Factors affecting the perception of noise-vocoded speech: stimulus properties and listener variability

Carolyn McGettigan

Thesis submitted for the degree of Doctor of Philosophy

University College London, November 2007

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This thesis presents an investigation of two general factors affecting speech perception in normal-hearing adults. Two sets of experiments are described, in which speakers of English are presented with degraded (noise-vocoded) speech. The first set of studies investigates the importance of linguistic rhythm as a cue for perceptual adaptation to noise-vocoded sentences. Results indicate that the presence of native English rhythmic patterns benefits speech recognition and adaptation, but not when higher-level linguistic information is absent (i.e. when the sentences are in a foreign language). It is proposed that rhythm may help in the perceptual encoding of degraded speech in phonological working memory. Experiments in this strand also present evidence against a critical role for indexical characteristics of the speaker in the adaptation process.

The second set of studies concerns the issue of individual differences in speech perception. A psychometric curve-fitting approach is selected as the preferred method of quantifying variability in noise-vocoded sentence recognition. Measures of working memory and verbal IQ are identified as candidate correlates of performance with noise-vocoded sentences. When the listener is exposed to noise-vocoded stimuli from different linguistic categories (consonants and vowels, isolated words, sentences), there is evidence for the interplay of two initial listening 'modes' in response to the degraded speech signal, representing 'top-down' cognitive-linguistic processing and 'bottom-up' acoustic-phonetic analysis. Detailed analysis of segment recognition presents a perceptual role for temporal information across all the linguistic categories, and suggests that performance could be improved through training regimes that direct attention to the most informative acoustic properties of the stimulus. Across several experiments, the results also demonstrate long-term aspects of perceptual learning.

In sum, this thesis demonstrates that consideration of both stimulus-based and listener-based factors forms a promising approach to the characterization of speech perception processes in the healthy adult listener.

Acknowledgements

I feel extremely privileged to have had such an enjoyable time in my PhD studies, and there are several people I would like to thank for making these some of the best years of my life. First, I'd like to thank my supervisors, Professors Sophie Scott and Stuart Rosen. Sophie has been fantastically generous and unshakingly supportive of my ideas, and has done so much to encourage my development as a scientist. I would like to thank her for giving me the confidence and self-belief to get this far, and crucially for sharing my passions for speech, fashion jewellery and karaoke. Thanks to Stuart for all of his technical and mathematical advice, his generous giving of time, and his consistent optimism and good humour. I thank all of my colleagues, past and present, in the Speech Communication group at the ICN for being great friends and inspiring scientists. Thanks in particular to Frank Eisner, Jonas Obleser and Disa Sauter for technical advice and comments on earlier drafts of this thesis. Volker Dellwo deserves a special thanks for getting me on the right track with the linguistic rhythm measurements, and for helpful discussions along the way.

I would never have got to this point without the immeasurable love and support of my parents, Teddy and Pauline McGettigan, who have always believed in my abilities, even at the most stressful times. Thanks to Conor and Hannah for being the best housemates (and subjects) a PhD student could wish for - I miss you - and to Diarmuid for being a great friend throughout. Thanks to Clare for convincing me to pursue this PhD - my life would have been considerably less interesting had I gone against your advice. Finally, thanks to Rebecca for all your love and energy, and for bringing me so much happiness - I don't know how lucky I am.

I'd like to dedicate this thesis to my grandparents, Robert and Jane Andrews.

Contents

List of Tables

List of Figures

Chapter 1

Factors affecting the perception of noise-vocoded speech

Abstract

Speech distortion is a useful tool with which to challenge the speech perception system and investigate its component processes. This chapter reviews the use of noise-vocoded speech both as a simulation of a cochlear implant and as a means of studying speech recognition and perceptual adaptation in the normal-hearing population. A two-strand approach to further research is outlined, suggesting speech rhythm and individual differences as topics for future experiments.

1.1 Introduction

1.1.1 Speech perception and the redundancy problem

One of the most striking phenomena described in the literature on speech perception is the robustness of speech recognition in the face of considerable variability in the acoustic speech signal. The acoustic signal for even the simplest of utterances can vary dramatically across speakers due to differences in characteristics such as accent, gender, age, and vocal tract length. For example, Peterson and Barney (1952) found great variation across speakers in the frequencies of formants corresponding to the same vowel category. Other authors (B. Smith & Kenney, 1998; B. Smith, Kenney, & Hussain, 1996; Munson & Babel, 2005; Munson, 2004) have identified differences in speech production with age. There is also the potential for great variability within the utterances of the same speaker - the acoustic realization of speech stimuli can be considerably changed with an alteration in the speaker's rate of speech (J. Miller, Green, & Reeves, 1986; J. Miller & Baer, 1983; Summerfield, 1981; Port, 1979), emotional state (Laukka, Juslin, & Bresin, 2005; Banziger & Scherer, 2005), mental health (Cannizzaro, Harel, Reilly, Chappell, & Snyder, 2004) and even his/her level of intoxication (Pisoni, 1991). There are also a number of changes to the speech stimulus that are dependent on whether speech is conversational or formal (e.g. read aloud; Krause and Braida (2004); Liu, Bradlow, and Zeng (2004)). Liu et al. measured speech from talkers when they were asked to speak 'clearly' and when they spoke in a more conversational style. The authors found that clear speech had a slower overall rate and higher temporal amplitude modulations than conversational speech. Furthermore, they observed that clear speech was more intelligible than conversational speech. Krause and Braida (2004) carried out a comparison of clear and conversational speech when both were produced at normal speaking rates. Their analysis looked at global, phonological and phonetic levels in each speech type. At a global level, they found that increased energy in the 1000-3000-Hz range of long-term spectra and increased modulation depth of low frequency modulations of the intensity envelope were associated with clear speech. At lower levels of analysis, they found that the frequency of stop burst releases, the VOT (voice onset time) after word-initial voiceless stop consonants and short-term vowel spectra were amongst the phonetic and phonological properties which differed between clear and conversational utterances.

At the level of basic sounds in the language, the spectro-temporal description of a single consonant or vowel within an utterance is dependent not only on the factors listed above, but also on the sounds that come before and after it. For example, the first sound in 'heat' has a different energy distribution to the corresponding sound in 'hot', as the articulators are already preparing

for the following vowel sound. However, both of these word-initial sounds are recognised as /h/ by the listener. 'Coarticulation' of sounds, or 'sloppy speech' can lead to a mismatch between the intended message and the actual sounds articulated. For example, in English, in a phrase such as 'sweet girl', the alveolar plosive /t/ at the end of 'sweet' will often be velarised, leading to the phrase being realised more like 'sweek girl'. Despite such effects, the aid of context means that the listener is usually not confused by such speech patterns. Furthermore, there is evidence to suggest that the speech signal is not completely lacking in consistency. Moore (1997) reviews a number of studies that have identified relatively invariant acoustic cues some to phoneme identity. For example, stop consonants can be uniquely defined from a combination of features; rapidity of spectral change, abruptness of amplitude change, voicing cues and the relationship of the consonant's spectral shape to that of the following sound (Blumstein & Stevens, 1979; Stevens, 1980; Kewley-Port, Pisoni, & Studdert-Kennedy, 1983).

It is clear, from the normal listener's ability to comprehend speech with ease in the face of extensive variability, that there is a great deal of information in the speech stimulus that is redundant for the process of extracting meaning. Kluender, Coady, and Kiefte (2003) account for the robustness of speech perception by proposing that perceptual systems undergo "constant calibration to maximize sensitivity to change" (p. 59). The relative ease of speech perception in everyday situations is obviously advantageous to the listener, but from the point of view of the cognitive psychologist or psychoacoustician, it makes it very difficult to uncover the processes involved in perception. Historically, a popular method of investigating speech processing has been to put the system under time pressure, with tasks such as speeded lexical decision and speech shadowing (Marslen-Wilson & Warren, 1994; Marslen-wilson, 1985; Swinney, 1979). Placing subjects under a time limit to respond to stimuli increases difficulty, and produces a dependent measure of this difficulty in the form of reaction times. An alternative method of placing pressure on the speech recognition system is to remove some of the redundancy from the speech signal by distortion or degradation, and observe the effects on the listener's recognition performance (Remez, Rubin, Pisoni, & Carrell, 1981; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Altmann & Young, 1993). This is the topic of the next section.

1.1.2 Distorted speech as an investigative tool

The use of a distorted or degraded speech signal in a speech recognition task holds a number of advantages in the investigation of the human speech perception system. First, and critically, it places the system under sufficient pressure to make speech recognition difficult and partially

override the effect of the redundancy that is obviously present in the speech signal. For example, numerous studies employing artificial temporal compression of the speech stimulus have shown that intelligibility decreases as compression rate is increased (Dehaan, 1982; Foulke & Sticht, 1969; Heiman, Leo, Leighbody, & Bowler, 1986). Second, with a synthetic manipulation of the speech signal, investigators can attempt to identify the critical elements necessary for speech perception by removing or altering parts of the acoustic signal and assessing the relative impact on speech recognition. Remez et al. (1981) presented listeners with sentence stimuli composed only of sinusoid signals tracking the centre frequencies of the three main formants in the original speech signal. Listeners who were instructed to transcribe these distorted sentences were able to perform the task, despite rating the vocal source as 'unnatural'. Shannon et al. (1995) found relatively robust speech recognition using stimuli which retained amplitude envelope and broad spectral cues but removed virtually all fine spectral detail. Altmann and Young (1993) reported that they found 'virtually no loss of intelligibility' when spoken sentences were compressed to 50% of their original duration.

A third important outcome of studies using distorted speech is that they inform on the plasticity of the speech perception system i.e. the capacity of the system to adapt to an unfamiliar stimulus over a period of exposure to that stimulus. Many studies using distorted speech stimuli claim to have demonstrated 'perceptual learning' of the unfamiliar distortion. Schwab, Nusbaum, and Pisoni (1985) found that listeners' comprehension of synthetic speech improved significantly over a 10-day training period. Davis, Johnsrude, Hervais-Adelman, Taylor, and McGettigan (2005) found rapid improvement in recognition scores in normal listeners exposed to spoken sentences containing primarily temporal cues - listeners' recognition scores went from 0% to 70% words correct over exposure to 30 distorted sentences. Voor and Miller (1965) showed a significant increase in comprehension over the first 8 to 10 minutes of listening to speech that had been temporally compressed (or 'time-compressed'). Pallier, Sebastian-Galles, Dupoux, Christophe, and Mehler (1998) even showed that training with time-compressed sentences in one language resulted in perceptual learning that could be transferred to perception of time-compressed sentences in another language.

A challenging issue in interpreting the above studies is that of how to define 'perceptual learning'. One might expect that, in order to establish whether true learning has taken place rather than some attentional 're-tuning' process, there should be criteria in place in the literature with regard to the time-course and permanence of these perceptual changes. In his review of perceptual learning, Goldstone (1998) defines it as a process that "involves relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment and

are caused by this environment" (p. 586). Altmann and Young (1993) demonstrated long-term aspects of adaptation to time-compressed speech. They gave their subjects a second recognition test with time-compressed speech 12 months after their original testing session. At the time of the second test, the authors found better recognition performance in listeners with previous experience of the distortion than that achieved by naïve listeners tested in the same session. Dupoux and Green (1997) differentiated between short- and long-term aspects of adaptation to time-compressed speech by showing that a sudden change in compression rate or speaker can cause a small decrease in speech recognition performance, but not to naïve levels, indicating the retention of more global, long-term representations of the stimulus. Thus, it seems that these studies support the existence of learning with exposure to time-compressed speech. However, how does one deal with inconsistencies with regard to rate of adaptation across different studies? Davis et al. (2005) report rapid adaptation to their speech stimuli within minutes of exposure, whereas Schwab et al. (1985) observe improvements in perception of synthetic speech over a time-course of several days. This variability seems to be accepted in the literature; Goldstone (1998) acknowledges that there are numerous mechanisms of perceptual adaptation, while Atienza, Cantero, and Dominguez-Marin (2002) explain that differences in the time course over which perceptual learning takes place "can be explained by neural changes evolving within different temporal windows" (p. 138). Atienza et al. contrast receptive field modulation of cortical neurons, as the cause of rapid perceptual change, with cortical reorganization supporting slower behavioural changes. In summary, whilst it appears that the perceptual changes that have been reported with exposure to distorted speech are indeed evidence of perceptual learning, it is important to bear in mind the possibility that the time course and mechanistic componentry of perceptual change may differ across different stimulus types and training regimes.

The investigation of perceptual adaptation bears particular relevance to the real-world phenomenon of 'tuning in' to an unfamiliar accent, and this more naturalistic type of stimulus has been used in a number of recent studies (Clarke & Garrett, 2004; Weill, 2001). Clarke and Garrett (2004) showed that adaptation to a foreign speaker can be very rapid. They exposed English speakers to Spanish- and Chinese-accented speech, and found that, in a speeded probe word recognition task, processing speed is initially slowed for accented speech. However, this deficit diminished within 1 minute of experience with the foreign accent, and under some circumstances listeners had adapted to the accented speech within only two to four sentences of exposure. However, Evans and Iverson (2004) demonstrated some indications of more slowly evolving aspects of perceptual change in response to an accent. They showed that young adults from the north of England who had studied and lived for at least 1 year in southern England (London) showed some adjustment of perceptual vowel categorization when listening to the Standard Southern British English (SSBE),

the accent to which they were heavily exposed at university. Such adjustments were not shown by a control group of adults who had remained in northern England, who continued to use their native vowel categories.

Sebastian-Galles, Dupoux, Costa, and Mehler (2000) strongly advocate using adaptation techniques such as those used by Davis et al. (2005), Altmann and Young (1993) and Pallier et al. (1998) to investigate speech processing. In these studies, listeners experienced a 'training phase' - a period of exposure to distorted speech in which either passive or active listening was required - followed by a test phase demanding performance of an active listening task. By manipulating the training materials (e.g. in terms of linguistic content or acoustic structure) to create different training conditions while keeping the test phase constant, one can interpret group differences in test phase scores in terms of the relative adaptive usefulness of the materials used in training. Like Pallier et al. (1998), Sebastian-Galles et al. (2000) used a cross-linguistic approach of training listeners with time-compressed sentences in one language and testing with sentences in another language. Through investigation of different training-test language combinations, Sebastian-Galles et al. were able to tease apart lexical from phonological factors in the transfer of adaptation from one language to another. They write that 'such a discovery may also help us find out why subjects have trouble with the perception of known foreign languages and with the perception of their native language when it spoken with a foreign accent'.

1.2 Noise-Vocoded Speech - a cochlear implant simulation and an investigative tool

Noise-vocoding is a method of distorting speech that preserves temporal cues while greatly reducing spectral detail (Scott, Blank, Rosen, & Wise, 2000). This is achieved by dividing the original speech signal into frequency bands, extracting the amplitude envelope from each band and reapplying it to band-limited noise. The resulting stimulus can be described as sounding "like a harsh whisper" (Scott et al., 2000, p.2401) with a weak sense of pitch (Faulkner, Rosen, & Smith, 2000). The number of bands into which the speech signal is divided can be varied to change the severity of the distortion and hence the difficulty of speech recognition. Recognition of noise-vocoded speech improves logarithmically with the increase in the number of bands (Davis & Johnsrude, 2003; Shannon et al., 1995), as each additional band in a noise-vocoded stimulus contributes greater spectral resolution. Figure 1.1 shows the general scheme by which speech is converted into a noise-vocoded stimulus.

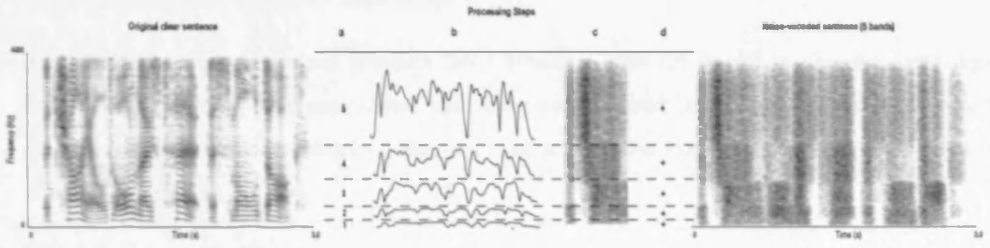


Figure 1.1: Creating noise-vocoded stimuli from clear speech. **Step a**: Clear speech is divided into several contiguous frequency bands, whose bandwidths correspond to equal distances on the basilar membrane (calculated using an equation by Greenwood(1990)). **Step b**: The amplitude envelope is extracted from each band and used to modulate wide-band noise within matching band limits (**Step c**). Finally, the ampitude-modulated noise-bands are combined to form the noise-vocoded speech stimulus (**Step d**). This figure has been adapted from Davis et al. (2005)

Research using noise-vocoded speech can be seen as two-pronged. First, in preserving amplitude envelope information while discarding fine spectral detail, the distortion simulates the transduction of the speech signal by a cochlear implant hearing device (Shannon et al., 1995). A cochlear implant is a hearing aid that converts acoustic sound energy into electrical stimuli to be transmitted to the auditory nerve (Rubenstein, 2004). A typical multichannel processor operates by separating the incoming sound into a number of channels via a set of bandpass filters. In an analogue processor, the dynamic range in each channel is reduced by compression before the signal is passed on to an array of electrodes in the cochlea. In a pulsatile processor, the envelope is extracted from the output of each bandpass filter, then compressed and used to modulate a current pulse before transfer of the signal to the auditory nerve via an electrode. Moore (2003) has reviewed the extent to which cochlear implants can replicate how sounds are 'coded' in the normal auditory system. In his conclusions, he identifies a number of weaknesses in the coding abilities of cochlear implants:

1. Cochlear implants bypass the fast-acting compression mechanism of the healthy cochlea that enables the large dynamic range of the normal ear.

2. Implants effectively provide fewer frequency channels than the normal ear, thus limiting the extraction of fine spectral information.

3. The normal ear is capable of coincidence detection across channels to code sound level, spectral shape and pitch. Cochlear implants are currently incapable of this.

4. Pitch perception is limited with cochlear implants because, unlike the healthy ear, the implant

cannot resolve low-frequency harmonics.

5. The healthy auditory system benefits from binaural cues for sound localization and signal detection in noise. Moore anticipates that bilateral cochlear implantation may be helpful for some cases of speech detection in noise, but that it is unlikely to return detection of interaural time differences to normal.

6. The healthy auditory system makes important use of cues from deflection of sounds from the pinnae of the out ear for sound localization. Cochlear implant microphones are located behind the ear so make no use of these cues. Furthermore, pinna cues require resolution of spectral detail at high frequencies, where cochlear implant performance is poor.

7. Some researchers claim that stochasticity and independent firing of individual neurons in the auditory nerve may be important, perhaps to enable detection of small changes in weak signals. Moore comments that cochlear implants may be able to achieve this independence through very high pulse rates for pulsatile processors, or by addition of low-level noise for analogues processors.

Thus, modelling of cochlear implants in simulations such as noise-vocoded speech with normal listeners can be a useful experimental tool in devising modifications to cochlear implant processing systems, with the aim of maximising their coding efficacy for use in everyday listening situations.

The second strand of research using noise-vocoded speech is to exploit its relative difficulty as a perceptual stimulus in the normal listener to investigate the components of the speech perception process in the healthy/intact auditory system.

1.2.1 Cochlear Implants - limits on speech recognition and musical enjoyment

A number of studies have described limitations of speech recognition ability with cochlear implants and their simulations. Valimaa, Maatta, Lopponen, and Sorri (2002) describes the difficulty that cochlear implant users have with certain consonants. Speech pitch is not well preserved in cochlear implants and their simulations. In tonal languages, like Chinese, lexical tone is phonemic and thus affects word meaning. Ciocca and colleagues (Ciocca, Francis, Aisha, & Wong, 2002) describe the difficulty that cochlear implant wearers experience in identifying lexical tones.

An important note to make at this point is that, while some authors have shown that cochlear

implant users can perform as well with their implant as normal listeners exposed to a simulation (Dorman & Loizou, 1998), others have shown a poorer performance by implant wearers relative to normal listeners. Fishman, Shannon, and Slattery (1997) observed no increase in average sentence and consonant recognition tests in cochlear implant users when the number of channels was increased from 4 to 20, suggesting that the cochlear implant listeners are unable to make full use of the spectral information presented through 20 channels. Friesen and colleagues (Friesen, Shannon, Baskent, & Wang, 2001) write that most cochlear implant-wearing participants "are not able to fully utilize the spectral information provided by the number of electrodes used in their implant" (p. 1150). Wei, Cao, and Zeng (2004) observed much lower performance scores in their cochlear implant users compared with normal listeners hearing a simulation. The authors speculate that 'neither temporal nor spectral cues have been adequately and appropriately extracted and encoded' (p.87) in the cochlear implants they tested. This phenomenon of cochlear implant listeners 'under-performing' compared to normal listeners with a simulation has been observed by other authors. For example, Davis et al. (2005) observed an improvement in speech recognition scores from 0% to 70% words correct over 30 noise-vocoded sentences in normal listeners - this is much more rapid than the adaptation described clinically in cochlear implant users (Clark, 2002; Dorman, Hannley, Dankowski, Smith, & McCandless, 1989; Tyler & Summerfield, 1996). Of course, cochlear implantees in these studies vary along factors such as their pre-implantation auditory and linguistic ability, the time of implantation and the amount of residual hearing. However, the differences found between the abilities of normal listeners and implantees suggests and indicates that factors such as central auditory processing, listening strategy and neurobiological recruitment or reorganisation may contribute to the difficulties experienced by cochlear implant wearers. These possible cognitive and anatomical factors add complexity to the process of working out how the listener adapts to their cochlear implant, and this needs to be accounted for in the training of new implantees.

Aside from speech recognition, the introduction of a cochlear implant to deaf patients presents them with other challenges. Music perception is difficult for cochlear implant wearers, and implant recipients have expressed a decrease in their enjoyment of listening to music after implantation. In studies with noise-vocoded simulations, Burns and colleagues (Burns, Sanborn, Shannon, & Fu, 2001) found that simple musical melody recognition required 7-8 bands for 50% recognition, while Z. Smith, Delgutte, and Oxenham (2002) found that over 32 bands were needed to recognise a melody in the presence of a competing melody. Leal and colleagues (Leal et al., 2003) found that 38% of cochlear implant patients in their study did not enjoy listening to music with their implant, while 86% listened to music less after implantation. Scheirer (1998) showed that, using a computerised beat tracker, some musical information could be gained from 6-band noise-vocoded

stimuli through extracting the tempo and beat. However, more recent studies in normal human listeners have shown that beat tracking with 6-band noise-vocoded music samples is significantly worse than with CD-quality samples (Collins & Cross, 2005).

Some studies of music perception in cochlear implant-wearing children are a little more positive. In a survey of children with cochlear implants, Gfeller and colleagues (Gfeller, Witt, Spencer, Stordahl, & Tomblin, 1998) found that a large proportion of children were involved in regular musical activity, and that this involvement, and enjoyment of music, correlated with the children's speech perception abilities. Nakata, Trehub, Mitani, and Kanda (2006) found that Japanese children with cochlear implants exhibited similar timing skills to normal-hearing children during singing, and that some of the implanted children could also, surprisingly, produce sung pitch as accurately as a normally-hearing child. The authors propose that rhythmic perception abilities in the children with cochlear implants may enable them to gain enjoyment from music. They also speculate that neural plasticity in these young implantees may facilitate perceptual learning of music with regard to pitch processing.

A striking phenomenon in the literature on cochlear implants is a large amount of variability in performance between implant recipients (Skinner, 2003; Munson, Donaldson, Allen, Collison, & Nelson, 2003; Sarant, Blamey, Dowell, Clark, & Gibson, 2001; Pisoni, 2000). Shannon, Galvin, and Baskent (2002) observed much greater variability in the baseline speech perception scores of their cochlear implant listeners compared with normal-hearing individuals listening to a cochlear implant simulation. van Wieringen and Wouters (1999) were able to divide their cochlear implant-wearing subjects into three performance groups according to consonant and vowel identification scores. They established that, while all the subjects seemed to use the same information in speech recognition, the better cochlear implant users were more able to make use of certain cues than the other implantees. Wei et al. (2004) comment on high variability in rate discrimination and tone recognition in cochlear implant users. Fu, Chinchilla, and Galvin (2004) found large variability in auditory gender discrimination scores in cochlear implant wearers. Hanekom and Shannon (1998) identified a method of differentiating cochlear implant users according to individual psychophysical tuning curves; they found that these tuning curves were generally sharper in the subjects with better speech recognition. This topic will be addressed further in a later section of this chapter.

1.2.2 Cochlear Implants - processing strategies for better speech recognition

After implantation, learning to use a cochlear implant for speech recognition can be a slow and difficult process, of the order of weeks or months (Clark, 2002; Dorman et al., 1989; Tyler & Summerfield, 1996). Studies of both cochlear implantees and normal-hearing listeners exposed to simulations attempt to achieve greater understanding of the perceptual consequences of implantation, and to establish ways of improving implants to better facilitate speech recognition.

Shannon et al. (1995) found that recognition of noise-vocoded sentences asymptotes at 3 or 4 bands, but this was after a considerable training period, and using very simple test sentences. In contrast, Davis and Johnsrude (2003) found that their materials gave sentence report scores of only medium intelligibility with 7-band noise-vocoded speech, while normal listeners in another study with vocoded speech (Loizou, Dorman, & Tu, 1999) required 8 bands to reach asymptotic sentence recognition performance. Dorman and Loizou (1997a) found asymptotes at 5 bands for sentence recognition, and 8 bands for recognition of isolated vowels in normal listeners. However, a more realistic listening situation is that the cochlear implant user will attempt to understand speech against background noise, featuring competing speakers or otherwise. Using sine-wave carriers in their cochlear implant simulations, Dorman and Loizou (1998) presented normal listeners with vocoded speech in noise. They found that a -2dB signal-to-noise ratio presentation required 20 bands for maximum performance, while +2dB signal-to-noise presentation required 12 bands. Thus, addition of noise to the listening environment places a large demand on the number of channels of resolution needed for speech recognition with a cochlear implant. Other studies have worked on mechanisms to improve speech recognition in these conditions. Nie, Stickney, and Zeng (2005) found that a cochlear implant simulation that included extraction and combination of frequency modulation with amplitude modulation information in each band improved the recognition of sentences in the presence of a competing voice by up to 71%. Turner, Gantz, Vidal, Behrens, and Henry (2004) found that preservation of residual low-frequency acoustic hearing aided perception of noise-vocoded speech against a background of other talkers in normal listeners, and benefited cochlear implant wearers.

An important clinical consideration with cochlear implants is the risk of frequency-place mismatch on the cochlea. The healthy cochlea exhibits frequency-place coding such that a certain point on the basilar membrane, and the auditory nerve fibres running from that point, will be most sensitive to a 'best frequency' or range of frequencies. The place coding is tonotopic, in that basal regions of the cochlea are more sensitive to high frequencies, while apical regions are more

sensitive to lower frequencies. Problems are often encountered in the attempt to fully insert the cochlear implant into the coil of the cochlea, resulting in an absence of electrodes at the apex of the cochlea where low-frequency sounds are transduced, and a consequent basal-ward shift in the placing of the implanted electrodes. This effectively creates an upward shift in the distribution of excitation, and a frequency mismatch between the analysed signal and the tonotopic placing of the electrodes. Several studies using cochlear implant simulations have simulated these spectral shifts by creating a mismatch between the analysis bands (the bandpass filters) and the carrier bands (noise bands/sinewaves). Dorman and Loizou (1997b) investigated the effects of the depth of cochlear implant insertion on speech perception, using a five-channel simulation with sinewave carriers. They found a significant effect of simulated insertion depth; speech recognition performance was normal at a simulated depth of 25mm, but was poorer at 23mm and 22mm. Shannon, Zeng, and Wygonski (1998) found that a simulated tonotopic shift in a 4-channel noise-vocoded signal resulted in a disruption of speech intelligibility, and that this decrease in performance occurred regardless of whether the relative distribution of the simulated electrode array was preserved; this indicated that absolute, not relative, place on the cochlea is represented centrally. A more recent study by Faulkner, Rosen, and Stanton (2003) simulated the effect of shallow cochlear implant insertion depth without distortion of frequency-place mapping. Using 8-band noise-vocoding, they preserved matching of the analysis bands to the centre frequencies of the simulated electrode sites at all simulated insertion depths. In this way, they could assess the effects of low-frequency information loss caused by shallow implant insertion into apical regions of the cochlea. Faulkner et al. looked at identification of consonants, vowels and sentences at simulated insertion depths of 17-25mm from the base of the cochlea. They found that insertion depths of 19mm or less were significantly more detrimental to speech recognition than the deeper insertions tested. However, Faulkner et al. (2003) did observe learning with experience in their study, particularly for the shallowest simulated insertion depth. Other studies of the effects of basalward shifting on the cochlea, with normal listeners (Rosen, Faulkner, & Wilkinson, 1999) and cochlear implant wearers (Fu & Shannon, 1999b; Fu, Shannon, & Galvin, 2002), support the suggestion that cochlear implant wearers have the ability to at least partially adapt to a basalward spectral shift.

Work has also been carried out on the relative importance of each frequency band in the noise-vocoded stimulus. In addition to the mismatch problems caused by improper insertion of the cochlear implant, frequency-place mismatch and distortion can be caused by cochlear pathology. For example, an implant patient may have local areas of neuronal damage and loss along the cochlea, which effectively cause a 'hole' in the frequency spectrum of the transmitted signal. Shannon et al. (2002) simulated these 'holes in hearing' in cochlear implant listeners, and in normal listeners with noise-vocoded simulation. The holes were created in basal, middle or apical parts

of the cochlea by eliminating electrodes or noise bands - the information that would have been communicated by these electrodes or noise bands was then treated in one of four ways:

1. Dropped from the signal
2. Assigned to electrodes/noise bands on the apical side of the hole
3. Assigned to electrodes/noise bands on the basal side of the hole
4. Split to both sides of the hole

Shannon et al. (2002) found that holes in the apical region i.e. the lower-frequency region were most damaging to speech perception. However, there was no advantage of redirecting information around the hole over dropping the information. Interestingly, Shannon et al. found a similar pattern of effects in normal listeners and cochlear implant users, which suggested that the loss of spectral information was the main cause of speech perception deficits, rather than processing or other differences between the groups. Apoux, Tribut, Debruille, and Lorenzi (2004) obtained contrasting results to Shannon et al.. They used 4-band noise-vocoded speech to assess the relative importance of temporal information for consonant identification in different frequency regions, on the basis that noise-vocoding forces listeners to primarily use temporal envelope cues in perception. Like Shannon et al. (2002), Apoux et al. (2004) simulated holes in hearing to look at frequency dependency, both in quiet and in noise. They found that temporal cues in the highest frequency band (3.5kHz and above; i.e. basal regions of the cochlea) were most important for consonant identification. They believe that the difference between their study and Shannon et al. may be linked to the fact that their (Apoux et al., 2004) study used only 4 bands of information, while Shannon et al. used 6 to 10 simultaneous electrodes/bands of information. Apoux and Bacon suggest that the greater spectral degradation in their stimuli led to the difference in frequency dependence. They propose that this finding might inform what might happen in some cochlear implant users, who cannot make full use of the number of electrodes of information provided to them and so experience a more spectrally-degraded signal than predicted from the resolving capabilities of the processor.

Some authors have investigated the possibility of expansion of the amplitude envelope to assist speech recognition with vocoded stimuli. Apoux, Crouzet, and Lorenzi (2001) applied a power-law transformation to temporal modulations below 500Hz and observed some benefit in recognition scores for vowel-consonant-vowel patterns in normal listeners, and an improvement in reaction times for both normal-hearing and hearing-impaired listeners. Fu and Shannon (1999a) described the

effects altering the acoustic dynamic range on phoneme recognition in cochlear implant users and normal listeners. They found that phoneme recognition could still be performed using a 4-channel processor (or simulation) when the dynamic range was reduced to 30dB. Below this, phoneme identification was impaired. However, centre clipping gave a small increase in performance in cochlear implant users when the phonemes were presented against background noise.

1.2.3 Noise-Vocoded Speech - training and perceptual learning studies

While many studies have looked at the immediate impact of stimulus alterations on recognition of noise-vocoded stimuli, others have concentrated on the process of adaptation to the stimulus. Rosen et al. (1999) observed a substantial improvement in normal listeners' performance with spectrally shifted speech after a few hours of training. Fu et al. (2002) reported significant improvements in speech perception in cochlear implant wearers over the first three weeks of experience with a new filter-to-electrode mapping.

Davis et al. (2005) carried out five key experiments exploring the perceptual learning of 6-band noise-vocoded speech in normal-hearing listeners in terms of both stimulus content and training regime. Experiment 1 demonstrated that perceptual learning of noise-vocoded speech can occur rapidly. Over 30 sentences of exposure (approximately 20 mins), normal listeners' sentence report scores improved from nearly 0% to 70% words correct. Experiments 2 and 3 showed that listeners benefitted significantly from a training regime in which they were given a second opportunity to hear each distorted sentence after being presented with the sentence identity in auditory or written form (a *'Distorted-Clear/Written-Distorted'* or *'DCD'* presentation). This training routine was more effectivet than one in which listeners heard a repetition of the distorted form *before* the clear version (a *'Distorted-Distorted-Clear/Written'* or *'DDC'* presentation). Auditory and written feedback provided equivalent facilitation for learning; the authors suggest that this shows that learning is driven by higher-level sentence information, and is not dependent on acoustic information. However, in a final experiment, Davis et al. found that, while real word information during training was necessary for maximally efficient learning, there was no advantage of semantic coherence in the training sentences. The authors concluded that lexical information drives perceptual learning of noise-vocoded speech. However, they also note that listeners' improvement in recognition during the experiments extended to words that they had not previously heard in noise-vocoded form; this suggests that perceptual changes are occurring at a prelexical level of processing. Hervais-Adelman, Davis, Taylor, Carlyon, and Johnsrude (2006) found that the locus of learning is not at the level of peripheral auditory processing, given that learning of noise-vocoded

sentences can be transferred across frequency regions. They showed that listeners who experienced a change in frequency range (from 50-1406Hz to 1593-5000Hz or *vice versa*) halfway through a set of 40 bandpass-filtered noise-vocoded sentences showed equivalent learning to those who experienced the same bandpass-filtered frequency range throughout. In summary, Davis et al. (2005) interpret their findings as indicating that the learning mechanism operates via feedback from a lexical level that alters perceptual processing at a prelexical level.

Davis et al. (2005) saw the clear presentation in the 'DCD' training routine as an effective top-down 'teaching signal' which could be used to map the acoustic stimulus onto linguistic representations during the second presentation of the distorted sentence on each trial. Hervais-Adelman, Johnsrude, Davis, and Carlyon (in press) have continued work on perceptual learning of noise-vocoded speech in normal-hearing adults, but they focus on learning of isolated word stimuli only. They propose that the training disadvantage for the 'DDC' training regime in Davis et al. may relate to short-term memory factors. Given that a distorted sentence is likely to be less readily or completely encoded in short term auditory and verbal memory than its undistorted equivalent, the clear feedback in the 'DDC' presentation can perhaps only be used in relation to a fading memory trace of the second distorted presentation, while in the 'DCD' presentation the clear sentence should be more robust in short-term memory by the time the second distorted version is played. Hervais-Adelman et al. (in press) propose that if the clear sentence in the 'DCD' routine is truly a top-down teaching signal, then the 'DCD' advantage should be replicated in a paradigm using only single words (which they claim will place a lesser load on short-term memory). Indeed, they find that the effect remains for the single word paradigm (Experiment 1).

In a second experiment, Hervais-Adelman et al. (in press) re-visit the question of the lack of training observed for noise-vocoded nonword sentences in Davis et al. (2005). The original study showed no training with nonword sentences, even when short-term memory constraints were taken into account by maintaining written feedback onscreen during the second distorted presentation (in the 'DCD' routine). However, Hervais-Adelman et al. (in press) acknowledge that additional factors, such as the difficulty of segmenting nonword sentences into pseudo-lexical units for the purposes of comparing these with the written versions, may have obstructed learning. This is a strong possibility, as the nonword sentences in Davis et al. were, even in undistorted form, slower and less 'naturalistic' rhythmically than the other sentence sets employed in the study[1] - the role of linguistic rhythm in learning will be explored in Chapters 4-6 of this thesis. To circumvent the issue of lexical segmentation, Hervais-Adelman et al. (in press) compare the training efficiacies of isolated words with that of isolated nonwords, using the 'DCD' training routine. They found that

[1] Personal observation, as an author on the paper

training was significant and equivalent for both stimulus types, suggesting that lexical information is perhaps only necessary for learning in certain contexts. Hervais-Adelman et al. conclude that while learning benefit from top-downs cues, these may not necessarily come from feedback. Davis et al. showed that learning of English noise-vocoded sentences could occur when no feedback was present at all. Hervais-Adelman et al. find that learning with isolated noise-vocoded words is much slower than that observed with the same number of words in sentences, and suggest that this is because a listener has a much richer set of cues and expectations when listening to sentences. In this way, a listener might be able to get a better sense of whether or not their sentence perception is successful on each trial (e.g. if their percept 'makes sense' semantically and syntactically) than they would have for a task involving recognition of isolated words. Therefore, Hervais-Adelman et al. conclude that what seems to be crucial for learning is 'the presence of some constraint on the interpretation of distorted speech that permits listeners to reinforce accurate pereptual hypotheses and make alterations that can correct inaccurate hypotheses'.

It is worth mentioning that the finding by (Davis et al., 2005) that certain types of stimuli can produce significant learning of noise-vocoded speech, while others cannot, is in contrast to the findings of some of the research in the field of auditory learning. Amitay, Hawkey, and Moore (2005) have found that performance on pure tone frequency discrimination can improve quickly and dramatically in normal-hearing adults. They have found that learning of this task is little affected by variation of the training regime, and is still observed (to a small extent) when training involves a task with no auditory stimulation at all (Amitay, Irwin, & Moore, 2006). They interpret these findings as suggesting that auditory learning may involve attentional processing as well as the expected sensory components, claiming potential applications for the improvement of listening skills in populations such as cochlear implant wearers. Logan, Lively, and Pisoni (1991) have shown that Japanese listeners can be trained to perceive the contrast between the English phonemes /r/ and /l/, which do not occur in the Japanese language, thus demonstrating the transfer of auditory learning to speech perception skills.

The approach of using separate training and test phases to compare the adaptive properties of different stimulus types has already been described. The findings of Davis et al. (2005) and Hervais-Adelman et al. (in press) have demonstrated the usefulness of feedback regimes within this general paradigm. However, one could claim that Davis et al.'s methods of training are ecologically invalid, as they involve no real human interaction and all the materials are pre-recorded, clear speech stimuli in isolated sentences. The day-to-day challenges facing a cochlear implant wearer are likely to involve communicative exchanges with other individuals in a conversational setting, either face-to-face or on the telephone.

Rosen et al. (1999) addressed some of these issues by using a technique called Connected Discourse Tracking (CDT) in training. This technique was designed by Defilippo and Scott (1978) and follows a format where a 'talker' reads segments to a 'receiver' from a prepared text. The receiver's task is repeat back what the talker has said. The talker cannot progress to the next segment until verbatim repetition has been given. However, the talker can use different strategies to achieve the correct response from the receiver. For example, he/she can change the way in which the segment of text is read by changing emphasis or speed, or he/she can reduce the length of the segment. In this way, progress through the material is made in an adaptive and cooperative way. Performance is measured as an 'absolute words per minute (wpm)' score, where this is the number of words read divided by the time elapsed in minutes. Rosen et al. (1999) used this technique in both audio-visual (where the receiver could see the talker through a glass panel) and audio-only contexts to train normal-hearing listeners with spectrally-shifted noise-vocoded speech (which simulated a basalward shift of 6mm on the cochlea). They found that, after nine 20-minute sessions of CDT with the spectrally-shifted stimuli, scores on speech recognition tests using the shifted speech had improved significantly.

M. Smith and Faulkner (2006) re-visited the issue of 'holes in hearing' (as investigated by Shannon et al. (2002)) in order to explore effects on perceptual learning. Smith and Faulkner simulated a 10mm hole in a mid-frequency region of the cochlea within a six-channel noise-vocoded cochlear implant simulation. They trained normal-hearing listeners, using CDT, on two schemes for reassignment of information from the 'hole' - one in which the missing information was assigned across the two bands bordering the hole, and one where the information was spread evenly across all six output bands. The results showed that performance improved over 3 hours of training for both reassignment conditions, to a greater extent than the improvement seen for a 'dropped' condition in which information from the hole region was dropped from the input. This study stands in contrast to that of Shannon et al. (2002), who found no advantage of reassignment. Smith and Faulkner suggest that, over time, preservation of acoustic information with warping is better than losing this information completely.

A disadvantage of CDT is that it is very labour intensive. Realistically, the level of one-to-one therapy needed using this approach with a cochlear implant user may not be convenient for the patient, nor implementable by the service provider. Other studies have considered the usefulness of computer training paradigms on perceptual learning in normal-hearing populations. Fu, Nogaki, and Galvin (2005) found that a training regime using spectrally-shifted, sinewave-vocoded (8-band) words led to significant improvements in recognition of shifted vowels. Their training task used words with CVC (*consonant-vowel-consonant*) structure in discrimination tasks in which the

listener was required to focus attention on the medial vowel. For example, an initial trial may ask the subject to label the auditory stimulus by choosing between two stimuli that were acoustically quite different (e.g. "said" and "sued"), but as performance improves, the onscreen discriminations first become acoustically-similar pairs (e.g. "said" and "sad"), followed by increases in the number of onscreen options (to a maximum of 6) with further improvement in performance. Fu et al. found significant improvement in vowel identification test performance after five such training tasks. Notably, performance also improved significantly on isolated consonant recognition after training with words, even though the within-trial discriminations in the training only required participants to perform medial vowel discriminations. In a follow-up study, Nogaki et al. (2007) found that the rate of training did not affect the significance of this improvement in performance. In their study, all participants received five 1-hour sessions of training, but the three groups experienced different rates - 1 session per week, 3 sessions per week or 5 sessions per week. The authors note, however, that there was considerable variability in the amount of learning exhibited by individuals. They ascribe this to variable levels of enthusiasm and involvement in a difficult task, and contrast this with the keener sense of urgency shown by cochlear implant patients, for whom successful training has important consequences for their quality of life.

Stacey and Summerfield (2007) recently built upon the study by Fu et al. (2005), using similar computer-based training paradigms with words and sentences to train normal-hearing listeners exposed to spectrally-shifted noise-vocoded speech. For words, quasi-minimal pairs were presented onscreen as response options for the identity of the auditory stimulus. Accuracy feedback was given onscreen for each trial, and an incorrect response meant that the trial would be repeated until a correct response was achieved. For sentence training, acoustic sentence presentation was followed by presentation of six written words randomly position on the computer screen, where the participant's task was to select the three words that had been present in the auditory stimulus. Selection of an incorrect word resulted in repetition of the sentence, and this pattern continued until all three words were correctly selected. Once all three words were correctly identified, a written version of the full correct sentence was displayed followed by a final acoustic presentation, after which the participant was asked once more to pick out the three keywords in the sentence. Importantly, in order to address findings such as those by Amitay et al. (2006), Stacey and Summerfield included a visual control task, which operated in the same way as the auditory training but replaced auditory presentations with distorted visual presentations of target stimuli. This enabled them to account for the proportion of learning that could be attributed to 'incidental', or procedural, learning. The intensive auditory training with words and sentences produced significant improvements in auditory recognition of spectrally-shifted noise-vocoded sentences, consonants (although not after sentence training) and vowels. On several occasions, the amount of learning with auditory stimulation for

consonants and vowels was not significantly greater than for the visual control task. However, there was an overall advantage for auditory training, which was found to be long-lasting (on the order of weeks) and therefore led the authors to conclude that computer-based auditory training is a potentially valuable means of rehabilitation in cochlear implant users.

1.2.4 Noise-Vocoded Speech and the brain - neuro-imaging studies

While many studies have been devoted to noise-vocoded speech purely as a simulation of cochlear implant stimulation, others have used the distortion as a tool to uncover processing in the brain of the normal-hearing listener. In a PET (positron emission tomography) imaging study, Scott et al. (2000) compared brain responses to 4 types of speech stimuli: normal speech, 6-band noise-vocoded speech, and spectrally rotated versions of these stimuli. Spectrally-rotated speech is made by inverting high and low frequency information about a chosen frequency value - essentially, the speech spectrum is 'flipped' around this frequency. In this way, the spectral and temporal complexity of the original speech stimulus is preserved. This form of distorted speech is only intelligible to some listeners after considerable training (Rosen, Finn, & Faulkner, 2002), whereas normal speech and 6-channel noise-vocoded speech are more readily intelligible. Thus, in the four types of stimulus used in the Scott et al. study, two were intelligible (normal speech and noise-vocoded speech), and two were unintelligible. Furthermore, each of the unrotated conditions was matched by its rotated counterpart in terms of spectral and temporal structural complexity. Subtraction of the activations for the unintelligible stimuli from activations for the intelligible stimuli gave brain responses to intelligibility - these were found in the left anterior superior temporal sulcus (STS). Subtraction of the activations for noise-based stimuli (noise-vocoded speech and rotated noise-vocoded speech) from the speech-based stimuli (undistorted speech and its rotated counterpart) gave responses in the right anterior temporal lobe. The authors suggested that this area may be a neural correlate of the perception of speech melody or intonation, as the speech-based stimuli possess more dynamic pitch variation relative to the noise-based stimuli. By subtracting the rotated noise-vocoded speech condition from the other three, Scott et al. found areas responsive to phonetic information, but not necessarily to intelligent speech, in the left superior temporal gyrus (STG), lateral and anterior to the primary auditory cortex, and in posterior left STS. Narain et al. (2003) carried out an fMRI study using the same four speech conditions as Scott et al. (2000). A conjunction analysis found activation for intelligible speech in the left anterior STS (as in the Scott et al. study), but Narain et al. also found a second activation site in the dorsal posterior margin of the left temporal lobe, within classical Wernicke's area. This supports the classical view that Wernicke's area is important in speech comprehension.

Davis and Johnsrude (2003) carried out an fMRI study of adaptation to distorted speech using segmented speech (created by dividing the speech waveform into short chunks at fixed intervals and replacing even-numbered chunks of speech with a signal-correlated noise version of the original speech ; Bashford, Warren, and Brown (1996)), noise-vocoded speech and speech in noise. These three types of distorted speech stimuli were each presented at three different levels of severity to produce a range of intelligibility ratings and sentence report scores. Normal undistorted speech and signal-correlated noise were also included as a maximally intelligible stimulus and an unintelligible stimulus, respectively. Using a correlational approach, Davis and Johnsrude identified areas responding to speech intelligibility by looking for a positive correlation between the BOLD response and the listeners' sentence recognition scores. They found responses along the length of the left STG and middle temporal gyrus (MTG) and less extensively in the corresponding areas of the right hemisphere, as well as in the left inferior frontal gyrus (LIFG) and the body of the left hippocampal complex. Within these areas, the authors found areas of sensitivity to the differences between the distortion types in the STS bilaterally, thus indicating form-dependent responses to intelligibility. The authors suggest that these areas are involved in acoustic analysis of the speech signal. Form-independent responses to intelligibility were found in anterior MTG bilaterally, left posterior STS, LIFG, left hippocampus and left precuneous - the authors suggest these areas are involved in higher-level linguistic, non-acoustic processing of speech. The paper also identifies both form-dependent and form-independent areas that showed 'compensation for distortion' i.e. greater activation for the three distorted stimuli types than for undistorted speech and signal-correlated noise. Hervais-Adelman, Johnsrude, Carlyon, and Davis (2007) recently carried out an fMRI study in which they attempted to identify the neural correlates of perceptual adaptation to noise-vocoded words. They were unable to find any correlates of behavioural improvement in speech recognition, but did identify a similar fronto-temporal network of activations, corresponding to effortful speech comprehension, as that observed by Davis and Johnsrude.

Sharp, Scott, and Wise (2004a) investigated semantic processing within inferior temporal cortex (IT) in control subjects and in aphasic patients with infarction involving the superior temporal sulcus (STS). They proposed that infarction of STS, which is involved in speech perception, disrupts feedforward of perceptual information from this area to IT, an area believed to be involved in accessing word meaning. They modelled this effect in normal listeners by presenting them with degraded speech in the form of 8-channel noise-vocoding. It was found that patients listening to clear speech and normal listeners listening to noise-vocoded speech showed impairment in performance on a spoken semantic task, and a reduction in activity in the left anterior fusiform gyrus in left IT. In another study, Sharp, Scott, and Wise (2004b) used 8-channel noise-vocoded speech in a similar way to observe the involvement of the prefrontal cortex in semantic processing. They found

that degradation of speech in a semantic task reduced activity seen in left rostral prefrontal cortex (RPFC) with clear speech, but caused activity in right dorsolateral prefrontal cortex (DLPFC). This DLPFC activation was inversely proportional to task performance. The authors suggest that left RPFC is involved in 'extensive semantic elaboration' while right DLPFC is recruited as monitoring demands from the stimuli increase.

Obleser, Wise, Dresner, and Scott (2007) carried out an fMRI investigation of the neural correlates of 'top-down' and 'bottom-up' processing in speech perception, using noise-vocoding (to 2, 8 and 32 bands) to manipulate acoustic or 'bottom-up' information and semantic predictability (high or low predictability sentences) to manipulate linguistic or 'top-down' information. At intermediate levels of sound degradation (8-band noise-vocoded speech), Obleser et al. identified a left-lateralized network of brain areas including prefrontal cortex, angular gyrus and the posterior cingulate, whose activity and connectivity increased with an increase in semantic predictability. This gave a clear demonstration of top-down influences on speech intelligibility, operating in higher-order cortical regions than those associated with basic auditory processing.

Noise-vocoded stimuli have also been used in the identification of neural substrates of human speaker discrimination. In an analysis of fMRI-derived BOLD responses to 1-, 6- and 32-channel noise-vocoded words spoken by six different males, J. Warren, Scott, Price, and Griffiths (2006) found a bilateral network including posterior and middle parts of the superior temporal sulcus corresponding to the interaction of spectral resolution and voice source (i.e. a change in speaker). The authors claim these areas reflect the detailed extraction of voice information based on the available spectro-temporal detail.

Thus, noise-vocoded speech has proven to be a highly valuable tool in the investigation of the neural correlates of a range of speech, voice and linguistic behaviours. However, studies to date have not identified the neural correlates of perceptual learning with this distortion type.

1.3 Experiments in this thesis

Stimulus properties and listener variability

The studies described so far in this chapter have given an overview of the increasing use of noise-vocoded speech as a cochlear implant simulation and as a means of investigating speech recognition and perceptual learning, both behaviourally and neurally, in the normal-hearing population. The

central measures in these studies (with some exceptions) have mainly been concerned with the level of speech recognition achievable with an implant or simulation. Some of the studies focus on basic speech recognition, while others concentrate on exploring the mechanisms of perceptual learning, that is, the improvement in speech recognition performance over time and further exposure. However, the presented data suggest two overall factors driving performance outcomes: (1) the properties of the speech stimulus, and (2) inter-listener variability in the skills required to perform the task.

Shannon, Fu, and Galvin (2004) agree that differences in listening difficulty can be task-based or listener-based. In a meta-analysis of several studies using cochlear-implant simulations, Shannon et al. point out that *"the number of spectral channels required for speech recognition depends on the difficulty of the listening situation"*(p.50), where difficulty can emerge from the informational impoverishment of the distorted stimulus or the limits on the listener's abilities. For example, melody identification is more challenging than simple sentence recognition with pitch-impoverished cochlear implant simulations, and performance of speech recognition tasks with distorted stimuli is more difficult for non-native speakers of a language than for native speakers.

The approach taken in this thesis follows the view that speech perception is a behaviour emergent from the interaction of the perceptual skills of the listener with the content (acoustic and linguistic) of the stimulus. The following sections outline two proposed themes, one listener-based and one task-based, for experimental investigation using noise-vocoded speech.

1.3.1 Stimulus properties: The role of rhythm and timing

Like music, spoken language has rhythmic structure which can be exploited by the listener. Proponents of the *isochrony hypothesis* (Abercrombie, 1967; Pike, 1945) claimed that there is measurable temporal regularity in speech. The approach distinguished between 'stress-timed' languages as those that possess roughly equivalent durations between successive *stressed syllables* (e.g. English, German and Dutch), while 'syllable-timed' languages were those for which the duration of successive syllables was thought to be relatively constant (e.g. Spanish, Italian, French). More recent studies where these durations have been measured have disproven the specific claims of the isochrony hypothesis (e.g. Dauer, 1987a). However, other authors have shown that the 'classes' espoused by the orginal hypothesis stand when rhythm is described using different metrics, which are based on the durations of consonantal and vocalic intervals in speech (Dellwo, Steiner, Aschenberner, Dankovičová, & Wagner, 2004; Ramus, Nespor, & Mehler, 1999; Grabe & Low, 2002).

Several authors have written on the importance of rhythm as a cue for segmenting the continuous speech stream into words. Cutler and colleagues (Cutler & Butterfield, 1992; Cutler & Norris, 1988) have presented evidence to suggest that English-speaking listeners use speech stress to segment the signal into words by using the hypothesis that strong syllables occur at word onsets in English. M. Smith, Cutler, Butterfield, and Nimmo-Smith (1989) showed that English listeners could use cues from the duration of syllabic portions of speech to locate the position of word boundaries in noise-masked speech. The type of segmentation strategy used has been found to be dependent upon the listeners' first language, with French listeners using the syllable as the unit of segmentation (Mehler, Dommergues, Frauenfelder, & Segui, 1981), and Japanese listeners using the mora (Otake, Hatano, Cutler, & Mehler, 1993). Interestingly, listeners will use the strategy corresponding to their native language, even if they are listening to another language (Cutler, Mehler, Norris, & Segui, 1986; Otake et al., 1993). There is also evidence from the developmental literature in support of speech rhythm as an important segmentation cue used by infants (Nazzi & Ramus, 2003). Nazzi, Bertoncini, and Mehler (1998) showed that newborn infants are able to discriminate between languages from different rhythmic classes.

While the *rhythmic segmentation hypothesis* put forward by Cutler and colleagues places emphasis on 'bottom-up' auditory processing (based on stress and durational information) in identifying the onsets of lexical units in the speech stream, other authors would challenge the simplicity of this interpretation. Mattys and colleagues put forward a hierarchical structure of factors affecting segmentation, placing higher-level linguistic knowledge such as semantic context and lexical probability (i.e. the process of segmenting the stream to produce words rather than nonwords) at the top of this hierarchy, and metrical stress at the bottom (Mattys, White, & Melhorn, 2005). Davis and Johnsrude (2007) also describe the multiplicity of cues to segmentation, and the difficulty of implementing this in models of speech recognition. However, there is general agreement that, under conditions of reduced lexical clarity through distortion or addition of noise, sublexical factors such as stress and timing (and acoustic-phonetic factors including coarticulation and the use of experience-dependent statistical knowledge about word boundary phonotactics) become more prominent (Cutler & Butterfield, 1992; Davis, Marslen-Wilson, & Gaskell, 2002; Mattys et al., 2005).

Other authors have put forward opinions as to the role of rhythmic structure in listening to complex auditory stimuli such as music and speech. Jones and Yee (1993) interpret the importance of rhythm within an attentional framework, in which temporal regularities in speech may be important in the generation of *expectations* about speech content. In this scheme, violation of expectations, through for example pausing or the stretching of inter-stress intervals across syntactic

phrase boundaries, alert the listener to sentence structure. However, Jones and Yee do not offer specific mechanisms by which these rhythmic regularities might be extracted by the listener. Where violation through pausing may alert the listener to phrase boundaries, the authors suggest that a 'domain-specific' knowledge of rhythmic patterns, for example learned knowledge that stressed syllables are likely to signal word boundaries, enables further segmentation of the stream into lexical units. Hence, the *rhythmic segmentation hypothesis* might be seen as attentional in nature. A slightly different opinion is given by Boltz (1998). Her study of memory for melodies showed that 'incoherence' between pitch and rhythmic information impairs melody recall compared with situations where the two aspects of the melody cohere. Boltz claims that this incoherence (which, as suggested by Jones with reference to speech stress, is surely detectable with domain-specific experience) impairs the success of perceptual encoding of the to-be-remembered melody. Citing studies that identify similarities between speech and music, Boltz posits that similar issues of coherence may affect the remembering of speech. Thus, Boltz presents a working memory framework for rhythm in perception of complex auditory stimuli.

In the literature so far, there are no direct studies of the importance of speech rhythm in adaptation to noise-vocoded speech. However, there are several reasons why rhythm may be an important cue for perception and adaptation in noise-vocoded stimuli:

1. The nature of the noise-vocoded stimulus - Rosen (1992) offers a three-layer framework to describe the temporal information in speech. Noise-vocoded speech primarily preserves the slow-moving amplitude envelope information (fluctuations at rates from around 2 to 50Hz) from the original speech stimulus (Shannon et al., 1995), and Rosen describes how cues from this envelope are influential to a considerable extent in perception of manner of articulation, speech tempo/rhythm and syllabicity, and to a lesser extent for perception of voicing, voice quality and stress.

2. Previous studies with cochlear implants and distorted speech - Numerous studies have described the importance of temporal information for speech perception (van der Horst, Leeuw, & Dreschler, 1999), particularly when the speech stimulus is degraded (Shannon et al., 1995; Salomon, A and Espy-Wilson, CY and Deshmukh, O, n.d.; van Tasell, Soli, Kirby, & Widin, 1987). It is possible that this could extend to speech rhythm. In a melody perception task with cochlear implant users and normal listeners using a simulation (Kong, Cruz, Jones, & Zeng, 2004), it was shown that cochlear implant users depended more on the rhythmic cues than the pitch cues. Gfeller and Lansing (1992) found that cochlear implant wearers were more accurate at the Rhythm subtest of the Primary Measures of Music Audiation (PMMA; Gordon (1986)) than they were on the Tonal subtest. More recently, Meister et al. (2007)

found that, in prosodic descrimination tasks, cochlear implant users rarely made errors on discriminations based on temporal structure, while they performed worse that normal-hearing listeners on discriminations based on variations on amplitude or pitch. Nakata et al. (2006) found that performances of songs from implanted children and normal-hearing children were similar in rhythm, while on pitch the deaf children made errors of pitch range and direction.

Evidence for a possible role for rhythm in perceptual adaptation comes from previous studies using time-compressed speech (Mehler et al., 1993; Pallier et al., 1998; Sebastian-Galles et al., 2000), in which the authors observed transfer of adaptation between languages of the same rhythmic class (e.g. from Dutch to English) but not across class (e.g. from French to English), indicating a possible role for language rhythm as a cue to this adaptation process.

3. Evidence from studies of dyslexia. A number of studies with dyslexic listeners have indicated that there are problems with temporal processing in this population. Muneaux and colleagues (Muneaux, Ziegler, Truc, Thomson, & Goswami, 2004) observed problems with beat and tempo detection in dyslexics. Goswami and colleagues have carried out extensive work showing weaknesses in beat detection in dyslexic children (Goswami et al., 2002), namely in their abilities to make duration discriminations and to discriminate amplitude envelope rise times. Thomson, Fryer, Maltby, and Goswami (2006) recently showed that such deficits in temporal processing continue into adulthood in dyslexic individuals. As a stimulus that preserves temporal information while discarding much of the spectral detail, the possibility exists that noise-vocoded speech could be used as a diagnostic tool in uncovering timing-related speech processing problems such as those exhibited in dyslexia.

Another important factor in the motivation to study the role of rhythm and timing in noise-vocoded speech perception comes from the personal observation of a certain lack of rhythmic 'naturalness' in the nonword sentences used by Davis et al. (2005). The lack of training with nonword sentences in Davis et al. (2005), followed by significant training with isolated nonwords in Hervais-Adelman et al. (in press) leaves open the question of the necessity of lexical information in adaptation to noise-vocoded speech. This is particularly confusing in light of evidence of learning with lexically unfamiliar sentence materials in studies using time-compressed speech (Altmann & Young, 1993; Pallier et al., 1998; Sebastian-Galles et al., 2000).

Whilst it is acknowledged that linguistic rhythm is carried by several cues in speech, the experiments of this thesis will measure rhythm primarily along a temporal hypothesis, based on measurement of the durations of consonantal and vocalic intervals in speech. The reasoning for this is that the nature of the noise-vocoded stimulus means that durational and tempo informa-

tion are extremely well transferred, even at low numbers of bands/channels, while the conveyed information on stress is weakened by a lack of pitch. Thus, even though syllable stress survives noise-vocoding, it is the temporal aspects of this which are most faithfully preserved and should therefore occupy the focus of studies on speech rhythm within noise-vocoded stimuli.

1.3.2 Listener variability: Individual differences

Variability in speech recognition amongst cochlear implant users has been described above. This section outlines previous attempts to identify correlates of this variabilty, in implanted children and adults.

Pisoni and colleagues have carried out extensive investigation into the factors predicting speech processing ability in children with cochlear implants (Cleary, Pisoni, & Geers, 2001; Cleary, Pisoni, & Kirk, 2002; Dillon, Pisoni, Cleary, & Carter, 2004; Dillon, Burkholder, Cleary, & Pisoni, 2004; Pisoni, 2000; Pisoni & Geers, 2000; Pisoni & Cleary, 2003). They point towards cognitive differences between implanted children as the root of their varied ability with implants, emphasising the importance of including "process" measures of children's capacity for learning, memory and attention in addition to the more traditional audiological outcome measures (Pisoni, 2000). Pisoni and Geers (2000) tested 43 children with cochlear implants who were relatively similar in demographic background. They found that there was a strong correlation between verbal digit span and measures of speech perception, speech production, language development and reading skills. Pisoni and Cleary (2003) unpacked the verbal digit span task into its component parts, namely working memory capacity and verbal rehearsal speed, by accompanying the digit span task with measures of children's speech rates as an indicator of rehearsal speed. With demographic factors such as age of onset of deafness and duration of implant use partialled out, it was found that as much as 20% of variability in speech perception amongst paediatric cochlear implant users might be related to maintenance and retrieval of representations of spoken words in working memory (i.e. verbal/subvocal rehearsal), while 7% could be accounted for by differences in working memory capacity.

In adults, a number of studies have pointed towards cognitive factors as predictors of performance with a cochlear implant. Knutson et al. (1991) found that performance in a visual monitoring task in postlingually deafened adults predicted their audiological performance 18 months after receiving a cochlear implant, while standard measures of verbal and performance IQ had no predictive value. The monitoring task was intended to measure the listener's ability to rapidly

detect and respond to features within sequentially arranged information, as this was posited as an important in processing incoming signals through the implant. However, it could be claimed that the task which emerged as the strongest predictor of audiological outcome loads on similar properties to those involved in the verbal digit span task. The visual monitoring task in Knutson et al. involved the subject viewing a sequence of single digits on a computer screen and pressing a button when he or she had seen an even-odd-even sequence. This involved holding the last two digits in working memory, and updating the memory of these two digits when each new digit is presented. It is intuitive to assume that, even though the stimuli were visual, verbal encoding and rehearsal would have taken place as a means of encoding the items in working memory. The same visual monitoring task also emerged in a later multivariate analysis to find a predictive index of audiological outcome 9 months after cochlear implantation of postlingually deafened adults (Gantz, Woodworth, Knutson, Abbas, & Tyler, 1993).

More recently, Collison, Munson, and Carney (2004) attempted to identify cognitive and linguistic performance correlates of spoken word recognition in a group of 15 adult cochlear implant users. They measured participants' scores on one standardized test of nonverbal cognitive ability, a test of expressive vocabulary and a test of higher-level linguistic ability. Correlations between scores on the cognitive-linguistic measures and performance on monosyllabic word recognition and a word gating task (where the participant makes guesses as to the word identity while being played progressively longer chunks of the word from onset) offered no strong relationships, although almost all were in the predicted direction (better scores on one task associated with better scores on the other). Collison et al. (2004) explain this null result in terms of the heterogeneity of their listening population, in terms of the exact implanted device used by each listener, age, aetiology and duration of deafness. Lyxell et al. (1998) had rather more success in their study, in which they assessed the relationship between pre-implantation performance on a battery of cognitive-linguistic tasks (including letter name matching, visual lexical decision, semantic categorization of words, rhyme judgements (visual), and two tests of memory span for linguistic materials) and the level of reported speech comprehension after 12 months of implant use. They found that the strongest speech comprehenders (those who could use the telephone or understand a conversation with a familiar speaker without seeing their face) were those who showed the greatest evidence of preserved phonological representations in the lexical decision and rhyme judgement tasks. The degree of evidence for preserved phonological representations was negatively correlated with the number of years of deafness, suggesting that these representations degrade over time; however, there were exceptions to this. Similarly, most of the better speech comprehenders gave better performance on the pre-implantation test of short-term memory span, yet two of these better comprehenders gave weak memory span scores. Lyxell et al. write that this indicates that a good working memory

capacity may not be necessary for good performance with an implant, as long as there are strong mental representations of speech sounds. They point to investigation of the factors enabling the preservation of phonological representations as one of the key questions for future research.

There have been several previous studies on the contributing factors to inter-individual variation in speech processing ability in normal-hearing adults. Surprenant and Watson (2001) explicitly attempted to identify auditory spectro-temporal factors underlying speech and nonspeech perception in normal-hearing college students. They ran a battery of speech and nonspeech auditory tasks on a large number of students, and ran a factor analysis on the scores. It was found that speech and nonspeech tasks loaded on different factors; the authors suggested that this either signified that speech and nonspeech processing were truly orthogonal, or that there were problems with the choice of tests in the study. Supplementary measures of general intelligence loaded on a factor with auditory temporal order tasks, but with none of the other perceptual tasks. Surprenant and Watson (2001) suggest that this may inform the theories of Tallal and colleagues (e.g. Tallal, 1980) with regard to the role of temporal processing in developmental language disorders, in that these temporal processing skills may be associated with general cognitive ability rather than linguistic processing specifically.

With regard to test choice, Surprenant and Watson (2001) point out that psychoacoustic measures such as those used in their nonspeech tasks often measure the limits of sensitivity, while speech processing involves the discrimination of many suprathreshold changes. There is also a difference in attentional set-point or focus for these two task sets - the nonspeech tasks, which were spectro-temporal discrimination tasks, may have encouraged what Surprenant and Watson call 'analytic listening' for a very subtle acoustic change. In contrast, the speech perception tasks may have encouraged more global listening. Another issue with Surprenant and Watson's (2001) study is that all of the speech perception tasks measured detection of speech tokens (consonant-vowel syllables, words or sentences) against white noise. It is possible that this type of task engages more attentional than perceptual mechanisms, in that the listener is trying to detect a signal in noise rather than interpret a coherent distorted stimulus, and perhaps therefore should not be the only task type in a speech perception battery. In defence of this potential weakness, Watson, Qiu, Chamberlain, and Li (1996) found a significant correlation between speech-reading ability and recognition of speech in noise in normal listeners, thus indicating a possible modality-independent, linguistic processing commonality between these two speech perception tasks. Furthermore, Surprenant and Watson (2001) cite Stankov and Horn's (1980) 'Speech Perception Under Distraction/Distortion (SPUD)' factor, which Stankov and Horn identified in a similar factor analysis on a range of auditory tasks attempting to characterize the structure of auditory capabilities in humans. The tasks

loading on this factor included speech recognition against (i) another talker and (ii) Cafeteria noise, plus recognition of speech that had been temporally expanded or compressed, or spoken in an unusual way. This appears to suggest that ability to recognise speech in noise can be generalized to other speech perception tasks. However, importantly the SPUD factor did not include all the tests of distorted speech perception in Stankov and Horn's battery, and so did not signify a comprehensive speech processing factor. Stankov and Horn suggest that it may, rather, have represented detection of a speech signal against similar speech sounds, or that it may be an index of subjects' tolerance of unpleasant-sounding stimuli. van Rooij, Plomp, and Orlebeke (1989) designed a more comprehensive test battery to investigate factors predicting speech perception performance in the elderly. The battery comprised auditive and cognitive tests at a range of processing levels, from simple pure-tone threshold measurement and frequency selectivity to IQ tests and sentence-picture matching. However, they had also used detection of speech in noise as their measure of speech perception. In their control group of young subjects, they found that the analysis was limited by 'remarkably small' differences in speech perception abilities. It is therefore clear that selection of the right speech perception tasks is important in the test battery approach to investigating individual differences.

By extending the test set used in Surprenant and Watson (2001) to comprise 19 auditory tasks (which now also included a test of environmental sound recognition), Kidd, Watson, and Gygi (2007) have made a more recent attempt to address individual variability in auditory abilities. In a factor analysis, the speech tasks (comprising speech-in-noise recognition tasks for nonwords, words and sentences) loaded most heavily on one factor, which was also loaded upon by the environmental sounds. Kidd et al. (2007) call this the 'Familiar sounds factor', and relate it to the SPUD factor described above. They put forward three possible (non-mutually-exclusive) skills that this factor may represent:

1. Rapidity of access to representations of familiar sounds in the brain e.g. lexical entries for words
2. Problem-solving or guessing strategies to identify a 'whole' from fragmented/impoverished input
3. Attentional strategies using experienced-based knowledge of the most useful spectro-temporal information for identifying, detecting and discriminating familiar sounds

Kidd et al. (2007) acknowledge that the independent factor-loading of the speech and environmental sounds tasks from spectro-temporal tasks in their study does not mean that these lower-level

auditory factors are not important for familiar sound perception, but that they do not account well for the variability in performance of the higher-level tasks. However, Kidd et al. point out that more degraded speech stimuli, such as noise-vocoding, may place heavier demands on lower level auditory sensitivities. An important weakness in the large-scale studies of Surprenant and Watson (2001) and Kidd et al. (2007) is the omission of basic measures of cognitive ability. In both studies, the authors represent general 'intellectual capabilities' using the SAT school examination scores from participants. They then use these scores in correlation analyses with the auditory processing factors. On both occasions, the correlation coefficients between these scores and other factors were weak - however, Kidd et al find a significant (at $p<.05$) relationship between verbal SAT scores and the familiar sounds factor. The authors in both studies suggest that more specific tests of cognitive ability (e.g. memory, attention) could be useful in further characterizing speech (and familiar sound) recognition capabilities.

Another aspect of variability in speech perception is that associated with perceptual learning. Golestani and colleagues have carried out work on the neural correlates of learning a non-native phonetic discrimination in English speakers (Golestani, Paus, & Zatorre, 2002; Golestani & Zatorre, 2004; Golestani, Molko, Dehaene, Lebihan, & Pallier, 2007). In a study of brain structure using MRI and diffusion tensor imaging (DTI) followed by voxel-based morphometry analysis, Golestani et al. (2007) found that listeners who were quicker to learn the unfamiliar discrimination had greater white matter density and overall larger size in left Heschl's gyrus, a larger asymmetry between left and right parietal lobes (where left is larger), and more inferiorly located right insula and Heschl's gyrus, compared with slower learners. In a functional imaging study, behavioural improvement in the non-native discrimination was associated with changing activations in frontal and left parietotemporal speech areas (Golestani & Zatorre, 2004).

