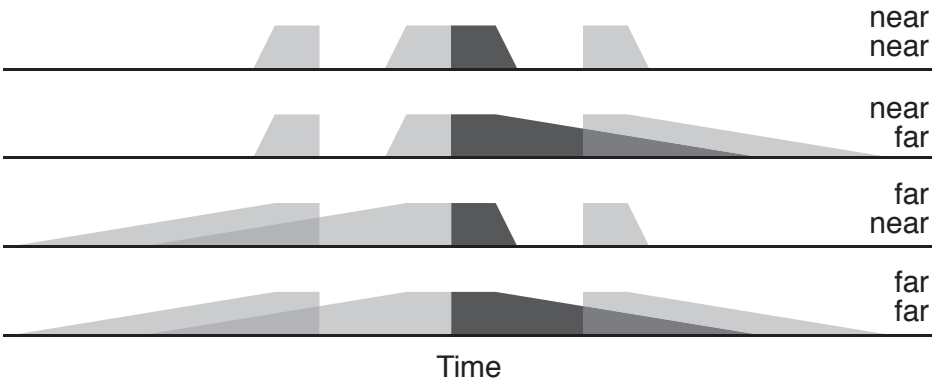


Figure 5.8: Schematic illustration for reverse reverberation stimuli. The test-word (dark) and surrounding context (light) were independently reverberated at 'near' or 'far' room distances (depicted with short or long tails respectively) to give, from top to bottom: near-near, near-far, far-near and far-far *context-test* distance. Fig. 5.8a shows the time-reversed reverberation conditions in from Watkins (2005a, Experiment 5). Fig. 5.8b shows the corresponding conditions in Experiment H2, two of which were previously plotted for a real stimulus in the fourth and fifth panels of Figure. 5.5. Preceding context reverberation and test-word reverberation conditions are implemented identically in both experiments, but reverberation of the *following* context differs. In Watkins, reverberation from the following context impinges (backwards) on the test-word in time reversed reverberation conditions. Therefore, the level of reflected energy in the test-word varied with the context distance. In Experiment H2 the final context portion is presented with time-forward reverberation in all conditions, thus the level of reflected energy in the test-word portion of the signal does not vary with the context distance.

It is interesting to note here that Watkins and Raimond (2013) have recently reported that compensation effects may include 'intrinsic' influences – those which originate from material located within the test-word itself (including its reverberation tail) – in addition to the more commonly studied 'extrinsic' effects (like those seen in Experiments H1 and H2) which originate in the preceding speech context. The stimuli in the study by Watkins and Raimond used the 'sir-stir' continuum words, but presented these test items in isolation, without the speech context. It is uncertain therefore whether such an effect would be replicable with continuous speech, particularly using a source of highly variable speech material such as the AIC where the test and context portions of the signal vary unpredictably from trial to trial. The following experiment therefore asks whether an intrinsic compensation effect is apparent when a source of extrinsic information is available in addition via the preceding context words.

5.5 Experiment H3: An intrinsic effect

Experiments H1 and H2 demonstrated an 'extrinsic' effect of compensation mediated by the preceding speech context (termed extrinsic since it was brought about by factors outwith the test-word in question). On the other hand, Watkins and Raimond (2013) reported an 'intrinsic' compensation effect[1], so-called since it originates from reverberation information due to the signal content of the test-word itself. However, in these studies reverberant test-words were presented in isolation (i.e., in silence). Experiment H3 now asks whether intrinsic information plays a role when extrinsic information is additionally present.

In Experiment H3 the preceding context phrase received three different treatments: near-distance reverberation and far-distance reverberation (replicating the baseline conditions of Experiments H1and H2), and thirdly, a silencing treatment which removed the preceding context cues and gave conditions similar to Nielsen and Dau (2010) and Watkins and Raimond (2013).

The test-word in Experiment H3 was potentially subject to two different types of processing. As in the earlier experiments, the test-word was first reverberated at either the near or far room distance. Secondly, the reverberant test-word was gated in some conditions following the method of Watkins and Raimond. By shortening the reverberation tail that follows the test-word's final vowel, the intrinsic information content was reduced. Stimuli are described in detail below, and selected

[1] Indeed, Nielsen and Dau (2010) have shown listener results consistent with this effect, though their analysis did not represent the data in this fashion.

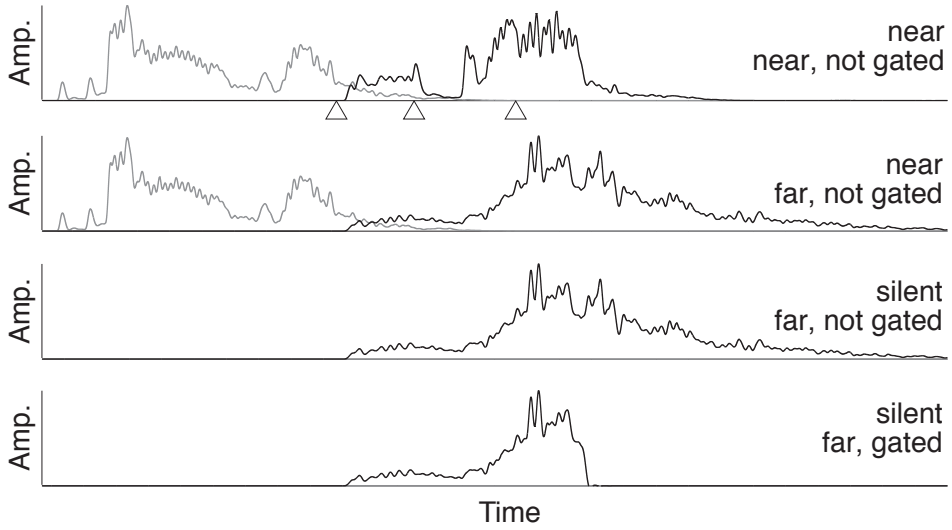


Figure 5.9: Illustration of selected stimuli in Experiment H3. Context phrase *(light line, upper label)* and test-word *(dark line, lower label)* were independently reverberated at 'near' or 'far' distances as before *(upper half)*. In other conditions the context was silenced *(lower half)* and the test-word was gated *(bottom panel)*. Other details are the same as Fig. 5.1.

conditions are displayed in Figure 5.9 to depict the separate context silencing and test-word gating processes.

Watkins and Raimond reported a gating effect when the test-word was presented in isolation (i.e. in silence, without a context), however, they did not investigate this effect for conditions in which gated test-words had speech precursors. As detailed in § 2.4.1, the carefully controlled stimuli of the 'sir-stir' continuum provide a sensitive paradigm in which to investigate such effects yet despite this, the gating effect they measured, though robust, was relatively small. The implication of this finding is that a ceiling effect may exist: in cases where an extrinsic compensation effect arises from the speech precursor (i.e., in the far-distance context condition), then there may be no remaining headroom in which to measure a further intrinsic effect exposed by the gating process.

To avoid this likely ceiling effect, Experiment H3 therefore asks whether the intrinsic effect exposed by the test-word gating process is present for naturalistic speech utterances with near-distance and silent precursor conditions only[1]. However, in order to maintain uncertainty about the level of reverberation occuring in a given

[1]Planned comparison tests are used to address the data conditions of interest (excluding the far-distance data) following the initial ANOVA data summarisation stage.

trial, far-distance context conditions were also included in the experimental stimuli as before.

If apparent, the intrinsic effect would be measurable in the following way: in conditions where the reverberation tail of the test-word promotes some intrinsic effect of compensation, listeners would make more confusions when that tail is removed by gating, and fewer consonant confusions when the test-word's reverberation tail remains intact. Moreover, if an effect of gating *is* apparent (i.e., if listeners' ability to identify the test-word consonant reduces in gated conditions), then it would suggest that the reverberation tail following the test-word's final vowel contributes to perceptual compensation even though it occurs some time *after* the stop consonant that distinguishes the test-word.

Crucially, it should be noted that this experimental design does not gauge the full size or importance of *all* intrinsic sources of reverberation information. Instead, the gating operation evaluates only the contribution of the vowel-end tail. Other aspects of reverberant test-words, notably the tail that follows the test-word's consonant, are not evaluated at present but should not be discounted since they are located closer in time to the frication part of the sound.

5.5.1 Stimuli

Utterances selected for Experiment H3 were similar in form to those of the earlier experiments, however, in this case they were truncated to remove the final context word. The stimuli now were of the form

[CW1][CW2][TEST]

which facilitated the independent manipulation of the reverberation elicited by the test-word and the context.

Additionally, to gather more consonant confusion data from each participant, the set of test-word vowels was expanded further in this experiment. However, since the perception of [p] and [k] (but not [t]) has been shown to depend on the following vowel (Liberman et al., 1952), care was taken to ensure that the following vowel would have similar effects across the new set of test-words. The vowels {[eɪ], [iː], [ɛ], [ɪ], [æ], [ɝ]} were finally selected from the AIC since coarticulatory variation is not prominent among front vowels. The last of these is the vowel from the 'sir-stir' paradigm that was previously used in Experiments H1 and H2. The corpus contained appropriate data (i.e., all four test-words) with two other vowels in addition to the six finally selected. The vowel labelled [ɑ] was rejected, however, because it was spoken inconsistently by the 20 talkers (Wright (2005) reports frequent mergers of the two back vowels [ɑ] and [ɒ]). Additionally, the vowel [oʊ]

was not included since it was the only back vowel remaining in the group. Using the same initial consonant groups as in the previous study, the current experiment thus employed 480 AIC utterances (20 talkers $\times$ 4 consonants $\times$ 6 vowels).

The process of locating word boundaries was partially automated in Experiment H3 due to the large number of utterances involved, but considerable care was taken to position word boundaries in order that the resulting speech stimuli sounded naturally spoken after truncation and reverberation. The AIC transcripts were initially expanded into phone sequences using the Carnegie Mellon University pronunciation dictionary[1]. Subsequently, a hidden-Markov model-based automatic speech recognition system[2] was then used in conjunction with TIMIT-trained monophone acoustic models (Lee and Hon, 1989) to force-align each phone sequence with its corresponding speech signal. This allowed the test and context regions of each sound file to be identified. To overcome quantisation errors (due to the 10 ms frame rate of the recogniser), the word boundaries were subsequently checked using Praat (Boersma and Weenink, 2010) and amended by hand where necessary.

After lowpass filtering throughout at 4kHz as before, same- and mixed-distance stimuli were again created following the reverberation processing methods previously described (cf. § 5.3.1), with scaling factors were calculated across CW1, CW2 and TEST in order to ensure that the level of context and test portions was balanced in mixed-distance conditions. Two such conditions are illustrated in the upper two panels of Figure 5.9.

Silent context conditions are illustrated in the third panel of Figure 5.9. Here, the preceding context words CW1 and CW2 were omitted and silent intervals, SIL, of equal duration were introduced so that the utterances now comprised [SIL][SIL][TEST]. This further increased the uncertainty in the temporal location of the test-word since the preceding context varied in duration for each utterance (ranging from 0.23 s to 1.24 s, with a mean duration of 0.65 seconds). As a result, any quasi-semantic cues from the preceding pronoun and verb were removed in these conditions.

The gating process applied to the test stimulus was based on that used in Watkins and Raimond (2013) and is illustrated in the final panel of Figure 5.9. The gating function was created using the right-hand-side of a Hann window of 10 ms duration[3], and was applied to 'near' and 'far' reverberated versions of the test-word, with the function time-aligned to begin its descent at the end of the test-word (at the

[1] Carnegie Mellon University pronunciation dictionary, v. 0.7a, accessed 1 Jul 2010 at http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] HTK v. 3.4.1, accessed 1 Jul 2010 at http://htk.eng.cam.ac.uk

[3] The gate duration used in Watkins and Raimond (2013) was comparatively shorter at just 1 ms.

position in the AIC utterance at which the following context word was truncated). Hence, the reverberant tail following the test-word's final vowel was cropped off without shortening the test-word beyond its initial unreverberated duration. Other aspects of reverberant test-word, particularly the reverberation tail following the test-word's consonant, remained intact and provided reflected energy that was temporally proximate to the crucial frication part of the sound.

Finally, the twelve versions of each spoken utterance were equalised in RMS level, the headphone correction was applied and the sound files were saved as previously described in § 5.3.1. The set of sound stimuli for Experiment H3 thus comprised 5760 sound files (480 utterances $\times$ 3 context conditions $\times$ 2 test-word distances $\times$ 2 gate conditions).

5.5.2 Participants

Sixty participants from among the student and staff population of Sheffield University were recruited by email for Experiment H3, and were compensated for their time. A further 10 people participated but were discounted from further analysis. In one case this was due to a reported hearing impairment. In the remaining 9 cases this was through failure to meet the inclusion criterion at the control condition (achieving above 90% correct responses at the near-context, near-test distances, with full reverberation tails following the test-words).

5.5.3 Procedures

Stimuli were once again partitioned among listeners to avoid repetition of an item in different experimental conditions that would otherwise increase association between the test and context portions of the phrase and thereby assist identification of the test-word. Every participant heard 480 different phrases, comprising 40 items in each of 12 experimental conditions. Vowels were divided evenly across the listener group, and stimuli rotated among participants so that each listener heard every test consonant either once or twice in each condition and the set was balanced overall. In cases where listeners heard the same test consonant twice in a given experimental condition, the two instances were from different phrases (and thus were spoken by different talkers). In this experiment, participants were not required to identify the test-word completely. Rather, they were instructed to identify the initial part of the word by choosing among buttons labelled 's', 'sk', 'sp' or 'st'. For each participant, the 480 stimuli were presented in a randomised order in a single session. Participants were encouraged to take short breaks whenever needed, and the experiment was typically completed in around 25 minutes. Other aspects of stimulus presentation were as described in section 5.3.3.

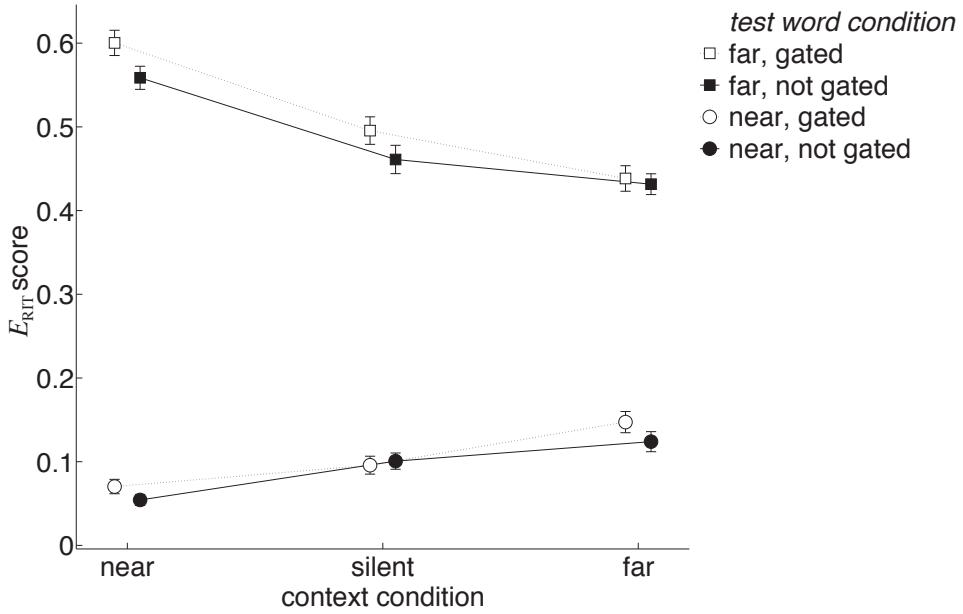


Figure 5.10: Mean and standard error of 60 participants' E_{RIT} scores in Experiment H3. Conditions in which the reverberation tail following the test-word was removed by gating are shown with white markers. The not-gated conditions that preserve intrinsic information are shown with black markers.

5.5.4 Results

Participants' responses were once again recorded in confusion matrices and analysed in terms of their information transfer characteristics. Overall, the main extrinsic compensation effect of Experiments H1 and H2 was replicated in baseline conditions, as shown by the mean and standard errors of participants' E_{RIT} scores in Figure 5.10. Firstly, increasing the test-word distance from near to far gave rise to a large increase in the number of consonant confusions in the listener data. Secondly, in comparison with the near-distance context condition, extrinsic compensation at the far-distance context condition brought about a reduction in the number of misidentifications recorded for both gated and not-gated stimuli. Since the final context word was omitted from all stimuli used in this experiment, extrinsic information from the context portion *following* the test-word was clearly not required in order to achieve perceptual compensation for the effects of reverberation.

Following the linear contrast previously used in Experiment H2 and in Watkins and Raimond (2013), participants responses were again transformed prior to analysis in this experiment. As before, the difference between participants' scores for the two levels of test-word reverberation was calculated using Δ_{RIT} defined in Equa-

tion 5.5, and used to characterise the compensation effect: constancy is considered to be greater when this difference is small.

This linear contrast avoids a number of potential confounds that might otherwise have been a concern in data analysis. Looking left to right[1] in Figure 5.10, the temporal uncertainty of the test-word increases between the near-distance and silent context conditions. There is substantial evidence that listeners are more sensitive to a sound if they know when it is likely to occur, as shown by the effect of temporal uncertainty on signal detection (Egan et al., 1961) and gap detection (Green and Forrest, 1989). Therefore the move from near-distance to silent context conditions would be anticipated to bring about an increase in the overall number of consonant confusions since the silent context cannot cue the location of the test-word. The reduced degree of temporal uncertainty at far distance contexts might suggest a decrease in error on continuing towards the right from silent to far contexts (cf. fig. 5.10). However, in this comparison the presence of overlap masking (Nábělek et al., 1989) from the context in the far condition would be likely to bring about an increase in consonant confusions. Concurrently, extrinsic compensation effects would be expected to decrease the overall error rates of far test-words.

At each context and gate condition, difference scores Δ_{RIT} were therefore calculated for each participant as their RIT at the near distance test-word minus their RIT at the far distance test-word (cf. Equation 5.5). Means and standard errors of this measure are shown in Figure 5.11 (*left*).

Participants' difference scores were analysed with a 2-way repeated measures ANOVA (all within-subject factors), using one factor for test-word gate condition (with two levels: gated, not-gated) and a second factor for preceding context condition (with three levels: near, far, silent). Mauchley's test showed that conditions of sphericity were met. An overall reduction in Δ_{RIT} was observed when moving from left to right in Figure 5.11 which suggested that constancy increased in silent-context conditions, and increased further in far-distance contexts. This large, extrinsic compensatory effect showed in the data as a significant factor for context type, with $F_{(2,118)} = 90.61$, and $p < 0.001$. There were no other significant F ratios.

It was argued above that a ceiling effect is likely to prevent gating effects from being measurable in conditions with far-distance contexts (where an extrinsic compensation effect dominates). The far-distance listener data points were therefore excluded in the planned comparison stage (Howell, 1982, Chapter 12). Moreover, since here we are interested in the effect of gating, the remaining silent and near

[1] Since the horizontal axis describing context treatment is non-ordinal in this experiment, the left-to-right positioning of context types was selected purely for convenience.

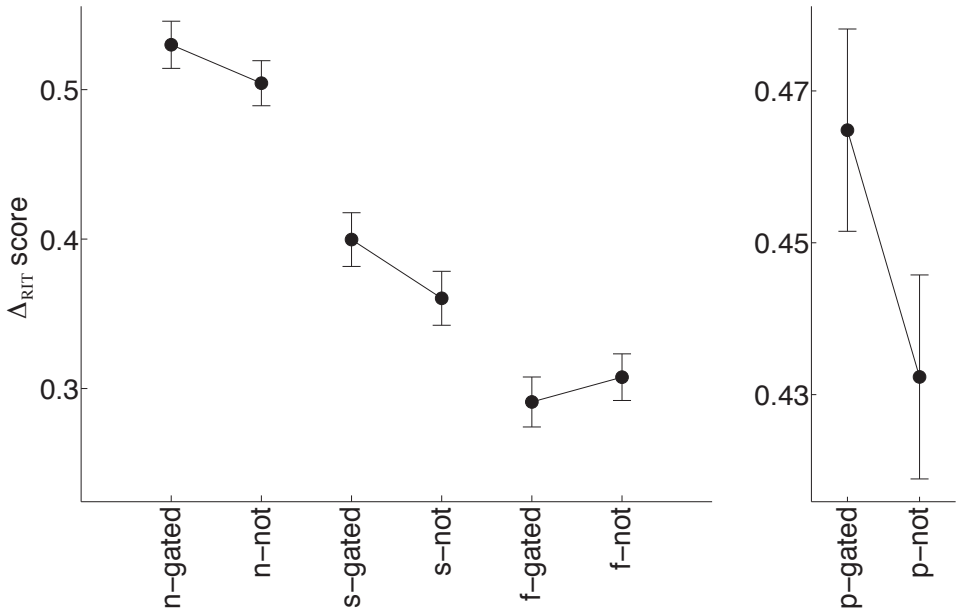


Figure 5.11: *Left* – Mean and standard error for 60 participants' Δ_{RIT} scores for near (*n*), silent (*s*) and far (*f*) contexts at each gate condition in Experiment H3. *Right* – Pooled results (*p*) show some effect of gating on the test-word's vowel, suggesting that intrinsic effects may have helped to disambiguate far distance test-words in the near- and silent-context conditions. Note that the ordinate scales of the two panels differ.

context conditions were pooled to undertake the linear contrast depicted in Figure 5.11 (*right*). Here, a paired-samples t-test revealed that there *was* some effect of test-word gating in these conditions, with $t_{(119)} = 2.43$ and $p = 0.017$. Thus, despite the high variability of the speech stimuli used, this dataset nonetheless constitutes further evidence in support of a role for intrinsic information which seems to help the listener to identify reverberant test-words in the near- and silent-context conditions.

