Acoustic cue weighting strategy and the impact of training for cochlear implant users and normal hearing listeners with acoustic simulation

Yue Zhang

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

Speech Hearing and Phonetic Sciences

University College London

2017

I, Yue Zhang, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This project investigates the impact and plasticity of perceptual cue weighting strategy for normal hearing (NH) listeners with cochlear implant (CI) acoustic simulation and CI users. It is hypothesised that how listeners allocate perceptual attention on different speech cues is related to how accurately and effectively they can restore the phonemic structures from the acoustic inputs. Therefore, it can be beneficial to use auditory training to guide listeners' attention to the more reliable and informative cues for their own specific language, in order to improve speech recognition and ease listening effort.

The first part of this project investigated the impact of perceptual weighting strategy for both groups of listeners. Listeners' sentence recognition and pupillary response (taken as a measure of listening effort) were measured. They were then taken into the same model with listeners' acoustic cue weighting ratio and auditory sensitivity to explore their relation.

The second aim of this project was to examine the possibility of using auditory training to change listeners' acoustic cue weighting strategy towards an optimal one. A distributional training method was used here, with the sampling of spectral and temporal cues in the training word stimuli manip-

ulated in a way that only the spectral cue followed a bimodal distribution that resembled the natural speech. This was to increase the saliency of the spectral cue, in order to direct listeners' attention to the spectral cue. Sentence recognition performance in quiet, acoustic cue weighting strategy and pupillary responses were measured before and after the training to examine the effectiveness of the training.

Findings from this project will extend the current understanding on CI users' perceptual cue weighting strategy and provide inspiration for a more comprehensive rehabilitation scheme for CI users.

Acknowledgements

5 years ago, I left home for my first international flight to start a new journey; and 5 years later, I am wrapping up this journey in 210 pages. Although I cannot say that this journey is only happiness, I can say that it has changed me into a better person: more grateful, more resilient, more aware and more optimistic. But this acknowledgement is not about me, it's about all the people who have made my change and this dissertation possible.

I will not be able to say enough thanks to Stuart Rosen, for his undying support, patience and attention. Without him, I would have learned so much less and struggled so much more in both my research and life.

I would also like to thank other faculty members and researchers in Speech, Hearing and Phonetic Sciences (SHaPS): Paul Iverson, for his expertise and advice; Valerie Hazan, for her help on applications; Bronwen Evans, for her knowledge and help; Debi Vickers, for her valuable feedback; Tim Green, for his help on experiments; Andrew Clark, for his great patience.

And a special thank you to Emma Brint, Laurianne Cabrera, Axelle Calcus, Mauricio Figueroa Candia, Faith Chiu, Cristiane Hsu, Daniel Kennedy-Higgins, Hao Liu, Albert Lee, Gisela Tome Lourido, Tim Schoof, Jieun

Song, Kurt Steinmetzger, Helen Willis, Katharina Zenke, and Jose Joaquin Atria (who have made this dissertation much tidier and easier to write). It is my great honour to have your company and friendship.

For my parents who are behind my every step helping me in all possible ways, you deserve way more than I can ever give back and I love you way more than I can ever say.

Finally, thank you Yoann Ferrand, to accept me and love me for who I am. You have given me support, courage, and a vision of a happy future.

Writing this acknowledgement brings back many good memories, and makes me realise how lucky I have been. It is impossible to put all these emotions down to words. I will continue my journey, with all my passion and curiosity. This will be the best and only way to thank you for all your support and love.

Contents

List of Figures

List of Tables

Part 1

Introduction

Chapter 1

Introduction

Since their first introduction in the early 1980s, multi-channel cochlear implants (CI) have significantly improved the speech perception performance and life quality of profoundly and totally deaf patients (Budenz et al. 2011; Clark, Clark, & Furness 2013; Mosnier et al. 2015; Sladen et al. 2017). However, many challenges still remain. Current postoperative speech recognition measures typically report an average of approximately 70% for sentence recognition in quiet, and significantly impaired performance in noise (Firszt et al. 2004; Gifford et al. 2008; Di Nardo et al. 2010; Sladen & Zappler 2015). There is also substantial variability among individuals in both speech performance and life quality (Hirschfelder et al. 2008; Capretta & Moberly 2016a; Sladen et al. 2017). In order to maximise post-lingually deaf CI users' benefits from the device, a variety of rehabilitation strategies have been developed to provide users with intensive auditory training on the degraded and distorted auditory inputs through the CI. Typically, they focus on improving listeners' speech recognition performance. This might not be sufficient to characterise the difficulties experienced by CI

users. High levels of listening effort and fatigue are reported by hearing impaired (HI) listeners and CI users, and their ratings of life quality are not correlated with their speech recognition (McCoy et al. 2005; Bologna et al. 2013; Sladen & Zappler 2015; Capretta & Moberly 2016a; Sladen et al. 2017). Therefore, without a training program that aims to ease listening effort and a post-training assessment that measures the change in listening effort, there is no guarantee that CI listeners will have a more effortless speech communication after the rehabilitation, even if the speech recognition scores have improved.

This project intends to investigate the effect of an auditory training method that aims to utilise listeners' sensitivity to the statistical features of speech cues in order to adopt a better listening strategy. It is hypothesised that how listeners prioritise certain acoustic cues in speech is related to their speech recognition performance and listening effort. Arguably, allocating cognitive resources on the speech cue that is most informative for restoring phonemic structure and most reliable across different styles of speech should help listeners to map the acoustic inputs to their phonological representations better and quicker. Therefore, the training that encourages listeners to adopt this strategy should be beneficial both behaviourally and cognitively.

This chapter will firstly review past studies on normal hearing (NH) and post-lingually deaf CI listeners' acoustic cue weighting strategy and its possible relation with listeners' general speech performance, in order to illustrate the potential benefit of a certain weighting strategy. The sec-

ond section will look at the plasticity of listeners' acoustic cue weighting strategy and speech recognition, in order to explore the possibility of integrating the adjustment of the weighting strategy into the auditory training for CI users.

Finally, this chapter will provide an introduction to the project, explaining its rationale and potential contribution to the existing studies.