The aim of experiments in this thesis will be to carry out studies in an adult population of listeners with healthy hearing, using noise-vocoded speech as a stimulus that will yield a wide range of performance, potentially in both overall recognition scores and in the rate of improvement over time. There are some previous acknowledgements of considerable variability in performance amongst normal-hearing listeners exposed to cochlear implant simulations (Nogaki, Fu, & Galvin, 2007; Stacey & Summerfield, 2007), both in naïve performance levels and the amount of improvement with auditory training. Stacey and Summerfield (2007) measured the relationship between baseline performance and the amount of improvement for sentences, consonants and vowels in three of their experiments. Four of the nine correlations were significant, and indicated that the poorer initial listeners were those who exhibited most learning. This relates to a finding by Amitay et al. (2005) that those listeners initially giving the poorest frequency discrmination thresholds around

1kHz were those who improved most dramatically with more experience of the task. This issue will be addressed in the experiments of this thesis.

If there are systematic auditory, cognitive and/or anatomical factors that can account for the variability in speech recognition performance with noise-vocoded speech, identifying these factors could potentially assist in the design of future experiments employing vocoding. The experiments of the thesis will attempt to identify cognitive and linguistic functions that account for variability, after the suggestions of Watson and colleagues. Chiu, Eng, Strange, Yampolsky, and Waters (2002) have touched upon this area of research by investigating the role of working memory span and reading ability in perception of noise-vocoded speech in normal-hearing adults. They took three measures of working memory span (for digits, letters and sentences) and one measure of reading ability (the American National reading Test; ANART) from normal-hearing adults and correlated scores with recognition of 4-band noise-vocoded speech, finding the strongest positive correlation with reading ability, and the weakest with digit span. Eisenberg, Shannon, Martinez, and A Boothroyd (2000) measured correlations between recognition of 4- and 8-band noise-vocoded speech and total score on the digit span task (both undistorted and vocoded with 8 noise-bands) for a group of normal-hearing children (N=26) and adults (N=10). They found Pearson coefficients of $r<0.30$ for these correlations, suggesting a weak relationship between digit span and speech perception. Therefore, despite indications of an important role for working memory in speech perception from the cochlear implant literature in adults and children, the data from normal-hearing listeners is not conclusive. Experiments in this thesis will address such unresolved issues.

1.4 Summary

This chapter has presented an overview of the use of noise-vocoded speech in the investigation of speech perception, with clinical and non-clinical applications. The growing wealth of literature employing this stimulus distortion technique justifies its thorough investigation as a research tool, in order that it can be used more efficiently in future studies. Two topics of research interest, one stimulus-based and one-listener based, are suggested as themes for new research. Chapter 2 sets out an overview of the experiments carried out along these two lines of investigation, which will be presented and discussed in the remaining chapters of the thesis.

Chapter 2

Overview of the thesis

2.1 Introduction

The experiments contained in this thesis follow two lines of investigation. The first (described in Chapters 3-6) focuses on expanding our current knowledge of the perceptual properties of noise-vocoded speech and their implications for speech recognition and adaptation (Study 1 and Experiments 2-5). The general aim of this set of experiments is to assess the effects of sublexical information, particularly linguistic rhythm, on perception of, and adaptation to, noise-vocoded sentences. Specifically, this formed an investigation of the perceptual properties of linguistic rhythm and timing in the noise-vocoded stimulus, in which fine spectral detail is degraded but temporal information is relatively well preserved.

The second strand of experiments (described in Chapters 7-9) in the thesis use noise-vocoded speech as a tool to investigate listener-based variability in speech perception and adaptation. Individual variability has been well documented in the outcomes of cochlear implantation, but is also becoming a topic of interest for studies with the normal-hearing population. Experiments 2a, 6 and 7 quantify individual variability in recognition of noise-vocoded sentences, and explore cognitive correlates of performance. Experiment 8 investigates variability across different linguistic categories; sentences, words and segments (consonants and vowels).

2.1.1 The role of rhythm and timing

Study 1 is a post-hoc analysis of rhythm in the training materials used in Experiment 5 of Davis et al. (2005). We had noticed that the nonword sentences in our experiment, which gave no training in perception of noise-vocoded sentences, sounded slightly laboured and 'unnatural'. Study 1 uses measures of rhythm described by Ramus et al. (1999), Grabe and Low (2002) and Dellwo et al. (2004) to assess whether rhythmic differences really set the nonword sentences apart from the other training materials used in our previous study.

Experiment 2 employs a training-test paradigm (in which listeners experience an exposure phase before being tested on recognition of noise-vocoded speech) to address two research questions. The first, and more general of the two, re-visits the questions of whether any benefit can be gained from training with noise-vocoded sentences that lack familiar lexical information. The second, more specific question of interest in Experiment 2 is whether the particular timing properties of spoken sentences are important in learning to understand noise-vocoded sentences. These questions are addressed in a between-subjects design, using 5-band noise-vocoded sentences, with four different

training conditions - English, Dutch, Italian and No Training (Control) - followed by test on English materials. While foreign-language materials do not maintain the English phonotactic legality of Davis et al.'s (2005) nonword sentences, the use of recordings from native speakers preserves naturalness in the utterance. Hence, use of foreign language training should offer a test of whether any learning can be achieved from a period of exposure to naturalistic, spoken material with no meaningful content (to the monolingual listener) that has been degraded through noise-vocoding. As Dutch and English are from the same rhythm "class", while Italian is from a different class, a comparison of the training efficacies of Dutch and Italian allows assessment of the importance of linguistic rhythm in adaptation.

Experiment 2 uses native speakers to record all training and test sentences, thus necessitating a speaker change between Training and Test phases of the experiment. Experiment 3 investigates the effects of a speaker change in adaptation to 5-band noise-vocoded sentences, employing the same sentence corpus and procedure as Experiment 1. This experiment uses English sentences only, spoken by the two corresponding voices from Experiment 2. In a simple between-subjects design, the Test Phase sentence recognition performance is compared for listeners who hears two different speakers and those that experience only one speaker in the experiment. In order to assist in the interpretation of Experiment 3, Experiment 4 directly tests the disciminability of the two speakers in a new group of participants.

Experiment 5 adopts an alternative approach to investigating rhythm and timing in perception of noise-vocoded sentences by testing listeners' recognition of noise-vocoded sentences (4-band) in which English lexical, syntactic and semantic content is preserved, and only rhythm is changed. English noise-vocoded sentences are presented in two rhythmic styles, recorded by a single speaker: 'natural' stress-timed rhythm, and an artificial,'metronomic' rhythm that would be unexpected in naturally spoken English. Unlike Experiments 2 and 3, Experiment 5 does not adopt a Training-Test paradigm. It is intended to establish whether or not there were any perceptual differences between the two rhythmic styles, and thus the experiment acts as a precursor for a potential training study. Using a within-subjects design, sentences from each rhythm class are presented in a randomized fashion. Listeners' recognition scores and adaptation trajectories are compared for Natural and Metronomic sentences.

2.1.2 Individual differences

Experiment 2a is so called because it addresses individual variability in speech perception within the noise-vocoded sentence recognition data from Experiment 2. The overall aim of this experiment was to identify possible starting points for more controlled investigation of individual differences in noise-vocoded speech perception in later experiments. The test battery comprises three further measures of speech recognition (speech-in-noise recognition, speech-reading, and written report of undistorted spoken sentences), a measure of complex non-speech auditory processing (amplitude modulation (AM) detection), measures of verbal and performance IQ (WAIS-III Vocabulary and Matrices tests), and a measure of rhythm perception (the Seashore Test of Rhythm Perception). The battery measures are used in correlational and regression analyses with the noise-vocoded sentence recognition scores from Experiment 2.

Experiments 6 and 7 take a more focused approach to measuring individual differences and follow on from the findings of Experiment 2a with respect to possible cognitive correlates of performance. An important change in approach adopted for these experiments is the use of variable intelligibility levels (in terms of the number of noise bands) for the noise-vocoded test sentences, such that variability in performance is no longer quantified in terms of the number of words correctly reported, but rather in terms of the amount of spectral resolution needed for a certain threshold recognition performance.

Experiment 6 adopts an adaptive tracking approach to measurement of sentence recognition thresholds, and the rate of improvement in thresholds associated with adaptation processes, for each listener. The listeners are also tested on two measures of working memory (Digit Span and Nonword Memory Test) and a measure of vocabulary size (British Picture Vocabulary Scale).

Experiment 7 adopts a constant measures approach to threshold measurement for recognition of noise-vocoded sentences in order to facilitate the fitting of logistic performance functions to individual data sets. Approximately 2 months after competing Experiment 6, a subset of the participants are re-tested on recognition of noise-vocoded sentences. This time the items are equally distributed across 10 intelligibility levels and presented in randomized fashion. Thresholds and slopes from performance curves are correlated with scores on the cognitive tasks from Experiment 6, and a new measure of Backward Digit Span. The change in performance from Experiment 6 to Experiment 7 is also assessed for evidence of long-term perceptual learning.

Experiment 8 assesses the effects of systematic variation in linguistic content on individual variability in noise-vocoded speech recognition. A constant measures approach is used to quantify

speech recognition and perceptual learning for High Predictability Sentences, Low Predictability Sentences, Monosyllabic Words, Consonants and Vowels, in five separate recognition tasks. The resulting data set allows several lines of analysis. Individual scores from the five tasks are used in correlational and common factor analyses to characterize the relationships between the different linguistic stimulus types. As both sessions of Experiment 8 employ precisely the same methodology, this enables an analysis of the relative retention of learning/adaptation across the different linguistic stimulus types, at group and individual levels. The data set also offers the opportunity to unpack the low-level aspects speech sound recognition by allowing for an Information Transfer Feature Analysis on consonant and vowel recognition.

2.2 Summary

The following chapters of the thesis present nine experiments designed to investigate the perception of noise-vocoded speech in normal-hearing adults. Study 1 and Experiments 2-5 deal with the role of linguistic rhythm, and other stimulus properties, in noise-vocoded sentence perception. Experiments 2a, 6, 7 and 8 explore variability in noise-vocoded speech recognition and perceptual adaptation, using an individual differences approach to the data in order to characterize the cognitive and acoustic-phonetic processes involved.

Although the experiments of the thesis were designed under two headings, there are of course many instances of overlap between the two sets of experiments, and the evolution of each strand has naturally been informed by developments in the other. For this reason, it is hoped that the thesis should not be read as a description of two separate projects, but rather a theoretically-motivated approach to speech perception as a complex human behaviour emergent from the interaction of the acoustic signal and the listening brain.

Chapter 3

Stimulus properties: Linguistic rhythm in Davis et al. (2005)

Abstract

This chapter addresses the finding by Davis et al. (2005, Experiment 5) that perceptual learning of noise-vocoded sentences cannot be achieved from training with phonotactically-legal nonword sentences, and in turn the conclusion that real word information is necessary in order for significant learning to occur. The BonnTempo package of analysis tools and multi-linguistic speech samples (Dellwo et al., 2004) is used to compare the rhythmic properties of the Davis et al. training materials with those from the traditional 'rhythmic classes'. The results indicate a tendency toward 'syllable-timing' in the Nonword sentences in Davis et al. (2005), in contrast with the more standard 'stress-timing' of the other training conditions. The conclusions of Davis et al. (2005) are reviewed and the challenges of task design and stimulus selection discussed.

3.1 Introduction

3.1.1 The problem with nonsense

As discussed in Chapter 1, the most recent evidence on perceptual adaptation to English noise-vocoded sentences indicates that sentences must contain familiar lexical information in order to be learnable (Davis et al., 2005). In our experiment (Davis et al., 2005), listeners were tested on recognition of twenty 6-band noise-vocoded sentences in English. There were five conditions. Four groups of listeners heard a training phase of twenty noise-vocoded sentences before being tested, while a fifth group experienced no training. We found that training with Nonword Sentences such as *'Cho tekeen garund pid ga summeeun'* produced Test Phase performances that were no better than those exhibited by listeners who received no training. 'Jabberwocky' sentences like *'The tekeen garund to the sumeeun'*, in which real function words had been restored, provided more training but this advantage was not significant in all analyses. Training with Syntactic Prose (e.g. *'The effect supposed to the consumer'*, where the sentences are legal in English but make no apparent sense) and Normal Prose (e.g. normal English sentences such as *'The police returned to the museum'*) produced significantly higher test scores than found in the control group, indicating that these sentence types are effective training materials. The finding of no training with nonwords is at odds with a previous observation with perceptual learning of distorted speech. Altmann and Young (1993) found that recognition of time-compressed sentences is significantly better after training with Jabberwocky sentences than in a control condition with no training. The extent of the training observed with Altman and Young's Jabberwocky sentences (which they called 'nonsense' sentences, but which contained real function words and morphological endings as in the Jabberwocky condition of Davis et al. (2005)) was equivalent to the training obtained with full English sentences. Davis et al.'s finding also sits uneasily with another study on perceptual adaptation to isolated noise-vocoded words (Hervais-Adelman et al., in press), in which training was equivalent with nonword and word training stimuli.

How can the lack of training with nonwords be explained? The interpretation in Davis et al. (2005) was that adaptation is driven by top-down influences from the lexical level of processing. Hence, sentences that contained real words - Jabberwocky, Syntactic Prose and Normal Prose - provided training, but there was no advantage of including higher-level semantic context (i.e. there was no difference between Syntactic Prose and Normal Prose). A possible conflation of added sentential complexity in Davis et al. is the relative loading on working memory in the four training sentence types: Nonword, Jabberwocky, Syntactic Prose and Normal Prose. In the Davis et

al. study we provided feedback in all four training conditions. This took the form of a clear (i.e. undistorted) repetition of each training sentence followed by a repetition of its distorted form. Thus, each trial adopted a *'DCD'* format - *Distorted-Clear-Distorted.* This allowed the listener to make an initial attempt to comprehend the stimulus, before providing knowledge of the sentence content and giving the listener a chance to map these linguistic representations back onto the distorted version. In this paradigm, a nonword sentence containing no real word information is bound to load more heavily on working memory than a regular English sentence with full syntactic and semantic coherence. In Davis et al. (2005), we defended against the possibility that working memory load could have interfered with adaptation by showing that, for the nonword sentences, giving written feedback that stayed onscreen during the distorted repetition of the sentence produced the same result as giving auditory feedback.

The fact that Altmann and Young (1993) found training with time-compressed Jabberwocky that was equivalent to that obtained with full English sentences indicates one or both of two things for adaptation to noise-vocoded sentences as investigated by Davis et al. (2005) et al. (2005): (1) that there is something intrinsically different about the processing of time-compressed and NV sentences and (2) the type of feedback used has differential efficacy across different sentence types, which may be unrelated to memory load (note that Altmann and Young (1993) provided no feedback on their training sentences). The finding from Hervais-Adelman et al. (in press) that noise-vocoded nonwords can provide some training for noise-vocoded word recognition furthermore suggests that, within the perceptual learning of noise-vocoded speech, there is something different about adaptation to sentences and words.

3.1.2 The segmentation challenge

What is the nature of the difference between words and sentences, as applied to perceptual learning of noise-vocoded speech? An initial interpretation could be that a sentence is neither linguistically, nor acoustically, a string of single words, and thus is unlikely to be processed as such - hence the differences in the adaptive properties of these two stimulus types. The observation of no difference between the Syntactic Prose and Normal Prose conditions in Davis et al. (2005) suggest that semantic predictability is not critical to the learning process in noise-vocoded speech.[1] Along with the finding of no learning with nonword sentences, it appears that the listeners in Davis et al. (2005) were indeed learning on the basis of word-to-word mapping of the clear to the distorted

[1]However, inspection of the data from Experiment 5 of Davis et al. suggests that participants in the Normal Prose condition may have reached ceiling during the test phase, thus reducing the difference between their scores and those of the participants in the Syntactic Prose condition.

sentence versions, and not necessarily using overall meaning, in the training phase. Therefore, all training conditions except the nonword sentences (and the Naïve condition) resulted in some learning. So, were the listeners effectively treating the training sentences as a list of words that could be held in working memory during feedback, regardless of semantic predictability? If so, why do the results of Davis et al. (2005) not mirror those of Hervais-Adelman et al. (in press), who found that learning could take place in the absence of meaning?

An important aspect of feedback in Davis et al. (2005) is to consider not just whether the listener could hold feedback information in working memory long enough to map sentence content back onto its vocoded representation, but how this mapping process may have taken place. In other words, despite being able to hold nonword sentences in memory during feedback, there may have been aspects of the stimuli that made it difficult to align the clear and distorted representations on a syllable-by-syllable, or word-by-word, basis. In other words, the nonword stimuli may have been memorable but difficult to segment. This is acknowledged by Hervais-Adelman et al. (in press) in their discussion of the findings of Davis et al. (2005). An aspect of the training sentences in Davis et al. that was not intentionally manipulated was gross syntactic structure. The order and syllabicity of content and function words in each of the Normal Prose sentences was preserved across the corresponding sentences in the other training conditions - by preserving both real word categories (Syntactic Prose), function words only (Jabberwocky) or the overall prosodic contour (Nonword sentences). The preservation of syntactic structure in sentences has two important consequences. First, it enables the extraction of coherent meaning from the sentence, and use of this information in feedback processing. However, the evidence from Davis et al. suggests that coherent sentential meaning had no influence on learning. The second advantage of preserved syntax in the training sentences would have been to cue segmentation of the speech stream i.e. the detection of word onsets and offsets. For example, the function words in a sentence would have cued the listener to the location of content words, enabling word-to-word mapping even when the content words had no obvious meaning, in themselves or in combination. For the Jabberwocky, Syntactic Prose and Normal Prose conditions, the lexical marking of content word onsets (via function word position and morphological endings) would have been assisted by prosodic cues, as an attempt was made to match the prosodic contours of each Normal Prose sentence in its other forms. For example, prosodic features such as vowel reduction, the shortening of function words and the addition of stress and lengthening to content words, would have contributed to the listeners identification of content words as perceptual 'islands' in the speech stream.

In Davis et al. (2005), where the speaker attempted to read each item with the same prosody in each condition, prosodic matching would have been most crucial in the Nonword condition, as in

the absence of real function words the prosody of the sentence provides the main set of segmentation cues. At the time of running the experiment, it was noted that the nonword sentences sounded 'unnatural' in their timings (or rhythm) and overall pitch contours, such that successive syllables sounded more evenly timed than is natural in spoken British English. It is feasible that, due to the pitch-impoverished nature of noise-vocoded stimuli, the timing/rhythmic properties of such stimuli may be perceptually more important than in their undistorted versions, and that disruption of these properties may have contributed to their lack of training efficacy. In Study 1 of this thesis, this proposition is addressed in a post-hoc analysis of the rhythmic properties of the Davis et al. (2005) training sentences. A rhythmic difference between the Nonword sentences and the other conditions would suggest that this factor, rather than the absence of real word information, may have disrupted the potential for learning with noise-vocoded nonword sentences.

3.2 Study 1

3.2.1 Measuring linguistic rhythm

Historically, linguistic rhythm was described in terms of two overall 'classes' - 'stress-timed' and 'syllable-timed' - and languages were classified accordingly (Abercrombie, 1967; Pike, 1945). In the crudest terms, stress-timed languages were deemed to be those with relatively constant durations between successive stressed syllables, whereas syllable-timed languages were described as those in which the durations of successive syllables were relatively constant. Hence, stress-timed languages like English, Dutch and German were thought to have a 'dotted' rhythm, while syllable-timed languages like French, Spanish and Italian were said to have more of a 'machine-gun' rhythm. However, since then evidence has been shown to counter descriptions that focussed on isochrony in speech timing. Several studies found considerable variability in the inter-stress intervals of 'stress-timed' languages, and evidence to suggest that syllable duration is far from constant in 'syllable-timed' languages (Dauer, 1987a, 1987b; Roach, 1982).

More recent descriptions of rhythm classification have turned to *variability* in durations, rather than isochrony. Measurements of rhythm now focus on lower-level durational properties of utterances - namely, the durations of vocalic and consonantal intervals in the speech. Ramus et al. (1999) labelled speech stimuli from a range of languages for vocalic and consonantal (or 'intervocalic') intervals. From these data, the authors extracted two measures, which they claimed best classified languages according to rhythm. These are '%V', which is the percentage of the speech

containing vocalic (vowel) material, and ΔC, which is calculated as the standard deviation of the durations of consonantal intervals in the speech. These were measured by labelling speech samples to identify the vocalic and inter-vocalic (consonantal) intervals. Equations 3.1 and 3.2 shows the equations used to calculate these metrics. In these equations, 'v' and 'c' refer to vocalic and consonantal/intervocalic, respectively.

$$\%V = \frac{100(\sum_{i=1}^{n_v} v_i)}{\sum_{i=1}^{n_{cv}} cv_i}$$

v = v-interval duration

cv = cv-interval duration

n_v = total number of v-interval samples

n_{cv} = total number of c- and v-interval samples

(3.1)

$$\Delta C = 100\sqrt{\frac{n\sum_{i=1}^{n} c_i^2 - (\sum_{i=1}^{n} c_i)^2}{n(n-1)}}$$

c = duration of c-interval

n = total number of sampled c-intervals

(3.2)

Ramus and colleagues saw the $\%V$ as a useful classifying metric as it would set apart the 'stress-timed' languages that featured vowel reduction, like English, from the 'syllable-timed' languages in which this did not feature e.g. French, Spanish. Similarly, the ΔC measure would be greater for languages like English, German and Dutch, which feature complex consonant clusters, and hence a wide range of syllable complexities, than for languages like Italian and Spanish that feature much simpler syllabic structures. In their study of the rhythmic differences between British English and Singapore English (which is thought to exhibit more syllable-timed rhythmic properties), Low, Grabe, and Nolan (2000) proposed an alternative to the measures proposed by Ramus et al. (1999). They put forward a Pairwise Variability Index (PVI) to measure vocalic variability, in which variability was calculated on the basis of changes in vocalic durations between successive

pairs of vocalic intervals. Low et al. claim that this forms a better differentiation between languages according to their perceived rhythmic class. The index is normalised for speech rate, as the authors found that rate affected the PVI across several languages. Equation 3.3 shows the equation used to calculate their normalized Pairwise Variability Index ($nPVI$) based on vocalic interval durations.

$$nPVI = 100\frac{\sum_{v=1}^{n-1}(\frac{x_v - x_{(v+1)}}{(x_v + x_{v+1})/2})}{n-1}$$

n = number of v-intervals sampled

x = duration of v-interval

(3.3)

In a later study, two of the authors of Low et al. (2000) acknowledged that, as for Ramus et al. (1999), a second metric would be needed in classification of language rhythm (Grabe & Low, 2002). This is to account for the fact that some languages, like Polish and Catalan, do not easily fall into one 'class' or the other. For example, in Polish the amount of variability of vocalic intervals is relatively low (in alignment with syllable-timed languages) but the availability of complex consonant clusters in this language results in a high variability of intervocalic intervals. So, Grabe and Low proposed a second Pairwise Variability Index for intervocalic intervals. Grabe and Low elected not to normalize this metric as they believed it was more difficult to tease apart the effects of speech rate on the different segments that may be contained in an intervocalic interval. In contrast, vocalic intervals generally comprise one vowel that becomes longer or shorter with changes in speech rate. The equation for this raw Pairwise Variability Index based on intervocalic durations is shown in 3.4 below.

$$rPVI = \frac{\sum_{c=1}^{n-1}(x_c - x_{(c+1)})}{n-1}$$

n = number of c-intervals sampled

x = duration of c-interval

(3.4)

Grabe and Low (2002) claimed that their $rPVI$ and $nPVI$ measures offer a more sophisticated classification of languages than the approach taken by Ramus et al. (1999). However, in this chapter both pairs of measures are employed in the measurement of linguistic rhythm. Study 1 calculates mean values of %V, ΔC, $nPVI$ and $rPVI$ for each of the four sentence sets (Nonword, Jabberwocky, Syntactic Prose, Normal Prose) used in the training phase of Experiment 5 of Davis et al. (2005). In line with the hypothesis that rhythmic 'naturalness' is important in the adaptation to noise-vocoded speech, it is predicted that the Nonword sentences from Davis et al. (2005) will align with languages traditionally viewed as 'syllable-timed', while the remaining conditions will exhibit rhythmic measures closer to those of traditionally 'stress-timed' English.

3.2.2 Method

Materials

The materials were 80 sentences taken from the training phase stimulus set of Experiment 5 of Davis et al. (2005). The sentences fell into four sets of 20, according to the training conditions Nonword, Jabberwocky, Syntactic Prose and Normal Prose. The Normal Prose sentences comprised two matched blocks of 10 sentences of 6 to 13 words in length (M= 8.7 words), which were equated for mean length in words and duration in seconds, and matched for naturalness and imageability (Rodd, Davis, & Johnsrude, 2005). The Syntactic Prose items were constructed in such a way as to create sentences of syntactic coherence but relatively little semantic coherence. The items were created from the Normal Prose sentences by replacing all content words (nouns, verbs and adjectives) with replacement items of the same word class, with the same number of syllables and similar lexical frequency. The Jabberwocky sentences were created by replacing all content words with phonotactically legal nonwords of the same number of syllables, while leaving the function words (including pronouns and adverbs) in their original forms from the Normal Prose sentences. Finally, the Nonword sentences were created by replacing all words in the Normal Prose sentences with phonotactically legal nonwords of the same number of syllables. Below are written examples of the four sentence types. It should be pointed out that, for each training item, the sentences of the Nonword and Jabberwocky conditions shared the same nonword replacements for the content words. An audio example of each sentence type is available on the CD accompanying this thesis.

Normal Prose	*The police returned to the museum*
Syntactic Prose	*The effect supposed to the consumer*
Jabberwocky	*The tekeen garund to the sumeeun*
Nonword Sentence	*Cho tekeen garund pid ga sumeeun*

In Davis et al. (2005), every attempt was made to record all four versions of each test sentence with a similar overall rate and prosodic contour. However, as noted in the Introduction, this was difficult to achieve and was (particularly for the Nonword Sentences) detectable perceptually, hence the motivation for the current study. However, it is also worth mentioning that the sentence sets differed significantly in duration (Mean durations - Normal Prose: 2.0secs, Jabberwocky: 2.3secs, Syntactic Prose: 2.4secs, Nonword Sentences: 2.6secs). As mentioned above, the rate of speech can have an effect on measures of durational variability. This issue will be addressed in the Results section.

Labelling sentences

The sentences were concatenated into blocks of twenty, corresponding to each condition of the Davis et al. (2005) experiment. For each block, a two-tier textgrid in PRAAT (Boersma & Weenink, 2005) was used in conjunction with a spectrogram of the speech to label the onset and offset timepoints of consonantal and vocalic intervals, and the onsets and offsets of syllables. Figure 3.1 shows an example of a labelled sentence in PRAAT - note that pauses in the utterances were also labelled. A vocalic interval was taken to be the period between the offset of one consonant (or group of consonants/non-vowels) and the onset of the next consonant (or group of consonants/non-vowels). A consonantal interval was seen as the interval between the offset of one vowel and the onset of the next vowel (i.e. reflecting the duration of the intervening consonants), and included semivowels and glides. Any pauses were labelled as such; these pause intervals often split vocalic or consonantal intervals.

The criteria for labelling were a collection of phonological and acoustic standards, grounded in the intuitive interpretation of the sounds and the spectrogram. Generally, if a given interval was expected to be vocalic (according to the sentence's written version), and there was sufficient auditory and spectrogram evidence to support this, it was labelled as a vocalic interval. The same was true for consonantal intervals. One labeller was responsible for labelling all the materials from Davis et al. (2005) used in the current study, therefore these should be labelled sufficiently consistently to assess the materials' relationship to each other along measures of linguistic rhythm.

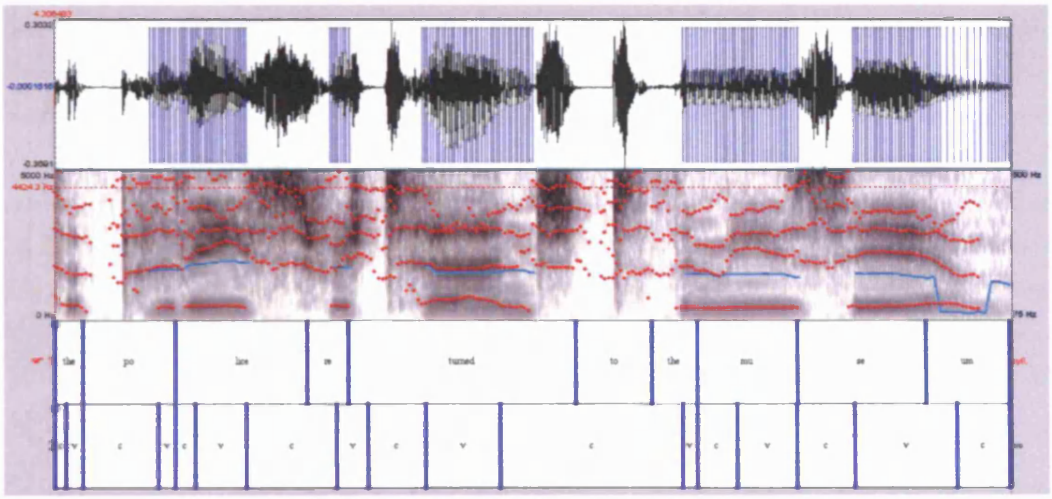


Figure 3.1: An example of a labelled sentence in PRAAT. The first 'tier' beneath the spectrogram shows labels for the onsets and offsets of syllables; the second tier corresponds to labelled consonantal (c) and vocalic (v) intervals.

For the purposes of statistical comparisons between the different sentence sets, the sentence blocks described above, with their corresponding labels, were also divided into individual labelled sentences. Each of these was then analysed for several rhythmic metric values, as described below.

Analysing Speech Rhythm

The BonnTempo Tools and Corpus

The four rhythm metrics - %V, ΔC, $nPVI$ and $rPVI$ - were calculated according to the equations shown in Equations 3.1 to 3.4 above, with the use of a set of analysis tools developed by Dellwo and colleagues at the University of Bonn (Dellwo et al., 2004). Pause intervals in the labelled files were not entered into the analyses. However, where a pause occurred, the speech samples on either side of the pause interval were treated as part of different phrases. Pauses most often occurred at the ends of sentences, but also occasionally within sentences. The BonnTempo tools are programmed within PRAAT, and include both a direct means of analysing labelled text files and a graphical interface for display of results. Furthermore, the BonnTempoCorpus of recorded spoken material forms one of the largest databases available for speech rhythm analysis (Dellwo et al., 2004), with recordings of a large number of native speakers of English, Czech, French, German and Italian,

and of L2 (non-native) speakers of English, French and German (e.g. German speaking English, French speaking German). Figure 3.2 below shows a plot of the Ramus et al. (1999) metrics, %V and ΔC, for four of the core corpus languages in the BonnTempo Corpus (BTC). The languages are represented in different colours, labelled with the language and speaker abbreviation - 'Dd' corresponds to 'German (Deutsch) speaking German', 'Ee' to 'English speaking English', 'Ii' to 'Italian speaking Italian' and 'Ff' to 'French speaking French'. The plot shows a clear separation of the languages along the traditional rhythm 'classes'. In the default display setting, the languages are represented by 5 data points, corresponding to the mean values at five 'intended' rates of articulation - 'very slow', 'slow', 'normal', 'fast' and 'very fast'. The data points are joined in order of increasing speed, with the language label (e.g. 'Dd') next to the slowest rate.

In the current experiment, each of the four conditions was analysed with only one speech rate (the rate at which the sentences were originally recorded for Davis et al. (2005)).

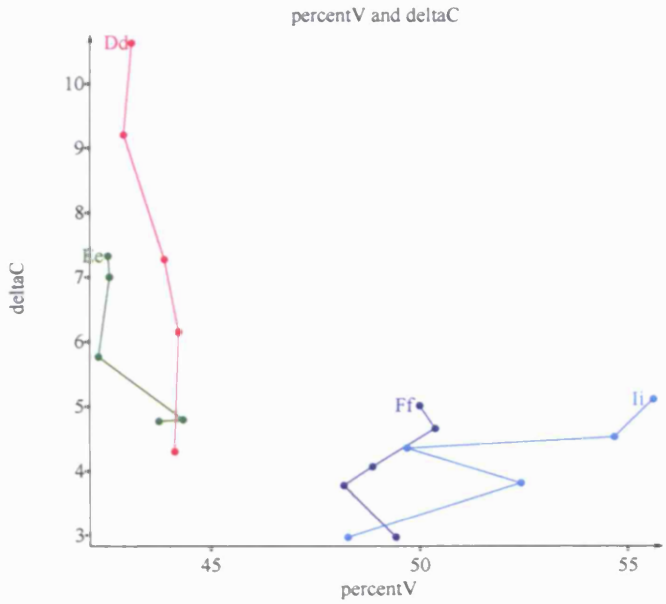


Figure 3.2: The rhythmic properties of the BTC languages as plotted in PRAAT, using the Ramus et al.(1999) measures.

3.2.3 Results and Discussion

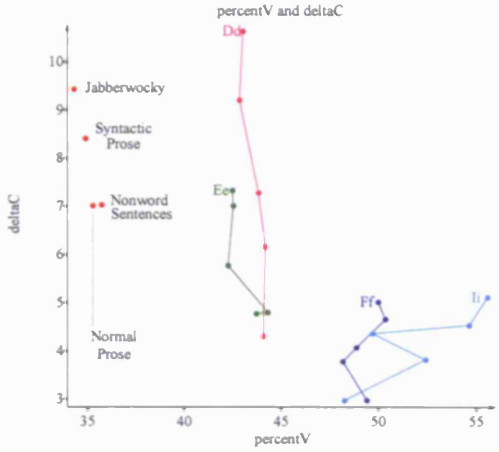
The mean values for the four sentence sets (in 20-sentence block form) were plotted alongside the BTC materials in plots of ΔC versus %V, and of $rPVI$ versus $nPVI$. Table 3.1 presents the mean values for each metric in each condition, and the respective plots are shown in Figures 3.3(a) and 3.3(b).

Table 3.1: Mean values from a blocked analysis of the four sentence conditions.

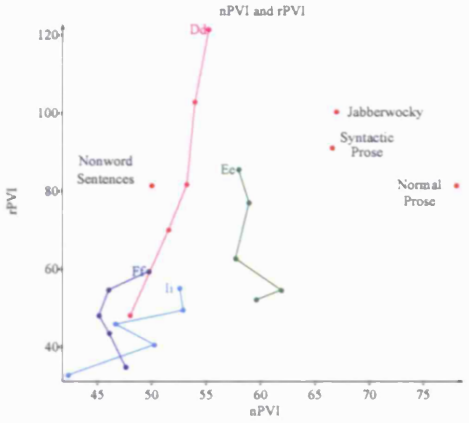
	%V	ΔC	nPVI	rPVI
Normal Prose	35.3	7.0	78.1	81.4
Syntactic Prose	34.9	8.4	66.7	91.1
Jabberwocky	34.3	9.4	67.1	100.4
Nonword Sentence	35.8	7.0	50.0	81.4

The plot of ΔC vs %V (after Ramus et al. (1999)) indicates that there is little difference between the four conditions in terms of the %V metric. In all four conditions, the value for %V is much lower than those observed for English and German (the 'stress-timed' languages) in the BonnTempo Corpus, indicating more 'extreme' stress-timing. This may be due to differences between the Davis et al. (2005) materials and the spoken passage used by Dellwo and colleagues for the BonnTempoCorpus. In contrast, there is some separation between the four sentence sets along the ΔC scale. However, on this scale, the Nonword Sentences and Normal Prose sentences are equivalent, with the Jabberwocky and Syntactic Prose sentences exhibiting higher levels of consonantal variability. This finding goes against the hypothesis that the Nonword Sentences would exhibit more 'syllable-timed' properties in contrast to the other conditions, as this should have produced a significant difference between the Nonsense and Normal Prose conditions.

The plot of $rPVI$ against $nPVI$, preferred by Grabe and Low, presents a clearer picture of the results. According to the $nPVI$ metric, which is a rate-normalised measure of variability in vocalic intervals, the Nonword Sentences are aligned with the more syllable-timed languages of French and Italian. As was the case for the %V measurements, the other three conditions (Jabberwocky, Syntactic Prose, Normal Prose) exhibit 'extreme' stress-timed values for nPVI. This is the first piece of supportive evidence that there is a rhythmically based difference between the Nonword Sentences and the other training conditions in Davis et al. (2005). However, similarly to the ΔC values, the rPVI measure of consonantal variability shows equivalent values for the Nonsense Sentences and Normal Prose. These values are in the region of the English values from the BonnTempo corpus,



(a) Using Ramus et al. (1999) measures



(b) Using Grabe and Low (2002) measures

Figure 3.3: Rhythmic properties of the Davis et al. (2005) materials plotted against BonnTempoCorpus languages.

while the Jabberwocky and Syntactic Prose sentences show levels of variability that are higher than the means observed from the BonnTempo speakers.

For the purposes of statistical comparison between the four conditions, individual values were calculated for each sentence. Table 3.2 shows the mean values, by condition, obtained through this approach. The values for %V and ΔC are much like those shown in Table 3.1 above. However, the values of nPVI and rPVI, whilst patterning similarly, are numerically quite different from those obtained from the blocked analysis. This is likely to be a result of the greater effect on the PVI metric of chopping the speech sample up into shorter portions, as this measures is calcuated on the basis of successive interval pairs. Values for %V, ΔC, $nPVI$ and $rPVI$ were compared in a set of repeated-measures ANOVA analyses. Repeated-measures analyses were used as the four conditions are related, by design and utterance, on an item-by-item basis. The analyses produced a significant effect of Condition for ΔC ($F(3, 57) = 4.38$, $p = .008$, $\eta^2 = .187$, power = .849), $nPVI$ ($F(3, 57) = 10.43$, $p = .000$, $\eta^2 = .354$, power = .998), and $rPVI$ (Wilks' Lambda $F(3, 17) = 9.20$, $p = .001$, $\eta^2 = .619$, power = .984). Sidak-corrected post-hoc comparisons were significant between Jabberwocky and Normal Prose sentences for ΔC ($p = .027$); for $nPVI$, there were significant comparisons between Nonword Sentences and all three of the other conditions (Jabberwocky: $p = .004$; Normal Prose: $p = .000$; Syntactic Prose: $p = .043$). While the results for $nPVI$ are in line with the predictions of a difference in the Nonword Sentences, the ΔC and $rPVI$ values are not quite in alignment, as these assign the Nonword and Normal Prose sentences similar values when these are predicted to be most different (along the 'naturalness' hypothesis).

It would be difficult to obtain values for consonantal variability that lie in the range of the classic 'syllable-timed' languages from test materials that are based on English phonotactics. The reason for this lies in the range of syllable complexities available in English, which are most strongly pronounced in the available consonant clusters (the classic example of consonantal complexity in English being the word *'strengths'*). This will automatically result in quite high consonantal variability scores for any English-based stimuli. However, another variable which can interact with durational variability measures in speech is the rate of articulation. It is intuitive that, if the speech rate is high, and all intervals are consequently reduced, this produces a reduced range in interval durations and hence lower variability. This has been documented by Dellwo and colleagues in the relationship between ΔC and rate (Dellwo & Wagner, 2003), and the relation of $rPVI$ to rate (Dellwo, 2007). In both cases, the variability is reduced at higher speech rates. In Davis et al. (2005), there were small differences in overall duration across the four sentence sets (see Method), despite the authors' attempts to control for speech rate in their recordings. This resulted in the Nonword Sentences producing the slowest speech (average sentences duration: 2.6 seconds) while the Normal Prose sentences were the fastest (average duration: 2.0 seconds). This may have had implications for the findings above. For example, in the comparison between the four conditions in terms of $rPVI$ and ΔC values, the possibility that the Nonword Sentences exhibit

lower consonantal variability (and hence are more 'syllable-timed' than the other conditions) may have been masked by the fact that these sentences were also produced at the slowest rate. Similarly, a fast rate of speech in the Normal Prose condition may have produced relatively lower consonantal variability than predicted for this sentence set.

Table 3.2: Mean values from an items-based analysis of the four sentence conditions.

	%V	ΔC	$nPVI$	$rPVI$
Normal Prose	35.1	7.0	105.1	75.1
Syntactic Prose	35.0	8.4	99.2	90.5
Jabberwocky	34.5	9.1	95.9	94.6
Nonword Sentence	35.8	7.0	82.3	76.2

The possible effect of rate on rhythmic measures was investigated further. A measure of speech rate was extracted for each condition, by block (of 20 concatenated sentences), and by individual sentence. Speech rate was measured as the laboratory speech rate for CV intervals, where 'CV' refers to all consonantal and vocalic intervals from the labelled files. The mean block values, in units of intervals per second, obtained for the four training conditions from Davis et al. (2005), were 8.679 (Nonword Sentences), 8.501 (Jabberwocky), 8.867 (Syntactic Prose) and 10.179 (Normal Prose). There are five different 'intended' speech rates adopted by the speakers in the BTC - very slow, slow, normal, fast and very fast. For comparison, the mean English speech rates in the BTC (in intervals/second) are 8.727 (very slow), 9.501 (slow), 10.756 (normal), 11.636 (fast) and 13.183 (very fast). Thus, the Normal Prose condition in Davis et al. (2005) was of a rate intermediate between the BTC 'slow' and 'normal', while the rates for the other three conditions were close to the 'very slow' BTC rate. A repeated-measures ANOVA was conducted for speech rate on the individual items across the four conditions of Davis et al. (2005). The mean values, when the conditions were analysed by sentence, were 8.532 intervals/second for Jabberwocky, 8.787 for Nonsense Sentences, 9.021 for Syntactic Prose and 10.228 for Normal Prose. The effect of Condition in a repeated-measures ANOVA on these measures was significant ($F(3, 57) = 10.81$, $p = .000$). Post-hoc pairwise comparisons, with Sidak correction, showed that the Normal Prose sentences were spoken at a significantly faster rate than each of the other conditions (Jabberwocky: $p = .000$; Nonsense Sentences: $p = .001$; Syntactic Prose: $p = .001$). Thus, there is an indication that the unexpected match in consonantal variability observed between Normal Prose and Nonword Sentences above was not due to particular slowness of the Nonword Sentences, but rather the increased speed of utterance of the Normal Prose sentences.

In order to re-assess the rhythmic properties of the Davis et al. (2005) materials without the conflation of rate effects, a new, rate-normalised version of the rPVI index was calculated. Although this was warned against by Grabe and Low (2002), Dellwo (2007) found that the normalized version ('$rPVInorm$'), along with %V, provides the clearest distinction between languages of different rhythmic classes.[2]. Equation 3.5 below shows the equation used to calculate the rate-normalized rPVI for consonantal variability.

$$rPVInorm = 100\frac{\sum_{c=1}^{n-1}\left(\frac{x_c - x_{(c+1)}}{(x_c + x_{c+1})/2}\right)}{n-1} \tag{3.5}$$

n = number of c-intervals sampled

x = c-interval duration

Figure 3.4 shows a new plot of the four training conditions in Davis et al. (2005) against the results of the BonnTempoCorpus (for English, Dutch, French and Italian) for the vocalic nPVI and the consonantal rPVInorm. This plot shows a much clearer distinction between the Nonword Sentences and the other three conditions of the Davis et al. experiment, with the mean nPVI and rPVInorm values for Nonword Sentences lying deep within the range of the classic 'syllable-timed' languages of Italian and French. Again, a repeated-measures ANOVA was run with rPVInorm as the dependent measure and Condition as the within-subjects factor, using rPVInorm values from the 20 individual sentences in each condition. The effect of Condition was non-significant ($F(3,57) = 1.85$, $p = .149$, $\eta^2 = .089$, power = .454), as were the post-hoc, Sidak-corrected pairwise comparisons between conditions. However, the emergent mean values for each Condition show a much more readily interpretable pattern, aligning in the order Nonword Sentences ($rPVInorm = 55.276$), Jabberwocky (60.701), Normal Prose (61.804), Syntactic Prose (64.376). Thus, the data show that the sentences that are likely to have proven more difficult to read (Nonword and Jabberwocky) are those showing consonantal variability more in the range of

[2]N.B. Dellwo (2007) also puts forward a rate-normalized alternative to the ΔC metric, in which the durations of the consonantal intervals undergo a natural log transformation before being entered into Equation 1.2. However, as the $nPVI$ metric has so far provided the most sensible (and already rate normalized) description of the Davis et al. materials, it is more convenient to re-measure consonantal variability with a metric based on the consonantal counterpart of $nPVI$

the syllable-timed languages (moreso for the Nonword Sentences). In the blocked-sentence analysis, the patterning of increasing $rPVInorm$ values follows the pattern of test phase sentence recognition scores obtained in the original Davis et al. (2005) noise-vocoded speech perception study (which increased in the order Nonword Sentences, Jabberwocky, Syntactic Prose, Normal Prose). For the $nPVI$ mean values, the means based on individual sentences pattern with the intelligibility results, while this only applies for the Nonword and Normal Prose conditions in the pattern of blocked-sentence means (lowest and highest $nPVI$ and intelligibility scores, respectively).

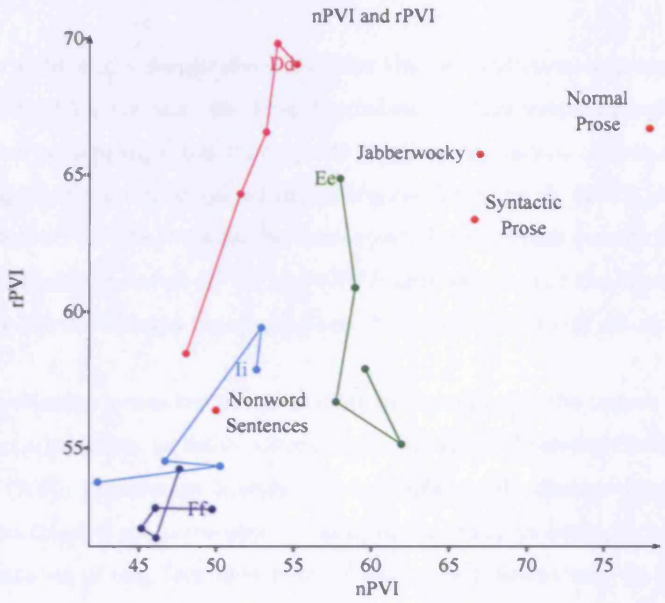


Figure 3.4: Plot featuring rate-normalized values for the intervocalic rPVI.

3.3 Summary

Overall, this analysis of the rhythmic properties of the sentences in Davis et al. (2005) strongly indicates that there are rhythmic differences between the four sets of sentences used as training conditions in Experiment 5 of their study. Initial assessments and analyses using recently established approaches to linguistic rhythm measurement produced mixed results. On the %V scale, there was little separation of the conditions, and the results for ΔC and $rPVI$ produced differences that were difficult to interpret - particularly the close similarity between these metrics for

the Nonsense Sentences and Normal Prose conditions, which were predicted to exhibit the greatest difference in rhythmic 'naturalness'. In contrast, the rate-normalized measure of vocalic interval variability, $nPVI$, showed a significant differrence between the Nonword sentences and the other three sentence sets. A subsequent analysis of speech rates in the training conditions from Davis et al. (2005) showed significant differences between the Nonword Prose sentences and the other conditions, and inspired use of a rate-normalized version of the consonantal PVI, the $rPVInorm$, as advised by Dellwo (2007). Using these two metrics ($nPVI$ and $rPVInorm$), there was a clear separation between the Nonword Sentences and the other three sentence sets, which was supported numerically rather than statistically for the $rPVInorm$ metric.

There was some slight disagreement between the two analytical approaches to the data in the current experiment i.e. the analysis of the conditions by block versus by individual sentence. However, the overall patterning of results supports positive correlations of both $nPVI$ and $rPVInorm$ with the intelligibility scores obtained in the original Davis et al. (2005) study. This relationship is most pronounced for the Nonword Sentences and Normal Prose conditions, where the Nonword Sentences Exhibited the lowest $nPVI$ and $rPVInorm$ values, and the lowest recognition scores in Davis et al., while the Normal Prose exhibited the highest scores on all counts.

Thus, the evidence presented in the current study supports the notion that there is a role for rhythm in the adaptation to noise-vocoded speech, as the Nonword Sentences condition of the Davis et al. (2005) experiment is shown to be rhythmically distinct from the other three, and is the only condition of the noise-vocoded adaptation study to have produced no learning. With the current data set of only four data points (one for each condition), we cannot clearly establish whether there is a significant linear relationship between rhythm and training efficacy. However, the observations in Davis et al. that the group exposed to Jabberwocky sentences exhibited a training efficacy that was not equivocally superior to naïve or Nonword trained groups, while the group trained with Syntactic Prose produced test results that were statistically indistinguishable from the group trained with Normal Prose, suggests that rhythm cannot account for all of the variability in the data. Indeed, Davis et al.'s conclusion that lexical information is involved in adaptation need not necessarily be ruled out by the findings in this chapter. However, the fact that there may be some role for sublexical phonological information in the adaptation to noise-vocoded speech brings us back to the opening question of this chapter - can adaptation take place in the absence of real word information? In other words, while Davis et al. showed that lexical information is certainly *useful* for adaptation, is it necessary?

Having shown the potential importance of rhythmic naturalness in training stimuli in this

chapter, and having observed that achieving naturalness is a difficult task with unfamiliar materials (such as nonword sentences in Davis et al. (2005), we must look to alternative materials with which to test the potential roles of sublexical information in adaptation to noise-vocoded speech. The most appealing solution would be to train listeners with foreign language stimuli, where we can obtain naturalistic recordings of sentences that have little to no semantic or lexical familiarity to the listeners, and assess their training efficacy in comparison to meaningful sentences in the listeners' native language. A further dimension to this study, picking up on the potential role for rhythm uncovered in the current experiment, would be to compare the training efficacies of a selection of unfamiliar foreign languages varying in their rhythmic similarity of the listeners' native language. For example, Dutch and German come from the same historical rhythm class as English, while French and Italian belong to the 'syllable-timed' category of languages. For test performance in English, a role for rhythm in the adaptation to noise-vocoded speech would predict more effective training from stress-timed languages than from syllable-timed languages. This design is employed in Experiment 2 of this thesis, which is described in Chapter 4.

Chapter 4

Stimulus properties: A cross-linguistic study of perceptual adaptation

Abstract

The following chapter addresses the question of whether perceptual adaptation to noise-vocoded sentences can occur in the absence of lexical information. Recognition scores on 10 English noise-vocoded sentences were compared for monolingual English listeners trained with noise-vocoded sentences in Italian, Dutch and English. A Control group experienced no training. The cross-linguistic design also allowed investigation of whether listeners used linguistic rhythm as an adaptation cue. This was tested by comparing the training efficacy of the Italian training items (syllable-timed) with that of Dutch and English (both stress-timed). The results indicate that the Italian and Dutch groups gave slightly lower recognition scores than the Control group, and that only English gave effective training. The findings are discussed with reference to issues of task procedure, stimulus properties and individual variability.

4.1 Introduction

The results of Study 1 indicate that a lack of rhythmic 'naturalness' may indeed be implicated in the lack of learning from Nonword Sentences in Davis et al. (2005). However, there is still sufficient evidence from Davis et al. (2005) to suggest that the presence of real words in the training sentences is certainly an important factor affecting the amount of learning that can be achieved. The question is whether we can conclude that there is nothing to be learned from nonword stimuli, even when their rhythmic properties are more naturally aligned with what is expected from English. To investigate this, we turn to foreign languages as naturalistic stimuli that can lack meaning for the non-native listener.

The use of foreign languages in perceptual learning studies

The potential for perceptual learning from foreign languages has been developed in a series of experiments using time-compressed speech (Altmann & Young, 1993; Mehler et al., 1993; Pallier et al., 1998; Sebastian-Galles et al., 2000). Time-compressed speech is constructed by removing pitch periods from the original speech stimulus, such that the speech is made faster without affecting the overall pitch or the relative durational properties of the consonantal and vocalic intervals. Using this approach, speech samples can be prepared that exhibit apparent speech rates that would be extremely difficult to achieve organically. However, in other respects the speech is natural in quality. Such stimuli challenge the speech perception system, but recognition performance can improve over a relatively short time-frame (i.e. a few sentences) - this has been demonstrated for listeners of Spanish, Catalan, French and English (Mehler et al., 1993).

Interest in adopting cross-linguistic paradigms emerged from the attempt to identify a role for sublexical phonological information in the adaptation to time-compressed speech. A study by Altmann and Young (1993) had shown that English-speaking listeners given a period of exposure to time-compressed 'Nonsense Sentences' (equivalent to the Jabberwocky described in Chapter 3 of this thesis) produced test phase recognition scores with English time-compressed sentences that were equivalent to those obtained after training with English. Furthermore, Mehler et al. (1993) had shown that monolingual Spanish speakers could receive training from time-compressed Catalan sentences that they didn't understand. Thus, there was evidence that higher-level meaning is not necessary for adaptation. However, other studies provided evidence that learning wasn't simply the use of low-level acoustics, either. Dupoux and Green (1997) conducted an adaptation study using English time-compressed sentences with English-speaking listeners, in which they introduced

an abrupt change in speaker after the first 10 sentences. Although there was a small numerical advantage for listeners who did not experience the speaker change over those who did, this only approached significance. Further evidence came from cross-linguistic studies with both monolingual and bilingual speakers of English and French, where it was shown that learning could not be transferred from one language to the other for time-compressed sentences (Altmann & Young, 1993; Mehler et al., 1993; Pallier et al., 1998). Thus, even when understanding is present, there is transfer between some languages but not others. Sebastian-Galles et al. (2000) point out that the lexical and phonological differences between French and English (which do not show transfer) are greater than those between Spanish and Catalan (which do show transfer). Therefore, in the case of bilingual speakers of French and English, there may be more of a tendency to adopt a different 'listening modes' for each language, hence reducing the possible transfer between the two. The greater interest, Sebastian-Galles et al. (2000) claim, is in cross-linguistic studies with monolingual listeners.

Two studies that pave the way for the experiment in this Chapter are those by Pallier et al. (1998) and Sebastian-Galles et al. (2000). Pallier et al. showed, in the first three experiments of their study, that adaptation to time-compressed speech could be transferred between Spanish and Catalan, in bilinguals (Expt 1) and monolingual Spanish speakers (Expt 2), but that English-French bilinguals received no transfer of adaptation between these two languages (Expt 3). This led the authors to posit that phonological information provided the locus for learning, pointing out that French and English are phonologically quite different while Spanish and Catalan are, relatively, much more similar. Experiment 4 of their study examined the question of phonological similarity across languages. Three groups of monolingual English listeners experienced a 'habituation phase' of 10 sentences, compressed to 50% of their original duration, followed by test on 5 sentences compressed to 40% of their original duration. The groups heard either French, Dutch or English in the habituation phase, followed by test on English sentences. A fourth group of listeners, in the control condition, had no habituation and went straight into the test phase. Pallier et al. (1998) found that habituation with English produced the greatest amount of adaptation, while performance in the French group was numerically (but not significantly) worse than control. Training with Dutch, however, produced an intermediate level of adaptation that, in post-hoc comparisons, was shown to be marginally greater than French and marginally less than English. The study concluded that 'pre-lexical' processing is involved in speech processing, and that it is similarities in pre-lexical representations that facilitate transfer of adaptation with time-compressed speech. A possible explanation to account for the patterning of adaptation transfers, based in phonological properties of speech, is linguistic *rhythm*. When we consider the rhythm class to which each of the languages in Pallier et al. (1998) belongs, we find that Spanish, Catalan and

French are syllable-timed, while Dutch and English are syllable-timed. Hence, the non-transferring language pair (English and French) may not have exhibited transfer of adaptation due to the fact that the languages in the pair come from different rhythmic classes.

The theory that listeners use strategies based on their own language to segment the speech stream was not new at the time of these cross-linguistic studies with time-compressed speech. Cutler and Mehler (1993) proposed, on the basis of the developmental literature and a programme of experiments on speech segmentation in adults, that French listeners use the syllable as the segmentation unit, while English listeners use stress patterns and Japanese listeners use the mora (a sub-syllabic speech unit). With a rhythm-class based approach in mind, Sebastian-Galles et al. (2000) built upon the findings of Pallier et al. (1998). They point out that a weakness of Experiment 4 of the Pallier et al. (1998) study is that the similarity between Dutch and English is not limited to sub-lexical phonology. As both languages are Germanic, there is considerable lexical overlap between them, which exceeds that between French and English (which would arise through borrowings from Romance languages in English). Thus, the learning effect for Dutch may have been driven by lexical information only. Sebastian-Galles et al. identify a neat means of overcoming this design flaw in their study of adaptation to time-compressed sentences in monolingual Spanish listeners. In Experiment 2 of their study, they identify Greek as a habituation language from the same rhythmic class as Spanish (the native language of their participants), but which bears little lexical overlap, as Greek belongs to the Hellenic languages whereas Spanish is a Romance language. In this experiment, they found that test performance on compressed Spanish sentences was just as good after habituation with compressed Greek sentences as it was after habituation with compressed Spanish. Thus, it seems that lexical overlap is not of primary importance in cross-linguistic transfer of adaptation to compressed sentences.

However, there remain exceptions to this putative rule. In Experiment 1 of their study, Sebastian-Galles et al. (2000) compared transfer of adaptation to Spanish from a selection of syllable-timed languages (Spanish, Italian, French), English (stress-timed) and Japanese (mora-timed). While they found no transfer from English and Japanese, as expected, and significant transfer from Italian and Spanish, there was no transfer from French. The authors posit that there are properties of French which set it further apart from other syllable-timed languages, particularly in how it is segmented. French has fixed stress (on the last syllable of all content words) while the other languages in Experiment 1 exhibit variable stress patterns. A 'deafness' to lexical stress in French listeners has been documented (Dupoux, Pallier, Sebastian, & Mehler, 1997). Furthermore, French has a larger vowel set size than the other syllable-timed languages in Sebastian-Galles et al.. Hence, the authors warn against a simplistic rhythm-based interpretation of their results, yet they

still acknowledge that rhythm plays a role alongside other, as yet uncharacterized, phonological properties.

Timing in noise-vocoded speech and implications for cross-linguistic studies

As described in Chapter 1, noise-vocoding is a process which degrades fine spectral detail in speech while preserving its temporal properties (Shannon et al., 1995). The resulting stimulus is impoverished in terms of pitch but with an intact envelope. In the cochlear implant literature, several studies have pointed toward an importance of rhythm when listening to and producing music (Gfeller & Lansing, 1992; Kong et al., 2004; Nakata et al., 2006; Meister et al., 2007). Furthermore, the outcomes of the post-hoc analysis in Study 1 of this thesis suggest that rhythmic factors may been involved in the adaptation to noise-vocoded sentences in Davis et al. (2005).

The aim of the current study is to re-visit the question of whether perceptual adaptation to noise-vocoded speech can take place in the absence of understanding. Specifically, the intention is to assess the role of linguistic rhythm in sub-lexical adaptation processes, if these are present. The approach taken by Pallier et al. (1998) and Sebastian-Galles et al. (2000) presents a promising means of addressing these questions. Furthermore, due to the spectrally-impoverished nature of noise-vocoded speech, it is posited that some of the other phonological factors, such as vowel space and set size, should not pose as great a conflating effect as might have been possible in the studies with time-compressed speech.

The current experiment will take the form of a near-replication of Experiment 4 of Pallier et al. (1998). Native English speakers are divided into four groups: three receive a pre-exposure phase of 10 noise-vocoded sentences before test on 10 noise-vocoded English sentences, while a fourth group receives the test without pre-exposure. The three pre-exposure languages (one per group) are English (native stress-timed), Dutch (foreign stress-timed) and Italian (foreign syllable-timed). Italian is included rather than French due to the concerns raised in Sebastian-Galles et al. (2000) regarding this language. It is predicted that English will produce the greatest amount of adaptation (as represented by Test Phase speech recognition scores), with Dutch providing an intermediate training and French giving no training benefit over the control condition (the group receiving no pre-exposure).

4.2 Experiment 2

4.2.1 Method

Participants

Sixty-four monolingual speakers of English (aged 18-40, 19 male), with no hearing, speech or language problems, took part in the experiment. Participants were recruited from the UCL Department of Psychology Subject Pool and the Institute of Cognitive Neuroscience participant database.

Materials

Questionnaire

A questionnaire (see Appendix A) was constructed to assess participants' experience and competency in foreign languages and music. Its primary purpose was to filter listeners' into the appropriate condition of the sentence recognition task, as listeners with considerable proficiency in either of the forsign languages used would violate the requirements of monolingualism. The questionnaire comprised several questions, including free response items such as: *Do you speak any other languages?* and multiple choice questions including: *Can you sing a familiar melody without accompaniment?* with response options *Yes*, *No* and *Not sure*. The questionnaire was also used to collect information on the participant's date of birth and regional accent of English.

Speech Perception Task

All sentences used in the experiment came from the LSCP multilingual corpus (compiled by Nazzi et al. (1998), which was previously used in Pallier et al. (1998) and donated in written form by Christophe Pallier for use in the current study. There were four sentence sets: Italian, Dutch, English Training and English Test. Each set comprised 2 sentences at each of 16, 17, 18, 19 and 20 syllables in length. An example sentence from the English set in this corpus is *'The committee will meet this afternoon for a special debate"*. As the training efficacies of the Italian, Dutch and English Training sentences were being directly compared in the experiment, these sets were most closely matched in item selection. Table 4.1 shows the length (in words and syllables) and the

mean durations (in seconds) of the four sentence sets. The experimental materials were recorded in a soundproof, anechoic chamber. Recordings were made on a Digital Audio Tape recorder (Sony 60ES) and fed to the S/PDIF digital input of an M-Audio Delta 66 PC soundcard. The files were then downsampled at a rate of 44100Hz to mono .wav files with 16-bit resolution using Cool Edit 96 software (Syntrillium Software Corporation, USA). The recordings were divided into a separate .wav file for each sentence. The English Training and Test sentences were recorded by two female native speakers of Standard Southern British English (SSBE). The Dutch and Italian sentences were recorded by female, native speakers of these languages who were recruited from the University of London research community. An example sentence from each of the training conditions can be found on the CD accompanying the thesis.

An extra set of five English sentences, to be used for task habituation in the speech perception task, was recorded by a female speaker of English from Northern Ireland. The reasoning behind using several English speakers was as follows. In the foreign language conditions, it was thought preferable to use monolingual native speakers rather than Italian-English and Dutch-English bilinguals. On a practical level, finding and recruiting speakers with the relevant bilingual status for this study would have proven much more difficult than finding monolingual native speakers. Even with accomplished non-native speakers of English, it would be difficult to control for the relative L1 (first language) and L2 (second language) proficiencies between speakers. A particular danger would be that, in non-native speakers, there could be 'pollution' of the L2 speech output with the the rhythmic properties of the L1, which would have proven difficult for the design of the current experiment. Therefore, it was decided that the English sentences for use in the English Test phase should be spoken by a native speaker. As this introduced a necessary speaker change in the Dutch and Italian conditions, this was balanced in the English condition by including a different speaker for the English Training sentences. The inclusion of the speaker change across all conditions also meant that the same English speaker could be used in all four conditions of the experiment for the Test phase, thus maximising the comparability of the conditions. The inclusion of a third English speaker for the habituation phase of the experiment was to rule out any possibility that improved performance with the noise-vocoded sentences could be ascribed to learning the indexical characteristics of one or other of the speakers in the habituation phase - this would have been most relevant for the English or Control condition, depending on when the habituation speaker was presented again.[1]

[1]Despite the scientific reasoning for inclusion of a different speaker in habituation, it would have been preferable to feature another speaker of SSBE, as the presence of a Northern Irish accent may have affected the expectations of listeners for the later phases of the experiment. However, the accented recordings were the only set available, aside from the two SSBE recordings, at the time of constructing the task. Given the length of the LSCP corpus items, it was decided that it was more important to include a habituation phase than to rule out its inclusion based on

Table 4.1: Basic properties of the experimental sentences.

	No. of Syllables		No. of Words		Duration(sec)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Dutch	18	1.49	9.8	1.40	2.96	0.27
Italian	18	1.49	8.9	1.52	2.96	0.25
English Training	18	1.49	10.9	1.20	2.99	0.26
English Test	18	1.49	12.2	1.40	3.29	0.33

Rhythmic Properties of Training and Test Sentences

On the basis of the findings of Dellwo (2007) and the post-hoc analysis of the Davis et al. (2005) sentences in Study 1 of this thesis, it was decided that $nPVI$ for vocalic intervals and rPVInorm (Dellwo, 2007) for intervocalic intervals would form the most suitable metrics for measurement of linguistic rhythm in this experiment. Each of the 10 selected items in the four sentence sets was labelled for vocalic, consonantal, syllable and pause intervals in PRAAT (Boersma & Weenink, 2005). As in Experiment 1, metrics were calculated for blocked versions of each condition, and for individual sentences. The mean values for the blocked and individual sentence approaches are shown in Table 4.2. Inspection of this table suggests a distinct difference between Italian and the other three conditions on the $nPVI$ metric, but no real difference between the conditions on the $rPVInorm$. A plot of the two measures for the four sentences sets against the BonnTempoCorpus languages is shown in Figure 4.1. While this shows that all of the test sets lie in the regions expected for their rhythm class, there seems to be little to separate them on intervocalic variability. This may be due simply to speaker variability, with the Italian speaker exhibiting a high level of variability for her language, while the other speakers exhibit relatively low variability in theirs. Two univariate ANOVAs were run using the individual sentence data - one for each metric - with Condition as the between-subjects factor. For $nPVI$, there was a significant effect of sentence set ($F(3, 36) = 13.42$, $p = .000$, $\eta^2 = .528$, power = 1.00). In post-hoc, Sidak-corrected comparisons, there were significant differences between English Training and Italian ($p = .000$); Dutch and Italian ($p = .003$); English Test and Italian ($p = .018$); English Training and English Test ($p = .022$).

the issue of accent. Furthermore, as the practice recordings were made by the experimenter, it was hoped that any difficulty with the accent would be overcome during the preamble and delivery of instructions before the experiment - Clarke and Garrett (2004) have shown that adaptation to an unfamiliar accent can occur after as little as two sentences of exposure.

There were no significant differences between English Training and Dutch ($p = 0.123$), nor between English Test and Dutch ($p = 0.983$). Interestingly, in both approaches (by Block and by Item) to the measurement of $nPVI$ in this experiment, the emergent values are much higher for the English Training condition than the English Test. This matter will be addressed in the Discussion. A univariate ANOVA using $rPVInorm$ measures for the sentence sets showed no significant effect of Condition ($F<1$), nor did it show any significant post-hoc comparisons between conditions (with Sidak correction).

Table 4.2: Mean values from a blocked and item-based analysis of the four sentence conditions.

	$nPVI$		$rPVInorm$	
	Block	Items	Block	Items
Dutch	64.0	79.0	57.3	56.7
Italian	45.6	62.4	58.4	57.1
English Training	73.3	89.3	58.1	57.1
English Test	61.3	76.0	57.6	54.9

Creating Noise-Vocoded Stimuli

Each of the 30 training sentences and the 10 test sentences were transformed into 5-band noise-vocoded versions. Davis et al. (2005) used 6-band stimuli, and there is evidence to suggest that the participants in that study had reached ceiling after 40 sentences. Furthermore, a pilot version of the current experiment obtained test phase speech recognition scores around 90% for both English and Dutch conditions, and so was thought to be too easy.

The noise-vocoding transformation was performed in PRAAT. Each sentence was passed through a set of five analysis filters (using Hann filters with smoothing set to 1/10 of the upper band frequency) covering the frequency range from 70Hz to 4kHz. Filtering divided the input waveform into five bands, whose bandwidths represented equal durations along the basilar membrane, and which were determined in accordance with the Greenwood (1990) equation relating filter position (on the basilar membrane) to best frequency. The amplitude envelope from each analysis filter was extracted through full-wave rectification and convolving with a Gaussian analysis window (Kaiser -20; sidelobes -190dB). These envelopes were then multiplied by a white noise, and passed through output filters (Hann) of matching bandwidth to the input filters. The output bands were matched to their original input bands for root-mean-square sound pressure level. The amplitude-modulated

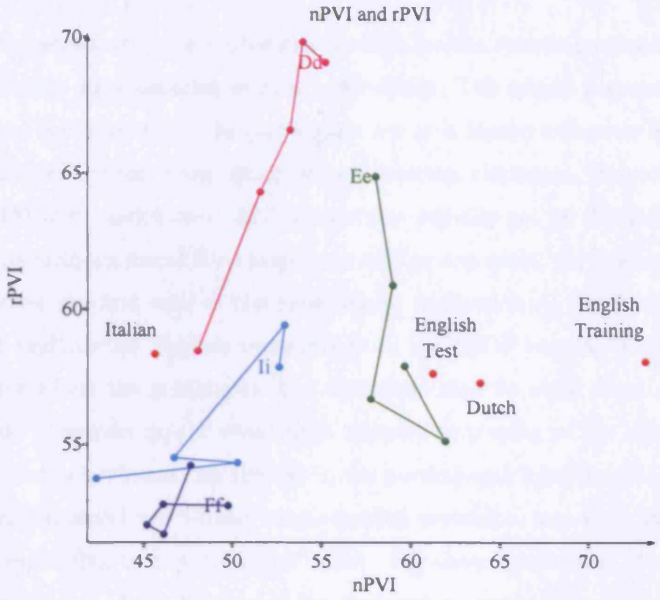


Figure 4.1: The rhythmic properties of the languages in the current experiment, as displayed in PRAAT (Boersma & Weenink, 2005) using the BonnTempo tools. The corresponding values for the BonnTempoCorpus languages are shown for comparison.

noise-bands were then summed together and low-pass filtered at 4kHz.

Design and Procedure

A between-subjects design was employed, in which 64 participants were assigned to one four training conditions (16 to each) - Italian, Dutch, English and Control (no training) - based on their experience and competency with foreign languages. Language experience was assessed via the questionnaire, which was given to the participant before the speech perception task. Any participant who indicated considerable experience (equivalent to A Level or above) to either or both of Dutch and Italian (or similar languages such as German and Spanish, respectively) was deliberately not included in the relevant foreign language condition(s). However, assignment of participants to conditions was done at random for participants whose questionnaires indicated

little to no relevant foreign language experience.