5.5.5 Interim discussion

Stimulus conditions in Experiment H3 were designed further investigate the results of Nielsen and Dau (2010) and Watkins and Raimond (2013).

To explain their data, Nielsen and Dau put forward a 'modulation masking' theory which proposed that the dip cueing the /t/ in a reverberant 'sir-stir' continuum test-word could be made less apparent (i.e., masked) by a preceding context, provided that that context contained a sufficient degree of modulation. By this theory, the near distance context can be thought to have a large degree of modulation (i.e. little

reflected energy to smooth over the spectro-temporal gaps), which would thereby induce substantial masking of the /t/ in the far-reverberant test-word. The near distance context would thus promote more 'sir' responses from the listeners (and, by inference, a greater degree of confusion in the AIC data). On the other hand, the far distance has a lesser degree of modulation since the higher degree of reverberation has somewhat smoothed the signal, and filled gaps with reflected energy. The lower modulation content of the far context condition would therefore promote less masking of the /t/ dip, and permit more 'stir' responses (and again by inference, fewer consonant confusions). For silent contexts, where there is no modulation and thus no modulation masking, Nielsen and Dau's proposal would predict a well-defined plosive dip, resulting in still fewer confusions in the AIC data. However, a different pattern of responses emerged in the listener results of Experiment H3. Here, far test-word consonant confusions were indeed less frequent for silent contexts than for near-distance contexts, but confusions were actually reduced still further by the presence of a far-distance context. The current result therefore provides further evidence[1] that the 'modulation masking' theory of Nielsen and Dau (2010) does not explain the compensation for reverberation paradigm.

Listener data in Experiment H3 supports an idea that perceptual compensation for reverberation may be influenced by several factors. An improvement in far test-word recognition accuracy was observed when moving from near context to silent context conditions which clearly cannot be attributed to the extrinsic effects elucidated in Experiments H1 and H2 since the preceding context cues have been removed by the silencing operation. Rather, the increase in performance can be attributed to an intrinsic compensatory influence. By examining the intrinsic influence from tails at the end of the far-reverberated test-word's vowel[2] it transpired that identification errors tended to be reduced in not-gated conditions with near and silent contexts. This suggests that the test-word's tail played a role in the identification of the *preceding* consonant when intrinsic and extrinsic information were placed in conflict. That is, when listeners were presented with an ambiguously reverberant stimulus, they appeared to use intrinsic information to help resolve the uncertainty.

Moreover, since this experiment indicates that some compensation may arise in silent-context conditions, it may in addition cast earlier datasets in a new light, in particular for experiments where silence has been used as a 'control' condition

[1] Watkins and Raimond (2013) have suggested that listener data of Nielsen and Dau (2010) was affected by having presentation of only far-distance test-words presented, since listeners could predict the level of test-word reverberation from one trial to the next.

[2] Indeed, tails from the test-word's initial consonant might form a second intrinsic influence.

against which a reverberated speech carrier has been contrasted (cf. for example Brandewie and Zahorik, 2013; Nielsen and Dau, 2010; Ueno et al., 2005).

Experiment H3 has therefore permitted observation of compensation mechanisms acting in both a time-*forward* fashion (i.e. due to the extrinsic context appearing prior to the test-word) and in a time-*reverse* fashion (i.e. due to factors arising in the final reverberant tail of the test-word occurring after the consonant which distinguished the test-word). It is apparent, however, from Figure 5.10 that the effects due to the test-word gating process are small in comparison to those mediated by the preceding context. In conclusion therefore, while intrinsic sources of information should not be discounted, Experiment H3 suggests that compensation for reverberation is rather strongly informed by extrinsic sources of information. The following experiment therefore investigates the time course of this extrinsic compensation effect.

5.6 Experiment H4: Investigating the time course of the extrinsic perceptual compensation effect

In Experiments H1, H2 and H3, the context words preceding the test-word varied in duration from trial to trial; in other words, the amount of extrinsic information available to listeners was not constant. Regardless of this, each of these experiments found that inconsistent reverberation in the context and test regions of the signal brought about a degradation in consonant identification for far-reverberated test-words, and that this degradation could be alleviated by increasing the level of context reverberation to match that in the test-word.

By varying the duration of the context speech that is reverberated at the far-distance, Experiment H4 now investigates the time course of the extrinsic perceptual compensation effect. In the previous experiments, the speech context preceding the test-word was wholly reverberated at either far- or near-distance. Here, the context speech is divided into two regions; the first part is reverberated at the near distance, and the second part (just prior to the test-word) is reverberated at the far distance. The boundary between these two regions is then varied as an experimental condition in order ask how much of the context must be far-reverberated to bring about an effect of compensation on a far-reverberated test-word.

5.6.1 Stimuli

As in the earlier experiment, Experiment H4 used speech material from the AIC, low-pass filtered at 4 kHz to reduce ceiling effects in consonant identification. In

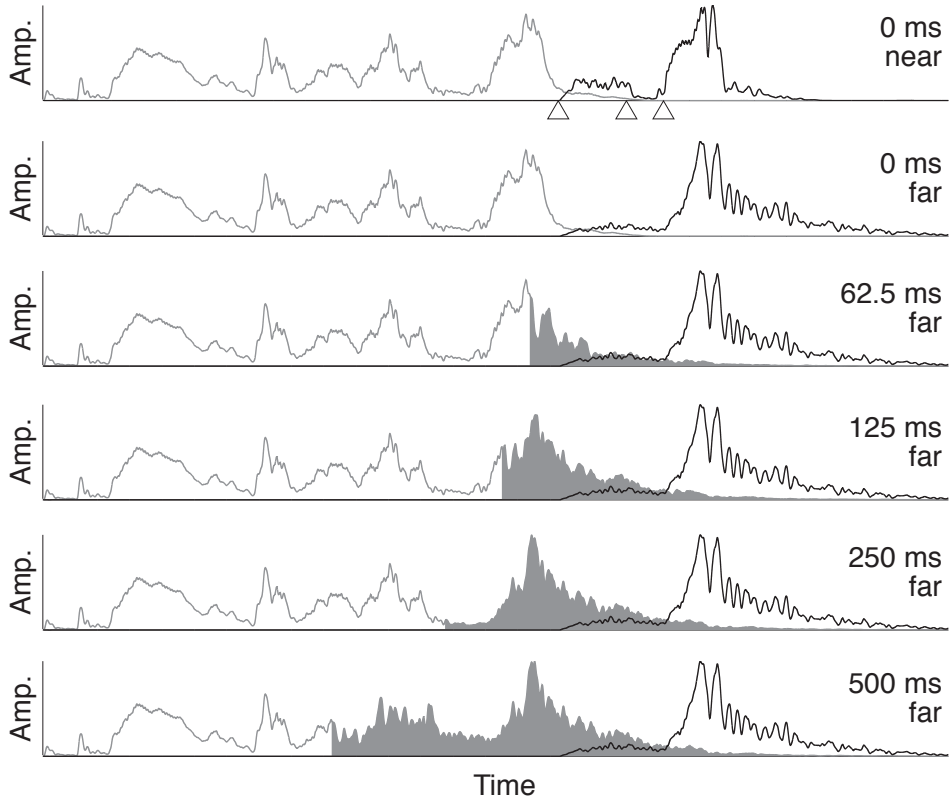


Figure 5.12: Illustration of selected stimulus conditions for one of the 100 utterances used in Experiment H4. The test-word *(dark line, lower condition label)* is preceded by the context *(light line)* which is divided into an initial near-reverberated part and a subsequent far-reverberated part. The temporal position of the boundary is varied between these parts, so that less or more of the context immediately prior to the test-word is far-reverberated *(shaded area, upper condition label)*. Other details remain as described for Fig. 5.1.

this experiment the stimuli and task were simplified, however, to reduce effects of a potential nuisance variable (consonant closure duration, which differs on average for [k], [p] and [t]). Using a two-alternative forced choice (2-AFC) design, listeners therefore identified just 's' or 'st' at the initial position of the test-word. Five following vowels were used to complete the test-words: {[eɪ], [iː], [ɛ], [ɪ], [æ]}, and the word-initial [s] sound was preserved from the earlier experiment design.

Using methods described in section 5.5.1, word boundaries were located to identify portions of the signal belonging to the test-word and context regions. However, in Experiment H4 the utterances were reordered and spliced so that all of the context words preceded the test-word. Utterances were now of the form

[CW3][CW1][CW2][TEST]

and therefore maximised the duration of the context before the test-word while retaining plausible phrases e.g., "daily we think stir". By limiting the number of corpus talkers to exactly half of those available, the resulting 100 utterances (10 talkers $\times$ 2 consonants $\times$ 5 vowels) had preceding contexts of around one second duration or longer. Four utterances fell slightly short of this target, however, and resulted in preceding context of 994, 979, 959 and 933 ms duration respectively.

The initial portion of the (rearranged) context phrase was always reverberated with the near-distance room impulse response in Experiment H4. Thereafter, a portion of the context just prior to the test-word was reverberated at the far-distance. The duration of this far-distance portion was controlled with a window of nominal duration 0, 62.5, 125, 250 or 500 ms, as depicted in the shaded regions of Figure 5.12. In practice, the window length was modified on an utterance-by-utterance basis so that the window edges were always aligned with zero-crossings in the audio signal. This ensured that reverberation of the context did not introduce any audible discontinuities in the signal. The duration of the far-context portion thus differed slightly from the nominal window length in almost all cases, but this variation was typically small. Across the whole set of stimuli, the mean deviation was 48.9 samples from the nominal window length, corresponding to approximately 1 ms at the 48 kHz sample rate in use.

As in earlier experiments, the near- and far-distance portions of the context were recombined with the test-word using the RMS balancing techniques outlined in 5.3.1, to create the same- and mixed- distance phrases. Finally, the overall RMS level was equalised across stimulus conditions, the headphone correction was applied throughout. The total set of sound files for Experiment 3 thus comprised 1000 stimuli (100 utterances $\times$ 2 test distances $\times$ 5 context window durations).

5.6.2 Participants

Forty participants were recruited for Experiment H4 via a university-wide email, and were compensated for their time. A further 5 people took part but were discounted from analysis. Two of the excluded participants reported hearing losses, and had considerable difficulties recognising test-words in all conditions. The remainder did not achieve the criterion for inclusion (above 90% correct responses for near-distance test-words at the 0 ms far-distance window condition).

5.6.3 Procedures

Stimuli were once again partitioned across the listener group (with 10 listeners per group) in order to avoid any association between test-word and context phrase that might otherwise assist recognition of the test-word. Talkers were divided evenly across participants so that each listener heard 10 utterances at each of the 10 experimental conditions, and each stimulus was repeated 4 times. Every test vowel was used in each experimental condition, once with the preceding 's' and once with the 'st' word initial consonants. Again, participants identified the test-word by clicking on either an 's' or 'st' alternative on the computer's screen, but were not required to identify the test-word's vowel. Stimuli were presented in a randomised order in a single session lasting around 20 minutes, and participants were encouraged to take breaks as often as they wished. Other aspects of stimuli presentation were carried out as described in section 5.3.3.

5.6.4 Results

Participants' responses for Experiment H4 are shown in Figure 5.13, presented in terms of the proportion of 's' responses at each test-word distance and context-window condition. These results were analysed with a two-way repeated measures ANOVA, using one factor for test-word distance (with two levels: near, far) and a second factor for the duration of the far-distance context preceding the test-word (with five levels: 0, 62.5, 125, 250, 500 ms). Mauchley's test again showed no violation of sphericity. The two-way interaction between factors for test distance and far-reverberated context duration was found to be significant ($F_{(4,156)} = 22.13$, $p < 0.001$) as were both main effects; test-word distance ($F_{(1,39)} = 75.96$, and $p < 0.001$) and far-context duration ($F_{(4,156)} = 25.55$, and $p < 0.001$). A linear trend test (using a least-squares method) across the log-spaced window-duration conditions showed a linear decrease in the number of 's' responses with increasing duration of the far-reverberated context window length ($F_{(1,39)} = 82.90$, $p < 0.001$).

5.6.5 Interim discussion

Experiment H4 asked how much of the context must be reverberated at the far distance in order to find a effect of perceptual compensation on a far-reverberated test-word. The current experiment was monaural, and thus complements a recent study on binaural compensation mechanisms by Brandewie and Zahorik (2013), Figure 5.13 suggests a clear trend; for a task in which listeners were required to determine whether the test-word began with an initial 's' or 'st', the number of

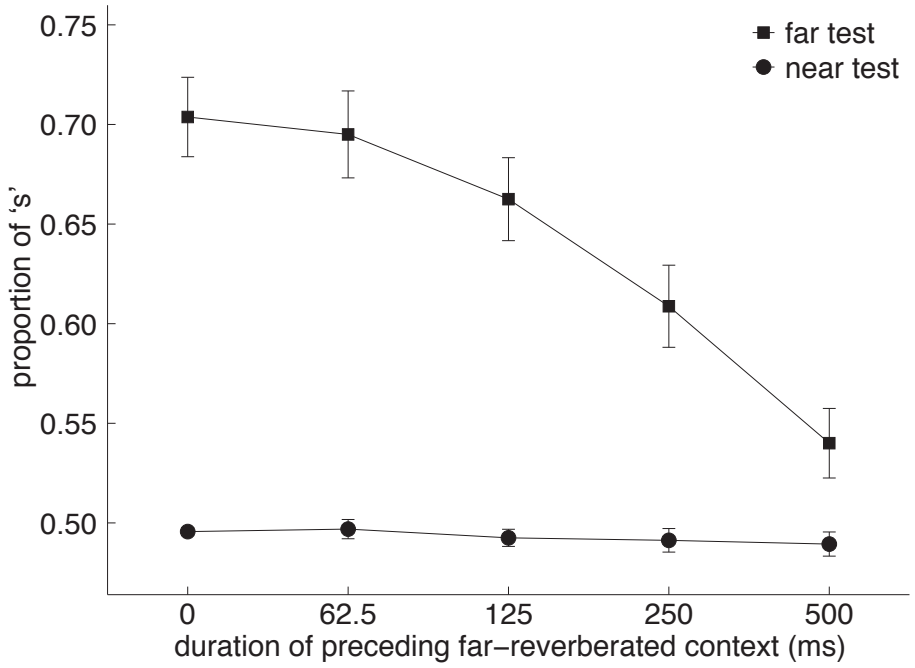


Figure 5.13: Mean and standard error of 40 participants' scores in Experiment H4. The lower line reports near-distance test-word scores, where listeners made few errors (i.e., the proportion of initial /s/ consonants reported is close to 0.5) and the data showed no dependency on the duration of the far-reverberated context. The upper line reports the far-distance test-word scores. For the zero-length far-context window condition, where the test-word has more reverberation than the context, listeners often misclassified 'st' as 's'. As the duration of the far-reverberated part of the context increased, fewer misclassifications were made and the proportion of /s/ consonants reported decreased.

incorrect responses consistently decreased as the duration of the far-reverberated context portion increased.

For consistency with the three previous experiments, Experiment H4 once again used utterances drawn from the AIC, however the final context word was spliced onto the beginning of the utterance in this experiment in order to maximise the duration of the preceding context. This process ensured that the preceding context duration was at least 1 second in most cases, which was sufficient to allow a near-reverberated portion (of 500 ms or longer) to be followed by a far-reverberated portion of *at most* 500 ms in duration. This allowed examination of the compensation effect across more rapid timescales than the shortest condition observed in Brandewie and Zahorik (2013), where an 850 ms context was sufficient to observe listeners' improvement in a binaural listening task. However, the maximum time resolution of Experiment H4 was insufficient to determine whether a *further* decrease in consonant identification error would have arisen from a longer far-

reverberated context portion (i.e. if the window duration exceeded 500 ms), thus is not clear from this data whether compensation is yet 'complete' in the longest time-condition studied.

One option to increase the length of the speech material presented before the test-word would have been to pad short utterances with additional speech material. However, this would have disrupted the consistent form of the utterances (which otherwise had exactly three context words preceding the test-word) and might have directed listeners' attention towards the context rather than the test-word itself (cf. Ueno et al., 2005, Experiment 1). To avoid this, context lengths were instead maximised by selecting the 10 talkers which had the fewest utterances of short durations in the corpus. By this method, it is likely that talkers with a slow speaking rate were preferentially selected. With fast and slow versions of the 'sir-stir' paradigm, Watkins has previously shown that speech at a faster rate was more influenced by the test-word reverberation, and resulted in a greater shift in the category boundary between the near-near and near-far conditions (see Watkins, 2005a, fig. 3). Hence by using slower talkers, Experiment H4 used a set of utterances that are likely to be relatively robust to the effects of reverberation, and a still larger effect of perceptual compensation than that observed in Figure 5.13 might have been apparent were the faster talkers selected instead. Be that as it may, inclusion of the slower talkers is more akin to every-day life, since talkers tend to slow down in reverberant rooms (Black, 1950) and listeners tend to prefer slower speech (Moore et al., 2007).

5.7 General discussion and conclusions

The human listener data presented in Chapter 5 has demonstrated monaural effects of compensation for reverberation with naturally spoken speech material. Table 5.3 presents a summary of the experiments in which data was gathered, achieving in total nearly 55,000 responses for around 10,000 stimuli. As has been shown, the collected data replicates the main findings of the 'sir-stir' continuum experiments undertaken by Watkins (2005a), in which listeners were asked to identify test-words where the impression of a [t] was created artificially (by superimposing the temporal envelope of a spoken 'stir' on top of a recorded 'sir' utterance). Moreover, the consonant confusion data collected here extends Watkins' paradigm by examining a slightly wider range of test-words embedded in short phrases where the talker and context vocabulary varies from trial to trial. The following section presents a meta-analysis of the combined data, not in a statistical sense (ie., there is no hypothesis to test), but rather in an observational sense in order to highlight some note-worthy features of the dataset that appear to be common across the four experiments undertaken but which have not hitherto been discussed.

Table 5.3: Data coverage of listener responses across all experiments, where experimental conditions are abbreviated according to the following alphabetical list. Totals are not shown for the number of conditions and number of utterances drawn from the Articulation Index Corpus (AIC) since there is a degree of overlap between experiments.

C – number of conditions in which the context words were presented;
F – number of low-pass filter cutoff frequency conditions;
R – number of reverberation directions in the context portion;
S – number of speech directions in the context portion;
T – number of test-word distances;
V – number of voices heard (talkers);
W – number of test-words presented.

Experiment	Conditions	AIC phrases	Stimuli	Participants	Responses
H1	20 (2T×2C×5F)	80 (20V×4W)	1,600	60	4,800
H2	16 (2T×2C×2S×2R)	80 (20V×4W)	1,280	64	5,120
H3	12 (2T×3C×2G)	480 (20V×24W)	5,760	60	28,800
H4	10 (2T×5C)	100 (10V×10W)	1,000	40	16,000
TOTAL	–	–	9,640	224	54,720

5.7.1 Meta-analysis of the human listener data

Grand summary confusion matrices are presented in Tables 5.4 and 5.5 for the baseline situation most resembling 'normal' listening, for the experiments with four (Experiments H1–H3) and two (H4) response alternatives respectively. The same data is also visualised in Figure 5.14 using the familiar RIT error grids. In contrast to earlier plots which showed the mean and standard error of the participants' individual E_{RIT} values, however, the value of E_{RIT} is calculated here from the aggregated confusion matrix in each experimental condition (using equations 5.1 through 5.4 as before).

Figure 5.14 shows, for the summary data aggregated across all participants, a replication of the main findings seen in individual listeners in the earlier analysis. As expected, the lowest error rates are seen in the least reverberant conditions (lower-left in each panel), at near-near *context-test* distance. Shown in the leftmost columns of Tables 5.4 and 5.5, responses at near-near show very few confusions, in part because this is the baseline conditions in each experiment that most resembles 'normal' listening, and in which the 90% correct inclusion criterion was applied.