1.1 Acoustic cues and listeners' perceptual weighting strategy

1.1.1 Acoustic cues and perceptual weighting

To understand speech, listeners need to recover linguistic structure from an acoustic speech signal. Components of the acoustic signal relevant to phoneme identities are referred to as cues (Repp 1982). Acoustic cues can come from the spectral, temporal, or amplitude structure in the speech signal (Lisker & Abramson 1964). Using these cues to recover the linguistic structure is not a straightforward process since there is no one-to-one mapping from an acoustic input to a phonemic representation (Liberman et al. 1967). Lisker (1978) catalogued as many as 16 acoustic cues that could characterise the English plosive voicing distinction. Accordingly, listeners rarely use one acoustic cue to make phonemic decisions. Instead, multiple co-occurring cues in the acoustic inputs are used to identify speech sounds (Lisker 1978; Bailey & Summerfield 1980; Best et al. 1981).

Nevertheless, not all acoustic cues that contribute to phonemic identity are

perceptually equivalent to listeners in determining the perceptual category of a sound, a procedure referred to as cue weighting. Listeners' cue weighting strategy depends on a few factors.

Typically, different languages use different sets of acoustic cues for categorising speech sounds. This specialisation in cue weighting strategy takes a significant amount of time for listeners to develop. It could start as early as the first year after birth and continue through the first decade of life until the strategy matches that of adults. For instance, 4-month-old infants displayed better general voice-onset-time (VOT) boundary discrimination than native French speakers, and at 8 months they showed increased sensitivity to the VOT boundary in French (Hoonhorst et al. 2009). 3-year-old children weighted formant transitions more than fricative-noise spectra cue in categorising syllable-initial fricatives compared to adults, but at 6 years old shifted to more weighting to the noise spectra cue with no significant difference from adults (Nittrouer 2002). Once this developmental shift in perceptual attention is completed, the weighting strategy will be rather robust and dominate listeners' perception of acoustic cues, making the processing of foreign language sound patterns difficult (Bradlow et al. 1999; Iverson et al. 2003). Therefore, at least within a language community, listeners share the use of a similar set of acoustic cues and a similar pattern of allocating perceptual attention to acoustic cues, although with some individual variabilities (Hazan & Rosen 1991).

Furthermore, certain features of acoustic cues also explain listeners' perceptual weighting strategy. Firstly, listeners' perceptual sensitivity to an

acoustic cue could be constrained by the auditory processing system. The auditory encoding of an acoustic cue is not linearly related to its physical properties meaning that equal physical steps do not correspond to equivalent changes in percept (Kuhl & Miller 1975; Stevens 1989; Kuhl 1991). Therefore, changes in some acoustic dimensions could be perceptually more salient than others. For instance, onset asynchrony differences less than 20ms in both speech and nonspeech stimuli are not well resolved in human auditory system (Jusczyk et al. 1980; Sinex & McDonald 1989; Sinex et al. 1991; Simos & Molfese 1997). This auditory constraint might heavily influence listeners' categorical perception of voicing since it provides a cut off on the VOT continuum, making the VOT cue more salient than other acoustic cues in identifying voicing categories (Pisoni 1977; Holt et al. 2004). Secondly, acoustic cues also vary in their distributions and variances within- and between-category, which gives a different reliability of acoustic cues in delimiting speech categories (Holt & Lotto 2006; Toscano & McMurray 2010; Idemaru & Holt 2011). For instance, VOT values in American English do not overlap across voicing categories, making VOT cue a more robust and informative marker of category identity (Lisker & Abramson 1964; Keating 1984). As a result, listeners would allocate more importance to this cue to facilitate categorisation. Meanwhile, the large within-category variance might decrease the informativeness of an acoustic dimension, since it increases the likelihood of overlaps between categories for different speakers or speaking styles, and also introduce more uncertainties during the phonemic mapping. Typically for the English tense and lax

vowel contrast, the primary acoustic cue is the spectral shape and the secondary cue is the vowel duration (Lehiste & Peterson 1961; Hillenbrand et al. 1995; Watson & Harrington 1999; Hillenbrand et al. 2000; Leung et al. 2016). Compared to lax vowels, tense vowels are produced with more extreme articulatory target positions and slower articulatory movement into and away from the target positions. Therefore, they have a larger vowel space with more peripheral formant frequencies and less dynamic formant trajectories. Vowel duration is also an important difference between English tense and lax vowels since the articulators need longer time to reach the more extreme target positions of tense vowels. However, this difference in vowel duration is more impacted by the contextual factors such as consonantal context and speaking style than the spectral features, potentially making it a more variable and hence less reliable cue for the vowel identity (Nearey & Assmann 1986; Port 1981; Gopal 1990; Leung et al. 2016).

Also, listeners' perceptual weighting strategy is affected by the acoustic features of the speech input. Speech communication rarely happens in a quiet room and between speakers with the same accents. The acoustic and distributional features of speech cues could be significantly affected by signal degradations, speakers' physiological features, dialect, etc (Allen et al. 2003; McMurray & Jongman 2011; Babel & Munson 2014; Iverson et al. 2006; Winn et al. 2012). Therefore, listeners should also be able to detect the mismatch between the immediate acoustic inputs and the phonemic representations and adjust the use of different acoustic cues to maximise speech perception performance. Indeed, various studies have shown that

adult NH listeners with an established cue weighting strategy could still change their relative attention based on the acoustic integrity and distribution of speech cues. Normally in English, spectral cues are informative and robust as to vowel identity: listeners could identify vowels even with only the synthesised stable formant section or snapshots of the vowel's onset and offset (Jenkins et al. 1983; Nearey & Assmann 1986; Hillenbrand & Nearey 1999). However, after noise-vocoding, a technique for manipulating the level of spectral detail in the speech signal and simulating CI speech processors, spectral cues are significantly affected. Noise-vocoded speech is created by dividing the speech signal into frequency bands (analogous to the individual electrodes in a CI) and then applying the extracted amplitude envelope in each frequency range to band-limited noise. This procedure introduces great spectral resolution reduction and spectral smearing, so the informativeness and reliability of the spectral dimension are significantly compromised. Formant structure is less distinctive and salient through the transmission of between-band amplitude differences, and can even be represented in inconsistent and unnatural frequency regions if further spectral shifting is implemented to simulate relatively shallow CI electrode insertion. Dynamic formant movements are also replaced by the noise within the channel (Dorman et al. 1997; Fu & Shannon 1999; Rosen et al. 1999; Davis et al. 2005; Roberts et al. 2010; Zhou et al. 2010). Vowel duration, in comparison, is untouched by this manipulation. To accommodate this significant change in the input signals, NH listeners adjust their original cue weightings. For instance, without any degradation, listeners typically