Participants were first given verbal instructions for the speech perception task by the experimenter, followed by more detailed written instructions. The speech perception task fell into three sections, named Sections A-C. The participant sat at a laptop computer and was provided with a pen and an answer sheet with which to give written responses. Stimuli were presented over Sennheiser HD25-SP headphones, and volume was initially set at the same level for all participants. If the participant found the stimuli too loud or too quiet, they were allowed to make small adjustments after the first trial of the experiment. In Section A, the Practice Phase, the participant heard 5 undistorted English sentences from the LSCP corpus. Each sentence was played once only, after which the participant had unlimited time to write down as much as they could of the sentence. The participant could then trigger the playing of the next sentence by pressing the space bar on the keyboard. In Section B, the participants from the training conditions Italian, Dutch and English heard ten 5-band noise-vocoded sentences, also from the LSCP corpus, in the language corresponding to the condition name. For these conditions, this section corresponded to the Training Phase. As in Section A, the task was to write down as much of the sentences as they could, although the instructions explained that the distorting effects of the vocoding would make this very difficult. In Section C (called Section B for the Control group), all participants (including those in the Control condition) heard a set of 10 noise-vocoded sentences in English. These ten sentences were the same for every participant, and formed the Test Phase in all four conditions. The reason for inclusion of 10 test sentences rather than 5 (as in Pallier et al. (1998)) is that previous studies have encountered considerable variability in sentence recognition scores with noise-vocoded speech (Davis et al., 2005; Nogaki et al., 2007; Stacey & Summerfield, 2007) - including 10 test sentences is hoped to increase the statistical power in the current experiment.

None of the participants who experienced the training sentences was given explicit instructions about the language of the sentences presented in Section B. This was done in order that the listeners would attend to the training sentences with the expectation that they were in English, and not disengage from the task on the basis that the stimuli were in a foreign language. There was no feedback of sentence content given on any of the trials, despite the fact that feedback is a good enhancer of perceptual learning in recognition of noise-vocoded sentences (Davis et al., 2005). The reason for this was two-fold. First, presentation of feedback in the training conditions would alert the participants to the fact that a foreign language was present and may encourage them to disengage from the task. Second, the lexical overlap between the languages use in this study is imbalanced - English and Dutch have considerable overlap while Italian is linguistically more distinct from the other two. Therefore, should Dutch provide better training than Italian in the

study, the presentation of feedback might reflect the lexical commonalities of the languages rather than their rhythmic similarities. This is at odds with the intention of the experiment, which was to assess whether listeners would attend to the rhythmic properties of sentences in order to access word boundaries in the noise-vocoded signal.

In Section A, the sentences were presented in a fixed order for all participants. In Sections B and C (where relevant), the sentence lists were randomized.

4.2.2 Results

For the purposes of this experiment, participants' responses on the questionnaire were only used to filter participants into a suitable condition of the speech perception task.

In Section A of the speech perception task, responses were scored as Proportion Words Correct. Scoring included all function words and keywords. A conservative scoring system was implemented, where only exact matches were accepted i.e. there was no allowance for morphological deviations or deletions due to number agreement. However, participants were not penalised if they included extraneous words or syllables. The mean recognition scores across the 5 practice sentences was calculated for each participant. Despite the presence of a regional accent, the similarly high mean scores in the habituation phase for the four conditions (Italian: $M = 0.96$, $SD = 0.04$; Dutch: $M = 0.97$, $SD = 0.03$; English: $M = 0.97$, $SD = 0.03$; Control: $M = 0.96$, $SD = 0.05$) were enough to indicate that the participants could perform the basic task of listening to sentences and giving written report of their content with sufficient accuracy. For Section B responses, response scoring was only carried out for those participants who heard English noise-vocoded sentences (English and Control groups). A more liberal scoring system was adopted here, where deviations in tense and number agreement on nouns (i.e. if the participant reported 'men' when the actual keyword was 'man') and verbs (i.e. if the participant reported 'carries' or 'carried' when the correct word was 'carry') were allowed. The reasoning behind this approach was to allow for errors that may have resulted from the participant's attempts to report a grammatically correct sentence for each item. For example, if the participant hears the first keyword in '*the cup hangs on a hook*' as 'cups', then he/she may choose to report 'hang' as the second keyword, in order to maintain number agreement. The same system was adopted for Section C responses across all participants. For each listener, three scores were calculated for use in analysis; Overall Test Phase Sentence Recognition (mean Proportion Words Correct score over the 10 sentences in Test Phase - Section B of Control condition; Section C of the remaining conditions), Block 1 Test Phase Sentence Recognition (the

mean score for sentences 1-5 of the Test Phase) and Block 2 Test Phase Sentence Recognition (the mean score for sentences 6-10 of the Test Phase). The equivalent scores were calculated for Training Phase (Section B) performance in the English training condition. Training Phase scores were not available for one of the listeners due to a computer error that resulted in loss of the file detailing the order of item presentation in this phase.

Figure 4.2 shows the Test Phase results for participants in the four conditions of Experiment 2. The analysis of the Test Phase performance data had two objectives. The first aim was to test for evidence of adaptation or perceptual learning, as demonstrated by Davis et al. (2005). The second was to assess whether the different training conditions had a significant effect on the rate of this adaptation. The Figure indicates that performance in all four conditions improved from the first half of the test phase to the second. It also appears that English training provided the highest mean Test Phase score, while the foreign language training offered no benefit over the Control condition, where training was absent. In fact, mean scores on the Dutch and Italian conditions were numerically lower than in the Control group. The two effects of interest were tested in a repeated-measures ANOVA, with Block as a within-subjects factor (using Block 1 and Block 2 Test Phase scores) and Condition as a between-subjects factor (with levels English, Italian, Dutch and Control). Arcsine transformed proportion scores were used in the ANOVA. The overall effect of Block was significant ($F(1,60) = 24.36$, $p = .000$, $\eta^2 = 0.289$, power = .998), indicating that perceptual adaptation/learning took place over the course of the Test Phase. The effect of Condition reached only marginal significance ($F(3,30) = 2.36$, $p = 0.080$, $\eta^2 = 0.106$, power = 0.564), suggesting that there was no difference between the training conditions over the whole experiment. Sidak-corrected post-hoc comparisons across conditions were all non-significant, with only the comparison between English and Italian reaching marginal significance ($p = .094$). The Block x Condition interaction did not reach significance ($F(3,60) = 1.84$, $p = .149$, $\eta^2 = .084$, power =.455), which suggests that learning in the Test Phase progressed at a similar rate across conditions. However, on the basis that between-condition differences would have been greatest in the first block of the Test Phase, a univariate ANOVA was run using only Block 1 Test Phase scores to re-assess the effect of Condition. This analysis produced a significant effect of Condition ($F(1,63) = 3.80$, $p = 0.015$). There was a significant post-hoc comparison (Sidak-corrected) between English and Italian (p=.019), and there were comparisons of marginal significance between English and Dutch ($p = .097$) and between English and Control ($p = .070$). All other post-hoc comparisons were non-significant. This analysis should be treated cautiously as the item presentation was not suitably counterbalanced, and so there may have been an uneven distribution of items in the first 5 test sentences. However, it gives support to the prediction that the English condition would yield significantly better Test Phase scores than Control, due to the exposure to

twice as many English noise-vocoded sentences in the former condition.

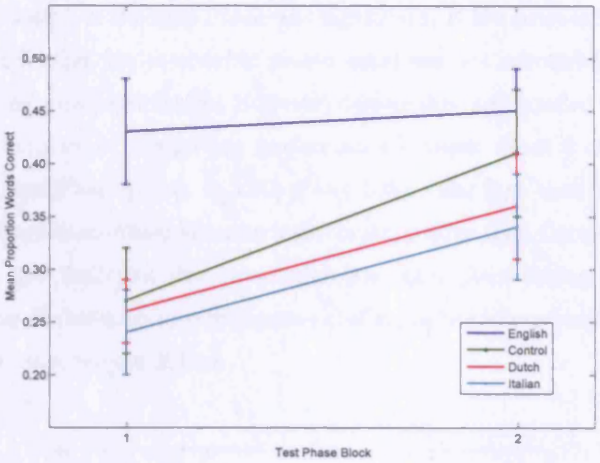


Figure 4.2: Plot of the results of the cross-linguistic adaptation study. Error bars show ±1 standard error of the mean.

Inspection of Figure 4.2 indicates that the lack of significant difference between the English condition and the other three conditions in the analysis of the Overall Test Phase scores may have resulted from a ceiling effect in the data. Although not of statistical significance, the improvement from Block 1 to Block 2 of the Test Phase is much more modest in the English condition than for the other three participant groups, suggesting that improvement in performance slows down after the first 10 sentences of exposure. Hence, the Control group, being naïve to noise-vocoding in the Test Phase, exhibit a more dramatic rate of improvement. By extension of this idea, the similar rate of improvement seen in the Test Phase performances of the Dutch and Italian groups to that of the naïve listeners in the Control group offers further support that the listeners in the foreign language training conditions could only learn significantly when exposed to noise-vocoded sentences in their own language.

To explore the possibility of a slowing in learning in the Test Phase of the English condition, mean scores on the two Blocks of this phase were compared with the same listeners' Training Phase performances. A plot of these values is shown in Figure 4.3, using data from the fifteen participants for whom Training Phase scores were available. This indicates, as anticipated, a slowing in the rate of learning in the Test Phase compared with the Training Phase. However, a striking observation is that there is a considerable drop in performance at the beginning of the Test Phase. This suggests

either a disruptive effect of the change in speaker on perceptual learning, or some other factor such as a mismatch in basic speaker intelligibilities and/or item difficulties across the two sentence sets. A paired-samples t-test was run to test whether the difference between scores in Block 2 of the Training Phase and Block 1 of the Test Phase was significant. It is acknowledged that item effects are not completely controlled for, as order of presentation was not counterbalanced by 5-sentence blocks in the original experimental design. However, despite this, a two-tailed paired-samples t-test showed a signficant decrease in recognition performance between Block 2 of the Training Phase and Block 1 of the Test Phase ($t(14) = 3.30$, $p = 0.005$). The fact that the mean Test Phase performance in the English condition is numerically much greater than Control performance offers sufficient support to the prediction that adaptation has taken place during the English Training Phase, but the finding of this drop in performance challenges the interpretation of the Test Phase scores for the foreign language conditions.

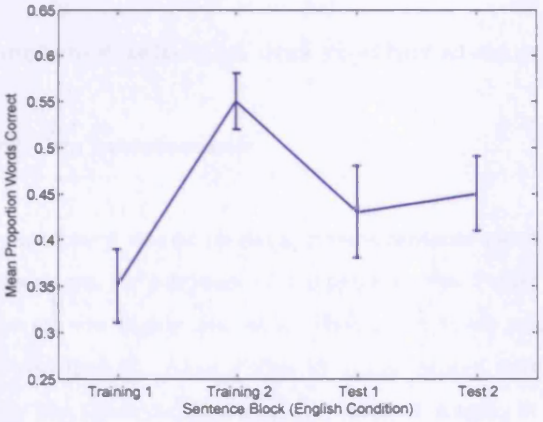


Figure 4.3: Mean performance throughout Training and Test in the English condition of Experiment 2.

4.2.3 Discussion

The results show little support for either of the hypotheses put forward in the Introduction. First, the lack of significant difference in the training efficacies of Italian and Dutch noise-vocoded sentences suggests that there is no advantage offered by Dutch and hence, that linguistic rhythm may not be an important cue in adaptation to noise-vocoded speech. Second, the finding that neither of the foreign languages offered any advantage over a Control condition, in which listeners received

no Training Phase exposure to vocoded speech, adds further support to the suggestion from Davis et al. (2005) that top-down influences of lexical information in the training materials are necessary in driving adaptation to noise-vocoded sentences.

There are many issues emergent from the results that merit discussion. These are varied and overlapping, but can be best thought of in four overall themes:

1. Problems with sentence selection and rhythm measurement
2. The influence of task demands, instructions and participant expectations
3. The differences between noise-vocoding and time-compression
4. Individual variability in the listening population.

Problems with sentence selection and rhythm measurement

The challenges of rhythm measurement

The emphasis in this experiment was on choosing spoken sentence materials that were as naturalistic as possible. Furthermore, for purposes of comparison with Pallier et al. (1998), employing the LSCP sentence corpus was highly desirable. Having set these priorities, control over other aspects of the stimuli was limited. After Pallier et al., sentences were matched across the four sentence sets (primarily the three training sets) for sentence length, in terms of both number of syllables and duration in seconds. This was done purely by selecting closely matched items, rather than altering any of the recordings. In matching for duration, we were indirectly controlling for speech rate in syllables/second - Dellwo and colleagues argue that this is potentially flawed as different languages naturally differ in speech rate, and this in turn affects the rhythmic properties of languages (Dellwo et al., 2004; Dellwo, 2007). Despite this, mean measurements of linguistic rhythm in this experiment (using $nPVI$ and $rPVInorm$ scores on each block of 10 sentences, and for individual sentences) indicated that the four speakers fell into their expected rhythm 'classes', with the Dutch and English speakers in the 'stress-timed' region of plots, while Italian appeared more 'syllable-timed'. This was supported by significant differences between Italian and the other languages on the $nPVI$ measure. However, it was perhaps short-sighted to consider only the mean values of the rhythm measures, as the structure of the task meant that sentences were presented individually with considerable intervening silences (while the participants wrote their responses). Therefore, for the purposes of 'tuning in' to linguistic rhythm, the individual sentence is key in

this experiment. With the emphasis on naturalness, we were not able to ensure that every item in the 'stress-timed' conditions was distinctly different in rhythm from those in the Italian condition. A further problem is that, despite both conditions exhibiting strongly 'stress-timed' rhythmic patterning, the English Training sentences gave considerably higher $nPVI$ values than the English Test set. This difference in rhythm may have contributed to the dip in performance between Training and Test phases of the English condition, rather than the disruption being caused by changes relating to the vocal tracts and vowel spaces of the two speakers. This was not a concern for the transition from Dutch to English, as these sentences were better matched rhythmically than the English Test and English Training sets.

Another issue in the sentence selection employed in this experiment, which could also relate to the potentially sub-optimal rhythmic contrasts between conditions, is that Pallier et al. (1998) do not state which particular sentences from the LSCP corpus they used in the training blocks of their experiment (which they call the 'habituation' sentences). There may be some sentences that are instrinsically more strongly representative of the language's rhythmic class, and Pallier et al., and Sebastian-Galles et al. (2000) could have serendipitously benefitted from this in their sentence selection. For example, most speakers of British English would be inclined to read *'It's easy to tell the depth of a well'* (one of the items from the IEEE sentence corpus (IEEE, 1969)) with a certain regular, stress-timed, meter. In contrast, *'Jump the fence and hurry up the bank'* may result in much more rhythmic variability across speakers. Regardless of this possibility, the fact that these two previous studies have turned up significant adaptation results that fall in line with an interpretation based on linguistic rhythm, suggests that the null result obtained in the current experiment (i.e. a lack of adaptation with Dutch sentences) is more likely to reflect (1) a lack of role for linguistic rhythm in adaptation to noise-vocoded speech or (2) masking of an underlying rhythmic effect due to other sources of variability e.g. a sudden change in speaker.

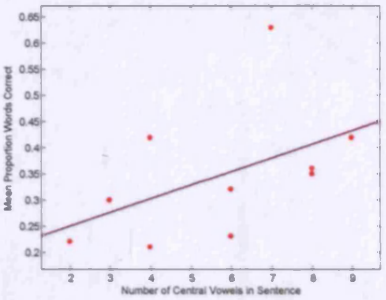
Item Effects in the Perception of Noise-Vocoded Speech

In selecting such a small number of quite linguistically complex sentences from the LSCP corpus, it would have been virtually impossible to match items along linguistic parameters such as lexical frequency, semantic predictability and syntactic complexity. Hence, these factors were not accounted for in stimulus selection for the current experiment. However, it is well established that such properties of linguistic stimuli can affect recognition of spoken material (Obleser et al., 2007; Hannemann, Obleser, & Eulitz, 2007). The choice of sentence corpus for the current study also made any attempt at phonetic matching across or within sentence sets unattainable, yet the noise-

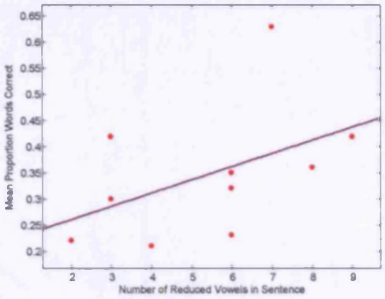
vocoding procedure is certain to differentially affect the intelligibility of different speech segments. As discussed above, the rhythmic properties of individual sentences, whilst lying within the range of values associated with their rhythmic 'class', still exhibit variability.

The presence of item effects is apparent from inspection of the Test Phase recognition data across all 64 participants in the current experiment. Across the group, the average Test Phase score was 0.35, ranging from 0.21 for the least intelligible sentence to 0.63 for the most intelligible. An exploratory analysis of the factors driving this variability was attempted by running a set of two-tailed Bivariate Pearson's and Spearman's correlations between the mean recognition score for each sentence and a set of quantifiable variables describing the sentence. These variables are listed in Appendix B, and cover phonetic content, complexity and rhythmic metrics for each sentence. Bearing in mind that there are only 10 items in the set, these analyses are interpreted with caution. However, several variables produced correlation coefficients of 0.4 and above with sentence recognition scores. Some of the correlations, when plotted, do not appear to be meaningful. However, Figure 4.4 shows a summary of the more readily interpretable correlations. The analyses indicate a positive correlation between sentence recognition scores and (1) the number of Central Vowels in the sentence (Pearson's $r = .490$) (2) the number of Reduced vowels in the sentence (Pearson's $r = .467$). Both Central and Reduced categories contain reduced vowels, which may assist in the identification of the rhythmic outline of the sentences in terms of stressed and unstressed syllables. Negative correlations were identified between sentence recognition scores and three descriptive variables: (1) the number of voiced plosives in the sentence (Pearson's $r = -.573$), (2) the number of open vowels in the sentence (Pearson's $r = -.494$) and (3) the number of diphthongs in the sentence (Pearson's $r = -.432$). With the very small number of data points in each correlation, any conclusions drawn from this post-hoc analysis must be very tentative. There seem to be some indications, quite sensibly, that more complex sentences (with more diphthongs and complex consonant clusters) are more difficult to understand. It remains to be tested whether the suggested relationships with sentence stress (as measured by the number of reduced vowels) would also be borne out in an analysis of a much larger item set.

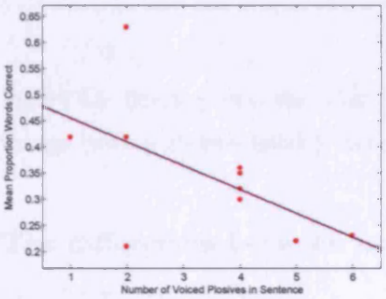
The scatterplots in Figure 4.4 clearly show that one sentence is much more intelligible than the others across the listening population. This sentence - *Seven paintings of great value have recently been stolen from the museum* - had a mean intelligibility score of 0.63. It is unclear what contributes to this distinct advantage in intelligibility over the other sentences, and whether it may be linguistic as well as phonetic in origin. Its outlying presence in the set of 10 sentences may have unduly influenced the outcome of the Pearson's correlation analyses above. However, all of the above correlation coefficients remained greater than 0.4 when this highest-scoring sentence



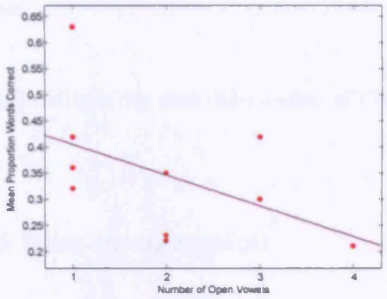
(a) Central Vowels



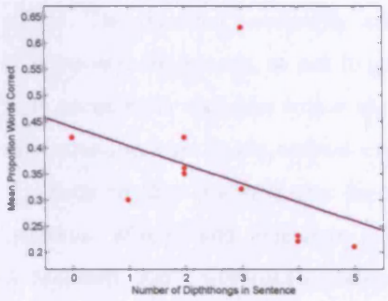
(b) Reduced Vowels



(c) Voiced Plosives



(d) Open Vowels



(e) Diphthongs

Figure 4.4: Scatterplots illustrating the correlation between recognition scores and phonetic properties of the sentences.

was removed from the analysis. In the case of the correlation between recognition scores and the number of diphthongs in the sentence, the removal of this sentence results in a considerable increase in the correlation coefficient, to $r = -.834$. The scatterplot is shown in Figure 4.5 below.

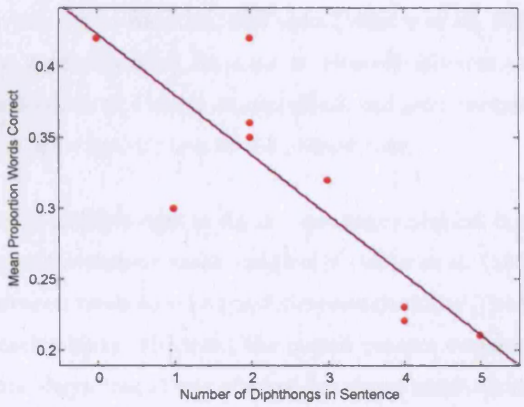


Figure 4.5: Scatterplot of the relationship between sentence intelligibility and the number of diphthongs (with high-intelligibility item removed)

The differences between noise-vocoding and time-compression

Importantly for the current data set, the initial exploration of item variability above indicated no significant relationship of sentence intelligibility with direct measures of linguistic rhythm such as $nPVI$. This does not necessarily indicate that there is no role for linguistic rhythm in perception of noise-vocoded speech, as our hypothesis is based on between-class differences in rhythm rather than parametric variation within the same class. However, it is also acknowledged that the rhythmic measures used in the current experiment are durational only, while it is widely accepted that linguistic rhythm is also carried by other properties of the stimulus, such as stress and intonation patterns. Mattys and colleagues (Mattys et al., 2005; Mattys, Melhorn, & White, 2007; Mattys & Melhorn, 2007) advocate an integrated approach to the issue of speech segmentation that considers a downward hierarchy of cues from sentential and lexical context to acoustic factors such as stress, duration and coarticulation. In this experiment, there was no attempt to control for these additional cues to rhythm and segmentation, and thus the study adopts a rather simplistic view. However, this was deliberate, and the approach can be defended in terms of the acoustic nature of the noise-vocoded stimulus. Mattys and colleagues acknowledge a flexibility in the hierarchy of segmentation cues, where reduced access to the linguistic content of the sentence (e.g. through addition of white noise) increases the relative significance of acoustic cues to segmentation. With only 5 bands, the noise-vocoded sentences in the current experiment are very heavily distorted, reducing the listener's certainty of the linguistic content and thus presenting conditions in which

acoustic cues to segmentation are likely to be more prominent. Furthermore, a recent study on the perception of prosodic cues by cochlear implant users (Meister et al., 2007) found that implantees performed similarly to normal-hearing listeners on prosodic discriminations based on temporal structure, but worse with contrasts based on amplitude and pitch variations. Hence, a durational hypothesis was thought to be appropriate in the present case.

Related to the issue of multiple cues to rhythm and segmentation, it is possible that the reason for the difference between the current result and that of Pallier et al. (1998) simply lies in the basic acoustic differences between noise-vocoding and time-compression. Time-compression results in a reduction in speech intelligibility. However, the overall percept remains highly *speech-like*. This stands in contrast to the degrading effects of noise-vocoding, which results in stimuli that are still recognisable as speech but are considerably altered in quality (particularly at levels of degradation such as the one used in the current experiment). For example, the time-compression algorithm preserves the pitch of the original stimulus while noise-vocoded stimuli are greatly impoverished in this respect. Thus, it is possible that time-compressed sentences simply provide sufficient information about lingustic rhythm to make this an important adaptation cue for these stimuli, while the primarily durational and amplitude-based rhythmic information in noise-vocoding is not enough. An alternative, more general explanation, is that the better speech-like quality of time-compressed stimuli makes these materials better adaptive stimuli overall. As well as being richer in spectral detail, time-compressed items potentially offer more immediate ecological validity for normal-hearing participants, in simulating fast speech, than that achieved by the whispered quality of noise-vocoding. Overall, time-compressed items are therefore likely to be more *familiar* and more easily encoded perceptually than their noise-vocoded equivalents, which may have implications for the extents to which learning of these two different stimulus types can be generalized across languages i.e. the greater challenge to perceptual encoding provided by noise-vocoding means that adaptation to this stimulus requires greater contextual support, for example in terms of familiar linguistic information.

The influence of task demands, instructions and participant expectations

The outcomes of the experiment, in particular the very low overall test phase recognition scores, suggest that the study may have benefitted from a little more direction in the instructions, despite the potential compromises in terms of ecological validity. As will be discussed in some more detail in the next Section, there was considerable individual variability in scores within conditions, and it is possible that this would have been reduced by giving listeners more detailed and pointed

guidelines for the performance of the task. If not ideal, this would at least have demonstrated whether listeners are capable of making use of rhythmic cues, whether or not this would have been done automatically i.e. without instruction.

Post-test debriefing of the participants presented a very interesting outcome of the decision not to instruct the participants as to the presence of foreign language stimuli. Very few of the listeners in the relevant conditions noticed, or hypothesised, that the training sentences were not in English. Furthermore, some listeners failed to make any more comment about the foreign training conditions other than that the constituent items were 'more difficult' than the Test Phase sentences. Such comments persisted even when the debriefing encouraged the listeners to comment on how the sections of the experiment may have been different from each other. Unfortunately, the debriefing session was not standardised across participants and so there was no useable record of their responses. Therefore, all commentary here is from the recollections and observations of the experimenter. However, it is striking how the content information provided to the listener can very much determine their expectations. A recent study by Magnuson and Nusbaum (2007) on the effects of speaker variability on speech perception exemplifies this very phenomenon. The authors assessed the effect of top-down, cognitive expectations on reaction times to probe words presented in a list. For certain pairs of voices or synthetic speakers, they found that reaction times are slower in blocks with a mixture of trials by the two speakers versus blocks of trials from one speaker only. However, in a further experiment with two synthetic voices differing only in fundamental frequency that had previously not produced such a 'speaker variability effect' on reaction times, an effect was produced in a group of listeners who were instructed that they would hear two different speakers. A control group, which was told to expect one voice that was sometimes altered in pitch, did not show a speaker variability effect. Thus, in the presence of the same physical stimuli, the instructions given to participants, and hence their cognitive expectations, determined the outcome of the experiment. It is difficult to anticipate how effective a change in instruction would be in the current experiment, but the observation that listeners effectively assumed the presence of English language throughout is perhaps evidence that top-down expectations are quite important when listening to degraded speech. This evidence for the importance of top-down processing is indirectly supported by the findings of Davis et al. (2005) and Obleser et al. (2007) who show evidence for the importance of lexical and semantic information, respectively, in their studies of perception of noise-vocoded sentences. Experiments involving manipulation of expectations may improve our understanding of the extent to which these top-down effects are effective. In the context of the current experiment, this could involve manipulating, separately and together, the expectation of foreign language stimuli and the instruction to attend to linguistic rhythm.

A further concern regarding participant expectations relates back to the presence of a UK regional accent in the five undistorted practice sentences presented to all participants at the beginning of the experiment. Without clear instruction that the test sentences would be in a standard Southern British English (SSBE) accent, listeners may have been under the impression that they would encounter the same accented voice in the distorted sentences as was experienced in the habituation trials.

Returning to issues concerned with the procedure employed in this experiment, a possible manipulation to improve participants' recognition scores and learning trajectories, and to potentially reduce variability, would be to introduce feedback to the Training and Test Phase trials. This could take the form of the most successful feedback regimes in Davis et al. (2005) - for each trial, participants would give their full response before receiving the sentence content in either written or undistorted spoken form, followed by a second playing of the distorted sentence. This is a manipulation that would also require explicit instruction regarding the presence of foreign languages in the relevant conditions. This was avoided in the current experiment for several reasons. Importantly, Pallier et al. (1998) did not give feedback in their task. Furthermore, the provision of feedback could conflate the intended rhythmic manipulations with the differing lexical similarities of Italian and Dutch with English. As Dutch has much greater lexical overlap with English, any benefit over Italian could be attributed to this rather than their rhythmic differences. fThe second problem was overcome by Sebastian-Galles et al. (2000), who identified Greek as a syllable-timed language showing little lexical overlap with their test language, which was Spanish. A lexically suitable stress-timed language to be paired with English is Arabic. For reasons of lexical overlap, it was hoped that Arabic could be used in place of Dutch in the current experiment. Indeed, transcriptions of Arabic sentences were available in the LSCP corpus. However, it was difficult to find a suitable speaker amongst the University College London community who could read the particular dialect used in the corpus sentences.

There is a further dimension to the consideration of feedback in the design of the current experiment. In Chapter 3, it was suggested that the key problem with Nonword sentences in Davis et al. (2005) may not have been simply whether listeners had sufficient short-term memory capacity to remember them during feedback, but whether they were able to make appropriate use of sound-to-meaning mapping processes in feedback. The hypothesis put forward in Chapter 3 is that the rhythmic non-naturalness of the Nonword Sentences made it more difficult for listeners to identify word onsets, even when presented with an onscreen written version throughout the repetition of the sentence to ward off the possibility of forgetting the feedback content. If we posit that some sort of feedback is important in the perceptual learning of noise-vocoded speech, in a

way that may not be so critical for learning of time-compressed speech (which still sounds like 'real speech' in quality, but not in rate), then it is possible that much of the learning takes place during feedback and is therefore dependent on the conditions of this feedback. In this way, where speech rhythm may have been integral to the accurate mapping of sound to content in Davis et al. (2005), it may not play such a strong role when no feedback is provided, as in the current study.

If we accept that some real word information might be necessary for learning from noise-vocoded stimuli, yet there may still be a role for linguistic rhythm, there are two ways in which the experiment in this chapter can be modified. The first, which prioritises naturalness in the stimuli, is to identify a native accent of English that exhibits syllable-timing, and compare its training efficacy with that of a standard stress-timed variety of English. Such a comparison is possible, as Singapore English is a native tongue that has measurable differences from Standard Southern British English, in the direction of syllable-timing (Deterding, 2001; Grabe & Low, 2002; Low et al., 2000). As a planned follow-up to the current experiment, recordings of the English Training sentences were made with a female undergraduate student from Singapore who was recruited from the University College London community. Unfortunately, when the Singapore English stimuli were measured along the $nPVI$ and $rPVI$ metrics, they lay firmly in the 'stress-timed' region and were insufficiently different from the English Training stimuli for use. A replication of the current experiment's design incorporating Singapore English would have necessitated a speaker change between Training and Test. As described in the Results section of this Chapter, this may prove problematic as there is evidence of a disruption to learning across the speaker change in the English condition of the current experiment. An exploration of the effect of changing speaker is therefore the topic of the next chapter. The second option to test the role of rhythm within perception of noise-vocoded British English would be to use the same speaker to make recordings of naturally-timed (i.e. 'stress-timed') and 'syllable-timed' versions of the sentences. This approach is adopted in Experiment 5.

Overall, there are several points on which criticism can be levelled at the procedure and instructions employed in the current experiment. However, the study forms a good starting point for exploration of some these interesting issues concerning participant expectations and awareness.

Individual Variability in the listening population

One of the most impressive findings in this experiment, and indeed perhaps the strongest obstructing factor to the manipulation of interest, is the considerable variability in the scores of

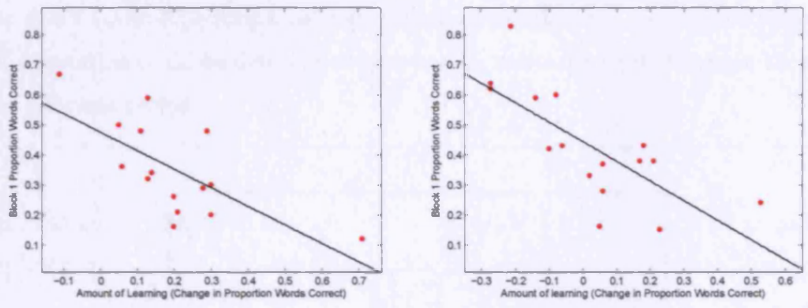
individual listeners, within and between conditions. The descriptive statistics for the overall Test Phase recognition scores in the four conditions, as shown in Table 4.3, demonstrate this variability clearly. What is most striking is that some listeners in the Dutch, Italian and Control conditions are evidently producing recognition scores that are equivalent to, or greater than, listeners in the English condition who have had exposure to 10 extra English sentences. Despite the criticisms of the experimental instructions and procedure explored above, the participants were all given a readily understandable and unambiguous basic instruction - to write down all that they could from each sentence. Had the task instructions been more complicated, for example involving a direction to attend to linguistic rhythm, the variability in outcome may have been more easily ascribed to the participants' misunderstanding of the instructions than to genuine differences in perception. However, if we assume that the instruction was sufficient to lead all listeners to use the same basic listening strategy, then the observed variability indicates that there are differences between individuals in core elements of speech perception - from basic hearing acuity to phonological working memory and verbal intelligence.

Table 4.3: Descriptive statistics illustrating individual variability in Experiment 2.

	Mean	Min	Max	IQR
English	.44	.19	.72	.17
Dutch	.31	.00	.52	.17
Italian	.29	.03	.56	.32
Control	.34	.00	.69	.35

If there is such variability in the overall Test Phase recognition performance, is there similar variability in the rate of perceptual adaptation exhibited in the Test Phase? Furthermore, is this related to the overall level of performance? The rate of adaptation is slightly more difficult to measure on an individual level, as sentence presentation was randomized in the Test Phase. Given the considerable item effects described above, these could significantly impact on individual learning trajectories. A score for 'Amount of Perceptual Adaptation' was calculated for each listener by subtracting the proportion scores for Block 1 from those for Block 2. The mean value for this variable was 0.09, with a range from -0.27 to +0.53, indicating that some listeners may have encountered difficult items late in the Test Phase that had the effect of reducing Block 2 scores. A two-tailed Pearson's correlation between Block 1 scores and Amount of Perceptual Adaptation gave a significant negative correlation (Pearson's $r = -.355$, $p = .004$), indicating that those listeners with higher scores at the beginning of the Test Phase exhibit the smallest amounts of learning. When the data were split by Condition, this significant correlation held

only for the English condition (Pearson's $r = -.743$, $p = .001$, 2-tailed). This may reflect the possible ceiling effect in learning identified earlier in the analyses. However, the correlation is also highly significant for the Training Phase performances in the English condition (Pearson's $r = -.692$, $p = .004$), which formed the first 10 noise-vocoded sentences encountered by this group of listeners. Figure 4.6 shows a scatterplot of the two significant correlations for the English group. Due to the evident item effects, we cannot draw any firm conclusions as to the true nature of the relationship between 'baseline' perceptual ability and the amount of adaptation shown by the participants in this experiment. The primary design requirement for formal investigation of this relationship, if using the LSCP sentences, would be to expose all listeners to the same order of presentation in the Test Phase. However, the correlations observed between initial performance and the amount of learning for the participants in the English group offers an interesting picture of individual differences in perceptual learning performance. Later experiments of the thesis will re-visit this question.



(a) Training Phase (b) Test Phase

Figure 4.6: Scatterplots illustrating the relationship between initial performance and amount of learning for participants in the English condition of Experiment 2.

An interesting observation regarding the range of overall scores obtained in the current experiment is the number of listeners achieving very low scores. Seven listeners obtained Test Phase scores less than 0.1 (equivalent to approximately 1 word per sentence), while fifteen produced proportion scores of 0.2 or less (less than 2.5 words per sentence). From observation of the scatterplot in Figure 4.7, which shows the relationship between initial Test Phase performance and learning for participants in all four conditions, it can be seen that several of the lowest-scoring individuals overall also exhibited the smallest amounts of learning in the Test Phase. It is possible that the 5-band noise-vocoded stimuli, which are quite strongly distorted, are simply too difficult for some listeners to extract any intelligible information from them. Not only does this result in very low recognition scores initially, but it also prevents any improvement in performance over time. This

indicates that there is a floor effect in the data. The presence of a possible ceiling effect at the other end of the spectrum of abilities is more difficult to detect in the current data set. However, in order to properly quantify and characterize variability in speech recognition and perceptual adaptation, it is important to rid the data of floor and ceiling effects where possible. In the case of noise-vocoded speech, there is a readily available means of adjusting task difficulty to accommodate the ability of the individual listener. The number of noise bands in the stimulus is logarithmically related to intelligibility (Davis & Johnsrude, 2003; Shannon et al., 1995, 2004). Thus, increasing the number of bands in the test stimulus presents a means of lifting a participant's performance above floor. Similarly, a listener achieving ceiling recognition scores at a certain level can be brought to a lower level of performance by decreasing the number of bands. Methods traditionally used in psychophysics, such as adaptive tracking, manipulate task difficulty in this way in order to quantify individual performance in terms of the stimulus difficulty level needed to achieve a criterion score. For example, in the case of noise-vocoded sentence recognition, performance could be tracked item-by-item, with the stimulus difficulty adjusted online to follow 50% words correct. This would give an overall performance score for speech recognition. A corresponding value for the amount of adaptation could be determined post-hoc by calculating the change in threshold over a pre-determined time period.

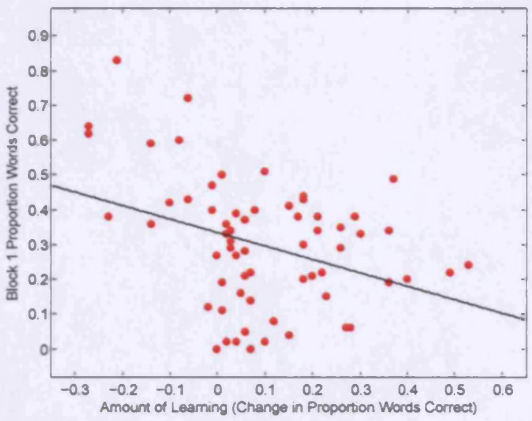


Figure 4.7: Scatterplot of the relationship between initial Test Phase performance and the amount of learning for all participants in Experiment 2.)

Chapter 1 gave an overview of the studies to date that have attempted to identify processing correlates of variability in speech perception in the normal-hearing population. Several pieces of evidence point toward higher-level cognitive contributions to speech recognition scores (Kidd et al.,

2007; Surprenant & Watson, 2001; Chiu et al., 2002). Furthermore, previous studies investigating the spectro-temporal processing correlates of speech recognition may have used auditory tasks that encouraged a low-level, 'analytic' listening mode that may be quite different to that used to identify speech sounds (Surprenant & Watson, 2001). Later experiments in the current thesis will attempt to approach the question of performance correlates through careful test selection based on previous evidence and reasoned estimation of the auditory skills required to process noise-vocoded stimuli.

4.3 Summary

The results of the current experiment were not as predicted. There was no evidence that noise-vocoded speech could be learned in the absence of meaningful lexical information. This result, in turn, limited the interpretation of the potential role of linguistic rhythm in adaptation, as manipulated through the difference in rhythmic class between Dutch and English ('stress-timed') and Italian ('syllable-timed'). Several factors made it difficult to establish whether the lack of effect was truly due to an absence of an effect of rhythm. First, and most importantly, overall performance was very low across all conditions. This indicates that the participants simply found the task so difficult, even for English noise-vocoded sentences, that to expect them to extract any useable information from the foreign language conditions was too ambitious. Second, there is evidence from the English language condition of a possible effect of changing speaker between Training and Test phases. It cannot be concluded whether this is based on item effects, on the indexical differences between the speakers' vocal tracts or even due to within-class rhythmic differences between the speakers. A further experiment is needed to address the role of the speaker in adaptation to noise-vocoded speech. Third, there was a large amount of variability exhibited in participants' test scores, both within and between conditions of the experiment. This may indicate that, despite the uniform instructions given to all participants, individuals may have differed in the listening strategies they adopted, and furthermore their ability to make use of certain cues in the stimuli.

Despite the complexity of the experiment, the extent of which has made its interpretation very difficult, it has opened up several interesting avenues for further experiments. In order to continue an investigation of the role of linguistic rhythm, as assessed by the cross-linguistic technique, the first step should be to gain a better hold on the potential effects of speaker change on adaptation to noise-vocoded sentences. Should the speaker change effect prove significant, the basic design of the current experiment would need to be reviewed. Should it emerge that the speaker change effect in the current experiment was driven by item effects, then counterbalancing of items should overcome this. An impressive 'side-effect' of the current experiment was the considerable inter-

individual variability both in overall sentence recognition scores and in the rate of improvement over the course of the Test Phase. That such considerable variability was found within a seemingly homogeneous young adult population presents the potential for noise-vocoded speech to be used as a tool to systematically investigate and characterize individual differences in speech perception.

Chapter 5

Stimulus properties: The role of the speaker

Abstract

Experiment 3 tests an issue arising from Experiment 2 - namely, the potentially disruptive effect of changing speaker during a period of exposure to noise-vocoded sentences. Forty listeners were exposed to two blocks of 10 noise-vocoded sentences from the same corpus as used in Experiment 2. Half of the participants experienced a change in speaker after the first block of sentences, while the other half of participants heard the same speaker across the full stimulus set. The results indicate that, despite a numerical advantage in the Block 2 performance of listeners who had experienced no speaker change, the difference between the groups was non-significant. The results are discussed in light of item effects, task demands and the effects of discriminability of the speakers (Experiment 4).

5.1 Introduction

Perceptual Learning - Transfer across speakers

The aim of this experiment was to address the possibility that the change in speaker between Training and Test in Experiment 2 may have disrupted adaptation to the noise-vocoded speech, as indicated by a significant drop in recognition scores at the start of the Test Phase in the English condition of that experiment. Put in other words, this experiment asks whether perceptual learning of noise-vocoded speech can be generalized from one speaker to another.

Studies of perceptual learning in speech have taken two quite different experimental approaches. One approach has assessed learning in terms of the listeners' improvement in recognition of speech stimuli that have been globally altered e.g. through time-compression (Altmann & Young, 1993; Mehler et al., 1993; Pallier et al., 1998; Sebastian-Galles et al., 2000), foreign accent (Clarke & Garrett, 2004; Weill, 2001), noise-vocoding (Davis et al., 2005; Hervais-Adelman et al., in press) and synthetic speech (Schwab et al., 1985). Another more recent approach, first used in a study by Norris, McQueen, and Cutler (2003), concentrates on the retuning of listeners' percepts of specific speech sounds. Norris et al. (2003) presented a fricative that was midway between [f] and [s] to a group of Dutch listeners. Some of the listeners were pre-exposed to the ambiguous fricative in lexical contexts favouring [f], while others heard the ambiguous phoneme in lexical contexts favouring [s], and another group of participants heard this sound in the context of nonwords. On a post-exposure categorization of items on an [ef]-[es] continuum, listeners who had previously heard [f]-biased presentations were more likely to categorize ambiguous phonemes as [f], while those from the [s]-biased exposure phase were more likely to label these sounds as [s]. No such effect was seen for the listeners who had heard the ambiguous sound in nonwords. Hence, Norris et al. (2003) showed that a period of exposure to sounds in a certain context could alter the listeners' percept, in a way that could simulate the process of 'tuning in' to a foreign accent.

Both of the approaches mentioned above have been used to investigate the *generalizability* of perceptual learning mechanisms. Put simply, if a listener can tune in to a particular accent, or an unusual pronunciation of a phoneme (e.g. a lisp), for one speaker, does the perceptual advantage remain for another speaker exhibiting similar speech patterns (note that for the particular purposes of the current experiment, the most relevant studies are those which assess the transfer of learning from one speaker to another i.e. *individual talker normalization*, rather than the effects of multitalker variability on a certain process (cf J. Warren et al., 2006; Magnuson & Nusbaum, 2007). Dupoux and Green (1997) carried out an experiment using time-compressed speech in which

they introduced a sudden change in speaker after 10 sentences of exposure to compressed sentences from one speaker. Although they found that there was an immediate dip in performance (which approached significance, within-subjects) after a speaker change, this dip was not to baseline levels and the overall effect of a speaker change was non-significant. This indicated that listeners were not using indexical speaker characteristics to tune in to compressed speech, and in turn therefore suggested that learning could be generalised across the two speakers (one male, one female) used in their experiment. Bradlow and Bent (2003) carried out a similar study involving transcription of Chinese-accented English by native US English listeners. In contrast to Dupoux and Green (1997), they found no transfer of learning when there was a change in speaker between training and test phases of the experiment.

Studies from the Norris et al. (2003) approach to perceptual learning have also produced some seemingly contradictory results. Eisner and McQueen (2005) found that training with fricatives produced by a female speaker did not transfer perceptual learning to a fricative continuum generated from a male speaker. Kraljic and Samuel (2005) found a more confusing result in their study of training with ambiguous fricatives, where perceptual learning generalized from female to male listeners, but not from male to female. They found that the female training stimuli were spectrally more similar to the male test (categorization) stimuli than the male training items were to the female test fricatives. Therefore, transfer of learning was based on acoustic similarity. The authors offer this as a possible explanation for their finding of full transfer from male training to female test items for perceptual learning of stops on a continuum from /d/ to /t/, as the differentiation of these consonants relies upon temporal cues that are more likely to be equivalent across speakers of different sex. Hence, Kraljic and Samuel (2005) propose an 'acoustically-grounded' mechanism for perceptual learning. Relating this proposal back to perceptual learning of globally-distorted speech stimuli, we can perhaps come to an explanation of the contradictory results of Dupoux and Green (1997) and Bradlow and Bent (2003). Time-compression is a primarily temporal manipulation involving the speeding of speech in the presence of preserved pitch and the relative durations of vocalic and intervocalic intervals. Assuming that listeners attend to the timing aspects of speech in the perceptual learning of compressed stimuli, and that these patterns should be relatively similar across male and female speakers of the same language (compared to spectral properties), then introducing a change in speaker during the learning period may not prove a threat to the learning trajectory. The small dip in performance observed by Dupoux and Green (1997) may have been nothing more than a global 'shock' reaction to the change in the speaker rather than reflecting an intrinsic change in the learning process. In contrast, the perception of foreign accent may be carried by non-native rhythmic properties in the speech, which may be similar across speakers, but there will also be important non-native aspects to vowel pronunciation. Relative formant frequencies

may be perceptually quite important in the normalization process for a single speaker of Chinese-accented English. These spectral properties of speech can vary substantially across speakers, and therefore the introduction of a speaker change for this type of speech 'distortion' may prove more desctructive to learning than for time-compressed stimuli.

The current experiment adopts the approach taken by Dupoux and Green (1997) in their study of perceptual adaptation to time-compressed speech. The training-test paradigm and materials used in Experiment 2 will be employed to compare the Test Phase recognition scores of listeners who experience a different speaker in Training and Test with Test scores from listeners who experience the same speaker throughout both phases. With only five bands, the noise-vocoded materials used in the experiment are greatly impoverished in spectral detail, yet rich in temporal envelope cues. Thus, the potential to discriminate the two speakers according to pitch and vowel space is likely to be very limited (see Gonzalez and Oliver (2005) for data on gender and identity processing with noise-vocoded speech). Given the low spectral resolution, and according to our interpretation above for the results of Dupoux and Green (1997) - which assumes that the two current talkers (both native speakers of British English) exhibit sufficiently similar speech rhythm - it is predicted that there should be no disruptive effect of a speaker change on perceptual learning.

5.2 Experiment 3

5.2.1 Method

Participants

Forty speakers of English (aged 18-40, 15 male), with no reported language or hearing problems, were recruited from the UCL Department of Psychology Subject Pool for participation in the experiment. Participants were assigned randomly to eight different conditions of the experiment (see below).

Materials

The sentences used were as in the English condition of Experiment 2. For the purposes of explaining the design of the current experiment, the 10 Training Phase sentences from Experiment 2 were

labelled Set A, and the 10 Test Phase sentences labelled Set B. All sentences were available in recordings made by the two English speakers from Experiment 2. These were both female speakers with Standard Southern British English accents. Speaker A, a university lecturer with extensive phonetic training and experience in making audio recordings, was 38 years old at the time of recording. Speaker B, a student of Speech Therapy, was 22 years old. There were no obviously distinguishable differences in accent or pronunciation between the two speakers; however, Speaker A's pitch was noticeably lower than that of Speaker B. The mean durations of the sentences in the Set A and Set B recordings were well matched across the two speakers, as shown in Table 5.1.

Table 5.1: Mean sentence durations (in seconds) for Set A and Set B across the two speakers.

	Set A		Set B	
	M	*SD*	*M*	*SD*
Speaker A	3.00	0.45	3.29	0.33
Speaker B	2.99	0.26	3.31	0.24

The five Practice Phase sentences were as in Experiment 2. Recording, digitizing and vocoding routines (for 5-band noise-vocoded stimuli) were as described for Experiment 2. The sentence lists were normalised for peak amplitude in PRAAT.

Design and Procedure

The design employed was between-subjects. As for Experiment 2, the task fell into three sections: a 5-item Practice Phase with undistorted speech, followed by two separate blocks of 10 noise-vocoded sentences. The main manipulation of interest was the Same/Different factor - 20 listeners experienced the same speaker for both Blocks of the experiment, while 20 heard a different speaker in each Block (i.e. half of the items spoken by Speaker A and half by Speaker B). The conditions were counterbalanced for Speaker Order and Sentence Block Order, thus giving a total of eight conditions. Sounds were played from a laptop computer through Sennheiser HD25-SP headphones. Volume settings were fixed at the same comfortable level for all listeners using the QuickMix software, version 1.06 (Product Technology Partners, Cambridge, UK). The listener heard each sentence in turn, and was instructed to write down as much as he/she could, using pen and paper. The experiment was 'self-timed' - listeners took as much time as needed to make their response, then triggered the beginning of the next trial by pressing the space bar on the keyboard. Each sentence was played once only.

5.2.2 Results

The sentence report responses were marked as in Experiment 2. The Practice Phase was marked in terms of the proportion of words reported correctly, using a conservative marking scheme that did not make allowances for deviations in inflectional morphology. The noise-vocoded sentence report responses were marked according to the number of words (function and content) correct, with a liberal marking scheme that allowed for morphological deviations that were due to tense and/or number agreement. For each participant, performance was recorded in terms of the proportion words correct for each item. Average scores were calculated for each of the two blocks. These scores then underwent arcsine transformation for use in statistical analysis.

The aim of the analysis was to test whether there was a statistically significant effect of a change in speaker on the relative improvement in performance from Block 1 to Block 2 - note that, unlike Experiment 2, Block here refers to a set of 10 consecutive sentences. The mean Block 1 scores were 0.37 ($SD = 0.17$) for the Different condition and 0.38 ($SD = 0.18$) for the Same condition, while the Block 2 means were 0.47 ($SD = 0.16$) for the Different condition and 0.50 ($SD = 0.14$) for listeners in the Same condition. Figure 5.1 shows a plot of the mean scores for each block in the two conditions.

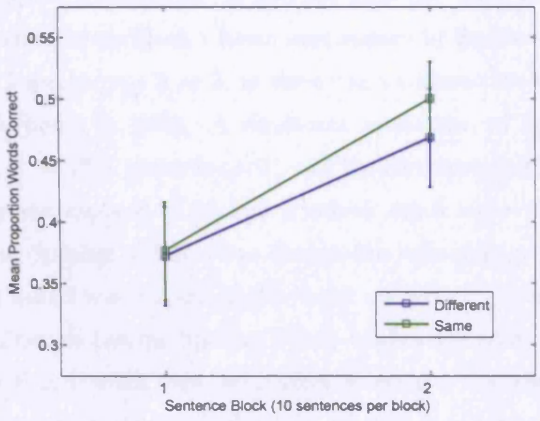


Figure 5.1: Plot of overall results in Experiment 3. Error bars show ±1 standard error of the mean.

As with Experiment 2, there is considerable variability in the scores. However, the mean scores indicate a numerical effect of the speaker change, as the listeners in the Different condition produced slightly lower Block 2 scores than listeners in the Same group. Overall, the Block 2 results are

higher than the equivalent mean score for the Test Phase in the English condition of Experiment 2. This may reflect possible Item or Speaker effects in Experiment 2, which have been accounted for by counter-balancing in the current experiment.

In order to formally test the effects of speaker change, and associated effects of speaker and items, a repeated-measures ANOVA was run using arcsine-transformed proportion scores for Block 1 (first 10 sentences) and Block 2 (final ten sentences). Block was the within-subjects variable and Condition (Same/Different), Speaker Order, and Sentence Order were between-subjects variables. There was a significant effect of Block ($F = 50.84$, $p = .000$, $\eta^2 = .614$, power = 1.00), suggesting that performance improves significantly between Block 1 and Block 2. The between-subjects effect of Condition was non-significant ($F<1$). There was also a significant effect of Speaker Order ($F = 9.35$, $p = .004$, $\eta^2 = .226$, power = .843). This suggests that there was an imbalance in intelligibility across the two speakers that interacted with adaptation to the noise-vocoded stimuli. However, the possibility of these effects was taken into account with counterbalancing. The critical result was the interaction of Condition with Block. If significant, this would suggest that changing speaker in the middle of the experiment has a significant effect on adaptation (i.e. the difference between Block 1 and Block 2). However, this interaction was found to be non-significant ($F<1$).

The effects of Speaker Order and Sentence Order were further explored through observations of the marginal means and interactions. It appears that the effect of Speaker Order reflected greater intelligibility scores when Block 1 items were spoken by Speaker B. This was not affected by whether the Block 2 speaker was A or B, as there was no interaction with Condition ($F = 1.16$, $p = .290$, $\eta^2 = .035$, power = .181). A significant interaction of Speaker Order with Block ($F = 15.72$, $p = .000$, $\eta^2 = .329$, power = .970), and the corresponding marginal means, indicate that listeners who are first exposed to Speaker B exhibit much higher Block 1 recognition scores than listeners who hear Speaker A first. Even though this advantage persists into Block 2 scores, the listeners who first hear Speaker A effectively 'catch up' with the other group by Block 2, such that it appears that listeners hearing Speaker B first exhibit less adaptation. This suggests that the speech of Speaker B is so much more supportive of learning that the subjects who heard this speaker first in the experiment approached a kind of ceiling, or upper asymptote, performance even by the end of Block 1. A significant interaction of Sentence Order and Block ($F = 4.27$, $p = .047$, $\eta^2 = .118$, power = .518) suggests that listeners who hear Sentence Set B in Block 1 exhibit the greater improvement from Block 1 to Block 2. Interestingly, there is a significant interaction between Speaker Order and Sentence Order, which indicates that the most favourable set of conditions for high recognition scores is to encounter Sentence Set B, spoken by Speaker B, in Block 1. In contrast, the least favourable combination is to have Sentence Set B spoken

by Speaker A in the first block. These two factors also gave a marginally significant three-way interaction with Block ($F = 3.60$, $p = .067$, $\eta^2 = .101$, power = .453). Given that Sentence Set B can lead to both the best and worst scores from a Block 1 position, depending on the speaker, it seems that Speaker Order is the dominant between-subjects effect in this experiment.

It was decided that a clearer interpretation of the results would be obtained by calculating mean scores for Sentence Set and Speaker separately without considering order effects. A table of means for Speaker vs. Sentence Set is shown in Table 5.2.

Table 5.2: Recognition scores, by Speaker and Sentence Set.

	Speaker A		Speaker B	
	M	*SD*	*M*	*SD*
Sentence Set A	0.42	0.16	0.47	0.17
Sentence Set B	0.31	0.13	0.52	0.16

It is clear that Speaker B is much more intelligible than Speaker A, but that the advantage interacts with Sentence Set. It appears that there are distinct differences in the way each speaker produces the sentences, and that this has a greater impact on intelligibility for Sentence Set B than Sentence Set A. Experiment 2 adopted a fixed order of Speaker and Sentence Set. In the English condition of Experiment 2, which is most comparable to the current experiment, Block 1 featured Sentence Set A and Speaker B, while Block 2 featured Sentence Set B and Speaker A. The means in Table 5.2 partly explain why there was only a small increase from Block 1 to Block 2 in Experiment 2.

It is clear from the data that any effect of a speaker change in the current experiment is certainly not devastating enough to reduce performance to baseline levels. In Dupoux and Green's (1997) study of perceptual adaptation to time-compressed speech, they found no evidence of an effect when they compared recognition scores on the 5 sentences before the speaker change to scores on the 5 following sentences. It was only when they narrowed in to the first two sentences after the change that they noticed a dip in performance, which approached significance and occurred only for those listeners who had experienced a change in speaker. It is possible that rapid adaptation within the first few sentences of the new speaker in the current experiment is the reason for the absence of a speaker change effect in the analysis using Block means. So, a similar analysis was run for the current data set, using the recognition scores from the first block of 10 sentences and the first two items from Block 2. Figure 5.2 shows a plot of the means, by Block and Condition. Indeed,

there appears to be a numerical drop in mean performance directly after Block 1 for the listeners who experienced a change in speaker, but not for those for whom the speaker stayed the same (who show a very slight improvement). However, there is considerable variability in performance. A repeated-measures ANOVA, with Block as the within-subjects variable and Condition, Speaker Order and Sentence Order as the between-subjects variables, was run on arcsined-transformed versions of these data, and showed non-significant effects of Block ($F<1$), Condition ($F<1$), and a non-significant interaction of Block and Condition ($F<1$). Two further paired-sample t-tests were run comparing scores before and after the speaker change point for each Condition separately. Both tests gave non-significant results with a corrected signifcance level of $p = .025$ (Different: $t(19) = .946$, $p = .356$; Same: $t(19) = -.299$, $p = .769$). Although the Different condition gave a larger t statistic, it is far outside significance. Inspection of the raw data supports this result, as only 11 of the 20 participants in the Different condition showed a drop in mean performance immediately after the speaker change.

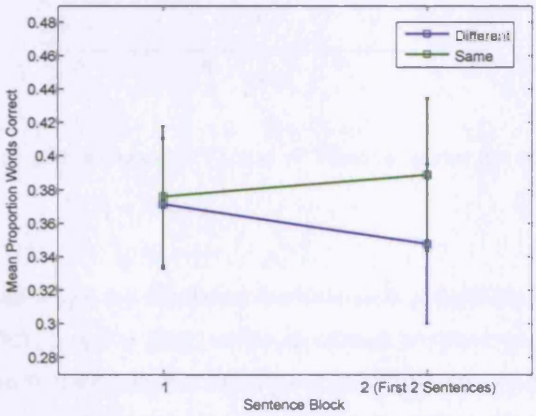


Figure 5.2: Plot of the change in mean recognition scores from Block 1 to the first two items in Block 2. Error bars show ±1 standard error of the mean.

Returning to issues discussed in the Introduction, and more central to Experiment 2, the four sentence sets used in the current experiment were labelled for vocalic and consonantal interval durations using PRAAT, in order to calculate the rhythmic properties of the materials. Despite the assumption that the two speakers in the current experiment should exhibit sufficiently similar within-class rhythmic properties to preclude any effect of rhythm on transfer of learning between the speakers, it remains that the English Training and Test stimuli from Experiment 2 were found to be significantly different on the size of the $nPVI$ metric (while still remaining in the 'stress-timed'

region). Values of $nPVI$ and $rPVInorm$ were calculated for each sentence from the two speakers in the current experiment, and entered into univariate ANOVA analyses with the sentences divided into different 'Conditions' according to the Speaker and Sentence Set employed in Experiment 3. Figure 5.3 shows a plot of both rhythm metrics for the four sets. As before, higher $nPVI$ and $rPVInorm$ values indicate more extreme stress-timing properties in the spoken materials.

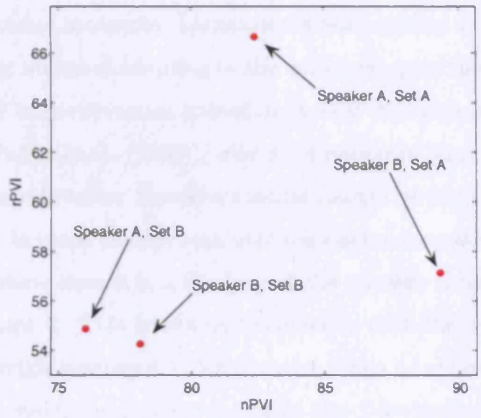


Figure 5.3: Scatterplot of the mean nPVI and rPVInorm values for the four sentence blocks (2 speakers, 2 sets).

An ANOVA with nPVI as the dependent variable gave a significant effect of Condition ($F = 3.21$, $p = .034$, $\eta^2 = .211$, power = .691), with a significant post-hoc comparison (Sidak-corrected) between Speaker A Set B and Speaker B Set A (p=.041). The second ANOVA, using rPVInorm as the dependent variable, also gave a significant effect of Condition ($F = 3.46$, $p = .026$, $\eta^2 = .224$, power = .728). In this analysis, there was a significant post-hoc comparison (Sidak-corrected) between Speaker A Set A and Speaker B Set B (p=.044), and a marginally significant comparison between between Speaker A Set A and Speaker A Set B (p=.062). Along a stricter rhythmic hypothesis of perceptual adaptation to noise-vocoded sentences, where an overall change in rhythm, even within-class, is predicted to disrupt learning, one should therefore expect that the these pairs of sentence sets should offer poorer improvements from Block 1 to Block 2 than other combinations where the differences in rhythm are non-significant between blocks. Unfortunately, there was not enough power in the data set to formally test the effect of rhythm on adaptation. However, the signficant post-hoc comparisons above, and the plot in Figure 5.3 indicate distinct rhythmic differences between Sentence Set A and B, across and within speakers. Given that every condition of the current experiment involved a change from Set A to Set B (regardless of the presence/absence

of a speaker change), these rhythmic differences may have had implications on learning that were untapped by the current experimental manipulation.

5.2.3 Discussion

It seems, from this data set, that there is no major effect of a sudden change in speaker on adaptation to noise-vocoded sentences. Listeners are not latching on to indexical characteristics of the speaker as a primary means of adapting to the sentences, even though there is a small numerical indication of a drop-off in performance immediately after the change. This informs the results of Experiment 2 (and of Pallier et al. (1998)), where the change of language from Training to Test was conflated with a change in speaker. Should a speaker change be disruptive to adaptation, this could have masked any possible small adaptive value of the foreign languages in Experiment 2. However, the current results indicate that it is unlikely that the speaker change interfered sufficiently with adaptation in Experiment 2. This leaves us to conclude that the most likely reason for a lack of adaptation with the foreign languages In Experiment 2 was a lack of meaningful linguistic content in the non-English materials, and possibly a lack of direct feedback (e.g. of undistorted or written sentence content) to assist learning. Based on the findings of Davis et al. (2005), it seems most likely that it is the absence of English words in the training sentences that prevented adaptation. The current results replicate the finding of Dupoux and Green (1997), who showed that perceptual adaptation to time-compressed speech is minimally disrupted by a sudden change in speaker. So, while Experiment 2 failed to find a cross-linguistic transfer of adaptation that had previously been demonstrated with time-compressed speech, Experiment 3 identifies a similarity in the perceptual learning processes involved with time-compressed and noise-vocoded sentence stimuli.

The discussion section of Chapter 4 addressed the possible influence of several procedural and design factors on the outcomes of Experiment 2. Amongst these was the possible disrupting effect of changing speaker on the trajectory of learning. The current experiment has addressed this factor, but some of the others remain. The main objectives of Experiment 2 were to test for perceptual learning in the absence of understanding (i.e. with foreign language stimuli) and to investigate the role of linguistic rhythm in this process by comparing the training efficacies of stress- and syllable-timed foreign languages. In Experiment 2, measurements of the rhythmic properties of the training and test materials of the English condition identified a significant difference in the $nPVI$ values (which represent variability of vocalic interval durations) between these two sentence sets. Although it was not possible to test the statistical effect of this difference on the rate of learning in the current experiment, measurements indicated more extreme stress-timing properties

in Sentence Set A than Set B across both speakers. This is an important point to consider in the design of these studies. In the collection of materials for Experiment 2, it was assumed that division of languages according to traditional linguistic rhythm class would be sufficient. Given that the current experiment served partly as a direct follow-up to the speaker change issue in Experiment 2, the choice of sentence materials had to remain the same. However, it is possible that this was an oversight in the design, as we have not yet tested the effects of systematic 'within-class' changes in rhythm on perception. The questions motivating Study 1 and Experiment 2 of this thesis placed more emphasis on rhythmic 'naturalness' than precise rhythmic properties. However, if the listener is using the rhythmic profile of the sentences as a segmentation and adaptation cue rather than a mere marker of naturalness, a systematic difference in rhythm between two testing blocks - as observed in both Experiments 2 and 3 - could have an effect on learning, even if this is a within-class difference. Global changes in linguistic rhythm may frequently arise as a consequence of a change in speaker, and so could be seen as an indexical characteristic. Taking this view, the fact that there was a change from Sentence Set A to Set B in all conditions of the current experiment (as items could not be repeated) means that the rhythmic differences between these item sets may have introduced a false cue to speaker change even in those conditions where the speaker remained the same for both sentence blocks. Hence, some of the effect of the deliberate speaker change (in the relevant conditions) could have been masked.

Another aspect of this discussion goes back to a point raised in the Discussion of Chapter 4; that is, that rhythmic properties of stimuli are not just affected by overall rhythm class, or the characteristics of the speaker, but also by the linguistic content of particular utterances. In Chapter 4 there was a suggestion that the significant cross-linguistic transfer of training seen with Dutch time-compressed speech in Pallier et al. (1998) may have arisen through the fortuitous selection of Dutch 'habituation' items that were particularly well-match rhythmically to the English items in the test phase, and in contrast this transfer may have been missed with noise-vocoded speech in Experiment 2 of this thesis through the selection of rhythmically unsuitable training items. A similar criticism may be levelled at the current experiment - the allocation of sentences to Set A and Set B may have come with a coincident group difference in sentence rhythm.

The above discussion of the rhythmic differences across sentence sets raises an important challenge to the research question at hand - if linguistic rhythm is so sensitive to item-to-item variations within rhythm class, then we would be forced to conclude that it is surely not a useful cue to perception. However, previous studies have shown that listeners can use rhythmic information effectively to identify word onsets in difficult listening conditions (M. Smith et al., 1989). There was nothing unusual about the way the sentences of the current experiment sounded (before distortion) when

read aloud by either speaker. This was not the case for the Nonword sentences in Davis et al. (2005) which were analysed in Experiment 1 of the thesis. In the current experiment, even in the face of coincidental rhythmic differences between the sentence sets, any variability in rhythm should not have presented much of a challenge to the course of perceptual adaptation. Along the argument of 'naturalness', which motivated the current experiment, it was assumed that all within-class variability in rhythm should be perceived as natural and expected within a sample of utterances from a language.

The next step in the process of investigating the role of linguistic rhythm in noise-vocoded sentence perception should be, however, to modify the experimental approach. To this point in the thesis, an emphasis has been placed on 'naturalness' in the stimuli, and with this has come an assumption that may not necessarily hold; that rhythmic variability within linguistic class should not threaten perception to the extent of cross-class variability. Furthermore, the use of foreign languages to investigate adaptation without lexical cues imposed certain constraints on design, such that a speaker change became necessary in Experiment 2, and the relatively small number of available items in the LSCP corpus limited the potential to match Sentence Sets along all the desired parameters (including rhythm). In the next experiment, some degree of naturalness must be sacrificed in order to get to the heart of the research question. It is important to have systematic control of the rhythms of the presented sentences, and to avoid extraneous design factors such as using multiple speakers. With naturalness no longer a priority, the simplest solution is to adopt a design where the same speaker produces all the stimuli, reading the test materials in both a natural manner (with no alteration on linguistic rhythm) and in a rhythmically-altered manner that can be systematically applied to all materials.

The results of the current experiment relate back to a study by Magnuson and Nusbaum (2007) which was mentioned in the Discussion of Chapter 4. Two interesting findings from their study bear relevance to the current discussion. First, they found that pairs of talkers that were more difficult to discriminate produced a smaller speaker variability effect (i.e. a cost to processing in the presence of changing speaker identity) in their word monitoring study. Second, they found that the size of the speaker variability effect for a pair of similar speakers was larger when the participants expected two speakers rather than one. A similar effect of expectations was observed by Newman and Evers (2007) in the context of a speech shadowing study, in which participants had to shadow a target voice (i.e. through immediate verbal repetition of this target talker's speech) while ignoring a distractor voice. Newman and Evers (2007) found that in a group of listeners who were all familiar with the target voice, those who were explicitly told the identity of the speaker before the experiment made fewer shadowing errors than those for whom the familiarity was implicit

(i.e. those who were not warned of the speaker's identity). In the current experiment, the listeners were not given any indication that they may be presented with stimuli from more than one speaker. Furthermore, although the information was not formally recorded, the vast majority of participants did not notice the change in speaker (when present) when questioned during the debriefing session. As soon as the task had finished, and before debriefing, several participants asked what had happened in the experiment, seemingly with no idea of the experimental manipulation. Given the findings of powerful top-down expectation effects in Magnuson and Nusbaum (2007) and Newman and Evers (2007), the lack of expectation of a change in this experiment, plus the fact that the presence of two speakers remained unnoticed by the participants, even post-hoc, may have been enough to prevent the emergence of a significant speaker change effect with the current stimuli. To address part of this issue, the next experiment directly measures the discriminability of the two speakers using the vocoded materials from Experiment 3, using a new set of listeners. A weak discriminability would add support to the theory that the listener's performance is less affected by a speaker change if he or she cannot consciously detect the difference between the speakers.

5.3 Experiment 4

5.3.1 Method

Participants

Participants were 9 adults (aged 18-40, 4 female), with no reported hearing, speech or language problems. Seven of the participants were native speakers of British English, while two were highly competent non-native speakers (whose native languages were German and Portuguese). The two non-native speakers were also experienced with noise-vocoded stimuli; however, they had not been previously presented with items from the LSCP corpus. All other participants were naive to noise-vocoded stimuli.

Materials

The materials used were the 20 items (where item refers to the linguistic content of the sentences and not the recorded tokens) used in Experiment 3, each spoken by the two speakers used in that experiment. Thus, there were 40 sentences in total. The sentences were all noise-vocoded to

5-bands as described in the Method section of Chapter 4.

Design

The experiment followed a basic 'same/different?' paradigm, in which participants were presented on each trial with a pair of items and asked to judge whether they had been produced by the same speaker or different speakers. There were 40 trials in total, of which 20 featured a single speaker while the other 20 featured two speakers. The items were counterbalanced for the speaker order (half of trials began with Speaker A, half with Speaker B) and sentence order (each item occurred twice in position 1 of a trial, and twice in position 2). To meet the design requirements, each sentence appeared twice from each speaker (once in position 1 and once in position 2). No trial featured a repetition of the same item. To minimise the ability of listeners to use recency information to make judgements on items that had occurred on previous trials, the list was manually altered to separate adjacent repetitions of items (i.e. the same item in position 2 of trial *n* followed by position 1 of trial *n+1*). This type of repetition only occurred once in the final ordering of items, which was the same for all participants, and in which the two trial types ('Same' and 'Different') were interleaved in a pseudorandom, unpredictable fashion.

Procedure

Participants sat at a laptop computer wearing headphones (Sennheiser HD25). They read onscreen instructions which told them that they would be presented with pairs of sentences, and that for each pair they would have to make a judgement of 'same' or 'different' (according to whether they heard one or two speakers, respectively). They were warned that the sentences would be distorted and that they should ignore the sentence content and instead attend to the voices. Responses were made by keypress, and participants received no feedback until after the last trial, when they received an onscreen accuracy score as a percentage. The participants were not told how many speakers would feature in the experiment to prevent them employing strategies to learn speaker identity or make discriminations based on specific tokens.

5.3.2 Results and Discussion

To account for response bias, scores were collated in terms of Hits, Misses, False Alarms and Correct Rejections, and used to calculate d' scores of speaker discriminability for each participant.