The far-near condition has not received much attention thus far, however the data in all four experiments clearly show the effect of overlap-masking in the lower-right data-point of each panel in Figure 5.14 (Nábělek et al., 1989). Here, the near-reverberated test-word is more frequently misidentified when it is preceded by a far-reverberated context, due to the prolongation of reflected energy from the

Table 5.4: Confusion matrices summarising 184 participants' responses in Experiments H1, H2 and H3. Rows correspond to the stimuli presented; columns record the responses. Reverberation conditions are labelled as *context-test* distance. Results presented in the near-near condition are for the baseline situation most resembling 'normal' listening (i.e., that in which the 90% correct inclusion criterion was applied). Few confusions were recorded here. In the near-far condition, listeners frequently misreported the stop consonants [sk], [sp] and [st] as 's', however these confusions were largely resolved in the far-far condition. Confusions for [s] stimuli rose in near-far conditions, but rose further still in the far-far condition (largely towards 'st' responses) so did not display the same compensatory effect. The final row shows the proportion ($\propto$) of responses per category of stimulus presentation. Values under 0.01 are not shown, and values may not total 1.00 due to numerical rounding.

	near-near					near-far					far-far				
		's'	'sk'	'sp'	'st'		's'	'sk'	'sp'	'st'		's'	'sk'	'sp'	'st'
H1	[s]	60	0	0	0	[s]	56	1	0	3	[s]	52	1	0	7
	[sk]	0	60	0	0	[sk]	9	46	3	2	[sk]	2	52	0	6
	[sp]	0	0	60	0	[sp]	27	3	27	3	[sp]	4	3	47	6
	[st]	0	0	0	60	[st]	23	2	1	34	[st]	2	0	0	58
H2	[s]	80	0	0	0	[s]	71	1	2	6	[s]	65	1	1	13
	[sk]	0	80	0	0	[sk]	15	63	2	0	[sk]	2	71	1	6
	[sp]	0	0	80	0	[sp]	17	8	52	3	[sp]	2	8	64	6
	[st]	0	0	0	80	[st]	23	1	0	56	[st]	4	5	0	71
H3	[s]	584	6	3	7	[s]	558	4	16	22	[s]	477	17	31	75
	[sk]	2	596	1	1	[sk]	70	502	6	22	[sk]	22	550	12	16
	[sp]	0	0	580	20	[sp]	228	40	243	89	[sp]	57	35	391	117
	[st]	6	9	2	583	[st]	300	24	16	260	[st]	83	38	35	444
H1+H2+H3	[s]	724	6	3	7	[s]	685	6	18	31	[s]	594	19	32	95
	[sk]	2	736	1	1	[sk]	94	611	11	24	[sk]	26	673	13	28
	[sp]	0	0	720	20	[sp]	272	51	322	95	[sp]	63	46	502	129
	[st]	6	9	2	723	[st]	346	27	17	350	[st]	89	43	35	573
$\propto$	[s]	.98	-	-	-	[s]	.93	-	.02	.04	[s]	.80	.03	.04	.13
	[sk]	-	.99	-	-	[sk]	.13	.83	.01	.03	[sk]	.04	.91	.02	.04
	[sp]	-	-	.97	.03	[sp]	.37	.07	.44	.13	[sp]	.09	.06	.68	.17
	[st]	-	.01	-	.98	[st]	.47	.04	.02	.47	[st]	.12	.06	.05	.77

Table 5.5: Confusion matrices summarising 40 participants' responses in Experiment H4. Other details are the same as described for Table 5.4 above.

	near-near			near-far			far-far		
		's'	'st'		's'	'st'		's'	'st'
H4	[s]	791	9	[s]	775	25	[s]	712	88
	[st]	2	798	[st]	351	449	[st]	152	648
$\propto$	[s]	.99	.01	[s]	.97	.03	[s]	.89	.11
	[st]	-	1.	[st]	.44	.56	[st]	.19	.81

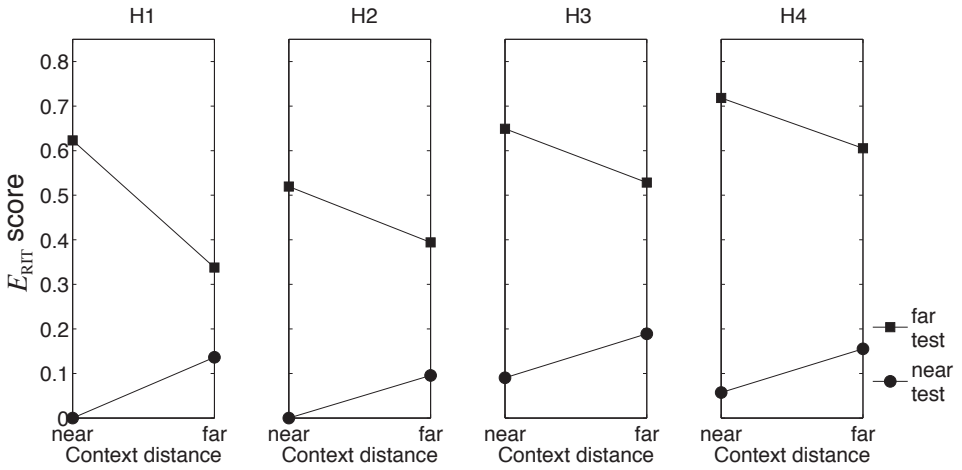


Figure 5.14: Compensation for reverberation is apparent in the aggregated summary confusion matrix data for baseline conditions in each experiment undertaken. Lowest error rates are found in the least reverberant condition (lower-left data-points in each panel), at near-near *context-test*. At far-near, the reverberant context causes overlap-masking in the test-word and error rates increase a little (lower-right data-points). More striking though is the effect of reverberation on the test-word (upper-left data-points): error rates are substantial in the near-far condition. In every experiment, the main finding underpinning the compensation for reverberation paradigm is replicated: error rates for a far-reverberant test-word reduce when a greater degree of reverberation is present in the context as well (upper right data-points).

context into the portion of the signal in which the test-word is located (Watkins, 2005a; Watkins et al., 2011).

More striking, however, is the effect of reverberation on the test-word itself: substantial error rates are seen in the near-far condition (the upper-left data-point in each panel in Figure 5.14). In this case there is little contribution to overlap masking from the context since it is at the low-level reverberation condition (the near source-receiver distance). However, overlap-masking protrudes from the reverberated [s] portion of the test-word into the region defining stop consonants (where present); moreover there is self-masking within the phoneme pertaining to the [k], [p] or [t] itself (Nábělek et al., 1989). As discussed earlier, listeners frequently misreported the [sk], [sp] and [st] word-initial clusters as 's' in this condition.

The final rows in Tables 5.4 and 5.5 show the full extent of these misidentifications at near-far distance across the four experiments by quantifying the proportion ($\propto$) of responses per category of stimulus presentation. For [st] presentations in near-far conditions, close to half of the total responses are consonant confusions in favour of 's' (47% in Experiments H1–H3, 44% in H4). In Experiments H1–H3, this accounted for $\simeq$ 89% of the misidentifications made in that condition. The

[sp] stimuli see over a third of responses toward 's' (37%), accounting for $\simeq$ 65% of the misidentifications in this case. The [sk] are substantially more robust in the near-far condition, with almost twice the correct identifications the [sp] stimuli just discussed. Nonetheless, 's' responses again constitute the majority of the misidentifications for [sk] stimuli, standing at around 73% of the errors.

Finally, the effect of compensation for reverberation can be seen at the far-far distance in every experiment performed (upper-right data-point in each panel of Figure 5.14): here, an increased level of reverberation in the preceding context appears to assist listeners in identifying reverberant test-words. Although the acoustic aspects of reverberation have not fundamentally changed, the perceptual implication of the reverberant context is quite different whether the test-word is at the near or far distance. That is, the far-distance context reverberation still causes the prolongation of reflected energy into the test-word, an effect usually termed overlap-masking which was previously seen to result in poorer listener performance. Now, however, the degrading effect of this energetic masking is outweighed overall by the beneficial compensation effect. This is seen most clearly at the foot of the right-hand column in Table 5.4, where the frequency of occurrence of [st], [sp] and [sk] stimuli being reported as 's' reduces from 47, 37 and 13% respectively to only 12, 9 and 4% of the total number of responses overall.

5.7.2 A different pattern in [s] stimuli responses

It is interesting to note, however, that this pattern is not observed in the data for all consonants: confusions for the [s] stimuli do not display the same compensatory effect. For [s] presentations, the number of correct identifications falls in the near-far condition, but now continues to fall still further in the far-far condition (in all experiments). A possible explanation for this finding is that during the far-distance context, listeners become primed to a high level of reverberation being present. Thus they are likely to be sensitive even to small dips in the temporal envelope, and may therefore over-estimate any degree of signal modulation as a stop closure[1]. There is a strong favouring of the 'st' responses in the data, amounting to approximately 88, 87 and 61% of the errors in Experiments H1, H2 and H3 respectively (the corresponding value of 100% was inevitable given the 2AFC design of the listener task in Experiment H4). A similar result was reported

[1] Indeed, a similar effect might be expected for the modelling simulation described in Chapter 4, where far-distance context reverberation gives rise to an increased level of efferent attenuation which in turn reduces the response in the auditory nerve. Typically this increases the dynamic range in the signal since the response to low-level portions of the stimuli is reduced below threshold; thus gaps are revealed in the signal which may (mistakenly, in this case) give rise to the impression of a stop consonant closure.

by Gelfand and Silman (1979), who studied confusions between 20 reverberant consonants and noted that errors regarding place of articulation were often toward alveolar consonants.

For [s] stimuli errors in the far-far condition, the clear preference for the 't' response (over 'k' or 'p') can perhaps be attributed to the high-frequency spectral prominence that defines both the alveolar consonants, [s] and [t] (cf. Figure 4 in Allen and Li, 2009). The velar consonant [k] and bilabial consonant [p], on the other hand, contain more prominent in mid- and lower- frequency energies respectively, and are thus less likely to be confused with the [s] stimuli. An alternative explanation for the same finding might suggest a lexical preference for words with 'st' arising from a phonotactic point of view, here analysed according to the pronunciations of words within the dataset used in Lanchantin et al. (2013). Excluding the cases in which [s] appears in the word-final position, nearly half (44.0%) of the remaining 151,744 occurrences of [s] are immediately followed by one of the three stop consonants studied. Among these phoneme pairs, [st] accounts for 77.4% of the data observed, with [sp] and [sk] appearing on 12.1% and 10.4% of occasions, respectively[1].

Since the analyses performed in this study assess the entire confusion matrix by reducing it to a single numerical value, these methods cannot distinguish between the two different effects observed in the data, i.e. the decrease in identification resulting for [s] at far-far, alongside the increase in identification rates for the combined [k], [p], [t] group. The meta-analysis findings suggest therefore that the RIT calculation in Experiments H1, H2 and H3 are likely to have *under*-estimated the compensation effect that would be seen in an analysis that considered only the participants' responses to stop consonant stimuli presentations. On the other hand, the analysis of Experiment H4 took the proportion of 's' responses itself as the dependent measure (analogous to the method used in Watkins' original studies). Here, the rapidity of the build-up of the compensation effect may have been *over*-estimated as a result of the continued deterioration in recognition of [s] stimuli in the far-far conditions (which would have contributed to the overall lowering of the number of 's' responses counted in this condition).

5.7.3 Concluding remarks

The series of four experiments reported in this chapter complements the work of other researchers in the field by providing evidence that monaural exposure to a re-

[1] I am grateful to Oscar Saz for computing the phoneme-pair statistics in the Lanchantin et al. (2013) dataset.

verberant environment is sufficient to bring about a significant improvement in consonant identification in speech material recorded by 20 talkers. Measured across 224 left ears, these effects clearly do not involve the types of inter-aural processing that are thought to underpin binaural hearing process. As such, they cannot be directly attributable to the raft of binaurally advantageous listening processes that are observed in reverberant listening tasks, such as 'echo-suppression' resulting from precedence effect buildup (cf. Brandewie and Zahorik, 2010; Zahorik et al., 2009). Instead, the listener data gathered in this study is consistent with a compensation mechanism that appears to enhance the amplitude modulation in reverberant signals (Zahorik et al., 2012), and which has been attributed to temporal envelope constancy (Kuwada et al., 2012; Watkins et al., 2011).

While Experiments H1 to H3 all hinted that these monaural compensation mechanisms are fast-acting (i.e. occur within an utterance), the time course of the large extrinsic effect was only directly investigated for the first time in Experiment H4. Here, the majority of consonant confusions were correctly resolved after only half a second of consistent far-distance reverberation on the preceding context (it was not determined whether *longer* contexts would increase performance further still). Relatedly, a study by Brandewie and Zahorik (2013) has recently reported that contexts of 850 ms bring about a performance improvement in a binaural listening task (conversely, in that study *shorter* timescales were not tested). At the other end of the scale, a long-term binaural improvement was reported by Shinn-Cunningham (2000) to take place over several hours exposure to a single room reflection pattern. It would appear from the pattern of results in current experiments, however, that the monaural compensation effects studied here are not primarily related to the refined learning of a particular room characteristic. Indeed, compensation effects have previously been demonstrated despite various distortions to the fine-structure of the room's reflection pattern, for example by presenting context and test-word with impulse responses recorded in different rooms (Watkins, 2005a), or by reversing the polarity of a randomly selected half of the samples in the impulse response (Watkins et al., 2011). Thus it would appear that the rapidly-acting monaural constancy mechanism acts to promote an initial, fast calibration to a new listening environment.

Experiments H2 and H3 replicated and extended experiments by Watkins (2005a) and Watkins and Raimond (2013) respectively. Experiment H2 confirmed two important findings; firstly, that the constancy mechanism does not require linguistic comprehension of the speech signal, and secondly, that time-reversed rooms disrupt compensation for reverberation. By slightly altering the implementation of the reverberation processing from conditions used in Watkins (2005a, Experiment 5), Experiment H2 additionally suggested that that the monaural compensation effect was dominated by information arising (or failing to arise) from treatment of the

preceding context rather than from the backward protrusion of reverberation from the following context into the test-word region. The contribution to compensation of the region following the test-word was therefore probed in more detail in Experiment H3. Watkins and Raimond had reported an intrinsic compensation effect which was mediated through information originating from the test-word's final tail, but had carried out their study only for test-words presented in isolation. Extending this, Experiment H3 assessed the contribution of intrinsic and extrinsic information when test-words were presented after a preceding speech context, and found that intrinsic cues contributed something to compensation effects even where there was a longer context available to the listener, and even for highly variable speech stimuli where the talker and speech content differs from trial to trial.

The basic aim of the present experiments was to investigate whether the monaural compensation effect found by Watkins (2005a) generalises to conditions where the variability among sounds is more similar to everyday listening. It clearly does. Watkins' listeners' task relied on stimuli that created the perception of a [t] by a process of amplitude-modulation. The current series of experiments instead used utterances in which the stop consonant is signalled by naturally occurring acoustic-phonetic cues. These experiments have demonstrated a large compensation effect overall, but have shown that this monaural constancy effect may not generalise to the full range of natural speech sounds (indeed, the meta-analysis reveals that the [s] may behave somewhat differently from the plosive consonants studied)[1].

Nonetheless, these results appear to be ecologically relevant for two main reasons that were previously discussed in § 2.3.4. The first of these relates to the high rate of occurrence of stop consonants in real speech. Indeed, Mines et al. (1978) reported that [t], [k] and [p] account together for 10.67% (respectively: 5.78, 3.10, 1.79%) of all phonemes encountered in American English casual conversational (including the vowels). Secondly, since the consonants studied here are among those most vulnerable to the effects of reverberation (Drullman et al., 1994b; Gelfand and Silman, 1979; Nábělek et al., 1989), the experiments above are directly concerned with the very parts of the speech signal that are the most awkward to hear in real reverberant listening situations.

The speech material selected for Experiment H4 was controlled for temporal effects of the preceding context, however the frequency content was not controlled in the same way. Allen and Li (2009) reported that frequency regions around 4 kHz are likely to contribute most significantly to identification of the [t] when it is present in test-words. Moreover, Watkins et al. (2011) reported that compensa-

[1] This raises a query in regard to the exact nature of effects reported recently for the binaural constancy effect (e.g., Brandewie and Zahorik, 2013; Srinivasan and Zahorik, 2013), and suggests it may be interesting to more closely study the speech materials used in such works.

tion for reverberation appears to work in a band-by-band manner, and that the level of compensation achieved would be similarly dependent upon these important frequency regions in the neighbouring context words. Taken together, these findings imply – provided the trial is identified based on the test-word's stop consonant – that a context phrase that is rich in sibilants and stops of a matched frequency region to the test-word (e.g., "first people detect") would promote a higher degree of compensation than would a phrase without (e.g., "now you remember"). Future work would be required to examine the implications of this phonetic variation on the time course of the constancy mechanism.

Chapter summary

This chapter has demonstrated perceptual compensation for the effects of reverberation in an ecologically relevant listening task. Listeners' ability to identify the consonant in a reverberated test-word was strongly influenced by the level of reverberation in the preceding speech context. When the reverberation in the context and test-word were consistent, an extrinsic compensation effect was apparent which helped listeners to monaurally identify the test-words' consonant. This effect, though not reliant on phonetic processing of the speech signal, broke down when time-reversed reverberation was applied to the context phrase. A smaller intrinsic compensation effect similarly appeared to aid identification of reverberant test items for stimuli which did not promote extrinsic compensation, namely those with silent or near-distance speech carriers. Subsequently, the time course of the extrinsic compensation mechanism was investigated by applying the same reverberation to the test-word and to a proportion of the preceding context. Consonant identification improved as the reverberated portion of the context increased in duration, up to the maximum tested duration of 500 ms. Finally, a meta-review of the listener data across the four experiments revealed that these compensatory effects were largely restricted to the stimuli containing the unvoiced stop consonants.

Chapter 6

Conclusions

Contents

Chapter overview

Perceptual compensation for reverberation was explored in this thesis by means of a computational modelling study and psychophysical experiments. This chapter summarises the main findings of these two strands of work, and discusses their mutual implications and interdependencies. The relationship between current findings and those of other researchers working on similar topics is discussed. Alternative perspectives are also considered, leading to an examination of the wider relevance of the research. Potential criticisms of the studies are addressed, and some suggestions for future work are presented.

6.1 Original contributions

To introduce the research topic and clarify the focus of experimental work in this thesis, existing studies which queried the effects of reverberation on human- and machine-speech identification were reviewed in Chapter 2, where perceptual compensation for reverberation was examined in detail. Auditory mechanisms that may potentially contribute to human listeners' robustness in reverberation were subsequently examined in Chapter 3. Chapters 4 and 5 of this thesis then presented two strands of original work investigating perceptual compensation for reverberation.

1. A candidate model of perceptual compensation for reverberation was proposed, based on the hypothesis that efferent-inspired auditory modelling may help improve speech identification in reverberant environments. The model provided a qualitative match to human response data from experiments presented in Watkins (2005a), in which perceptual compensation occurred for time-forward and time-reversed speech signals, but was disrupted when time-reversal was applied to the impulse response instead.

2. A series of psychoacoustic experiments showed that the monaural compensation effect that was originally demonstrated in Watkins (2005a) could be generalised from synthetic test-words to naturally spoken test-words. This compensation effect was found to derive from the context *and* test-word portions of the signal, to be fast-acting, and to break down when the temporal direction of reverberation was reversed.

Original findings from the modelling studies and listening experiments are summarised below.

6.1.1 Efferent-inspired model of perceptual compensation for reverberation

The model developed in this thesis is underlain by the proposal that medial olivocochlear (MOC) unmasking may assist with perceptual compensation for the effects of late reverberation. This proposal relies on the similarity of the effects of late reverberation and of additive noise, both of which can be understood to raise the noise floor of a signal, and thereby reduce its dynamic range. While a large body of evidence now points to the involvement of MOC efferents in noise-robust listening, it remains to be investigated physiologically whether (or how) the efferent system can assist in the processing of reverberant speech. Thus, while speculative, it seems plausible that MOC efferents may be partially responsible for human robustness to reverberation since they are implicated in recalibrating the sound-encoding behaviour of the cochlea in complex listening environments.

The computational model presented in this thesis embedded the efferent model of cochlear processing by Ferry and Meddis (2007) in a closed-loop structure, and automatically adjusted the level of efferent activity in response to the level of reverberation detected in the signal. The model was then challenged to replicate human response data from Watkins (2005a), in which sentences including 'sir' and 'stir' test-words were identified monaurally in a variety of experimental conditions.