weight the spectral cue heavier than the duration cue for vowel recognition. This is shown in Iverson et al. (2006) as a greater damage to the undegraded vowel recognition when removing vowel's formant movements. In comparison, when the vowel durations were equated across stimuli, there was less decrease in vowel recognition. This suggested that NH listeners were more affected by the damage on the spectral than the duration cue, hence, a greater perceptual weighting of the spectral cue. Similarly, Winn et al. (2012) showed that when performing a tense and lax undegraded vowel categorisation task, NH listeners altered their responses to a greater extent when stimuli were changing in formant structure and formant dynamics, compared to the duration. This suggested that listeners were relying more on the differences in the spectral dimension to perform vowel categorisation. With 8-band noise-vocoded speech, NH listeners typically decrease the perceptual weighting of the formant structure and dynamics cue. This is shown in Winn et al. (2012) as a decrease in the reliance on the formant structure and dynamics, and an increase in reliance on the vowel duration. This suggested that with the degradation of spectral cues, listeners relied less on them since they gave less information about vowel identity. The duration cue retained the same probabilistic relation with the vowel identity, therefore, listeners increased the importance of the duration cue in making the phonemic decision. Note that this adjustment couldn't compensate fully the spectral degradation, since vowel recognition decreased in general. But rather, it used the remaining categorical differences between tense and lax vowels to maximise performance in the task. Iverson et al. (2006) showed

no significant difference in the impact of formant dynamics and duration manipulations on vowel recognition after 8-band noise-vocoding. But when more spectral degradation was applied (4-band and 2-band), listeners didn't show significant changes in vowel recognition in the condition with formant dynamics eliminated, suggesting that listeners decreased their use of the spectral cue possibly because the degradation left little information in the spectral dimension. They also showed no change when vowel duration was equated, suggesting that NH listeners didn't increase the use of vowel duration when less spectral information was available. This difference with Winn et al. (2012) in the use of the duration cue might be due to the different nature of the task. Vowel length is not a phonemic contrast for English vowels, therefore, it might not contain much information to vowel identities but still remain useful for telling apart tense and lax vowels. While listeners could pay no attention to the duration cue in a multiple choice recognition task without compromising the performance, in a two-choice categorisation task everything that could help to tell apart the two tokens will have much more perceptual importance. This also illustrates the importance of task requirements when investigating acoustic cue weighting, since the same acoustic dimension in the same signal may be more heavily weighted in one perceptual task but less in another (Holt & Lotto 2006).

This plasticity of NH listeners' cue weighting can also be found when the distribution of speech cues was manipulated. For instance, when the distribution of the VOT cue that differentiates word-initial voicing was manipulated to have smaller variance, NH listeners showed steeper cate-

gorisation function slopes, compared to the condition with larger variance (Clayards et al. 2008a). It seemed that listeners were able to both perceive the distributional change in the speech input and adjust their use of the VOT cue in categorising the new acoustic inputs. In a more extreme case, Idemaru and Holt (2011); Lehet and Holt (2017) created an artificial accent with the distribution of the Fo cue reversed in its correspondence to the voicing and VOT cues. Along with VOT, Fo is used by listeners as an acoustic cue to voicing: typically voiced consonants have lower Fos than the voiceless consonants (Kohler 1982, 1984). In the artificial accent, low Fo was coupled with longer VOT and high Fo was coupled with shorter VOT for the 'beer' - 'deer' and 'pier' - 'tier' word contrast. After a short period of exposure, listeners showed less difference in voiceless responses between low and high Fo, suggesting that they used the Fo cue less to categorise voicing since its new distribution was unfamiliar. However, unlike previous studies, listeners were not able to incorporate the new distribution of Fo cues by associating low Fos with voiceless consonants and high Fos with voiced consonants, even after 5 days of training. This shows that although NH listeners are sensitive and flexible to the short-term changes in the acoustic signal, this plasticity couldn't overwrite the long-term acoustic cue encoding and cue weighting. Considering that the VOT distribution tends to have less within-category variation due to the auditory perceptual discontinuity mentioned above, Liu and Holt (2015) created a similar artificial accent for tense and lax vowels in American English. Tense vowels were coupled with short vowel durations and lax vowels were coupled with long vowel du-

rations. Similarly, NH listeners down-weighted the use of vowel duration for vowel categorisation after a short period of exposure, suggesting that even for more variant speech categories, listeners still maintain similar sensitivity to the statistical features of the acoustic cues. Note that this line of studies manipulated the non-primary cue, and used the primary cue to guide the word recognition and adjust the weighting of the non-primary cue. It is likely that since the probabilistic relation between the primary acoustic cue and speech category is intact, there is enough information to guarantee reliable identification and the best strategy would be to ignore other insignificant and noisy cues. However, if the acoustic integrity and distribution of the primary cue are damaged, finding a coping strategy might be more complex and requires more than just acoustic inputs. Top-down lexical tuning might be necessary to firstly re-establish the speech category, and the weighting on both cues might need to be adjusted based on the amount of distinctive information left and their probabilistic relation with the speech category.

1.1.2 Cue weighting strategies of post-lingually deaf CI users

Modern cochlear implants have multiple electrodes that are inserted into the scala tympani to try to recreate the normal tonotopic distribution of information along the cochlea. Cochlear implant speech processors typically filter speech signals into multiple frequency bands, and within each band the relatively slowly varying temporal envelope is extracted and is used to modulate a high-rate train of electrical pulses. The processed outputs are

then presented to electrodes spaced along the tonotopic locations along the cochlear.