These are shown in Table 5.3 below. The mean d' score was 0.38, with a standard deviation of 0.26 indicating considerable individual variability. A one-sample t-test indicated, however, that the d' scores obtained were significantly greater than zero (which would indicate no discriminability) - $t(8) = 4.42$, $p = .002$[1]. Given that most psychophysical discrimination tests aim for a d' score of 1, while the maximum possible d' is 6.93, the results may suggest that the two speakers are not completely indicriminable, but the discrimination still remains very difficult. In light of the findings regarding expectations by Magnuson and Nusbaum (2007) and Newman and Evers (2007), the fact that the discrimination is so weak when participants are aware of the presence of more than one speaker and are attempting to discriminate them directly, this makes it quite unlikely that the unexpected speaker change in Experiment 3 had any perceptual effect on the participants. Given the literature on cochlear implant simulations, this is not surprising, as Gonzalez and Oliver (2005) found that discrimination based on gender alone was only around 70% with 5-channel noise-vocoded speech samples.

Table 5.3: Individual d' scores for discriminability of the two speakers in Experiment 3.

Participant	d'
1	0.08
2	0.52
3	0.16
4	0.37
5	0.28
6	0.23
7	0.42
8	0.39
9	0.96

A word should be said on the variability in performance with the speaker discrimination task. In the case of Participant 1, the speakers were almost completely indiscriminable, while performance for Participant 9 approached acceptable levels for psychophysical studies. Participant 9 was by far the best performer on this task, and also the most experienced with noise-vocoded speech. However, the majority of this experience was in the construction and performance of tasks to assess intelligibility of sentence content rather than speaker identification. During debriefing it emerged that Participant 9 had thought that both speakers in the discrimination task were male - this is perhaps some indication that his experience with the distortion type did not transfer any

[1]This significant result remained when the two experienced listeners were removed from the analysis

conscious perception of speaker differences, despite his good performance score.

5.4 Summary

The results of the two experiments described in this chapter suggest that the relatively modest increase in performance from Training to Test in the English condition of Experiment 2 was more likely a result of item effects than a destructive effect of a sudden speaker change on perceptual adaptation. While there is some suggestion that this may have been due to rhythmic difference between the sentence sets, this cannot be confirmed. Further evidence against the possibility of an effect of the speaker change came from a direct test of the discriminability of the two speakers (with noise-vocoded stimuli) showing very low d' scores across a sample of 9 listeners.

Future studies on the issue of transfer of learning across speakers in noise-vocoded speech should consider both the discriminability of the speakers presented, and the particular expectations of the listeners in the perceptual learning experiment. Eisner and McQueen (2005) showed that splicing male vowels onto female fricatives in the categorization phase of their experiment allowed transfer of learning from early exposure to the same female fricatives, despite the fact that the categorization stimuli were perceived as coming from a different speaker (i.e. male). Therefore, in their case, the effects of acoustics outweighed those of expectations. The current experiment assessed the effect of a speaker change in perceptual learning of 5-band noise-vocoded sentences. Increasing the spectral detail by vocoding to slightly higher band numbers will make available spectral cues that could contribute to speech intelligibility yet may still be insufficient to perform speaker discrimination. It is at such levels of spectral detail that the interaction of acoustics and expectations may emerge for this distortion type.

Chapter 6

Stimulus properties: The role of linguistic rhythm in English

Abstract

Experiment 5 assesses the effect of altered speech rhythm on the intelligibility of noise-vocoded English sentences. A within-subjects design with 24 participants directly compared recognition of 30 naturally-timed sentences with 30 sentences that were read in a 'metronomic' style. The results indicate a highly significant effect of the rhythmic manipuation on overall sentence intelligibility, but an equivocal effect on the rate of learning. A significant correlation between noise-vocoded sentence recognition and performance on the Seashore Rhythm Perception Test suggested a role for working memory capacity in recognition of degraded speech stimuli. The findings are discussed in terms of the challenges placed on perceptual encoding by noise-vocoding.

6.1 Introduction

The combined results of Experiments 2-4 support the main conclusion of Davis et al. (2005), that real word information is necessary in training materials to produce significant perceptual learning of noise-vocoded sentences. Before re-visiting the question of rhythm, therefore, real lexical information must be restored to the experimental materials.

A different beat - practical issues in altering linguistic rhythm

The appeal of using foreign languages in Experiment 2 was to maintain a naturalness in the spoken materials that had (as shown in Study 1) previously not been achieved with nonword sentences. One means of testing the role of rhythm using only English materials would be to contrast the intelligibility of two dialects of English that differ in their rhythmic properties. For example, Singapore English has been described as syllable-timed (Deterding, 2001; Grabe & Low, 2002; Low et al., 2000) and thus would form a rhythmic contrast to Standard Southern British English, which is stress-timed. Furthermore, the evidence from Experiment 3 suggests that the use of two speakers has a negligible effect on adaptation to noise-vocoded sentences. Unfortunately, as described in Chapter 4, the recordings of Singapore English that we obtained showed very similar rhythmic properties to British English and fell well within the range of the 'stress-timed' languages on measures of linguistic rhythm. It is likely that this was a consequence of the fact that the speaker had been in full-time residence and education in London for some time, and hence did not have such a strong accent as may be exhibited by individuals still resident in Singapore. It was anticipated that the same characteristics might apply to many of the Singaporean individuals within the University community. Therefore, it was decided not to obtain recordings from other local Singaporean speakers, as this would potentially be very time-consuming, with no guarantee of success. The only remaining alternative is to obtain alternative rhythmic conditions through the manipulation of sentences uttered by one speaker. In line with the majority of the available testing population, this speaker should be a native speaker of Standard Southern British English.

In breaking the rule of 'naturalness' in manipulating the rhythm of native speech, we are faced with the decision of whether to alter speech synthetically or organically i.e. whether to physically manipulate pre-recorded, naturally-timed materials, or to encourage the speaker to produce the chosen rhythmic changes herself. In the interest of maintaining some degree of naturalness (in terms of whether these rhythms could be produced organically), the latter option is favoured. Furthermore, it should be relatively straight-forward to get a speaker to produce speech that

adheres to the overall percept of 'syllable-timed' speech, that is, speech that sounds metronomic, as if all syllables in the sentence are of an equivalent duration. This approach may be assisted by using a speaker who has some experience of learning a language in which duration is phonemically relevant, for example Japanese. The rhythmic properties of the resultant sentence materials can then easily be measured using the approach taken in Study 1 and Experiments 2 and 3, in order to select the best examples for test.

Further design considerations - Item Difficulty, Individual Differences and the Aims of the Experiment

The interpretation of the results of earlier experiments in this thesis was made difficult by two apparent sources of variability - item effects and individual differences amongst the listeners. The sentences from the LSCP corpus (Nazzi et al., 1998), used in Experiment 2, are individually quite long (15-21 syllables), and while they could still be recognized with high accuracy in undistorted form, the recognition scores were not 100% in all cases. Furthermore, the aim of this experiment is to amass a sizeable sample of good examples of 'syllable-timed' sentences, so it is advantageous to record a large sample from which to pick the best exemplars after their rhythm has been measured. As the LSCP corpus offers a total of only 36 items in English, this limits the degree of selectivity that could be employed. A suitable alternative sentence corpus is the Harvard IEEE sentence set (IEEE, 1969), which comprises 720 items (72 lists of 10 sentences each) of relatively low predictability, each containing five keywords.

The striking individual variability in performance exhibited in both Experiment 2 and Experiment 3 forms the theme for the second strand of investigation in this thesis, as described in Experiments 2a, 6, 7 and 8. The previously observed variability also has important implications for the design of the current experiment. The between-subjects design in Experiments 2 and 3 meant that relatively large numbers of participants (64 and 40, respectively), needed to be included in order to achieve acceptable levels of statistical power on the between-subjects comparisons. In the current experiment, the primary aim is to test whether or not linguistic rhythm has an effect on the basic recognition of noise-vocoded sentences. Therefore, a within-subjects design is used. If the 'natural' sentences are shown to be more intelligible than the 'metronomic' sentences, then a further study could address, as in Experiments 2 and 3, whether pre-exposure to more intelligible materials produces superior training than pre-exposure to naturally-timed noise-vocoded sentences. However, a within-subjects design in which each participant is tested on recognition of both timing categories allows a comparison of the two sentence sets' intelligibility with increased power, while

also allowing a limited interpretation of the rates of learning exhibited for each sentence type.

In order to address the evident variability across listeners in recognition of noise-vocoded sentences, the current experiment features an additional test to provide a measure of individual variability on a cognitive task. The Seashore Rhythm Perception Test (Seashore, Lewis, & Saetreit, 1960) was chosen for this experiment as a measure of rhythmic processing capability, although historically it has also been routinely used as a test of attention and working memory (Ben-Yehudah & Ahisaar, 2004; Halstead, 1947). The trials of the Seashore test each involve a discrimination judgement on a pair of rhythmic sequences, played one after the other. It is predicted that better performance on the Seashore test will be associated with higher recognition scores on the noise-vocoded sentences, on the basis that sentence perception under difficult listening conditions should involve many of the processes involved in the Seashore task, for example sustained attention and memory for sequences of auditory information.

6.2 Experiment 5

6.2.1 Method

Participants

Participants were 24 normally-hearing adults (aged 18-40, 11 male) who spoke English as their first language, and who reported no speech, language or hearing problems. Participants were recruited from the UCL Department of Psychology Subject Pool. Half of the participants were randomly assigned to Version A of the Sentence Recognition Task, and the other half to Version B.

Materials

Sentence Recognition Task The test materials comprised sentences from the Harvard IEEE sentence corpus (IEEE, 1969). Each featured five 'keywords', for example: *'GLUE the SHEET to the DARK BLUE BACKGROUND'*. Sentences were recorded by a female native speaker of Standard Southern British English (aged 25). Each sentence was recorded in two modes - a 'natural' mode, in which the speaker was encouraged to read the sentences with natural rhythm and intonation, and a 'metronomic' mode, in which the speaker was instructed to assign equal duration to each syllable

whilst attempting to maintain natural intonation, as if she were speaking in time with a metronome. The speaker was not a phonetician but had extensive musical training and recent experience of learning a syllable-timed language (Indonesian).

Ninety sentences were recorded in the session, each uttered in both styles. Syllable onset and offset labels were added to each recorded sentence in PRAAT (Boersma & Weenink, 2005). A subset of 60 sentences was then chosen for use in the experiment according to a rate-normalised variation coefficient for syllable duration (*'varcoSyll'*, as calculated using Equation 6.1 (after Dellwo, 2007)) of the items. The reason for using a syllable duration metric, despite the evidence that it does not reliably differentiate languages on the basis of linguistic rhythm, is that it is the metric which most closely measures the feature that the speaker was asked to manipulate in the 'metronomic' versions of the sentence materials. Furthermore, the aim of the current experiment is not contrast different languages, but timing within the same language. For the final 60 items, the mean varcoSyll scores for the 'natural' and 'rhythmic' versions were 60.13 (SD = 6.11) and 31.17 (SD = 5.55), respectively. Further, all 'natural' recordings gave varcoSyll scores of over 50, while none of the 'metronomic' sentences had a varcoSyll score exceeding 42. A univariate ANOVA comparing *varcoSyll* scores across the two conditions produced a strongly significant effect of Condition ($F(1,59)$ = 1521.05, p = .000, η^2 = .963, power = 1.00), demonstrating that the 'metronomic' sentences demonstrated significantly less variability in syllable duration. In addition, the significant effect of Condition was maintained for subsequent univariate ANOVA analyses with vocalic $nPVI$ ($F(1,59)$ = 45.15, p = .000, η^2 = .434, power = 1.00) and intervocalic $rPVInorm$ ($F(1,59)$ = 65.89, p = .000, η^2 = .528, power = 1.00) as dependent variables, again indicating stronger tendencies toward syllable-timing in the 'metronomic' sentences. Table 6.1 shows the means and standard deviations for these three rhythm measures.

$$varco_{syll} = \frac{\Delta syll.100}{mean_{syll}}$$

where

$$\Delta syll = 100\sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}}$$

$mean_{syll}$ = mean duration of syllabic intervals

x = duration of syllabic interval

n = total number of sampled syllable intervals

(6.1)

Table 6.1: Mean rhythmic properties of the test sentences.

	varcoSyll		nPVI		rPVInorm	
	M	*SD*	*M*	*SD*	*M*	*SD*
Natural	60.13	6.11	96.87	13.84	60.19	13.92
Metronomic	31.17	5.55	81.71	11.31	45.33	14.53

An inevitable consequence of the required rhythmic manipulation in recording the sentences was that the 'metronomic' recordings were much slower than their 'natural' counterparts. The mean duration of the 'metronomic' items was 3.35 seconds (s.d. 0.42 sec) in comparison with a mean of 2.17 seconds (s.d. 0.30 sec) for the 'natural' recordings. It was decided to match the two recording styles item-for-item for duration, and in the interest of aiming for a speech rate as close to the speaker's natural rate as possible, all the rhythmic items were converted to the original file duration of their 'natural' counterparts. This was done by time-compression using the PSOLA algorithm (Charpentier & Stella, 1986) in PRAAT. To control for the degrading effects of this transformation, each of the 'natural' sentences was slowed by 17.5% and then returned to their original duration - both transformations were performed using PSOLA. The average time-compression for the 'metronomic' sentences was to 65% of the original duration, hence a 35% change, so half of this seemed a fair degree by which to initially lengthen the 'natural' sentences. If anything, the two transformations of the 'natural' sentences should be more destructive to the clarity of the speech signal than the single shortening of the already more clearly articulated 'metronomic' recordings.

Sentences were recorded and digitised as described in the Method section of Chapter 4. The sentence lists were normalised for peak amplitude in PRAAT and each item then noise-vocoded with 4 bands using the scheme described in the Method of Chapter 4. A distortion level of four bands was chosen to compensate for the likelihood that the IEEE corpus sentences would be slightly easier to recognise than the LSCP materials (Nazzi et al., 1998) used in Experiments 2 and 3, by virtue of being shorter and less complex syntactically. Thus, it was anticipated that a more degraded distortion level than that employed in the previous experiments would bring the mean recognition scores to a level similar to that obtained in those experiments. An example sentence in each rhythmic 'style' is included in the CD accompanying this thesis.

Seashore Rhythm Perception Task

Each trial of the Seashore Rhythm Perception Test involves a same/different judgement on a pair of rhythmic sequences. Each sequence comprises several notes of equal pitch, temporally ordered such as to produce a rhythmic musical phrase in duple time (i.e. where there could be 2 or 4 beats to each bar/measure) with a pulse of approximately 2Hz. The task falls into three Blocks - A, B and C - of 10 trials each. There are five 'Same' trials and five 'Different' trials in each block. However, the duration of the trials increases across the blocks, with a mean duration of 4.02 seconds ($SD = 0.02$) in Block A, 5.11 seconds ($SD = 0.02$) in Block B and 6.16 seconds ($SD = 0.01$) in Block C (each trial contains 200ms of silence at its beginning and end).

Design and Procedure

Each participant performed the tasks in the same order, completing the Sentence Recognition Task first and the Seashore Rhythm Perception Task second.

Sentence Recognition Task

The Sentence Recognition Task employed a within-subjects design in which each listener heard 30 'natural' and 30 'metronomic' 4-band noise-vocoded sentences. For each item, the participant heard the sentence once via Sennheiser HD25-SP headphones, and was then required to report the sentence content using the computer keyboard. The presentation level of the sentences was fixed at the same, comfortable level for all participants, using the QuickMix programversion 1.06 (Product Technology Partners, Cambridge, UK). Sentence report was self-timed; when the participant has completed his/her response for an item, a press on the space bar of the keyboard triggered the next sentence.

There was no overlap of items in the 'natural' and 'rhythmic' conditions, that is listeners heard a different set of sentences in each condition. To counterbalance for possible item effects, the sentences were divided into two sets - A and B. These were obtained by simply splitting the final item list down the middle. Half of the participants - those assigned to Version A - heard set A in the 'natural' condition, while the other half of participants - those assigned to Version B - heard set A in the 'rhythmic' condition. Items were presented to the participants via DMDX presentation software (University of Arizona, AZ). The order of presentation of all 60 items was pseudo-randomised such that no more than two items in a row came from the same condition.

Seashore Rhythm Perception Task

The Seashore task was run as a speeded judgement task to increase difficulty. Listeners were asked to press the left mouse button (positioned at the base of the laptop PC keyboard) with their left index finger to indicate that the two items in the sequence pair were the same, and to press the right mouse button with their right index finger to indicate a difference between the members of the sequence pair. Participants were encouraged to respond quickly and accurately, and were provided with feedback of their accuracy and speed after every trial.

The three Blocks of the test were presented in order from A to C, with the same order of items for each participant. Progression through each block was automatic and triggered by the participant's response, or after 8 seconds in the absence of a response. Listeners were given the opportunity to take a break between blocks. The test was resumed by pressing the space bar on the computer keyboard.

6.2.2 Results

Task scoring

Sentence Recognition Task

The sentences were scored in terms of the number of keywords reported correctly. As in Experiments 2 and 3, a relaxed marking scheme was adopted, where deviances in inflectional morphology, in the presence of the appropriate word stem, were not scored as incorrect. A score of Proportion Keywords Correct was calculated for each listener in each condition - this was further divided into scores for three chronological blocks of each condition, based on the total number of keywords correct for the first, second and third group of 10 sentences encountered in each condition.

Seashore Rhythm Perception Task

Each item was scored as correct or incorrect, and the total number of errors for each block, and for the overall task, were calculated for each participant. Errors included an incorrect response, and any 'time-out' trials (where the participant failed to give a response within 2 seconds of the stimulus offset).

Sentence Recognition Task - Effects of rhythm and learning

Figure 6.1 shows a plot of the mean recognition scores, in terms of the proportion of keywords correctly reported, for each of the sentence types, divided across the three blocks of 20 items (10 in each condition, in each Block).

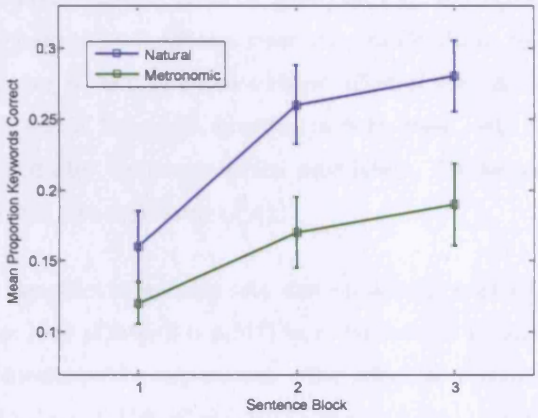


Figure 6.1: Mean sentence recognition scores across condition and block. Error bars show ±1 standard error of the mean.

The figure indicates a clear numerical separation of recognition scores for the natural and metronomic sentences, with the natural sentences showing greater intelligibility overall. However, the overall shape of the function characterising the rate of improvement over the three chronological sections of the Sentence Recognition Task appears similar for the two conditions, indicating a lack of effect of the rhythmic change on adaptation to noise-vocoded sentences (see below). The mean proportion of keywords correctly reported, of the 150 presented in each condition, was 0.23 (SD = 0.11) for the 'natural' sentences and 0.16 (SD = 0.09) for 'metronomic' sentences. As observed in other experiments of the thesis, there was considerable inter-individual variability in the scores, indicated by the standard deviations in the two conditions. The highest score for the natural sentences was 0.49, while that for the metronomic sentences was 0.42, while the corresponding minima were 0.06 (Natural) and 0.03 (Metronomic).

A repeated-measures ANOVA, with the within-subjects factors of Condition ('metronomic' versus 'natural') and Block (with three levels for each chronological third of the experiment) and the between-subjects factor of Version (A or B, according to which sentence set appeared in the

'metronomic' condition), was run to test for the effect of rhythmic disruption on the intelligibility of 4-band noise-vocoded sentences. The dependent variables of Mean Proportion Keywords Correct for Blocks 1 to 3 underwent an arcsine transformation before entry into the analysis. The ANOVA showed a significant effect of Condition ($F(1,22) = 45.71$, $p = .000$, $\eta^2 = 0.675$, power = 1.00) and of Block ($F(1,44) = 23.08$, $p = .000$, $\eta^2 = 0.512$, power = 1.00), but no significant effect of Version ($F<1$). There was no significant interaction of Condition by Block ($F(2,44) = 1.55$, $p = .224$, $\eta^2 = 0.066$, power = 0.312), which suggests that the two sentence types are learned at the same rate. There was, however, a significant interaction of Condition by Version ($F(1,22) = 5.14$, $p = .033$, $\eta^2 = .190$, power = .583) reflected a bigger effect of rhythm for Version B than Version A. However, the effect was in the same direction in both cases, with the metronomic sentences being less well recognised over the course of the experiment. Furthermore, the between-subjects main effect of Version was non-significant ($F<1$).

To further explore the effect of learning rate, three post-hoc repeated-measures ANOVAs (with a corrected significance level of $0.05/3 = 0.017$) were carried out to assess the effect of Condition in each of the three blocks of the experiment. The effect of Condition was non-significant in Block 1 ($F(1,22) = 2.63$, $p = 0.119$, $\eta^2 = 0.107$, power = 0.341), but significant in both Block 2 ($F(1,22) = 12.50$, $p = 0.002$, $\eta^2 = .362$, power = 0.922) and Block 3 ($F(1,22) = 18.41$, $p = .000$, $\eta^2 = 0.456$, power = 0.984). Interestingly, the F statistic and effect size become progressively larger from Block 1 to Block 3, suggesting that there is indeed a difference in the rate of adaptation between the two conditions (i.e. that listeners adapt more quickly to the 'natural' condition than the 'metronomic' condition).

Seashore Rhythm Perception Task and relationship to Sentence Recognition

The aims of the combined analysis of the Sentence Recognition and the Seashore Rhythm Perception task were:

1. To investigate whether there is a relationship between performance on the Seashore task and overall performance in the Sentence Recognition task.
2. To investigate whether there is a relationship between performance on the Seashore task and the extent of the effect of rhythmic disruption on sentence recognition (i.e. the difference between 'natural' and 'metronomic' conditions).

Looking at the Seashore Rhythm Perception Task in isolation first, the average number of errors across the listening population, by Block and overall, are shown in Figure 6.2. The increase in the number of errors from Block A to Block C reflects an overall effect of increased working memory load, as the rhythmic sequences increase in duration across the three blocks of the test.

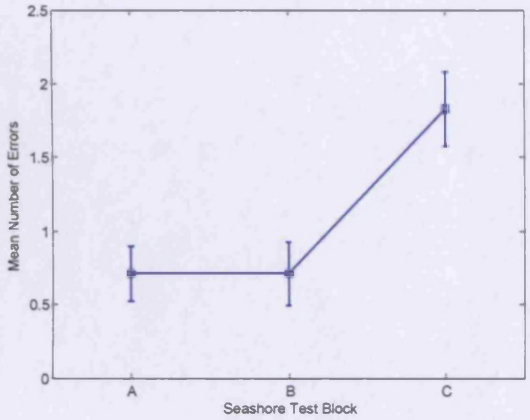
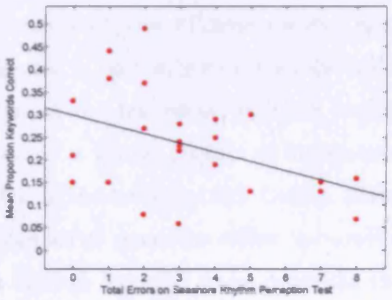


Figure 6.2: Mean error scores in the Seashore Rhythm Perception Test. Error bars show ±1 standard error of the mean.

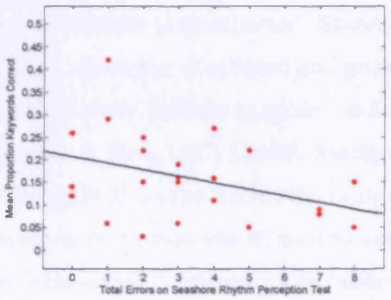
The following variables were entered into two-tailed, bivariate Pearson's correlations: Mean Proportion Keywords Correct in the 'natural' condition, Mean Proportion of Keywords Correct in the 'metronomic' condition, Overall Mean Proportion Keywords Correct (both conditions), Learning Effect for 'natural' sentences (increase in proportion scores from Block 1 to Block 3), Learning Effect for 'metronomic' sentences (increase in percentage points from Block 1 to Block 3), Overall Learning Effect (both conditions), Size of the Rhythmic Disruption (difference between the overall 'natural' and 'metronomic' scores, as a proportion of the 'natural' score), Total Errors on Seashore Task. Within the Sentence Recognition Task, significant correlations emerged between the Mean Proportion of Keywords Correct from the two conditions (Pearson's $r = .839$, $p = .000$), and between each of these measures and the Overall Mean Proportion Keywords Correct (Natural: Pearson's $r = .967$, $p = .000$; Metronomic: Pearson's $r = .950$, $p = .000$). There were no significant correlations involving the Learning Effect for natural sentences, other than the expected correlation with the Overall Learning Effect (Pearson's $r = .663$, $p = .000$). However, the Learning Effect for metronomic sentences was significantly correlated with all Mean Proportion Keywords Correct scores (Natural: Pearson's $r = .603$, $p = .002$; Metronomic: Pearson's $r = .616$, $p = .001$; Overall: Pearson's $r = .634$, $p = .001$). However, inspection of scatterplots revealed that these

positive correlations were created by one extreme data point. The scores for Size of the Rhythmic Disruption correlated significantly with the Mean Proportion Keywords Correct on the metronomic sentences (Pearson's $r = -.473$, $p = .020$). However, this correlation is also made significant by only one data point - the single listener for whom the metronomic sentences were more intelligible than the natural sentences.

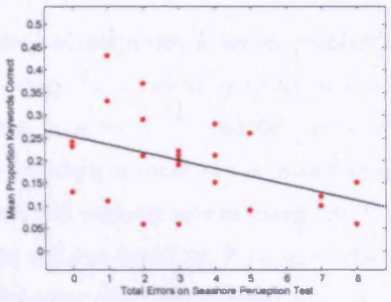
Turning to the relationship between the Sentence Recognition Task and the Seashore Task, there were significant and marginally-significant correlations between the Total Errors on the Seashore Task and the Mean Proportion Keywords Correct scores on the Sentence Recognition Task (Natural: Pearson's $r = -.436$, $p = .033$; Metronomic: Pearson's $r = -.374$, $p = .072$; Overall: Pearson's $r = -.425$, $p = .038$). Figure 6.3 shows the scatterplots for these significant correlations.



(a) Natural Sentences



(b) Metronomic Sentences



(c) Overall Sentence Recognition Performance

Figure 6.3: Scatterplots illustrating the correlations between performance on the Sentence Recognition Task and the Seashore Rhythm Perception Test.

The structure of the Seashore test offers a useful means of investigating the role of working memory capacity in the current experiment. The Seashore task comprises trials at three mean durations - approximately 4 seconds, 5 seconds and 6 seconds - which are collected together by

block. For each of the blocks (A-C) of the Seashore test, two-tailed Pearson's correlations were run between the number of errors and sentence recognition scores (i.e. the mean of all 30 sentences in each condition). For both 'natural' and 'metronomic' sentences, the size (and two-tailed significance) of the correlaton coefficient increases across the three blocks of the Seashore task (Natural: Block A - $r = -.149$, $p = .487$; Block B - $r = -.371$, $p = .074$; Block C - $r = -.428$, $p = .037$; Metronomic: Block A - $r = -.087$, $p = .687$; Block B - $r = -.300$, $p = .155$; Block C - $r = -.415$, $p = .044$).

6.2.3 Discussion

The results of the current experiment present evidence of a significant decrement in sentence intelligibility when the rhythm of noise-vocoded sentences is different than expected. Standard Southern British English is normally described as 'stress-timed', where the durational and prosodic properties of stressed and unstressed syllables are proposed by some authors to guide the listener to the informational i.e. lexical content of the sentence (Cutler & Foss, 1977; Cutler, Norris, & Williams, 1987; Cutler & Butterfield, 1992; Cutler, 1994a, 1994b). In this experiment, the same speaker produced examples of stress-timed (or 'natural') sentences, and examples of 'metronomic' sentences bearing a rhythm intended to be more like the 'syllable-timed' languages. The 'metronomic' sentences produced lower overall recognition scores than the 'natural' items, hence suggesting that linguistic rhythm is indeed a useful cue in the recognition of noise-vocoded speech.

How can the current results be interpreted along established theories of the role of rhythm in speech? The segmentation strategy put forward by Cutler and colleagues indicates an attentional role for the processing of rhythm in speech. The listener can, in effect, attend to linguistic rhythm in order to 'grab hold' of the linguistic content of the incoming speech signal. From an early age, this theory claims that listeners will segment speech using the strategy of their dominant language (e.g. stress-based for English, syllable-based for French; as described in Cutler, 1994b). Cutler (1994b) goes as far as to suggest that rhythmic segmentation is the child's means of breaking into the speech stream for the first time in order to develop their lexicon. Cutler and Foss (1977) suggest that the temporal organization of speech guides attention and enables predictions to be made as to the location of upcoming accents (and therefore important elements) in the speech stream.

Noise-vocoding compromises phonetic intelligibility, through 'smearing' of fine spectral detail. According to Cutler and colleagues, this should direct attention to the rhythmic structure of the sentence, and hence rhythmic segmentation. In the current experiment, half of the sentences ex-

hibited the expected 'stress-timed' rhythmic patterning of English, and so were directly accessible to the proposed native rhythmic segmentation strategy for English listeners. However, the 'metronomic' sentences showed a non-native patterning of syllable durations and hence would not be as readily segmented according to the strategy that some author propose is developed in early childhood. This could explain why the metronomic sentences were more difficult to understand in the current experiment, and produced lower mean recognition scores.

Could rhythm have an alternative role in the process of noise-vocoded speech recognition? There are some important aspects of the experimental design to consider. Very few of the listeners in the current experiment could, during debriefing, describe what it was that made some sentences more difficult than others. In other words, they had not noticed the unusual timing patterns on 50% of the trials. Of course, Cutler and colleagues do not make any requirement that rhythmic segmentation need be a conscious process, and on the basis of its proposed early acquisition we might expect that it would operate subconsciously anyway. If we consider the conclusions of Davis et al. (2005) and assume a lexical focus of attention in the listener, one might suggest that the listeners in the current experiment were consciously attending for real word information and not paying attention to rhythmic structure. An alternative interpretation, which is not mutually exclusive from the attentional hypothesis, is that whether or not the 'metronomic' rhythm obstructs the segmentation of the speech stream, it also affects the extent to which the listener can hold the sentence online in working memory. In each trial of the current experiment, the participant listened to a sentence and then wrote down as much of it as he/she could. Essentially, this is a recall task, albeit with no delay between stimulus and response - the listener was presented with a sentence that was unfamiliar and asked to immediately reproduce its linguistic content from memory.

With noise-vocoded speech, the degraded nature of the stimulus means that the mapping of acoustic-phonetic information to linguistic representations in the brain is not as automatic as it may be for clear speech. Burkholder, Pisoni, and Svirsky (2005) showed that digit spans in normal listeners were shorter when the digit stimuli had been degraded to simulate a cochlear implant with 8 electrodes than when they were presented undistorted. However, through separately measuring isolated digit intelligibility under distortion for each listener, Burkholder et al. were able to use intelligibility as a predictor of digit spans under degraded conditions. With only intelligibility as a predictor, they found a close agreement between the predicted and observed digit spans. This suggested that item errors during perceptual encoding are driving problems with recall of material. Therefore, it is important that the listener makes the most of the degraded input by fully encoding it in working memory to allow the subsequent mapping of sound to representations to take place. Leading on from this, an unexpected rhythmic pattern in a to-be-remembered spoken sentence

may compromise the encoding process. The role of temporal patterning in the encoding of to-be-remembered material is not new. It is a well-documented finding that the temporal grouping of an unfamiliar sequence of items, through the addition of extra pauses to a regularly-timed sequence to create several subsequences, facilitates sequence recall (Michon, 1964; Ryan, 1969a, 1969b). In anecdotal terms, one can relate this finding to the grouping of digits in telephone numbers when communicating them verbally to another individual, or when committing them to memory.

A very interesting study from the field of music cognition describes the effect of pitch and rhythm coherence on listeners' memory for melodies. Boltz (1998) carried out a study in which listeners were exposed to a learning phase, which contained musical phrases that either exhibited coherence between pitch and rhythmic accents, or were incoherent in this respect. After this phase, listeners were tested on their ability to estimate the total duration, reproduce the rhythm and recall the pitch of each melody. It was found that, for coherent melodies, accuracy on the memory tasks was similar no matter which dimension(s) the listeners had been instructed to attend during the learning phase - pitch, rhythm/duration, or both. However, for incoherent melodies, accuracy was dependent on the attended dimension(s) during the learning phase; for example, listeners who had attended to rhythm during learning exhibited lower scores on the pitch recall task. Boltz interprets the results in terms of a *structural remembering approach*, in which the coherence of temporal (i.e. rhythm) and non-temporal (i.e. pitch) information affects the success of encoding and thus subsequent recall of a melody. Although her theory is not intended to describe memory for speech, Boltz (1998) acknowledges studies that have shown similarities between music and speech (Jackendoff & Lerdahl, 1982; Jackendoff, 1989; J. Martin, 1972, n.d.), and proposes that the encoding of temporal and non-temporal information may produce similar results for speech as found in her study. More recent studies have, indeed, identified within-culture similarities in the rhythmic properties of language and music (e.g. for French and English: Patel & Daniele, 2003; Patel, 2006). Relating this back to the issue of noise-vocoded sentence perception, Boltz's (1998) theory would predict that, even though the listeners may not have noticed the abnormal rhythm of the 'metronomic' sentences during the task, and may be attending more closely to the non-temporal features of the speech stream (i.e. lexical information, as suggested by Davis et al., 2005), the fact that the rhythm of 'metronomic' sentences forms an 'incoherent' relationship to the linguistic content (according to what is expected for Southern Standard British English) presents difficulty for the perceptual encoding of the speech material and hence limits the accuracy of its subsequent recall.

Although not shown in the main ANOVA analysis in the current experiment, a set of post-hoc comparisons suggested that the effect of rhythm became more pronounced throughout the course

of the Sentence Recognition Task. In other words, recognition performance on the 'metronomic' sentences, while improving over the course of the task, was improving at a slower rate than observed for the 'natural' sentences. Boltz also has something to say on the matter of learning with regard to temporal and non-temporal information in the stimulus. She states that 'a structurally coherent event may not appear to be so when it is initially encountered' (1998, p. 1090). However, Boltz suggests that, after a sufficient period of learning, the coherence will become apparent and hence assist remembering. So it may be with the perception of noise-vocoded sentences in this experiment - in the first block of testing (20 sentences), there was no significant difference between the recognition scores for the 'metronomic' and 'natural' sentences. However, in the second and third blocks, the difference between the conditions widened. This, in Boltz interpretation, reflects the listeners' recognition, conscious or otherwise, of a coherence between the temporal and non-temporal information in the natural sentences, while the 'metronomic' sentences remain to some extent incoherent. However, it is not the case that the 'metronomic' sentences cannot be learned, as performance on these sentences did improve across the course of the experiment. It may be the case that some learning may be transferred from the improvements on the 'natural' sentences, for example the recognition of basic sound-to-representation mappings in the noise-vocoded stimuli. However, the incoherence of the metronomic sentences potentially placed a working memory-based limitation on the extent to which the learned information could be used to improve immediate sentence recall performance.

Further support for a working memory interpretation in the current experiment comes from the increase in the strength of correlations between performance on sentence recognition and the Seashore test as the Seashore trials become progressively longer in duration. However, the finding that these coefficients are not larger for the metronomic than the natural sentences suggests one of two things - either the fact that many of the metronomic scores are clustered around zero limits the range of scores for correlations, or working memory load is only telling part of the story.

6.3 Summary

The current experiment demonstrates an effect of altered linguistic rhythm on the recognition of noise-vocoded sentences, which impairs overall sentence intelligibility and slows the rate of adaptation. The cause of these impairments in performance is interpreted as a combination of increased difficulty in rhythmic segmentation of the speech stream, and a limitation on perceptual encoding caused by the incoherence of the temporal and non-temporal content of the metronomic sentences. The next step in this line of investigation would be to re-assess the effects of linguistic

rhythm in the current materials by reverting to the training-test approach taken in Experiments 2 and 3. This would illuminate whether the learning achieved in the 'metronomic' condition of the current experiment was due entirely to within-condition learning or whether these sentences benefitted from being interleaved with naturally timed items.

Given the considerable inter-individual variability in performance already observed in Experiments 2 and 3 of this thesis, a word should be said on variability with regard to the effect of rhythm in the current experiment. Having shown that rhythm has an effect at group level, an assessment of individual effects should give us an idea of the importance of linguistic rhythm as a general cue to noise-vocoded sentence perception. In the within-subjects design, 23 out of 24 listeners gave a lower recognition score on the 'metronomic' sentences than the 'natural' sentences. Thus, linguistic rhythm has some role in noise-vocoded sentence recognition for all listeners, but is it a critical factor in this process? Probably not, as the metronomic sentences were still recognised to some degree across the population. However, we can attempt to expore the *relative* importance of linguistic rhythm, even if it is not essential to perception. If linguistic rhythm were of great importance for noise-vocoded speech perception, then we might expect that those who use this cue most successfully will exhibit the highest scores in the population for 'natural' sentences. In turn, if the effect of disrupted rhythm cannot be circumvented, then these more successful listeners might be those most affected by the rhythmic manipulation (as those less skilled in the use of rhythmic segmentation may not be using rhythmic cues anyway so will be relatively unaffected). However, the Pearson's correlation between overall performance on the natural sentences and the effect of rhythm (as a proportion of the score on natural sentences) was non-significant. Hence, even if some listeners use rhythmic segmentation very successfully for coherent sentences, it is possible to either adapt to a new rhythmic pattern, or rely on other cues when the expected rhythm has changed. Again, these observations tie in with the overall interpretation discussed above with regard to the importance of coherence for the recall of sentences - the change in rhythm may not affect recall because it is in itself a primary cue, but because it affects the coherence of the sentence and thus the success of perceptual encoding.

Chapter 7

Listener variability: Correlates of speech recognition performance

Abstract

The study of individual differences in psychology is becoming increasingly important in our attempts to understand the behavioural and neural processes involved in certain abilities. This chapter motivates the study of individual variability in perception of noise-vocoded speech, and offers a brief review of the recent literature. Experiment 2a uses a battery of auditory, speech perception and general cognitive function tasks to identify significant correlates of individual variability in the noise-vocoded sentence recognition task described in Experiment 2. Sentence recognition data from thirty-three of the participants in Experiment 2 was entered into correlation and regression analyses with scores on the task battery. Significant correlations emerged with undistorted sentence repetition scores, vocabulary size and scores on a test of rhythm perception.

7.1 Introduction

One of the biggest challenges to the interpretation of results from Experiments 2-5 of this thesis has been the considerable individual variability in sentence recognition scores when listeners are presented with noise-vocoded materials. Table 7.1 shows the range of scores obtained by listeners in the different conditions of these experiments.

Table 7.1: Table summarizing the range of performances across Experiments 2-5 of the thesis. Data are proportion recognition scores for Experiments 2 (Test Phase), 3 (Block 2) and 5 (Overall), and d′ scores for Experiment 4.

Experiment	Condition	*Mean*	*Min*	*Max*	*IQR*
2	English	.44	.19	.72	.17
	Dutch	.31	.00	.52	.17
	Italian	.29	.03	.56	.32
	Control	.34	.00	.69	.35
3	Same	.50	.27	.75	.26
	Different	.47	.20	.73	.30
4	Speaker Discrimination	.38	.08	.96	.28
5	Natural	.24	.06	.49	.15
	Metronomic	.17	.05	.42	.13

The procedures employed in the sentence recognition tasks of Experiments 2, 3 and 5, including a lack of feedback to assist training and the deliberate decision to leave participants' expectations relatively open (e.g. by making no mention of foreign language stimuli in Experiment 2, nor of the potential change in speaker in Experiment 3), may have maximised the potential for individual variability in results. This is in contrast to the experiments of Davis et al. (2005), where the presence of feedback (via undistorted sentence repetitions) gave participants a much clearer idea of the structure of the task (e.g. speaker identity, linguistic content of the sentences) from the beginning and thus enabled clearer between-subjects comparisons with similar numbers of participants as used in the experiments of this thesis. However, this is not to say that the current studies should have included feedback, as there were theoretical motivations for the choice of procedures used. Furthermore, feedback is not necessarily indicative of better ecological validity. It is true

that we often have better contextual expectations in everyday listening situations. For example, a loudspeaker announcement in a busy train station is likely to contain the names of destinations, times and platform numbers. However, it is not that frequently that a noisy or muffled message will be repeated with perfect clarity (as in Davis et al. (2005)) in an everday setting - the train anncouncement may be repeated, but the ambient noise of trains and passengers is unlikely to completely disappear. Therefore, the extent of the variability described in the previous chapters of this thesis signals something worth investigating, despite the experimental conditions which generated it.

Individual Differences in psychology - A growing literature

The study of individual differences is not a new topic. Since the early days of psychology in the 19th century, much research has been occupied by the topic of intelligence and its constituent abilities (Carroll, 1993; Galton, 1869; Sternberg, 1977), and measures such as IQ, which quantify abilities purported to reflect intelligence, are routinely used in everyday life. From job recruitment procedures to mainstream entertainment, humans put great faith in the potential to compare individuals according to basic mental capabilities. However, in many studies in the field of experimental psychology, the experimental exercise of describing the processes involved in human abilities has focussed on group studies where individual differences are ignored amid large sample sizes. The individual differences approach to characterizing abilities through correlational analyses on different task scores claims to add to our understanding of cognitive processes by establishing relation or non-relation between different skills. These analyses prove the most enlightening when they associate functions previously thought to be diverse, and when they dissociate functions which were previously found to be inseparable experimentally.

Recent moves in the psychology of perception have attempted to re-introduce the study of individual differences as a standard experimental approach. As described in Chapter 1, Charles Watson and colleagues have made several attempts to characterize individual differences in auditory processing through the use of large auditory test batteries and factor analyses (Kidd et al., 2007; Surprenant & Watson, 2001) and through simpler correlations (Watson et al., 1996). This adopts the more domain-general approach of describing differences in a global 'ability'. Other studies have concentrated on using individual variability to uncover details on specific behaviours. In vision research, a recent study by Wilmer and Nakayama (2007) correlated measures of pre-saccadic and post-saccadic eye movements in the pursuit of a moving object (where accuracy is known to improve after this initial 'catch-up' saccade) with different measures of speed estimation (at fixation).

They found evidence that post-saccadic pursuit accuracy is driven by a different motion signal than pre-saccadic pursuit. Peterzell and colleagues have carried out a series of studies on contrast sensitivity, using individual differences in spatial frequency tuning functions to measure correlation between adjacent frequencies and hence identify 'covariance channels' across the tested frequency range. This allowed for the comparison of behaviours for different modulation types (red-green versus luminance gratings) and different age groups (Peterzell & Teller, 2000; Peterzell, Chang, & Teller, 2000). In visual object recognition, recent work by Yovel (2007) used a correlational analysis to assess whether individual visual object discrimination performance based on the spacing of constituent parts was independent of discrimination based on replacement of a constituent (e.g. altered spacing of eyes in a face versus replacement of the eyes with those from another face). For upright faces, there was a significant positive correlation between performance on the two types of discrmination, whereas for houses there was not. This suggests independent process of spacing and parts cues for houses, but interactive processing of these cues for faces.

The study of individual differences also allows for the identification of the neural substrates of cognition, through finding brain areas whose activations or size correlate with perceptual behaviour. Vogel and Machizawa (2004) have identified electrophysiological correlates of working memory capacity in adults; in a later study, the same group identified an index of the ability to exclude irrelevant items in a to-be-remembered sequence (Vogel, McCollough, & Machizawa, 2005). Using brain imaging techniques to measure structure and function, Golestani and colleagues have identified neural correlates of phonetic learning through correlational analyses and by comparing groups of fast and slow learners (Golestani et al., 2002; Golestani & Zatorre, 2004; Golestani et al., 2007). A recent voxel-based morphometry analysis indicated, amongst several findings, that listeners who faster to learn a non-native contrast (Hindi dental vs. retroflex consonant) greater volume in the left Heschl's gyrus area of auditory cortex, and a larger asymmetry in parietal lobe volume (where the left is greater than the right) compared with slower learners. There was also evidence for global changes in the positioning of right-hemisphere language-processing areas (Golestani et al., 2007). Functionally, correlational analyses indicated that faster learning, as tested behaviourally, is associated with greater processing efficiency in frontal language regions (Golestani & Zatorre, 2004).

Variability in Noise-vocoded speech recognition - Experiments of this thesis

The experiments described in this and following chapters are motivated by the observations of variability in the opening experiments of the thesis, and from previous experience of variability in recognition tasks with vocoded stimuli (Davis et al., 2005). As described above, individual variability in large data sets can be harnessed to great effect to uncover details of processes that might otherwise remain undetected with basic group comparisons of treatment against control. The use of noise-vocoded speech as a tool in perceptual studies is becoming more widespread. The neatly quantifiable method of vocoding with variable numbers of noise-bands, and the fact that it is a learnable stimulus, has made this method appealing for the investigation of speech perception (Shannon, 2007), and it has been used to great effect in several key studies in the general speech perception literature (Davis et al., 2005; Obleser et al., 2007; Scott et al., 2000; Shannon et al., 1995). In this respect, a greater understanding of variability in tasks using noise-vocoding will have important applications in the design of future studies. However, there is also important motivation for a general characterization of perception of noise-vocoded speech as a 'skill'. Noise-vocoding has been used in several studies using normal-hearing listeners to simulate the experience of a cochlear implant. Given the dramatic variability in speech perception and learning that has been documented for the implanted population (Munson et al., 2003; Pisoni, 2000; Sarant et al., 2001; Skinner, 2003), studies which attempt to characterize this variability are not only of interest but of clinical relevance. It is important to note, however, that the experiments of this thesis are not intended to make strong claims about (distorted) speech perception as a general ability (cf the approach of Kidd et al. (2007); Surprenant and Watson (2001); Watson et al. (1996). Noise-vocoding is only one method of speech distortion, which is physically quite different from other commonly-used methods such as sine-wave vocoding, time compression and addition of noise. Generalised claims will only be made if there is direct evidence for strong shared processing for noise-vocoded speech and several of these other speech stimuli.

The first experiment in this strand of the thesis takes a 'first pass' look at some of the possible factors involved in individual variability in noise-vocoded sentence recognition, through correlational and regression analyses. Having observed considerable variability in scores, betweeen and within conditions, on a pilot version of Experiment 2, it was decided that the final version of the experiment would be accompanied by a battery of tests measuring possible correlates of sentence recognition performance. Hence, the first experiment in this strand of the thesis is referred to as Experiment 2a. The subtests included in the battery were selected with the potential skill-set needed to recognise and learn noise-vocoded speech, rather than general speech stimuli, in mind.

Two of the tests came from the main content of Experiment 2. The main measure of interest is the overall recognition score from the Test Phase of the noise-vocoded sentence recognition task used in Experiment 2. The second score taken directly from the main design of Experiment 2 is a measure of Sentence Repetition in quiet, using the Practice Phase scores from that experiment to obtain a measure of undistorted sentence recognition with the same linguistic materials as used for the noise-vocoded task. In order to address the possibility that general intelligence might play a role, or indeed account for most of the variability, two relevant tests are included - a vocabulary test as a measure of verbal IQ, and a matrix reasoning task as a measure of performance IQ, both from the WAIS-III(UK) battery (Wechsler, 1997). It is predicted that verbal IQ will play a stronger role than performance IQ, given the evidence from Davis et al. (2005) for the importance of lexical information in adaptation to noise-vocoded speech, and the recent evidence for top-down contextual processing with noise-vocoded sentences of intermediate intelligibility (Obleser et al., 2007). In order to be able to address the possibility of domain-generality of the variability in noise-vocoded sentence perception, two further tasks are included involving speech perception under difficult conditions: recognition of sentences in noise, and speech-reading of sentences. Motivated by the outcomes of the large-scale studies by Watson and colleagues, which showed speech perception as an independent skill amongst low-level auditory tasks, it was decided not to place emphasis on basic auditory skills in the current battery. However, the fact that 5-channel noise-vocoded speech (as used in Experiment 2) is so greatly spectrally impoverished yet preserves the temporal envelope of speech opens up the possibility that listeners with greater skill in the use of envelope cues may produce higher sentence recognition scores. Thus, a test of amplitude modulation was included to assess this skill - an 8Hz modulating frequency was chosen as an intermediate value between the rates of envelope modulation for syllables (~3-4Hz) and segments (~16Hz). Finally, in order to address the role of rhythmic processing addressed in Experiments 2 and 5, the Seashore Rhythm Perception Test was included to complete the test battery.

This list of tests is by no means exhaustive and, in light of the results of earlier experiments in the thesis, certainly lacks some obvious inclusions, such as a simple test of working memory. However, the experiment was run alongside Experiment 2a and therefore should be seen as an introduction to a more detailed investigation of individual variability rather than a direct response to the results of all the preceding experiments of the thesis.

As a final word on the design of the current experiment, it must be borne in mind that it does not form a direct attempt to replicate the approaches of Watson and colleagues to variability in auditory processing. First, these experiments were conducted on a much larger scale and using more exhaustive assessments of auditory processing, which could not be replicated with the available

time and resources. Furthermore, the focus of attention in the current experiment is on one type of distorted speech - noise-vocoding - and so the selection of tests was motivated by directly relevant literature rather than an attempt to characterize general speech perception. With the emphasis on speech rather than more generalized auditory processing, the following studies do however offer a challenge to previous approaches to characterizing higher-level cognitive correlates of perception. Kidd et al. (2007); Surprenant and Watson (2001) included scores on school exams (the US Scholastic Aptitude Test) as measures of higher-level cognitive ability. As scores on these exams may be affected by the amount of preparation and practice performed by the student, and whether the student received coaching (Becker, 1990; Slack & Porter, 1980), the use of more direct tests at the time of administration of the remainder of the test battery may have been preferable. This is addressed in the present studies by taking new measures of more specifically chosen cognitive tasks.

7.2 Experiment 2a

7.2.1 Method

Participants

Thirty-three speakers of English (aged 18-40, 9 male) participated in the experiment. Testing occurred in the same session as the cross-linguistic noise-vocoded sentence recognition task described in Experiment 2, thus all the present participants were included in the Experiment 2 data set. None of the listeners presented with any known speech, hearing or language problems.

Materials

1. **Sentence Repetition** This task corresponds to the 'Practice Phase' of the sentence recognition task described in Experiment 2. Each listener heard clear recordings of 5 sentences from the LSCP corpus (Mean no. of syllables = 18).

2. **Noise-Vocoded Sentence Recognition** The test materials were ten 5-band noise-vocoded sentences in English, as described in the Method section of Chapter 4.

3. **Vocabulary task (verbal IQ)** This measure was acquired using the Vocabulary task from the Verbal subset of the WAIS-III(UK) test set (Wechsler, 1997). The test featured 33 single-

word items for verbal definition by the participant. These ranged from easy items such as *'winter'* and *'breakfast'* to more difficult words such as *'audacious'*, *'encumber'* and *'tirade'*.

4. **Non-verbal Reasoning (performance IQ)** This was assessed via the Matrices task of the WAIS-III(UK) performance subset. This task comprises 24 trials, each featuring an array of abstract pictures in which a missing picture is indicated by a black rectangle containing a question mark. Below this display is a set of potential 'replacement' pictures, of which one is most suitable to complete the array - this picture may complete an overall geometric pattern, or a sequence.

5. **Sentences-in-noise Recognition** Sentences were taken from the Bamford-Kowal-Bench (BKB) corpus (Bench, Kowal, & Bamford, 1979). They were available in .wav format, read by a female speaker of Standard Southern British English. Each sentence in the BKB corpus features three or four keywords keywords (shown here in upper case letters) and is of low semantic and syntactic complexity e.g. *'The CLOWN has a FUNNY FACE'* and *'THEY'RE BUYING some BREAD'*. Pronouns and content words are counted as keywords. Twenty sentences were available for use in the test, and all featured 3 keywords. The 'noise' was provided by a single .wav file of 'multi-talker babble'. The relative levels of the target sentence and babble were determined and assembled online, according to the trajectory of the adaptive track.

6. **Amplitude Modulation (AM) Detection** An adaptive tracking procedure was used to measure the 50% detection threshold for amplitude modulation of a white noise by an 8Hz sinewave. Ninety-nine stimuli were available along a logarithmic scale of modulation depths, ranging from 5% to 80% of the full depth of the sinewave. The comparison stimulus was the unmodulated white noise.

7. **Speechreading Task** Materials were video-only clips in .avi format of a female speaker of Standard Southern British English saying sentences from the BKB corpus. The speaker was camera-facing, with the head positioned centrally onscreen. Sixty-four sentences were included in task, in separate video files.

8. **Seashore Rhythm Perception Task** The design and procedure were as described in the Method section of Chapter 6.

   For the tasks using sentences from the BKB corpus (5 and 7), item lists were chosen such that there was no repetition of items between the tasks.

Design and Procedure

1. **Sentence Repetition** Each listener heard the same 5 sentences, in the same order, over headphones. After each sentence was played, the participant had to write down as much of the sentence as he/she could remember. Each sentence was played once only. There was no time limit on responses, and the listener had control, via keypress, over the time at which the next sentence was played.

2. **Noise-Vocoded Sentence Recognition** This task was run exactly as described in the Method section of Chapter 4. The 33 participants were sampled from all four conditions of the task (8 Dutch, 9 Italian, 9 English, 7 Control). The participant's task was to write down as much of each sentence as he/she could hear. Each sentence was played once only, and the pace of the experiment was controlled by the participant via keypress to trigger each trial.

3. **Vocabulary task (verbal IQ)** The participant was asked to give a full verbal definition of the test word in each trial. Test items were administered in the same order for all participants, beginning with Item 4. Items 1-3 were administered, in reverse order, only if the participant made a mistake on Items 4 or 5. For each trial, the listener was asked to give a full definition of the meaning of a word. Definitions were given to the experimenter via verbal descriptions and scored online by the experimenter. Testing was stopped when the participant obtained a score of zero on six consecutive trials.

4. **Non-verbal Reasoning (performance IQ)** In all tasks, the participant was asked to choose what they thought to be the most suitable picture to complete the sequence or array. There were three demonstration items (labelled A-C), which were not scored, and which were presented in the same order for all participants. Test items were also presented in a fixed order, starting from Item 4. As for the Vocabulary task, Items 1-3 were only presented if Items 4 or 5 were answered incorrectly. Testing was stopped if the participant gave incorrect responses on four consecutive trials, or on four out of five consecutive trials.

5. **Sentences-in-noise Recognition** An adaptive tracking procedure was used to characterize a threshold signal-to-noise ratio for identification of sentences in the presence of multitalker babble. The sentences were presented over headphones, in a fixed order for each listener. The listener was asked to give verbal report of as much of the sentence content as he/she could, while the experimenter scored their response, in terms of the number of keywords correct, by clicking labelled buttons shown on a GUI. The first sentence was presented at a signal-to-noise ratio (S:N) of +10dB, at which a correct answer is expected. In general, if the listener accurately identified all three keywords, the next stimulus was made more difficult. If the listener reported two out of three keywords, the next stimulus was presented at the same S:N,

while less accurate responses (1 or 0 correct keywords) resulted in the next stimulus being made easier (i.e. presented at a higher S:N). Hence, the procedure tracked a performance level of around 2 in 3 keywords correct. Several additional parameters were employed in the running of the track. The change in 'step size' from one stimulus to the next was varied over time, such that steps in the first downward run were 8dB, then after the first reversal (where the listener first scored less then 3 keywords correct; in other words, the 'turning points' of the track) became 5dB, then 2dB and remained at this level for the remainder of the track. In order that the thresholds would be calculated from a more stable portion of the track, the first two runs were discounted and only the remaining reversals were recorded. The number of presented items was limited to 20; however, the track stopped earlier if eight reversals had occurred (NB not including the first two runs).

6. **Amplitude Modulation Detection** An adaptive tracking procedure was used to measure detection thresholds for amplitude modulation (AM) of a noise-band by an 8Hz sinewave. This tracked AM detection at a level of approximately 79% trials correct. The track was run using the GLIMPSE (University College London, UK) software package, in an ABX paradigm. Participants sat at a computer, wearing headphones. A graphic presentation showed three cartoon frogs, each sitting on its own rock in a pond. Three stimuli were played, to correspond to a vocalisation from each frog (indicated by a simulatneous slight movement of the onscreen character in time with the sound). Two of the sounds were the unmodulated noise(the comparison stimulus), while one was modulated. The participant was instructed to choose the 'odd frog out' by mouse click, on the basis of sound alone. Feedback was given on each response, with correct selection shown by a 'check' mark onscreen and incorrect responses by an 'X'. The first three reversals were ignored for the analysis, and the number of following reversals limited to four. The initial step size was 15 points on the continuum, followed by 10.67, 6.33 and a final step size of 2. From the first reversal, the listener had to obtain three consecutive correct responses at each presented modulation depth before the track could move to a more difficult level.

7. **Speechreading Task** All 64 video clips were presented on a Dell desktop computer monitor in the same order for each participant, using DMDX presentation software. The participant's task was to write down as much of each sentence as he/she could, using pen and paper. As for tasks one and two, listeners' responses were not timed, and the participant could control the onset of stimuli by keypress. However, each stimulus could be viewed only once.

8. **Seashore Rhythm Perception Task** The design and procedure were as described in the Method section of Chapter 6, with one notable difference. In the current experiment, listeners were not given explicit instructions as to when they could make a response. Most participants

chose to wait until the second rhythmic sequence was complete, while a few made some responses earlier in the second sequence (probably as soon as they percived a difference).

7.2.2 Results

Scoring tasks

1. **Sentence Repetition** Performance on this task was calculated as the mean recognition score (in terms of proportion words correct) on the five Habituation Phase sentences from Experiment 2.

2. **Noise-vocoded Sentence Recognition** Individual scores were calculated as the mean Test Phase recognition performance (in terms of proportion words correct) on Experiment 2.

3. **Vocabulary Task** The participant's response to each item could be given a maximum of 2 points, with a point awarded for each piece of information the participant reported from a list of approved statements. A total score was calculated out of 66 (33 items) and converted to an age-normalized equivalent for each participant. Where Items 1-3 were not presented, it was assumed that these would be answered fully, with 2 points each.

4. **Non-verbal reasoning** Each item was scored as correct (1 point) or incorrect (zero) and a total score summed out of 24. Where Items 1-3 were not administered, it was assumed that these would be answered correctly. The total score was then converted to a normalized score using age-related norms.

5. **Sentences-in-noise Recognition** An overall threshold sentence recognition score was taken as the average of the last 8 reversal trials in the adaptive track and was expressed as a signal to noise ratio in dB, where a negative value indicated greater intensity of noise than signal.

6. **Amplitude Modulation Detection** A threshold measure representing the point on the continuum of stimuli where performance was approximately 79% correct was extracted for each participant by taking an average of the last four reversal trial values. These thresholds (T) were then converted into modulation depth proportion scores using the equation below.

$$T = MinMD.(\frac{MaxMD}{MinMD})^{\frac{(N-1)}{Tot-1}}$$

N = threshold continuum stimulus number

$MaxMD$ = Maximum modulation depth in continuum

$MinMD$ = Minimum modulation depth in continuum

Tot = Total number of items in stimulus continuum

7. **Speechreading Task** Individual scores were calculated as the total number of keywords correctly reported across the 64 sentences.

8. **Seashore Rhythm Perception Task** All incorrect responses, and responses made in excess if 2 seconds after the end of the stimululus pair, were counted as errors. The total number of errors was calculated for each participant.

Analyses

Table 7.2 shows descriptive statistics for the 33 participants' performances on the sub-tests of the battery. For the sub-conditions of the Noise-Vocoded Sentence Recognition test, please see Table 7.1.

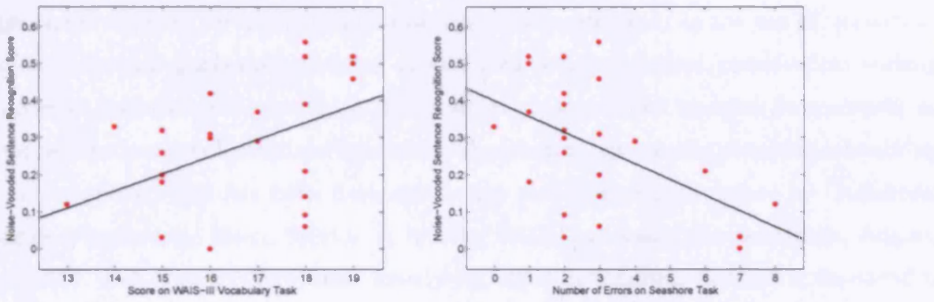
Table 7.2: Descriptive statistics for sub-test scores in Experiment 2a.

Task	*M*	*SD*	Best	Worst	*IQR*
Sentence Repetition (prop words correct)	.97	.04	1.00	.87	.06
Noise-Vocoded Sentence Recognition (prop words correct)	.31	.17	.56	.00	0.25
Vocabulary (age-normed score)	16.33	1.67	19.00	13.00	3.00
Matrices (age-normed score)	13.00	1.82	16.00	10.00	2.50
Sentences-in-noise Recognition (threshold S:N)	-2.48	.83	-4.25	-.50	1.13
AM Detection (threshold modulation depth)	.16	.08	.05	.44	.07
Speechreading (number keywords correct)	41.00	29.80	119.00	3.00	44.00
Seashore Rhythm Perception Task (number of errors)	3.12	2.16	.00	8.00	2.00

A step-wise multiple linear regression analysis was carried out, with scores on the Noise-Vocoded Sentence Recognition task as the dependent variable and the other scores, plus Age in years, as predictors. Dummy variables were entered to account for the Condition which each listener experienced in Experiment 2 (English, Dutch, Italian or Control). The analysis produced one significant model ($F = 4.91$, $p = .034$) with scores on the Sentence Repetition task as the only predictor. This model accounted for 10.9% of the variance. Two-tailed Pearson's correlations were also run between the subtest scores for all 33 participants. These produced significant correlations between Noise-Vocoded Sentence Recognition scores and scores on the Sentence Repetition task ($r = .370$, $p = .034$, 2-tailed), Vocabulary task ($r = .348$, $p = .047$, 2-tailed) and the Seashore task ($r = -.344$, $p = .050$, 2-tailed), all of which indicated that better performance on one task was associated with better performance on the other. There were also significant correlations between performance on the Sentence Repetition task (clear speech) and scores on the Vocabulary ($r = .435$, $p = .011$, 2-tailed) and Seashore ($r = -.422$, $p = .014$, 2-tailed) tasks.

As Experiment 2 had indicated an advantage for participants in the English condition, who

received 10 more noise-vocoded English sentences than participants in the other conditions, this factor may cloud the pattern of correlations above. Therefore, the same analyses were run for the 24 participants from the remaining conditions (Dutch, Italian, Control) of Experiment 2, as these can be assumed to be equivalent. There were signficant two-tailed correlations between noise-vocoded sentence recognition and scores on the Vocabulary ($r = .463$, $p = .023$, 2-tailed) and Seashore ($r = -.519$, $p = .009$, 2-tailed) tasks, but no longer with Sentence Repetition ($r = .319$, $p = 0.129$, 2-tailed). As above, the Sentence Repetition scores were also significantly correlated with Vocabulary ($r = .580$, $p = .003$, 2-tailed) and Seashore ($r = -.599$, $p = .002$, 2-tailed) task performance. Figure 7.1 shows the significant correlations involving noise-vocoded sentence recognition performance for this 24-participant sub-group.



(a) WAIS Vocabulary Test (b) Seashore Rhythm Perception Test

Figure 7.1: Scatterplots illustrating significant correlations between noise-vocoded sentence recognition scores and performance on other sub-tests of Experiment 2a.

7.2.3 Discussion

The results of the current experiment are, in some respects, theoretically intuitive and in line with previous findings. However, other results, particularly those that are noticeable in their absence, add complexity to the interpretation.

The finding that performances on both the noise-vocoded sentence recognition task and the clear sentence repetition task correlated significantly with Vocabulary and Seashore task scores indicates that these two sentence perception tasks share some component(s). The most immediate conclusion as to the nature of this shared processing is that it reflects verbal intelligence, as the Vocabulary task tests this directly while the Seashore task is known to load on working memory and attention. However, the finding that the two sentences tasks do not always correlate significantly with each

other suggests that there are considerable non-overlapping features of the tasks. In particular, that the noise-vocoded sentence scores produce weaker correlations with the Seashore and Vocabulary scores than the Sentence Repetition task indicates that the recognition of the distorted speech samples has added components. When the significant correlations involving noise-vocoded speech recognition were re-measured as partial correlations in which Sentence Repetition performance was controlled, the correlation with Vocabulary became marginally significant ($r = .361$, $p = .091$) and the correlation with Seashore task performance was weakened ($r = -.432$, $p = .039$). This indicates that part, but not all, of the relationships with these tasks can be accounted for by the basic requirements of sentence perception in quiet, such as working memory capacity and attention. This finding also arose from the linear regression analysis, which gave Sentence Repetition scores as the predictor in the only emergent model. The remainder of the relationships may tap more involved processes required for the perception of distorted speech, such as the use of 'top-down' contextual information to generate hypotheses about lexical sentence content, phonological working memory processes (including the encoding of distorted speech sounds and mapping from acoustic to linguistic representations), and sustained attention. The relationship between phonological working memory and vocabulary size has been described in the developmental literature by Gathercole and colleagues (Gathercole, Hitch, Service, & Martin, 1997; Gathercole, Service, Hitch, Adams, & Martin, 1999), and some authors have already identified correlations between noise-vocoded sentence recognition and working memory capacity (Eisenberg et al., 2000; Chiu et al., 2002).

The more striking findings in the current data set are the non-significant correlations where relationships were expected. Surprenant and Watson (2001) found that a factor analysis on individual scores on their battery of subtests gave a separate factor for their speech-in-noise and syllable identification tasks. Watson et al. (1996) went a step further by showing that this 'special' speech factor could be amodal by identifying a modest correlation between individual scores on auditory and visual-only (speech-reading) sentence recognition tasks. Most recently, Kidd et al. (2007) showed that this factor may not be restricted to speech, as a test of environmental sound recognition loaded onto the same factor as the speech tasks in their study. This lead them to propose a 'Familiar Sound Recognition (FSR)' ability to describe this specialised listening mode. However, in the current study, there is little evidence for relationship between the speech recognition tasks (noise-vocoded sentences, clear auditory sentences, sentences in noise, visual-only sentences (speech-reading)) beyond the significant correlation described above. There are several possible reasons for this. First, the participant sample for correlations is small ($N = 33/24$) compared with, for example, the Kidd et al. (2007) study with 340 participants. Therefore, any noise in the data stands to pose a much greater threat to the significance of correlations. This can be seen from the fact that the relationship between sentences-in-noise and speech-reading recognition scores appears

to demonstrate the appropriate direction ($r = -.266$; note that more negative S:N threshold scores indicate better performance), and a sensible supporting scatterplot (see Figure 7.2), yet fails to reach significance ($p = .209$, 2-tailed).

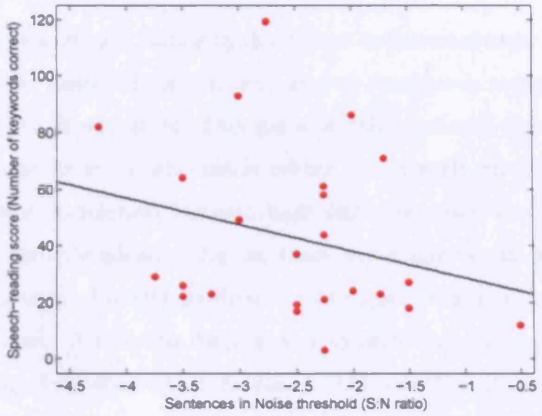


Figure 7.2: Scatterplot of Speech-reading and Sentences-in-noise scores (N=24).

A second potential factor in limiting the inter-relationship of speech recognition tasks in the current experiment is the linguistic difficulty of the materials, which varied from task to task. The sentences in the Sentence Repetition and Noise-vocoded Sentence Recognition tasks were taken from the LSCP corpus, as described in Chapter 4, and included items such as *The parents quietly crossed the dark room and approached the boy's bed* and *Finding a job is difficult in the present economic climate.* In contrast, the BKB corpus items used for the speech-reading and sentences-in-noise tasks were much shorter and less complex, e.g. *The glass bowl broke* and *The dog played with a stick.* There is evidence that, under degraded conditions, listeners make use of contextual information to generate lexical candidates for speech recognition (Kalikow, Stevens, & Elliott, 1977; Obleser et al., 2007; Hannemann et al., 2007). Furthermore, there is evidence that degradation of speech through noise-vocoding decreases working memory spans by increasing the difficulty of perceptual encoding (Burkholder et al., 2005). Thus, it is likely that the longer, less semantically predictable sentences of the LSCP are processed in a different manner than the short, semantically coherent BKB sentences under difficult listening conditions. As an extension of this, the degree of degradation being measured was not equivalent across the different speech tasks employed in the current experiment. In addition to differences in task difficulty, the qualitative perceptual experience of a 5-channel noise-vocoded sentence is quite different from an undistorted sentence, a sentence in noise, or a silent visual sentence. A primary reason for this is a difference in familiarity,

with noise-vocoded speech being perhaps less ecologically valid for normal-hearing listeners than the other conditions. For this reason, noise-vocoded speech may not fall so readily into an 'FSR' group with other speech tasks, even if all the tasks were matched for difficultly and employed the same linguistic materials.

Finally, another non-significant finding in the current experiment came from the lack of observed relationship between AM detection at 8Hz and any of the speech recognition measures. It was anticipated that the good preservation of temporal envelope cues in noise-vocoded speech would prove beneficial for those listeners who made better use of such cues, and hence this would be exhibited in a significant correlation between high scores on noise-vocoded sentence recognition and smaller AM detection thresholds. Again, there are a number of possible reasons why this relationship did not emerge. An 8Hz modulating frequency was used in the current experiment as an intermediate between 3-4Hz, the frequency of syllabic envelope cues in speech, and around 16Hz, the corresponding frequency for segments. If AM detection in humans is tuned for speech recognition and is thus frequency-specific, then using 8Hz in the current experiment may resulted in the failed detection of an existing relationship. However, the selection of 8Hz is defended in this instance because it was important to limit the duration of the testing session and allow for the inclusion of the other selected sub-tests. An alternative explanation for the lack of relationship is not that AM detection is of no importance in the perception of degraded speech, or noise-vocoded speech in particular, but simply that it does not account for the *variability* in this process in the normal-hearing adult population. The evidence of 'specialised' processing for speech by Watson and colleagues, and their inability to find convincing relationships between variability in speech recognition and auditory skills, plus the evidence for top-down influences in the current experiment and previous work on perceptual learning (Davis et al., 2005) and noise-vocoded speech recognition (Obleser et al., 2007), suggests that variability in noise-vocoded speech perception may more likely be driven by higher-level cognitive processes rather than differences in use of 'bottom-up' acoustic cues.