- Experiment M1 used human response data from two room-positions ('near' and 'far') in Watkins (2005a) to tune model parameters for two candidate metrics for efferent feedback regulation. The mean-to-peak ratio (MPR) metric effectively measured the dynamic range of the signal, and applied more efferent attenuation when a low dynamic range suggested that a high level of reverberation was present. Alternatively, the low-pass mask (LPM) measured the presence of reverberant tails at signal offsets. In this implementation, more efferent attenuation was applied when a mask defined by negative-going gradients (i.e., tails) in smoothed temporal envelopes of individual channels suggested that a high level of reverberation was present.

- Experiment M2 investigated stimulus conditions in which there was an abrupt change in reverberation condition mid-utterance, so that part of the sentence was heard from nearby and part from further afield. Whether driven with the MPR or LPM metric, the efferent-inspired circuit enabled the model to compensate for the effects of reverberation, replicating the effect observed for human listeners. When a low level of reverberation was detected in the preceding context, little attenuation was applied and the temporal dip signalling the 't' consonant was filled with reverberant energy, resulting in a high number of 'sir' responses for reverberant test-words. On the other hand, when a high level of reverberation was detected in the preceding context,

more attenuation was applied, and the dip signalling the 't' consonant was revealed. This resulted in a greater number of 'stir' responses across the continuum of test-words, and the restoration of the category boundary towards its original position.

- Experiment M2 also showed that compensation for reverberation persisted in the model (as it did for human listeners) when the linguistic content of the speech context was destroyed by time-reversing the speech utterance. This result was modelled by both the MPR and LPM measures.

- Experiment M3 asked whether compensation would occur in the model if the time-direction of the reverberation itself were reversed. Here, the LPM-driven model replicated the human response data, suggesting that the absence of reverberation decay tails at signal offsets is sufficient to block the compensation process in human listeners. On the other hand, the MPR was not able to simulate human listener data, suggesting that a reverberation measure based solely on dynamic range is not sufficient to explain the effects of compensation for reverberation.

6.1.2 Evidence for monaural compensation with naturalistic stimuli

Development of the auditory model exposed a number of gaps in our knowledge of how perceptual compensation arises in human listeners. A series of four psychoacoustic experiments addressed the most significant of these questions and, furthermore, provided sufficient evidence of the monaural constancy effect to show its relevance for everyday listening.

- Experiment H1 demonstrated a monaural compensation for reverberation effect (generalising the phoneme-continuum experiments of Watkins, 2005a) which holds relevance in ecologically valid listening environments where the talker, context speech and test-words varied from trial to trial. Perceptual compensation for reverberation was measured in terms of a reduction in the number of consonant confusions that listeners made. By using a range of low-pass filter conditions to assess the operational point of the experimental setup, Experiment H1 also obtained evidence suggesting that the compensation mechanisms works in a band-by-band fashion.

- Experiment H2 used naturally spoken materials to demonstrate that although the monaural compensation effect does not rely on the time-forward direction of a speech signal (and is thus not dependent on linguistic factors), it *is* dependent on the time-forward direction of the reverberation itself.

- Experiment H3 queried the temporal extent of the signal area which contributed to compensation, and showed that intrinsic information from the reverberation tail of the test-word's final vowel could influence the perceived identity of the consonant which preceded it. This indicates that perceptual compensation is influenced by both intrinsic and extrinsic factors.
- Experiment H4 queried the time course of the extrinsic compensation effect arising from the preceding context. Consonant identification improved as reverberation was applied to an increasing portion of the preceding context (up to 500 ms). This suggests a rapid compensation mechanism which allows listeners to gain robustness to reverberation within a second or so of environmental changes.
- Finally, a meta-analysis considering Experiments H1–H4 side-by-side showed that these compensatory effects were largely restricted to stimuli containing the plosive consonants [k], [p] or [t].

6.1.3 Implications of perceptual findings on auditory model development

Each perceptual experiment was designed to shed light on a different aspect of the auditory model, therefore the human response data has some very direct implications for future computational modelling work.

- Experiment H1 confirmed that the monaural constancy mechanism is relevant for recognition of naturally spoken materials, and also provided evidence of a band-by-band compensation effect. This result suggests efferent feedback can next be implemented in a channel-independent manner (as has recently been done, for example, by Clark et al., 2012).
- By altering the implementation of time-reversed reverberation on the following context from Watkins' original study (cf. Figure 5.8), Experiment H2 showed the compensation effect to be strongly influenced by the preceding extrinsic context. This suggests efferent attenuation should be driven by a metric that is sensitive to the time-direction of reverberation prior to the test-word.
- Experiment H3, however, suggests the model must be updated to include sources of intrinsic information deriving from within the test-word. Thus the windowed portion of signal which is assessed to determine the efferent control should extend both backwards into the preceding context and forwards throughout the duration of the test-word itself (including its reverberation tail).

- The temporal extent of the extrinsic compensation effect was probed more closely in Experiment H4. Results suggest that there is benefit in extending the time window backwards at least 500 ms into the preceding context, but may not require the full 1000 ms used at present (which was chosen to allow comparison with results shown in Beeston and Brown, 2010).

6.2 Relation to similar research

Perceptual compensation for reverberation has been discussed in this thesis in regard to both human listeners and machines. Psychoacoustic research into this phenomenon has increased significantly in the past decade, in parallel with greater interest in efferent computational modelling. However, as yet there is no significant linkage between the two domains.

6.2.1 Psychoacoustic investigation of compensation for reverberation

Experiments H1–H4 confirmed that the monaural compensation mechanism demonstrated in the 'sir-stir' continuum experiments of Watkins (2005a) is relevant with naturalistic speech material. This generalisation is important because doubt had been raised about the existence of a monaural mechanism on two fronts. Firstly, the majority of listeners tested in Brandewie and Zahorik (2010) did not improve from prior monaural room exposure, though this may have arisen due to the inclusion of a simultaneous background noise which required listeners to employ other hearing mechanisms. Secondly, Nielsen and Dau (2010) revisited the 'sir-stir' paradigm but were unable to demonstrate the compensation for reverberation effect with reverse-engineered room impulse responses (cf. § 2.4.2).

The current series of experiments also reveals that further claims made by Watkins and colleagues in regard to 'sir-stir' continuum stimuli are also relevant to naturally spoken materials heard monaurally. Using nonsense syllables as test-words in order to avoid semantic effects from the preceding context (Srinivasan and Zahorik, 2011), the compensation mechanisms was first shown not to rely on a linguistic component (as in Watkins, 2005a). Providing further evidence in support of Watkins (2005a, who used synthetic test-words) and Longworth-Reed et al. (2009, who used binaural presentation), the monaural compensation effect was also shown to be disrupted by time-reversed reverberation on real-speech contexts.

Compensation effects arising within the test-word itself, demonstrated for 'sir-stir' test words in Watkins and Raimond (2013), were also found to be relevant for real speech. Since this intrinsic compensation arises in silent-context conditions, this suggests that silence may not be a straightforward 'control' condition against

which to contrast a reverberated speech carrier (cf. Brandewie and Zahorik, 2013; Nielsen and Dau, 2010; Ueno et al., 2005).

Finally, the monaural time course of compensation for reverberation was examined directly, and found to act on a similarly rapid timescale to that of the binaural effect measured by Brandewie and Zahorik (2013). Together, the data suggest a compensation mechanism that appears to enhance the amplitude modulation in reverberant signals (Zahorik et al., 2012), and which has been ascribed to temporal envelope constancy (Kuwada et al., 2012; Srinivasan and Zahorik, 2014; Watkins et al., 2011).

6.2.2 Development of efferent-based computational models

Computational models of the auditory system evolve in a circular manner, being gradually refined or radically altered as novel hypotheses are posed and tested, and as further physiological or psychoacoustic datasets become available (Weintraub, 1985). A recent example of this way of working is given by van Dorp Schuitman et al. (2013) in a study which models the perception of room acoustics (though without any simulation of efferent activity in this case).

Since the MOC unmasking principle which underpins the current modelling effort is usually associated with *noisy* rather than reverberant listening experiments (cf. § 3.4.2), it is perhaps not surprising that further examples of efferent-models processing reverberant speech are not available for comparison. Yet, as Figure 2.3 depicted, the effects of late reverberation are somewhat comparable to those of additive noise, and thus MOC activation may be relevant (as described in § 3.4.3).

The majority of efferent models have been applied to the task of modelling speech-in-noise data. Indeed, the studies of Brown et al. (2010) and Clark et al. (2012) both simulated human recognition of speech in noise using the efferent-DRNL model of cochlear filtering by Ferry and Meddis (2007) which lies at the core of the computational work in this thesis. Additionally, noisy speech has been investigated by Lee et al. (2011) using the efferent-inspired model of Messing et al. (2009), and by Chintanpalli et al. (2012) using the model of Zilany et al. (2009).

6.3 Alternative perspectives

Via a raft of related phenomena grouped under the umbrella term 'echo-suppression' (including the precedence effect, Haas effect, law of the first wavefront, and localisation dominance), it has long been realised that exposure to room effects can assist listeners in *localisation* tasks, via the selective weighting of

acoustic cues based on time of arrival (cf. § 2.3). Such effects have been investigated with click stimuli and with speech (Hermansky et al., 2013; Litovsky et al., 1999; Nábělek and Robinette, 1978; Wallach et al., 1949), and have been modelled both monaurally and binaurally (Bürck and van Hemmen, 2007; Hummersone et al., 2013; Martin, 1997). Interestingly, it appears that while *spatial* information in the reflections is largely suppressed to provide a fused image, *non*-spatial details remain perceivable, and at times are reported by listeners as alterations in timbre, loudness, and so on (Bishop et al., 2014). More recently, researchers have also begun to describe the manner in which exposure to room effects can also assist listeners in *identification* tasks.

6.3.1 An internal model of a room

It has been proposed several times in the literature that human listeners' robustness to reverberation might arise from their ability to model rooms in their heads (Keen and Freyman, 2009; Menzer and Seeber, 2014; Shinn-Cunningham, 2000). One possibility is that listeners might build up internal geometrical representations of the various reflective surfaces in their environment, and mentally 'deconvolve' the associated signal and room effects. This theory is attractive since it would, for instance, allow the disruption of compensation observed in time-reversed reverberation conditions to be explained on the basis that listeners adapt only to *plausible* architectural arrangements. A study by Menzer and Seeber (2014) appears to support this theory, using stimuli in which high-order reflections were artificially scrambled in certain experimental conditions, but the conclusions are drawn from very limited data (a single listener).

However, other studies appear to argue against this room acoustic hypothesis. In particular, two experiments have reported compensation for reverberation despite various distortions to the fine-structure of the room's reflection pattern which ought to render any such internal models effectively useless. First, compensation occurred even when the context and test-word were convolved with impulse responses that had been recorded in different rooms (Watkins, 2005a). Second, compensation persisted even when half of the samples (randomly selected) in the impulse response had their polarity reversed before convolution with speech signals (Watkins et al., 2011).

6.3.2 Modulation masking

Section 5.5.5 discussed 'modulation masking', a theory which was put forward by Nielsen and Dau (2010) as an alternative to 'perceptual compensation' in order to account for the listener data of Watkins (2005a). Analogous to the forward

masking paradigm, their theory proposes that listeners may become accustomed to the degree of modulation present in a signal, and thus become less sensitive to it over time. The theory of Nielsen and Dau does sufficiently explain the 'sir-stir' continuum data presented in Watkins (2005a), but it cannot account for the consonant confusion data collected in Experiment H3.

According to the modulation masking theory, the relatively large modulation content of a near-distance context will mask the plosive consonant of a following test-word (i.e., make it more difficult to identify) if that test-word is reverberated at the far distance. This effect was observed in both datasets: more 'sir' responses arose in the near-far *context-test* distance in Watkins' data, and a larger number of consonant confusions occurred in Experiment H3. Conversely, the far-reverberated context induces less masking of the following test-word's consonant, and provokes fewer mis-identifications. Again, this is observed in both datasets. For silent-context conditions, no modulation forward masking is predicted. The theory of Nielsen and Dau thus predicts a well-defined plosive dip in silent-context conditions, which ought to give rise to still *fewer* mis-identifications. However, the opposite was observed in Experiment H3: a *greater* number of confusions arose for isolated test-words than for test-words preceded by the far-reverberated context.

6.3.3 No evidence of compensation in the ventral cochlear nucleus

The computational modelling studies outlined in Chapter 4 proposed a mechanism which simulated perceptual compensation for the effects of reverberation by means of MOC efferent activity. In the model, the level of reverberation present in the auditory nerve representation was assessed (via a metric which substituted the involvement of higher auditory centres), and a resulting efferent attenuation signal which was fed back to the inner ear. By such a scheme, perceptual compensation for reverberation was shown to alter the representation of the signal in the auditory nerve (cf. Figures 4.15c and 4.15d). It follows from this proposal that compensation effects ought to be 'visible' in signals recorded in subsequent layers of the central auditory system above the auditory nerve, the next of which is the cochlear nucleus (CN). However, a study by Lehtinen et al. (2011) did *not* find evidence of compensation for reverberation in the CN of anaesthetised guinea pigs.

There appear to be two possible interpretations of this negative result. Firstly, perceptual compensation for reverberation may be a more central auditory effect, unobservable in the auditory periphery. Recent experiments with unanesthetised rabbits have indeed demonstrated some monaural robustness to reverberation in the midbrain, e.g. in the inferior colliculus (Kuwada et al., 2012), as was discussed in § 3.3.2. Nonetheless, a second possibility remains, whereby the anaesthesia used in Lehtinen et al. (2011) might have prevented the efferent system from functioning

normally (Kuwada et al., 2012; Sayles et al., 2015), and thus prevented observation of the neural correlates signifying perceptual compensation for reverberation.

6.4 Wider relevance

The computational model of efferent function, in combination with the perceptual compensation for reverberation task, has implications in the area of hearing impairment, with additional relevance to reverberation-robust machine listening.

6.4.1 Hearing impairment

It has been established via cochlear implant (CI) simulations (Poissant et al., 2006), and CI listeners (Hu and Kokkinakis, 2014), that even relatively low levels of reverberation are particularly detrimental when heard through the implant. A possible explanation for this arises from the computational modelling work described in Chapter 4.

The computational model implements a mechanism of medial olivocochlear (MOC) unmasking which, it is proposed, assists a listener to accommodate the multitude of late-arriving reflections which are present in a reverberant environment. As depicted in Figure 3.1a, the auditory periphery usually consists of a series of processes undertaken in the outer and middle ear which translate acoustic vibration into motion of the basilar membrane (BM). From here, sound is encoded as an electrical signal by the hair cells in the inner ear which is then passed upwards into the auditory nerve.

For individuals with a peripheral hearing impairment (and for whom the auditory nerve is present), a CI is often a suitable prosthetic hearing device to use. The CI bypasses the acoustic processing usually undertaken by the outer and middle ear, delivering electrical signals directly to the auditory nerve via an electrode which is inserted into the cochlea (Rubinstein, 2004). Importantly, while this intervention replaces the afferent pathway in the peripheral system, it does not repair or replace any of the efferent processing that would ordinarily be undertaken at this stage of audition. Since, for CI users, sound is not transmitted via the action of the BM, any preserved action of the outer hair cells (OHCs) may be effectively wasted. As a result, the MOC system cannot provide the feedback necessary to the peripheral filters, placing higher decoding demands on more central auditory stages.

Un-aided hearing-impaired listeners, for whom it could be theorised that the MOC effect would still be operational via the OHCs, have shown similar compensation effects as the normal-hearing population (Zahorik and Brandewie, 2011). However,

from the above analysis it seems unlikely that cochlear implanted listeners would benefit in the same way. On the other hand, if CI users *do* demonstrate compensation for reverberation then it would suggest either that central auditory processes are sufficient to provide robustness to the effects of reverberation on speech identification, or that the speech processor worn with the CI is capable of replacing some of this function.

Interestingly, efferent processing has also recently been linked with hearing impairment in another way. Liberman et al. (2014) report that efferent feedback appears to slow cochlear ageing (in mice), because the reduction of cochlear gain that is imparted by efferent signals effectively protects the hair cells from noise-induced hearing loss.

6.4.2 Reverberation-robust machine listening

The most significant benefit of increased robustness to reverberation for machine listeners would be to allow distant speech recognition. At present, commercial ASR systems (such as those offered by Google and Apple) are limited by the close-talking position which is required of their users. Far-field systems, on the other hand, require techniques to compensate for a variety of acoustic unknowns, principally background noise, competing talkers and room reverberation. Engineering approaches have resulted in methods to successfully combat distortions introduced by noise and interfering speakers, yet reverberation still poses a serious challenge for most speech recognisers (cf. § 2.2).

One of the main approaches to reverberation-robust ASR is concerned with dereverberation, that is, recovering a cleaner signal representation than the original (reverberation-corrupted) observation. While this is not the aim of the auditory model described here, it could be argued that efferent-based processing is effectively removing reverberation from the auditory nerve representation. However, this interpretation does not specifically explain the understanding of perceptual compensation for reverberation which arose from Experiment H3 (which indicated instead that human listeners may have anticipated the presence of reverberation in the signal). Rather, in borrowing the MOC unmasking scheme for the effects of late reverberation suppression, the efferent-based processing bears more resemblance to spectral-subtraction (Boll, 1979). Using the suggested channel-dependent efferent implementation, the model would thus act similarly to a continually-updating multi-band spectral subtraction method (as has recently been suggested by Upadhyay and Karmakar, 2013).

The computational model in Chapter 4 focussed on demonstrating a specific psychoacoustic phenomenon and, if viewed in terms of speech recognition, can be

thought to involve a 'back-end' recognition engine which can identify only two words. In its present implementation, the computational auditory model would be far too computationally expensive to consider extending it for large vocabulary online speech recognition. Nonetheless, such a model could in principle be used as a front-end for an HMM-based or deep neural network-based recogniser. In the tasks against which it was tested, the processing undertaken in the LPM-based efferent-model 'front-end' was sufficiently close as to allow a good simulation of human speech identification. This indicates that a spectro-temporal excitation pattern (STEP) representation might, in theory, provide an enhanced feature representation, and might therefore allow an improvement in recognition (of stop consonants at least).

Auditory models capable of perceptual compensation for reverberation might alternatively help to make predictions of the likely intelligibility of speech signals heard in particular rooms (or positions within a room). A good model of intelligibility prediction is important, particularly for educational settings such as classrooms and lecture theatres where intelligibility is critical to the function of the room itself[1]. The commonly used STI method, however, has been shown not to predict listener intelligibility in certain conditions (see e.g., Longworth-Reed et al., 2009). Thus, although it is clearly functional for many useful cases, the STI cannot yet substitute for human intelligibility judgements outright. After a study of various reverberation-detection methods (Beeston and Brown, 2013), the LPM-based efferent model was tested for generalisation across six listener/talker positions in each of two rooms (cf. Figure 4.21), and examined using multiple speech datasets (cf. Figure 4.22). From this work it appears that the exact characteristic of the reflection pattern is not critical, provided that there is energy in sufficient regions of the input signal, and that the LPM-based model stands a good chance of maintaining its performance in unknown room conditions as well.

6.5 Potential criticisms

This section addresses some potential criticisms of the research presented, and discusses the computer model, the listening experiments, and some wider factors which relate to both strands of work.

[1] Another interesting approach to this topic is taken by Culling and colleagues, who use geometrical descriptions of simulated rooms (with interfering noise sources) to optimise intelligibility in a given space. The resulting benefit is that suggestions can then be made about where to stand at a party, where to sit in a restaurant etc. in order to maximise speech intelligibility (Culling et al., 2013; Jelfs et al., 2011; Lavandier and Culling, 2010).

6.5.1 Limitations of efferent modelling

In order to make further progress in efferent auditory modelling, a greater understanding of biological auditory systems and their activation in particular tasks is needed (a requirement raised by Guinan, 2014).

The relative scarcity of efferent-related data implies that many aspects of efferent-inspired auditory models remain under-determined at present. This can be seen with regard to the efferent-DRNL studies. The MOC unmasking phenomenon was simulated initially using an open-loop configuration in which the value of attenuation was manually specified (e.g. Brown et al., 2010; Ferry and Meddis, 2007), and has subsequently been set in a closed-loop configuration in which the attenuation control parameter was derived internally (e.g. Beeston and Brown, 2010; Clark et al., 2012). The model by Clark et al. (2012) further improves the implementation of efferent suppression in such a way that MOC activation may vary in strength across channels. However, none of these models has yet implemented the off-frequency effects described by Lilaonitkul and Guinan (2009) and Zhao and Dhar (2012), nor satisfactorily resolved the potential for ambiguity in the time-course datasets published by Cooper and Guinan (2003) and Backus and Guinan (2006).