Different acoustic cues are affected by this process, but compared to the temporal and amplitude cues, the spectral cues are significantly damaged in both the quantity and quality of the information they contained for speech recognition. Although the slowly varying temporal envelope cannot transmit spectral fine structure (>500 Hz), it still contains important linguistic information like the manner of articulation, voicing, periodicity and durational contrasts (Rosen 1992; Drullman 1995; Shannon et al. 1995). Studies manipulating the temporal structure of speech (envelope low-pass filtering, pulse rate, temporal reversal) also showed that temporal modulation over 20 Hz has little effect on speech recognition (Drullman et al. 1994; Shannon et al. 1995; Arai & Greenberg 1998; Fu & Shannon 2000; Fu & Galvin III 2001). Similarly, loudness mapping procedures in the CI device reduces the bigger acoustic amplitude range into the smaller electrical range of the electrodes, limiting the number of amplitude steps in the speech signal. Although an important cue to the syllable structure, amplitude structure has little correlation with the linguistic structure of speech, and the reduction of amplitude steps only affects phoneme recognition in quiet when there are less than 8 levels (Zeng & Galvin III 1999; Loizou et al. 2000). In comparison, spectral cues are informative and robust for speech perception, as mentioned above. However, CI users have only limited access to the detailed spectral structure in the signal, due to the limited number of effective spectral channels and interactions among those channels (Chatterjee & Shannon 1998; Friesen et

al. 2001). Meanwhile, incomplete insertion of the electrodes presents spectral information to the wrong populations of auditory nerve fibres, with a shift of spectral information to nerves that typically carry higher-frequency information. Studies show that speech recognition is significantly damaged if the shift is more than about 3mm, or if the frequency-to-place mapping is linearly expanded or compressed by more than 3mm (Dorman et al. 1997; Fu & Shannon 1999; Baskent & Shannon 2003). This spectral distortion along with the reduction of spectral resolution makes spectral cues contain less distinctive information about the linguistic structure, and distorts the correspondence between the spectral cues and the phoneme categories.

Note that the acoustic integrity is not the only factor that could constrain post-lingually deaf CI users' speech processing ability. While NH listeners might have similar auditory and cognitive abilities, CI users are more variable. For instance, the duration of deafness is correlated with the extent of peripheral neural degeneration (Nadol & Eddington 2006). So for CI listeners who experience a longer period of deafness and suffer more spiral ganglion cell degenerations, the electrical stimulation from the device might not even reach the brain. Some listeners might have 'dead regions ' on the cochlear where there are no functioning neurons, and the electric currents delivered to those regions would spread to the neighbouring neurons, causing more distortions to the tonotopic representations (Shannon et al. 2002; Moore 2004). And even if all the peripheral auditory system is relatively intact, functional cortical auditory networks might be re-wired after a long period of deafness. Due to progressive hearing loss, listeners' phonological

representations deteriorate with the lack of auditory inputs. For instance, listeners employ different strategies when performing a rhyme matching task prior to the implantation. After the implantation, listeners using the phonological information performed better, suggesting that preserving the phonological structure of the language helps listeners to regain phonological sensitivity with electric stimulation (Lazard et al. 2010).

Considering all these factors, it should not be a surprise if NH listeners and CI users have different sensitivity to acoustic cues or employ cues differently for speech processing. For instance, Dorman et al. (1991) compared CI and NH listeners' discrimination response to the place of articulation of stops with three cues: formant transitions, spectrum tilt at signal onset and the abruptness of spectral change. Using words with three steps on each dimension, they showed that NH listeners' responses were more significantly affected by the step change in formant transition than that in the spectrum tilt and abruptness, but CI listeners were more affected by the change in spectrum tilt and abruptness. Hedrick and Carney (1997) manipulated the formant onset amplitude and the vowel spectrum for /sɑ/ - /ʃɑ/ and /pɑ/ - /tɑ/ labelling. Comparison of d-prime values for each continuum showed that CI listeners were more sensitive to the amplitude cue. A ratio between two d-prime values was interpreted as the degree of cue integration, and the comparison between CI and NH listeners showed that CI listeners had less equivalent weighting on both cues. For vowel perception, Kirk et al. (1992) showed that adding the formant transition cue was less beneficial to vowel recognition than adding the vowel centres for CI users. This was in-

terpreted as listeners' lack of access to the dynamic spectral cue, although no auditory test on the spectral information was performed. Donaldson et al. (2013) also manipulated vowels' dynamic and stable spectral cues in a /dVd/ syllable, by attenuating the entire vowel centre to silence or deleting the vowel transition section (both with the vowel duration fixed). While NH listeners' vowel recognition performance was not affected by the loss of the dynamic cue in the absence of the duration cue, CI listeners' performance was significantly impaired, suggesting that both cues were weighted more strongly by CI users. Donaldson et al. (2015) added the vowel duration cue to the stimulus set, making six types of /dVd/ syllables that varied in vowel centre, formant transitions and duration. When the vowel centre was preserved, fixing duration led to a significant drop in performance for CI users but not NH listeners, suggesting the importance of the duration cue for CI users even when the formants were intact. When the formant transitions were preserved, fixing the vowel duration had no impact on either group, possibly because listeners' couldn't perceive the silence between the abrupt vowel edges as the vowel.

While these studies typically investigate listeners' use of acoustic cues by constructing stimuli that have one dimension fixed or eliminated, Winn et al. (2012) adopted a different approach by introducing gradient steps on each acoustic dimension investigated. It provided the possibility to investigate listeners' perceptual sensitivity to changes in acoustic cues and to quantify this sensitivity using a statistical model that was sensitive to individual variability. Winn et al. (2012) constructed multiple steps on the acoustic

dimensions between the tense - lax ('hit' 'heat') vowel and word-final voicing ('loss' 'laws') and combined them orthogonally. Listeners' categorical responses were then fit by mixed effect models using the acoustic steps. Results showed that in both contrasts, in comparison to NH listeners, CI listeners' response function for spectral cues was flatter, but their response function for the vowel duration cue was steeper. This suggested that CI listeners used the changes in the spectral dimension for speech categorisation to a lesser degree than NH listeners, but CI listeners used the changes in the temporal dimension to a greater degree than NH listeners. Note that all the previously mentioned studies didn't specifically assess CI users' spectral resolution on the acoustic stimuli, making it impossible to interpret the results entirely as the effect of listeners' perceptual weighting. It is possible that some listeners might not have sufficient spectral resolution to detect variation in the spectral dimension to use it for making further phonemic decisions. Therefore, both the minor impact of eliminating the spectral cue or flatter response function observed in these studies might be due to listeners' limited access to the spectral information.