7.3 Summary

The results of the current experiment present two topics for research in the remainder of the thesis. The first is the role of higher-level cognition in the recognition of, and perceptual adaptation to, noise-vocoded speech. The present results suggest a role for working memory, through the significant correlations observed between noise-vocoded sentence recognition and performance on the Seashore task. However, the reliability of the Seashore task has been questioned by some authors

(Charter & Webster, 1997; Sherer, Parsons, Nixon, & Adams, 1991). The next experiment of the thesis will address the issue by taking direct measures of working memory with more established and widely-used tests.

The second topic to be addressed is the role of linguistic information in the perception (i.e. recognition and perceptual adaptation processes) of noise-vocoded speech. Sentences are the commonly-used stimulus to measure speech recognition, but the results of the current experiment indicate that even different styles of sentence can produce quite different patterns of performance (as indicated by the lack of significant correlation between tasks using the LSCP and BKB sentences). With regard to variability in perceptual processes, it is of interest to see whether the top-down processing proposed by the current study and others can be fragmented, and whether by investigating all levels of linguistic processing - from segment to word to sentence - we can identify some role for lower-level processes in accounting for the variability in the normal-hearing listening population.

Although it has not formed part of the current discussion, an important finding in Experiment 2 was that some listeners were performing at floor on the noise-vocoded sentence recognition tasks. In the interests of uncovering all the potential variability in the listening population, with reduced threat of floor or ceiling effects, it is more appealing to describe performance at a certain threshold level (e.g. 50% correct) in terms of the amount of spectral clarity (i.e. the number of bands) needed to achieve the threshold score. Not only does this approach potentially give a more realistic account of individual variability, but it also allows for easier comparability of performance on different noise-vocoded speech perception tasks, as intended for the investigation of linguistic effects.

Chapter 8

Listener variability: Two experimental approaches

Abstract

This chapter addresses the issue of individual variability in recognition of, and adaptation to, noise-vocoded speech. Two experimental approaches - adaptive tracking (Experiment 6) and constant measures (Experiment 7) - were used to quantify individual differences in perception of noise-vocoded sentences. The speech recognition measures obtained were analysed for correlations with scores on tests of verbal IQ and working memory. Evaluation of the two approaches identified the constant measures method as the more informative of the two, while the correlations with cognitive measures suggest a possible role for phonological working memory for familiar and unfamiliar material in describing variability in perception of distorted speech.

8.1 Introduction

Individual Variability - The Problems for Adaptation Studies

When making between-group comparisons of perceptual adaptation effects (Davis et al., 2005; Pallier et al., 1998; Sebastian-Galles et al., 2000), it is important to take account of within-group variation in perceptual ability. Aside from the possibility of differing adaptation rates across listeners, there are potential differences in individual speakers' raw or baseline perceptual capabilities. If large variability occurs within a group, this can mask potential between-group comparisons.

There are several ways to address this issue. First, one could obtain a baseline speech recognition score before training, and then quantify the amount of learning as a percentage change in recognition performance from baseline to test phase. Davis et al.'s (2005) study of perceptual learning of 6-band noise-vocoded sentences used a paradigm comprising a training phase and a test phase to compare the training effects of feedback style and sentence content on adaptation to noise-vocoded sentences. Davis et al. (2005) have been criticised by Burkholder (2005) because they neglected to obtain a pre-training measure of noise-vocoded sentence recognition for each listener. Thus, one could argue that any test-phase effects of training condition could be a consequence of participant sampling, and may bear little reflection of the adaptive properties of the training materials.

However, a defence of the Davis et al. (2005) approach is that adaptation to noise-vocoded speech is a relatively rapid process. It is difficult to construct a pre-test of perception that is a reliable measure of baseline capabilities but still leaves scope for further learning. Davis et al. found that recognition scores for 6-band noise-vocoded sentence recognition went from 0% words correct to 70% words correct over only 30 sentences. In contrast, Burkholder (2005) observes 8-band recognition performance that does not reach ceiling, even after the participant has performed feedback on over 150 sentence items. This difference in performance is likely to reflect that fact that Burkholder's noise-vocoded stimuli had a range of 854Hz-11000Hz (the raised lower limit reflecting incomplete insertion of the electrode array of a cochlear implant), while Davis et al. divided their frequency spectrum in the range 50Hz-8000Hz. The lost information in critical low-frequency regions, which are useful for extraction of first formant information, in Burkholder's vocoding routine is likely to have made her stimuli more difficult than those used by Davis et al., despite her use of a larger overall range and number of bands. A more challenging stimulus set may have kept performance below ceiling in the Davis et al. study.

In both Davis et al. (2005) and Burkholder (2005), listeners were exposed to a fixed level of distortion - 6-band and 8-band noise-vocoding, respectively. This adds a further difficulty for both of the above studies' approaches to calculating and comparing training effects. Use of a fixed distortion level cannot overcome the possibility of floor or ceiling performance in some listeners. For both of these extremes of performance, the fixed-level approach could result in little to no training; a strong listener's performance may already be perfect at pre-test, while repetition of very difficult stimuli may be insufficient to facilitate any improvement in a weak listener's performance. Furthermore, any increase in performance from a very low starting level will be disproportionately large compared with the potential improvements for more average listeners, hence the effects of individual differences can potentially still disrupt the between-group comparisons of interest.

A second means of controlling for individual differences in group adaptation studies is to ensure that all listeners begin the training phase at a similar level of performance. If a fixed level of distortion is to be used, this can be done in two ways:

1. Training all listeners to a criterion performance on the chosen distortion level before dividing them into further training conditions. The weakness of this approach is that it will result in imbalance in the amount of exposure to the distortion across individuals - some listeners will take much longer than others to reach criterion performance.

2. Performing a pre-test using a range of distortion levels and choosing, for each participant, the level required to produce a criterion performance. However, this presents complications associated with inconsistencies in the acoustic stimulus presented to each individual, and across conditions, in any subsequent independent-groups design.

A third possible means of overcoming the disruptive effects of individual variability in adaptation studies is to obtain a more comprehensive profile of behaviour in the main experiment that covers the whole range of performance from floor to ceiling recognition, and to observe the effects of training across this range for each listener. This can be achieved easily by varying the number of bands in the noise-vocoded stimulus and presenting the participant with stimuli from a number of distortion levels. Shannon et al. (2004) performed a meta-analysis of several perception studies using cochlear implant simulations such as noise-vocoding that tested participants across a range of distortion levels. They showed that sigmoidal curves describing a logarithmic relationship between the number of bands and performance could be fitted to recognition data in tasks such as sentence recognition and melody identification, for different listening populations. Shannon et al. (2004) noted that increasing the difficulty of the listening situation, either through changing the stimulus, increasing task complexity or by changing the listening population (for example from

normal-hearing adult native speakers to children or non-native speakers), shifts the sigmoidal performance curve to the right i.e. listeners need greater stimulus clarity to achieve the same level of performance. In a study of adaptation effects, it is feasible to suggest that adaptation could be quantified in terms of the shift of these performance curves. This not only helps to control for different starting levels of performance across individuals, but may also contribute to the understanding of how different rates of learning may emerge on a listener-by-listener basis; perhaps certain distortion levels are more optimal for learning than others, so covering a range of levels can account for such effects.

Cognitive Correlates of Speech Perception

A cognitive factor which has been implicated in language learning and the development of speech recognition skills is phonological short-term memory i.e. memory for speech/verbal information. Baddeley, Lewis, and Vallar (1984) proposed a model of phonological short-term memory comprising two parts: a 'buffer' or memory store that can hold memory traces for a few seconds, and a subvocal rehearsal process (known as the 'phonological loop') that refreshes these memory traces. Two tests that have been used in numerous studies to assess phonological working memory are nonword repetition and digit span.

Gathercole and colleagues have carried out extensive investigation into the role of working memory in language development, and have developed the Children's Test of Nonword repetition (CNRep) as a measure of phonological working memory (Gathercole, Willis, Baddeley, & Emslie, 1994). In this task, the child hears a set of nonwords (i.e. non-lexical but phonotactically legal speech tokens with no semantic referents) of 2-5 syllables in duration, and is asked to repeat each nonword aloud. Children's performance on the CNRep has been positively correlated with vocabulary development/word learning, reading and speech comprehension in normally-developing young children. It has furthermore been shown that nonword repetition is *predictive* of the development of language skills in children. For example, CNRep scores at age 4 were found to be predictive of vocabulary size at age 5 (Gathercole & Baddeley, 1989). Recently, Gupta (2003) has provided evidence that the inter-relationship of word learning, nonword repetition and immediate serial recall that has been observed in children (Baddeley, Gatherole, & Papagno, 1998) persists into adulthood in the normal-hearing population. Importantly, Gathercole and colleagues (Gathercole et al., 1994) presented evidence suggesting that the CNRep gives a better account of language development in children than that achieved by Digit Span, the more conventional measure of phonological memory. Nevertheless, Digit Span is still a popular method of measuring phonological working memory.

Tests of memory span have also been investigated with regard to the outcomes of cochlear implantation. Pisoni and Geers (2000) found a significant positive correlation between digit span and speech perception in paediatric cochlear implantees, after demographic factors such as duration of defaness and duration of implant use had been partialled out. Pisoni and Cleary (2003) explored this relationship further. They measured forward and backward digit spans, speech rate and word recognition performance in 176 children fitted with cochlear implants. By analysing the inter-relationships of all the test scores, Pisoni and Cleary found that the correlation between digit span and word recognition approached zero when speech rate was partialled out. As speech rate was taken to reflect subvocal rehearsal speed by these authors, they concluded that the relationship between digit span and speech recognition was driven by the verbal rehearsal speed component of phonological working memory rather than the phonological store capacity. The overall conclusion from the analysis was that speech rate and digit span both have some memory capacity component in common, which relates them both to speech perception, but that speech rate has an extra predictive factor that relates it to speech perception when digit span effects are partialled out. Pisoni and Cleary (2003) claim that this factor is associated with maintenance and retrieval of phonological and lexical information from working memory i.e. verbal rehearsal. In a study measuring auditory digit span in normal-hearing adults listening to a cochlear implant simulation, Burkholder et al. (2005) ascribed the shortened digit spans to the poor encoding of digit identities.

Nonword repetition performance has also been thoroughly investigated in paediatric cochlear implant groups. Dillon, Pisoni, et al. (2004) found that nonword repetition performance was related to measures of speech perception, speech production and verbal rehearsal speed in phonological working memory in children with cochlear implants. However, they did not find a relationship with digit span measures. In a review of phonological working memory and speech processing, Jacquemot and Scott (2006) echo the original motivations of Gathercole et al. (1994) in pointing out that the use of digit span as a measure of phonological working memory may be sub-optimal, as the task involves memorizing lists of highly familiar items that have long-term semantic representations. They cite evidence from patient studies to suggest that semantic and phonological information may have separable stores in short-term memory (R. Martin, Shelton, & Yaffee, 1994; Romani & Martin, 1999). However, one would predict that verbal digit span and nonword repetition would still share some processes. With reference to the findings of Dillon, Pisoni, et al. (2004), it may be that nonword repetition was so challenging to the children for reasons other than loading on working memory that the relationship with digit span was overshadowed.

Few studies have addressed the role of cognitive factors in the outcomes of adult cochlear implantation, and the results from existing studies are unclear. While Knutson and colleagues

(Gantz et al., 1993; Knutson et al., 1991) showed that performance on a visual monitoring task is associated with post-implantation audiological success with a cochlear implant, the nature of the actual task used suggests that it placed processing demands on verbal encoding and rehearsal in phonological working memory. Thus, this task may not have been that different from more standard tests of phonological short-term memory. Lyxell et al. (1998) found evidence of a relation between preserved phonological representations and success with post-implantation speech comprehension in adult implant-wearers. However, several studies, such as that by Collison et al. (2004) have been limited by the effects of 'demographic' factors such as age at implantation and duration of deafness, which have precluded analysis of cognitive correlates of performance.

On the basis of the findings in research on language development and cochlear implantation, and on the outcomes of previous experiments in this thesis, it is predicted that phonological working memory should explain some of the variability in speech perception in normal-hearing adults exposed to a cochlear implant simulation. In this relatively small literature, a role has already been found for digit span as a correlate of perception of noise-vocoded speech, but the relationship has been weak (Chiu et al., 2002; Eisenberg et al., 2000). In the current study, the relationship will be re-tested. Alongside a test of noise-vocoded sentence recognition, measurements are made of participants' scores on verbal digit span and nonword repetition with clear (undistorted) speech. Digit span is included for historical reasons as a standard test of phonological working memory for familiar items, while nonword repetition is included as a measure of memory for unfamiliar phonological information. Where the highly familiar items of the Digit Span should gain automatic and complete access to working memory, nonword items place greater loading on the *encoding* of the acoustic input into phonological representations - what Gathercole et al. (1994) call 'Phonological Analysis'. In contrast, where the Digit Span tests a listener's capacity to hold online representations of several seconds of spoken information, the repetition of single nonwords (especially those with only 2 or 3 syllables) is likely to place less emphasis on this aspect of phonological memory. Recognition of degraded sentences places heavy demands on Phonological Analysis, while also requiring the online maintenance of several seconds of auditory-linguistic input. Therefore, it is predicted that both tests of phonological working memory in the current experiment should independently account for some of the variability in noise-vocoded sentence perception, while themselves sharing common processes.

Another cognitive measure will be taken in the current study. Vocabulary size will be measured as an indication of general verbal intelligence and linguistic knowledge. This is included to follow on from the finding of a significant correlation between noise-vocoded sentence recognition and performance on the WAIS-III Vocabulary task observed in Experiment 2a. Furthermore, its inclusion

also potentially addresses the claim from Davis et al. (2005) that perceptual learning of noise-vocoded sentences is driven by lexical information. If a participant has greater verbal intelligence and lexical knowledge, this may contribute to more efficient adaptation to noise-vocoded speech via the mechanism proposed by Davis et al. (2005). It is therefore predicted that participants with greater vocabularies will recognise more noise-vocoded words and exhibit greater improvements in performance over time.

The Current Study

The focus of this chapter is the quantification of individual differences in perception of, and adaptation to, noise-vocoded speech. The purpose of this exercise is twofold: (1) With relevance to the above discussion, it should inform the hitherto problematic issue of variability in group studies of perceptual adaptation and (2) As discussed in Chapter 1, variability in the normal population is of interest in itself as a relatively untapped area of the speech recognition literature. If a distortion such as noise-vocoding can effectively uncover the underlying variability in the normal-hearing population, it could be use as a tool to explore the correlates of this variability and develop a greater understanding of the factors that contribute to successful speech perception in the normal-hearing adult. More directly, as argued in Chapter 7, the growing literature employing noise-vocoding with normal-hearing populations motivates an investigation of the variability seen with tests using this particular method of speech distortion. In Experiment 6, an adaptive tracking procedure is used to extract a baseline recognition threshold for noise-vocoded speech, and a measure of adaptation over time, on a listener-by-listener basis. Experiment 7 adopts a constant measures approach, where the listener hears a fixed number of stimuli across a range of distortion levels, rather than being adaptively guided to a threshold level. For both methods, the individual measures of speech recognition and adaptation are correlated with working memory span, nonword repetition and vocabulary scores. It is predicted that better speech recognition and adaptation will be associated with greater digit spans and higher scores on the vocabulary and nonword repetition tasks.

8.2 Experiment 6

Adaptive tracking (Levitt, 1971) is a technique that is frequently adopted in the psychoacoustic literature as a method of perceptual threshold estimation. Figure 8.1 shows a representation of an adaptive track. The approach is to initially present the participant with a suprathreshold stimulus or discrimination, then to use the participant's response to guide the choice of stimu-

lus/discrimination for the next trial. In the most basic 'one down - one up' paradigm, a correct response results in the selection of a more difficult next trial, while an incorrect response is followed by an easier trial. The overall effect is that of focussing in on the participant's threshold performance level, making this a more time-efficient means of threshold estimation. The step size in difficulty from one trial to the next can be fixed or variable - for example, if the first four steps are large and subsequent steps smaller, the adaptive track can initially make quick progress to the level of interest without losing sensitivity in the later stages. The threshold estimate is calculated as the difficulty level at the point where the track is equally likely to go up or down i.e. the midpoint of the track.

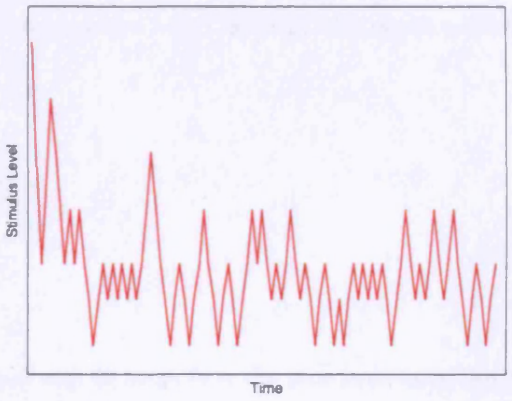


Figure 8.1: A schematic representation of an adaptive tracking procedure.

In this experiment, adaptive tracking is applied to the problem of individual differences in speech recognition and perceptual adaptation to extract indices of these skills for each listener. Noise-vocoded sentences will be available at a range of different distortion levels (1-20 bands), and the adaptive track will follow the number of bands needed for threshold sentence recognition of 50% sentences correct. While a threshold recognition score can be extracted from the overall equilibrium point of the track, any adaptation effect across the stimulus set can be estimated by fitting a straight line to the track and calculating its slope as a function of number of bands/time.

8.2.1 Method

Participants and Apparatus

Twenty-seven native speakers of English (aged 18-40, 11 male), with no known hearing or language difficulties, were tested. Participants were recruited from the UCL Department of Psychology Subject Pool. For the speech perception task, auditory stimuli were presented from a Dell personal computer through Sennheiser HD25-SP headphones. Volume settings were fixed at the same level for all listeners using the QuickMix software, version 1.06 (Product Technology Partners, Cambridge, UK). In tests where participants' responses were recorded, this was done using a PC-compatible microphone in conjunction with Cool Edit 96 (Syntrillium Software Corporation, USA) software.

Design and materials

Adaptive Track

All listeners were presented with 98 items from the Bamford-Kowal-Bench sentence corpus (Bench et al., 1979). All participants heard the same sentence set, in the same order. Each sentence in the set features three keywords (shown here in upper case letters) and is of simple syntatic and semantic structure e.g. *'The CLOWN has a FUNNY FACE'* and *'THEY'RE BUYING some BREAD'*. Pronouns and content words are counted as keywords. Audio recordings of the sentences were made by a female speaker of Southern British English as described in the Methods section of Chapter 4. However, the files were downsampled at a rate of 22050Hz for the current experiment. The recordings were divided into separate .wav files for each sentence and these were normalized for peak amplitude. Each sentence was noise-vocoded according to the general scheme described by Shannon et al. (1995), using MATLAB Version 7.0 software (The MathWorks, Inc., Natick, MA). The sentences were converted to noise-vocoded versions for all band numbers between 1 and 20 inclusive. To illustrate the effect of increasing band numbers on stimulus intelligibility, an example sentence vocoded at several different band numbers is included on the CD accompanying this thesis.

For each sentence, the input speech waveform was passed through a bank of analysis filters spanning the frequency range 100-5000Hz (each 6th-order Butterworth IIR filters, with 3 orders

per upper and lower side and frequency responses crossing 3dB down from the analysis peak), the number of which was determined by the desired number of output bands. the filter bandwidths represented equal durations along the basilar membrane, and were determined in accordance with the Greenwood (1990) equation relating filter position (on the basilar membrane) to best frequency. The amplitude envelope was extracted from each analysis filter by half-wave rectification and 4th-order Butterworth low-pass filtering at 400Hz. The extracted envelopes were multiplied by a white noise, then each was filtered through a 6th-order Butterworth IIR output filter identical to the analysis filter. The root-mean-square sound (rms) pressure level from each of the output filters was set to be equal to the rms level of the original analysis filter outputs. Finally, the amplitude-modulated output bands were summed together and low-pass filtered at 5kHz.

The scripts for the adaptive paradigm, including a graphic user interface, were written in MATLAB Version 7.0 (The MathWorks, Inc., Natick, MA).

Cognitive Tasks

Each participant was assessed on three tests of cognitive ability; forward digit span, vocabulary size, and nonword repetition. The forward digit span materials were taken from the Digit Span task within the Verbal subset of the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III(UK); Wechsler (1997). The vocabulary task was the British Picture Vocabulary Scale, Second Edition (BPVS-II; Dunn, Whetton, and Burley (1997). Each participant also performed the Nonword Memory Test (Gathercole & Baddeley, 1996). Items in the Nonword Repetition Test were recorded onto digital audio tape by a female speaker of Southern British English in an anechoic chamber, then transferred to PC and downsampled as described above.[1]

Procedure

All participants completed the audiometry screening first, followed by the adaptive paradigm. They then performed the cognitive tasks in the order: Digit Span, Nonword Memory Test, BPVS-II. The order of tasks was fixed in order to maximise comparability of individual performances.

[1]Note that none of the items in the supplementary cognitive tasks underwent the noise-vocoding transformation.

Audiometry

The aim of this test was to find the detection threshold for a pure tone at 500Hz, 1kHz, 2kHz and 4kHz. Stimuli lasting 1-3 seconds in duration were played through earphones from a Kamplex diagnoistic audiometer (model TA155). Presentation began at 40dB HL. The participant pressed a response button to indicate when he/she could hear the tone; the button press was indicated by illumination of an LED on the main panel of the audiometer. For every correct response (consistent with the onset and offset of the tone played) the presentation level was reduced by 10dB. For every inaccurate response, the level was increased by 5dB. The timing and duration of presentations was varied to minimize predictability. At each of the four presentation frequencies, threshold was taken as the lowest level at which responses occurred in at least half of a series of ascending trials, with a minimum of two responses required at that level.

Adaptive Track

Participants were tested individually in a sound-attenuated booth. The experimenter instructed the participant that he/she would hear a set of distorted sentences, and that some items would be more difficult to understand than others. The participant was instructed to listen carefully to each sentence and give immediate spoken report of whatever they had perceived, even if this was only one or two words. All recorded audio stimuli were presented from a Dell personal computer through Sennheiser HD25-SP headphones. Volume settings were fixed at the same comfortable level for all listeners using the QuickMix software, version 1.06 (Product Technology Partners, Cambridge, UK).

The first stimulus presented was noise-vocoded to 20-bands, and served as an example item. The experimenter viewed a graphic interface on the PC screen which featured three buttons, each one bearing a keyword from the sentence just played. For every keyword correctly reported, the experimenter clicked the corresponding button onscreen and this was registered as one point in the total score for the sentence.

The adaptive track used in this experiment had an overall 'one up - one down' structure, thus it measured a speech recognition threshold of 50% items correct. If the item was reported correctly (all 3 keywords correct), the next stimulus had fewer bands; if the item was reported incorrectly (2, 1 or 0 keywords correct), the next item had more bands than the previous. The exact distortion level for the next item was determined by a ratio division or multiplication - the nearest whole

number to the result of this calculation was used to select a stimulus with the corresponding number of bands. For the first run in the track (i.e. downward from 20 bands), the ratio was 2.0. For the second run (which began when the participant made his/her first mistake and the track changed direction), the ratio was reduced to 1.67 - this facilitated an increase in sensitivity as the track finds the general range containing the threshold level. From the next run, the ratio became 1.4, and remained so until the end of the experiment. There was no limit placed on the amount of times that the track could turn - it was intended that each participant heard all 98 sentences, in order to maximise inter-individual comparability of tracks. All 98 sentences (i.e. their linguistic content) were presented in the same order to each participant.

Cognitive Tasks

Forward Digit Span

In this test, the participant's task was to give immediate verbal report of a list of single digits between 1 and 9, which were read aloud by the experimenter. The test had 8 levels, one for each list length from 2-9 digits. Each level featured two items, of which the participant must report at least one correctly to proceed to the next level. A correct response must feature the correct digits in the same order as presented by the experimenter. The first level contained 2-digit lists and the list length increased by one digit with for each subsequent level. Administration of the task was terminated if the participant scored 0 on both items within the same level, or if all 16 items had been administered.

Nonword Memory Test

The Nonword Memory Test (Gathercole & Baddeley, 1996) comprises 28 nonword items, each item being an utterance of 2-5 syllables in duration that has no semantic referent but that is phonotactically legal in English e.g. *doduloppity*, *strunfabe*. On each trial, the participant listened to the item over headphones and immediately repeated it aloud. The participant's responses were marked online by the experimenter, and judged to be correct only for exact repetitions of the original items. However, consistent mispronunciations of certain phonemes, for example through lisping, were not treated as errors. No partial marks were awarded. Participants' responses were recorded using a desktop microphone and saved for later score checking.

British Picture Vocabulary Scale

In this test, the participant's task is to match the meaning of a word read by the experimenter to one of four pictures presented in a 2x2 array. The entire test comprises 168 English words (nouns, adjectives and verbs), which are divided into 14 Sets of increasing difficulty. As the test was designed to be administered to children as young as 3 years old, many of the initial items are relatively simple for the adult participant. For this reason, participants in the current study began the test with Set 9 (the Start point recommended for Ages 16-21). If one or fewer errors were made on Set 9, the participant progressed to Set 10. The test was terminated at the end of the first Set in which the participant made 8 or more errors - if this termination criterion was not met, administration continued until the last item of Set 14. In the exceptional case that the participant made more than 1 error in Set 9, Set 8 was administered next. If an error score of more than 1 was obtained on Set 8, the experimenter would move back to Set 7. This would continue until a Set with 1 or 0 errors was found - after running this Set, the test would resume at Set 10 and progress as normal.

8.2.2 Results

Audiometry

All participants scored within the normal range (±20dB HL) for 500Hz, 1kHz, 2kHz and 4kHz.

Scoring Sentence Report

All sentences were marked out of 3 for the number of keywords correctly reported. The marking scheme adopted was liberal and as described in the Results section of Chapter 4, where errors of number or tense on nouns and verbs were accepted as correct.

Scoring Cognitive Tasks

Forward Digit Span

The participant's score on this task was recorded as the Maximum Digit Span achieved i.e. the longest span at which the listener achieved at least one correct response.

Nonword Memory Test

The online scores for the task were double-checked by the experimenter and a second judge. Using agreed criteria, the two judges listened to the audio recordings made during testing and made adjustments to the original scores where these were judged to have been in error. Each participant's score on the task was recorded as the total number of correct responses out of the 28 items.

British Picture Vocabulary Scale

The participants' score was calculated as the Item Number of the last item administered, minus the total errors made.

Table 8.1 shows descriptive statistics for scores on these three cognitive tasks.

Table 8.1: Descriptive statistics for cognitive task scores in Experiment 6.

	M	SD	Min	Max	IQR
Digit Span	7.0	1.22	5	9	2.00
Nonword Memory Test	22.9	2.36	17	27	4.00
BPVS-II	153.6	6.64	138	166	10.00

Analysing the Adaptive Track

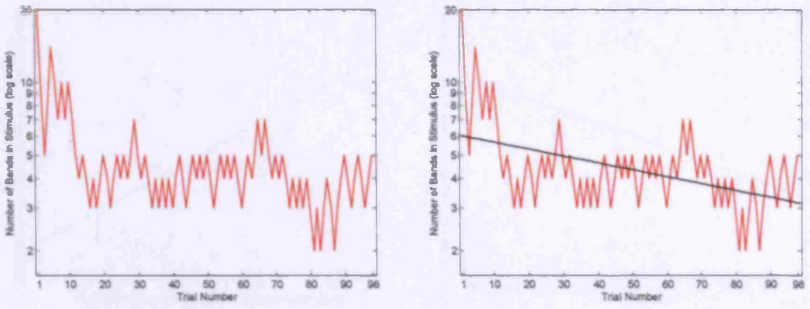
Figure 8.2(a) shows the adaptive track obtained across the 98 sentences for one of 27 participants tested; the figure shows the presentation level (i.e. log_{10}(number of bands) - the *log* is taken as this offers the best description of the relationship between the number of bands and the amount

of spectral resolution, and the structure of the adaptive track - for each item, in chronological order. For each participant, a number of values were extracted from this track as measures of the initial and final performance levels, overall performance level, and the rate of adaptation. For all calculations, the first 2 runs (i.e. the first downward run from 20 bands, and the upward run after the first reversal) were ignored, and the beginning of the third run was taken as the starting point of the track. This is because the effect of all participants having to begin with the same highly suprathreshold stimulus would act to reduce any individual differences in the initial phase of the track. The measure of Initial Speech Recognition Performance was the mean log_{10}(number of bands) of the first 20 stimuli (from the start of the third run), while Final Speech Recognition Performance was measured as the mean log_{10}(number of bands) of the final 20 stimuli. For these scores, better speech recognition performance corresponded to lower numerical values. Overall Speech Recognition Performance was calculated as the average log 10(number of bands) for all reversals in the track i.e. trials on which the track changed direction. The Rate of Adaptation was measured by fitting a straight line to all items from the start of the third run through to the last stimulus (as shown in Figure 8.2(b)) and calculating its slope as the change in the log_{10} number of bands over time. For this measure a more negative value corresponds to a greater rate of adaptation. Table 8.2 shows the descriptive statistics associated with these four performance scores, in terms of log_{10} values. In terms of numbers of bands, the range of scores obtained for Overall Speech Recognition Performance was 3.3 bands to 5.4 bands (mean 4.5 bands), the range for Initial Speech Recognition Performance was 4.0 bands to 6.9 bands (mean 5.5 bands), and the range for Final Speech Recognition Performance was 2.6 bands to 5.4 bands (mean 3.9 bands).

Table 8.2: Descriptive statistics for adaptive track performance measures. Original log 10 values are shown.

	M	SD	Min	Max	IQR
Overall Performance	.656	.045	.526	.735	.072
Initial Performance	.743	.068	.615	.841	.110
Final Performance	.587	.073	.410	.735	.105
Slope of Track	-.004	.001	-.004	-.002	.001

Bivariate Pearson's correlations were run on these measures. There were no specific *a priori* hypotheses about the interrelationships of the measures, so 2-tailed tests were run. It was found that there was a significant positive correlation between Initial Speech Recognition Performance and Final Speech Recognition Performance (Pearson's $r = .444$, $p = .020$, 2-tailed), suggesting that participants with a lower threshold at the start of the track had the lowest thresholds at the end



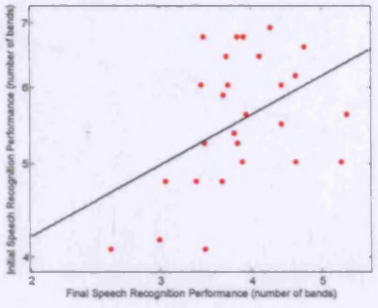
(a) Raw adaptive track

(b) With fitted line to measure rate of adaptation

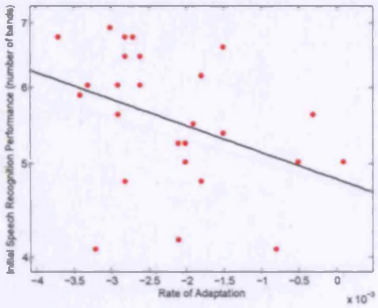
Figure 8.2: Plots showing example adaptive track for one participant in Experiment 6.

of the track. A significant negative correlation was found between the Initial Speech Recognition Performance and Rate of Adaptation (Pearson's $r = -.390$, $p = .044$, 2-tailed) suggesting that participants with an initially higher threshold had a greater rate of adaptation. A significant positive correlation was found between Final Speech Recognition Performance and Rate of Adaptation (Pearson's $r = .575$, $p = 0.002$, 2-tailed), suggesting that participants with the lowest threshold at the end of the track had adapted at the fastest rate. These three correlations are not mutually viable, as there are contradictions between them - this issue will be addressed in the discussion. However, a more informative measure of the relationship between baseline speech recognition and rate of adaptation may be the correlation between Overall Speech Recognition Performance and Rate of Adaptation, which was non-significant (Pearson's $r = .082$, $p = .685$, 2-tailed). Figure 8.3 shows scatterplots of the significant correlations.

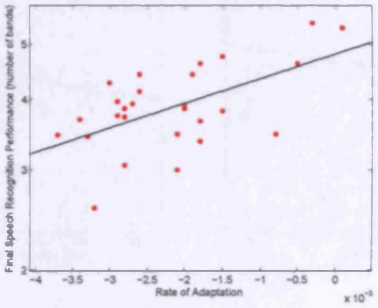
Bivariate Pearson's correlations were run between the four track measures and the participants' scores on the cognitive tasks. As there were hypotheses about the direction of these particular correlations, 1-tailed significance values were taken. All correlations between Initial Speech Recognition Performance and the cognitive measures were non-significant. A significant negative correlation was found between Final Speech Recognition Performance and scores on the BPVS-II (Pearson's $r = -.451$, $p = .009$, 1-tailed), Rate of Adaptation and scores on the BPVS-II (Pearson's $r = -.324$, $p = .049$, 1-tailed), and Overall Speech Recognition Performance and scores on the BPVS-II (Pearson's $r = -.323$, $p = .050$, 1-tailed). There was also a marginally-significant correlation between Overall Speech Recognition Performance and scores on the Nonword Memory Test (Pearson's $r = -.293$, $p = .069$). Figure 8.4 shows scatterplots of these four correlations.



(a) Initial and Final Speech Recognition Scores



(b) Initial Speech Recognition and Rate of Adaptation



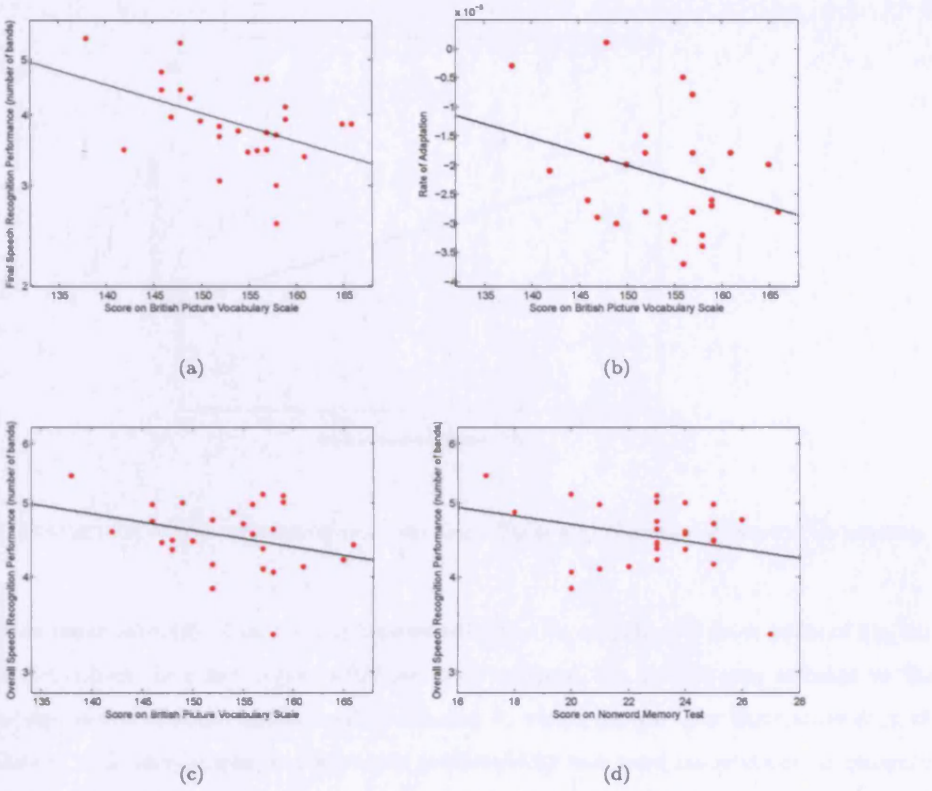
(c) Final Speech Recognition and Rate of Adaptation

Figure 8.3: Scatterplots showing significant correlations between the measures of speech recognition performance in Experiment 6.

All other correlations between the cognitive measures and the speech recognition data were non-significant. Within the cognitive tasks, there was a marginally-significant correlation between Maximum Digit Span and scores on the Nonword Memory Test ($r = .294, p = .068$). This is unsurprising, as both tests are presumed measures of phonological working memory. Figure 8.5 shows a scatterplot of the correlation between these two variables.

8.2.3 Discussion

It was hypothesised that performance on the speech recognition part of Experiment 6 would be significantly correlated with scores on the cognitive tasks. Based on the previous literature, it was expected that better speech recognition performances would correspond to a greater digit span,



(a) (b)

(c) (d)

Figure 8.4: Correlations between scores on the cognitive tasks and measures of speech recognition performance.

larger vocabulary and fewer errors on nonword repetition. For Initial Speech Recognition scores, there was no significant correlation with any of the cognitive test scores. However, significant correlations between Vocabulary and Final Speech Recognition Performance, and between Vocabulary and Rate of Adaptation, showed that listeners with larger vocabularies exhibited greater rates of adaptation and better Final Speech Recognition Scores.

How do we interpret the correlations between performance on the speech task and vocabulary size? The presence of a relationship between vocabulary size and overall speech recognition, adaptation rate and final speech recognition performance, but not with initial/baseline speech recognition, suggests that it is the processes of adaptation that are potentially most dependent on vocabulary size. From the nature of the speech recognition task, the most straightforward explanation for the correlation observed involves viewing vocabulary size as an index of Verbal IQ (vocabulary tests are routinely used as measures of verbal IQ; the WAIS-III Vocabulary subtest is included as a measure of Verbal Comprehension). When listening to heavily distorted sentences,

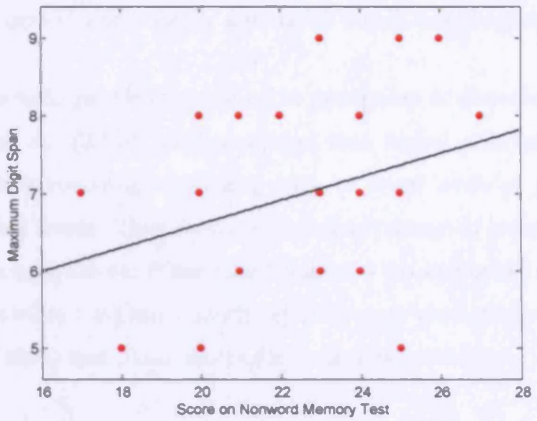


Figure 8.5: Scatterplot of the correlation between Digit Span and Nonword Memory Test scores.

particularly at lower numbers of bands, the listener will often be confident of some parts of his/her percept but not others. In order to give a full sentence response, the listener may attempt to 'fill in the gaps' by means of some higher-level processing in which he/she uses their knowledge of linguistic factors such as syntactic and semantic predictability and word associations to generate hypotheses about the words in the sentence. The products of this linguistic processing may feed back in a 'top-down' fashion to lower speech processing levels, i.e. at the level of word or individual speech sound recognition, and facilitate the recognition of words that the listener did not recognise immediately on hearing the sentence. Grant and Seitz (2000) showed that this 'top-down' processing becomes more prominent as acoustic degradation of the speech signal is increased. Thus, with noise-vocoded speech, there may also exist a dynamic process between higher to lower processing levels, where the listener tests their higher-level hypotheses against the auditory percept in working memory in an attempt to find a match. Efficient use of this processing loop may facilitate quicker and stronger adaptation to the noise-vocoded stimulus. Thus, listeners with more sophisticated linguistic knowledge (as exemplified in this experiment by a larger vocabulary) are better equipped to adapt quickly to distorted speech.

The predicted significant correlations between both measures of phonological working memory and speech recognition scores were not found. The only indication of a role for this variable was a marginally-significant correlation between Nonword Memory Test scores and Overall Speech Recognition Performance. Chiu et al. (2002) and Eisenberg et al. (2000) found that digit span was only weakly correlated with perception of noise-vocoded sentences by normal-hearing adults, and the current findings support this view. Alternatively, these null results could be the consequence

of the adaptive track's emphasis on whole sentence perception (in accepting only responses with 3 identified keywords as correct) and possible associated higher-level linguistic processing.

The role of higher-level linguistic information in perception of distorted speech was previously investigated by Davis et al. (2005), who concluded that lexical information in training stimuli drives adaptation to noise-vocoding by feeding back to lower levels of perception and effecting adaptive changes at these levels. They found no extra advantage of semantic context, as Normal Prose was no better than Syntactic Prose (see Chapter 3 for examples) as a training stimulus - however, there is a possibility that any underlying differences were masked by an apparent ceiling effect in Davis et al.'s (2005) test phase scores for these two conditions.

In this study, we have arrived at the issue of top-down processing from a different angle. Using the same set of sentences across all participants, this study has identified a potential relationship between linguistic knowledge and individual differences in adaptation to noise-vocoded speech. The findings in Davis et al. (2005) would suggest that this correlation reflects differences at the lexical level of processing, but without further studies we currently cannot tease apart the effects of top-down processing from the lexical level from higher-level processing on the level of sentence structure and semantics. A potential means of addressing whether higher-level semantics play a role in individual differences is to test listeners on perception of both Normal Prose and Syntactic Prose noise-vocoded sentences and look for differences in rate of adaptation and overall recognition performance between conditions, on a listener-by-listener basis. The problem with using Jabberwocky or Nonword sentences as test stimuli is that these contain highly unfamiliar items, which would put a load on working memory that may conceal the effects of the lexical manipulations.

No specific *a priori* hypotheses were put forward for the three different measures extracted from the adaptive track in this experiment (Initial Speech Recognition Performance, Final Speech Recognition Performance, Rate of Adaptation). However, we recognised that the relationship between baseline perceptual ability and rate of adaptation was potentially complex. This potential has been borne out in the correlations found between the three measures. The significant correlation between Initial Speech Recognition and Rate of Adaptation suggests that the listeners with better baseline speech recognition may have reached ceiling performance by the end of the track, thus allowing for a limited adaptation effect. The significant positive correlation between Initial and Final Speech Recognition scores suggests that the listeners who were strongest and weakest initially remained so at the end of the track, so the greater adaptation by poorer initial listeners was not enough to overcome baseline recognition weaknesses. However, the significant correlation between

Final Speech Recognition scores and Rate of Adaptation is at odds with this interpretation, as this suggests that the listeners with better final performance had exhibited greater adaptation. This doesn't fit; if the best listeners at the start of the experiment exhibited lower adaptation rates, and these are the same listeners who perform better at the end of the experiment, they cannot be the same group that creates the correlation between Final Recognition Score and Rate of Adaptation.

This pattern of correlations can be better understood by viewing Figure 8.3. It is clear from this Figure that there is no uniform relationship between baseline recognition of noise-vocoded speech and the rate of adaptation to the stimulus. Whilst there may be a tendency within the whole group for worse initial performers to exhibit greater adaptation, there are individuals who strongly contradict this pattern. For example, the listener with the best initial (and final) performance also exhibited the fourth largest rate of adaptation, while the listener with the worst Final Speech Recognition score exhibited the second lowest amount of adaptation - these listener's scores are strongly affecting the correlation between Final Speech Recognition Performance and Rate of Adaptation. From the current data set, it is therefore not possible to draw firm conclusions about the relationship between 'raw' speech recognition capabilities and the capacity for adaptation to a distorted speech stimulus. However, a very similar pattern of results has been shown in a previous study of auditory perceptual learning using measurement of pure-tone frequency discrimination thresholds in normal-hearing adult listeners (Amitay et al., 2005). For a group of participants who were trained on frequency discrimination about a 1kHz standard only, the listeners with higher initial thresholds ('poor' listeners) exhibited much more dramatic and rapid improvements in performance in the first 1500 trials than those who started with low thresholds ('good' listeners), but in post-test measurements the 'good' listeners still had lower thresholds (i.e. better performance) overall. Stacey and Summerfield (2007) also observe a pattern in listeners exposed to shifted noise-vocoded speech, where listeners with 'poor' baseline scores exhibiting greater learning. Further investigation of the interaction of starting performance and learning in noise-vocoded speech will be necessary to see if this relationship can be replicated in the current thesis.

The question of the relationship between baseline performance and adaptation is an important one in the field of distorted speech perception. It seems from this experiment that there is a complex relationship between the ability to perceive a difficult speech stimulus straight away and the capacity to adapt to the stimulus over time. A potential problem with the current design is that listeners can adapt to noise-vocoded speech quite quickly (Davis et al., 2005), and it is possible that several listeners may have achieved considerable adaptation even within the first 20 sentences of exposure. A way to address these issues in the current testing paradigm may be to present listeners with a more challenging form of speech distortion that requires much more

exposure to achieve the same level of adaptation as observed after short-term exposure to noise-vocoding. Spectrally-shifted noise-vocoded speech would be a good candidate stimulus - this is constructed similarly to the noise-vocoded speech described in the Method section of this chapter, but the noise-band carriers are shifted upwards in frequency relative to the original bands in the speech stimulus. This creates a frequency-to-place mismatch similar to that which results from incomplete insertion of the cochlear implant into the cochlea of the patient. This added factor in the distortion greatly increases its perceptual difficulty and the time-scale over which a normal-hearing listener can adapt, relative to unshifted noise-vocoding (Rosen et al., 1999). Using this stimulus in the context of adaptive tracking would probably require much longer sessions than used in the current experiment. However, the very slow adaptation would lend itself better to receiving relatively clean measures of initial speech recognition performance that better reflect the baseline speech recognition capabilities of the individual listeners in the presence of very slow adaptation, rather than the possible conflation of baseline ability and adaptation observed in this experiment. The current experiment did not make any attempt to distinguish the proportion of improvement in performance that is due to non-auditory 'task practice' from that which is due to true perceptual learning. Previous studies have shown that performance of auditory tasks can improve with visual-only training (Amitay et al., 2006; Stacey & Summerfield, 2007). Inclusion of a suitable visual baseline task would be desirable in future studies of learning.

Despite its ease of administration and the advantages of obtaining a relatively speedy measure of speech recognition threshold, the adaptive track used in the current experiment has a number of weakness that should be noted. In its more conventional use in the psychoacoustic literature, the adaptive track tests the participant's detection/discrimination on one dimension e.g. loudness, amplitude modulation with relatively simple stimuli. In this experiment, each item contained many parts i.e words of different classes, frequencies and morphologies within sentences of varying meaning and predictability. The track was set up to measure a threshold of 50% items correct - in this case, an item was an entire sentence. This shifts the emphasis of the track toward measurement of 'top-down listening' i.e. the 'gap-filling' sentence completion approach mentioned above, rather than toward a lower level of speech sound or word perception. This may go some way in explaining the significant correlation between the adaptive track measures with vocabulary scores, and perhaps the absence of significant correlations with digit span and nonword repetition.

Whilst measuring the proportion of complete sentences correct is still an acceptable means of measuring performance, it may communicate an incomplete picture. In the adaptive procedure adopted in this experiment, item scores of 2, 1 and 0 keywords correct were all treated as incorrect responses, and seen equally in terms of the movement of the track. This means that it would

have been possible for two listeners to end up with very similar adaptive tracks but have a large difference in the total amount of keywords correctly identified e.g. Listener A scores 2 in most of his/her incorrect sentences, while Listener B scores 1. It would be desirable to be able to extract a more sensitive measure of noise-vocoded speech recognition, in terms of the proportion of Keywords correct rather than whole sentences.

A way in which to obtain a cleaner measure of the speech recognition data in terms of Proportion Keywords Correct is look beyond the adaptive track to the raw scores on each item. A plot of the number of keywords correct against log_{10}(number of bands) characterizes performance across the full range, from floor to ceiling recognition. By fitting a logistic function to this data, one can then interpolate to find the number of bands corresponding to any performance level (See Figure 8.6). In order to estimate adaptation effects, the data can be split into blocks according to presentation order and the shift of the performance curve estimated in terms of the number of bands.

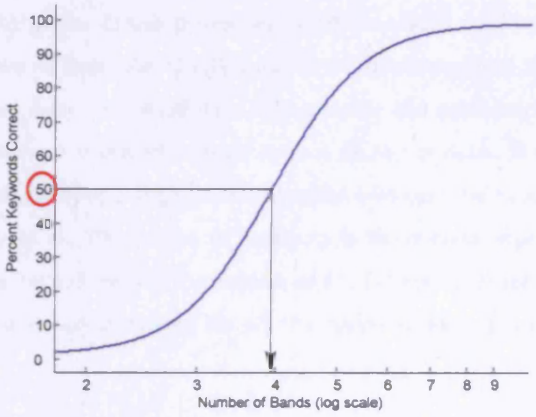


Figure 8.6: Diagrammatic representation of how logistic fits could be used to measure performance threshold in terms of the number of bands needed for a given speech recognition score e.g. for 50% Keywords Correct.

An analysis of this kind was carried out on the raw Keyword recognition scores from Experiment 6. Scores were organised in terms of the numbers of Keywords recognised and presented at each distortion level (number of bands) - distortion levels were entered into the analysis in terms of their log_{10} equivalents. Curve-fitting on each individual data sets (for all 98 sentences) was carried out using the psignifit software package (Wichmann and Hill (2001a, 2001b); available from `http://www.bootstrap-software.org/psignifit`). The equation used for fitting was:

$$(f(x:\alpha,\beta,\gamma,\lambda)) = \gamma + \frac{1-\gamma-\lambda}{1+e^{-(x/\alpha)^{\beta}}}$$

In the output of the fitting procedure, the α parameter corresponds to the curve's displacement along the abscissa (in this case, the log_{10}(number of bands) for 50% of maximum performance), and β is inversely proportional to the curve slope. The parameter γ corresponds to the base rate of performance (or the 'guessing rate'), while λ reflects the 'lapse rate' i.e. a lowering of the upper asymptote to allow for trials where the listener gives an incorrect response whose accuracy is not related to the stimulus level. The software takes a constrained maximum-likelihood approach to fitting, where all four variables are free to vary, but where, in this case, γ and λ are constrained between 0.00 and 0.05.

The goodness-of-fit of each fitted function given by psignifit is determined via the use of the deviance statistic for the curve, d, and Monte Carlo simulations. Using the parameter values obtained (for α, β etc.) in the fitting procedure, 1999 simulated data sets are generated and 90% confidence limits extracted from the distribution of the corresponding 1999 deviance statistics. If the value of d obtained from the fitted data falls outside the confidence limits generated by the simulated data sets, the curve is said to show a poor fit to the data. It was decided, on the basis of previous findings supporting a logistic relationship between the number of bands and speech intelligibility (Shannon et al., 2004), that all data sets in the current experiment would be included in subsequent analyses, regardless of the goodness-of-fit. However, for reference, Appendix C shows the results of the goodness-of-fit testing for all the curves used in the analyses described in this Chapter.[2]

For the purposes of the analysis, two measures were extracted from the curve-fitting process for each individual; α and β. Note that the α parameter describes, in log_{10} values, the point along the abscissa which equates to the performance level halfway between the lower and upper asymptotes (determined by the values of γ and λ). Hence, whilst it will be referred to as the '50% threshold' throughout this chapter (and the following chapter), this does not correspond directly

[2]The Experiment 6 Keywords curves show several poor fits across the participants. This is likely caused by violation of the assumption of independent trials made when entering individual scores for Keywords that came from the same original sentence. This violation would have accentuated the damaging effects of certain data points on fitting. If we consider the case of the adaptive track, there may be only one or two sentences presented at certain higher distortion levels. Should the listener entirely miss the content of one of these sentences, this forms one missed trial for the Sentences curve, but three trials for the Keywords curve. However, the improved Keyword curve fits for the constant measures approach in Experiment 7 justified continued use of Keyword data. This was important for the analysis of Experiment 8, where Word/Token recognition was used as the common measure across the different tasks.

to a numerical performance of 50% correct. A lower α indicates that a listener requires less spectral detail to achieve 50% of maximum performance, and hence is exhibiting more successful performance with noise-vocoded speech. As the β parameter is inversely related to the slope of the logistic curve, a smaller beta value corresponds to a steeper curve slope. For convenience of reading, these parameters will be referred to as *threshold* and *slope* (as opposed to α and β) throughout the chapter. Table 8.3 shows the descriptive statistics for the two extracted parameters for the 27 listeners who participated in Experiment 6. The maximum and minimum thresholds exhibited are lower than the corresponding Overall Speech Recognition scores obtained from the adaptive track. It was expected that this would be the case, as the track scores corresponded to performance on the more difficult task of recognizing whole sentences, whereas the curves were fitted to recognition of approximately 50% of Keywords within sentences.

Table 8.3: Descriptive statistics for logistic functions describing Keyword recognition. Original log values are shown.

	M	*SD*	*Min*	*Max*	IQR
50% Threshold α	.553	.051	.436	.658	.074
Slope Parameter β	.122	.026	.070	.169	.032

Individual threshold and slope values were entered into a two-tailed Pearson's correlation analysis to assess the relationship between curve position and slope in the sentence recognition task. There was a significant correlation between the threshold and slope scores (Pearson's $r = -.447$, $p = .019$, 2-tailed). Given the inverse relationship between threshold and slope, this suggests that listeners with lower thresholds had shallower performance functions than those with higher thresholds. This is not readily interpretable, as one might predict that a sharper slope would be exhibited by stronger listeners i.e. those who are making better use of an impoverished signal. A sharp slope could reflect something more than bottom-up resolving power, for example the incorporation of top-down processing mechanisms that enable a greater increase in recognition performance for each additional unit of spectral resolution when compared with a shallower curve. However, an alternative explanation could be that, in this experiment, those listeners with higher thresholds are unable to make good use of bottom-up information at low band numbers and so cannot progress from floor recognition performance until much greater spectral clarity is offered. When this threshold is reached, performance increases rapidly with further increases in spectral detail as higher-order processes can then 'kick in'. In contrast, the stronger listeners' ability to 'decode' the acoustic stimulus at lower levels means that the increase in their performance is more strongly influenced by the increase in spectral detail across the whole stimulus range tested. The

relationship between threshold and slope will be re-visited later in this chapter, and in further chapter of the thesis.

As with the measures of performance from the adaptive track, the threshold and slope scores were entered into one-tailed, bivariate Pearson's correlations with the scores from the cognitive measures (Total Score on BPVS-II, Total Score on Nonword Memory Test, Maximum Forward Digit Span). This analysis identified significant correlations between 50% threshold and scores on the Vocabulary task (Pearson's r=-.326, p=.048, 1-tailed), and between thresholds and Nonword Memory Test scores (Pearson's $r = -.348$, $p = .036$, 1-tailed), indicating that listeners with lower thresholds have higher scores on both cognitive tasks. The analysis also produced a significant correlation between Forward Digit Span and slope scores (Pearson's $r = -.336$, $p = .043$, 1-tailed), indicating that listeners with better performance on Forward Digit Span produce steeper psychometric functions. While this correlation with slope gives an intuitive result, a significant correlation between Nonword Memory Test scores and slope values took the opposite direction (Pearson's $r = .372$, $p = .028$, 1-tailed). All of these significant correlations are shown in Figure 8.7.

These results differ from the adaptive track scores in terms of correlations between speech recognition threshold measures and cognitive scores. However, as has been mentioned previously, the adaptive track follows a level of performance of 50% Sentences Correct, while the logistic functions fitted in the current analysis extract thresholds for 50% Keywords recognition. In order to make the analyses more comparable, curve fitting was repeated for each individual, this time with scores arranged in terms of the numbers of Sentences recognised and presented at each stimulus level. The resulting descriptive statistics are shown below in Table 8.4.

Table 8.4: Descriptive statistics for logistic functions describing Sentence recognition. Original log values are shown.

	M	*SD*	*Min*	*Max*	IQR
50% Threshold	.651	.045	.514	.741	.054
Slope Parameter	.119	.029	.060	.167	.048

The range of thresholds observed was 3.3 bands to 5.5 bands, with a mean of 4.5 bands. This corresponds closely to the values obtained for Overall Speech Recognition from the adaptive track, and offers reassurance that the curve-fitting approach is measuring the same underlying variables as the track. Further support comes from two-tailed Pearson's bivariate correlations that were run between the Overall Speech Recognition scores from the adaptive track, the 50% thresholds for

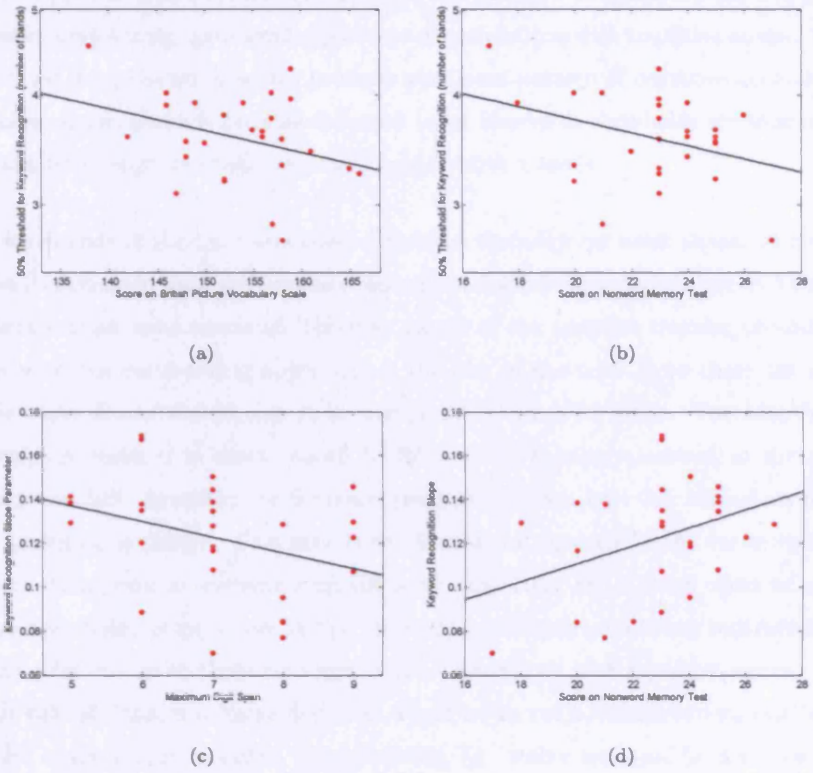


Figure 8.7: Correlations between scores on the cognitive tasks and curve-derived measures of Keyword recognition performance.

the Keywords logistic curves and the 50% thresholds for the Sentences curves - all of these were highly significant ($p<.001$), with Pearson's coefficients between $r = .893$ and $r = .953$.

As before, two-tailed bivariate Pearson's correlations were run between the threshold and slope scores - this time, the result was non-significant (Pearson's $r = .125$, $p = .534$). One-tailed bivariate Pearson's correlations were run with the scores collected from the cognitive tasks (Total Score on BPVS-II, Total Score on Nonword Memory Test, Forward Digit Span). In this analysis, the correlation between threshold and BPVS-II score was exactly as for the curve fit for 50% word recognition (Pearson's $r = -.326$, $p = .048$, 1-tailed), which also closely corresponds to the correlation obtained with Overall Speech Recognition scores from the adaptive track. For the Sentences performance curves, the correlation between threshold and Nonword Memory Test performance remained significant (Pearson's $r = -.346$, $p = .039$, 1-tailed). There were no significant correlations between thresholds and Digit Span performance, nor were there significant correlations between slope scores and any of the scores on the cognitive tasks. Thus, when measuring the same

performance criterion (Sentence Recognition), the two methods of analysis (adaptive tracking and psychometric curve-fitting) gave similar patterns of correlations with cognitive scores. The logistic functions fitted from Keywords scores produce a different pattern of cognitive dependencies from the Sentences curves, perhaps because the much lower Keywords thresholds are tapping a slightly different balance of cognitive behaviours at lower distortion levels.

Despite evidence of good correspondence between the adaptive track threshold measures and the thresholds extracted from the Sentence recognition logistic performance curves for Experiment 6 data, there remain some concerns. The very nature of the adaptive tracking procedure slightly is at odds with the curve-fitting approach, as the aim of the track is to characterize threshold only, while curve fitting should also yield interpretable values for slope. The adaptive track for each participant resulted in most stimuli in Experiment 6 being clustered at distortion levels surrounding the 50% threshold for Sentence recognition, with only few stimuli at other points along the performance range. This may have affected the outcome of the curve-fitting analysis, as a deviant data point at extreme stimulus levels may have had a large effect on curve slope. The analyses involving slope values in this experiment produced somewhat contradictory results, both in the relationship to thresholds and in the correlations with cognitive scores. In order to re-visit the role of slope, a suitable approach would be to run a second experiment in which the presentation of data is more suited to curve-fitting i.e. where an equal (and sufficient) number of stimuli are presented at several distortion levels, and across the whole range of performance (0-100% Keyword recognition). This approach of constant measures may also be more condusive to the investigation of perceptual learning with noise-vocoded stimuli. In the current experiment, all of the stimuli at higher band numbers were encountered early in the track - this conflation of level (number of bands) and time may have affected the curves. In contrast, an appropriate control of both the number and temporal ordering of stimuli in the constant measures approach should facilitate a more accurate investigation of learning through measurement of the changes in curve position and slope over time.

8.3 Experiment 7

This experiment uses an alternative technique to explore and quantify individual differences in recognition of, and adaptation to, noise-vocoded speech. The overall approach is to present the listener with a fixed number of stimuli at each of a number of distortion levels (quantified in terms of number of bands) and characterize a performance curve, or curves, that can be used for the extraction of numerous measures of recognition and adaptation to noise-vocoded sentences.

This was an attempt to overcome some of the weaknesses identified in Experiment 6 with regard to adaptive tracking, and in particular to improve measurement of the slope of the performance function.

The results of Experiment 6 indicated a complex relationship between baseline noise-vocoded speech recognition and amount of adaptation over the course of the adaptive track. This issue is re-visited in Experiment 7. The experiment is divided into two blocks, such that two corresponding curves can be fitted for each listener. This allows the amount of adaptation between the two blocks to be characterized in terms of the changes in performance curve position and shape.

It was decided to re-test the Experiment 6 participants in Experiment 7, in order to maximise comparability of the two experimental approaches. As in Experiment 6, speech recognition scores are examined for correlations with scores on cognitive tests. For the vocabulary and nonword repetition tasks, scores from Experiment 6 were used. The justification for this was that the items of the BPVS-II and Nonword Memory Test would be easier on re-test. For the BPVS-II, it is possible that participants could have looked up the correct solutions to more difficult items after the first test session. Furthermore, vocabulary size is often used as an index of premorbid intelligence in cases of brain injury or dementia and so it was assumed that vocabulary scores are sufficiently stable for a new vocabulary measure to be unnecessary. The Nonword Memory Test relies upon the fact that its constituent items are unfamiliar to the participant, and so to run the test again would not be informative due to the participants' previous exposure to the items. In contrast, the digit span already contains familiar items, and it is not expected that participants would be able to remember the precise order of digits in each sequence from one testing session to the next. Therefore, a new measure of Forward Digit Span was taken during the Experiment 6 testing session. A second reason for inclusion of the Forward Digit Span in Experiment 7 was to facilitate measurement of Backward Digit Span in this session. The two tests are usually run together in the WAIS-III Verbal subset, and the Forward Digit Span provides a good familiarisation for the Backward task (in which the participant is required to perform the additional task of recalling the digits in reverse order). The Backward Digit Span task is more difficult than the Forward equivalent, as the spoken information has to be manipulated online before the spoken answer is produced. Thus, having obtained a significant correlation between speech recognition and Forward Digit Span in Experiment 6, it was decided that the Backward Digit Span would help to unpack which elements of phonological memory are required in the noise-vocoded sentence recognition task - the basic attentional and capacity demands of the Forward span, or higher-order aspects of remembering as tested by the Backward span.

8.3.1 Method

Participants and Apparatus

Twenty native speakers of English (aged 18-40, 8 male), with no known hearing or language difficulties, were tested. These were returning participants from Experiment 6 (the remaining seven participants from Experiment 6 were unavailable for re-test). The mean delay between participation in Experiments 6 and 7 was eight weeks and six days (Range = 55-73 days, SD = 6.5). Apparatus and volume settings used in this experiment were as in Experiment 6.

Design and materials

Speech Recognition Task

Participants each heard 200 sentences from the Banford-Kowal-Bench (BKB) corpus. Sentences were divided into two blocks of 100, called Block A and Block B. The presentation order was counterbalanced across participants such that half the participants heard Block A followed by Block B, and the other half heard Block B followed by Block A. This was done to facilitate comparison of first and second blocks across all participants, as a group measure of adaptation/learning during the experiment. Only two of the sentences in the set had been included in Experiment 6 - the rest were new to the participant.

All sentences were made available in 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-, 12-, and 16-band noise-vocoded versions. The noise-vocoded stimuli were created as in Experiment 6. Each block of sentences was divided into 10 sub-blocks. Each sub-block featured one example from each of the 10 distortion levels. The levels were randomized within each sub-block and the list of items was randomized across the whole block.

Cognitive Tasks

All materials for the cognitive tasks were as in Experiment 6, with the addition of the Backward Digit Span task from the Digit Span test of the WAIS-III(UK) battery.

Procedure

All participants performed the first block of the speech recognition task, followed by the Forward and Backward Digit Span tasks, followed by the second block of the speech recognition task.

Speech Recognition Task

The biggest difference between the participant's task in Experiment 6 and Experiment 7 is that he/she was required to type responses in Experiment 7. This change was put in place for reasons of simplicity in administration of the test. For the adaptive track to work in Experiment 6, participant's responses had to be marked online and this was most easily facilitated by the participant giving spoken responses to the experimenter. The fast, online nature of the task also meant that only keyword responses could be recorded. In Experiment 7, the participants' full, typed responses were saved in a readily accessible text file format, where they could later be used for analysis of both the keyword and non-keyword responses. As the sentence materials used in the two experiments were relatively simple and all the participants were educated to degree level, it was not expected that converting from spoken to written format should result in significant performance changes.

Participants were tested individually in a sound-attenuated booth. The experimenter instructed the participant that he/she would hear a set of distorted sentences, and that some items would be more difficult to understand than others. The participant then viewed onscreen instructions which stated that he/she should listen carefully to each sentence and give immediate typed report of whatever they had perceived, typing their best guess if uncertain. After each sentence played, the participant typed his/her response in a text bar on the computer screen. The participant could then advance to the next item by pressing the Enter key on the keyboard. All recorded audio stimuli were presented from a Dell personal computer through Sennheiser HD25-SP headphones. For each block, the experiment continued at the participant's pace until all 100 stimuli had been delivered and the participant's responses recorded.

Unlike Experiment 6, in which the first item was a 20-band sentence that also acted as an example stimulus, the speech task in Experiment 7 could begin with any of the available distortion levels. However, as all of the current participants had also completed Experiment 6 and so it was assumed they were familiar with the nature of the noise-vocoding distortion and that the presentation level of the first stimulus would have been minimally disruptive.

Cognitive Tasks

Vocabulary and Nonword Memory Test scores were taken from the results of the BPVS-II and Nonword Memory Test administered in Experiment 6. The Forward Digit Span test was re-administered using the same procedure as in Experiment 6. Backward Digit Span was administered in the same way as the Forward Digit Span, with the main difference being that the participant was expected to report the digit lists in reverse order to that in which the experimenter read them. The maximum digit list length in the Backward Digit Span task was 8 (rather than 9 as for the Forward Digit Span).

8.3.2 Results

Scoring Sentence Report

Sentence report was marked in terms of Number of Keywords Correct. The marking scheme for sentence report was the same as that described for Experiment 6. The only extra consideration was to allow homophones to be marked correct (this is an issue for typed responses that did not occur for verbal responses in Experiment 6).

Scoring Cognitive Tasks

Performance scores on the British Picture Vocabulary Scale, Nonword Memory Test and Forward Digit Span were taken from Experiment 6. The Forward and Backward Digit Span performances in Experiment 7 were scored as in Experiment 6, in terms of Maximum Span, with separate scores for the Forward and Backward tasks.

Table 8.5 shows descriptive statistics for the cognitive measures gathered in Experiment 7.

Fitting Curves and extracting Performance Measures

For the purposes of the main analysis, the scores from the speech recognition task were collapsed across the two test blocks. For each participant, sentence report scores at each distortion level were summed to give a Total Keywords Correct score out of 60 (3 Keywords in each sentence,

Table 8.5: Descriptive statistics for cognitive task scores in Experiment 7.

	M	*SD*	*Min*	*Max*	*IQR*
Nonword Memory Test	23.20	2.26	17	27	2.75
BPVS-II	153.10	6.82	138	166	9.75
Forward Digit Span	7.10	1.07	5	9	2.00
Backward Digit Span	5.55	1.19	4	8	1.75

10 sentences in each block). These data were used to fit logistic functions to the data for each participant, according to procedures and constraints described in the Discussion of Experiment 6. For further comparison with Experiment 6 results, the data were also arranged for curve fitting in terms of Total Sentences Correct (out of 20) at each distortion level.

Table 8.6 shows the descriptive statistics for threshold and slope values obtained from the performance curves, for Keyword and Sentence recognition scores. For the purposes of comparison with the curves fitted in Experiment 6, the corresponding statistics for the Experiment 6 data are also shown - note that the Experiment 6 data shown here are only for those 20 participants who returned for re-test in Experiment 7.

Table 8.6: Descriptive Statistics for logistic performance curves in Experiment 6 and Experiment 7.

		Experiment 6					Experiment 7				
		M	*SD*	*Min*	*Max*	IQR	*Mean*	*SD*	*Min*	*Max*	IQR
Keywords	50% Threshold	.551	.055	.436	.658	.081	.448	.033	.364	.491	.049
	Slope Parameter	.123	.027	.070	.169	.033	.107	.017	.077	.150	.021
Sentences	50% Threshold	.649	.049	.514	.741	.049	.534	.039	.420	.588	.047
	Slope Parameter	.120	.030	.060	.166	.053	.103	.021	.064	.144	.030

The threshold and slope values from the two experiments were entered into two-tailed bivariate Pearson's correlation analyses to assess how closely the ordering of scores on these measures agreed across sessions. Thresholds for Sentence and Keyword recognition were significantly correlated across the two experiments (Sentences: Pearson's $r = .665$, $p = .001$; Keywords: Pearson's $r = .548$, $p = .012$). However, there was no significant correlation between the slope measures across Experiments 6 and 7. This gives an early indication that curve slope fitting may indeed have been affected by the distribution of data points used to generate fits in Experiment 6.