All the models described have presented a simplified picture of efferent processing, as discussed in § 4.7.2. Significiantly, the effects of the lateral olivocochlear (LOC) are not considered at all, despite the fact that the LOC efferents synapse on the afferent auditory nerve fibres themselves (Guinan, 2011) and their effects are likely therefore to be influential (Guinan, 2014). Further, ipsilateral and contralateral effects are yet to be implemented separately (Brown, 2011; Guinan, 2006). Finally, top-down effects of attention and experience, which are thought to mediate MOC activity *in addition* to the bottom-up sound-invoked reflex, remain entirely absent from all models at present. Discussion returns to this topic in § 6.5.4 below.

6.5.2 Methodological issues in the psychoacoustic studies

There are several potential criticisms of the psychoacoustic studies.

Excluded participants

A relatively high number of people were excluded from analysis because they did not meet the inclusion criteria by achieving 90% accuracy in the baseline experiment condition. Discounting three participants who were excluded because of

reported hearing losses, a further 35 people were excluded across the four experiments, representing 13.5% of the participants with normal-hearing.

The 90% inclusion criterion was selected on the basis of its use in similar studies of reverberant speech identification (see e.g., Nábělek and Robinson, 1982), and does not seem a high measure to achieve in the conditions thought to be 'easiest' (near-near *context-test*). However, as a result of the relatively small number of baseline stimulus presentations (particularly in Experiments H1 and H2), this led to the situation where participants could be excluded based on a single misidentification in the baseline condition.

A second possibility is that, due to the partitioning of sound files across participants, the excluded participants were presented with more 'difficult' stimuli to identify in these baseline experimental conditions. Studying noise-induced confusion patterns in the AIC stimuli, Phatak et al. (2008) reported that the intelligibility of consonants heard in background noise differs significantly when spoken by different talkers. Barker and Cooke (2007) also showed that speech produced by different talkers varies in intelligibility when heard in noise, and described a glimpsing-based model that predicts the intelligibility of different talkers. Initial investigation of the data gathered in Experiments H1–H4 suggests, similarly, that certain stimuli were particularly likely to cause participant exclusion. It remains for future study to investigate precisely *which* AIC talkers and test-words were frequently misidentified, and to investigate the acoustic properties of these utterances which may have led to their perceptual ambiguity.

Measures of compensation

Researchers appear to have struggled to find a good way to quantify compensation for reverberation: at least two methods were employed by Watkins (cf. Figure 2.11); multiple measures were used by Zahorik's group (cf. § 2.4.2); finally, several methods used in the Experiments H1–H4 were introduced in § 5.2.4.

Since the 2AFC response pattern in Experiment H4 was identical to that of Watkins (2005a)' listener task, the measure in that experiment (number of 's' responses) exactly mirrors the analysis of Watkins. In Experiments H1–H3, however, a 4AFC paradigm was used and participant responses were captured in 4×4 consonant confusion matrices. The measure of relative information transmitted (RIT) was selected as a convenient way in which to summarise the complex pattern of responses in a single number (Miller and Nicely, 1955; Smith, 1990). However, although convenient, this method does raise additional questions.

One potential criticism of the measure is that RIT is biased to overestimate information transfer for small samples (Miller and Nicely, 1955). This is a particular

issue in Experiments H1 and H2 where there are only 20 data points per 4×4 matrix. However, since it is the *pattern* of responses (relative differences between conditions) which is of interest rather than the absolute values, then the analysis remains valid (Sagi and Svirsky, 2008). The primary problem of the small number of data points per matrix shows itself here as a higher inherent variability in the RIT estimate, as a result of which it is more difficult to achieve statistically significant differences between the experimental conditions.

A potentially more serious criticism of the RIT measure was exposed in the meta-analysis undertaken in § 5.7.1. Since the RIT assesses the whole confusion matrix with a single number, it cannot separately account for the compensation effect which is seen for the [k], [p], [t] group and, concurrently, the *lack* of compensation which is observed for [s]. Earlier analysis (cf. § 5.7.2) revealed that this may have led to an under-estimation of the compensation effect observed for stop consonant stimuli in Experiments H1, H2 and H3, and an over-estimation of the speed of the effect for [t] stimuli in Experiment H4.

This level of detail is unfortunately not available in the published analyses by other researchers. Indeed, all studies so far have looked for only one compensation effect (i.e., have used one measure) at a time. Thus it remains to be investigated whether, for instance, the compensation effects reported in studies by Zahorik and colleagues are similarly reliant on effects occurring for one particular class of speech sound (e.g., plosive consonants), or whether the compensation effects they measure are spread more evenly across speech categories. To know the ecological validity of the compensatory mechanisms, the eventual goal should be to find a measure of compensation that can be applied to connected speech. This would allow investigations to examine compensation for reverberation with conversational speech, including natural 'difficulties' such as overlapping talkers which are a feature of daily experience.

Pre-filtering of speech materials

Section 5.3 argued that, in order to test effects of reverberation on speech identification, it is necessary to find the operating point where (i) some degradation in listener performance is reported and (ii) some recovery from that 'low-point' can be observed. However, human speech recognition is generally good, even in reverberant environments. The speech signal itself is highly redundant, and likely contains multiple cues for speech identification. For example, French and Steinberg (1947) demonstrated that speech is equally intelligible whether high-pass or low-pass filtered at around 1900 Hz, but that each 'half' of the filtered speech allowed c. 68% intelligibility alone.

Studies into compensation for reverberation have therefore typically introduced some further signal distortion in order to make the listeners' task harder, i.e. to guarantee that some misidentification will occur. Table 2.3 previously showed that Zahorik's group increased the difficulty of the listener task by introducing a simultaneous (spatialised) background noise. The psychoacoustic experiments presented in this thesis introduced low-pass filtering instead. Motivated by the study of Miller and Nicely (1955), a range of low-pass filter conditions were tested in Experiment H1, and the 4 kHz cutoff condition was then used in all subsequent experiments. The bonus of using low-pass filtering over spatialised noise is that it is unlikely to involve binaural hearing mechanisms which seemed to play a role in Brandewie and Zahorik (2010). However, the low-pass filtering does reduce the high-frequency variability in the signal which has been shown to be crucial for consonant perception.

It could be argued, therefore, that the inclusion of the low-pass filtering stage had an impact on the compensation results obtained. Appendix A below, however, presents preliminary results which argue against this criticism. Comparing unfiltered (wideband) and filtered (4 kHz) speech stimuli side-by-side, Figure A.1 shows that perceptual compensation for reverberation occurred in both conditions. Here, an increased level of reverberation in the context, resulting from an increase in context distance, brought about an improvement in the identification of the far-distance test-words, whether filtered or not. Although this pattern of responses was collected from only a small number of listeners (eight listeners, the minimum possible given the eight experimental conditions and partitioning of stimuli required), this pattern of responses suggests that the pre-filtering of speech materials is not required for demonstrating the monaural compensation for reverberation effect. The low-pass filtering stage could likely therefore be omitted from future experiments investigating perception of reverberated plosives with naturally spoken speech stimuli.

6.5.3 Simulated talkers and listening via head-phones

Since reverberant speech in the present experiments was created via convolution of room impulse responses with dry speech signals, the talker could not slow down or otherwise accommodate to the reverberation condition as would have happened in real life (Barker et al., 2013; Black, 1950). Since it has been shown that listeners tend to prefer slower speech in reverberant environments (Moore et al., 2007), it is likely that this may have contributed somewhat to the difficulty of the speech identification task.

Additionally, since sounds were delivered over a single headphone channel, listeners could not move their heads to resolve confusion as they might in a free-field

listening environment (cf. § 2.3.1). Interestingly, Kim et al. (2013) report that head movement is dependent on the task being undertaken; listeners moved their heads more to estimate source width than to judge timbre. It is unclear at present exactly what relevance this would have for compensation for reverberation. However, such head movements do appear to be relevant for identification (as well as for localisation), since intelligibility in a room also varies as a function of the listener's head orientation (Culling et al., 2013).

6.5.4 Multi-sensory integration and non-auditory factors

Another factor which has been not been addressed in either the psychoacoustic studies or the computational model, is that hearing is modulated by our other senses, and by a number of non-auditory processes such as attention and experience.

Since the brain uses multiple senses (and multiple strategies) to undertake tasks (Bolognini et al., 2007; London et al., 2012), it seems unlikely that compensation for reverberation will be unaffected by such multi-sensory integration and non-auditory factors. In particular, vision can strongly influence both speech identification (McGurk and MacDonald, 1976; Schroeder et al., 2008) and the spatial impression perceived in localisation tasks (Bishop et al., 2011, 2012; Murray and Spierer, 2011). Moreover, London et al. (2012) propose that attention to low-level spatial aspects can influence the formation of individual auditory objects, which has knock-on effects on our understanding of the acoustic scene at large, and also for speech perception in particular (Shinn-Cunningham et al., 2013).

While it is clear that MOC responses are modulated by attention, the principles governing these actions have not been elucidated in the literature. Since attention seeks to reduce the complexity of the scene analysis and focus on most relevant object at one time (Fritz et al., 2007), it is likely that so-called 'attention' effects will actually vary considerably from one task to the next. Perhaps as a result, Guinan (2010) hypothesised that MOC activity may increase when it is 'beneficial' for the listener task (and decrease otherwise). To model attentional effects in this manner, e.g. to simulate the task-related head movement in Kim et al. (2013), is well beyond the scope of the state-of-the-art techniques in efferent computational models at present.

Additionally, MOC effects appear to depend on listeners' experience, and to be trainable themselves. This has led researchers to suggest that the MOC suppression might *itself* be under the control of other 'higher' efferent signals (as shown in Figure 3.1b) from either the cortex (De Boer and Thornton, 2008) or the inferior colliculus (Brown, 2011). If the MOC were indeed implicated in reverberant

listening as this thesis hypothesises, then it may help to account for the benefits that musical training (which promotes highly skilled listening) appears to offer in certain listener tasks. In particular, musical training was seen to reduce listeners' sensitivity to reverberation in Bidelman and Krishnan (2010), and to enhance performance in a gap detection task in Mishra et al. (2014).

6.6 Future work

A few avenues for further research are explored in greater detail in this section.

6.6.1 Computational studies

As described above in § 6.5.1 above, a great deal of the limitations in efferent-modelling are due to an incomplete understanding of human auditory function as it is applied to particular tasks. However, the process of auditory modelling has itself proved useful for making predictions in regard to how a system *may* function, and for assisting with the design of psychoacoustic experiments which can gather data to validate proposed hypotheses.

For instance, one of the next suggested developments for the auditory model (cf. § 6.1.3) would be to implement efferent feedback within individual channels, so that the value of efferent attenuation applied in the DRNL filter differs across frequency as well as through time (cf. Figure 4.3). This band-by-band hypothesis has been suggested to be relevant in perceptual compensation for reverberation, seen first in listener responses to vocoded 'sir-stir' stimuli in Watkins et al. (2011), and later supported by consonant confusion data gathered in Experiment H1 above (cf. 5.3.5). Additionally, such a band-by-band approach has improved performance of efferent-based ASR in experiments by Lee et al. (2011) and Clark et al. (2012).

Additionally, it is expected that further improvements could be made to the model by simulating a greater range of experiments in which human responses have already been collected. Further 'sir-stir' experiments by Watkins and colleagues examined alternative speaking rates and different rooms (Watkins, 2005a), single-, multiple- and wide-band noise contexts (Watkins and Makin, 2007a), tonal contexts (Watkins and Makin, 2007c), and vocoded speech stimuli (Watkins et al., 2011). With some development of the 'back end' speech identification process in the computational model, detailed consonant confusion patterns could also be matched for the responses gathered using the Articulation Index Corpus (AIC) dataset in Experiments H1–H4, or world-level recognition performance could be modelled for the studies undertaken by Zahorik and colleagues (listed in Table 2.3 above).

For scalability and comparison of results among differing datasets, definition of a battery of test-stimuli and associated listener responses would be extremely helpful. Further, it is anticipated that techniques such as sensitivity analysis and uncertainty quantification would help to identify *which* model parameters it would be particularly beneficial to gather perceptual data on in future listening studies (Sacks et al., 1989). In addition to the macroscopic modelling tasks undertaken so far (where the computer simulation is trained to match the average of all listener responses), it would also be interesting to approach this on the microscopic level (cf. § 6.5.2), so that (i) responses to individual utterances and (ii) responses of individual listeners may be examined in detail (Cooke et al., 2006; Meddis et al., 2010).

6.6.2 Perceptual studies

Experiments H1–H4 took Watkins' monaural demonstration of compensation for reverberation from the 'sir-stir' continuum stimuli, and generalised this to the case of naturally spoken material. These experiments did not, however, generalise the reverberation conditions themselves which were used to demonstrate the constancy effect.

A number of different rooms (real and simulated) have been investigated by researchers interested in the compensation effect, however the relative distance between talker and listener has yet to be studied. The experiments in this thesis re-used the left channel of binaural room impulse responses that were originally recorded at 'near' and 'far' positions in an L-shaped room (as described in Watkins, 2005a). The far condition, measured at a source-receiver distance (SRD) of 10 m between the centre of the talker and listener heads suggests, for instance, listening to a lecturer at the front of a seminar room while being seated several rows back. Contrasting this, the near condition (at only 0.32 m SRD) represents listening to the person sat in the neighbouring seat. In future perceptual studies it would be of interest to examine perceptual compensation using different near- and far-SRDs, and to see whether differences in their relative position would make a measurable difference to the time course of the constancy effect. In this way, it could be asked whether listeners might take longer to adapt to a speaker that is heard from a greater distance away.

Background noise is another factor which appears to affect the speed of the compensation mechanism, and was reported by Brandewie and Zahorik (2013) to slow down compensation for reverberation for tasks in which listeners binaurally listened to target speech from straight ahead, with a noise-source presented from side-on. It remains to be tested whether this finding would also be true for monaural compensation, or whether it relied in part on the binaural presentation method

which was used. There is little noise occurring in the monaural experiments studied so far, since noise was removed from the impulse responses used by Watkins (cf. Figure2.4b), and was effectively absent in both the 'sir-stir' and AIC speech materials. Nonetheless, additive noise is of particular interest to the current modelling investigation due to the adoption of the noise-related MOC unmasking effect (cf. § 3.4.3). Noise itself does not prevent compensation, and indeed can promote it in cases where there is ample modulation across sufficient channels in the context (Watkins and Makin, 2007a). However, since the compensation mechanism appears to be effected in a primarily within-band manner (Watkins et al., 2011), it would be interesting in addition to consider effects of spectral centroid and bandwidth with such simultaneous noise sources.

Finally, the interaction between the two sides of the auditory system is another area which would be of particular interest in the context of the current computational modelling effort since, as Lopez-Poveda et al. (2013) point out, the two human cochleae are linked at the level of MOC efferent processing. Different groups of MOC efferents exist in different pathways, activated by sound occurring in the ipsilateral ear, contralateral ear or, for a small group of neurons, *either* ear (Brown, 2011; Guinan, 2006). If it can be assumed that these reflexes are sufficiently fast so as to occur within individual utterances, then it may be feasible to explore them within the current psychoacoustic experimental paradigm. For instance, it would be interesting to know whether compensation persists if the context is presented to one ear and the test-word occurs in the other. Further, in comparison to the monaural demonstration of the effect, it would be interesting to examine whether compensation is enhanced by a context which occurs in both ears when the test-word is heard in only one ear.

Concluding remarks

This thesis proposes a candidate model of compensation for reverberation in which efferent auditory processing is responsible for adjusting the sound-encoding behaviour of the cochlea, based on the level of reverberation detected in the environment. Perceptual studies indicate that the monaural compensation effect is likely to be relevant in everyday listening, assisting listeners to re-calibrate rapidly when encountering a new listening environment.

Appendix A

Pre-filtering of speech materials

Section 5.3 described an experiment in which perceptual compensation for the effects of reverberation was apparent in a consonant identification task using spoken material recorded by a range of talkers and with speech contexts which varied from trial to trial. The remaining experiments presented in Chapter 5 investigated further conditions in which compensation was either promoted or was blocked. However, in each of these experiments, speech stimuli were low-pass filtered at 4 kHz, replicating conditions which had promoted compensation for reverberation in Experiment H1. The following experiment, modelled closely on Experiment H1, asks whether compensation for reverberation may be demonstrated monaurally without the prior low-pass filtering of speech stimuli.

Experiment H5: Compensation for reverberation with wideband speech materials

Sections 2.3.3 and 2.3.4 argued that speech identification cues which are reliant on the perception of dips in the temporal envelope are particularly susceptible to the effects of reverberation. For these consonants, reflected energy that persists beyond the signal offset fills the periods of low energy which would otherwise indicate the plosive dip.

This observation underlies the 'sir-stir' continuum experiments of Watkins and colleagues as described in § 2.4.1. Here, an amplitude modulation cue altered the identity of test-words across a continuum by varying the depth of the modulation applied. When a stimulus with the synthetically introduced temporal dip was presented in a low-level of reverberation, listeners' typically identified the test-word

as ‘stir’. However, when the same stimulus was presented in a higher level of reverberation, listeners typically responded ‘sir’. In this way, reverberation can be understood to have ‘undone’ the amplitude modulation processing that was introduced.

It is not immediately obvious whether naturally spoken material would be systematically affected in a similar manner to that outlined by Watkins (2005a, b). Indeed, the relatively small number of speech identification errors typically reported in reverberant speech perception studies appears often to have influenced researchers to include additional factors which would make the listeners’ task somewhat harder (cf. Table 2.2). The same has been true for experiments examining compensation for reverberation. Zahorik and colleagues, for instance, addressed this issue by including a simultaneous noise source in addition to the reverberated speech conditions (used in 6 of their 8 studies listed in Table 2.3). Unfortunately, this somewhat muddied subsequent analyses as it was difficult to attribute observed results to solely reverberation- or noise-compensation strategies.

Following the observation of Miller and Nicely (1955) that cues regarding *place* of articulation are severely degraded in low-pass filtered speech, the studies presented in Chapter 5 used low-pass filtering (rather than noise) to increase the difficulty of the listeners’ task. Speech stimuli were low-pass filtered at a range of cutoff values in Experiment H1 to find a suitable operating point at which compensation for reverberation could be achieved with naturally spoken stimuli. From these results, the 4 kHz condition was selected for use in later experiments since, of those conditions tested which permitted a clear demonstration of compensation, it was the most ‘normal’ (i.e., least strongly filtered). Although the 4 kHz filtering is relatively mild, and indeed matches conditions frequently heard (e.g., in telephone conversations), it does nonetheless reduce the closeness of the experimental paradigm to listening to natural speech in real rooms. Experiment H5, presented in this appendix, therefore asks whether it was necessary to pre-filter the speech material at all.

Stimuli

Utterances selected for Experiment H5 were identical in form to those used in Experiment H1 above, comprising

[CW1][CW2][TEST][CW3]

in which reverberation conditions of the test-word (TEST) and the context words (CW) could again be independently manipulated. Using the same four initial consonant conditions, {[s], [sk], [sp], [st]}, the test-word set extended those used in

Experiment H1 with a second vowel in addition, {[æ], [ɝ]}. Again, all talkers from the corpus were used, thus the current experiment employed 160 Articulation Index Corpus (AIC) utterances (20 talkers × 4 consonants × 2 vowels).

As was described previously, stimuli in the filtered experimental condition were created by low-pass filtering with an 8^{th} order Butterworth filter at a cutoff frequency of 4 kHz matching the experimental condition used in Experiments H1–H4 above (cf. § 5.3.1). This process was omitted for the wideband condition. Same- and mixed-distance phrases were then created exactly by the methods described for Experiment H1, resulting in phrases as described by the illustration provided in Figure 5.1. Other details of stimulus creation followed the details provided in § 5.3.1. The set of sound files for Experiment H5 thus comprised 1280 stimuli (160 AIC utterances × 2 filter conditions × 2 context distances × 2 test distances).

Participants and procedures

A single listener group comprising eight individuals took part in this study. As in the earlier experiments, stimuli were partitioned among participants in order that each AIC phrase was heard in a single experimental condition (ensuring that there could be no association between the test and context portions of the phrase which might otherwise assist identification of the test-word). Every participant heard 160 different phrases, comprising 20 items in each of 8 experimental conditions. Vowels were divided evenly across the listener group, and stimuli rotated among participants so that each listener heard every test consonant five times in each condition, with the five instances being from different phrases (and thus were spoken by different talkers). As in the earlier experiments, participants identified the initial part of the word only, by choosing among buttons labelled ‘s’, ‘sk’, ‘sp’ or ‘st’. Other aspects of stimulus presentation were as described in section 5.3.3, and the experiment was typically completed in around 10 minutes.

Results and discussion

Preliminary results in Figure A.1 show that perceptual compensation for reverberation occurred in both the 4 kHz (filtered) and wideband (unfiltered) stimulus conditions. In both conditions, an increased level of reverberation in the context, resulting from an increase in context distance, brought about an improvement in the identification of the far-distance test-words.