This possibility was investigated in Moberly et al. (2014), where CI participants were asked to both label and discriminate synthesised stimuli as either /bɑ/ or /wɑ/. The syllable-initial stop and glide have similar onset and steady-state formant frequencies, since the articulatory gestures involved are essentially the same: lips are initially closed and then opened with the tongue reaching the target position of the vowel. But for the stop /b/, it takes less time to reach the steady-state values and the peak amplitude. The

time to reach the target formant frequencies was termed as the formant rise time (FRT) cue, and the time to reach the peak amplitude was termed as the rise time (ART) cue. Previous studies with NH listeners have shown that NH native English speakers use FRT as the primary cue (Nittrouer & Studdert-Kennedy 1986; Nittrouer et al. 2013). CI listeners were required to label four continua of syllables, with one of the cues fixed to the value of a typical /b/ or /w/ and another cue varying in multiple steps. Their responses were fitted with logistic regression and the coefficients of the acoustic cues were used as their perceptual weighting factors. Results showed that CI listeners in general had smaller weighting factors on the FRT cue and bigger weighting factors on the ART cue. They also conducted an AX discrimination task using stimuli with the same cue values but with formants replaced by sine waves. These stimuli were used as non-speech controls, and were intended to measure listeners' absolute sensitivity to the acoustic properties, without the influence of phonemic knowledge. Regression analysis found no significant relations between ART and FRT's cue weighting factors and the d-prime values of the sine-wave counterparts. Similar results were reported using other phoneme contrasts: word-final /p/ and /b/ that differ in vocalic duration (temporal cue) and syllable-final formant transition (dynamic spectral cue), and /s/ and /ʃ/ that differ in fricative-noise spectrum (static spectral cue) and syllable-initial formant transition (dynamic spectral cue) (Moberly et al. 2016). The cues' weighting factors were also not significantly related to the d-prime values for the sine-wave counterparts, for both CI and NH listeners. These studies seem to suggest that listeners'

decision on which acoustic cue to rely on for speech processing does not depend only on the auditory saliency of the acoustic cues through the CI device. Other factors might also play a role, for instance, their perceptual weighting strategy prior to the hearing loss. Moberly et al. (2016) also reported that half of the CI participants could not complete the spectral discrimination or labelling task, suggesting that a large number of CI users only have limited access to the spectral features of the acoustic inputs.It also highlights the importance to pre-screen listeners' auditory ability and take it into account when conducting perceptual weighting experiments for CI users. However, using sine wave speech as a non-speech control in both studies might have an unpredictable effect on the accuracy of auditory sensitivity measurements. It is still unclear whether CI listeners perceive a similar non-speech sound quality as NH listeners from the sine wave speech. While reducing the speech signals to only two or three frequency modulated sine waves destroys the speech naturalness for NH listeners, it might decrease the amount of smearing between electrodes by stimulating only a small number of electrodes and transmit better the spectral shape and movement for CI users.

Although there is little doubt that adult post-lingually deaf CI users and NH listeners have access to different acoustic cues, whether they weight differently the cues they have access to is less clear. Due to the hearing loss later in life, post-lingually CI users should have enough language exposures to develop a highly specific and robust perceptual weighting strategy like NH listeners. After implantation, listeners could either use the acoustic cues

that are immediately available, or apply the mature perceptual weighting strategies acquired for the specific language prior to the deafness. As mentioned above, Winn et al. (2012) demonstrated a simultaneous change in the weighting of the spectral and temporal cues, but it was unclear whether CI listeners adopted a different relative weighting strategy by switching to duration as the primary speech cue. This question was also not answered in Moberly et al. (2014, 2016), mostly due to the lack of a measurement that describes the relative reliance on acoustic cues. Lowenstein and Nittrouer (2015) demonstrated the importance of looking at such a measurement, by using the data from Nittrouer et al. (2013) that compared the coefficients of the FRT and ART cues in the /bɑ/ /wɑ/ contrast for 4, 5-year-olds and adults. While it seems that 4- and 5-year-olds weighted the FRT cue less than adults compared to the ART cue, a ratio of the two coefficients shows no significant difference between two groups, suggesting that the proportion of total perceptual weight allocated to each cue is similar across age groups. This method is also used in other speech perception studies (Giezen et al. 2010; McMurray & Jongman 2011; Moberly et al. 2014). Arguably, this relative measurement is relevant when comparing CI and NH groups, since similarly, they don't have the same amount of access to the acoustic cues. Looking only at their absolute weighting of speech cues might be insufficient to display the difference in the weighting strategy between two groups. Note that this cue weighting ratio involves measurements from different acoustic dimensions. It only provides information on listeners' relative reliance on different cues, while listeners' absolute use of acoustic cues is not included.

Therefore, both the weighting of acoustic cues and the ratio are informative and should be included in the analysis. Meanwhile, the much bigger variability in the cue weighting for CI than NH listeners also makes it difficult to compare two groups statistically. Typically in these studies, some CI listeners might have similar weightings as NH listeners, and some might have different weightings.