It can be seen from the means in Table 8.6 that threshold performance improved from Exper-

iment 6 to Experiment 7. This might be expected, as Experiment 7 provided participants with 200 more items with which to gain experience of listening to distorted speech. A more informative comparison, in order to assess the retention of adaptation between Experiments 6 and 7, is to compare the Final Speech Recognition scores (Expt2 - Last 20 trials) with Thresholds for Sentence Recognition in the first block of Experiment 7. The mean recognition thresholds are 3.9 bands and 3.5 bands, for these Experiment 6 and Experiment 7 values, respectively - a paired t-test comparison indicates that this difference is marginally significant ($t(19) = 2.01$, $p = .059$, 2-tailed). However, closer inspection of the range of scores shows that, while the upper end of the range of threshold scores improves from 5.4 to 4.2 bands, the lower end changes from 2.6 to 2.8 bands from Experiment 6 to Experiment 7. This indicates a possible levelling-off effect i.e. those participants exhibiting very low thresholds (and hence the best speech recognition performances) at the end of Experiment 6 have improved less by Experiment 7 (or even become slightly worse) because they have already fully adapted to the noise-vocoded stimulus. To investigate this further, a Pearson's correlation was run on Experiment 6 Final Speech Recognition scores and the amount of improvement between Experiment 6 and 7 (calculated by subtracting the Experiment 7 Block 1 Sentence Thresholds from Experiment 6 Final Speech Recognition scores). This gave a highly significant positive correlation (Pearson's $r = .842$, $p = .000$, 2-tailed); the corresponding scatterplot is shown in Figure 8.8.

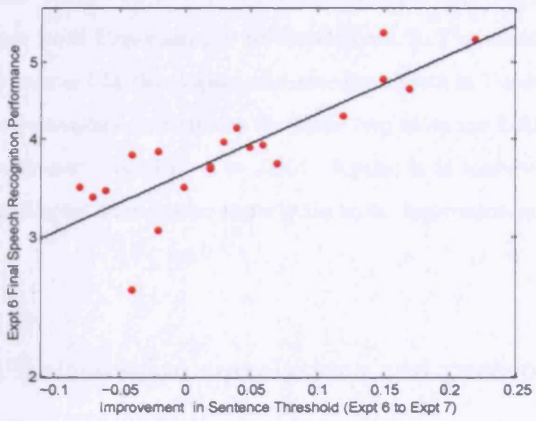


Figure 8.8: Relationship between Experiment 6 threshold performance and the improvement by Experiment 7. The improvement scores are shown in log_{10} format

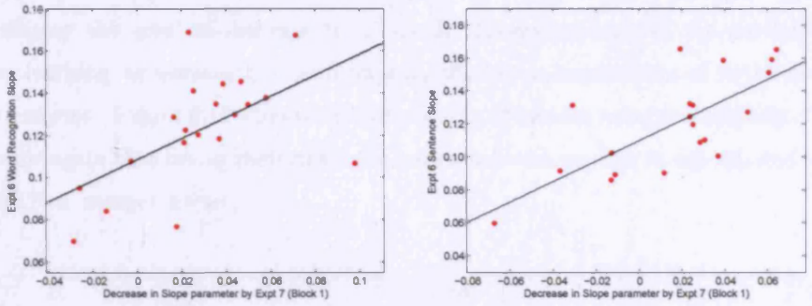
It is clear from this plot that there is a kind of ceiling, or 'asymptoting', effect in adaptation. Many of the participants with low thresholds from the end of Experiment 4 exhibit little to no

improvement in threshold in the first block of Experiment 7, while some participants' thresholds become slightly worse. In contrast, the participants who do not reach ceiling by Experiment 7 show an amount of improvement that is proportional with their performance in Experiment 6. This might suggest that, in cases where there is still room for improvement, these listeners are improving at roughly the same rate, despite their widely-ranging threshold scores. Again, this relates to the findings of Amitay et al. (2005) and Stacey and Summerfield (2007), that 'poor' listeners at baseline can improve more dramatically than 'good' listeners. However, the difference between these groups raises issues about what is actually being measured in the adaptation to noise-vocoded sentences, where listeners of high 'baseline' speech recognition ability can all improve at the same rate across a period of exposure and task practice. It seems that this question must be addressed directly in a context where adaptation can be slowed down sufficiently to avoid a ceiling effect - as previously suggested in the Discussion section of Experiment 6, the most intuitive way to do this in the context of noise-vocoding is to introduce a frequency-based shift, which has been shown to slow adaptation to a time-frame that approaches hours rather than minutes.

A similar comparison of Experiment 6 and Experiment 7 data was made for the slopes of the logistic functions in the two experiments, for curves fitted to the Sentence and Keyword data. Again, Experiment 7 scores were taken from the Block 1 curves only for these comparisons. For both Keyword and Sentence slopes, the parameter decreased significantly from Experiment 6 to Experiment 7 (Keywords: $t(19) = 2.379$, $p = .028$; Sentences: $t(19) = 2.542$, $p = .020$); hence, the slopes became steeper from Experiment 6 to Experiment 7. The scatterplots of Experiment 6 slope values versus Improvement in the slope parameter are shown in Figure 8.9, for both Keyword and Sentence plots - the two-tailed correlations for these two plots are both significant (Keywords: $r = .821$, $p = .000$; Sentences: $r = .742$, $p = .000$). Again, it is mainly those participants that showed steeper slopes in Experiment 6 who show little to no improvement, or a decrease in slope, in Experiment 7.

Experiment 7 - Within-session correlations and perceptual learning

Concentrating purely on the Experiment 7 results, correlations were run between the threshold and slope values for the Keywords and Sentences curves (Blocks collapsed). Unlike Experiment 6 data, there was no significant correlation between thresholds and slope parameters for either the Keywords or the Sentences curves, although the correlation assumed the same direction for Keywords as observed in the Experiment 6 data (Pearson's $r = -.337$, $p = .146$, 2-tailed; again, suggesting against intuition that higher thresholds correspond to steeper slopes).



(a) Keywords (b) Sentences

Figure 8.9: Relationship between Experiment 6 slope scores and the improvement by Experiment 7. Scores are shown in the original log_{10} format of the slope parameter.

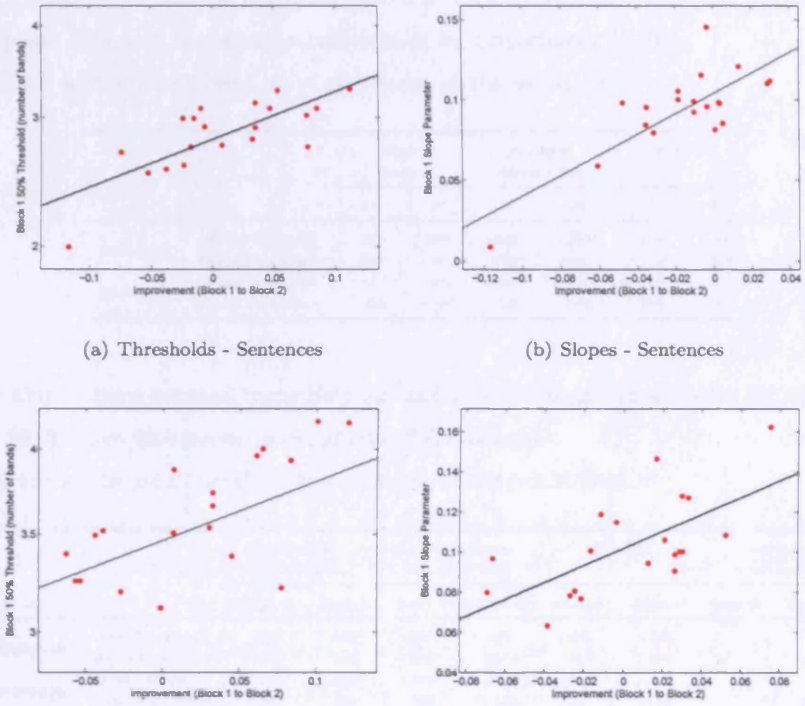
Following on from the observations of decreases in threshold and slope parameters (thus indicating increases in slope) between Experiment 6 and Experiment 7, and the emergence of apparent ceiling effect, analyses turned to the full data set in Experiment 7, to test for perceptual learning within this experiment from the first to second Block of 50 sentences. Repeated-measures ANOVA analyses were run with Speech Recognition score (threshold or slope measure, for Keywords and Sentences curves separately), with Block as the within-subjects variable and Sentence Order (AB or BA) as a between-subjects factor. Table 8.7 shows the descriptive statistics for the four dependent measures by Block. This indicates that perceptual learning involves a decrease in threshold over time, but the results for slope changes are contradictory between the Keyword and Sentence curves. However, none of the ANOVA analyses gave an effect of Block that was significant at the $p<.05$ level.

Table 8.7: Descriptive Statistics for logistic performance curves in Blocks 1 and 2 of Experiment 7.

		Block 1					Block 2				
		M	*SD*	*Min*	*Max*	IQR	*M*	*SD*	*Min*	*Max*	IQR
Keywords	50% Threshold	.453	.050	.300	.516	.058	.444	.038	.359	.503	.053
	Slope Parameter	.094	.027	.009	.145	.024	.110	.020	.080	.148	.028
Sentences	50% Threshold	.549	.046	.447	.621	.081	.526	.052	.383	.592	.066
	Slope Parameter	.103	.024	.063	.162	.099	.035	.028	.055	.162	.043

Inspection of the correlations and corresponding scatterplots between Block 1 performance and the amount of improvement from Block 1 to Block 2 offer an explanation for the non-significant ANOVAs. These show exactly the same pattern as exhibited for the changes between Experiment

6 and 7. For both thresholds and slopes, the participants who start with the highest values in Block 1 display the greatest decrease by Block 2. However, many of the participants exhibit little to no learning, or worsening in performance, offering an explanation as to the non-significant ANOVA analyses. Figure 8.10 illustrates this pattern of results using scatterplots. The findings indicate once again that better performance is associated with smaller thresholds and smaller slope parameters (i.e. steeper slopes).



(a) Thresholds - Sentences

(b) Slopes - Sentences

(c) Thresholds - Keywords

(d) Slopes - Keywords

Figure 8.10: Scatterplots of the relation between Block 1 performance and improvement across Experiment 7.

Relationship between the threshold and slope parameters

As for Experiment 6, bivariate Pearson's correlations were run between the curve properties (threshold and slope parameters) for individual speech recognition data across all of Experiment 7, and the scores on cognitive tasks (Total Score on BPVS-II, Total Score on Nonword Memory Test, Forward Digit Span Percentile Score, Backward Digit Span Percentile Score). As in the Experi-

ment 6 analyses, 1-tailed correlations were run based on the theoretical prediction of a significant relationship between speech recognition and performance on each of the additional cognitive measures. The results are shown in Table 8.9. To highlight any differences in the Experiment 6 data set when only the 20 participants who returned for Experiment 7 are considered, Table 8.8 directly compares the results of Experiment 6 correlational analyses for both the original 27 participants and the Experiment 7 subset.

Table 8.8: Correlations between cognitive tasks and speech recognition scores in Experiment 6, for all participants (27), and the subset who returned for Experiment 7 (20).
$*$ = significant at the $p<.05$ level, $**$ = significant at the $p<.01$ level.

		Digit Span		Nonword Memory Test		BPVS-II	
		27	20	27	20	27	20
Keywords	50% Threshold	.211	.223	-.348*	-.333*	-.326*	-.331
	Slope Parameter	-.336*	-.372	.372*	.435*	.110	.316
Sentences	50% Threshold	-.012	-.080	-.346*	-.378*	-.326*	-.311
	Slope Parameter	-.263	-.337	.164	.080	.093	.344

Table 8.9: Correlations between cognitive tasks and speech recognition scores in Experiment 6 and 7, for the 20 listeners who participated in both Experiments.
$*$ = significant at the $p<.05$ level, $**$ = significant at the $p<.01$ level.

		Forward Digit Span		Nonword Memory Test		BPVS-II		Backward Digit Span	
		Expt 6	Expt 7	Expt 6	Expt 7	Expt 6	Expt 7	Expt 6	Expt 7
Keywords	50% Threshold	.223	-.423*	-.333*	-.459*	-.331	-.234		-.199
	Slope Parameter	-.372	.194	.435*	-.026	.316	.186		.075
Sentences	50% Threshold	-.080	-.446*	-.378*	-.382*	-.311	-.147		-.205
	Slope Parameter	-.337	.100	.080	-.119	.344	.210		.118

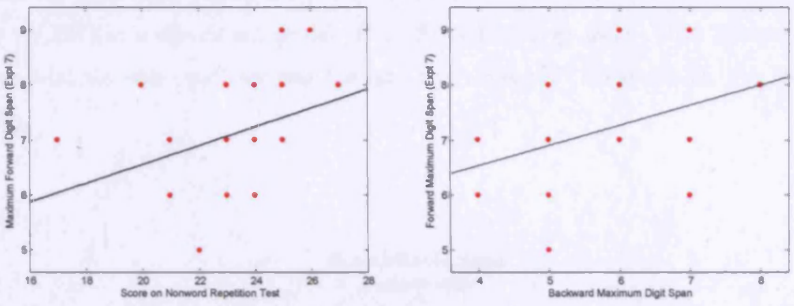
On first inspection, it appears that there is some inconsistency between the Experiment 6 behaviours of the original group of 27 participants and the 20-participant subset who participated in Experiment 7. For example, the correlation between scores on the BPVS-II and the 50% Threshold for Keyword recognition is significant for the 27 participants, but not for the 20-participant subset. However, it is more informative to compare correlation coefficients rather than probability values here, as the loss of 7 participants will certainly affect the significance of correlations. When looking purely at correlation coefficients, the relationship between Keyword thresholds and Vocabulary scores is closely matched for the two participant groups ($r = -.326$ for the 27-participant set, $r = -.331$ for the 20-participant set). In fact, all of the relationships of interest from the full participant set remain at least similarly strong when the set is reduced to 20. Some correlations become stronger for this subset; namely, the correlation between the thresholds and slopes for

Sentences recognition and scores on the Nonword Memory Test. There will be no discussion of other correlations that emerge as strong, or significant, within the 20-participant subset of Experiment 6, if these were weak for the 27 participants. The reasoning behind this is that, by virtue of the added statistical power for 27 participants, results from this full participant set should be taken as more indicative of the underlying relationships in the data.

On inspection of Table 8.9, we first consider those correlations of interest from Experiment 6. These are the correlation of Forward Digit Span and Word slope, the correlations of Nonword Memory Test scores with Word and Sentence thresholds and Word slopes, and the correlations of BPVS-II scores with Word and Sentence thresholds. The correlation between Forward Digit Span and the slope of the Word performance curve is reduced in Experiment 7 (Expt6: $r = -.372$; Expt7: $r = .194$, $p = .206$), as is the relationship between Nonword Memory Test performance and Word slopes (Expt6: $r = .435$; Expt7: $r = -.026$, $p = .456$). The correlations between BPVS-II scores and 50% thresholds for Keywords and Sentences are also reduced (Words: Expt6 $r = -.331$, Expt7 $r = -.234$; Sentences: Expt6 $r = -.311$, Expt7 $r = -.147$). In contrast, the correlations between Nonword Repetition and Word and Sentence thresholds become stronger in Experiment 7 (Words: Expt6 $r = -.333$, $p = .076$, Expt7 $r = -.459$, $p = .021$, 1-tailed; Sentences: Expt6 $r = -.378$, $p = .050$, Expt7 $r = -.382$, $p = .048$, 1-tailed). Interestingly, a new correlation emerges between Forward Digit Span and 50% threshold scores for both sentences and Keywords (Words: Pearson's $r = -.423$, $p = .028$, 1-tailed; Sentences: Pearson's $r = -.446$, $p = .024$, 1-tailed), suggesting that better performance on Forward Digit Span is associated with lower thresholds for noise-vocoded sentence recognition. This particular finding differs from the results of Experiment 6, which suggested a correlation between steep curve slopes and greater performance on the Forward Digit Span. There is no evidence of a correlation between Backward Digit Span performance and any of the measures of noise-vocoded sentence recognition.

In Experiment 7, there is a significant correlation within the cognitive tasks showing that longer Forward Digit Spans are associated with longer Backward Digit Spans (Pearson's $r = .408$, $p = .037$, 1-tailed). A marginally significant correlation also indicates that higher scores on the Nonword Memory Test are associated with better performance on the Forward Digit Span (Pearson's $r = .361$, $p = .059$, 1-tailed). Figure 8.11 shows the scatterplots of these two significant correlations.

The presence of a correlation between the two measures of phonological working memory prompted further analysis into how the three variables of interest - Speech Recognition Performance(Keywords and Sentences), Digit Span, Nonword Repetition - are related in this experiment.



(a) (b)

Figure 8.11: Scatterplots of the significant correlations between the cognitive tasks in Experiment 7.

In this case, the 50% thresholds are taken as the measures of Speech Recognition Performance. By running partial correlations between pairings of these variables (each time controlling for scores on the third variable), we can determine whether their correlations are due entirely to the third variable, or whether they represent sharing of a process that does not relate to this variable. This approach follows that taken by Pisoni and Cleary (2003) with measures of digit span, speaking rate and speech recognition in children with cochlear implants. It was decided to run 1-tailed correlations, as the direction of the relationships between the variables had been hypothesised a priori, and supported by the results of bivariate Pearson's correlations.

Figure 8.12 shows the results of the 1-tailed partial correlations in diagrammatic form (after the format used by Pisoni and Cleary (2003)). This shows that each of the phonological memory measures, Forward Digit Span and Nonword Repetition, makes an independent contribution to accounting for variability in noise-vocoded sentence recognition. The correlation between digit span and speech recognition performance is reduced to marginal significance when Nonword Memory Test scores are partialled out for both Keyword and Sentence measures. The relationship of the Nonword Memory Test to speech recognition performance, when digit span is partialled out, is inconsistent across speech perception measures. For keywords, the correlation remains marginally significant ($p = .065$), while for sentences, it becomes non-significant ($p = .136$). This suggests that, for perception at the level of Keyword recognition, nonword repetition is a better predictor of variability in speech perception scores than digit span. However, for whole sentence perception, digit span offers a better account of speech perception performance. These conclusions were supported by the results of two separate stepwise multiple linear regressions on Speech Recognition Performance, with Forward Digit Span and Nonword Memory Test as predictors. With Keyword

thresholds as the dependent variable, Nonword Memory Test emerged as the single predictor (Adjusted $Rsq. = .167$) in a significant model ($F(1, 18) = 4.81$, $p = .042$). With Sentence thresholds as the dependent, the sole predictor was Forward Digit Span ($F(1, 18) = 4.47$, $p = .049$; Adjusted $Rsq. = .155$).

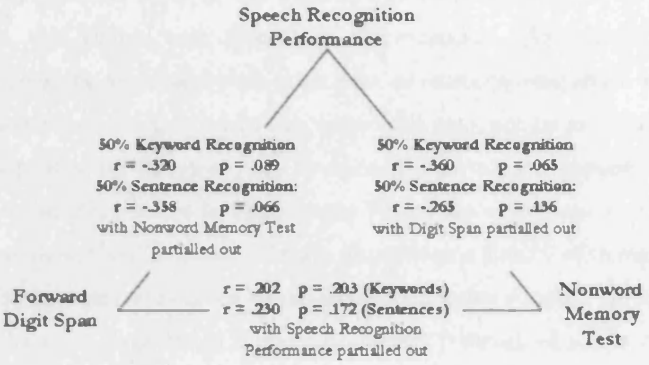


Figure 8.12: Partial Correlation coefficients between the three variables of interest in Experiment 7. Speech Recognition scores are in terms of log10 (number of bands) for 50% Keywords Correct and 50% Sentences Correct; Digit Span is represented by maximum span on Forward and Backward Digit Span (Experiment 7); Nonword Repetition is measured in terms of Total Score on the Nonword Memory Test

8.3.3 Discussion

The challenge with the curve-fitting approach to investigating individual differences is to decide upon which of the measures extracted from these curves is the most useful in offering a quantitative description of speech recognition and adaptation.

The main advantage of the constant measures technique was that the fitting of individual and group curves offered a visually interpretable comparative tool, from which two measures could be used to explore performance differences between individuals, and over time. Despite the differences between the adaptive tracking and constant measures metholodologies, a high correlation was found between curve-fitting measures extracted in Experiments 6 and 7, suggesting that these two

techniques are measuring the same underlying processes.

In Experiment 7, the outcomes of the correlations with cognitive scores differed from those in Experiment 6. First, vocabulary score (as measured by performance on the BPVS-II) was no longer a correlate of speech recognition performance, for both 50% Keywords Correct and 50% Sentences Correct measures. For the first of these performance measures, this null result could be interpreted in terms of the focus of the performance measure being taken away from whole sentence perception, and shifted to a lower level of perception. As posited in Experiment 6, linguistic knowledge may be associated with skills such as sentence completion and use of context, thus aiding whole sentence report, whereas this knowledge may not be as influential at the lower levels of perceptual processing. However, the absence of a correlation between vocabulary scores and 50% sentence recognition scores in Experiment 7 presents a problem for this interpretation. The explanation may lie in the differences between Experiments 6 and 7 with regard to adaptation. In Experiment 7 the listeners were more experienced with noise-vocoded speech, and inspection of the relationship between Experiment 6 and Experiment 7 threshold scores indicated a kind of ceiling effect. Crucially, Vocabulary scores were the only cognitive measure to correlate significantly with the rate of adaptation to noise-vocoded sentences in the adaptive track analysis of Experiment 6. Given that, for at least some of the listeners in this study, learning had begun to level off by the end of Experiment 7, there may have been less dependence upon the skills associated with vocabulary size in the current experiment.

In contrast, the relationship between nonword repetition and sentence recognition scores became stronger in Experiment 6. This was reassurance of the relationship between phonological working memory and speech perception. However, as acknowledged by Gathercole et al. (1994), successful execution of nonword repetition involves a number of stages. Dillon, Burkholder, et al. (2004) describe the stages of nonword repetition as (a) perceiving a completely novel sound pattern in auditory-only mode, without cues from pragmatics, semantics or lip-reading; (b) holding and verbally rehearsing the novel sound pattern in immediate phonogical memory; and (c) reassembling and translating the perceived sound pattern into an articulatory program to produce speech. Failure on the nonword repetition task could happen at any one of these stages. Further investigation would be needed to establish which aspects of phonological working memory are causing the differences in the Nonword Memory test scores across individuals in this experiment. Importantly, a task which is perceptual only, with no speech production or repetition component, would help to establish the contribution of speech production difficulties to the variability in nonword repetition scores in the present data set.

The Forward Digit Span scores provide the most confusing results of Experiments 6 and 7. Experiment 6 suggests that better performance on Forward Digit Span is associated with a steeper curve (for Keyword recognition), but the Experiment 6 data set also suggests that steeper curves are associated with higher word recognition thresholds. In contrast, Experiment 7 data suggest that higher Forward Digit Span scores are associated with lower Word recognition thresholds. It is initially unclear how these two results can relate to each other. However, the first step is to note that there is evidence of changes in both the threshold and slope parameters with perceptual learning - threshold parameter values became smaller (indicating lower thresholds) and slope parameters became smaller (indicating steeper slopes). From this, we interpret that an increase in slope of the performance curve is indicative of improved noise-vocoded speech recognition behaviour, as performance is expected to improve over time. Furthermore, the experimental design in Experiment 7 was more geared to achieving a better measure of curve slope than that obtained from the adaptive track data in Experiment 6. Therefore, regardless of the observed relationship between slope and threshold for the Experiment 6 data, the significant relationship of Forward Digit Span with slope indicates that higher scores on this task are associated with better noise-vocoded speech recognition performance. Similarly, the significant correlation between Forward Digit Span and 50% threshold (for Words and Sentences) suggests that listeners with greater Digit Spans perform better on noise-vocoded sentence recognition. An interesting element of the relationship between Digit Span and noise-vocoded sentence recognition is that, while there is evidence for a relationship between Forward Digit Span and speech recognition in both experiments, there is no evidence of a relationship between Backward Digit Span (only measured in Experiment 7) and speech recognition. This could be interpreted in terms of the differential demands of the Forward and Backward tasks on cognitive processes. The idea that these tests are not tapping variability in exactly the same cognitive processes is supported by the significant, but not complete, correlation between scores on the Forward and Backward Digit Spans in Experiment 7 (Pearson's $r = .408$, $p = .037$). It is possible that the Backward Digit Span is more demanding on executive control elements of working memory by requiring the listener to reverse the number sequence online, and that such post-encoding manipulations are not necessary in the repetition of short sentences such as those from the BKB corpus.

Part of the differing relationships of Forward Digit Span and noise-vocoded sentence recognition across the two experiments is likely to come from the poor re-test reliability of the Forward Digit Span measures in this study (Pearson's correlation of Expt6 and Expt7 Forward Digit Span Scores: r=.427, p=.030, 1-tailed). It is unclear why this correlation was so poor, as the tests were administered by the same experimenter in each Experiment, under identical test conditions. It is possible that Digit Span performance might fluctuate with listeners' attentiveness in the testing

session, and that this in turn might be related to the listeners' speech recognition performance. If so, a fluctuation of attentiveness might be reflected in a lack of correlation between Digit Span and speech recognition performance when cross-experiment correlations are run (e.g. Expt 6 Digit Span with Expt 7 Speech Recognition). This is not completely borne out when cross-experiment comparisons are made - there is still some evidence of a relationship between the Digit Span measures taken in one Experiment, and the speech recognition measures taken in the other. Unfortunately, it is outside the scope of this thesis to perform a more detailed investigation of the Forward Digit Span and its reliability as a measure of cognitive ability. What can be concluded from the pair of studies discussed in this chapter is that they have both, albeit in different ways, indicated a relationship between recognition of noise-vocoded speech and performance in the Forward Digit Span task. That no such relationship exists between Backward Digit Span and noise-vocoded speech recognition may indicate that the relationship of Forward Digit Span to noise-vocoded speech perception may have more to do with sustained attention and engagement with the task than any involved online manipulation of encoded phonological information, as is more heavily required in the re-ordering process of the Backward Digit Span.

We can conclude that Experiment 6 has succeeded in replicating the findings of Experiment 7, in identifying a relationship between two measures of phonological working memory (Forward Digit Span and the Nonword Memory Test) and noise-vocoded sentence recognition. As discussed above, the weakening of a relationship between vocabulary size and noise-vocoded speech recognition in Experiment 7 may be due to the fact that the skills associated with Vocabulary size may be more in demand during the learning/adaptation phase of performance, which for several participants has ended (or at least slowed down) by the end of Experiment 7.

In Experiment 7, both Forward Digit Span and Nonword Memory Test scores are significantly related to 50% speech recognition thresholds for both Keyword and Sentence performance curves, and marginally significantly related to each other (Pearson's $r = .361$, $p = .059$). Further analysis to explore the inter-relationship between these three variables produces two different pictures for perception of keywords within sentences on the one hand, and perception of whole sentences on the other. At a 'keyword level' of listening, it appears that while both digit span and nonword repetition both account for variability in speech recognition performance, it is nonword repetition which provides the slightly better account. In contrast, digit span provides the greater account of speech recognition variability when the participant is listening at 'sentence level' (i.e. when the speech recognition scores reflect correct recognition of whole sentences). One possible interpretation of these findings is to suggest that, although both Digit Span and the Nonword Memory Test load on phonological working memory, they differ in the extent to which they tap variability in different

aspects of this cognitive capacity. The Nonword Memory Test loads heavily on the mapping from unfamiliar speech input to phonological representations (what Gathercole et al. (1994) identify as 'Phonological Analysis'), while the use of highly familiar material (numbers from 1-9) in the Digit Span task reduces the load on this mapping process. In terms of the loading on subvocal rehearsal, the retention of 2-5 syllables of spoken material is perhaps less challenging for the adult listener than rehearsing a sequence of 6 or 7 digits read at a steady pace. This difference between the tests has implications for the success of perceptual encoding, which was discussed in Chapter 6. At lower levels of spectral resolution, the mapping from sound to representation is less certain than when greater acoustic detail is available. In the context of the current experiment, the mean 50% Keywords threshold was 2.8 bands, while for 50% Sentences, this value was 3.4 bands (this difference was significant in a paired t-test: $t(19) = 17.54, p = .000$). Therefore, at the point where participants are at half of maximum performance for Sentence recognition, the speech signal is less distorted. This potentially places less pressure on the perceptual encoding process, as sound-to-representation mappings may be more certain at this level of higher spectral resolution. Hence, a weaker relationship emerges between nonword repetition and speech recognition for Sentences thresholds. As Forward Digit Span is perhaps less sensitive to variability in perceptual encoding and more tuned to measuring the capacity for rehearsal of phonological information, its relationship to speech recognition measures remains relatively similar across distortion levels, in keeping with the reasonably consistent item durations in the BKB corpus.

The above interpretation should be treated with a little caution in respect of the claim that the Digit Span loads on memory for familiar items while the Nonword Memory Test does not. It cannot be claimed that the items of the Nonword Memory Test are entirely unfamiliar, as there are elements of 'wordlikeness' present in several of the nonwords. For example, some of the items contain sound patterns recognisable as morphological endings in English e.g. instadront*ally*, dexiptec*astic*. Previous studies of wordlikeness in nonword repetition tests has shown that items rated as having high wordlikeness (by normal-hearing adults) are more successfully repeated by children than those exhibiting low wordlikeness (Gathercole, Willis, Emslie, & Baddeley, 1991; Gathercole, 1995). Such ratings were unavailable for the Nonword Memory Test used in the current experiment, nor was there time to obtain ratings from a new group of participants. However, it is important to note that elements of similarity to real words may have allowed lexical representation to assist the phonological mapping process for some of the nonwords presented, even if this lexical facilitation was relatively much weaker than for the Digit Span.

An important theme in Experiment 7, which was touched upon earlier in the Discussion, is the 'ceiling effect', or slowing of learning, that became apparent in the sentence recognition data.

This will undoubtedly have affected the correlational analyses, as an important consequence of the effect was a narrowing of the range of scores (thresholds and slopes) across the participants. This cannot be ignored in the assessment of the results of Experiment 7, and this, of course, presents a difficulty for the comparability of Experiment 6 and Experiment 7 findings. However, importantly, Experiment 7 reaffirmed the presence of a relationship between speech recognition measures and scores on both Forward Digit Span and nonword repetition, and in doing this fulfilled the aims of the experiment. Where an effect was not replicated (i.e. for BPVS-II scores and speech recognition), there was a plausible explanation in terms of vocabulary size as a correlate of adaptation (which was limited in Experiment 7).

A final word should be said on the mathematical issues associated with the analysis of the Experiment 7 data. Logistic functions were fitted to the data in Experiment 7 in order to obtain speech recognition performance measures in terms of the number of bands needed to reach a recognition criterion with noise-vocoded speech. By fitting these functions to the data, performance across the whole range of recognition scores, from floor to ceiling recognition, has an effect on the criterion measure extracted. This gives a better measure of overall performance than a simple averaging of scores across all distortion levels, as the latter approach results in high scores at less distorted presentation levels 'smearing out' the individual differences at more difficult distortion levels. However, there are problems with fitting mathematical functions to behavioural data. With logistic regression, there are a number of assumptions that should not be violated. For example, the plotting of a logistic function of Proportion Keywords Correct across distortion levels in this study violates the assumption of independent measures, as the probability of getting a keyword correct in a sentence will often be influenced by whether other words in the sentence were recognized. In the case of noise-vocoded speech recognition, curve-fitting is made more complicated by the fact that noise-vocoded speech is a learnable stimulus and the listeners performance at each distortion level has the capacity to change across the course of the experiment. It was hoped, however, that the constant measures approach would offer a more interpretable investigation of both the position *and* shape of the speech recognition function across listeners than could be achieved through adaptive tracking. As the results stand, the data sets from Experiments 6 and 7 give a mixed view of whether slope is indeed an interpretable marker of performance variability.

8.4 Summary

The experiments described within this chapter have used two different methods to quantify individual differences in perception of noise-vocoded speech, and to identify cognitive correlates of

perceptual ability.

Overall, despite being a more labour-intensive technique, the constant measures approach described in Experiment 7 emerges as the more favourable. The plotting of individual performance functions offers a readier visual representation of individual differences, and the potential for a greater variety of measures to be extracted in terms of function shape and position. The challenge associated with this wider range of measures is to decide upon which are the more interpretable in terms of reflecting perceptual processes.

Both experiments in this chapter attempted to identify the cognitive correlates of individual differences in speech perception. It was predicted that better speech recognition scores would be correlated with a larger vocabulary score, greater digit span and better nonword repetition. Across the two experiments, and using the preferred approach of logistic curve-fitting to quantify performance, the variables that were most consistently associated with noise-vocoded speech recognition were performance on the Nonword Memory Test and the Forward Digit Span. Therefore, the current data set suggests a role for working memory as a correlate of perception of noise-vocoded speech in the normal-hearing population. However, it should be noted that the strength of the correlations between sentence recognition and Forward Digit Span were not much greater than those observed by Eisenberg et al. (2000).

Within the general finding of a role for phonological working memory, it seems that there are roles for memory for familiar items (as measured by Digit Span tasks) and memory for phonological information (as measured by both the Nonword Repetition Test and the Digit Span). Further studies need to address the component processes in these memory tasks. For example, do visual and spatial memory spans also correlate with noise-vocoded speech performance? Also, can we rule out the possibility that the variability in nonword repetition performance is driven by individual differences in motor planning or speech production rather than variability in phonological working memory?

An important issue which has emerged from both experiments described in this chapter is the relationship between 'baseline' speech perception capabilities and adaptation to a difficult stimulus over time. The results of the adaptive tracking procedure in Experiment 6 suggested that initial performance with noise-vocoded speech was not related to cognitive measures of vocabulary and working memory, while the amount of adaptation to the distortion during the experiment was related to these measures. However, this initial measure was based on 20 sentences, a sufficient sample to allow a significant amount of adaptation (Davis et al., 2005). Therefore, the measures of Initial Speech Recognition Performance in Experiment 6 will be, to a certain extent, already

contaminated with the effects of learning. A way to address this issue in future experiments would be to present listeners with a stimulus type to which they can adapt less readily, such as spectrally-shifted noise-vocoded speech. However, the general pattern of greater learning by those with poorer baseline performances is in line with the findings of Amitay et al. (2005) and Stacey and Summerfield (2007), and presents an important demonstration of individual differences in perceptual learning.

One of the more interesting findings in the current experiment is the overall observed retention of perceptual learning from Experiment 6 to Experiment 7. These experiments were run approximately 2 months apart. Altmann and Young (1993) tested a group of participants on recognition of time-compressed sentences, in two sessions spaced by 12 months. On the participants' second visit, the authors found that they gave significantly better speech recognition performance than a naïve group of listeners. The current findings indicate that a similar long-term aspect to learning of noise-vocoded speech. However, a replication of this effect using precisely the same methodology in the two testing sessions (including the number and distribution (in terms of distortion level) of items) is necessary to affirm this finding.

Chapter 9

Listener variability: Linguistic factors

Abstract

Experiment 8 addresses the question of how individual differences in speech recognition and perceptual learning are affected by the linguistic properties of the task materials. Twenty-eight native speakers of English participated in two sessions in which they were tested on recognition of noise-vocoded sentences, words and segments (consonants and vowels). The resulting data is analysed at both group and individual level, using a selection of different techniques to explore the processing of noise-vocoded speech in these different contexts.

9.1 Introduction

The experiments of the thesis to this point have concentrated on the recognition of, and perceptual adaptation to, noise-vocoded sentences. However, there are several reasons to look beyond sentences in characterizing individual differences in speech perception. In Experiment 2a of the thesis, individual scores on the recognition of noise-vocoded sentences from the LSCP corpus bore no apparent relationship to the recognition of sentences in noise or scores on a speech-reading task, where both the latter tasks used materials from the BKB corpus. These findings are at odds with those of Watson et al. (1996), who found a strong correlation between performance on auditory and visual speech perception, claiming a modality-independent source of individual variability. In Experiment 2a, the noise-vocoded sentences were much more complex than the BKB items along several parameters, including length and lexical, syntactic and semantic complexity. The lack of significant correlations between the two sentence types suggested that although there was a possibility that the skills needed to perform noise-vocoded sentence recognition are different from those needed on the other tasks, it was more likely that the effects of linguistic properties of the materials were masking the potential cross-modal relationships. There is evidence from numerous sources of linguistic effects on speech recognition. For example, word recognition in sentences is better when the sentences are highly predictable (Kalikow et al., 1977). This effect has been shown recently for recognition of noise-vocoded sentences. Using Kalikow et al.'s SPIN sentences, Obleser et al. (2007) showed a highly significant advantage for recognition of the final word in sentences like 'He caught a fish in his NET' over sentences like 'Sue discussed the BRUISE' (where the to-be-reported word is shown in capital letters) under conditions of intermediate signal degradation (noise-vocoding with 8 bands). A study by Grant and Seitz (2000) showed that use of 'top-down' contextual information in sentences varies across individuals. They presented 34 hearing-impaired listeners with filtered sentences from the IEEE corpus, and their constituent keywords in isolation, at three different intelligibility levels. Using Boothroyd and Nittrouer's (1988) equation explaining the relationship between word recognition in sentences and in isolation (See Equation 9.1), Grant and Seitz (2000) calculated individual k-factor scores at each intelligibility level. This k-factor represents the listener's ability to use semantic and morpho-syntactic information in the sentence to identify the words within it, with a high k-factor corresponding to better use of this context. Grant and Seitz observed considerable variability in the k parameter across their listening population. Moreover, they found that k became larger as the difficulty of the listening situation was increased (through filtering in the frequency domain).

Other studies have demonstrated the powerful influences of top-down information on speech perception at a lower level. R. Warren (1970) showed that replacing (not masking) a speech sound

in a word with a cough or a tone resulted in listeners believing that they really had heard the missing speech sound. In a later study, R. Warren and Warren (1970) showed that the exact speech sound perceived in this 'phonemic restoration' hallucination changed to fit sentence context. The aim of this strand of the thesis is to characterize individual differences in the performance of noise-vocoded speech recognition tasks, and in the ability to adapt to these distorted speech stimuli (perceptual learning). Given the previous demonstrations of strong top-down influences from the sentence and lexical level in the speech recognition process, and the results of Experiment 2a, it is difficult to differentiate the influences of 'top-down' and 'bottom-up' processes on individual variability, and how these change over time, using sentence recognition alone. Therefore, this experiment aims to describe individual differences across a range of linguistic levels - Sentence, Word and Segment (Consonants and Vowels).

Individual differences across linguistic levels have been explored before, in the cochlear implant literature. Rabinowitz, Eddington, Delhorne, and Cuneo (1992) tested 20 cochlear implant users on recognition of sentences (of differing difficulty), monosyllabic words, consonants and vowels. They identified strong correlations across all levels, from word recognition in sentences to isolated words, segments and their underlying phonetic features (scores for which were obtained using Information Transfer analyses (G. Miller & Nicely, 1955). However, they took the approach of comparing scores when certain top-down influences had been accounted for e.g. words-in-sentences scores were modified to take account of the overall k-factor of the sentence set before comparison with the scores on isolated words. This approach reflects the authors' interests in low-level signal processing in cochlear implant users, which is likely to be of much greater influence on speech recognition than contextual processing in that particular participant group. In the normal-hearing population, however, it is of interest to take these higher level processing factors into account. The results of Surprenant and Watson's (2001) study of individual variability in speech-in-noise recognition indicate that speech recognition processes are far from identical across different linguistic levels - Pearson's correlation coefficients between speech-in-noise recognition of CV-units, Words and Sentences and a clear-speech syllable identification task ranged from only 0.25 to 0.47 in their experiment. In the current experiment, the aim is to use a similar set of tasks to Rabinowitz et al. (1992), but to employ a different design in two respects. First, no adjustments will be made to individual recognition scores before cross-task correlations are assessed, such that the resulting correlation values will reflect all sources of contributing variability in the scores, both 'top-down' and 'bottom-up'. Second, an important element of the current study will be to investigate how the inter-relationship of the different tasks varies over time, with the effects of perceptual learning. It is of interest to observe whether retention of perceptual adaptation can be demonstrated over a long-term absence of exposure, as was indicated by the improvement in sentence recognition

scores between Experiments 6 and 7. While Experiments 6 and 7 employed slightly different methodologies from each other, the current study will improve the comparability of testing sessions by adopting the same design in each, and further allow us to investigate whether Words and Segments are also amenable to adaptation.

The current experiment tests normal-hearing listeners on perception of noise-vocoded sentences, words, consonants and vowels in five separate tasks. To test for long-term adaptation, the listeners make two visits to the lab, separated by at least 1 week. Sentence recognition is assessed using two sentence sets differing in overall complexity - the BKB and IEEE sentences - in order to re-visit the issue emergent from Experiment 2a (the lack of correlation across different sentence sets). Monosyllabic word recognition is tested using items from the Boothroyd (1968) AB lists, while recognition of isolated segments is tested separately for Consonants in a VCV ('Vowel-Consonant-Vowel') context and Vowels in a CVC ('Consonant-Vowel-Consonant') context. To allow the fitting of psychometric performance functions, tasks will feature a range of distortion levels (as quantified by the number of bands in noise-vocoded speech). It is hypothesised that individual scores across the five tasks will be significantly correlated, but that individuals' differing abilities to use top-down and bottom-up sources of information will limit the strength of these correlations. A set of planned analyses will attempt to describe the interaction of these levels of processing.

9.2 Experiment 8

9.2.1 Method

Participants

Participants were 28 native speakers of British English (aged 18-40, 12 male), who reported as non-bilingual, with no known language or hearing problems. All participants were recruited from the UCL Department of Psychology Subject Pool, and were naïve to noise-vocoded speech.

Materials

Listeners were tested on perception of 5 different stimulus types, all vocoded to 1, 2, 4, 8, 16 and 32 channels. The items were also available in undistorted form (which will be referred to as

'Clear Speech'), giving seven presentation conditions (or distortion levels) in total. Recording and vocoding routines were as described in the Method section of Chapter 8.

1. **Simple Sentences**. One-hundred-and-forty items from the BKB sentence corpus (Bench et al., 1979), as described in the Method section of Chapter 7. Each of the 140 items was available at the seven distortion levels.

2. **Low Predictability Sentences**. 140 items from the IEEE sentence corpus (IEEE, 1969), as described in the Method section of Chapter 6. Each of the 140 items was available at the seven distortion levels.

3. **Single Words**. 140 items from the phonemically-balanced Boothroyd AB lists (e.g. *gas*, *mice*, *whip*; Boothroyd (1968)). These are routinely used in audiological assessment. Each of the phonemically-balanced AB lists contains 10 CVC (consonant-vowel-consonant) test items, plus two practice items. Due to a problem with the recording procedure, the practice items from six lists had to be used to make up the full set of 140 items for the current experiment. This led to the inclusion of some non-CVC words - *oil* and *pour*, which have the structures VC and CV, respectively, and to the repetition of three items - *five*, *good* and *shop*.[1] Furthermore, a mis-recording of the word *bomb* as *bombs* meant that this item had CVCC structure in the test. Each of the 140 selected items was available at each of the seven distortion levels.

4. **Consonants**. Listeners were tested on perception of 17 consonants: b, d, f, g, ʤ, k, l, m, n, p, s, ʃ, t, v, w, j, z. One token of each consonant was recorded in the context */a:/-C-/a:/*, where *C* is a consonant e.g. *apa*, *aga*, *ala*. Each token was available at all of the seven distortion levels.

5. **Vowels**. Listeners were tested on perception of 17 vowels, including a combination of monophthongs and diphthongs: æ, eɪ, ɑ:, ɛ:, i:, iə, e, ɪ, aɪ, ɜ, ɒ, əʊ, u:, ɔ:, aʊ, ɔɪ, ʌ. One token of each vowel was recorded in the context */b/-V-/d/*, where *V* is the vowel e.g. *bad*, *beard*, *boyed*. Each token was available at all of the seven distortion levels.

[1] In the experiment, all repetitions were presented as different tokens of the same word. The two presentations of *good* and *shop* were split across the two sessions, while the repetition of *five* occurred within the same session.

Design

The listeners made two visits to the lab, separated by 7-15 days (N = 27: M = 10.44 days, SD = 2.69)[2], with the exception of one participant who could only return after 78 days. It was believed that 7 days would be a sufficient delay for episodic memory of the repeated items (Consonants and Vowels) to dissapate, as well as the immediate effects of task practice (i.e. those effects unrelated to the perceptual task).

The intention to assess the data in terms of group effects and individual differences meant that there was a trade-off between these two goals in the design. As the main motivation of the experiment, and this strand of the thesis, is to investigate individual differences, each session involved the same ordering of tasks - BKB, IEEE, Words, Consonants, Vowels - across all subjects. However, to facilitate analysis of group effects, there was randomisation of items within the subtests and counterbalancing of the order of presentation of sentence and word materials across the two sessions.

All stimulus presentation routines were programmed and run in MATLAB v7.1 (The Mathworks, Inc., Natick, MA). The design of the sub-tests was as follows:

Simple Sentences, Low Predictability Sentences and Words. Each session featured 70 items, with 10 at each distortion level. At the level of Testing Session, items were counterbalanced by labelling 70 of each stimulus type as Set A, and the other half as Set B. Half of the participant group (14 participants) received Set A items for the three tasks in Session 1, while the other half received Set B items in Session 1. Within each task, the order of presentation was pseudorandomised. The 70 items (i.e. their linguistic content) were completely randomized across the task but the task was constrained such that within each chronological block of 7 sentences there was an example from each distortion level. The distortion levels were, however, randomized within these

[2]Recruiting 28 participants who could guarantee two visits to the lab was incredibly difficult and time-consuming. Many more than this number participated in Session 1, but several data sets had to be abandoned due to the participant failing to report for Session 2. To maximise the chances of Session 2 attendance, a certain amount of flexibility in the exact timing of the return visit had to be allowed. This unfortunately lead to an imbalance of item presentation order and the length of time between sessions in the design, as presentation order was assigned to each participant in Session 1 when the exact timing of the second visit was not definite. The mean inter-session delay for participants receiving Set A items in Session 1 (excluding the outlier of 78 days) was 8.46 days (SD = 2.47), while for participants receiving Set B items in Session 1, this was 12.29 days (SD = 1.14). However, rigorous inspection of the data with 'Time Delay' as a covariate presented no obvious systematic effect of the size of the delay on the performance of the listeners (whether or not the subject with the 78-day inter-session delay was included). Therefore, none of the reported results in this chapter will include Time Delay as a factor or covariate.

sentence blocks. Hence, the listeners' exposure to each of the seven distortion levels was spread evenly across the task without the upcoming level being predictable within the blocks.

Consonants and Vowels. The Consonants and Vowels were tested in two separate tasks. Each of the tokens was repeated at all of the seven distortion levels, and the whole list of items was fully randomized. Hence, each task contained 119 items (17 tokens at each of 7 distortion levels). In these tasks, exposure to the distortion levels was not chronologically constrained as this was not included in the task script.

Procedure

All test materials were presented over Sennheiser HD25-SP headphones in a quiet room. The QuickMix software package (Version 1.06; Product Technology Partners, Cambridge, UK) was used to ensure the same comfortable volume setting for each participant. Participants were given informal face-to-face instructions by the experimenter before receiving detailed instructions on-screen. The listener was not given any example stimuli before the experiment began, although they were told that the stimuli would be distorted and that some examples would be more difficult to understand than others. It was felt that to present an example of each stimulus type in advance would give the participants too much practice, yet to present only one type may give an unfair advantage to that test, or bias the listener's expectations. The Sentence and Word tasks were open-set recognition tasks. The participant heard a stimulus played once over the headphones and then had the opportunity to type the item (Sentence or Word) content into an onscreen response bar. Responses were self-timed and the next stimulus was triggered by pressing space bar on the computer keyboard. The listener was encouraged to type as much as possible from what they heard (and that partial answers were acceptable), but were also told that it was fine to leave a blank response bar if the item was completely unintelligible. In constrast, the Consonants and Vowels tasks adopted a forced-choice paradigm. Each task had 17 alternatives, which were given to the participant on a printed sheet which remained in view for the duration of the task. The participant was told that they must give a response from the selection to every stimulus and leave no gaps - this was to facilitate the construction of confusion matrices in the analysis. The response alternatives for the two tasks were printed in exactly the format in which they would be heard ('*/a/-consonant-/a/*' and '*/b/-vowel-/d/*') and their text-to-pronunciation relations were unambiguous, with the only exception being one pair - '*bowed*' and '*bode*' - which was disambiguated orally by the experimenter before the Vowels task began. As with the Sentences and Words tasks,

the participants self-timed their responses. However, all five tasks contained no breaks and the participants were encouraged to progress through them at a steady rate.

9.2.2 Results

The data set for this experiment is large and multifactorial, with numerous possible lines of statistical analysis. For this reason, a structured, 3-stage approach to the data was planned.

1. **Main group effects of Task, Level and Session.** The first analysis aims to explore the group recognition data to identify the presence/absence of basic effects of Task (BKB, IEEE, Words etc.), Level (Number of Bands) and Session (Testing Session 1 or 2), and their interactions. This will be done using a repeated-measures ANOVA.

2. **Quantifying and Characterizing Individual Differences.** Experiments 2a, 6 and 7 explored the cognitive correlates of variability in recognition of noise-vocoded BKB sentences. The main motivation of the current experiment is to expand the investigation of variability by measuring recognition across a selection of vocoded stimulus types. This stage in the analysis is considered the most central to the research questions of the experiment, and consequently forms the bulk of the analysis. There are three phases to this analysis

   (a) *Calculating k-factors.* This analysis uses raw proportion scores on the Sentences and Words recognition tasks to calculate k factors (after Boothroyd and Nittrouer (1988)) for each listener that represent the use of sentential context in word recognition. In this way we can identify the range of this ability in the participant sample, as was done for speech-in-noise perception by Grant and Seitz (2000).

   (b) *Fitting psychometric functions - Correlational analysis.* The psignifit curve-fitting package Wichmann and Hill (2001a, 2001b) is used to fit logistic functions to group and individual recognition scores on the five tasks in the experiment. Pearson's correlations are run in order to establish a model of inter-relationship between the tasks. This is done for overall performance scores (two sessions collapsed), and for the two testing sessions separately.

   (c) *Fitting psychometric functions - Common Factor Analysis.* This is run on individual recognition threshold scores on the five tasks in order to explore the number and nature of the main processing factors loaded upon by the perceptual tasks. Separate factor analyses are run for the two test session data sets.

3. **Exploring Acoustic-Phonetic Perception** The forced-choice nature of the Consonant and Vowel tasks in this experiment allowed for confusion matrices to be constructed for each participant. Subsequent Information Transfer analyses are used to measure the amount of phonological feature information transferred in each task, for each listener. Multiple regression analyses allow for the assessment of the relative contribution of the different phonological features to the performance of the consonant and vowel tasks.

Data Scoring and Preparation for Analysis

For the sentences (BKB and IEEE), participants' responses were marked according to the number of keywords correctly reported in each sentence. These individual item scores were summed at each distortion level - 1, 2, 4, 8, 16, 32 bands and Clear Speech.[3] For Analysis 1 and 2(a), scores were organised in terms of the Proportion of Keywords Correct at each distortion level, by dividing the total number of keywords correct by the total number presented. For Analysis 1, these scores were then transformed using the arcsine transformation.

For Analysis 2(b) and (c), as in Experiments 6 and 7, individual data points were obtained by fitting logistic functions to the recognition data and extracting the 50% threshold and slope scores (in terms of log_{10}(no of bands)). These were constructed with the raw recognition data (Number Keywords Correct at each distortion level, along with the total number of keywords presented), but excluding the clear speech results, as there is not a straightforward relationship between clear speech and the 'number of bands' scale along which the vocoded speech conditions can be arranged. Data for Analysis 2(b) and (c) were arranged for the two testing session separately. For analyses on Overall Performance, additional logistic functions were fitted for the summed recognition scores across the two sessions. Thus, for each participant, a total of 15 psychometric functions was fitted (2 Session curves and 1 Overall Curve, for each of 5 tasks).

Responses in the Words recognition task were scored in terms of the Number of Items correct at the different distortion levels. Proportion Items Correct scores were arranged by distortion level for Analyses 1 and 2(a). The data were arranged for Analysis 2(b) and (c) as described above for the Sentence scores, with the only difference being that the Words test scores are in terms of Items Correct rather than Keywords Correct.

[3]For two participants, the computer crashed during the Session 1 run of the BKB sentences. However, as only the last 7 items were missed for one participant (one at each distortion level), and two for the other (one at 2 bands and one at 16 bands), their data were still used, with proportions calculated out of amended totals.

For Analyses 1 and 2, responses in the Consonants and Vowels tasks were scored and arranged as the Words. For Analysis 3, the scores were arranged as confusion matrices for each participant. An example confusion matrix is shown in Table 9.1 - in this format, the presented items label the rows, while the columns represent the participant's response. The number in any particular cell represents the number of presentations of the item labelling the row that were identified by the participant as the item labelling the column. For example, Table 9.1 shows that the consonant /ʃ/ was mis-identified as /s/ on 3 occasions.

Table 9.1: An example confusion matrix for a single participant in the Consonant perception task (sessions collapsed) - the row labels represent the presented stimuli and the columns contain frequency data for the responses given.

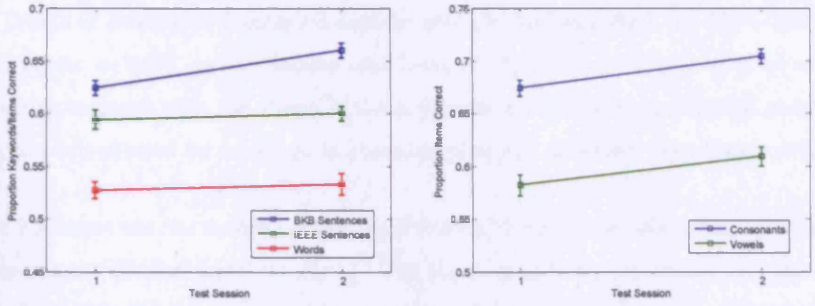
		Response															
		b	d	f	g	ʤ	k	l	m	n	p	s	ʃ	t	v	w	j	z
Stimulus	b	10	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
	d	1	6	0	1	1	1	0	0	0	2	0	0	0	0	0	0	0
	f	0	0	11	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	g	0	4	1	5	0	1	0	0	0	1	0	0	0	0	0	0	0
	ʤ	0	0	1	0	9	0	0	0	0	0	0	0	2	0	0	0	0
	k	0	0	0	0	0	11	0	0	0	1	0	0	0	0	0	0	0
	l	0	0	1	0	0	0	7	0	0	0	0	0	0	0	4	0	0
	m	0	0	0	0	0	0	1	7	0	0	0	0	0	0	4	0	0
	n	0	0	1	0	0	0	0	0	8	0	0	0	0	0	2	0	1
	p	0	0	0	0	1	0	0	0	0	8	0	0	2	0	1	0	0
	s	0	0	9	0	0	0	0	0	0	0	1	1	0	0	0	0	1
	ʃ	0	0	4	0	0	0	0	0	0	0	3	5	0	0	0	0	0
	t	0	0	0	1	0	2	0	1	0	1	0	0	7	0	0	0	0
	v	0	0	1	0	0	0	0	0	0	0	0	0	0	10	0	1	0
	w	0	0	1	0	0	0	1	0	1	0	0	0	0	1	8	0	0
	j	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4	6	1
	z	0	0	1	0	1	0	0	0	0	0	0	0	0	6	0	1	3

In addition to the above, overall group logistic functions were fitted for each task, for the separate testing session performances, and for performance collapsed across the two sessions. As above, the clear speech data were not included in these fits.

Analyses

Figure 9.1 represents the group mean scores in the five tasks, with error bars, for each testing session. As the sentence and word tasks were open set recognition tasks, while the Consonants and Vowels tasks were forced-choice, it helps to consider the figure in two parts. For the open-set tasks (Figure 9.1(a)), the difficulty of the task increases in the order BKB, IEEE, Words. There is a marked overall improvement between testing session 1 and 2 for the BKB sentences, but this improvement is more modest for the IEEE sentences and Words. A comparison of the closed-set tasks (Figure 9.1(b)) shows that performance on both improves between testing sessions, and that

recognition scores for the Consonants task are considerably greater than those for the Vowels task in both sessions.



(a) Open Set Recognition - Sentences and Words

(b) Closed Set Recognition - Consonants and Vowels

Figure 9.1: Mean Proportion Items/Keywords correct, with error bars showing ±1 standard error of the mean.

Analysis 1

Repeated-measures ANOVA

Data from the 28 participants were entered into two repeated-measures ANOVA analyses - one for the open-set tasks (BKB Sentences, IEEE Sentences, Words) and one for the closed-set tasks (Consonants and Vowels), with arcsine-transformed Proportion Correct scores as the dependent variable in each case. The within-subject variables were Task (BKB, IEEE, Word / Consonants, Vowels), Level (7 levels of spectral resolution - including undistorted speech) and Session (2 levels). Version (Item order AB or BA) was the between-subjects factor.

Open-Set Recognition Tasks

There was a significant effect of Task, reflecting the observed difference in difficulty across the different stimulus types ($F(2, 52) = 121.80$, $p = .000$, $\eta^2 = .824$, power = 1.00), and a significant effect of Level ($F(7, 156) = 2771.37$, $p = .000$, $\eta^2 = .991$, power = 1.00), reflecting the predicted increase in intelligibility with an increase in the number of bands. There was also a significant effect of Session ($F(1, 26) = 9.36$, $p = .005$, $\eta^2 = .265$, power = 0.838), indicating that learning took place

between the first and second testing sessions on each task. A significant Task*Session interaction ($F(2,52) = 4.89$, $p = .011$, $\eta^2 = .158$, power = .781) suggests that the relative differences between the tasks were altered over time/exposure, and therefore that the tasks improved to differing extents. Task and Level also interacted significantly (Wilks'Lambda $F(12,15) = 22.51$, $p = .000$, $\eta^2 = .947$, power = 1.00), as did Session and Level ($F(6,156) = 3.04$, $p = .008$, $\eta^2 = .105$, power = .902), which suggests that the shape of the psychometric relationship between number of bands and recognition is altered by a change in stimulus category, and with the passage of time.

Figure 9.2 shows the raw means for the Task*Session*Level interaction, which was non-significant in this experiment (Wilks' Lambda $F<1$). The figure clearly demonstrates the significant interaction of Task with Session, where there is visible improvement (leftward shift of the curve) in performance from Session 1 to Session 2 for the BKB sentences, but not for the other two tasks. What could explain this interaction? On a basic explanation, each of the sentences contains several words (including 3 keywords for each BKB sentence and 5 for each IEEE sentence). Therefore, with the same number of trials in each of the Words and Sentences tasks, perhaps the greater number of words of exposure facilitates greater learning for the sentences than the words. However, this explanation cannot account for the fact that overall performance on the IEEE sentences improves less than the BKBs, and the fact that the IEEE sentences give lower overall recognition scores than the BKBs. Hervais-Adelman et al. (in press) comment on this very issue in their recent study on perceptual adaptation to noise-vocoded words. They observed much slower improvement in noise-vocoded Word recognition than that observed for Sentences in a previous study (Davis et al., 2005), even when the number of words of exposure was taken into account. After around 120 words of exposure to 6-band noise-vocoded sentences (with feedback), performance on sentences was around 60% words correct (Davis et al., 2005), while for Hervais-Adelman et al. the same number of isolated words presented under equivalent feedback conditions gave an average score of only 39% correct. Hervais-Adelman et al. attribute this slower rate of learning to the absence of higher-order contextual information in isolated words. This seems a likely explanation for the current finding, and could also be extended to the difference in the extent of perceptual learning experienced for the IEEE and BKB sentences, where the BKB sentences offer greater semantic predictability. This is a demonstration that both top-down and bottom-up effects are involved in the perception of these distorted stimuli.

There was no significant effect of the between-subjects factor Version ($F(1,26) = 1.61$, $p = 0.216$), suggesting that the ordering of items did not affect the overall level of performance in the experiment. However, Version is involved in a significant 2-way interactions with Session ($F(1,26) = 10.93$, $p = .003$, $\eta^2 = .296$, power = .889) and in a 3-way interaction with Session and

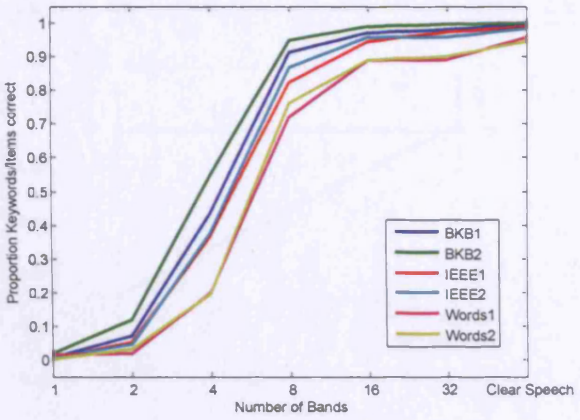


Figure 9.2: Raw means showing the interaction of Task, Session and Level for the open-set recognition tasks.

Level ($F(6, 156) = 5.72$, $p = .000$, $\eta^2 = .180$, power = .997). This suggests that there is some difference between the item sets, such that greater learning occurs for the presentation order BA than for the order AB. It appears that Version A participants score more highly than Version B participants in Session 1, but that there is no difference between the groups by Session 2. The three-way interaction of Version, Session and Level suggests that there is furthermore a difference in the change in shape of the performance function from Session 1 to Session 2 that is dependent on the order of item sets. Table 9.3 shows a plot of the raw mean scores that generated the two-way interaction between Version and Session, from which it is clear that the size of this effect in terms of the mean intelligibility scores is very small, as is the overall effect of Session when averaged across the five tasks.

Closed-Set Recognition Tasks

As for the open-set recognition tasks, the results of the ANOVA showed significant effects of Task ($F(1, 26) = 105.59$, $p = .000$, $\eta^2 = .802$, power = 1.00), Session ($F(1, 26) = 12.42$, $p = .002$, $\eta^2 = .323$, power = .924) and Level (Wilks' Lambda $F(6, 21) = 734.40$, $p = .000$, $\eta^2 = .995$, power = 1.00). Thus the closed-set data also indicate that the tasks differ in difficulty, that performance improves over time, and that there is a strong improvement in performance associated with increasing the spectral detail in the speech stimuli. There was a significant two-way interaction between Task and Level (Wilks' Lambda $F(6, 21) = 6.12$, $p = .001$, $\eta^2 = .636$, power

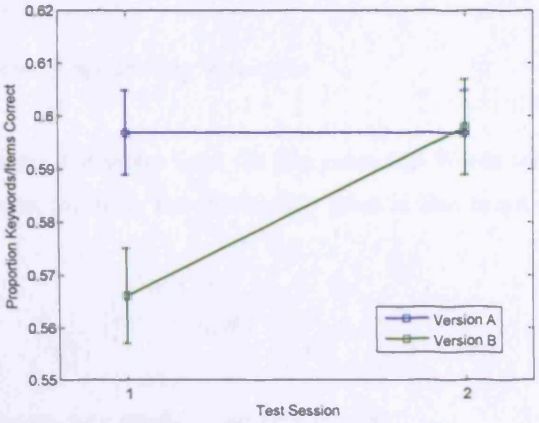


Figure 9.3: Showing the mean recognition scores for the Version*Session interaction. Error bars show ±1 standard error of the mean.

= .988), suggesting, as above, that there is variability in the shape of the recognition performance function across tasks. There were no significant three-way interactions, nor any effects involving the between-subjects variable, Version (F<1).

The above results are reflected in a plot of the raw means for the Task*Level*Session interaction, shown in Figure 9.4.

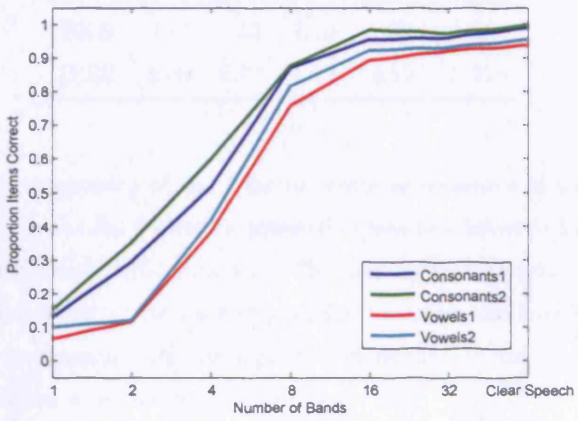


Figure 9.4: Raw means showing the interaction of Task, Session and Level for the closed-set recognition tasks.

Analysis 2

Individual Differences: Calculating k-factors

The untransformed proportion scores from the Sentences and Words tests were used to calculate individual k factor scores, following the relationship given in Boothroyd and Nittrouer (1988):

$$p_s = 1 - (1 - p_w)^k \tag{9.1}$$

p_s = probability of recognising words in sentence context

p_w = probability of recognising words in isolation

The exact value of each k-factor was obtained through least squares curve-fitting of p_s against p_w, and values were extracted separately for use of context in the BKB and IEEE sentences (for each level of distortion, excluding clear speech). The mean value of k was 3.1 for the BKB sentences, and 2.1 for the IEEE sentences indicating that the BKB sentences are easier to understand. The range of scores for each sentence type is shown in Table 9.2.

Table 9.2: Descriptive statistics describing k for the BKB and IEEE sentences.

	M	*SD*	*Min*	*Max*	*IQR*
BKB	3.10	1.53	1.19	7.66	1.61
IEEE	2.06	0.77	1.13	4.10	1.21

In order to test the consistency of the k factor scores as measures of the use of context and top-down processing, a two-tailed Pearson's correlation was run between individual k scores for use of context in the BKB and IEEE sentences. This was highly significant (Pearson's $r = .780$, $p = .000$), indicating that those listeners with higher BKB k scores also have higher IEEE k scores. This indicates that k is measuring the same process in relation to the two sentence sets. This correlation is represented as a scatterplot in Figure 9.5.

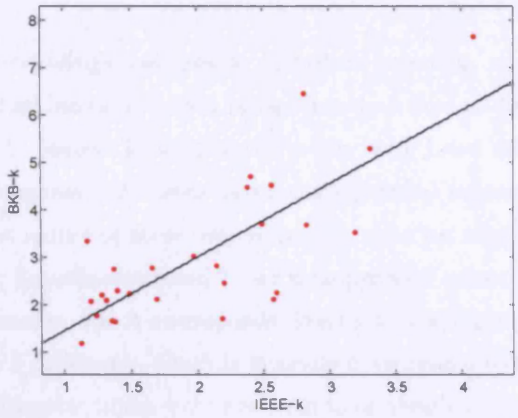


Figure 9.5: Scatterplot of the individual k factor scores for the BKB and IEEE sentence sets, as determined from the word recognition scores on the Words and Sentences tests.

Individual Differences: Correlations and Factor Analysis

An important aspect of the data plotted in Figure 9.2 is that, even with high levels of spectral resolution (32 bands and clear speech), mean performance on Words and Vowels recognition is well below 100%. This is likely to reflect task difficulty, and perhaps the poor ecological validity of an isolated word or segment recognition task - we routinely communicate in phrases and sentences, but rarely in *contextually unsupported* single words or monosyllables. Some of these errors at high band numbers may also reflect a certain level of inattentiveness, causing listeners to respond inaccurately despite adequate encoding of the incoming signal. In the psignifit package (Wichmann & Hill, 2001a, 2001b), the λ (lambda) parameter attempts to account for errors in attention at high signal levels by lowering the upper asymptote of the curve, which should in turn allow for better fits to the data. In turn, the threshold values extracted for 50% recognition in psignifit measure the point at which recognition performance is at half of the curve's maximum value, rather than those values giving exactly 50% - this should also account in some way for the overall differences in task difficulty in clear speech. This approach works on the assumption that errors at high signal levels are purely attentional, and not due to ineffectual signal processing. In a study of individual variability in any task, attentional factors are not of negligible importance, and it should be made clear that the presence of these influences merits investigation in speech perception studies. However, psychometric functions are very vulnerable to deviant points near asymptotic level, and so it was decided that the accuracy of curve-fitting should take greater priority in this

analysis.

The psignifit software package was used to fit logistic functions, as described in the Method of Chapter 8, to individual listeners' speech recognition data from all five tasks. Separate curves were fitted for Session 1, Session 2, and Overall scores, with Level converted to log_{10}(Number of Bands).[4]For the Consonants and Vowels tasks, the γ(gamma) parameter was adjusted to take account of the closed-set nature of these two recognition tasks i.e. that chance level performance would be 1 in 17. As for Experiments 6 and 7, two measures were extracted from each fitted curve. These were the α parameter, which corresponds directly to the log 10(number of bands) of the 50% threshold, and the β parameter, which is inversely proportional to the steepness of the curve slope. Throughout the Chapter, alpha will be referred to as 'threshold' and beta as 'slope' or 'slope parameter'. Table 9.3 shows descriptive statistics for the two measures of interest for the Overall fits across the five tasks, while Table 9.4 gives these for Session 1 and Session 2 curves separately. In both cases, the values are reported in their log 10 form. For the threshold parameters, the mean in terms of the number of bands is given in brackets.

Table 9.3: Descriptive Statistics for logistic performance curves in Experiment 8 (Overall).

		M	*SD*	*Min*	*Max*	IQR
Threshold (α)	BKB	.598	.056	.469	.723	.053
	IEEE	.671	.049	.581	.776	.077
	Words	.784	.069	.701	.988	.089
	Consonants	.511	.062	.379	.652	.085
	Vowels	.670	.071	.556	.865	.079
Slope (β)	BKB	.106	.027	.022	.167	.033
	IEEE	.120	.016	.094	.152	.026
	Words	.130	.045	.031	.234	.056
	Consonants	.205	.046	.139	.360	.051
	Vowels	.168	.047	.095	.266	.057

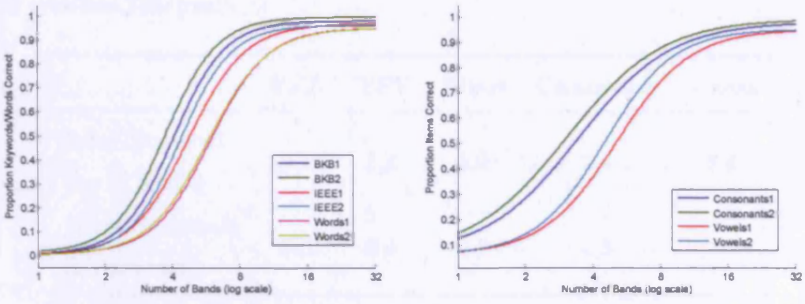
Figure 9.6 shows the logistic functions for pooled group recognition data on the first and second sessions of the open-set (Figure 9.6(a)) and the closed-set (Figure 9.6(b)) recognition tasks. It can be seen from the plots that the BKB sentences, as shown in the k-factor analysis above, are the easier of the two sentence sets, while the Words form the most difficult item set to recognise in the open-set tasks. In the closed-set tasks, the Consonants are easier to recognise than the Vowels. In

[4]Goodness-of-fit statistics are given in Appendix C

Table 9.4: Descriptive Statistics for logistic performance curves in Experiment 8 (by Session).