The pattern of results for the 4 kHz condition (Figure A.1, *left*) closely resembles those reported in Figure 5.2, and the major features of the compensation for reverberation paradigm are all replicated. At near-near *context-test* distances, few con-

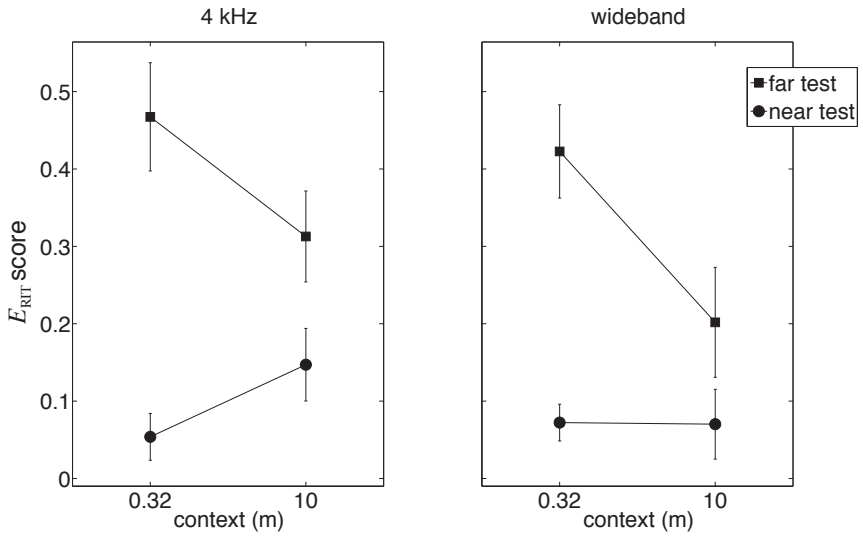


Figure A.1: Mean and standard error of 8 participants' E_{RIT} scores (cf. Equation 5.4) for the 4 kHz lowpass filter cutoff condition (*left*) and for wideband (unfiltered) speech (*right*). The lower line reports near-distance test-word scores, where listeners made few errors. The upper line reports the far-distance test-word scores. For both filter conditions, fewer misclassifications resulted when the reverberation in the context was increased to the far-distance condition.

sonants were misclassified, and the error measure, E_{RIT} (cf. Equation 5.4), was relatively low. In near-far stimulus conditions, a larger proportion of the stimuli were misclassified, and the E_{RIT} measure increased. However, in far-far stimulus conditions the test-words were easier for listeners to identify correctly, and the E_{RIT} measure decreased somewhat. In the near-far wideband stimulus condition (Figure A.1, *right*), a similar number of reverberant test-words were misclassified as in the 4 kHz condition. Moreover, the recovery due to the increased context-reverberation at far-far appears to be even larger than that which was observed in the 4 kHz condition.

Although collected from a small number of listeners, this pattern of responses suggests that the pre-filtering of speech materials is not required for demonstrating the monaural compensation for reverberation effect. The low-pass filtering stage could likely therefore be omitted from future experiments investigating perception of reverberated plosives with naturally spoken speech stimuli.

Glossary

2AFC two-alternative forced-choice.

AI Articulation Index.
AIC Articulation Index Corpus.
AM amplitude modulation.
AN auditory nerve.
ANOVA analysis of variance.
AR acoustic reflex.
ASR automatic speech recognition.

BM basilar membrane.
BRIR binaural room impulse response.

CI cochlear implant.
CN cochlear nucleus.

DRNL dual-resonance nonlinear.
DRR direct-sound to reverberant-sound ratio.

EDC energy decay curve.
EDT early decay time.
ERB equivalent rectangular bandwidth.

HERB harmonic dereverberation.
HMM hidden Markov model.

IC inferior colliculus.
IHC inner hair cell.
ILD interaural level difference.
ITD interaural time difference.

LOC lateral olivocochlear.
LPM low-pass mask.

MBPNL multiple band-pass non-linearity.
MFCC Mel-frequency cepstrum coefficient.
MOC medial olivocochlear.
MOCR medial olivocochlear reflex.
MPR mean-to-peak ratio.
MSE mean squared error.
MTF Modulation Transfer Function.

NMR noise-to-mask ratio.

OAE otoacoustic emissions.
OC olivocochlear.
OCB olivocochlear bundle.
OHC outer hair cell.
OME outer and middle ear.

PLP perceptual linear predictive.
PNCC power-normalized cepstral coefficient.
PSTH post-stimulus time histogram.

RIR room impulse response.
RIT relative information transmitted.
RMS root mean square.

SII Speech Intelligibility Index.
SNR signal-to-noise ratio.
SO superior olive.
SOC superior olivary complex.
SPL sound pressure level.
SR spontaneous rate.
SRD source-receiver distance.
SRM spatial release from masking.
SRR signal-to-reverberation ratio.
SRT Speech Reception Threshold.
STEP spectro-temporal excitation pattern.
STFT short-time Fourier transform.
STI Speech Transmission Index.
STMI spectro-temporal modulation index.

TFS temporal fine structure.
TM tectorial membrane.

WER word error rate.

References

A. Acero, L. Deng, T. Kristjansson, and J. Zhang. HMM adaptation using vector taylor series for noisy speech recognition. In *Proc Int Conf Spoken Lang Process (ICSLP)*, Beijing, China, October 2000.

E. H. Adelson. Lightness perception and lightness illusions. In *The new cognitive neurosciences*, pages 339–351. MIT Press, Cambridge, MA, second edition, 2000.

J. B. Allen. Image method for efficiently simulating small-room acoustics. *J Acoust Soc Am*, 65(4):943–950, 1979.

J. B. Allen. Cochlear modeling - 1980. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, pages 768–769, Atlanta, GA, May 1981.

J. B. Allen. Harvey Fletcher's role in the creation of communication acoustics. *J Acoust Soc Am*, 99(4):1825–1840, 1996.

J. B. Allen and F. Li. Speech perception and cochlear signal processing. *IEEE Signal Proc Mag*, 26(4):73–77, 2009.

T. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE T Acoust Speech*, 27(4):309–319, 1979.

M. Ardoint, C. Lorenzi, D. Pressnitzer, and A. Gorea. Investigation of perceptual constancy in the temporal-envelope domain. *J Acoust Soc Am*, 123(3):1591–1601, 2008.

I. Arweiler and J. M. Buchholz. The influence of spectral characteristics of early reflections on speech intelligibility. *J Acoust Soc Am*, 130(2):996–1005, 2011.

C. Avendano and H. Hermansky. Study on the dereverberation of speech based on temporal envelope filtering. In *Proc Int Conf Spoken Lang Process (ICSLP)*, pages 889–892, Philadelphia, PA, Oct. 1996.

B. C. Backus and J. J. Guinan, Jr. Time-course of the human medial olivocochlear reflex. *J Acoust Soc Am*, 119(5):2889–2904, 2006.

J. Barker and M. Cooke. Modelling speaker intelligibility in noise. *Speech Communication*, 49(5):402–417, 2007.

J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green. The PASCAL CHiME speech separation and recognition challenge. *Comput Speech Lang*, 27(3):621–633, 2013.

S. Bech. Timbral aspects of reproduced sound in small rooms. I. *J Acoust Soc Am*, 97(3): 1717–1726, 1995.

A. V. Beeston and G. J. Brown. Perceptual compensation for effects of reverberation in speech identification: a computer model based on auditory efferent processing. In *Proc 11th Conf Int Speech Commun As (INTERSPEECH)*, pages 2462–2465, Makuhari, Chiba, Japan, Sept. 2010.

A. V. Beeston and G. J. Brown. Modelling reverberation compensation effects in time-forward and time-reversed rooms. In *UK Speech Conference*, Cambridge, UK, Sept. 2013.

A. V. Beeston and G. J. Brown. Consonant confusions provide further evidence that time-reversed rooms disturb compensation for reverberation. In *Proc Forum Acust*, Krakow, Poland, Sept. 2014.

A. V. Beeston, G. J. Brown, and A. J. Watkins. Perceptual compensation for the effects of reverberation on consonant identification: evidence from studies with monaural stimuli. *J Acoust Soc Am*, 136(6):3072–3084, 2014.

G. M. Bidelman and A. Krishnan. Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Res*, 1355(C):112–125, 2010.

C. W. Bishop, S. London, and L. M. Miller. Visual influences on echo suppression. *Curr Biol*, 21(3):221–225, 2011.

C. W. Bishop, S. London, and L. M. Miller. Neural time course of visually enhanced echo suppression. *J Neurophysiol*, 108(7):1869–1883, 2012.

C. W. Bishop, D. Yadav, S. London, and L. M. Miller. The effects of preceding lead-alone and lag-alone click trains on the buildup of echo suppression. *J Acoust Soc Am*, 136(2): 803–817, 2014.

J. W. Black. The effect of room characteristics upon vocal intensity and rate. *J Acoust Soc Am*, 22:174–176, 1950.

J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT Press, Cambridge, MA, revised edition, 1997.

P. Boersma and D. Weenink. Praat, version 5.0.40. http://www.praat.org/, 2010. [Online; accessed 1 Jul 2010].

S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE T Acoust Speech*, 27(2):113–120, 1979.

N. Bolognini, F. Leo, C. Passamonti, B. E. Stein, and E. Ladavas. Multisensory-mediated auditory localization. *Perception*, 36(10):1477–1486, 2007.

R. H. Bolt and A. D. MacDonald. Theory of speech masking by reverberation. *J Acoust Soc Am*, 21(6):577–580, 1949.

J. S. Bradley, R. D. Reich, and S. G. Norcross. On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility. *J Acoust Soc Am*, 106(4):1820–1828, 1999.

J. S. Bradley, H. Sato, and M. Picard. On the importance of early reflections for speech in rooms. *J Acoust Soc Am*, 113(6):3233–3244, 2003.

E. J. Brandewie and P. Zahorik. Prior listening in rooms improves speech intelligibility. *J Acoust Soc Am*, 128(1):291–299, July 2010.

E. J. Brandewie and P. Zahorik. Adaptation to room acoustics using the Modified Rhyme Test. *Proc Meet Acoust*, 12:050007, 2012.

E. J. Brandewie and P. Zahorik. Time course of a perceptual enhancement effect for noise-masked speech in reverberant environments. *J Acoust Soc Am*, 134(2):EL265–EL270, 2013.

A. W. Bronkhorst. The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acta Acust united Ac*, 86(1):117–128, 2000.

G. J. Brown, R. T. Ferry, and R. Meddis. A computer model of auditory efferent suppression: implications for the recognition of speech in noise. *J Acoust Soc Am*, 127(2): 943–954, 2010.

M. C. Brown. Anatomy of olivocochlear neurons. In D. K. Ryugo, R. R. Fay, and A. Popper, editors, *Auditory and vestibular efferents*, pages 17–38. Springer, New York, NY, 2011.

I. C. Bruce, M. B. Sachs, and E. D. Young. An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. *J Acoust Soc Am*, 113(1):369–388, 2003.

M. Büchler, S. Allegro, S. Launer, and N. Dillier. Sound classification in hearing aids inspired by auditory scene analysis. *EURASIP J Appl Si Pr*, 18:2991–3002, 2005.

M. Bürck and J. L. van Hemmen. Modeling the cochlear nucleus: a site for monaural echo suppression? *J Acoust Soc Am*, 122(4):2226–2235, 2007.

S. Campanini and A. Farina. A new Audacity feature: room objective acoustical parameters calculation module. In *Linux Audio Conference*, Parma, Italy, Apr 2009.

X.-J. Cao and D. Oertel. Auditory nerve fibers excite targets through synapses that vary in convergence, strength, and short-term plasticity. *J Neurophysiol*, 104(5):2308–2320, 2010.

L. H. Carney. A model for the responses of low-frequency auditory-nerve fibers in cat. *J Acoust Soc Am*, 93(1):401–417, 1993.

A. Chabot-Leclerc, S. Jørgensen, and T. Dau. The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction. *J Acoust Soc Am*, 135(6):3502–3512, 2014.

A. Chesnokov and L. SooHoo. Influence of early to late energy ratios on subjective estimates of small room acoustics. In *105th Audio Eng Soc Conv*, number 4857, San Francisco, CA, Sept. 1998.

T. Chi, P. Ru, and S. A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am*, 118(2):887–906, 2005.

A. Chintanpalli, S. G. Jennings, M. G. Heinz, and E. A. Strickland. Modeling the antimasking effects of the olivocochlear reflex in auditory nerve responses to tones in sustained noise. *J Assoc Res Otolaryngol*, 13(2):219–235, 2012.

N. R. Clark, G. J. Brown, T. Jürgens, and R. Meddis. A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise. *J Acoust Soc Am*, 132(3):1535–1541, 2012.

M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition (L). *J Acoust Soc Am*, 120(5):2421–2424, 2006.

N. P. Cooper and J. J. Guinan, Jr. Separate mechanical processes underlie fast and slow effects of medial olivocochlear efferent activity. *J Physiol*, 548(1):307–312, 2003.

N. P. Cooper and J. J. Guinan, Jr. Efferent-mediated control of basilar membrane motion. *J Physiol*, 576(1):49–54, 2006.

L. Couvreur and C. Couvreur. Blind model selection for automatic speech recognition in reverberant environments. *J VLSI Sig Proc Syst*, 36(2–3):189–203, 2004.

R. M. Cox, G. C. Alexander, and C. Gilmore. Intelligibility of average talkers in typical listening environments. *J Acoust Soc Am*, 81:1598–1608, 1987.

K. E. Cullen and L. B. Minor. Semicircular canal afferents similarly encode active and passive head rotation: implications for the role of vestibular efference. *J Neurosci*, 22 (RC226):1–7, 2002.

J. Culling, M. Lavandier, and S. Jelfs. Predicting binaural speech intelligibility in architectural acoustics. In J. Blauert, editor, *The technology of binaural listening*, Modern Acoustics and Signal Processing, pages 427–447. Springer, Berlin, 2013.

J. F. Culling and E. R. Mansell. Speech intelligibility among modulated and spatially distributed noise sources. *J Acoust Soc Am*, 133(4):2254–2261, 2013.

J. F. Culling, K. I. Hodder, and C. Y. Toh. Effects of reverberation on perceptual segregation of competing voices. *J Acoust Soc Am*, 114(5):2871–2876, 2003.

J. F. Culling, B. A. Edmonds, and K. I. Hodder. Speech perception from monaural and binaural information. *J Acoust Soc Am*, 119(1):559–565, 2006.

K. N. Darrow, S. F. Maison, and L. M. C. Cochlear efferent feedback balances interaural sensitivity. *Nat Neurosci*, 9(12):1474–1476, 2006.

C. J. Darwin. Perceiving vowels in the presence of another sound: constraints on formant perception. *J Acoust Soc Am*, 76(6):1636–1647, 1984.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE T Acoust Speech*, 28(4):357–366, 1980.

E. de Boer. Travelling waves and cochlear resonance. *Scand Audiol Suppl*, 9:17–33, 1979.

J. De Boer and A. R. D. Thornton. Neural correlates of perceptual learning in the auditory brainstem: efferent activity predicts and reflects improvement at a speech-in-noise discrimination task. *J Neurosci*, 28(19):4929–4937, 2008.

L. Deng, J. Droppo, and A. Acero. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE T Speech Audi P*, 13(3):412–421, May 2005.

S. Devore and B. Delgutte. Effects of reverberation on the directional sensitivity of auditory neurons across the tonotopic axis: influences of interaural time and level differences. *J Neurosci*, 30(23):7826–7837, 2010.

S. Devore, A. Ihlefeld, K. Hancock, B. G. Shinn-Cunningham, and B. Delgutte. Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain. *Neuron*, 62(1):123–134, 2009.

A. Di Scipio. Sound is the interface: from interactive to ecosystemic signal processing. *Organised Sound*, 8(3):269–277, 2003.

N. Ding and J. Z. Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol*, 107(1):78–89, 2011.

N. Ding and J. Z. Simon. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J Neurosci*, 33(13):5728–5735, 2013a.

N. Ding and J. Z. Simon. Robust cortical encoding of slow temporal modulations of speech. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, editors, *Basic aspects of hearing: physiology and perception*, pages 373–381. Springer, New York, 2013b.

N. Ding and J. Z. Simon. Cortical entrainment to continuous speech: functional roles and interpretations. *Front Human Neurosci*, 8(311), 2014.

J. Droppo and A. Acero. Environmental robustness. In J. Benesty, M. M. Sondhi, and Y. A. Huang, editors, *Springer handbook of speech processing*. Springer-Verlag, Berlin,

2008.

R. Drullman, J. M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J Acoust Soc Am*, 95(5):2670–2680, 1994a.

R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am*, 95(2):1053–1064, 1994b.

J. R. Dubno, J. B. Ahlstrom, X. Wang, and A. R. Horwitz. Level-dependent changes in perception of speech envelope cues. *J Assoc Res Otolaryngol*, 13(6):835–852, 2012.

N. I. Durlach. Equalization and cancellation theory of binaural masking-level differences. *J Acoust Soc Am*, 35(8):1206–1218, 1963.

B. A. Edmonds and J. F. Culling. The spatial unmasking of speech: evidence for better-ear listening. *J Acoust Soc Am*, 120(3):1539–1545, 2006.

J. P. Egan, G. Z. Greenberg, and A. I. Schulman. Interval of time uncertainty in auditory detection. *J Acoust Soc Am*, 33(6):771–778, 1961.

M. Elhilali and S. A. Shamma. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *J Acoust Soc Am*, 124(6):3751–3771, 2008.

M. Elhilali, T. Chi, and S. A. Shamma. A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Commun*, 41(2–3):331–348, 2003.

D. P. W. Ellis and R. J. Weiss. Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, Toulouse, France, May 2006.

S. Ellison and P. Germain. Optimizing acoustics for spoken word using active acoustics. *Proc Meet Acoust*, 19:015073, 2013.

K. Eneman, J. Duchateau, M. Moonen, D. van Campernolle, and H. van Hamme. Assessment of dereverberation algorithms for large vocabulary speech recognition systems. In *Proc 8th Eur Conf Speech Commun Technol (EUROSPEECH)*, pages 2689–2692, Geneva, Switzerland, Sept. 2003.

A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *108th Audio Eng Soc Conv*, number 5093, Paris, France, Feb. 2000.

N. H. Feldman and T. L. Griffiths. A rational account of the perceptual magnet effect. In *Proc 29th Conf Cogn Sci Soc (COGSCI)*, pages 257–262, Nashville, TN, Aug. 2007.

R. Ferry and R. Meddis. A computer model of medial efferent suppression in the mammalian auditory system. *J Acoust Soc Am*, 122(6):3519–3526, 2007.

H. Fletcher and R. H. Galt. The perception of speech and its relation to telephony. *J Acoust Soc Am*, 22(2):89–150, 1950.

M. Fletcher, J. de Boer, and K. Krumbholz. Is overshoot caused by an efferent reduction in cochlear gain? In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, editors, *Basic aspects of hearing: physiology and perception*, pages 65–72. Springer, New York, 2013.

N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *J Acoust Soc Am*, 19(1):90–119, 1947.

J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma. Auditory attention – focusing the searchlight on sound. *Curr Opin Neurobiol*, 17(4):437–455, 2007.

G. I. Frolenkov, M. Atzori, F. Kalinec, F. Mammano, and B. Kachar. The membrane-based mechanism of cell motility in cochlear outer hair cells. *Mol Biol Cell*, 9(8):1961–1968, 1998.

K. Furuya and A. Kataoka. Robust speech dereverberation using multi-channel blind deconvolution with spectral subtraction. *IEEE T Audio Speech*, 15(5):1579–1571, 2007.

M. Gales, S. Young, and S. J. Young. Robust continuous speech recognition using parallel model combination. *IEEE T Speech Audi P*, 4:352–359, 1996.

W. F. Ganong, III and R. J. Zatorre. Measuring phoneme boundaries four ways. *J Acoust Soc Am*, 65(2):431–439, 1979.

B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. Perceptual Computing - Technical Report 280, MIT Media Lab, 1994.

J.-l. Gauvain and C.-h. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE T Speech Audi P*, 2:291–298, 1994.

S. A. Gelfand. *Hearing: an introduction to psychological and physiological acoustics*. Dekker, New York, second edition, 1990.

S. A. Gelfand and S. Silman. Effects of small room reverberation upon the recognition of some consonant features. *J Acoust Soc Am*, 66(1):22–29, 1979.

E. Georganti, N. Dillier, and J. Mourjopoulos. Measuring perception of coloration due to early reflections in binaural room impulse responses. In *Proc Forum Acust*, Krakow, Poland, Sept. 2014.

E. L. J. George, J. M. Festen, and T. Houtgast. The combined effects of reverberation and nonstationary noise on sentence intelligibility. *J Acoust Soc Am*, 124(2):1269–1277, 2008.