1.1.3 Perceptual weighting strategy and speech perception

This great variability in CI listeners' acoustic cue weighting has been found to be related to their speech recognition performance. Even with an older CI system, Kirk et al. (1992) has shown that CI listeners' open-set word recognition was positively correlated with their vowel recognition performance in the condition with only the dynamic spectral cue, but not with the condition with only the vowel centre. But as commented above, since there was no assessment of listeners' auditory abilities, this correlation might only suggest the relation between listeners' spectral resolution and word recognition. More controlled recent studies showed a similar pattern. Moberly et al. (2014) showed that CI listeners' word recognition performance was correlated with the weighting factor on the FRT cue, but not with the weighting factor of the ART cue or auditory sensitivity to the non-speech counterparts. The same relations were reported in Lowenstein and Nittrouer (2015), for both NH adults and 8- and 10-year olds using an eight-band noise vocoder on the same stimuli. For the word-final /p/ - /b/ contrast used in Moberly et al. (2016), the weighting factor of the dynamic spectral cue was positively related to listeners' word recognition, and again

no such correlation for the weighting factor of the duration cue or to auditory sensitivity to the non-speech stimuli. For the /s/ - /ʃ/ contrast, no significant correlation was found, possibly due to the small number of CI listeners passing the pre-screen requirement. Winn and Litovsky (2015) also reported a similar relation. CI listeners' word recognition score was significantly correlated with their weighting of the formant cue in a /ba/-/da/ categorisation task, but not with their weighting of the spectral tilt cue. However, although in these studies listeners' auditory acuities were measured, they were not taken into the same model with acoustic cue weightings in explaining the variability in word recognition. Therefore, it is unclear whether the perceptual weightings would still remain significant after taking away the part of the variability in word recognition explained by listeners' differences in auditory sensitivity.

Nevertheless, all these studies shared a similar pattern in that CI listeners who had a similar cue weighting strategy as NH listeners had better word recognition. This might at first sound counter-intuitive. NH listeners typically weight the spectral cue more than the duration cue in these contrasts, but since CI listeners only have limited access to the spectral dimension, it does not sound beneficial to put most perceptual attention on a compromised cue. However, considering that the weighting strategy of a typical NH listeners is developed over years and highly specific to their native language, it should allocate more importance on acoustic cues that are most useful and robust in that language. Therefore, it is likely that even when the informativeness of this primary cue is compromised, it might still

be more reliably associated with speech categories than other cues. Putting more perceptual attention on this damaged primary cue will not yield the same level of performance as NH listeners, but it will give the best possible speech performance level with the remaining information.

In summary, listeners' perceptual weighting of acoustic cues is a result of both a long-term and language-specific perceptual weighting strategy, and the features of the perceived speech. Post-lingually deaf CI users will have developed an optimal cue weighting strategy prior to the hearing loss, but this strategy is altered after the implantation due to the differences in the electric hearing. However, the benefits of this weighting strategy for speech recognition might still be preserved for some listeners, even though the access to the primary speech cue in some phonemic contrasts has been significantly compromised. Therefore, there might potentially be a benefit to train listeners to re-tune to this optimal strategy for the specific language, in order to utilise the most useful and informative speech cues to maximise speech performance.

1.2 Rehabilitation for CI users

1.2.1 Speech training for CI users

Most CI users take 6 months or more to make sense of the new signals via the implants, and for some listeners this process of adaptation requires more than daily passive listening (Manrique et al. 1997; Dillon et al. 2013). A variety of active and intensive rehabilitation strategies have been devel-

oped to improve listeners' speech performance.

A number of studies used CI acoustic simulations to train NH listeners' degraded speech perception. Typically, noise-vocoded speech and noise-vocoded speech with spectral shifting were used. They present different challenges to NH listeners for adaptation. Although both methods reduce the spectral resolution of the speech signal, spectral shifting involves spectral distortion that introduces phonological mismatch between the acoustic inputs and listeners' phonemic representations. This mismatch requires listeners to re-establish the phonological association, and therefore is found in past studies to take longer time to adapt to compared to noise-vocoded speech (Dorman et al. 1997; Skinner et al. 2002; Finley & Skinner 2008; Di Nardo et al. 2010).

Rosen et al. (1999) trained NH listeners with 4 band and 6.46mm shifted noise-vocoded speech for over ten sessions (about three hours), with a highly interactive audio-visual Connected Discourse Tracking (CDT) (De Filippo & Scott 1978) procedure. Both the talker and listeners were engaged during the training: they faced each other through a glass partition in two adjacent sound-proofed rooms and worked together to maximize the correct level of listeners' verbal repetitions of speech materials. Listeners' were always presented with the processed discourses from the talker, while the talker received undistorted feedback from the listener. After the training, listeners' speech test scores (BKB sentences, vowel and consonant recognition) improved significantly, but only the performance of consonants reached a similar level with the unshifted noise-vocoded speech at the

end of the training. A detailed analysis of consonant features showed that only the place of articulation perception was still significantly worse with shifted than unshifted speech after ten sessions. In terms of the speed of improvements, listeners improved faster in sentence recognition. A later study also showed that a similar computer-based training with pre-recorded materials was similarly effective (Faulkner et al. 2012). This seemed to suggest that speech materials with rich contextual information were most effective in helping listeners to recover from the spectral distortions and degradations, but only to a certain level even with intensive training.

Also training NH listeners but with a different CI acoustic simulation (8 band 8mm shifted sine-vocoded speech), Fu, Nogaki, and Galvin III (2005) compared the change in vowel categorisation of four different training groups: group 1 received no training; group 2 were trained with the same vowels as in the testing but spoken by different talkers; group 3 were trained with different sets of medial vowels through an adaptive scheme, in which listeners started from the easiest vowel contrasts to smaller vowel differences as they improved; group 4 were trained with isolated sentences. Vowel categorisation showed no significant improvement after the training for group 1 and, surprisingly, group 4, and a larger as well as faster improvement for group 2 than group 3. It seemed that vowel recognition benefited most from the multi-talker repetitions.Note that the post-training speech performance was only measured in phoneme categorisation. Therefore, it is unclear whether listeners will improve similarly in the recognition of sentences that are more realistic speech materials in daily communication.