		Session 1					Session 2				
		M	*SD*	*Min*	*Max*	IQR	*M*	*SD*	*Min*	*Max*	IQR
Threshold (α)	BKB	.624	.055	.477	.743	.086	.577	.061	.463	.706	.071
	IEEE	.686	.077	.580	.860	.133	.657	.067	.542	.793	.106
	Words	.794	.087	.632	.989	.125	.767	.089	.620	.937	.156
	Consonants	.542	.083	.386	.674	.142	.476	.069	.285	.619	.071
	Vowels	.728	.092	.556	.957	.093	.677	.070	.556	.852	.096
Slope (β)	BKB	.099	.047	.020	.224	.050	.102	.033	.015	.167	.026
	IEEE	.122	.020	.080	.158	.033	.109	.023	.055	.140	.036
	Words	.125	.063	.019	.288	.095	.107	.056	.006	.215	.072
	Consonants	.197	.076	.037	.460	.068	.203	.042	.097	.294	.051
	Vowels	.172	.058	.069	.310	.069	.163	.053	.094	.314	.078

all cases, the performance function is shifted leftward in the second Session, indicating that the listeners needed less spectral resolution to achieve threshold performance on their second visit to the lab. In most cases, this leftward shift is also acompanied by a visible increase in the slope of the curve, indicating a sharpening of the 'tuning function' as a result of perceptual adaptation to the noise-vocoded stimuli.



(a) Open Set Recognition - Sentences and Words

(b) Closed Set Recognition - Consonants and Vowels

Figure 9.6: Psychometric functions describing group performance on the speech recognition tasks.

Figure 9.7 shows the logistic functions fitted to individual participants' recognition data for the IEEE sentences. This illustrates the extent of the variability in the position and shape of the

performance curve across listeners.

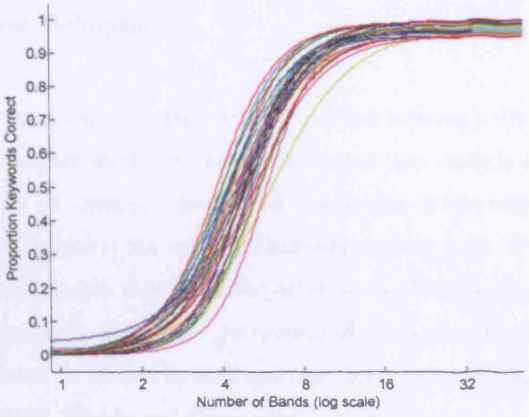


Figure 9.7: Logistic functions for individual listeners in the IEEE sentence recognition task (sessions collapsed)

To illustrate the range of performance, Table 9.5 shows the range of thresholds obtained in the sample population for the 5 different tasks (for Overall performance). These are the Min and Max values given in Table 9.3 in terms of numbers of bands.

Table 9.5: Number of bands for 50% threshold performance, shown for the best and worst listener in each task (sessions collapsed).

	BKB	IEEE	Words	Consonants	Vowels
Best Threshold (no of bands)	2.9	3.8	5.0	2.4	3.6
Worst Threshold (no of bands)	5.3	6.0	9.7	4.5	7.3

The fitting of psychometric functions and the extraction of thresholds allows comparisons to be drawn across individual scores on all five open- and closed-set tasks together. The 50% threshold and slope values for the five tasks were entered into a Pearson's correlation analysis - this was done separately for Session 1, Session 2 and Overall (sessions collapsed) data. In each case, a 1-tailed analysis was chosen as it was hypothesised that a negative relationship between performance on

the different tasks would be unlikely.[5]

Overall data (Sessions Collapsed)

Several significant positive 1-tailed correlations were found between pairs of tasks. Table 9.6 shows the pattern of correlations across the five tasks. It is clear that there is close inter-relationship on performance of the tasks - a listener who is good at one task is also likely to be good at another. However, the set of correlations is not complete between all task pairs. Performance on the Vowels task in particular is more weakly related to the other tasks. This perhaps indicates that different cues are needed to recognise these items - perception of noise-vocoded vowels is likely to be more reliant on effective resolution of the limited spectral detail than is perception of the other four stimulus sets (BKB, IEEE, Words and Consonants), in which more effective use can be made of the well-preserved temporal cues in noise-vocoding.

Table 9.6: Pearson's coefficients for the one-tailed correlations between Overall 50% threshold scores on the five tasks.
$*$ = significant at the $p<.05$ level, $**$ = significant at the $p<.01$ level.

	BKB	IEEE	Words	Consonants	Vowels
BKB	1.00	.486**	.456**	.389*	.195
IEEE		1.00	.509**	.321*	.156
Words			1.00	.541**	.302
Consonants				1.00	.386*
Vowels					1.00

The pattern of correlations within the slope parameter scores presented a contrast to that for the threshold scores, in that there were no significant correlations observed between tasks. However, there were some significant correlations across the two parameter types. The slope parameters from the Words test were involved in significant (1-tailed) positive correlations with the threshold scores

[5]For the individual differences analyses, within-Session correlations come from 14 Set A datapoints and 14 Set B datapoints on each measure. It is acknowledged that the group ANOVA in Analysis 1 gave evidence of a difference in performance between Set A and Set B in Session 1; however, lengthy exploration of the pooled intelligibility of individual items did not yield any helpful clues as to the source of possible item effects. Furthermore, we cannot say with certainty how much of this effect of Version was due to item differences and how much was due to underlying individual differences between participants. The purposes of Analysis 2 is to focus on individual variablity, therefore item effects will not be considered.

from the IEEE sentences (Pearson's $r = .341$, $p = .038$), Words (Pearson's $r = .711$, $p = .000$), Consonants (Pearson's $r = .428$, $p = .012$) and Vowels (Pearson's $r = .327$, $p = .045$). These suggest that a steeper Words recognition curve (i.e. with a smaller slope parameter) corresponds to listeners with lower thresholds on most of the recognition tasks. There were also significant correlations between the Words thresholds and slope values for Consonants (Pearson's $r = -.339$, $p = .039$), the Consonants thresholds and the slope values for Vowels (Pearson's $r = .389$, $p = .020$), and between the Vowels thresholds and Vowels slope values (Pearson's $r = .366$, $p = .028$). Interestingly, the correlation between the Consonants test's slope and Words threshold scores is in a different direction from the others described, suggesting that a higher threshold is associated with a more sharply-tuned curve (i.e. a steeper slope). From inspection of the group curves, it is clear that the consonants task is the easiest at very low numbers of bands, and so for many participants the curve may already be on the steeper portion at only 1 band, while for other tasks there is a lower asymptote reached between 1 and 2 bands. This may contribute to this seemingly anomalous correlation for the Consonants task, where some participants may exhibit a lower asymptote followed by a sharp rise in performance (high threshold, steep curve), while other listeners with high scores at low bands numbers may consequently show a more gradual increase in performance (low threshold, shallower curve).

Subsequent analyses looked at the pattern of results in the separate Testing Sessions. As there were no significant correlations within the slope parameters on the five tasks for the Overall performance curves, it was decided to analyse only the relationships within the threshold scores, and between the thresholds and slope parameters, for the Session 1 and Session 2 data sets.

Session 1 data

Threshold values from Session 1 data underwent the same analyses as the data from the collapsed sessions. One-tailed Pearson's correlations showed a weaker pattern of relationships than exhibited by the collapsed data. However, an interesting pattern of significant and marginally-significant correlations emerged, which suggested close relationship between Sentences and Words on one hand, and Words and Segments (Consonants and Vowels) on the other. The results are shown in Figure 9.8, which for ease of interpretation includes only those correlations that emerged as significant (at $p<.05$) or marginally significant (i.e. at $p<.10$).

The pattern of positive relationships suggests two independent 'levels' of processing - higher-level 'linguistic' processing employed for perception of sentential and lexical information, and a lower-level listening mode used for acoustic-phonetic discrimination. Assuming that these 'levels'

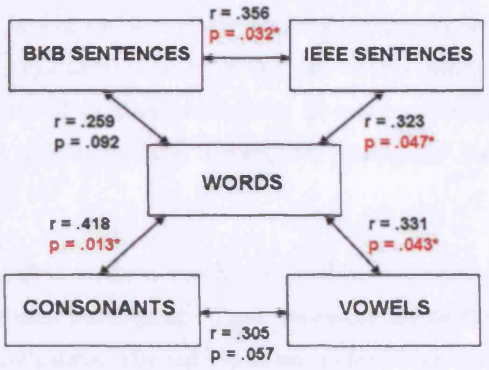


Figure 9.8: One-tailed Pearson's correlations between task thresholds in Session 1 of the experiment.

are orthogonal,[6] a common factor analysis was run on the threshold data, with maximum likelihood extraction and varimax rotation. The rotated factor matrix is shown in Table 9.7, for those factors producing eigenvalues above 1.

Table 9.7: Rotated factor matrix from the Common Factor Analysis on Session 1 threshold data from the five tasks.

	Factor 1	Factor 2
BKB	-.038	**.605**
IEEE	.065	**.593**
Words	**.705**	**.469**
Consonants	**.558**	.055
Vowels	**.562**	-.141

As suggested by the pattern of one-tailed correlations, two components were extracted in the factor analysis. In the rotated matrix, the first component accounted for 22.60% of the variance, while the second component accounted for 19.21%. The pattern of loading, with Sentences and Words on one component and Words, Consonants and Vowels on the other, is concordant with a

[6]Further analyses were run using oblique rotation, which allows for correlation of factors. However, these also suggested that the extracted factors were orthogonal.

'dual processing' mechanism for linguistic and acoustic-phonetic information. The monosyllabic words recognition task is likely to load on both components as a listener could feasibly place equal attentional focus on lexical and acoustic recognition to identify single word tokens presented out of context. The statistical independence of the two factors means that a particular listener may exhibit any combination of weighting strengths. In other words, a listener with a high 'top-down' weighting does not necessarily have a high/low 'bottom-up' weighting when listening to noise-vocoded stimuli.

To further explore the roles of these two factors within individual listeners, regression factor scores were calculated for each participant. These represent the relative loadings on each factor exhibited by each individual's data. The individual scores for Factor 2 were entered into Pearson's correlation analyses with the k-scores calculated in Analysis 2(a) for the BKB and IEEE sentences. This was intended to give a sense of whether Factor 2 is indeed tapping 'top-down' processing, as is measured by the k scores. The analyses gave a marginally significant negative correlation between Factor 2 scores and BKB-k values (Pearson's $r = -.338$, $p = .079$, two-tailed) and a significant negative correlation with the IEEE-k scores (Pearson's $r = -.410$, $p = .030$, two-tailed; see Figure 9.9). The correlation is negative because the relationship between 50% thresholds and performance is negative (a smaller threshold means less spectral resolution is needed to reach threshold score) while that between k scores and the use of context is positive (a higher k score indicates a more effective use of context). Thus listeners with more negative Factor 2 loadings are those who exhibit low thresholds on the Words and Sentences tasks. This finding offers support to the notion that Factor 2 is more concerned with top-down than bottom-up processing. However, k-values and factor scores are not completely equivalent in this case as Factor 2 potentially involves top-down effects of lexical information as well as higher-order linguistic processes, whereas k-factors concern processing above the lexical level. Importantly, there were no significant correlations between the k values and Factor 1 loading scores, for either BKB-k (Pearson's $r = -.076$, $p = .702$, two-tailed) or IEEE-k (Pearson's $r = -.002$, $p = .993$, two-tailed).

The correlations between threshold and slope parameters show significant correlations between the slope of the Words curves and threshold performance on the Words (Pearson's $r = .499$, $p = .003$), Consonants (Pearson's $r = .488$, $p = .004$) and Vowels (Pearson's $r = .444$, $p = .009$) tasks. The correlation between the IEEE thresholds and Words slopes was of only marginal significance (Pearson's $r = .275$, $p = .078$). As for the Overall data, there is a significant correlation between the slopes and thresholds on the Vowels task (Pearson's $r = .453$, $p = .008$). The overall impression from these correlations is that steeper slopes (as indicated by smaller β parameter values) are associated with lower thresholds for 50% recognition scores.

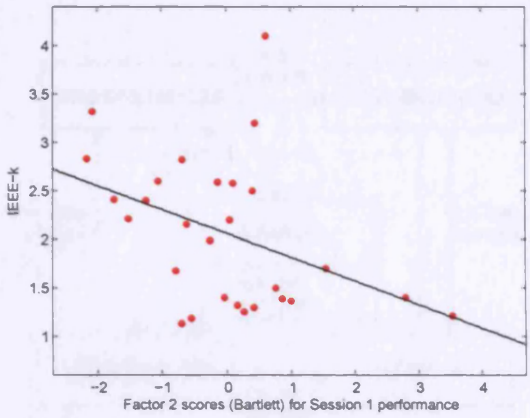


Figure 9.9: Scatterplot of the correlation between individual Factor 2 scores and k values for use of context in the IEEE sentences.

It therefore seems that, on initial exposure to noise-vocoded speech, the participant can adopt two different 'modes' of listening in recognition tasks, reflecting the influences of top-down and bottom-up processing. The more homogeneous pattern of correlations from the Overall data suggest that, over time, this two-level approach changes, perhaps reflecting a lesser need to employ more analytical listening as the mappings between the vocoded signal and phonological representations are learned.

Session 2 data

One-tailed Pearson's correlations on the Session 2 task thresholds did not concord with the Overall or Session 1 results. The results are shown in Figure 9.10. As for Session 1, only significant and marginally significant correlations are shown.

The pattern of correlations no longer fits the processing framework suggested by the Session 1 data, where one could identify a hierarchical arrangement of tasks in terms of increasing linguistic information from Segment to Word to Sentence. It appears that, with the passage of time and associated perceptual learning and improvements in task practice, the style of listening adopted by the participants has undergone some change. Whereas the word level of processing seemed to occupy the central role in processing for the Session 1 data, the sentence level appears to be involved in most significant and marginally significant relationships in Session 2. This may reflect

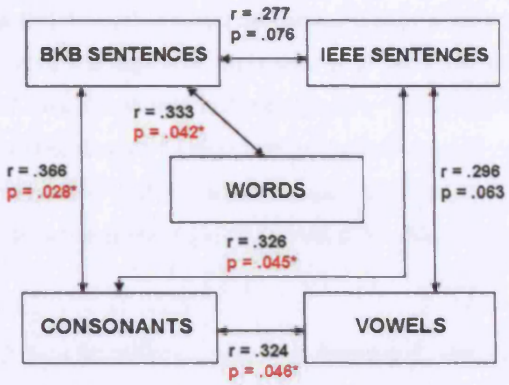


Figure 9.10: One-tailed Pearson's correlations between task thresholds in Session 2 of the experiment.

a general change of attentional focus, where the listener has become familiarised with the acoustic characteristics and acoustic-phonetic mappings of the stimulus, and can therefore begin to listen at a more global level, treating all stimuli in a similar way. For further help in interpreting this pattern of correlations, a Common Factor Analysis (maximum likelihood extraction) was run on the data, with Varimax Rotation. The rotation converged on two components; however, these components appear to be quite different from those identified in the corresponding analysis for Session 1 data. The results of the analysis are shown in Table 9.8.

Table 9.8: Results of the Common Factor Analysis on Session 2 threshold data from the five tasks.

	Factor 1	Factor 2
BKB	**.520**	**.344**
IEEE	**.545**	-.034
Words	.014	**.946**
Consonants	**.642**	.047
Vowels	**.491**	.051

In this analysis, Factor 1 accounts for 24.41% of the variance, where Factor 2 accounts for a further 20.38%. It is difficult to interpret the two components: Factor 1 is loaded on by all tasks but Words, while the second component is loaded upon only by the BKB sentences and the Words,

with the words almost entirely loaded on this task. Factor 1 may be seen to form a general 'speech listening' mode, with a more global attentional focus that can be applied to all the tasks. However, the absence of the Words task from this factor demands further interpretation. From inspection of Figure 9.6, it can be seen that the degree of improvement in performance from Session 1 to Session 2 varies across tasks, and that the Words task exhibits very little improvement. The Words task remains the most difficult stimulus set to recognise in the current experiment, across both session. Therefore, it could be interpreted that, for some reason, identification of monosyllabic words does not improve sufficiently to achieve more global listening on this task (cf Hervais-Adelman et al. (in press)).

Overall, the Session 2 data fit neither a predicted framework, nor an easily interpretable alternative. The results from Session 1 suggest that the initial perception of noise-vocoded speech is dominated by the hierarchical framework displayed in Figure 9.8, but Session 2 data suggest that, with perceptual learning, this pattern is broken down. The pattern of results in Session 2 may reflect differences emerging from the varying *learnabilities* of the five noise-vocoded stimulus sets, where some tasks are more easily learned than others (see Hervais-Adelman et al. (in press)). The changing pattern may also come as a result of the emergence of 'ceiling' or asymptoting effects in the data. Table 9.4 shows that *the range of scores decreases* on all five tasks, with the lowest thresholds in the population (i.e. best performances) decreasing less than the highest thresholds over time. This indicates that the best listeners in Session 1 have less room for improvement than weaker listeners, who can continue to learn - this topic will be discussed further in this section. That some listeners are reaching the upper asymptote of performance, or at least showing a slowing in the rate of improvement, by Session 2 is bound to reduce the strength of correlations between threshold scores across tasks.

Returning to the relationship between the threshold and slope parameters, for Session 2, there were six significant correlations. There were significant within-task, one-tailed correlations between the thresholds and slopes on the Words (Pearson's $r = .492$, $p = .004$), Consonants (Pearson's $r = -.392$, $p = .020$) and Vowels tasks (Pearson's $r = .364$, $p = .028$). There were also significant cross-task correlations between the slope of the BKB performance functions and the thresholds on the IEEE task (Pearson's $r = .450$, $p = .008$), and the Words thresholds and Vowels slopes (Pearson's $r = .397$, $p = .018$). There were also marginally significant correlations between the Words thresholds and Vowels slopes (Pearson's $r = .271$, $p = .081$) and within-task between the BKB thresholds and slopes (Pearson's $r = -.257$, $p = -.093$). In the main, the observed pattern of significant correlations indicates that low thresholds are associated with steep slopes, that is both indicators of good performance are seen together. However, there is not a convincing picture

of the relationship of thresholds to slopes overall, and perhaps some indication that this may be influenced by the structure of the task (see the discussion of the Consonants task above).

Perceptual adaptation and retention of learning - Comparing Sessions 1 and 2

An important element of the design of the current experiment was to investigate whether perceptual adaptation to noise-vocoded speech could be retained over a delay of at least 7 days. This was to address the finding by Altmann and Young (1993) that their listeners retained what they had learnt about time-compressed speech when re-tested approximately 1 year after their first testing session. A criticism which can be levelled at the design of the current experiment is that it is difficult to draw firm conclusions on the source of learning in individual tasks, as these were performed in the same order in both testing sessions. If we demonstrate an improvement in performance between Session 1 and Session 2 on the BKB sentences, which came first in each session, we cannot disentangle whether this improvement is due to retention of adaptation from the later tasks within Session 1 (IEEE, Words, Consonants, Vowels), or due to some consolidation process between the two testing sessions. However, the fixed ordering of tasks in Session 1 was a necessary step to ensure comparability of individual scores in this session.

Two repeated-measures ANOVAs, one for threshold scores and one for slopes, were run to assess the degree of improvement on the five tasks between Sessions 1 and 2. The within-subjects factors were Session and Task, and the between-subjects factor Version. The ANOVA on threshold scores (see Table 9.4 for descriptive statistics) gave a significant effect of Session ($F(1,26) = 35.094$, $p = .000$, $\eta^2 = .574$, power = 1.000), the expected significant effect of Task ($F(4,104) = 117.18$, $p = .000$, $\eta^2 = .818$, power = 1.000), and a non-significant interaction of these two factors ($F<1$), indicating that the degree of improvement was not significantly different across tasks. Importantly, there was also no significant effect of Version ($F<1$), nor did this factor interact with Session ($F(1,26) = 1.33$, $p = .260$, $\eta^2 = .049$, power = .199). The non-significant interaction of Session and Task was perhaps unexpected as the Session 1 and Session 2 group curves for the Words task were barely distinguishable, in contrast with the clear spacing of the two for other tasks. Overall, however, these individual data give a clear picture of improvement in performance between the two sessions, and thus an indication that learning of noise-vocoded speech can be at least retained for 1-2 weeks after an initial period of exposure. Inspection of the descriptive statistics for Slope parameters gives a much more complicated picture, with some of the tasks (BKB sentences and Consonants) showing a decrease in the mean slope (i.e. an increase in the Slope parameter) between Session 1 and Session 2. Furthermore, all five tasks showed considerable variability in slope. The

forced-choice nature of the Consonants and Vowels tasks clearly has an effect on their slopes, producing much more gradual improvements with the introduction of greater spectral detail. For this reason, the dependent variable was split across two separate ANOVAs on slope scores. The first included slope values from the open-set recognition tasks (BKB, IEEE and Words). This found non-significant effects of Session ($F(1,26) = 2.80$, $p = .106$, $\eta^2 = .097$, power = .364) but a significant effect of Task (Wilks' Lambda $F(2,25) = 3.54$, $p = .044$, $\eta^2 = .220$, power = .604). The interaction between Task and Session was non-significant ($F<1$). There was no effect of Version ($F(1,26) = 1.11$, $p = .301$, $\eta^2 = .041$, power = .174), nor a significant interaction between Version and Session ($F<1$). However, there was a significant three-way interaction between Session, Task and Version ($F(2,52) = 3.93$, $p = .026$, $\eta^2 = .131$, power = .683), suggesting some difference in the relative amounts of slope change exhibited by the different tasks across Versions. The corresponding ANOVA on slope parameters from the Consonants and Vowels tasks gave a non-significant effect of Session ($F<1$) and a non-significant interaction of Session and Task (F<1), but a significant effect of Task ($F(1,26) = 6.00$, $p = .017$, $\eta^2 = .188$, power = .655). The effects of Version, and its interaction with Session, were non-significant ($F<1$).

Overall, there is clear evidence of group improvement in thresholds between Session 1 and Session 2, but none for slopes. This clearer pattern of results for the thresholds is in line with the findings of the correlation analyses across tasks as described above, which gave interpretable patterns within the task thresholds but not within the slopes measures. However, the significant differentiation of tasks along the slope parameter is not without interest, as it challenges the finding by Shannon et al. (2004), that a sigmoid curve with fixed slope (thus having threshold as the only freely varying parameter) could fit, with high accuracy, task performance across a range of tasks and listening populations. These tasks bridged sentence, melody and complex music perception, and Shannon and colleagues fitted curves to scores from groups of children, normal-hearing native listeners and non-native listeners, obtaining excellent fits across all groups and tasks.

The relationship between 'baseline' listening ability and the capacity to learn presented an interesting finding in Chapter 8 of the thesis, where there were indications that poorer initial listeners exhibited the largest amounts of learning while better listeners' performance may have reached asymptote. In the current experiment, two-tailed Pearson's correlations between the Session 1 thresholds/slopes and the improvement (in terms of the lowering of thresholds and slope parameters) in these values between Sessions 1 and 2, presented clear positive correlations for all five tasks. These correlations are shown in tabular form below (Table 9.9 and Table 9.10) In other words, those listeners with high thresholds and shallow slopes in Session 1 are those who show most improvement by Session 2.

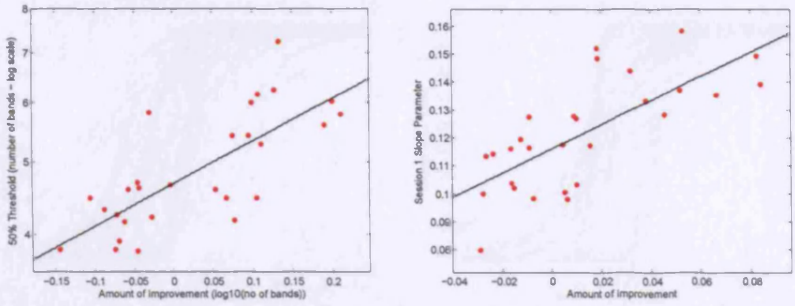
Table 9.9: Pearson's correlations (two-tailed) of Session 1 thresholds with the change in threshold from Session 1 to Session 2.

	Pearson's r	p (1-tailed)
BKB	.485	.009
IEEE	.751	.000
Words	.676	.000
Consonants	.654	.000
Vowels	.686	.000

Table 9.10: Pearson's correlations (two-tailed) of Session 1 slopes with the change in slope from Session 1 to Session 2.

	Pearson's r	p (1-tailed)
BKB	.802	.000
IEEE	.726	.000
Words	.736	.000
Consonants	.843	.000
Vowels	.622	.000

Figure 9.11 shows scatterplots of these correlations for performance scores on the IEEE sentence set. Across a number of the tasks, those who showed the best performance in Session 1 become worse on the same task by Session 2. This may reflect some sort of fluctuation around the upper asymptote of performance in these listeners, or some item differences between Set A and Set B. However, the general pattern of correlations holds when the participants in Version A and Version B are treated separately - those with higher Session 1 thresholds and slope parameters exhibit greater learning. Furthermore, item effects are not involved in the Consonants and Vowels tasks, these tasks having the same items in Session 1 and Session 2, and these tasks show the same strong positive correlation between initial performance and amount learned. It seems, therefore, that the only listeners who exhibit significant perceptual learning are those whose performance was sub-optimal to begin with.



(a) 50% Thresholds

(b) Slope parameters (smaller parameter = steeper slope)

Figure 9.11: Scatterplots of Session 1 scores against amount of improvement by Session 2 for IEEE sentence recognition.

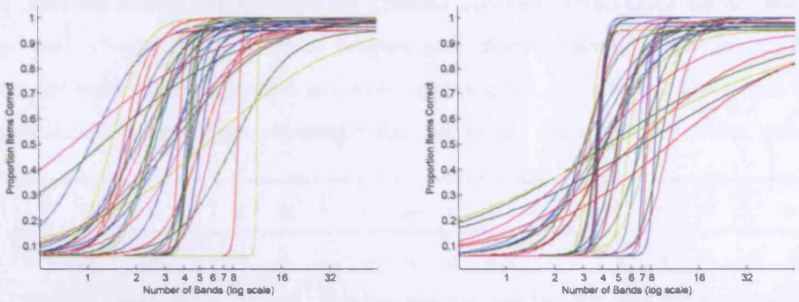
Analysis 3

Information Transfer Analysis of Acoustic-Phonetic Perception

It is highly likely that there is variability in the extent to which the noise-vocoding transformation degrades individual speech sounds, and that this may extend to variability in the shape of the intelligibility functions for each consonant and vowel. In order to explore this in detail, information transfer feature analyses were run on individual listeners' recognition scores in the Consonants and Vowels tasks. However, in order to first gain a sense of how the segments may differ in their individual intelligibility functions, group logistic functions (by session) were fitted using the psignifit package. The group functions for Consonants (Figure 9.12(a)) and Vowels (Figure 9.12(b)) are shown below.

These figures clearly show a widely varying pattern of results, in terms of the position and shape of the individual performance curves, and perhaps give a much more convincing argument against Shannon et al.'s (2004) 'one slope fits all' claim. However, we must acknowledge that these curves come from a closed-set recognition task, which by definition makes the task more like a discrimination than a basic identification. This has implications for the shape of the curve, particularly for the more confusable stimuli.

The forced-choice nature of the Consonant and Vowel tasks means that the data could be arranged into confusion matrices like the one shown in Table 9.1, for use in an Information Transfer



(a) Consonants (b) Vowels

Figure 9.12: Group performance functions for individual speech segments on Sessions 1 and 2 of Experiment 8.

Analysis. These included collapsed frequency data for all distortion levels except clear speech. However, despite the clear instructions of a forced-choice procedure, the fact that the participants' responses were made by typing the answers, rather than by selecting onscreen response options, meant deviance from the response constraints was possible. This could take the form of omitted responses (which often occurred at particularly difficult distortion levels) or responses from outside the closed list. As a consequence, all data sets that included any omissions or deviations from the forced-choice options were not included in the Information Transfer analysis.

Consonants

A total of 14 data sets were entered into the Information Transfer analysis for Overall consonant recognition. The feature matrix used included Voicing, Place and Manner, and is shown below in Table 9.11.

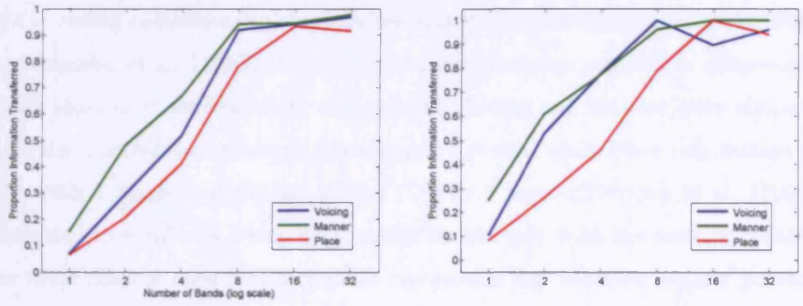
Information transfer feature analyses were run, using an ordinary information transfer analysis within the FIX (Feature Information XFer, University College London, UK) analysis package, for the group confusion data (split across two sessions) and for individual confusion matrices (sessions collapsed). For the particular set of consonants used, there were 0.937 bits of information available for Voicing, 2.542 bits for Place of articulation and 2.095 bits for Manner of articulation.

Table 9.11: Feature matrix for information transfer analysis of the Consonants task. For Voicing, the '+' and '-' signs correspond to present and absent voicing, respectively. For Manner, *plos*=plosive, *fric*=fricative, *aff*=affricate, *app*=approximant, *nas*=nasal. For Place, *bil*=bilabial, *alv*=alveolar, *lad*=labiodental, *paa*=postalveolar, *vel*=velar, *lav*=labialized velar, *pal*=palatal.

	b	d	f	g	ʤ	k	l	m	n	p	s	ʃ	t	v	w	j	z
Voicing	+	+	-	+	+	-	+	+	+	-	-	-	-	+	+	+	+
Manner	plos	plos	fric	plos	aff	plos	app	nas	nas	plos	fric	fric	plos	fric	app	app	fric
Place	bil	alv	lad	paa	vel	alv	bil	alv	bil	alv	bil	paa	alv	lad	lav	pal	alv

Group Data

The amount of information transferred for Voicing, Place and Manner (as a proportion of the amount input for each of the features) was recorded for group confusion matrices constructed at 1, 2, 4, 8, 16 and 32 bands, for Session 1 and Session 2 separately. The data are plotted in Figure 9.13.



(a) Session 1 (b) Session 2

Figure 9.13: Feature information transfer for group data on Consonants task.

The plots give a readily interpretable visual representation of the 'cue-trading' behaviour of the listeners as spectral information is manipulated, and as a result of perceptual learning. In Session 1, Manner of articulation appears to be the best transferred feature at low spectral resolutions (1, 2, and 4 bands), with Voicing next best and Place of articulation worst. At higher resolutions, the three features are equivalent. By Session 2, transfer of information on all three features has improved. However, the improvement is most marked for Voicing, which becomes the best transferred feature. The improvement is least marked for Place of articulation, which remains the

least well transferred feature at low spectral resolution.

What do these two plots suggest about which acoustic features are used in the recognition of noise-vocoded consonants, and how this changes over time? It should be noted that the three phonetic features of Voicing, Place and Manner are not completely independent of each other, and there is likely to be some degree of overlap in the corresponding acoustic features. Dorman et al. (1990) tested identification of consonants by cochlear implant patients. They reasoned that, given the good temporal resolution of cochlear implants, envelope-borne information would be well transferred while the poor resolution offered by a small number of electrodes (6 in the device tested in their study) would limit the transfer of spectral information. Envelope information potentially cues listeners to voicing and manner, while transmission of place information is dependent on temporal fine structure (fluctuation rates from around 600Hz to 10kHz) and spectral shape i.e. the ability to resolve formants in the frequency domain. In cochlear implants and their simulations, frequency resolution can be very poor in the region of formant frequencies, such that both F1 and F2 may be represented by the output of only one channel/electrode. Even if the first two formants can be resolved, the ability to differentiate one speech sound from the other can depend on within-formant transitions in frequency, for example in the discrimination of /b/ and /d/. The ability to make discriminations based on formant-carried frequency information in noise-vocoded speech will depend on the ability of listeners to compare the relative amplitude outputs of the different bands. A study by Shannon et al. (1995) with normal-hearing listeners exposed to noise-vocoded speech demonstrated that, after several hours of exposure, Voicing and Manner were almost completely transferred from spectral resolutions of 2 bands and upwards, while Place information transfer was around 30% with 2 bands and did not exceed 70% by 4 bands. Dorman et al. (1990) point out that the amount of transferred Place information should vary with the amount transferred about Manner, as some manner cues facilitate place recognition e.g. frication manner potentially allows relatively easy discrimination between /s/ and /ʃ/, as /ʃ/ displays a more wideband noise than /s/.

The more marked proportional improvement in reception of Voicing information than for the other two features in the current task can perhaps be explained by considering the acoustic nature of the noise-vocoded stimulus. Voicing can be weakly signalled by slow-moving envelope fluctations - for example, through detection of the longer silent periods in voiceless than voiced plosives, or in the greater amplitude of voiced compared to voiceless obstruents. However, voicing is also signalled by periodicity, that is, temporal regularity in the speech waveform carried by fluctuations primarily between around 50 and 500Hz (Rosen, 1992). This information is reasonably well preserved after the vocoding scheme used in the present experiment, where the amplitude envelope was low-pass

filtered at 400Hz. The improvement shown for Voicing at low band numbers could reflect the participants' increased ability to use the available temporal information in the stimulus to assist performance in the absence of cues to place of articulation that are more dependent on spectral resolution. However, voicing information is also carried by cues to overall spectral balance, as voicing is weighted toward low frequencies. These cues become apparent as soon as at least a second band of information is added to the noise-vocoded stimulus.

Inspection of the group confusion matrices for the six distortion levels (1,2,4,8,16 and 32 bands) shows that, in accordance with Dorman et al. (1990) and Shannon et al. (1995), even with 1-band stimuli listeners are able to quite accurately categorise stimuli according to manner, but not by place. For example, the /f/ consonant is quite well identified at 1 band (5 out of 14 correct in Session 1), while the /s/ and /ʃ/ consonants are misidentified as /f/ (same manner - frication - but different place of articulation). A similar pattern of results is seen amongst the voiced stops (/b/, /d/ and /g/), which are well identified as a group from spectral levels of 2 bands upwards, but are often confused within category. These confusions are inconsistent across band numbers. For example, at 4 bands, /d/ is wrongly identified as /b/ as often as it is correctly identified - however, at 8 and 16 bands /d/ is identified with 100% accuracy, while at 32 bands one quarter of instances of /d/ are misidentified as /g/. The purpose of the current study is not to explain every pair of confusions, but to look at overall patterns of performance, particularly with regard to changes across time. The overall impression given by comparing the Session 1 and Session 2 group confusion matrices at each distortion level is that there are no dramatic changes in the way listeners perform. By Session 2, the listeners show greater numbers of correct identifications, but the most prominent confusions shown in Session 1 remain to nearly the same extent as before. Thus, the added accurate identifications are generally not gained from resolution of the confusions but rather from 'mopping' up of noise that came from more spurious misidentifications in Session 1. So, it seems that over the course of perceptual learning, the listeners are making better use of the available information in the signal rather than starting to listen in a completely new way.

Individual Listeners

The proportion of information transferred (as a proportion of the amount of information input for each feature) for Voicing, Place and Manner (sessions collapsed) was collected for each listener. Table 9.12 shows descriptive statistics associated with information transfer for these three features. Two-tailed Pearson's correlations run between information transfer scores on each of the features gave a marginally significant positive correlation between Voicing and Manner (Pearson's $r = .476$,

$p = .086$).

Table 9.12: Descriptive statistics describing the proportion of information transferred for phonetic features in the Consonants task.

	M	*SD*	*Min*	*Max*	*IQR*
Voicing	.542	.072	.385	.637	.094
Manner	.665	.040	.590	.735	.066
Place	.470	.031	.404	.514	.052

Individual data points (for proportion information transferred) were entered into a step-wise multiple linear regression analysis on the listeners' Proportion Correct scores for the Consonants task. The regression analysis found that the model that accounted for the greatest proportion of the variance included Place of articulation as the single predictor. This model accounted for 61.2% of the variance and was highly significant ($F(1, 12) = 21.47$, $p = .001$). It should be noted that Place of articulation is the feature for which most information was available in the Vowels task (2.542 bits). This suggests that discrimination of place information is central to accurate performance in consonant labelling. Therefore, it is perhaps no surprise that variability in reception of this feature drives performance in the Consonants task.

The results of Analysis 2(b) suggested that there was a connection between consonant perception and perception of monosyllabic words and vowels. In this analysis, data from the 14 participants in the Information Transfer analysis were entered into 2-tailed Pearson's correlations with threshold scores on the five tasks (from Analysis 2(b)). These analyses produced a significant negative correlation (Pearson's $r = -.756$, $p = .002$) between the amount of information transferred about Place of articulation and the 50% threshold scores for the Consonants task. The negative correlation indicates that greater information transferred is related to lower threshold scores (i.e. less spectral detail needed to achieve threshold performance), and re-affirms the importance of place of articulation information shown in the results of the linear regression. There were also significant negative correlations between the amount of information about Manner of articulation and the threshold scores on the Consonants task (Pearson's $r = -.596$, $p = .024$) and on the IEEE sentences task (Pearson's $r = -.597$, $p = .024$), and between the amount of Voicing information and thresholds on the Vowels task (Pearson's $r = -.634$, $p = .015$). There was also a marginally signficant correlation between the amount of information transferred on Manner and the thresholds on the Vowels task (Pearson's $r = -.528$, $p = .052$). However, it may not be the case that these five correlations are special. Most of correlations between the three features and

the five tasks were negative in direction, and it may simply be the relatively small number of data points ($N = 14$) that prevented the emergence of more significant correlations.

In order to assess the effects of perceptual learning on information transfer, and the relative importance of the different feature types for consonant recognition, two further step-wise linear regressions were carried out, this time using information transfer data from Session 1 and Session 2 separately to predict consonant recognition scores in each session. The descriptive statistics for feature-based information transfer in each of the two sessions are shown in Table 9.13. In Session 1, there were significant correlations between information transfer for Voicing and Manner (Pearson's $r = .608$, $p = .021$) and for Place and Manner (Pearson's $r = .653$, $p = .011$). By Session 2, only one correlation, between Voicing and Manner, reached marginal significance (Pearson's $r = .480$, $p = .082$).

Table 9.13: Descriptive statistics describing the proportion of information transferred for phonetic features in separate sessions of the Consonants task.

	Session 1					Session 2				
	M	*SD*	*Min*	*Max*	*IQR*	*M*	*SD*	*Min*	*Max*	*IQR*
Voicing	.498	.089	.333	.660	.141	.604	.111	.430	.840	.126
Manner	.646	.045	.556	.729	.064	.723	.064	.642	.839	.094
Place	.483	.042	.428	.569	.050	.512	.039	.432	.579	.048

For Session 1, the emergent significant model ($F(1, 12) = 11.69$, $p = .005$) featured Place of articulation as the sole predictor and accounted for 45.1% of the variance in the data. By Session 2, a model containing Place as the only predictor (F(1,12) = 17.78, $p = .001$) accounted for 56.3% of the variance, while a second model with Place and Manner as predictors ($F(2, 11) = 25.4$, $p = .000$) accounted for 79% of the variance. This pattern of results strongly indicates that variability in Place of articulation information transfer is key in predicting basic noise-vocoded Consonant recognition, both when the stimuli are initially encountered and after perceptual learning. As for the group data, we gain the sense that listeners are not drastically changing what they listen out for in noise-vocoded speech, but that they are making better use of the existing information as they gain more experience with the stimuli. An overall weakening of the inter-correlations of the information transfer scores from the three Features by Session 2 reflects the listeners' abilities to make more independent use of these cues after achieving sufficient improvement in Place resolution.

Vowels

A total of 14 data sets were entered into the Information Transfer analysis for Overall vowel recognition. The feature matrix used included vowel Height, Backness, Roundedness, Length, and whether the vowel is a Monophthong or Diphthong. The matrix is shown in Table 9.14.

Table 9.14: Feature matrix for information transfer analysis of the Vowels task. For Height, *o*=open, *no*=near-open, *om*=open-mid, *m*=mid, *cm*=close-mid, *nc*=near-close, *c*=close. For Backness, *b*=back, *nb*=near-back, *c*=central, *nf*=near-front, *f*=front. For Roundedness, *y*=rounded and *n*=unrounded. For Length, *s*=short and *l*=long. For Diphthong, *y*=diphthong and *n*=monophthong. Dashes indicate the separation of the diphthong descriptions into monophthongal elements, in temporal order.

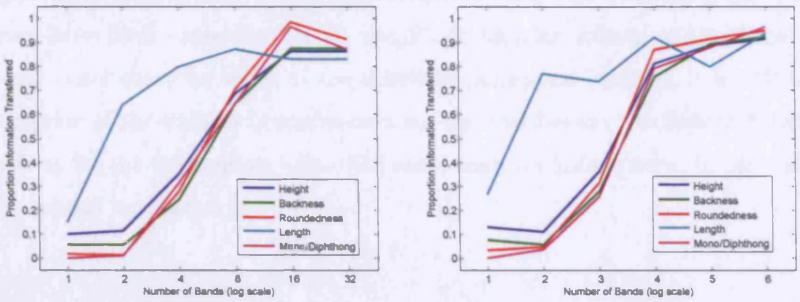
	æ	eɪ	ɑ:	ɛ:	i:	iə	e	ɪ	aɪ	ɜ	ɒ	əʊ	u:	ɔ:	aʊ	ɔɪ	ʌ
Height	no	cm-nc	o	om	c	nc-m	cm	nc	o-nc	om	o	m-nc	c	om	o-nc	om-nc	om
Backness	f	f-nf	b	f	f	nf-c	f	nf	f-nf	c	b	c-nb	b	b	f-nb	b-nf	b
Roundedness	n	n	n	n	n	n	n	n	n	n	ny	n	y	ny	ny	yn	n
Length	s	l	l	l	l	l	s	s	l	l	s	l	l	l	l	l	s
Diphthong?	n	y	n	n	n	y	n	n	y	n	n	y	n	n	y	y	n

Information transfer feature analyses were run, using the FIX analysis package, for the group data (split across two sessions) and for individual listeners (sessions collapsed and for separate sessions). In all analyses, there were 3.264 bits of information available for vowel Height, 2.816 bits for Backness, 1.452 bits for Roundedness, 0.874 bits for Length and 0.937 bits for Mono- versus Diphthong status.

Group Data

The amount of information transferred for Height, Backness, Roundedness, Length and Mono/Diphthong status (as a proportion of the amount input for each of these features) was recorded for group confusion matrices constructed at 1, 2, 4, 8, 16 and 32 bands, for Session 1 and Session 2 separately. The data are plotted in Figure 9.14.

These plots show that vowel Length information is the best transferred (as a proportion of the information input about this feature) of the five features at low spectral resolutions (1,2 and 4 bands), with the other features more closely bunched. At greater spectral resolutions (16 and 32 bands), however, the amount of Length information transferred levels off while the other features



(a) Session 1 (b) Session 2

Figure 9.14: Feature information transfer for group data on Vowels task.

reach 100%. The most sizeable improvements between Session 1 and Session 2 happen at 8 bands for Height, Backness and Roundedness. This stands in contrast to the Consonants task, where the greatest improvements occurred at 2 and 4 bands for all features - however, this is a reflection of the fact that the Consonants task was performed better overall and so feature information transmission was near-perfect at 8 bands in that task. The greatest improvements for the Length feature occur at 1 and 2 bands in the Vowels task, and this is, proportionally, the most improved feature between Session 1 and 2.

As for the consonants, it is clear that the transmission of frequency-based information, as is required for identification of Height, Backness, Roundedness and Diphthongs, is a limiting factor in recognition of noise-vocoded vowels, and that these four features are closely related in terms of recognition. Inspection of the group confusion matrices adds a little more detail to the picture. From the lowest level of spectral resolution, there are very few errors of vowel length assignment but considerable errors on the other features. In 2-band noise-vocoded vowels, the introduction of a second band of information enables comparisons of the energy outputs of these two bands to facilitate resolution of spectral shape. In the current experiment, this resulted in an identification bias toward the 'bard' vowel, which was the chosen response in 183 of the 476 responses at 2 bands. The introduction of further bands improved performance on most vowels, in accordance with the prediction that greater frequency resolution should facilitate the discrimination of vowels through resolution of formants. However, two vowel pairs remained poorly identified - these were the 'bared' and 'bird' vowels, and the 'bed' and 'bid' vowels (n.b. these confusions were directionally biased, with 'bird' and more likely to involved in these misidentifications than the other two vowels). The difficulties here may have emerged from these items' status as central vowels, with neither characteristically high nor low F1 and F2 frequencies. Thus, these two pairs remain difficult to

distinguish at levels where other vowels are recognised with high accuracy. This indicates that listeners may have been depending on F1 and F2 to identify vowels, and perhaps making less use of F3 and other cues. In terms of the effects of perceptual learning, it is difficult to give a simple description of the changes in confusion matrices from Session 1 to Session 2, but the overall impression is as for the Consonants - that the same main confusions occur in both sessions while the noise is reduced by Session 2.

Individual Data

Individual data were collected for the amount of information transferred for each feature - descriptive statistics are shown in Table 9.15 for the collapsed session data. Over both sessions, two-tailed correlations between information transfer scores on the features were positive and significant between Height, Backness, Roundedness and Diphthong status (Pearson's correlations between 0.667 and 0.971, all significant at $p<.01$).

Table 9.15: Descriptive statistics describing the proportion of information transferred for phonetic features in the Vowels task.

	M	SD	Min	Max	IQR
Height	.432	.044	.356	.523	.061
Backness	.377	.044	.310	.466	.061
Roundedness	.384	.043	.296	.446	.044
Length	.662	.173	.253	.857	.291
Diphthong	.392	.052	.197	.405	.055

Individual scores were entered into a step-wise multiple linear regression, with the dependent variable as Proportion Correct scores on the Vowels task. The feature which came out as the single strongest predictor of performance was the Height feature. This accounted for 86.4% of the variance in performance on the Vowels tasks, and produced a significant regression model ($F(1, 12) = 83.51$, $p = .000$).

The output of the Information Transfer analysis was entered in a two-tailed Pearson's correlation analysis with participants' threshold scores from the five tasks. Unsurprisingly, there were significant negative correlations between the amount of information transfer for vowel Height, Backness, Roundedness and Diphthong status and the thresholds on the Vowel task (Height: Pear-

son's $r = -.897$, $p = .000$; Backness: Pearson's $r = -.870$, $p = .000$; Roundedness: Pearson's $r = -.542$, $p = .045$; Diphthong: $r = -.654$, $p = .011$). There were also significant negative correlations between the amount of information transferred regarding vowel Length and threshold scores on the BKB sentences ($r = -.564$, $p = .036$), and the Words task ($r = -.663 p = .010$). Of marginal significance was the relationship between the IEEE sentences task and information transferred about vowel length ($r = -.512$, $p = .061$).

The significant relationships with Length support the proposal that timing and rhythmic information are of importance in perception of noise-vocoded speech, which formed the theoretical motivation for Chapters 3 to 6 of the current thesis. The between-subjects variability in the reception of Length information (displayed in Tables 9.15 and 9.16) shows that some listeners are making poor use of this information despite the fact that it is readily transmitted in noise-vocoding. This finding is similar to that obtained by Iverson, Smith, and Evans (2007), who measured information transfer for vowel length in cochlear implant users and normal-hearing listeners listening to a cochlear implant simulation. Both listening groups in the Iverson et al. study showed sub-optimal information transfer. The authors propose that, given the excellent preservation of durational information in noise-vocoding, participants should be able to show 100% information transfer for length, even at low spectral resolutions. Therefore, while the evidence suggests that timing and rhythm may be important for successful perception of some forms of noise-vocoded speech, listeners may require more guidance and training in order to make better use of durational cues.

The emergence of vowel Height as the sole predictor of Vowels task thresholds indicates an important role for F1 frequency in this recognition task. However, observation of the plots in Figure 9.14 indicates that there is little to separate Height from the Backness, Roundedness and Mono/Diphthong features. Furthermore, the scores for proportion of transferred information on these four features are highly intercorrelated across individuals, while the Length feature is not significantly related to any of the others. The prominence of Height as a predictor may also be related to the fact that the input feature matrix provides most information on Height (3.264 bits), and so the extraction of this information, and variability in this, might be expected to be more influential than for the other features.

As for the Consonants analysis, step-wise multiple linear regressions were carried out on Vowel recognition scores using information transfer data from Session 1 and Session 2 separately. The descriptive statistics for information transfer for each feature in the two sessions are shown in Table 9.16. Bivariate correlations were measured between the individual information transfer scores in each session. In Session 1, significant correlations were as observed for the Overall

(sessions collapsed) data, with Pearson's correlation coefficients between 0.615 and 0.954. However, the correlations between Roundedness and Height, and Roundedness and Backness, were only significant at $p<.05$. In Session 2, there were two significant correlations, between Backness and Height (Pearson's $r = .841$, $p = .000$) and between Roundedness and Length (Pearson's $r = -.621$, $p = .018$). There was also a marginally significant positive correlation of Mono/Diphthong status and Roundedness (Pearson's $r = .474$, $p = .087$). The negative correlation between Length and Roundedness is of interest as it suggests that those listeners making more use of Length information are those making less use of Roundedness information in Session 2. As for Voicing information in the Consonants task, it is Length information which shows the greatest improvement in the mean information transfer score from Session 1 to Session 2. This perhaps indicates, as in the Consonants task, that listeners are learning by making better use of the available temporal information, which then allows for improved performance on features that depend more on spectral resolving skills. The negative correlation of Roundedness and Length might suggest that some listeners attend to formant resolution, namely resolution of F3, at a cost to the reception of temporal information.

Table 9.16: Descriptive statistics describing the proportion of information transferred for phonetic features in separate sessions of the Vowels task.

	Session 1					Session 2				
	M	*SD*	*Min*	*Max*	*IQR*	*M*	*SD*	*Min*	*Max*	*IQR*
Height	.456	.048	.357	.560	.056	.493	.039	.435	.543	.081
Backness	.392	.050	.298	.498	.061	.424	.040	.360	.488	.069
Roundedness	.379	.047	.289	.450	.071	.422	.063	.314	.554	.091
Length	.632	.183	.232	.879	.286	.707	.181	.280	.929	.254
Diphthong	.318	.065	.144	.407	.095	.345	.054	.240	.440	.066

For Session 1, the emergent significant model ($F(1,12) = 41.25$, $p = .000$) featured vowel Backness as the single predictor and accounted for 75.6% of the variance in the data. By Session 2, several possible models emerged. These are shown in Table 9.17 below. The overall indication is that vowel Height is the most important of the features, as indicated in the analysis for the collapsed sessions, but that by Session 2 there is weaker correlation of the variability expressed for the five features, allowing more independent contributions to regression models. The basis of this change could take the form of more sophisticated use of cues to resolve F1 and F2, thus allowing use of other cues such as F3 frequency (which conveys information about Roundedness) and overall spectral energy peaks to give improved performance later in Session 1 and in Session 2.

Table 9.17: Significant regression models for Session 2 Information Transfer data.

Model	Predictors	F	Sig	Adj. R^2
1	Height	42.56	.000	.762
2	Height, Length	32.09	.000	.827
3	Height, Length, Diphthong	34.32	.000	.885
4	Height, Length, Diphthong, Backness	53.68	.000	.942

9.2.3 Discussion

The analyses described above are detailed and several; thus the Discussion will also include summaries of the overall findings from the different sections of the Results section. It is hoped that this will demonstrate the importance, and the particular difficulties, of employing multiple analysis techniques.

Analysis 1 - Group Data

A repeated-measures ANOVA on the proportion recognition scores for the Sentences and Words tasks found significant effects of all three within-subjects factors of Task, Level and Session, plus significant interactions of Task*Session, Task*Level and Session*Level. These findings suggest a rich pattern of behaviours, where three tasks of varying difficulty (increasing in the order BKB, IEEE, Words) exhibit differently-shaped performance functions of recognition against spectral resolution. Overall performance improves on all tasks, but the Task*Session interaction suggests that this improvement occurs at different rates for the different tasks. Finally, the Sesssion*Level interaction suggests a change in the shape of task performance functions over time, but the lack of a three-way interaction with Task indicates that the function slopes maintain the same inter-relationship across tasks.

For the closed-set recognition data (Consonants and Vowels tasks), there were also significant effects of Task, Session and Level. However, in this case the only significant interaction was between Task and Level. This suggests that while the two tasks are of differing difficulty, there is no change in the relative difficulty levels over time i.e. performance improves on the two tasks at a similar rate. The data also indicate that there is no significant change in performance function shape over time on these tasks.

There are two important findings from this basic analysis that merit discussion at this point. First, it is an important finding that, with a mean intervening period of 10 days, subjects' overall recognition performance improved significantly. Inspection of the mean recognition scores for the significant interaction of Task and Session indicates that the extent of this mean improvement in proportion scores was small overall, and unlikely to be significant for all tasks - scores improved by less than 1% overall for the IEEE sentences and the Words. However, mean performance did not decrease on any of the tasks, and the Task*Session interaction actually provided reassurance that the improvements in performance are not simply due to generalised task practice, as this should affect all tasks equally. Furthermore, the true extent of improvements is likely to be limited by the presence of many data points with proportion scores of 1 (for items with higher spectral resolutions) 'smearing out' the changes at lower numbers of bands. Therefore, this initial finding paved the way for more focussed exploration of threshold changes in Analysis 2.

The second important finding from the group analysis was the indication of a difference in performance function *shape* across the tasks. This speaks to the claim by Shannon et al. (2004) that, regardless of the test materials or the listening situation, performance functions of recognition against spectral resolution can be fitted with the same slope. This topic will be addressed further in the discussion of Analysis 2.

Analysis 2 - Individual Differences

In this analysis, attention was turned to characterizing individual differences in performance as a function of spectral resolution, task and time. Two different curve-fitting analyses were performed. The first approach used curve-fitting to relate Words and Sentences (for BKB and IEEE sets separately) recognition scores in order to extract individual measures of the use of top-down, contextual information in speech perception, known as the k-factor. The second approach (Analysis 2(b) and 2(c)) used the fitting of logistic speech recognition functions to recognition data from individual tasks, and the extraction of performance measures in the form of parameters describing 50% thresholds and slopes.

Thresholds - interactions and perceptual learning

The psignifit package was used to fit individual performance functions to recognition data, from which individual indices of performance were extracted in the form of 50% recognition thresholds

and slope parameters. The extraction of the 50% threshold values allowed greater comparability of performance across all five tasks than could be achieved in Analysis 1. For the Overall (sessions collapsed) data set, there were strong correlations between the threshold scores on all the different tasks. However, performance on the Vowels appeared to be slightly set apart from the other four tasks. Given the known physical effects of noise-vocoding on speech stimuli - degradation of spectral detail in the presence of a relatively well-preserved envelope - it is perhaps unsurprising that variability on a task which relies heavily on the resolution of formant frequencies (as demonstrated by the results of the Information Transfer analysis) may not covary with that on tasks in which the listener can make more effective use of envelope cues (i.e. the Words, Consonants and Sentences). Previous studies by van Ooijen (1996) and Cutler, Sebastian-Galles, Soler-Vilageliu, and van Ooijen (2000) showed that when listeners are presented with a nonword such as *eltimate* and asked to replace one phoneme to create a new word, the listeners are more likely to change the vowel (e.g. to produce *'ultimate'*) than the consonant (e.g. to produce *'estimate'*). This finding holds across several languages. A possible explanation is that vowel production is naturally more variable than consonant production, giving listeners greater expectations of variability and misperceptions with these speech sounds and hence making them more likely to replace vowels to correct mishearings. The results of the current analysis may reflect this increased perceptual 'vulnerability' of vowel sound identification in speech perception.

An interesting pattern of results emerged when 50% thresholds were analysed separately for Session 1 and Session 2 data. Following on from the significant effect of Session in Analysis 1, an ANOVA on Session 1 and 2 threshold values showed statistically equivalent (significant) decreases in 50% threshold between Session 1 and 2 across all tasks. Intuitively, a decrease in threshold indicates that, with perceptual learning, less spectral resolution is required to reach 50% recognition performance. This supports the finding of overall improvement across Session in Analysis 1, but now without the Task*Session interaction. It is difficult to give an explanation for why this interaction did not present itself with the individual data, as the group performance functions for Session 1 and 2 suggested that improvement on the Words was much less marked than that for the BKB and even the IEEE sentences. However, it must be noted that the curve-fitting analysis differed in a number of ways from Analysis 1, for example excluding the data for undistorted speech, taking account of attentional slips through employment of the *lambda* parameter, and generally offering a more focussed account of the logistic relationship of spectral resolution to recognition performance. Beyond all these potential factors, it may simply have been variability in the individual learning performances at lower band numbers that prevented the Task*Session interaction from reaching significance. This contradiction between Analysis 1 and 2 highlights the complicated task of interpreting multiple analyses; on the other hand, the finding

in both Analysis 1 and 2 of a general improvement in performance over time is an important affirmation of the presence of perceptual learning in this study.

Further exploration of individual threshold scores enabled some characterisation of the nature of perceptual learning in this study. While the Overall data had shown rich inter-correlation of tasks, the Session 1 data indicated 'listening modes' at play - one loading on Sentences and Words, and the other loading on Words and Segments (Consonants and Vowels). This was supported by a common factors analysis. A significant correlation between IEEE-k scores and factor scores for the Sentences and Words factor further supported the hypothesis that this factor represented the use of top-down processing in noise-vocoded speech recognition. However, the involvement of both Words and Sentences recognition scores on this factor indicated that 'top-down' may mean more than the use of context in this case, for example the use of top-down influences of lexical expectancies on segment recognition. In contrast to the Session 1 results, there was no ready explanation for the pattern of correlations seen for Session 2 data. However, the factor analysis showed a separation of the Words task from performance on the other stimulus types. It seems that perceptual learning of the sound-to-representation mappings in noise-vocoded speech allowed a fusion of the two processing modes into one generalised listening approach to vocoded stimuli. However, some combination of the limited improvement over time, and the marked difficulty of the Words compared with the other tasks, meant that performance of this task may have called upon different strategies during Session 2.

'Asymptote' effects and item effects - a discussion of the experimental design

As mentioned in the Methods section of this Chapter, the facilitation of both group and individual analyses required a certain balance in the experimental design. To allow assessment of the relationship of individual performances across the different tasks, task order was the same in both testing sessions. However, to allow the interpretation of effects of the main manipulations of Task, Session and Level at a group level, counterbalancing and randomisation of items were employed. The counterbalancing of items should not have presented a challenge to individual differences such as the correlations by Session in Analysis 2 if the two item sets were found to be equivalent in difficulty. However, a Version by Session interaction in Analysis 1 indicated that the Words and Sentences in Set A may be easier than those in Set B. It was decided, in the interests of statistical power, to continue with Analysis 2 without splitting the participant group by Version. The presence of item effects in noise-vocoded speech perception experiments is perhaps unavoidable.

Earlier experiments of the thesis using items from the LSCP corpus clearly demonstrated that some items were routinely better recognised than others in 5-band noise-vocoding. Hervais-Adelman et al. (in press) note that, despite matching for factors such as number of phonemes, number of syllables, wordform frequency and imageability, there is considerable variability in the intelligibility scores for their groups of 6-band noise-vocoded words. This variability is likely to have emerged through the combined effects of the controlled variables and the noise-vocoding transformation of the acoustic signal, which is not uniformly disruptive to the intelligibility of speech sounds (see the results of Analysis 3 in this chapter). Hervais-Adelman et al. (in press) were able to overcome the potential problems of item variability by using the item recognition scores from one experiment as a parameter for item matching in the following experiment. However, the sheer number of items and distortion levels involved in the current study, combined with the evident variability across listeners with vocoded items, would have made such a 'pre-calibration' pilot study prohibitively labour-intensive. The randomization process involved in the current study also prevented any interpretable post-hoc screening for 'troublesome' items, as items (i.e. the linguistic content) did not occur with equal frequency across distortion levels for the group data. However, with noise-vocoding being used increasingly frequently as a speech perception tool in experimental psychology and phonetics, the creation of intelligibility-calibrated noise-vocoded item sets for use in such experiments would be a worthy endeavour. In respect of the potential item set imbalance in the current experiment, it is acknowledged that item effects cannot be entirely ruled out in the study. While the design compromises were necessary for the overall aims of the experiment, some of the findings put forward here (such as the correlations by Task and Session) should be re-tested in a design devoted entirely to individual differences analyses and employing a fixed presentation order on all counts (task, item, distortion level) to each participant.

For both threshold and slope data, the participants exhibiting better Session 1 performances were those who showed the least improvement by Session 2. This may reflect either an 'asymptoting' effect or a slowing of learning in the better listeners. The finding held across both Versions of the experiment (and hence both orders of presentation). This replicates a similar result in Chapter 8 and once again provides corroboration with previous findings of individual differences in perceptual learning of speech (Stacey & Summerfield, 2007) and auditory tasks (Amitay et al., 2005). In future investigations of the relationship between baseline speech recognition and perceptual adaptation capabilities, it would be desirable to slow down the learning process by making the stimulus more difficult (e.g. introducing a frequency shift; Rosen et al. (1999)) and look for replication of the pattern of variability seen in Experiments 6-8. As with the potential effects of items, this effect may also be partly attributable to the requirements of the current experimental design, as the fitting of logistic curves demanded sufficient numbers of stimuli at each distortion

level but also may have provided sufficient exposure for performance to have reached asymptote in some listeners. The identification of this 'ceiling' effect also informs the interpretation of the Session 2 correlations, where the reduced range of individual scores introduced by ceiling effects in the better listeners restricts the potential for observation of significant relationships across tasks.

Different slopes for different folks? Addressing the claims of Shannon et al. (2004)

An important strand of this analysis was an exploration of the role of slope in task performance functions. With reference to Shannon et al. (2004), this experiment did not directly compare the goodness-of-fit achieved with fixed vs. variable slope. However, in finding a significant difference in slope based on Task within the open- and closed-set subgroups, the current data set challenges the inference from Shannon et al. (2004) that slope steepness may not be informative.

The current findings present a less clear picture of the role of curve slopes compared with that for thresholds. For individual performances describing the Overall, Session 1 and Session 2 data sets, there were indications, via significant correlations, that low threshold scores (i.e. better performance scores) were associated with steep curve slopes. The repeated finding that good Session 1 performers exhibited lesser gains by Session 2 while weaker Session 1 listeners showed larger improvements, indicated that perceptual learning also leads to a sharpening of the performance function. An initial interpretation would be to suggest that those listeners who can make better use of the incoming acoustic information will make better 'gains' from each addition of spectral detail (corresponding to an increase in the number of bands), thus producing a sharper curve. However, there were no significant correlations between individual slope values across the tasks for the overall data set, nor was there was there a significant group improvement in slope from Session 1 to Session 2. This uncertain role of slope for individual listeners may reflect difficulties in fitting procedures, where the slope is more sensitive to deviant data points than the 50% threshold - it may be that more trials would have been needed to obtain satisfactory estimates. Inspection of the group functions for individual items in the consonants and vowels tasks also indicates that slope might also depend on task structure (open-set versus closed-set recognition).

Despite the apparent lack of systematicity in the role of slope and its relation to thresholds and task, it is clear from the available data that curve slope is not a negligible parameter in describing overall performance. However, for the current data set, it appears that threshold values are the more robust measures of performance.

Analysis 3 - Bottom-up processing in noise-vocoded speech perception

The outcomes of information transfer analyses on the Consonants and Vowels recognition data are generally in agreement with the findings of several previous studies using noise-vocoded speech (Dorman et al., 1990; Dorman, Loizou, & Rainey, 1997; Dorman & Loizou, 1998; Iverson et al., 2007; Shannon et al., 1995). However, the current study enabled the assessment of two extra dimensions: the effect of perceptual learning on the extraction of feature information, and the relationship of feature processing to performance on the five speech recognition tasks.

The important overall finding is that good performance on segment recognition tends to depend on the same cues in noise-vocoded speech that would be attended in identification of segments in clear speech. Therefore, despite the good preservation of Voicing information at low spectral resolutions in the Consonants task, it is Place of articulation information - which is most likely to be carried by formant transitions in the lower frequency regions of the spectrum - that best predicts performance on the task. Similarly, the potentially complete transmission of vowel Length than more formant-dependent features such as Height and Backness does not prevent the latter features from driving recognition of noise-vocoded Vowels. However, it is in the relationship between feature extraction and performance on the other speech recognition tasks that a particular role for temporal information emerges. The finding of a significant relationship between good transfer of vowel Length information and better threshold scores on the Words and Sentences tasks offers support to the claims made throughout this thesis of the potential importance of timing information in recognition of noise-vocoded stimuli. Furthermore, the effects of perceptual learning reflected marked improvements in transmission of those features that are well-conveyed by the preserved temporal information in noise-vocoding - namely, Voicing and Length. However, the variability in reception of these well-preserved temporal characteristics of speech in noise-vocoding suggests that listeners may need direction and training in order to make the most of the useable temporal information in the stimulus.

9.3 Summary

The present experiment formed a thorough investigation of speech recognition and perceptual learning in the context of five speech perception tasks of varied linguistic content. There is convincing evidence of perceptual learning for Sentences, Words, Consonants and Vowels that survives across a delay of one week or more, in the absence of any specific training procedures. There appear

to be two independently-varying 'levels' of processing at work in the initial perception of these difficult stimuli - a 'top-down' listening mode making use of contextual and lexical information in the speech, and a 'bottom-up' mode in which the focus of attention is on lower-level acoustic-phonetic discriminations and sound-to-representation mappings. Over time, performance on all tasks may be achieved by employment of more general mechanisms once sound-to-representation mappings have been learnt. The evidence suggests that this generalised listening may result from more successful use of existing temporal information in noise-vocoded stimuli. However, some listeners fail to take advantage of this information and continue to attend to the same cues as used in perception of undistorted speech - training regimes involving directed attention to durational features of noise-vocoded speech may help to improve performance.

Chapter 10

General Discussion

10.1 Factors affecting the perception of noise-vocoded speech

This thesis has presented a two-strand approach to investigating the recognition of, and perceptual adaptation to, noise-vocoded speech. The approach was inspired by Shannon et al. (2004), who addressed roles for both stimulus-based and listener-based sources of variability in the difficulty of a speech perception task. This section summarises the outcomes relating to each of the two research questions, via a summary of the results of the experimental work.

10.1.1 Stimulus properties - the role of rhythm

This research question emerged directly from a personal observation of rhythmic 'unnaturalness' in the nonword sentences used in Davis et al. (2005), which were the only sentences not to provide evidence of significant training for perceptual adaptation to noise-vocoded speech. In Study 1 of the current thesis, an analysis of the rhythmic properties of the training materials used in Davis et al. was carried out following recent methods that use durational analysis of consonantal and vocalic intervals within the speech (Dellwo et al., 2004; Dellwo, 2007; Grabe & Low, 2002; Ramus et al., 1999). It was found that rate-normalized PVI (pairwise variability index) measures of consonantal and vocalic durational variability offered a clear separation of the Nonword sentences from the other sentence sets (Jabberwocky, Syntactic Prose and Normal Sentences) on both metrics, although this was significant only along the vocalic measure. Direct comparison of the measured values for the Davis et al. sentences with those in the BonnTempoCorpus (Dellwo et al., 2004) suggested that the nonword sentences exhibit tendencies toward the 'syllable-timed' class of languages, thus indicating not only a numerical difference but also a cross-class difference in rhythm between the nonword items and the more typically 'stress-timed' sentences of the other training conditions. Thus, it was concluded that the rhythmic 'unnaturalness' of nonwords sentences may indeed have prevented the observation of any training effect for these sentences in Davis et al. (2005).

Following from the findings of Study 1, Experiment 2 attempted to find a more 'naturalistic' means of testing the role of linguistic rhythm in perceptual adaptation. Several previous studies had shown cross-linguistic transfer of adaptation to time-compressed speech (Mehler et al., 1993; Pallier et al., 1998; Sebastian-Galles et al., 2000). Importantly, this transfer was found, in most cases, to occur only between languages of the same rhythmic class, and not between those from different rhythmic classes. Moreover, this transfer could occur even in the absence of understanding. Given that temporal information is very well preserved in noise-vocoded speech, a cross-linguistic design presented an opportunity to test the null result of Davis et al. (2005) along a rhythmic hypothesis.

It was predicted, for a group of monolingual English listeners, that a Training Phase of exposure to ten 5-band noise-vocoded Dutch (stress-timed) sentences would provide significantly greater training than exposure to equivalent stimuli spoken in Italian (syllable-timed), based on the fact that the Test Phase language, English, was stress-timed. After Pallier et al. (1998), the stimuli were presented without feedback, and the participants were given no explicit instruction about the presence of foreign language materials. The results of the experiment indicated that the groups trained with foreign language stimuli received no training advantage over those who received no training at all. A fourth group of participants experienced training with English noise-vocoded sentences. This condition was expected to provide a significantly better Test Phase performance than control. However, despite a numerical advantage for the English condition, this was not significant over the whole Test Phase. Further exploration of this condition across both Training and Test phases indicated a significant drop in sentence recognition performance at the beginning of the Test Phase, at the same point where a new speaker was introduced to all the training conditions. This suggested that a processing cost incurred with the change in speaker between the Training and Test phases may have masked any possible learning effect in the foreign language conditions.

To address the possibility of a 'speaker change effect' in adaptation noise-vocoded speech, Experiment 3 used English stimuli only, within the same paradigm as Experiment 2, to compare Test Phase performances of listeners who experienced a change in speaker after training with those who heard the same speaker throughout. Similarly to the findings of Dupoux and Green (1997), who studied the same question with time-compressed speech, there was a small but non-significant dip in performance directly after the change in speaker, but this was unlikely to be sufficient to have completely masked learning in Experiment 2. This was supported by the finding in Experiment 4 that the two English speakers in Experiments 2 and 3 were only weakly discriminable with 5-band noise-vocoding. According to the findings of Magnuson and Nusbaum (2007), the lack of conscious expectation or detection of the speaker change in Experiment 2 is unlikely to have allowed this change to affect performance. It emerged that the most likely reason for the drop in performance observed in the English condition of Experiment 2 was item difficulty effects, where the Training set was of overall greater intelligibility than the Test set. A post-hoc analysis of sentence rhythm indicated that these item effects may have had a rhythmic basis, but there was insufficient power to explore this statistically in Experiment 3.

The combined evidence of Experiments 2-3 suggested, in line with the conclusions of Davis et al. (2005), that real word information is necessary to enable learning with noise-vocoded stimuli. Therefore, Experiment 5 re-visited the question of linguistic rhythm in the presence of full English

test sentence materials. A set of 60 sentences was recorded by the same speaker in two ways: once with 'natural' rhythm, and once with 'metronomic' rhythm (which approximated syllable-timing). Four-band noise-vocoded versions of these sentences were presented to English listeners in a within-subjects design, where participants heard 30 sentences from each rhythmic condition in a randomised presentation order. Listeners gave significantly better recognition scores for the 'natural' sentences than the 'metronomic' items, but showed learning in both conditions. While main analyses did not indicate a difference in the rate of learning between the two conditions, post-hoc comparisons indicated that improvement in recognition was slower for the 'metronomic' sentences. A significant relationship of individual sentence recognition performances (in both conditions) to the number of errors on the Seashore Test of Rhythm Perception indicated a role for working memory in the perception of noise-vocoded speech. After the work of Boltz (1998) on the recall of melodies, it was concluded that the effect of the 'metronomic' timing in Experiment 5 may have been to impair the perceptual encoding of the noise-vocoded sentences in working memory.