O. Ghitza. Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Comput Speech Lang*, 1(2):109–130, 1986.

O. Ghitza. Temporal non-place information as a front-end for speech recognition in a noisy environment. *J Phonetics*, 16(1):109–123, 1988.

O. Ghitza. Using auditory feedback and rhythmicity for diphone discrimination of degraded speech. In *Proc 16th Int Congr Phonetic Sci*, Saarbrücken, Germany, Aug. 2007.

C. Giguère and P. C. Woodland. A computational model of the auditory periphery for speech and hearing research. II. Descending paths. *J Acoust Soc Am*, 95(1):343–349, 1994.

C. Giguère, A. J. Bosman, and G. F. Smoorenburg. Automatic speech recognition experiments with a model of normal and impaired peripheral hearing. *Acta Acust United Ac*, 83(6):1065–1076, 1997.

B. W. Gillespie and L. E. Atlas. Strategies for improving audible quality and speech recognition accuracy of reverberant speech. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, pages 676–679, Hong Kong, Apr. 2003.

B. W. Gillespie, H. S. Malvar, and A. F. Florencio. Speech dereverberation via maximum-kurtosis subband adaptive filtering. *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, 6:3701–3704, May 2001.

D. Giuliani, M. Omologo, and P. Svaizer. Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation. *Proc 4th Int Conf Spoken Lang Process (ICSLP)*, 1:1329–1332, Oct. 1996.

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

B. R. Glasberg and B. C. J. Moore. Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *J Acoust Soc Am*, 108(5):

2318–2328, 2000.

H. E. Gockel and H. Colonius. Auditory profile analysis: is there perceptual constancy for spectral shape for stimuli roved in frequency? *J Acoust Soc Am*, 102(4):2311–2315, 1997.

J. L. Goldstein. Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering. *Hear Res*, 49(1–3):39–60, 1990.

F. Gomez, V. Saase, N. Buchheim, and R. Stoop. How the ear tunes in to sounds: a physics approach. *Phys Rev Applied*, 1:014003, 2014.

D. M. Green and T. G. Forrest. Temporal gaps in noise and sinusoids. *J Acoust Soc Am*, 86(3):961–970, 1989.

K. G. Gruters and J. M. Groh. Sounds and beyond: multisensory and other non-auditory signals in the inferior colliculus. *Front Neural Circuits*, 6:96, 2012.

J. J. Guinan, Jr. Olivocochlear efferents: anatomy, physiology, function, and the measurement of efferent effects in humans. *Ear Hear*, 27(6):589–607, 2006.

J. J. Guinan, Jr. Cochlear efferent innervation and function. *Curr Opin Otolaryngol Head Neck Surg*, 18(5):447–453, 2010.

J. J. Guinan, Jr. Physiology of the medial and lateral olivocochlear systems. In D. K. Ryugo, R. R. Fay, and A. Popper, editors, *Auditory and vestibular efferents*, pages 39–81. Springer, New York, 2011.

J. J. Guinan, Jr. Olivocochlear efferent function: issues regarding methods and the interpretation of results. *Front Systems Neurosci*, 8:142, 2014.

J. J. Guinan, Jr and M. L. Gifford. Effects of electrical stimulation of efferent olivocochlear neurons on cat auditory-nerve fibers. III. Tuning curves and thresholds at CF. *Hear Res*, 37(1):29–45, 1988.

S. Handel. *Perceptual coherence: hearing and seeing*. Oxford University Press, New York, 2006.

R. W. Harris and D. W. Swenson. Effects of reverberation and noise on speech recognition by adults with various amounts of sensorineural hearing impairment. *Audiology*, 29(6): 314–321, 1990.

P. Heil. Coding of temporal onset envelope in the auditory system. *Speech Commun*, 41 (1):123–134, 2003.

K. S. Helfer. Binaural cues and consonant perception in reverberation and noise. *J Speech Hear Res*, 37(2):429–438, 1994.

K. S. Helfer and R. A. Huntley. Aging and consonant errors in reverberation and noise. *J Acoust Soc Am*, 90(4):1786–1796, 1991.

K. S. Helfer and L. A. Wilber. Hearing loss, aging, and speech perception in reverberation and noise. *J Speech Hear Res*, 33(2):149–155, 1990.

D. Henderson, X. Zheng, and S. McFadden. Cochlear efferent system - a factor in susceptibility to noise? In D. Henderson, D. Prasher, K. R, R. Salvi, and R. Hamernik, editors, *Noise-induced hearing loss*, pages 127–139. Noise Research Network Publications, London, 2001.

H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J Acoust Soc Am*, 87 (4):1738–1752, 1990.

H. Hermansky. Should recognizers have ears? *Speech Commun.*, 25:3–27, 1998.

H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE T Speech Audi P*, 2 (4):578–589, 1994.

H. Hermansky, J. R. Cohen, and R. M. Stern. Perceptual properties of current speech recognition technology. *Proc IEEE*, 101(9):1968–1985, 2013.

M. J. Hewitt and R. Meddis. An evaluation of eight computer models of mammalian inner hair–cell function. *J Acoust Soc Am*, 90(2 Pt 1):904–917, 1991.

T. Hidaka, Y. Yamada, and T. Nakagawa. A new definition of boundary point between early reflections and late reverberation in room impulse responses. *J Acoust Soc Am*, 122(1):326–332, 2007.

J. R. Hopgood and P. J. W. Rayner. Blind single channel deconvolution using nonstationary signal processing. *IEEE T Speech Audi P*, 11(5):476–488, 2003.

C. Horton, M. D'Zmura, and R. Srinivasan. Suppression of competing speech through entrainment of cortical oscillations. *J Neurophysiol*, 109(12):3082–3093, 2013.

T. Houtgast and H. J. M. Steeneken. The Modulation Transfer Function in room acoustics as a predictor of speech intelligibility. *J Acoust Soc Am*, 54(2):557–557, 1973.

T. Houtgast and H. J. M. Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J Acoust Soc Am*, 77(3): 1069–1077, 1985.

D. C. Howell. *Statistical Methods for Psychology*. Duxbury Press, Boston, MA, 1982.

Y. Hu and K. Kokkinakis. Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners. *J Acoust Soc Am*, 135(1):EL22–EL28, 2014.

C. Hummersone, R. Mason, and T. Brookes. A comparison of computational precedence models for source separation in reverberant environments. *J Audio Eng Soc*, 61(7/8): 508–520, 2013.

T. Irino and R. D. Patterson. A time-domain, level-dependent auditory filter: the Gammachirp. *J Acoust Soc Am*, 101(1):412–419, 1997.

ISO226. Acoustics – normal equal-loudness-level contours. ISO ISO 226:2003(en), International Organization for Standardization, Geneva, Switzerland, 2003.

H. Javed and P. Naylor. Development and evaluation of an improved reverberation decay tail metric as a measure of perceived late reverberation. In *UK Speech Conference*, Edinburgh, UK, June 2014.

S. Jelfs, J. F. Culling, and M. Lavandier. Revision and validation of a binaural model for speech intelligibility in noise. *Hear Res*, 275(1-2):96–104, May 2011.

S. G. Jennings, M. G. Heinz, and E. A. Strickland. Evaluating adaptation and olivocochlear efferent feedback as potential explanations of psychophysical overshoot. *J Assoc Res Otolaryngol*, 12(3):345–360, 2011.

T. L. Johnson and W. Strange. Perceptual constancy of vowels in rapid speech. *J Acoust Soc Am*, 72(6):1761–1770, 1982.

H. G. Jones, K. Koka, J. Thornton, and D. J. Tollin. The sound source distance dependence of the acoustical cues to location and their encoding by neurons in the inferior colliculus: implications for the duplex theory. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, editors, *Basic aspects of hearing: physiology and perception*, pages 273–282. Springer, New York, 2013.

S. Jørgensen and T. Dau. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J Acoust Soc Am*, 130(3):1475–1487, 2011.

P. X. Joris and P. H. Smith. The volley theory and the spherical cell puzzle. *J Neurosci*, 154(1):65–76, 2008.

H. Kallasjoki, J. F. Gemmeke, K. J. Palomäki, A. V. Beeston, and G. J. Brown. Recognition of reverberant speech by missing data imputation and nmf feature enhancement. In *REVERB challenge workshop in conjunction with ICASSP 2014 and HSCMA 2014*, pages 1–8, Florence, Italy, May 2014.

R. Keen and R. L. Freyman. Release and re-buildup of listeners' models of auditory space. *J Acoust Soc Am*, 125(5):3243–3252, 2009.

D. T. Kemp. Stimulated acoustic emissions from within the human auditory system. *J Acoust Soc Am*, 64(5):1386–1391, 1978.

A. Kern, C. Heid, W. H. Steeb, N. Stoop, and R. Stoop. Biophysical parameters modification could overcome essential hearing gaps. *PLoS Comput Biol*, 4(8):e1000161, Aug. 2008.

C. Kim and R. Stern. Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, pages 4101–4104, March 2012.

C. Kim, R. Mason, and T. Brookes. Head movements made by listeners in experimental and real-life listening activities. *J Audio Eng Soc*, 61(6):425–438, 2013.

D. O. Kim. Active and nonlinear cochlear biomechanics and the role of outer-hair-cell subsystem in the mammalian auditory system. *Hear Res*, 22(1–3):105–114, 1986.

D.-S. Kim, S.-Y. Lee, and R. M. Kil. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE T Speech Audi P*, 7(1): 55–69, 1999.

B. E. D. Kingsbury. *Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments*. PhD thesis, University of California, Berkeley, 1998.

B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Improving ASR performance for reverberant speech. In *Robust speech recognition for unknown communication channels (RSR-1997)*, pages 87–90, Pont-à-Mousson, France, Apr. 1997.

B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Commun*, 25(1):117–132, 1998.

K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE T Audio Speech*, 17(4):534–545, 2009.

R. E. Kirk. *Experimental design: procedures for the behavioral sciences*. Brooks-Cole Publishing Company, Belmont, CA, 1968.

D. Klatt. Prediction of perceived phonetic distance from critical-band spectra: a first step. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, pages 1278–1281, Paris, France, May 1982.

M. Kleinschmidt, J. Tchorz, and B. Kollmeier. Combining speech enhancement and auditory feature extraction for robust speech recognition. *Speech Commun*, 34(1–2):75–91, 2001.

A. H. Koening, J. B. Allen, D. A. Berkley, and T. H. Curtis. Determination of masking level differences in a reverberant environment. *J Acoust Soc Am*, 61(5):1374–1376, 1977.

K. D. Kryter. Methods for the calculation and use of the articulation index. *J Acoust Soc Am*, 34(11):1689–1697, 1962.

P. K. Kuhl. Speech-perception in early infancy: perceptual constancy for spectrally dissimilar vowel categories. *J Acoust Soc Am*, 66(6):1668–1679, 1979.

P. K. Kuhl. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Percept Psychophys*, 50(2):93–107, 1991.

S. Kuwada, B. Bishop, and D. O. Kim. Approaches to the study of neural coding of sound source location and sound envelope in real environments. *Front Neural Circuits*, 6:42, 2012.

P. Ladefoged and D. E. Broadbent. Information conveyed by vowels. *J Acoust Soc Am*, 29 (1):98–103, 1957.

P. Lanchantin, P. Bell, M. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, and M. Seigel. Automatic transcription of multi-genre media archives. In *Workshop on Speech, Language and Audio in Multimedia (SLAM)*, pages 26–31, Marseille, France, Aug. 2013.

T. Langhans and H. W. Strube. Speech enhancement by nonlinear multiband envelope filtering. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, pages 156–159, Paris, France, May 1982.

M. Lavandier and J. F. Culling. Prediction of binaural speech intelligibility against noise in rooms. *J Acoust Soc Am*, 127(1):387–399, 2010.

M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin. Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. *J Acoust Soc Am*, 131(1):218–231, 2012.

LCNDEC. Laurent Clerc National Deaf Education Center, Gallaudet University. http://clerccenter.gallaudet.edu, 2009. [Online; accessed 20 April 2009].

K. Lebart, J. M. Boucher, and P. N. Denbigh. A new method based on spectral subtraction for speech dereverberation. *Acustica*, 87(3):359–366, 2001.

C.-y. Lee, J. R. Glass, and O. Ghitza. An efferent-inspired auditory model front-end for speech recognition. In *Proc 12th Conf Int Speech Commun As (INTERSPEECH)*, pages 49–52, Florence, Italy, Aug. 2011.

D. Lee, D. Cabrera, and W. L. Martens. The effect of loudness on the reverberance of music: reverberance prediction using loudness models. *J Acoust Soc Am*, 131(2):1194–1205, 2012.

K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE T Acoust Speech*, 37(11):1641–1648, 1989.

C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Comput Speech Lang*, 9:171–185, 1995.

S. K. Lehtinen, A. P. Raimond, A. Stasiak, A. J. Watkins, and I. M. Winter. No evidence for neural correlates of perceptual compensation for room reverberation in the ventral cochlear nucleus. *Int J Audiol*, 50(10):762, 2011.

X. Li and R. E. Pastore. Perceptual constancy of a global spectral property: spectral slope discrimination. *J Acoust Soc Am*, 98(4):1956–1968, 1995.

H. Liao and M. J. Gales. Issues with uncertainty decoding for noise robust automatic speech recognition. *Speech Commun*, 50(4):265–277, 2008.

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

M. C. Liberman. Auditory-nerve response from cats raised in a low-noise chamber. *J Acoust Soc Am*, 63(2):442–455, 1978.

M. C. Liberman, L. D. Liberman, and S. F. Maison. Efferent feedback slows cochlear aging. *J Neurosci*, 34(13):4599–4607, 2014.

W. Lilaonitkul and J. J. Guinan, Jr. Reflex control of the human inner ear: a half-octave offset in medial efferent feedback that is consistent with an efferent role in the control of masking. *J Neurophysiol*, 101(1):1394–1406, 2009.

R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *J Acoust Soc Am*, 106(4):1633–1654, 1999.

F.-h. Liu, R. M. Stern, X. Huang, and R. Acero. Efficient cepstral normalization for robust speech recognition. In *Proc ARPA Speech and Natural Language Workshop*, pages 69–74, 1993.

T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo. Disentangling preference ratings of concert hall acoustics using subjective sensory profiles. *J Acoust Soc Am*, 132(5):3148–3161, 2012.

S. G. Lomber and S. Malhotra. Double dissociation of 'what' and 'where' processing in auditory cortex. *Nat Neurosci*, 11(5):609–616, 2008.

S. London, C. W. Bishop, and L. M. Miller. Spatial attention modulates the precedence effect. *J Exp Psychol Hum Percept Perform*, 38(6):1371–1379, 2012.

L. Longworth-Reed, E. J. Brandewie, and P. Zahorik. Time-forward speech intelligibility in time-reversed rooms. *J Acoust Soc Am*, 125(1):EL13–EL19, 2009.

E. A. Lopez-Poveda and R. Meddis. A human nonlinear cochlear filterbank. *J Acoust Soc Am*, 110(6):3107–3118, 2001.

E. A. Lopez-Poveda, E. Aguilar, P. T. Johannesen, and A. Eustaquio-Martn. Contralateral efferent regulation of human cochlear tuning: behavioural observations and computer model simulations. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, editors, *Basic aspects of hearing: physiology and perception*, pages 47–54. Springer, New York, 2013.

R. F. Lyon. A computational model of filtering, detection and compression in the cochlea. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, volume 7, pages 1282–1285, Paris, France, May 1982.

R. Maas, A. Thippur, A. Sehr, and W. Kellermann. An uncertainty decoding approach to noise- and reverberation-robust speech recognition. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, pages 7388–7392, May 2013.

G. A. Manley. Cochlear mechanisms from a phylogenetic viewpoint. *Proc Natl Acad Sci USA*, 97(22):11736–11743, 2000.

G. A. Manley and C. Köppl. Phylogenetic development of the cochlea and its innervation. *Curr Opin Neurobiol*, 8(4):468–474, 1998.

G. L. Marshall. An analysis procedure for room acoustics and sound amplification systems based on the early-to-late sound energy ratio. *J Audio Eng Soc*, 44(5):373–381, 1996.

K. D. Martin. Echo suppression in a computational model of the precedence effect. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1997.

J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow. To separate speech. In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, *Machine learning for multimodal interaction*, volume 4892 of *Lecture Notes in*

Computer Science, pages 283–294. Springer Berlin Heidelberg, 2008.

A. McEwan and A. van Schaik. An alternative analog VLSI implementation of the Meddis inner hair cell model. In *Proc Int Symp Circuits Systems (ISCAS)*, volume 4, pages 928–931, 2004.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *J Acoust Soc Am*, 79(3):702–711, 1986.

R. Meddis. Simulation of auditory-neural transduction: further studies. *J Acoust Soc Am*, 83(3):1056–1063, 1988.

R. Meddis. Auditory-nerve first-spike latency and auditory absolute threshold: a computer model. *J Acoust Soc Am*, 119(1):406–417, 2006.

R. Meddis, M. J. Hewitt, and T. M. Shackleton. Implementation details of a computation model of the inner hair cell auditory nerve synapse. *J Acoust Soc Am*, 87(4):1813–1816, 1990.

R. Meddis, L. P. O'Mard, and E. Lopez-Poveda. A computational algorithm for computing nonlinear auditory frequency selectivity. *J Acoust Soc Am*, 109(6):2852–2861, 2001.

R. Meddis, W. Lecluyse, C. M. Tan, M. R. Panda, and R. Ferry. Beyond the audiogram: identifying and modeling patterns of hearing deficits. In *The neurophysiological bases of auditory perception*, pages 631–640. Springer, New York, 2010.

R. Meddis, W. Lecluyse, N. R. Clark, T. Jürgens, C. M. Tan, M. R. Panda, and G. J. Brown. A computer model of the auditory periphery and its application to the study of hearing. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, editors, *Basic aspects of hearing: physiology and perception*, pages 11–19. Springer, New York, 2013.

F. Menzer and B. U. Seeber. Does reverberation perception differ in virtual spaces with unrealistic sound reflections? In *Proc Forum Acust*, Krakow, Poland, Sept. 2014.

D. P. Messing. *Predicting Confusions and Intelligibility of Noisy Speech*. PhD thesis, Massachusetts Institute of Technology, 2007.

D. P. Messing, L. Delhorne, E. Bruckert, L. D. Braida, and O. Ghitza. A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise. *Speech Commun*, 51(8):668–683, 2009.

X. Meynial and O. Vuichard. Objective measure of sound colouration in rooms. *Acta Acustica*, 85(2):101–107, 1999.

G. Miller and P. Nicely. An analysis of perceptual confusions among some english consonants. *J Acoust Soc Am*, 27(2):338–352, 1955.

M. A. Mines, B. F. Hanson, and J. R. Shoup. Frequency of occurrence of phonemes in conversational english. *Lang Speech*, 21(3):221–241, 1978.

S. K. Mishra, M. R. Panda, and C. Herbert. Enhanced auditory temporal gap detection in listeners with musical training. *J Acoust Soc Am*, 136(2):EL173–EL178, 2014.

B. C. J. Moore. Parallels between frequency selectivity measured psychophysically and in cochlear mechanics. *Scand Audiol Suppl*, 25:139–152, 1986.

B. C. J. Moore. *An Introduction to the psychology of hearing*. Academic Press, London, fifth edition, 2004.

R. Moore, E. Adams, P. A. Dagenais, and C. Caffee. Effects of reverberation and filtering on speech rate judgment. *Int J Audiol*, 46(3):154–160, 2007.

D. C. Mountain. Changes in endolymphatic potential and crossed olivocochlear bundle stimulation after cochlear mechanics. *Science*, 210(4465):71–72, 1980.

M. M. Murray and L. Spierer. Multisensory integration: what you see is where you hear. *Curr Biol*, 21(6):R229–R231, 2011.

E. Murugasu and I. Russell. The effect of efferent stimulation on basilar membrane displacement in the basal turn of the guinea pig cochlea. *J Neurosci*, 16(1):325–332, 1996.

A. K. Nábělek and P. A. Dagenais. Vowel errors in noise and in reverberation by hearing-impaired listeners. *J Acoust Soc Am*, 80(3):741–748, 1986.

A. K. Nábělek and J. M. Pickett. Reception of consonants in a classroom as affected by monaural and binaural listening, noise, reverberation, and hearing aids. *J Acoust Soc Am*, 56(2):628–639, 1974.

A. K. Nábělek and L. Robinette. Influence of the precedence effect on word identification by normally hearing and hearing-impaired subjects. *J Acoust Soc Am*, 63(1):187–94, 1978.

A. K. Nábělek and P. Robinson. Monaural and binaural speech perception in reverberation for listeners of various ages. *J Acoust Soc Am*, 71(5):1242–1248, 1982.

A. K. Nábělek, T. Letowski, and F. Tucker. Reverberant overlap- and self-masking in consonant identification. *J Acoust Soc Am*, 86(4):1259–1265, 1989.