Stacey and Summerfield (2008) compared the effects of another three training schemes on sentence and phoneme categorisation (with the 8 band and 6mm shifted noise-vocoded speech). The word training group was delivered as two-alternative word choosing tasks; the sentence training group was similar to group 4 in Fu, Nogaki, and Galvin III (2005) of using pre-recorded isolated sentences, although presented in a specific distorted-clear-distorted sequence; the phonetic training group was delivered with single vowel and CV syllable discrimination tasks starting from pairs with easy phonemic contrasts to harder ones. For all groups, the training started only after participants reached an asymptotic level of performance with test materials and was then run for ten sessions. There was no significant effect of phonetic training on sentence recognition, and no significant effect of sentence and phonetic training on consonant and vowel categorisation. This was interpreted as demonstrating the importance of lexical labels for training, since the most ineffective phonetic training condition provided no information on the lexical level. Although perceptual learning experiments using noise-vocoded words (Davis et al. 2005) and spectrally-shifted vowels (Li & Fu 2007) also demonstrated the importance of a lexical level feedback, after controlling for the differences in working memory, Hervais-Adelman et al. (2008) reported similar results for lexical and phonological feedback. The worse performance of the phonetic training condition in Stacey and Summerfield (2008) might also be due to the difference between the training and testing stimuli. The vowels in the phonetic training were only present in two phonetic contexts (isolated and word-final) and consonants only at

the beginning of the word, while vowels and consonants in the testing were in the middle of a word. Similarly, Dahan and Mead (2010) reported no generalisation between onset and coda of noise-vocoded words, suggesting the importance of including phonetic variants in training.

The training studies above using CI acoustic simulations experimented with different training methods and provided important insights into how NH listeners adapt to the degraded and distorted speech signals. The studies above show that using training materials with more contextual and lexical information, and using an interactive procedure with intensive training sessions seem to benefit listeners sentence recognition and generalisation after the training. However, for CI users, the amount of speech degradation and distortion is much more severe, and listeners' hearing, cognitive and demographic conditions are more variable. Therefore, it is important to examine whether CI listeners are also responsive to these training methods.

Fu, Galvin, et al. (2005) trained ten CI users with an adaptive method similar to group 3 in Fu, Nogaki, and Galvin III (2005). Their pre-training baseline performances were measured multiple times over a 1- or 2-week period, so that listeners were fully familiar with the training procedure and stimuli. Based on their baseline levels, poorer-performing listeners were firstly trained with an adaptive 3-forced-choice phoneme discrimination task, with the acoustic differences between phonemes (F1, F2 and duration values for vowels; voicing, manner and place of articulation for consonants) increased with the wrong response and decreased with the correct response. As listeners improved on the discrimination, they were trained with the

phoneme categorisation task, and with increasing number of word choices as they improved (up to 6 words). All together, they were trained 1 hour per day, 5 days per week, and a retest every two weeks with adjustments to the training program to target the worst performing tasks. After the training, CI listeners' vowel, consonant and sentence recognition score improved significantly. Listeners were also different in terms of the speed of improvement.

Wu et al. (2007) also trained Mandarin-speaking CI users using a similar adaptive procedure, and saw a significant improvement in vowel, consonant and tone recognition. This seemed to suggest that a training procedure that was tailored to individual performances and improvements could benefit CI users even after they've reached their asymptotic speech levels. Stacey et al. (2010) performed the same word and sentence training as in Stacey and Summerfield (2008) on one group of CI users who varied in age, speech processors and speech performance for three weeks, five days per week and one hour per day. The results showed significant improvements in IEEE (but not BKB) sentence recognition and consonant (but not vowel) categorisation after the training. A large variability was also observed between each training session. Compared to the Stacey and Summerfield (2008), there was little evidence that CI listeners benefited as much as NH listeners from the training, possibly due to the need of a longer training time and a greater between-session variability.

Miller et al. (2016) trained nine postlingually deaf CI users with a /ba/, /wa/, /da/ and /ja/ identification task. Training stimuli were spoken by eight na-

tive American English speakers and more speakers were only added if the identification correct rate was over 90%. Pre- and post-training assessments were the same set of stimuli, but spoken by another four speakers. After the training, listeners' phoneme categorisation scores improved significantly, even for phonemes produced by different talkers. But similarly, there was a large variability in both listeners' rate and level of improvement. Note that although stimuli were chosen based on their spectral and temporal differences, no adaptive scaffolding was applied to the acoustic differences between phonemes, but only on the number of speakers. Also, only four sets of phonemes were used in the training and testing, so it was unsure whether the speaker generalisation observed in the results was applicable for bigger speech units.

Fu and Galvin (2008) reported a pilot study, which trained listeners word and sentence recognition in multi-talker babble noise adaptively: if listeners failed to reach a certain level, the noise level decreased and vice versa. Both auditory and visual feedback were provided, and the SNR level was adjusted in each trial based on listeners' previous responses. After the training, listeners' phoneme categorisation performance increased in both steady speech-shaped and multi-talker babble noise, but only the listener trained with sentences had an improvement in a post-training sentence recognition task. Although with only two CI users, this experiment compared the effect of training on speech recognition in different types of noise. This could be specifically important for CI users, since they had significantly worse speech performance in background noise.

Ingvalson et al. (2013) trained five CI users' word, phrase and sentence recognition in multi-talker babble noise adaptively. Their pre- and post-training performance were assessed with QuickSIN and HINT sentences in noise. Results showed an improvement in the noise tolerance level for both sentence types. Even when training with digits in noise, Oba et al. (2011) showed significant improvement in ten CI users on digit recognition in speech-shaped noise, HINT in multi-talker babble and IEEE sentences in multi-talker babble for moderately difficult SNRs. They also maintained these gains at 1-month post-training. But no improvement in HINT recognition in speech-shaped noise or in IEEE sentence recognition in multi-talker babble at extremely difficult SNRs was observed. Note that in these three studies, no control group and no randomisation on multi-talker babble were included in the design. This makes it difficult to assess the real effect of the training, since it is likely that listeners could learn the pattern of noise after some exposures and improve by expecting the right 'dips' in the signals.

With more CI users, Schumann et al. (2015) had fifteen CI users in the training group and twelve in the control group. The training group received VCV and CVC syllables in a closed-set identification task format as training. If the subject responded incorrectly, the next target syllable would be more different in the voice gender, speech rate and sound type, so the task would be easier. No training was provided for the control group. After the training, improvements of sentence recognition in the moderate noise condition (5 dB SNR) were significantly greater for the training group than for the control group, but not in the condition with a difficult noise level (0 dB SNR).

There was also an improvement in syllable recognition in quiet after the training, but the size of improvement depended on the actual syllable.