The results of Study 1 and Experiments 2-5 suggest two initial conclusions regarding the role of linguistic rhythm in the perception of noise-vocoded speech. First, perceptual adaptation on the basis of linguistic rhythm cannot be achieved in the absence of comprehensible linguistic content. Therefore, regardless of the rhythmic properties of the nonword sentences in Davis et al. (2005), these are unlikely to provide effective training in a Training-Test paradigm. Second, when lexical and higher-level linguistic content is present, rhythmic naturalness is advantageous for sentence recognition and perceptual adaptation; however, the effect of rhythm in this context is limited and unlikely to reflect a critical role for this stimulus property in perception. It seems that, under the particular conditions of investigation explored in the present experiments, listeners certainly do not attend to linguistic rhythm, but instead are attempting to identify familiar units of speech at higher levels - the results of Davis et al. (2005) would suggest that they listen for words i.e. at the lexical level.

It is difficult to say with great confidence which is the most likely mechanism for observed effects of linguistic rhythm seen in Experiment 5. The overall level of difficulty of the four-band sentences used in Experiment 5 meant that listeners regularly failed to give full-sentence answers, and so responses could not be analysed for segmentation errors - a greater frequency of segmentation errors in the 'metronomic' sentences may have indicated that rhythm was being used as a segmentation cue (after the *rhythmic segmentation hypothesis* of Cutler and colleagues). However, whether or not rhythmic cues are used for the purposes of segmentation, a more general working memory account remains appealing. Hervais-Adelman et al. (in press) acknowledges that the auditory memory trace generated by a distorted sentence is likely to be less richly encoded and more

quickly fading than that for an undistorted, clearly spoken sentence. In Davis et al. (2005), this has implications for the usefulness of feedback routines in which clear sentence content must be mapped back onto a recent distorted presentation, which may have already faded in echoic memory. On a basic interpretation, any factor which obstructs encoding or weakens the memory trace of a distorted sentence will reduce the likelihood of its full recognition by the participant. So, while manipulations of linguistic rhythm in Experiment 5 were damaging to sentence recognition, these rhythmic alterations may not have directly targeted a specific perceptual mechanism. Rather, they instated an 'incoherence' between the higher-order sentence content and its expected acoustic carrier signal.

10.1.2 Listener Variability in recognition of, and adaptation to, noise-vocoded speech

The question of individual variability in perception of noise-vocoded speech is relatively untouched in the literature. However, its investigation was motivated in the current thesis by the well-documented variability in the outcomes of cochlear implantation, and from personal experience (as an author on Davis et al. (2005) and documented observations (Nogaki et al., 2007; Shannon et al., 2004; Stacey & Summerfield, 2007) of variability in normal-hearing listeners exposed to cochlear implant simulations.

Experiment 2a of the current thesis exploited the variability observed within the cross-linguistic experiment described in Experiment 2 to make an initial assessment of candidate processing correlates of individual sentence recognition performance, using a battery of auditory, cognitive and speech recognition tasks. The results pointed to the importance of 'top-down' cognitive processing in noise-vocoded sentence perception, through significant correlations with performance on the Seashore Rhythm Perception Test (a test of working memory and sustained attention) and vocabulary size (a measure of verbal IQ). A second important outcome of this experiment was the lack of relationship between speech recognition tasks using different linguistic stimulus sets. These two overall findings motivated the remaining experiments of the thesis, which explored cognitive correlates of noise-vocoded sentence recognition (Experiments 6 and 7), and investigated the effects of linguistic content on perception of noise-vocoded stimuli (Experiment 8).

Experiments 6 and 7 compared two approaches to harnessing individual differences in the investigation of noise-vocoded sentence perception - adaptive tracking (Experiment 6) and constant measures (Experiment 7). Both methods attempted to overcome floor and ceiling effects in the

data by expressing performance in terms of the number of bands needed to achieve a threshold level of performance. Correlations between sentence recognition performance from the adaptive track and scores on cognitive tasks in Experiment 6 indicated a role for verbal IQ (as measured by Vocabulary size) in perceptual adaptation to noise-vocoding. The measures extracted directly from the adaptive track reflected whole sentence perception. However, re-analysis using a logarithmic curve-fitting approach allowed the extraction of thresholds based on the number of substituent keywords correcty recognised in the sentences. In this analysis, significant correlations suggested that higher scores on the Nonword Memory Test and Forward Digit Span were associated with lower thresholds and steeper slopes, respectively.

Overall, curve-fitting provided a more readily interpretable representation of individual performance in Experiment 6. However, the negative correlation between threshold and slope parameters (suggesting that lower thresholds are associated with shallower curves) was unexpected, and did not fit easily with the finding that greater Forward Digit Span scores were associated with steeper slopes. Therefore, Experiment 7 repeated the curve-fitting approach with a constant measures distribution of data points that was expcted to be more suitable for slope estimation. Experiment 7 also used the same participants who completed Experiment 6, in order that some assessment of perceptual learning between testing sessions could be made. In this study, the correlations between vocabulary size and sentence recognition were no longer significant, again supporting the interpretation in Experiment 6 of a role for this aspect of verbal IQ in perceptual adaptation, where listeners with more sophisticated linguistic ability might be able to engage more in the task of 'filling in the gaps' in the distorted percept. There were significant correlations between speech recognition scores, Nonword Memory Test and Forward Digit Span scores, but this time only with speech recognition thresholds. Notably, though, the strengths of these correlations with phonological memory scores (in terms of the size of the coefficient) exceeded those reported by Eisenberg et al. (2000), who reported coefficients of $r<0.3$ between noise-vocoded sentence recognition and digit span scores. Looking over all the experiments of the thesis, the positive correlations betweeen sentence recognition and the Seashore Rhythm Perception Test (Experiments 5 and 2a), the Nonword Memory Test (Experiments 6 and 7) and the Forward Digit Span (Experiment 7) strongly suggest a role for short-term memory in accounting for the variability in noise-vocoded sentence recognition.

An important aspect of Experiment 7 was to assess whether improvements in sentence recognition from Experiment 6 could be preserved over a long-term period (around 2 months). Analyses showed significant changes in the threshold and slope parameters of speech recognition performance curves, in which values for both parameters decreased. These results indicated that perceptual

learning resulted in fewer bands being necessary to reach threshold recognition, and in an increase in the steepness of the performance function. Importantly, they suggested that perceptual adaptation to noise-vocoded speech involves the construction long-term changes in acoustic-to-phonetic mappings in the brain. Similar results have been reported by Stacey and Summerfield (2007), but over a shorter time range (9-18 days). These authors also note a significant relationship between baseline recognition scores and the amount of learning, where poorer initial performers tend to be those who exhibit the greatest learning - this relationship has also been documented in auditory perceptual learning by Amitay et al. (2005). This finding was replicated in Experiments 2, 6 and 7. In Experiment 7, the relationship was demonstrated for both thresholds and slopes. The significant increase in slope steepness provided a better understanding of the possible interpretative value of a slope measure in terms of perceptual learning. However, the direction of the relationship between threshold and slope was as observed in Experiment 6. Thus, it is apparent that the usefulness of slope measures is limited in these studies, and should be treated with caution.

Experiment 8 also addressed the question of the long-term nature of perceptual learning, through analysis of listeners' performances in two testing session spaced 1-2 weeks apart. In each testing session, the listener performed five noise-vocoded speech recognition tasks: two for sentence recognition, and one each for isolated words, consonants and vowels. In a curve-fitting analysis, there were small but significant improvements across session in threshold, but not for slopes. However, the negative relationship between baseline performance and the amount of improvement was significant for thresholds and slopes, and provided another replication of the observations made by Stacey and Summerfield (2007) and by Amitay et al. (2005). The overall pattern of results also suggested a significant effect of the linguistic content of the stimulus on both thresholds and slopes. The evidence for different slopes across the tasks challenges the inference by Shannon et al. (2004) that a fixed slope can describe performance on any tasks, and by any listening population, on perceptual tasks with noise-vocoded speech. Despite the sometimes contradictory results obtained in the current study with regard to the role of slope, it seems that there is sufficient evidence that it is a useful measure in the characterization of performance with these tasks.

Some of the more interesting results from Experiment 8 come from the use of individual differences analyses to unpack the processing demands of noise-vocoded speech perception. First, correlational analyses suggested that the listener initially employs two 'levels' of processing when exposed to noise-vocoded speech - a 'top-down', cognitive-liguistic mode and a 'bottom-up' acoustic-phonetic mode - but that listening can become more generalised with perceptual learning. Information transfer analysis on consonant and vowel recognition showed that individual variability in perception of Place of articulation offers the best account of initial consonant recognition scores,

while with learning the perception of Manner of articulation also begins to play a predictive role. For Vowels, the perception of acoustic features related to formant discrimination was most predictive of vowel recognition scores throughout both sessions. However, significant correlations between the amount of information transferred relating to vowel length and performance on sentence and word recognition re-affirms some of the findings described in the previous section with regard to a role for timing information in the perception of noise-vocoded speech.

The use of individual differences analyses has provided evidence for the importance of cognitive factors, particularly working memory, in the recognition of noise-vocoded speech, and allowed the relative influences of 'top-down' and 'bottom-up' factors to be unpacked. The experiments in both strands of the thesis have provided several demonstrations for perceptual learning in the absence of feedback, on both short- and long-term time-scales, and made numerous replications of the observation that those listeners who begin with the poorest performance are generally those who show greatest improvements (Amitay et al., 2005; Stacey & Summerfield, 2007).

10.1.3 Key outcomes, issues and future directions

This section identifies the most important findings of the thesis, and discusses some of the main challenges encountered. Within each section, suggestions are made for improvements to the current design, and proposals set out for future studies.

Working memory and speech perception

Several of the data sets in this thesis directly point to a role for short-term memory in perception of noise-vocoded sentences. Experiments 6 and 7 both gave significant correlations between noise-vocoded sentence report scores and scores on the Nonword Memory Test and Forward Digit Span, while Experiments 2a and 5 show significant correlations between sentence recognition and short-term memory capacity as measured by the Seashore Rhythm Perception Test.

The proposal that phonological short-term memory (as measured in this thesis by the Nonword Memory Test and Digit Span) has a role in speech perception is certainly not new. In a detailed review, Jacquemot and Scott (2006) summarise extensive and convincing evidence of the importance of short-term memory processes in speech perception and production, yet they also point out that little attention has been given to working memory in models of spoken language perception. The authors present a model of short-term working memory that features two phonological buffers

- one for phonological input and one for output, plus reciprocal connections between them. As the current thesis data set stands, there is no conclusive way to differentiate whether the source of variability in phonological short-term memory that relates to noise-vocoded speech perception is the input buffer (speech encoding), the output buffer (speech production), or in the connection from one to the other (the conversion of input information to motoric representations for output). Basic comparison of the structure of the two main phonological working memory tasks used in Experiment 6 and 7 offers the beginnings of an explanation. First, nonword repetition involves repetition of unfamiliar items with no semantic referent, whereas the to-be-remembered items in the Forward Digit Span are highly familiar. Second, the Forward Digit Span is primarily a measure of capacity, whereas nonword repetition is a measure of the ability to encode and repeat phonological information with high accuracy. The results of Experiment 7 indicated that each of these two tasks made an independent contribution to sentence recognition scores. Therefore, we might conclude that variability in both the capacity and the mechanisms of phonological short-term memory contribute to individual differences in noise-vocoded sentence perception. There remains the challenge of determining which specific abilities are being tapped by the Nonword Memory Test. A simple method of differentiating between input and output buffers as the source of variability is to run a test parallel to a repetition task in which speech production is not involved in task performance, for example a nonword matching task. Further, a test of capacity using, for example, measurement of maximum span for repetition of nonword lists would also form a suitable task for correlation with repetition of noise-vocoded sentences.

Hervais-Adelman et al. (in press) consider the limiting effects of distortion on the ability to preserve an auditory memory trace, and the knock-on effects on their feedback regimes. There are two implications for processing, which are not mutually exclusive. The first is that the distortion slows or limits the faithfulness of encoding of the auditory information in working memory, and the second is that the encoded message cannot be actively maintained in the phonological buffer long enough to be compared against a feedback stimulus, which in the case of Davis et al. (2005) and Hervais-Adelman et al. (in press) was an undistorted or written version of the to-be-recognised item. Hervais-Adelman et al. claim that using isolated words would overcome the limitations of short-term memory and allow for a more reliable assessment of whether undistorted presentations truly acted as a top-down 'teaching signal' for perceptual learning. To the extent that single words place less stress on working memory capacity, the authors' claim is likely to hold. However, there may still remain problems with the encoding of a memory trace for a short stimulus such as a single word, if that word is heavily distorted. Burkholder et al. (2005) were able to predict digit span scores with noise-vocoded items using only the intelligibilities of the isolated digits, suggesting that it is the faithfulness of encoding that limits memory rather than a general effect on capacity

for distorted acoustic information. If encoding is compromised for this highly familiar, closed set of items, it will surely be so for isolated noise-vocoded words. There is a possibility that, under severe distortion and during the first seconds of exposure to noise-vocoded stimuli, the auditory memory trace may occupy a 'pre-phonemic' status as an 'echoic' memory - the auditory equivalent of the highly-detailed, fast-fading sensory memory in vision. In many cases, this trace may have faded before phonological encoding was possible.

Another potential means of unpacking the short-term memory effects at play in the perception of distorted speech is via brain imaging studies. Jacquemot and Scott (2006) cite previous studies that have identified different candidate cortical regions as buffers for speech input and output. The posterior superior temporal sulcus, supramarginal gyrus and medial planum temporale form a possible input buffer system, while the left inferior frontal gyrus, inferior motor cortex and anterior insula are proposed in an output buffer network. Variability in the activation of input and/or output buffer regions of cortex that correlates with noise-vocoded sentence recognition may help to describe the role of phonological short-term memory in this speech perception task.

Perceptual learning - how does it happen?

Collison et al. (2004) write that

> The process of matching a variable acoustic signal to an invariant phoneme or syllabic representation requires individuals to make probabilistic matches between a variable input and relatively invariant representations in long-term memory (p. 497).

This was essentially the task faced by the normal-hearing listeners exposed to noise-vocoded speech in the experiments of the current thesis. With perceptual learning, the mapping of the variant to invariant signals becomes more efficient over time, and so speech recognition scores improve - but what are the mechanisms that allow this to happen? Hervais-Adelman et al. (in press) offer the following explanation for the mechanisms of perceptual learning, which they believe to operate under top-down, cognitive influence:

> We hypothesise that the presence or absence of external feedback may not be so crucial as the presence of some constraint on the interpretation of distorted speech that permits listeners to reinforce accurate perceptual hypotheses and make alterations that can correct inaccurate hypotheses

This is a helpful way of looking at the problem of explaining how perceptual learning takes place. We must accept that, in the vast majority of speech perception studies, listeners *expect* to hear speech, usually by virtue of the instructions given to them before a test begins, for example that they will be expected to repeat sentences or make lexical decisions on monosyllabic stimuli. With these expectations, the listener will automatically attempt to interpret auditory stimuli as speech, and with this comes the basis of the 'teaching signal' that Hervais-Adelman et al. describe in their paper. The most powerful demonstration of this was that by Remez et al. (1981), who played listeners the sentence *"Where were you a year ago?"* in sinewave speech. Some participants were asked to describe their percept, without being told anything about it. These listeners described unrelated collections of beeps, whistles and 'science fiction sounds'. However, members of another set of listeners instructed to transcribe a sentence were more likely to recognise some of the content. Effects of attention and expectation will be discussed in more detail in the next section. However, the general expectation of speech input is relevant for the current topic.

In the experiments of the current thesis, there was no feedback on any of the trials, and in many cases the stimuli presented were very heavily distorted, yet listeners showed perceptual adaptation in every noise-vocoded speech recognition task. So, they must have been using some 'constraints', as Hervais-Adelman et al. suggest, to drive the formulation of hypotheses. In the case of sentences, there are numerous lines along which the listener might constrain his or her responses - whether or not they can hear real words, whether their percept has syntactically appropriate structure, whether it makes sense semantically. By imposing these constraints, trial by trial, the listener can gain a sense of whether their response is likely to be accurate, even without feedback. With the consonants and vowels recognition tasks used in Experiment 8, the listener has a slightly different set of constraints. Each of the tasks involves recognition from a closed set of stimuli, thus responses from outside this set are guaranteed to be incorrect and the acoustic matching of sounds to representations can be done on the basis of learned discriminations. Most of the presented stimuli in the vowels task had a lexical entry (e.g. 'bid', 'bead', 'bad'), therefore any non-lexical percept was likely to be one of only a few possible responses, or incorrect. However, each of the tasks also had its own challenges to recognition; for example, the sentences were longer than the materials in the other tasks so placed greater load on memory capacity, and much of the information regarding vowel identity is carried in the formant frequencies, which are very poorly represented in noise-vocoding at low band numbers. For the isolated words recognition task, the balance of constraints and challenges is perhaps least favourable to learning. Despite the opportunity to place constraints of lexical status on responses, there were no effects of syntax or semantics to narrow down the set of possible answers, and with monosyllabic stimuli the probability of choosing a neughbour of the target was often quite high. Although the statistical analysis of thresholds in

Experiment 8 suggested that performance improved by the same degree across all tasks, the words task was, by some margin, the most difficult initially and remained so by the second session of testing. Given that this task had considerable scope for improvement, it might have been expected that performance with the words would improve most dramatically.

A possible effect of task structure on perceptual learning is the range of spectral clarities (in terms of number of bands) used in any particular experiment. One might posit that a task featuring some item with high spectral resolution (e.g. 32-band or undistorted stimuli) might offer the listener a better idea of the linguistic properties of the stimulus set, or provide easier examples of the acoustic-to-phonetic mappings needed to achieve perceptual learning. Alternatively, the inclusion of many levels of spectral resolution, in particular undistorted examples, may be distracting to the 'attentional set' used for perception and thus reduce the amount of learning compared to a design in which the number of bands is fixed. Due to a lack of matching in the linguistic content and number of stimuli across the different experiments of the current thesis, this could not be analysed directly. However, Golomb, Peelle, and Wingfield (2007) recently investigated the effect of intervening undistorted stimuli on adaptation to time-compressed sentences and found no difference in the amount of learning compared with presentation of compressed sentences only. In a recent study using noise-vocoded sentences (Obleser, McGettigan, Alba-Ferrara, & Scott, under review), we found no effect of the range of band numbers used (2, 8 and 32 bands versus 6, 8 and 10 bands), nor of the inclusion of undistorted items, on the recognition of 8-band noise-vocoded sentences.

An important finding in the individual differences analyses of the current data set is the relationship between baseline noise-vocoded speech recognition and the amount of perceptual learning exhibited. This relationship was shown by Amitay et al. (2005) in relation to perceptual learning of frequency discrimination around 1kHz, and more recently by (Stacey & Summerfield, 2007) in normal-hearing listeners undergoing auditory training with spectrally-shifted, noise-vocoded speech. Neither of these two previous studies offers an explanation for why this pattern emerges. The current thesis demonstrates the same pattern of results in four experiments (Experiments 2, 6, 7 and 8) [1]. Perhaps the result is to some extent intuitive - those who start with the poorest scores have the greatest room for improvement. However, the converse scenario is one in which initial performance is so bad that listeners are unable to improve at all - a floor effect. An important aspect of this pattern is that, even after learning, the poorer baseline listeners still tended to be worse than the 'better' listeners at baseline. So, it's not simply a case of a numerical effect. Stacey

[1] This pattern may also have been apparent in Experiments 3 and 5, but did not form part of the analyses for those studies

and Summerfield acknowledge that there is still considerable variability within this overall pattern - this is also observed in the current thesis. Stacey and Summerfield suggest that other factors in addition to baseline performance are likely to be at play in the variability in the amount of learning involved. This is entirely reasonable, but there remains the challenge of how these sources of learning can be disentangled experimentally. Some progress may come from examining the neural correlates of baseline performance and learning. This could be achieved by imaging the brain (e.g. using fMRI) at baseline and *during* learning, to identify those regions whose activity is associated with higher recognition scores at baseline from those whose *change in activity* is associated with improvements in speech recognition. A candidate speech stimulus for this type of study is spectrally-shifted, noise-vocoded speech, which has a slower time-course of learning than unshifted versions (Rosen et al., 1999). There remains, of course, the distinct possibility that these processes will have very similar neural loci, and it is acknowledged that studies to date (Hervais-Adelman et al., in press; Narain et al., 2003) have failed to capture activation changes associated with the learning process.

Task-specific attention and the effects of participant expectations.

It is acknowledged that overall performance in the tasks of this experiment will have been affected by intra-individual fluctuations in arousal and attentional engagement. The very basic task structure of self-timed sentence and item report, without feedback, was important for the design of the current experiments, and may have made the findings of significant long-term perceptual learning more impressive, but perhaps left the data much more susceptible to disengagement effects than would have been the case for the more involved procedures adopted by Fu et al. (2005), Stacey and Summerfield (2007) and M. Smith and Faulkner (2006). Replications of the current experiments with a more engaging testing procedure would certainly be of interest. However, this section of the discussion addresses more specific task-related effects of attention - that is, the effects of listener expectations on the impact of an experimental manipulation.

Experiments 2, 3 and 5 of the current thesis involved signal manipulations of which the listener was not made aware in their task instructions. In Experiment 2, listeners in two of the three training conditions were exposed to noise-vocoded sentences in a language that they did not understand. No feedback was provided, nor was the participant told to attend to any particular aspect of the speech. In this experiment, the only manipulation that had any effect on perceptual learning was the removal of English lexical information - a further rhythmic distinction between the two foreign-language training conditions did not separate them perceptually. Interestingly, hardly any

of the participants noticed, or even speculated, that they had heard foreign-language sentences. In Experiment 3, half of the listeners experienced a change in speaker between the Test Phase and Training Phase. Despite a small indication of a drop in performance in the first two sentences after the change occurred, none of the listeners noticed the change. In a separate test of the discriminability of the two speakers (Experiment 4), a participant with considerable experience of noise-vocoded speech recognition performed considerably better than the other participants, yet found the task very difficult and failed to identify the correct gender for the speakers. In Experiment 5, a manipulation of linguistic rhythm impaired overall sentence recognition scores and gave evidence for slowed perceptual learning. However, in this case none of the participants noticed the manipulation. It must be emphasised that the debriefing process was neither formally administered nor recorded, therefore interpretations of participants' awareness in these experiments must remain speculative at this stage. However, it is striking that some manipulations can affect speech recognition and learning in the absence of awareness, while others may require it.

There are similar examples from the literature on perceptual learning in speech. In their study of perceptual adaptation to spectrally-shifted speech, Fu et al. (2005) noticed that targeted training of medial vowel discrimination with single word stimuli generalized strongly to consonant perception, despite the identification of consonants being irrelevant for performance of the training task. Eisner (2006) carried out a series of experiments related to the findings of Norris et al. (2003) of lexically-driven perceptual learning of phonemic identity. In their original experiment, Norris et al. (2003) exposed listeners to an ambigious phoneme between [f] and [s] in the items of a lexical decision experiment. If the ambiguous phoneme's occurrence was confined to [f]-biased lexical contexts, participants were more likely to label items in an [f]-[s] continuum as [f] in a later categorization task (with the complementary result found for exposure in [s]-biased lexical contexts). However, Eisner (2006) and McQueen, Norris, and Cutler (2006) found that exposure phases in which the interpretation of the ambiguous phoneme was not so critical, or indeed entirely irrelevant, to task performance, the same degree of perceptual learning still took place. Eisner interprets these findings using a model proposed by Seitz and Watanabe (2005) that accounts for task-irrelevant perceptual learning in vision. This model suggests that task-irrelevant features can be learned if their occurrence is temporally coincident with internal reinforcement signals generated from a correct task response (e.g. target identification). In this way, the focus of attention can be placed on a task-relevant stimulus property but subliminal learning of other dimensions can occur coincidentally.

In the case of the current experiments, participants were not given any explicit instructions about task-relevant aspects of the noise-vocoded speech, yet by the nature of the speech signal,

all presented features can be seen as temporally coincident. For sentence recognition, listeners are likely to have tried to identify meaningful lexical units in the speech, in the absence of any directed instruction. Should these meaningful units have been absent, as in the foreign-language conditions of Experiment 2, the listener received no reinforcement and therefore there was no opportunity for coincident learning of rhythmic information (which was outside the focus of attention). However, when potentially reinforcing information was restored through the use of English sentences in Experiment 5, the properties of unattended linguistic rhythm were allowed to contribute to the learning process. Follow-up studies to Experiments 2-5 should explore the effects of participant instruction and expectations by making it possible for listeners to focus their attention on the variable of interest. Further, the data should be amenable to assessment of the participant's perception of this variable. A limitation on the interpretation of the results of Experiment 5 was that listener responses were often incomplete, and so the data set could not be assessed for segmentation errors versus more general errors of perceptual encoding. A testing procedure similar to that used in the sentence training regime of Stacey and Summerfield (2007), in which the listener is forced to give a complete response from a selection of alternatives, presents a plausible modification.

10.2 Conclusion

The experiments of this thesis took a dual approach to investigation of the perception of noise-vocoded speech, in which equal time was devoted to stimulus- and listener-based factors. Experiments investigating the role of linguistic rhythm in recognition of, and perceptual adaptation to, noise-vocoded sentences indicated that while this factor may influence perception, its effect is limited in the context of generalised speech comprehension where meaningful linguistic content seems to present the necessary 'teaching signal' to drive learning. Further experiments on the effects of directed attention and participant expectations are required to develop this research question. The use of individual differences analyses in the current thesis identified direct relationships between measures of phonological short-term memory and the recognition of noise-vocoded sentences, and provided evidence of long-term aspects of perceptual learning of noise-vocoded speech in a variety of linguistic contexts. Behavioural replications incorporating more stringent controls on arousal and attentional engagement may yield even stronger results. Furthermore, future experiments in neuroimaging may provide greater insight into the locus of working memory effects and contribute greater understanding of the relationship between baseline abilities and the capacity to learn.

References

Abercrombie, D. (1967). *Elements of General Phonetics.* Edinburgh: University Press.

Altmann, G., & Young, D. (1993). Factors affecting adaptation to time-compressed speech. In *EUROSPEECH '93* (p. 333-336).

Amitay, S., Hawkey, D., & Moore, D. (2005). Auditory frequency discrimination learning is affected by stimulus variability. *Perception & Psychophysics, 67*(4), 691 - 698.

Amitay, S., Irwin, A., & Moore, D. (2006). Discrimination learning induced by training with identical stimuli. *Nature Neuroscience, 9*(11), 1446 - 1448.

Apoux, F., Crouzet, O., & Lorenzi, C. (2001). Temporal envelope expansion of speech in noise for normal-hearing and hearing-impaired listeners: effects on identification performance and response times. *Hearing Research, 153*(1-2), 123-131.

Apoux, F., Tribut, N., Debruille, X., & Lorenzi, C. (2004). Identification of envelope-expanded sentences in normal-hearing and hearing-impaired listeners. *Hearing Research, 189*(1-2), 13-24.

Atienza, M., Cantero, J., & Dominguez-Marin, E. (2002). The time course of neural changes underlying auditory perceptual learning. *Learning & Memory, 9*(3), 138 - 150.

Baddeley, A., Gatherole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158-173.

Baddeley, A., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 36*(2), 233-252.

Banziger, T., & Scherer, K. (2005). The role of intonation in emotional expressions. *Speech Communication, 46*(3-4), 252 - 267.

Bashford, J., Warren, R., & Brown, C. (1996). Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration. *Perception & Psychophysics, 58*(3), 342 - 350.

Becker, B. (1990). Coaching for the Scholastic Aptitude Test - Further synthesis and appraisal. *Review of Educational Research, 60*(3), 373 - 417.

Bench, J., Kowal, A., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology, 13*(3), 108-112.

Ben-Yehudah, G., & Ahisaar, M. (2004). Sequential spatial frequency discrimination is consistently impaired among adult dyslexics. *Vision Research, 44*, 1047-1063.

Blumstein, S., & Stevens, K. (1979). Acoustic invariance in speech production - evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66*(4), 1001 - 1017.

Boersma, P., & Weenink, D. (2005, February). Praat: Doing phonetics by computer (version 4.3.02) [Computer software].

Boltz, M. (1998). The processing of temporal and nontemporal information in the remembering of event durations and musical structure. *Journal of Experimental Psychology - Human Perception and Performance, 24*(4), 1087 - 1104.

Boothroyd, A. (1968). Developements in speech audiometry. *Sound, 2*, 3-10.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America, 84*(1), 101-114.

Bradlow, A., & Bent, T. (2003, Aug 3-9). Listener adaptation to foreign-accented speech. In *Proceedings of the 15th international congress of phonetic sciences.* Barcelona, Spain.

Burkholder, R. (2005). *Perceptual learning of speech processed through an acoustic simulation of a cochlear implant* (Tech. Rep. No. 13). Indiana University.

Burkholder, R., Pisoni, D., & Svirsky, M. (2005). Effects of a cochlear implant simulation on immediate memory in normal-hearing adults. *International Journal of Audiology, 44*(10), 551 - 558.

Burns, E., Sanborn, E., Shannon, R., & Fu, Q. (2001). Perception of familiar melodies by implant users. In *Proceedings of the conference on implantable auditory prostheses* (p. 81). Pacific Grove, CA.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., & Snyder, P. (2004). Voice acoustical measurement of the severity of major depression. *Brain and Cognition, 56*(1), 30 - 35.

Carroll, J. (1993). *Human cognitive abilities: A survey of factor-analytical studies.* NY: Cambridge University Press.

Charpentier, F., & Stella, M. (1986). Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proceedings of ICASSP 86* (p. 2015-2018).

Charter, R., & Webster, J. (1997). Psychometric structure of the Seashore Rhythm Test. *Clinical Neuropsychologist, 11*(2), 167-173.

Chiu, P., Eng, M., Strange, B., Yampolsky, S., & Waters, G. (2002, May). *On learning to recognize spectrally reduced speech: Ii. individual differences* [Poster presented at the 143rd meeting of the Acoustical Society of America]. Pittsburgh, PA.

Ciocca, V., Francis, A., Aisha, R., & Wong, L. (2002). The perception of cantonese lexical tones by early-deafened cochlear implantees. *Journal of the Acoustical Society of America, 111*(5), 2250 - 2256.

Clark, G. (2002). *Learning to understand speech with the cochlear implant* (M. Fahle & T. Poggio, Eds.). Cambridge, MA: MIT Press.

Clarke, C., & Garrett, M. (2004). Rapid adaptation to foreign-accented english. *Journal of the Acoustical Society of America, 116*(6), 3647 - 3658.

Cleary, M., Pisoni, D., & Geers, A. (2001). Some measures of verbal and spatial working memory in eight- and nine-year-old hearing-impaired children with cochlear implants. *Ear and Hearing, 22*(5), 395 - 411.

Cleary, M., Pisoni, D., & Kirk, K. (2002). Working memory spans as predictors of spoken word recognition and receptive vocabulary in children with cochlear implants. *Volta Review, 102*(4), 259 - 280.

Collins, N., & Cross, I. (2005). Beat tracking and reaction time. In *10th rhythm perception and production workshop* (p. 81). Alden Biesen, Belgium.

Collison, E., Munson, B., & Carney, A. (2004). Relations among linguistic and cognitive skills and spoken word recognition in adults with cochlear implants. *Journal of Speech Language and Hearing Research, 47*(3), 496 - 508.

Cutler, A. (1994a). The perception of rhythm in language. *Cognition, 50*(1-3), 79 - 81.

Cutler, A. (1994b). Segmentation problems, rhythmic solutions. *Lingua, 92*(1-4), 81 - 104.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation - evidence from juncture misperception. *Journal of Memory and Language, 31*(2), 218 - 236.

Cutler, A., & Foss, D. (1977). Role of sentence stress in sentence processing. *Language and Speech, 20*, 1 - 10.

Cutler, A., & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics, 21*(1-2), 103 - 108.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of french and english. *Journal of Memory and Language, 25*(4), 385 - 400.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology - Human Perception and Performance, 14*(1), 113 - 121.

Cutler, A., Norris, D., & Williams, J. (1987). A note on the role of phonological expectations in speech segmentation. *Journal of Memory and Language, 26*(4), 480 - 487.

Cutler, A., Sebastian-Galles, N., Soler-Vilageliu, O., & van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition, 28*, 746-755.

Dauer, R. (1987a). Phonetic and phonological components of language rhythm. In *Proceedings of the 11th International Congress of Phonetic Sciences* (p. 447-450). Tallinn, Estonia.

Dauer, R. (1987b). Stress-timing and syllable-timing reanalysed. *Journal of Phonetics, 11*, 51-69.

Davis, M., & Johnsrude, I. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience, 23*(8), 3423-3431.

Davis, M., & Johnsrude, I. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research, 229*(1-2), 132-147.

Davis, M., Johnsrude, I., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of noise-vocoded speech. *Journal of Experimental Psychology:General, 4*(2), 254-264.

Davis, M., Marslen-Wilson, W., & Gaskell, M. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28*(1), 218-244.

Defilippo, C., & Scott, B. (1978). Method for training and evaluating reception of ongoing speech. *Journal of the Acoustical Society of America, 63*(4), 1186 - 1192.

Dehaan, H. (1982). The relationship of estimated comprehensibility to the rate of connected speech. *Perception & Psychophysics, 32*(1), 27 - 31.

Dellwo, V. (2007). *Influences of speech rate on acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence.* Unpublished doctoral dissertation, University of Bonn.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovičová, J., & Wagner, P. (2004). The bonntempo-

corpus and bonntempo-tools. a database for the combined study of speech rhythm and rate. In *Proceedings of the 8th International Conference on Spoken Language Processing.* Jeju Island, Korea.

Dellwo, V., & Wagner, P. (2003). Relations between Language Rhythm and Speech Rate. In *Proceedings of the 15th International Congress of Phonetic Sciences* (p. 471-474). Barcelona, Spain.

Deterding, D. (2001). The measurement of rhythm: a comparison of singapore and british english. *Journal of Phonetics, 29*(2), 217 - 230.

Dillon, C., Burkholder, R., Cleary, M., & Pisoni, D. (2004). Nonword repetition by children with cochlear implants: Accuracy ratings from normal-hearing listeners. *Journal of Speech Language and Hearing Research, 47*(5), 1103 - 1116.

Dillon, C., Pisoni, D., Cleary, M., & Carter, A. (2004). Nonword imitation by children with cochlear implants - Consonant analyses. *Archives of Otolaryngology - Head & Neck Surgery, 130*(5), 587 - 591.

Dorman, M., Hannley, M., Dankowski, K., Smith, L., & McCandless, G. (1989). Word recognition by 50 patients fitted with the symbion multichannel cochlear implant. *Ear and Hearing, 10*(1), 44 - 49.

Dorman, M., & Loizou, P. (1997a). Mechanisms of vowel recognition for ineraid patients fit with continuous interleaved sampling processors. *Journal of the Acoustical Society of America, 102*(1), 581-587.

Dorman, M., & Loizou, P. (1997b). Speech intelligibility as a function of the number of channels of stimulation for normal-hearing listeners and patients with cochlear implants. *American Journal of Otology, 18*(6), S113-S114.

Dorman, M., & Loizou, P. (1998). The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels. *Ear and Hearing, 19*(2), 162-166.

Dorman, M., Loizou, P., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America, 102*, 2403-2411.

Dorman, M., Soli, S., Dankowski, K., Smith, L., Parkin, J., & McCandless, G. (1990). Acoustic cues for consonant identification by patients who use the Ineraid cochlear implant. *Journal of the Acoustical Society of America, 88*(5), 2074-2079.

Dunn, L., Whetton, C., & Burley, J. (1997). *British Picture Vocabulary Scale, Version-II.* London: NFER-Nelson.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology - Human Perception and Performance, 23*(3), 914 - 927.

Dupoux, E., Pallier, C., Sebastian, N., & Mehler, J. (1997). A destressing "deafness" in french? *Journal of Memory and Language, 36*(3), 406 - 421.

Eisenberg, L. S., Shannon, R. V., Martinez, A. S., & A Boothroyd, J. W. ad. (2000). Speech recognition with reduced spectral cues as a function of age. *Journal of the Acoustical Society of America, 107*(5), 2704-2710.

Eisner, F. (2006). Lexically-guided perceptual learning in speech processing. In *MPI Series in Psycholinguistics.* Wageningen, Netherlands: Ponsen & Looijen bv.

Eisner, F., & McQueen, J. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224 - 238.

Evans, B., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern british english sentences. *Journal of the Acoustical Society of America, 115*(1), 352-361.

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 108*(4), 1877-1887.

Faulkner, A., Rosen, S., & Stanton, D. (2003). Simulations of tonotopically mapped speech processors for cochlear implant electrodes varying in insertion depth. *Journal of the Acoustical Society of America, 113*(2), 1073 - 1080.

Fishman, K. E., Shannon, R. V., & Slattery, W. H. (1997). Speech recognition as a function of the number of electrodes used in the speak cochlear implant speech processor. *Journal of Speech Language and Hearing Research, 40*(5), 1201-1215.

Foulke, E., & Sticht, T. (1969). Review of research on intelligibility and comprehension of accelerated speech. *Psychological Bulletin, 72*(1), 50 - .

Friesen, L., Shannon, R., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America, 110*(2), 1150 - 1163.

Fu, Q., Chinchilla, S., & Galvin, J. (2004). The role of spectral and temporal cues in voice gender

discrimination by normal-hearing listeners and cochlear implant users. *Journal of the Association for Research in Otolaryngology, 5*(3), 253 - 260.

Fu, Q., Nogaki, G., & Galvin, J. (2005). Auditory training with spectrally shifted speech: Implications for cochlear implant patient auditory rehabilitation. *Journal of the Association for Research in Otolaryngology, 6*(2), 180 - 189.

Fu, Q., & Shannon, R. (1999a). Effect of acoustic dynamic range on phoneme recognition in quiet and noise by cochlear implant users. *Journal of the Acoustical Society of America, 106*(6), L65 - L70.

Fu, Q., & Shannon, R. (1999b). Effects of electrode location and spacing on phoneme recognition with the nucleus-22 cochlear implant. *Ear and Hearing, 20*(4), 321 - 331.

Fu, Q., Shannon, R., & Galvin, J. (2002). Perceptual learning following changes in the frequency-to-electrode assignment with the nucleus-22 cochlear implant. *Journal of the Acoustical Society of America, 112*(4), 1664 - 1674.

Galton, F. (1869). *Hereditary Genius: An Inquiry into its Laws and Consequences.* London: Macmillan.

Gantz, B., Woodworth, G., Knutson, J., Abbas, P., & Tyler, R. (1993). Multivariate predictors of audiological success with multichannel cochlear implants. *Annals of Otology, Rhinology and Laryngology, 102*(12), 909 - 916.

Gathercole, S. (1995). Is nonword repetition a test of phonological memory or long-term knowledge - it all depends on the nonwords. *Memory & Cognition, 23*(1), 83 - 94.

Gathercole, S., & Baddeley, A. (1989). Evaulation of the role of phonological STM in the development of vocabulary in children - a longitudinal study. *Journal of Memory and Language, 28*(2), 200 - 213.

Gathercole, S., & Baddeley, A. (1996). *The Nonword Memory Test.* University of Durham.

Gathercole, S., Hitch, G., Service, E., & Martin, A. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology, 33*(6), 966 - 979.

Gathercole, S., Service, E., Hitch, G., Adams, A., & Martin, A. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology, 13*(1), 65 - 77.

Gathercole, S., Willis, C., Baddeley, A., & Emslie, H. (1994). The Children's Test of Nonword Repetition: a test of phonological working memory. *Memory, 2*(2), 103-127.

Gathercole, S., Willis, C., Emslie, H., & Baddeley, A. (1991). The influences of number of syllables and wordlikeness on children's repetition of nonwords. *Applied Psycholinguistics, 12*(3), 349 - 367.

Gfeller, K., & Lansing, C. (1992). Musical perception of cochlear implant users as measured by the primary measures of music audiation - an item analysis. *JOURNAL OF MUSIC THERAPY, 29*(1), 18 - 39.

Gfeller, K., Witt, S., Spencer, L., Stordahl, J., & Tomblin, B. (1998). Musical involvement and enjoyment of children who use cochlear implants. *Volta Review, 100*(4), 213 - 233.

Goldstone, R. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585 - 612.

Golestani, N., Molko, N., Dehaene, S., Lebihan, D., & Pallier, C. (2007). Brain structure predicts the learning of foreign speech sounds. *Cerebral Cortex, 17*(3), 1701-1708.

Golestani, N., Paus, T., & Zatorre, R. (2002). Anatomical correlates of learning novel speech sounds. *Neuron, 35*(5), 997 - 1010.

Golestani, N., & Zatorre, R. (2004). Learning new sounds of speech: reallocation of neural substrates. *NeuroImage, 21*(2), 494 - 506.

Golomb, J., Peelle, J., & Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *Journal of the Acoustical Society of America, 121*(3), 1701 - 1708.

Gonzalez, J., & Oliver, J. (2005). Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *Journal of the Acoustical Society of America, 118*(1), 461 - 470.

Gordon, E. (1986). *Primary measures of music audiation and the intermediate measures of music audiation: Music aptitutde tests for kindergarten and first, second, third and fourth grade children.* Chicago: G.I.A. Publications, Inc.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., et al. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of The National Academy of Sciences of the United States of America, 99*(16), 10911 - 10916.

Grabe, E., & Low, E. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Papers in laboratory phonology 7* (p. 515-546). Berlin: Mouton de Gruyter.

Grant, K., & Seitz, P. (2000). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *Journal of the Acoustical Society of America, 107*(2), 1000-1011.

Greenwood, D. (1990). A cochlear frequency-position for several species - 29 years later. *Journal of the Acoustical Society of America, 87*(6), 2592 - 2605.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *Quarterly Journal of Experimental Psychology Section A - Human Experimental Psychology, 56*(7), 1213 - 1236.

Halstead, W. (1947). *Brain and intelligence.* Chicago: University of Chicago Press.

Hanekom, J., & Shannon, R. (1998). Gap detection as a measure of electrode interaction in cochlear implants. *Journal of the Acoustical Society of America, 104*(4), 2372 - 2384.

Hannemann, R., Obleser, J., & Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research, 1153*, 134 - 143.

Heiman, G., Leo, R., Leighbody, G., & Bowler, K. (1986). Word intelligibility decrements and the comprehension time-compressed speech. *Perception & Psychophysics, 40*(6), 407 - 411.

Hervais-Adelman, A., Davis, M., Taylor, K., Carlyon, R., & Johnsrude, I. (2006). Perceptual learning of vocoded speech: Where does it occur? Exploiting generalisability to find the locus of change. *Journal of Cognitive Neuroscience, Suppl.*

Hervais-Adelman, A., Johnsrude, I., Carlyon, R., & Davis, M. (2007). Effortful comprehension of noise vocoded speech recruits a fronto-temporal network. *Journal of Cognitive Neuroscience, Suppl.*

Hervais-Adelman, A., Johnsrude, I., Davis, M., & Carlyon, R. (in press). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception & Performance.*

IEEE. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*(3), 225-246.

Iverson, P., Smith, C., & Evans, B. (2007). Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration. *Journal of the Acoustical Society of America, 120*(6), 3998-4006.

Jackendoff, R. (1989). A comparison of rhythmic structures in music and language. In P. Kiparsky & G. Youmanns (Eds.), *Phonetics and phonology: Vol 1. rhythm and meter* (p. 15-44). New York: Academic Press.

Jackendoff, R., & Lerdahl, F. (1982). A grammatical parallel betwen music and language. In M. Clynes (Ed.), *Music, mind and brain* (p. 83-117). New York: Plenum Press.

Jacquemot, C., & Scott, S. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences, 10*(11), 480 - 486.

Jones, M., & Yee, W. (1993). *Attending to auditory events: the role of temporal organization.* NY: Oxford University Press.

Kalikow, D., Stevens, K., & Elliott, L. (1977). Development of a test of speech-intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61*(5), 1337 - 1351.

Kewley-Port, D., Pisoni, D., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America, 73*(5), 1779 - 1793.

Kidd, G., Watson, C., & Gygi, B. (2007). Individual differences in auditory abilities. *Journal of the Acoustical Society of America, 122*(1), 418 - 435.

Kluender, K., Coady, J., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication, 41*, 59-69.

Knutson, J., Hinrichs, J., Tyler, R., Gantz, B., Schartz, H., & Woodworth, G. (1991). Psychological predictors of audiological outcomes of multichannel cochlear implants - preliminary findings. *Annals of Otology, Rhinology and Laryngology, 100*(10), 817 - 822.

Kong, Y., Cruz, R., Jones, J., & Zeng, F. (2004). Music perception with temporal cues in acoustic and electric hearing. *Ear and Hearing, 25*(2), 173 - 185.

Kraljic, T., & Samuel, A. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141 - 178.

Krause, J., & Braida, L. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America, 115*, 362-378.

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion, 19*, 633-653.

Leal, M., Shin, Y., Laborde, M., Calmels, M., Verges, S., & Lugardon, S. e. a. (2003). Music perception in adult cochlear implant recipients. *Acta Oto-Laryngologica, 123*, 826-835.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49*(2), 467 - .

Liu, E., S amd Del Rio, Bradlow, A., & Zeng, F. (2004). Clear speech perception in acoustic and electric hearing. *Journal of the Acoustical Society of America, 116*, 2374-2383.

Logan, J., Lively, S., & Pisoni, D. (1991). Training Japanese Listeners to Identify English /r/ and /l/ - A 1st Report. *Journal of the Acoustical Society of America, 89*, 874-836.

Loizou, P., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America, 106*, 2097-2103.

Low, E., Grabe, E., & Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in singapore english. *Language and Speech, 43*, 377 - 401.

Lyxell, B., Andersson, J., Andersson, U., Arlinger, S., Bredberg, G., & Harder, H. (1998). Phonological representation and speech understanding with cochlear implants in deafened adults. *Scandanavian Journal of Psychology, 39*(3), 175-179.

Magnuson, J., & Nusbaum, H. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology - Human Perception and Performance, 33*(2), 391 - 409.

Marslen-wilson, W. (1985). Speech Shadowing and Speech Comprehension. *Speech Communication, 4*, 55-73.

Marslen-Wilson, W., & Warren, P. (1994). Levels of Perceptual Representation and Process in Lexical Access - Words, Phonemes, and Features. *Psychological Review, 101*, 653-675.

Martin, J. (n.d.). Aspects of rhythmic structure in speech perception. In J. Evans & M. Clynes (Eds.), *Rhythm in psychological, linguistic and musical processes* (p. 79-98). Springfield, IL: Charles C Thomas.

Martin, J. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review, 79*(6), 487 - 509.

Martin, R., Shelton, J., & Yaffee, L. (1994). Language processing and working-memory - neuropsychological evidence for separate phonological and semantic capacities. *Journal of Memory and Language, 33*(1), 83 - 111.

Mattys, S., & Melhorn, J. (2007). Sentential, lexical, and acoustic effects on the perception of word boundaries. *Journal of the Acoustical Society of America, 122*(1), 554 - 567.

Mattys, S., Melhorn, J., & White, L. (2007). Effects of syntactic expectations on speech segmentation. *Journal of Experimental Psychology - Human Perception and Performance, 33*(4), 960 - 977.

Mattys, S., White, L., & Melhorn, J. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology - General, 134*(4), 477 - 500.

McQueen, J., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech, 49*, 101 - 112.

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The Syllable's Role in Speech Segmentation. *Journal of Verbal Learning and Verbal Behaviour, 20*, 298-305.

Mehler, J., Sebastian, N., Altmann, G., Dupoux, E., Christophe, A., & Pallier, C. (1993). Understanding compressed sentences - the role of rhythm and meaning. *Annals of the New York Academy of Sciences, 682*, 272 - 282.

Meister, H., Tepeli, D., Wagner, P., Hess, W., Walger, M., wedel, H. von, et al. (2007). Experiments on prosody perception with cochlear implants. *HNO, 55*(4), 264 - 270.

Michon, J. (1964). Temporal structure of letter gruops and span of perception. *Quarterly Journal of Experimental Psychology, 16*, 232-240.

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

Miller, J., & Baer, T. (1983). Some Effects of Speaking Rate on the Production of /b/ and /w/. *Journal of the Acoustical Society of America, 73*, 1751-1755.

Miller, J., Green, K., & Reeves, A. (1986). Speaking Rate and Segments - A Look at the Relation Between Speech Production and Speech Perception. *Phonetica, 43*, 106-115.

Moore, B. (1997). *Speech Perception.* London: Academic Press.

Moore, B. (2003). Coding of sounds in the auditory system and its relevance to signal processing and coding in cohclear implants. *Otology & Neurotology, 24*, 243-254.

Muneaux, M., Ziegler, J., Truc, C., Thomson, J., & Goswami, U. (2004). Deficits in beat perception and dyslexia: evidence from French. *Neuroreport, 15*, 1255-1259.

Munson, B. (2004). Variability in /s/ production in children and adults: Evidence from dynamic measures of spectral mean. *Speech Language and Hearing Research, 47*, 58-69.

Munson, B., & Babel, M. (2005). The sequential cueing effect in children's speech production. *Applied Psycholinguistics, 26*, 157-174.

Munson, B., Donaldson, G., Allen, S., Collison, E., & Nelson, D. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability. *Journal of the Acoustical Society of America, 113*(2), 925 - 935.

Nakata, T., Trehub, S., Mitani, C., & Kanda, Y. (2006). Pitch and timing in the songs of deaf children with cochlear implants. *Music Perception, 24*(2), 147 - 154.

Narain, C., Scott, S., Wise, R., Rosen, S., Leff, A., Iversen, S., et al. (2003). Defining a left-lateralized response specific to intelligible speech using fmri. *Cerebral Cortex, 13*(12), 1362 - 1368.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology - Human Perception and Performance, 24*(3), 756 - 766.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication, 41*, 233-243.

Newman, R., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics, 35*(1), 85 - 103.

Nie, K., Stickney, G., & Zeng, F. (2005). Encoding frequency modulation to improve cochlear implant performance in noise. *IEEE Transactions on Biomedical Engineering, 52*, 64-73.

Nogaki, G., Fu, Q., & Galvin, J. (2007). Effect of Training Rate on Recognition of Spectrally Shifted Speech. *Ear & Hearing, 28*(2), 132-140.

Norris, D., McQueen, J., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204 - 238.

Obleser, J., McGettigan, C., Alba-Ferrara, L., & Scott, S. (under review). Interaction of acoustic and cognitive processes in the perception of degraded speech. *Perception & Psychophysics.*

Obleser, J., Wise, R., Dresner, M., & Scott, S. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience, 27*(9), 2283 - 2289.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or Syllable - Speech Segmentation in Japanese. *Journal of Memory and Language, 32*, 258-278.

Pallier, C., Sebastian-Galles, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition, 26*(4), 844 - 851.

Patel, A. (2006). Musical rhythm, linguistic rhythm, and human evolution. *Music Perception, 24*(1), 99 - 103.

Patel, A., & Daniele, J. (2003). An empirical comparison of rhythm in language and music. *Cognition, 87*(1), B35 - B45.

Peterson, G., & Barney, H. (1952). Control Methods Used in A Study of the Vowels. *Journal of the Acoustical Society of America, 24*, 175-184.

Peterzell, D., Chang, S., & Teller, D. (2000). Spatial frequency tuned covariance channels for red-green and luminance-modulated gratings: psychophysical data from human infants. *Vision Research, 40*(4), 431 - 444.

Peterzell, D., & Teller, D. (2000). Spatial frequency tuned covariance channels for red-green and luminance-modulated gratings: psychophysical data from human adults. *Vision Research, 40*(4), 417-430.

Pike, K. (1945). *Intonation of American English.* Ann Arbor: University of Michigan Press.

Pisoni, D. (1991). Effects of Alcohol on Speech - Acoustic Analysis of the Exxon-Valdez Tapes. *Journal of Psycholinguistic Research, 20*, 524-525.

Pisoni, D. (2000). Cognitive factors and cochlear implants: Some thoughts on perception, learning, and memory in speech perception. *Ear and Hearing, 21*(1), 70 - 78.

Pisoni, D., & Cleary, M. (2003). Measures of working memory span and verbal rehearsal speed in deaf children after cochlear implantation. *Ear and Hearing, 24*(1), 106S - 120S.

Pisoni, D., & Geers, A. (2000). Working memory in deaf children with cochlear implants: Correlations between digit span and measures of spoken language processing. *Annals of Otology, Rhinology and Laryngology, 109*(12), 92 - 93.

Port, R. (1979). Combinations of Timing Factors in Speech Production. *Journal of the Acoustical Society of America, 65*, S33.

Rabinowitz, W., Eddington, D., Delhorne, L., & Cuneo, P. (1992). Relations among different measures of speech reception in subjects using a cochlear implant. *Journal of the Acoustical Society of America, 92*(4), 1869-1881.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*(3), 265-292.

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). Speech-perception without traditional speech cues. *Science, 212*(4497), 947 - 950.

Roach, P. (1982). On the disctinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (Ed.), *Linguistic controversies* (p. 73-79). London: Edward Arnold.

Rodd, J., Davis, M., & Johnsrude, I. (2005). The neural mechanisms of speech comphrehension: fMRI studies of semantic ambiguity. *Cerebral Cortex, 15*(8), 1261-1269.

Romani, C., & Martin, R. (1999). A deficit in the short-term retention of lexical-semantic information: Forgetting words but remembering a story. *Journal of Experimental Psychology: General,*

128(1), 56 - 77.

Rosen, S. (1992). Temporal information in speech - acoustic, auditory and linguistic aspects. *Philosophical Transactions of The Royal Society of London Series B - Biological Sciences, 336*(1278), 367 - 373.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 106*(6), 3629 - 3636.

Rosen, S., Finn, R., & Faulkner, A. (2002). Plasticity in speech perception: spectrally-rotated speech, revisited. In *Association for Research in Otolaryngology Midwinter Meeting.* St Petersburg Beach, FL.

Rubenstein, J. (2004). How cochlear implants encode speech. *Current Opinion in Otolaryngology & Head and Neck Surgery, 12*, 444-448.

Ryan, J. (1969a). Grouping and short-term memory: different means and patterns of grouping. *Quarterly Journal of Experimental Psychology, 21*, 137-147.

Ryan, J. (1969b). Temporal grouping, rehearsal and short-term memory. *Quarterly Journal of Experimental Psychology, 21*, 148-155.

Salomon, A and Espy-Wilson, CY and Deshmukh, O. (n.d.). Detection of speech landmarks: Use of temporal information. *Journal of the Acoustical Society of America, 115*, 1296-1305.

Sarant, J., Blamey, P., Dowell, R., Clark, G., & Gibson, W. (2001). Variation in speech perception scores among children with cochlear implants. *Ear and Hearing, 22*(1), 18 - 28.

Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America, 103*, 588-601.

Schwab, E., Nusbaum, H., & Pisoni, D. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27*, 395-408.

Scott, S., Blank, C., Rosen, S., & Wise, R. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123*, 2400 - 2406.

Seashore, C., Lewis, D., & Saetreit, J. (1960). *Seashore measures of musical talents (revised).* NY: Psychological Corporation.

Sebastian-Galles, N., Dupoux, E., Costa, A., & Mehler, J. (2000). Adaptation to time-compressed speech: Phonological determinants. *Perception & Psychophysics, 62*(4), 834 - 842.

Seitz, A., & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive*

Sciences, *9*(7), 329 - 334.

Shannon, R. (2007). Understanding hearing through deafness. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(17), 6883 - 6884.

Shannon, R., Fu, Q., & Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Oto-Laryngologica*, *124*, 50 - 54.

Shannon, R., Galvin, J., & Baskent, D. (2002). Holes in hearing. *Journal of the Association of Research in Otolaryngology*, *124*, 50-54.

Shannon, R., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303-304.

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

Sharp, D., Scott, S., & Wise, R. (2004a). Monitoring and the controlled processing of meaning: Distinct prefrontal systems. *Cerebral Cortex*, *14*, 1-10.

Sharp, D., Scott, S., & Wise, R. (2004b). Retrieving meaning after temporal lobe infarction: The role of the basal language area. *Annals of Neurology*, *56*, 836-846.

Sherer, M., Parsons, O., Nixon, S., & Adams, R. (1991). Clinical validity of the Speech-Sounds Perception Test and the Seashore Rhythm Test. *Journal of Clinical and Experimental Psychology*, *13*(5), 741-751.

Skinner, M. (2003). Optimizing cochlear implant speech performance. *Annals of Otology, Rhinology and Laryngology*, *112*(9), 4 - 13.

Slack, W., & Porter, D. (1980). The Scholastic Aptitude Test - a critical appraisall. *Harvard Educational Review*, *50*(2), 154 - 175.

Smith, B., & Kenney, M. (1998). An assessment of several acoustic parameters in children's speech production development: longitudinal data. *Journal of Phonetics*, *26*, 95-108.

Smith, B., Kenney, M., & Hussain, S. (1996). A longitudinal investigation of duration and temporal variability in children's speech production. *Journal of the Acoustical Society of America*, *99*, 2344-2349.

Smith, M., Cutler, A., Butterfield, S., & Nimmo-Smith, I. (1989). The perception of rhythm and word boundaries in noise-masked speech. *Journal of Speech and Hearing Research*, *32*(4), 912 - 920.

Smith, M., & Faulkner, A. (2006). Perceptual adaptation by normally hearing listeners to a simulated "hole" in hearing. *Journal of the Acoustical Society of America, 120*(6), 4019-4030.

Smith, Z., Delgutte, B., & Oxenham, A. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416*, 87-90.

Stacey, P., & Summerfield, A. (2007). Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech. *Journal of the Acoustical Society of America, 121*(5), 2923 - 2935.

Stankov, L., & Horn, J. (1980). Human Abilities Revealed Through Auditory Tests. *Journal of Educational Psychology, 72*, 21-44.

Sternberg, R. (1977). *Intelligence, Information Processing, and Analogical Reasoning.* NY: Erlbaum.

Stevens, K. (1980). Acoustic Correlates of Some Phonetic Categories. *Journal of the Acoustical Society of America, 68*, 836-842.

Summerfield, Q. (1981). Articulatory Rate and Perceptual Constancy in Phonetic Perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1074-1095.

Surprenant, A., & Watson, C. (2001). Individual differences in the processing of speech and non-speech sounds by normal-hearing listeners. *Journal of the Acoustical Society of America, 110*(4), 2085-2095.

Swinney, D. (1979). Lexical Access During Sentence Comprehension - (Re)Consideration of Context Effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645-659.

Tallal, P. (1980). Auditory temporal perception, phonics and reading disabilities in children. *Brain and Language, 9*, 182-198.

Thomson, J., Fryer, B., Maltby, J., & Goswami, U. (2006). Auditory and motor rhythm awareness in adults with dyslexia. *Journal of Research in Reading, 29*, 334-348.

Turner, C., Gantz, B., Vidal, C., Behrens, A., & Henry, B. (2004). Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing. *Journal of the Acoustical Society of America, 115*, 1729-1735.

Tyler, R., & Summerfield, A. (1996). Cochlear implantation: Relationships with research on auditory deprivation and acclimatization. *Ear and Hearing, 17*, S38-S50.

Valimaa, T., Maatta, T., Lopponen, H., & Sorri, M. (2002). Phoneme recognition and confusions with multichannel cochlear implants: Consonants. *Journal of Speech, Language and Hearing*

Research, 45, 1055-1069.

van der Horst, R., Leeuw, A., & Dreschler, W. (1999). Importance of temporal-envelope cues in consonant recognition. *Journal of the Acoustical Society of America, 105*, 1801-1809.

van Ooijen, B. (1996). Vowel mutability and lexical selection in English: Evidence from a word reconstruction task. *Memory & Cognition, 24*, 573-583.

van Rooij, J., Plomp, R., & Orlebeke, J. (1989). Auditive and cognitive factors in speech-perception by elderly listeners. 1. Development of test battery. *Journal of the Acoustical Society of America, 86*(4), 1294 - 1309.

van Tasell, D., Soli, S., Kirby, V., & Widin, G. (1987). Speech Wave-Form Envelope Cues for Consonant Recognition. *Journal of the Acoustical Society of America, 82*, 1152-1161.

van Wieringen, A., & Wouters, J. (1999). Natural vowel and consonant recognition by Laura cochlear implantees. *Ear and Hearing, 20*, 89-103.

Vogel, E., & Machizawa, M. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428*(6984), 748 - 751.

Vogel, E., McCollough, A., & Machizawa, M. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature, 438*(7067), 500 - 503.

Voor, J., & Miller, J. (1965). The effect of practice on the comprehension of worded speech. *Speech Monographs, 32*, 452-455.

Warren, J., Scott, S., Price, C., & Griffiths, T. (2006). Human brain mechanisms for the early analysis of voices. *NeuroImage, 31*, 1389-1397.

Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science, 167*(3917), 392-393.

Warren, R., & Warren, R. (1970). Auditory illusions and confusions. *Scientific American, 223*, 30-36.

Watson, C., Qiu, W., Chamberlain, M., & Li, X. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *Journal of the Acoustical Society of America, 100*(2), 1153 - 1162.

Wechsler, D. (1997). *WAIS-III - Administration and Scoring Manual.* San Antonio, TX: Psychological Corporation.

Wei, C., Cao, K., & Zeng, F. (2004). Mandarin tone recognition in cochlear-implant subjects. *Hearing Research, 197*, 87-95.

Weill, S. (2001). *Foreign accented speech: Adaptation and generalization.* Unpublished master's

thesis, Ohio State University.

Wichmann, F., & Hill, N. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293 - 1313.

Wichmann, F., & Hill, N. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*(8), 1314 - 1329.

Wilmer, J., & Nakayama, K. (2007). Two distinct visual motion mechanisms for smooth pursuit: Evidence from individual differences. *Neuron*, *54*(6), 987 - 1000.

Yovel, G. (2007). Faces and objects are processed by independent mechanisms: Evidence from Individual differences. In *Vision Sciences Society Annual Meeting.* Sarasota, FL.

Appendix A

Experiment 2 Questionnaire

Questionnaire

Native language (please also give the region where you learned to speak e.g. Northern Ireland, Yorkshire):

Do you speak any other languages?

Language	**Standard (e.g. GCSE, A-Level or equivalent)**	**Age when you started learning the language**

Are there certain aspects of these languages in which you are particularly strong (e.g. reading, writing, listening, speaking)? Please give separate answers for each language.

Are there certain aspects of these languages in which you are weak, or that you find more difficult (e.g. reading, writing, listening, speaking)? Please give separate answers for each language.

Do you play any musical instruments?

Instrument	Standard (indicate an equivalent Grade)

Can you sing a familiar tune without accompaniment?

Yes No Not sure

Can you sing a familiar tune without accompaniment?

Yes No Not sure

Can you easily recognize a familiar melody if it is played without accompaniment (e.g. whistled)?

Yes No Not sure

Can you clap the rhythm of a familiar melody without accompaniment?

Yes No Not sure

Do you find it easy to clap along accurately with the beat of a sing/piece of music?

Yes No Not sure

Would you be interested in taking part in a brain-scanning experiment? Yes/No

Appendix B

Variables in phonetic analysis of Experiment 2 test sentences

Total number of words
Total number of syllables
Number of mono-syllabic words
Number of bi-syllabic words
Number of tri-syllabic words
Number of four-syllable words
Number of five-syllable words

Number of voiceless plosives
Number of voiced plosives
Number of voiceless fricatives
Number of voiced fricatives
Total number of consonants

Number of front vowels
Number of central vowels
Number of back vowels
Number of close vowels
Number of mid vowels
Number of open vowels
Number of reduced vowels
Number of diphthongs
Total number of vowels

Number of 2-consonant clusters
Number of 3-consonant clusters

Sentence rhythm: $nPVI$
Sentence rhythm: $rPVInorm$

Appendix C

Goodness-of-fit statistics for psychometric functions

Table C.1: Deviance statistics for each curve fitted in Experiments 6 and 7. † indicates those deviance statistics that lie outside the 90% confidence limits generated through Monte Carlo simulations in psignifit.

Participant	Experiment 6		Experiment 7					
			Keywords			Sentences		
	Keywords	Sentences	Block 1	Block 2	Overall	Block 1	Block 2	Overall
1	21.61†	9.34	7.56	7.66	10.26	6.85	8.09	11.53
2	28.19†	6.27	11.29	11.08	5.15	10.13	5.90	8.19
3	4.18	4.82						
4	5.26†	2.83	5.30	5.80	3.70	9.93	3.88	4.47
5	13.54†	4.77						
6	4.27	7.31	16.33†	8.95	20.35†	5.91	5.06	9.90
7	9.60	8.86	8.63	10.22	9.29	6.78	4.82	4.87
8	11.27†	1.69						
9	7.56	10.42						
10	27.67†	6.04	5.05	9.99	6.18	4.62	4.07	4.93
11	9.84†	6.38	5.91	9.38	4.61	5.77	3.25	4.53
12	6.60	2.23						
13	10.44†	7.56	16.36†	7.20	18.41†	11.04	5.53	10.35
14	17.10†	5.85	9.29	8.45	8.89	5.35	8.39	5.74
15	13.98†	7.37	19.87†	9.02	5.38	15.37†	8.61	5.01
16	7.78	2.19	4.39	8.26	11.20	4.13	3.19	4.12
17	5.11	0.71	12.00†	3.05	5.27	1.80	7.23	3.97
18	2.29	2.42	7.28	29.56†	10.18	13.13†	6.09	4.73
19	6.95	2.54	11.04	25.33†	31.27†	7.49	10.33	13.28†
20	8.26	9.72†	14.58†	8.59	7.85	8.63	3.53	6.05
21	3.36	3.47	6.39	3.77	6.49	3.53	1.75	3.21
22	11.91	3.65						
23	6.39	0.93	9.73	5.95	7.46	9.93	6.67	5.68
24	15.80†	7.89	12.28	12.08	11.54	5.55	10.05	3.77
25	16.79†	4.22	20.85†	16.78†	20.96†	4.93	13.69	6.93
26	6.39	4.34	8.90	1.88	6.82	6.08	3.35	5.27
27	3.13	2.28						

Table C.2: Deviance statistics for each curve fitted in Experiment 8. † indicates those deviance statistics that lie outside the 90% confidence limits generated through Monte Carlo simulations in psignifit.

	BKB			IEEE			Words			Consonants			Vowels		
Participant	Session 1	Session 2	Overall	Session 1	Session 2	Overall	Session 1	Session 2	Overall	Session 1	Session 2	Overall	Session 1	Session 2	Overall
1	1.44	2.23	2.08	1.91	4.51	1.37	1.41	2.05	2.07	1.16	1.95	1.22	0.68	1.91	2.05
2	2.52	0.62	1.48	3.70	5.17	5.57	0.75	5.93	3.29	3.61	2.42	1.54	0.97	3.11	1.59
3	1.73	2.76	1.63	2.78	3.38	0.08	3.93	6.29	8.96†	2.03	1.77	2.06	4.16	1.16	4.37
4	2.24	1.31	3.35	8.41†	11.49†	17.98†	1.71	3.82	1.56	2.17	3.70	5.56	5.18	1.95	3.51
5	4.16	3.23	0.90	4.46	2.83	3.48	2.57	2.76	5.13	3.31	0.54	2.70	5.10	4.15	6.49
6	4.67	1.04	5.26	11.55†	3.80	19.20†	2.47	9.47†	10.63†	6.21	4.27	4.63	2.03	3.99	1.78
7	1.46	2.81	3.95	2.63	8.30	2.39	3.57	9.91†	5.85	1.79	1.54	2.59	1.34	4.62	0.99
8	3.16	1.37	3.88	2.13	3.78	2.94	3.71	1.73	4.45	2.42	3.59	3.35	7.16	1.92	0.82
9	0.75	0.63	0.98	2.03	3.95	7.14†	3.60	3.03	3.84	3.07	1.32	3.70	1.12	1.66	1.22
10	4.49	3.08	6.25†	3.68	2.58	2.08	3.50	1.63	3.27	0.08†	1.70	1.05	1.81	1.35	0.70
11	8.30†	2.68	1.17	9.90†	3.40	6.32	1.49	1.82	1.93	1.32	2.06	3.12	3.99	2.55	1.99
12	13.09†	4.73	12.97†	6.20	1.22	6.81	8.71†	4.48	9.83†	2.61	3.54	3.25	2.27	3.75	3.41
13	2.26	2.20	1.39	4.09	5.63	3.44	6.58	11.45†	12.64†	3.52	1.55	4.67	10.75†	5.77	14.24†
14	0.94	2.98	2.92	2.92	1.68	5.06	2.17	2.04	4.26	2.12	4.18	5.48	0.63	3.62	2.48
15	1.01	4.82	0.73	1.89	4.94	1.39	9.30†	2.04	6.26†	1.11	2.02	2.43	3.48	0.43	3.13
16	6.52	0.79	4.22	11.97†	0.90	9.95†	5.51	1.02	6.72†	1.31	1.35	1.70	4.97	2.57	3.64
17	1.67	2.73	4.21	1.89	9.36†	5.71	11.21†	3.31	16.38†	3.96	5.42	7.43	7.15	1.06	4.75
18	2.90	1.48	4.44	11.18†	4.05	20.40†	3.23	1.66	3.91	1.15	3.13	3.24	0.95	0.52	2.36
19	1.93	2.67	2.70	8.17†	4.60	2.26	4.40	8.09†	4.34	6.08	1.98	3.96	4.71	3.03	3.29
20	2.48	0.75	2.19	6.95†	2.83	7.34	2.42	5.92†	4.61†	5.28	2.95	2.89	0.72	8.64†	4.02
21	2.22	1.35	2.28	13.08†	4.54	5.28	1.74	0.56	0.70	2.00	1.86	2.62	5.58	3.50	7.32
22	10.32†	5.85	15.17†	2.69	5.13	0.38	3.84	1.74	2.97	2.53	1.35	1.54	1.28	2.43	2.10
23	4.58	1.02	2.89	3.69	2.59	2.90	0.85	5.67	5.05	3.00	2.29	2.60	2.47	1.04	2.27
24	9.11†	0.92	5.88	12.52†	6.86†	15.61†	2.63	1.58	3.32	3.70	3.17	5.37	2.41	4.56	2.68
25	0.87	4.67	2.92	3.41	3.13	1.10	1.74	1.75	0.45	4.57	0.79	4.38	4.74	2.78	3.11
26	5.06	2.88	7.89†	8.00†	2.12	2.37	3.02	1.69	1.15	2.76	2.23	3.73	2.90	2.13	4.92
27	0.58	0.21	0.60	8.14†	7.49†	12.53†	1.58	3.71	2.74	3.16	2.58	5.76	1.38	3.91	4.02
28	5.57	3.87	0.30	0.69	5.90†	2.85	1.73	2.27	3.09	4.02	3.19	6.29	3.24	6.84	7.76