A. K. Nábělek, Z. Czyzewski, and L. A. Krishnan. The influence of talker differences on vowel identification by normal-hearing and hearing-impaired listeners. *J Acoust Soc Am*, 92(3):1228–1246, 1992.

A. K. Nábělek, Z. Czyzewski, and H. J. Crowley. Cues for perception of the diphthong /ɑi/ in either noise or reverberation. Part I. Duration of the transition. *J Acoust Soc Am*, 95(5):2681–2693, 1994.

A. K. Nábělek, A. Ovchinnikov, Z. Czyzewski, and H. J. Crowley. Cues for perception of synthetic and natural diphthongs in either noise or reverberation. *J Acoust Soc Am*, 99 (3):1742–1753, 1996.

T. Nakatani, M. Miyoshi, and K. Kinoshita. One microphone blind dereverberation based on quasi-periodicity of speech signals. In *Proc 16th Adv Neur In (NIPS)*, volume 16, pages 1417–1424, 2003.

T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE T Audio Speech*, 18(7): 1717–1731, 2010.

P. A. Naylor, E. A. P. Habets, J. Y.-C. Wen, and N. D. Gaubitch. Models, measurement and evaluation. In P. A. Naylor and N. D. Gaubitch, editors, *Speech dereverberation*, pages 21–56. Springer, London, UK, 2010.

J. B. Nielsen and T. Dau. Revisiting perceptual compensation for effects of reverberation in speech identification. *J Acoust Soc Am*, 128(5):3088–3094, 2010.

Y. Nikulin, K. Miettinen, and M. M. Mäkelä. A parameterized achievement scalarizing function for multiobjective optimization. TUCS Technical Reports 969, Turku Centre for Computer Science, Turku, Finland, 2010.

T. Nishiura, Y. Hirano, Y. Denda, and M. Nakayama. Investigations into early and late reflections on distant-talking speech recognition toward suitable reverberation criteria. In *Proc 8th Conf Int Speech Commun As (INTERSPEECH)*, pages 1082–1085, Antwerp, Belgium, Aug. 2007.

J. Nix and V. Hohmann. Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE T Audio Speech*, 15(3):995–1008, 2007.

K. J. Palomäki, G. J. Brown, and J. Barker. Techniques for handling convolutional distortion with 'missing data' automatic speech recognition. *Speech Commun*, 43(1–2): 123–142, 2004.

R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. Spiral VOS final report part A: the auditory filter bank. Internal Report 2341, MRC Applied Psychology Unit, Cambridge, 1988.

D. Pelegrín-García, B. Smits, J. Brunskog, and C.-H. Jeong. Vocal effort with changing talker-to-listener distance in different acoustic environments. *J Acoust Soc Am*, 129(4): 1981–1990, 2011.

R. Petrick, X. Lu, M. Unoki, M. Akagi, and R. Hoffmann. Robust front end processing for speech recognition in reverberant environments: utilization of speech characteristics. In *Proc 9th Conf Int Speech Commun As (INTERSPEECH)*, pages 658–661, Brisbane, Australia, Sept. 2008.

S. A. Phatak, A. Lovitt, and J. B. Allen. Consonant confusions in white noise. *J Acoust Soc Am*, 124(2):1220–1233, 2008.

J. O. Pickles. *An introduction to the physiology of hearing*. Academic Press, second edition, 1988.

C. Pike, T. Brookes, and R. Mason. Auditory adaptation to loudspeaker and listening room acoustics. In *135th Audio Eng Soc Conv*, number 8971, New York, NY, Oct. 2013.

C. J. Plack. *The sense of hearing*. Routledge, illustrated edition, 2005.

S. F. Poissant, N. A. Whitmal, and R. L. Freyman. Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. *J Acoust Soc Am*, 119(3):1606–1615, 2006.

B. Raj, M. Seltzer, and R. Stern. Reconstruction of missing features for robust speech recognition. *Speech Commun*, 43(4):275–296, 2004.

B. Rakerd and W. M. Hartmann. Localization of sound in rooms, II: the effects of a single reflecting surface. *J Acoust Soc Am*, 78(2):524–533, 1985.

D. K. Reed and S. van de Par. Characterization of a binaural modulation perception. In *Proc Forum Acust*, Krakow, Poland, Sept. 2014.

D. K. Reed, A. Kohlrausch, and S. van de Par. Limited effect of binaural modulation on monaural modulation sensitivity. In *Proc Forum Acust*, Krakow, Poland, Sept. 2014.

M. S. Regnier and J. B. Allen. A method to identify noise-robust perceptual features: application for consonant /t/. *J Acoust Soc Am*, 123(5):2801–2814, 2008.

T. Reichenbach and A. J. Hudspeth. Dual contribution to amplification in the mammalian inner ear. *Phys Rev Lett*, 105:118–102, 2010.

T. Reichenbach and A. J. Hudspeth. The physics of hearing: fluid mechanics and the active process of the inner ear. *Rep Progr Phys*, 77(7):076601, 2014.

R. E. Remez. Critique: auditory form and gestural topology in the perception of speech. *J Acoust Soc Am*, 99(3):1695–1698, 1996.

W. S. Rhode. Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. *J Acoust Soc Am*, 49(4):1218–1231, 1971.

L. Robles and M. A. Ruggero. Mechanics of the mammalian cochlea. *Physiol Rev*, 81(3): 1305–1352, 2001.

J. E. Rose, J. E. Hind, D. J. Anderson, and J. F. Brugge. Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey. *J Neurophysiol*, 34(4): 685–699, 1971.

J. T. Rubinstein. How cochlear implants encode speech. *Curr Opin Otolaryngol Head Neck Surg*, 12(5):444–448, 2004.

A. Rupp, A. Spachmann, A. Dettlaff, and R. D. Patterson. Cortical activity associated with the perception of temporal asymmetry in ramped and damped noises. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, editors, *Basic aspects of hearing: physiology and perception*, pages 427–433. Springer, New York, 2013.

I. J. Russell and E. Murugasu. Medial efferent inhibition suppresses basilar membrane response to near characteristic frequency tones of moderate to high intensities. *J Acoust Soc Am*, 102(3):1734–1738, 1997.

D. K. Ryugo. Introduction to efferent systems. In D. K. Ryugo, R. R. Fay, and A. Popper, editors, *Auditory and vestibular efferents*, pages 1–16. Springer, New York, 2011.

K. Saberi and D. R. Perrott. Cognitive restoration of reversed speech. *Nature*, 398:760, 1999.

W. C. Sabine. *Collected papers on acoustics*. Harvard University Press, Cambridge, MA, 1922.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statist Sci*, 4(4):409–423, 1989.

E. Sagi and M. A. Svirsky. Information transfer analysis: a first look at estimation bias. *J Acoust Soc Am*, 123(5):2848–2857, 2008.

M. Sayles and I. M. Winter. Reverberation challenges the temporal representation of the pitch of complex sounds. *Neuron*, 58(5):789–801, 2008.

M. Sayles, C. Füllgrabe, and I. M. Winter. Neurometric amplitude-modulation detection threshold in the guinea-pig ventral cochlear nucleus. *J Physiol*, 591(13):3401–3419, 2013.

M. Sayles, A. Stasiak, and I. M. Winter. Reverberation impairs brainstem temporal representations of voiced vowel sounds: challenging 'periodicity-tagged' segregation of competing speech in rooms. *Front Systems Neurosci*, 8:248, 2015.

C. E. Schroeder, P. Lakatos, Y. Kajikawa, S. Partan, and A. Puce. Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci*, 12(3):106–113, 2008.

M. R. Schroeder. New method of measuring reverberation time. *J Acoust Soc Am*, 37(3): 309–412, 1965.

A. Sehr, M. Gardill, and W. Kellermann. Adapting hmms of distant-talking asr systems using feature-domain reverberation models. In *Proc Eur Signal Process Conf*, pages 540–543, Glasgow, Aug. 2009.

A. Sehr, R. Maas, and K. W. Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition. *IEEE T Audio Speech*, 18(7): 1676–1691, 2010.

S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *J Phonetics*, 16(1):55–76, 1988.

R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.

B. Shinn-Cunningham, D. R. Ruggles, and H. Bharadwaj. How early aging and environment interact in everyday listening: from brainstem to behavior through modeling. In

B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, editors, *Basic aspects of hearing: physiology and perception*, pages 501–510. Springer, New York, 2013.

B. G. Shinn-Cunningham. Learning reverberation: considerations for spatial auditory displays. In *Proc Int Conf Auditory Display (ICAD)*, pages 126–134, Atlanta, GA, Apr. 2000.

J. H. Siegel and D. O. Kim. Efferent neural control of cochlear mechanics? Olivocochlear bundle stimulation affects cochlear biomechanical nonlinearity. *Hear Res*, 6(2):171–182, 1982.

A. Smith. On the use of the relative information transmitted (RIT) measure for the assessment of performance in the evaluation of automated speech recognition (ASR) devices. In *Proc 3rd Australasian Speech Science and Technology Association Conference (ASSTA)*, pages 368–373, Melbourne, Australia, Nov. 1990.

Z. M. Smith, B. Delgutte, and A. J. Oxenham. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876):87–90, 2002.

H. Spoendlin. Innervation patterns in the organ of Corti of the cat. *Acta Oto-Laryngol*, 67 (2-6):239–254, 1969.

T. S. Sridhar, M. C. Brown, and W. F. Sewell. Unique postsynaptic signaling at the hair cell efferent synapse permits calcium to evoke changes on two time scales. *J Neurosci*, 17(1):428–437, 1997.

N. K. Srinivasan and P. Zahorik. The effect of semantic context on speech intelligibility in reverberant rooms. *Proc Meet Acoust*, 12:060001, 2011.

N. K. Srinivasan and P. Zahorik. Prior listening exposure to a reverberant room improves open-set intelligibility of high-variability sentences. *J Acoust Soc Am*, 133(1):EL33–EL39, 2013.

N. K. Srinivasan and P. Zahorik. Enhancement of speech intelligibility in reverberant rooms: role of amplitude envelope and temporal fine structure. *J Acoust Soc Am*, 135 (6):EL239–EL245, 2014.

G. C. Stecker and E. R. Hafter. An effect of temporal asymmetry on loudness. *J Acoust Soc Am*, 107(6):3358–3368, 2000.

H. J. M. Steeneken and T. Houtgast. Physical method for measuring speech-transmission quality. *J Acoust Soc Am*, 67(1):318–326, 1980.

R. M. Stern. Applying physiologically-motivated models of auditory processing to automatic speech recognition. In *3rd Int Symp Auditory Audiol Res (ISAAR)*, Nyborg, Denmark, Aug. 2011.

R. M. Stern and N. Morgan. Hearing is believing: biologically inspired methods for robust automatic speech recognition. *IEEE Signal Proc Mag*, 29(6):34–43, 2012.

S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *J Acoust Soc Am*, 8(3):185–190, 1937.

W. Strange, J. J. Jenkins, and T. L. Johnson. Dynamic specification of coarticulated vowels. *J Acoust Soc Am*, 74(3):695–705, 1983.

E. A. Strickland and L. A. Krishnan. The temporal effect in listeners with mild to moderate cochlear hearing impairment. *J Acoust Soc Am*, 118(5):3211–3217, 2005.

Q. Summerfield. Articulatory rate and perceptual constancy in phonetic perception. *J Exp Psychol Human*, 7(5):1074–1095, 1981.

C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis. A revised model of the inner-hair cell and auditory-nerve complex. *J Acoust Soc Am*, 111(5):2178–2188, 2002.

T. Takiguchi, M. Nishimura, and Y. Ariki. Acoustic model adaptation using first-order linear prediction for reverberant speech. *IEICE Trans Inform Syst*, E89-D(3):908–914, 2006.

J. Tchorz and B. Kollmeier. A model of auditory perception as front end for automatic speech recognition. *J Acoust Soc Am*, 106(4):2040–2050, 1999.

S. Thomas, S. Ganapathy, and H. Hermansky. Recognition of reverberant speech using frequency domain linear prediction. *IEEE Signal Proc Let*, 15:681–684, 2008.

A. Tsilfidis and J. Mourjopoulos. Blind single-channel suppression of late reverberation based on perceptual reverberation modeling. *J Acoust Soc Am*, 129(3):1439–1451, 2011.

D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis. Speech enhancement based on audible noise suppression. *EEE T Speech Audi P*, 5(4):497–514, 1997.

B. Tuller, P. Case, M. Ding, and J. A. S. Kelso. The nonlinear dynamics of speech categorization. *J Exp Psychol Human*, 20(1):3–16, 1994.

K. Ueno, N. Kopčo, and B. G. Shinn-Cunningham. Calibration of speech perception to room reverberation. In *Proc Forum Acust*, Budapest, Hungary, Aug. 2005.

N. Upadhyay and A. Karmakar. An improved multi-band spectral subtraction algorithm for enhancing speech in various noise environments. *Procedia Engineering*, 64:312–321, 2013. Int Conf Design Manufacturing.

P. van Dijk, H. P. Wit, and J. M. Segenhout. Spontaneous otoacoustic emissions in the European edible frog (Rana esculenta): spectral details and temperature dependence. *Hear Res*, 42(2–3):273–282, 1989.

J. van Dorp Schuitman, D. de Vries, and A. Lindau. Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model. *J Acoust Soc Am*, 133(3):1572–1585, 2013.

B. D. van Veen and K. M. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Mag*, 5(2):4–24, 1988.

G. von Békésy. The variation of phase along the basilar membrane with sinusoidal vibrations. *J Acoust Soc Am*, 19(3):452–460, 1947.

H. Wallach, E. B. Newman, and M. R. Rosenzweig. The precedence effect in sound localization. *Am J Psychol*, 62(3):315–336, 1949.

A. J. Watkins. Perceptual compensation for effects of reverberation in speech identification. *J Acoust Soc Am*, 118(1):249–262, 2005a.

A. J. Watkins. Perceptual compensation for effects of echo and of reverberation on speech identification. *Acta Acust United Ac*, 91(5):892–901, 2005b.

A. J. Watkins and S. J. Makin. Perceptual compensation for reverberation in speech identification: effects of single-band, multiple-band and wideband noise contexts. *Acta Acust United Ac*, 93(3):403–410, 2007a.

A. J. Watkins and S. J. Makin. Steady-spectrum contexts and perceptual compensation for reverberation in speech identification. *J Acoust Soc Am*, 121(1):257–266, 2007b.

A. J. Watkins and S. J. Makin. Perceptual compensation for reverberation: effects of 'noise-like' and 'tonal' contexts. In B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. L. Verhey, editors, *Hearing–from sensory processing to perception*, pages 533–540. Springer, Berlin, 2007c.

A. J. Watkins and A. P. Raimond. Perceptual compensation when isolated test words are heard in room reverberation. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, editors, *Basic aspects of hearing: physiology and perception*, pages 193–201. Springer, New York, 2013.

A. J. Watkins, S. J. Makin, and A. P. Raimond. Constancy in the perception of speech when the level of room-reflections varies. In J. M. Buchholz, T. Dau, J. Dalsgaard, and T. Poulsen, editors, *Binaural processing and spatial hearing. ISAAR—International Symposium on Auditory and Audiological Research*, pages 371–380. The Danavox Jubilee Foundation, Ballerup, Denmark, 2010a.

A. J. Watkins, A. P. Raimond, and S. J. Makin. Room reflections and constancy in speech-like sounds: within-band effects. In *The neurophysiological bases of auditory perception*, pages 439–447. Springer, New York, 2010b.

A. J. Watkins, A. P. Raimond, and S. J. Makin. Temporal-envelope constancy of speech in rooms and the perceptual weighting of frequency bands. *J Acoust Soc Am*, 130(5): 2777–2788, 2011.

M. A. Webster, M. A. Georgeson, and S. M. Webster. Neural adjustments to image blur. *Nat Neurosci*, 5(9):839–840, Sept. 2002.

M. Weintraub. *A theory and computational model of auditory monaural sound separation*. PhD thesis, Stanford University, California, 1985.

C. J. C. Weisz, E. Glowatzki, and P. A. Fuchs. Excitability of Type II cochlear afferents. *J Neurosci*, 34(6):2365–2373, 2014.

T. Weller, J. M. Buchholz, and V. Best. Modelling binaural detection of speech stimuli in complex reverberant environments. In *Proc Forum Acust*, Krakow, Poland, Sept. 2014.

L. A. Westerman. *Adaptation and recovery of auditory-nerve responses*. PhD thesis, Syracuse University, 1985.

A. Westermann, J. M. Buchholz, and T. Dau. Binaural dereverberation based on interaural coherence histograms. *J Acoust Soc Am*, 133(5):2767, 2013.

F. L. Wightman and D. J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J Acoust Soc Am*, 91(3):1648–1661, 1992.

J. Wright. Articulation index. Technical report, Linguistic Data Consortium, Philadelphia, 2005.

M. Wu and D. L. Wang. A one-microphone algorithm for reverberant speech enhancement. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, volume 1, pages 844–847, Hong Kong, Apr. 2003.

Z. Yang and D. Purves. The statistical structure of natural light patterns determines perceived light intensity. *Proc Natl Acad Sci USA*, 101(23):8745–8750, 2004.

G. K. Yates and D. L. Kirk. Cochlear electrically evoked emissions modulated by mechanical transduction channels. *J Neurosci*, 18(6):1996–2003, 1998.

B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendano, and H. Hermansky. Enhancement of reverberant speech using LP residual. In *IEEE Int Conf Acoustics Speech Signal Process (ICASSP)*, volume 1, pages 405–408, Seattle, WA, 1998.

T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. *IEEE Signal Process Mag*, 29(6):114–126, 2012.

E. D. Young and M. B. Sachs. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *J Acoust Soc*

Am, 66(5):1381–1403, 1979.

S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK book*. Cambridge University Engineering Department, revised for htk version 3.4 march 2009 edition, 2009.

P. Zahorik. Assessing auditory distance perception using virtual acoustics. *J Acoust Soc Am*, 111(4):1832–1846, 2002.

P. Zahorik. Perceptually relevant parameters for virtual listening simulation of small room acoustics. *J Acoust Soc Am*, 126(2):776–791, 2009.

P. Zahorik and P. W. Anderson. Amplitude modulation detection by human listeners in reverberant sound fields: effects of prior listening exposure. *Proc Meet Acoust*, 19(1): 050139, 2013.

P. Zahorik and E. Brandewie. Perceptual adaptation to room acoustics and effects on speech intelligibility in hearing-impaired populations. In *Proc Forum Acust*, pages 2167–2172, Aalborg, Denmark, June 2011.

P. Zahorik and F. L. Wightman. Loudness constancy with varying sound source distance. *Nat Neurosci*, 4(1):78–83, 2001.

P. Zahorik, E. J. Brandewie, and V. P. Sivonen. Spatial hearing in reverberant rooms and effects of prior listening exposure. In *International Workshop on the Principles and Applications of Spatial Hearing (IWPASH)*, Zao, Miyagi, Japan, Nov. 2009.

P. Zahorik, D. O. Kim, S. Kuwada, P. W. Anderson, E. J. Brandewie, and N. K. Srinivasan. Amplitude modulation detection by human listeners in sound fields. *Proc Meet Acoust*, 12:050005, 2011.

P. Zahorik, D. O. Kim, S. Kuwada, P. W. Anderson, E. J. Brandewie, R. Collecchia, and N. K. Srinivasan. Amplitude modulation detection by human listeners in reverberant sound fields: carrier bandwidth effects and binaural versus monaural comparison. *Proc Meet Acoust*, 15:050002, 2012.

X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney. A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression. *J Acoust Soc Am*, 109(2):648–670, 2001.

W. Zhao and S. Dhar. Fast and slow effects of medial olivocochlear efferent activity in humans. *PLoS ONE*, 6(4):e18725, 2011.

W. Zhao and S. Dhar. Frequency tuning of the contralateral medial olivocochlear reflex in humans. *J Neurophysiol*, 108(1):25–30, 2012.

M. S. A. Zilany and I. C. Bruce. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *J Acoust Soc Am*, 120(3): 1446–1466, 2006.

M. S. A. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney. A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. *J Acoust Soc Am*, 126(5):2390–2412, 2009.

M. S. A. Zilany, I. C. Bruce, , and L. H. Carney. Updated parameters and expanded simulation options for a model of the auditory periphery. *J Acoust Soc Am*, 135(1): 283–286, 2014.

P. M. Zurek. Binaural advantages and directional effects in speech intelligibility. In G. A. Studebaker and I. Hockberg, editors, *Acoustical factors affecting hearing aid performance*. Allyn and Bacon, Boston, MA, 1993.

E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *J Acoust Soc Am*, 33(2):3637–3642, 1961.

J. J. Zwislocki-Moscicki. *Theorie der Schneckenmechanik*. PhD thesis, Technischen Hochschule in Zürich, 1948.