In general, intensive speech training for CI listeners seems to be able to improve their speech perception abilities, even after their initial adaptation period. However, due to the small number of CI participants, limited stimulus randomisation, no use of control groups in the design, and lack of comprehensive assessment post-training, it remains unclear whether the training effect observed was due to a real change in listeners' speech perception ability, or familiarisation to the test procedure or materials. Meanwhile, although high listening effort and fatigue during speech communication is reported by CI users, no study yet has assessed the impact of training on this aspect (Alhanbali et al. 2017).

Listening effort was defined in Pichora-Fuller et al. (2016) as the 'deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out listening tasks'. This concept is based on a limited-capacity resource model in which ongoing cognitive operations engage a given percentage of total cognitive capacity (Kahneman 1973; Rudner 2016). High task demands or low input signal quality, for instance, background noise or the use of a CI, will require additional resources to be allocated to maintain successful speech processing. This reallocation of resources is perceived subjectively as an increased effort and can produce decrements on other cognitive activities (Pichora-Fuller et al. 2016; Rudner 2016). For CI users, degradation of input signals and deficiencies in speech processing and cognitive abilities all require extra cognitive resources for understanding speech. Ide-

ally, a rehabilitation scheme should enhance CI listeners' abilities to extract meaning from the speech signals. In that case, speech recognition would require proportionally less cognitive resources afterwards, freeing up cognitive 'space' for other mental activities. Studies have shown that CI listeners' quality of life ranking improved after aural rehabilitation, and this trend is not significantly related to listeners' speech recognition levels (Capretta & Moberly 2016a; Harris et al. 2016). Intuitively, a subjective ranking of life quality should be related to the effort involved in speech communication, but previous studies with NH and HI listeners have shown that subjective scales couldn't reliably predict fatigue and don't correlate with objective measurements (Zekveld et al. 2011, 2013; Alhanbali et al. 2017). Therefore, a direct and objective measurement of listening effort would be useful, for instance, pupillometry, which has been used in many NH and HI studies. The pupillary response has been associated with many cognitive processes, and is used as an index for cognitive processing load in different types of tasks, for instance short-term memory (Kahneman & Beatty 1966; Peavler 1974), mental arithmetic (Hess & Polt 1964; Bradshaw 1968), physical (Zénon et al. 2014), near-threshold stimuli perception (Hakerem & Sutton 1966) and attentional tasks (Hillyard et al. 1973; Parasuraman 1979). As the task becomes more demanding, a larger pupil dilation will be evoked. Typically in the auditory tasks, when a task requires more cognitive resources in the same time interval, for instance with lower signal-to-noise ratios (SNRs), divided attention and spectral degradation, mean pupil dilation is larger (Zekveld et al. 2011; Koelewijn et al. 2012; Zekveld & Kramer 2014; Koelewijn

et al. 2015; Winn et al. 2015). The maximum dilation in that time window can also be taken as an index of the maximum processing load, and peak latency as an index of processing time (Zekveld et al. 2011). Therefore, comparing the pupillary responses during listeners' sentence recognition task before and after the training will provide us insights into whether CI listeners also benefit from the training to use the limited amount of cognitive resources more efficiently for speech recognition. Arguably, it should be used alongside speech recognition scores as an assessment of the efficacy of CI training, since both the level and the efficiency of speech perception are important outside the laboratory in the daily communication and for real CI users. A rehabilitation scheme that enables CI users to have higher speech recognition scores is promising, but if it comes at the expense of a greater listening effort after training, that might be more detrimental to the CI users' daily life.

1.2.2 Training acoustic cue weighting strategy

There is also no study yet to train CI listeners' cue weighting strategy. As reviewed above, training materials used either high-level linguistic units like sentences or words and guided listeners' learning with their rich lexical and contextual information, or minimal pairs to attract listeners attention to the acoustic details of phonemes. The potential problem with using the first type of material is that it is unsure how exactly listeners improved and what they changed to achieve better speech recognition. And the potential problem with the second is that stimuli typically change in many different dimensions, making it unclear to listeners which acoustic feature to track.

As reviewed in the first section, listeners' acoustic cue weighting strategy could be related to listeners' speech perception abilities and possibly processing speed (Tillman et al. 2017). Arguably, if CI users' cue weighting strategy could be trained towards a better one that allocates more perceptual importance on the most informative and robust speech cues, their speech perception performance would improve after the training. Also, the training could decrease their listening effort, since listeners would invest most cognitive resources to the most reliable features of the language and restores the phoneme structure from acoustic inputs more effectively.

One training method that utilises listeners' plasticity in acoustic cue weighting strategy by manipulating the distributions of speech cues is termed distributional training. It has seen some success in adjusting NH adult listeners' acoustic cue weighting for better identification of speech categories in a foreign language. Typically in distributional training, listeners hear a randomly presented series of stimuli that vary in steps along an acoustic dimension. Each stimulus on the continuum is presented with a certain frequency, such that some values along the acoustic dimension appear more often than others. In this way, listeners hear a certain distribution of speech sounds. For instance, in a bimodal distribution of speech cues, listeners will hear stimuli with values near the means of the two modes more frequently than the stimuli with marginal values. In a uniform distribution, listeners will hear stimuli with values near the means of the modes as frequently as other values. After the training, the listeners exposed to the bimodal distribution should have better recognition score

along the acoustic dimension than those listeners exposed to the uniform distribution, although the same set of stimuli appears in both distributions. The bimodal distribution could induce the perception of the two modes as exemplars of two different speech sound categories, while the uniform distribution couldn't facilitate the formation of speech categories with no 'clusterings' along the acoustic dimension.

For example, Gulian et al. (2007) exposed native Bulgarian speakers to bimodal distributions of the Dutch vowel contrasts /ɑ/–/a:/ and /ɪ/–/i/, which these listeners tend to perceive as the single Bulgarian vowels /a/ and /i/ respectively. After the training, listeners exposed to a bimodal distribution classified the vowels in each contrast more accurately than before the training. Escudero et al. (2011); Wanrooij and Boersma (2013); Wanrooij et al. (2013) presented Spanish-speaking listeners with bimodal distributions of Dutch /ɑ/–/a:/. The two Dutch vowels differ both in their formant structure (/a:/ has higher first and second formants) and duration (/a:/ is longer). Dutch native-speaking listeners rely primarily on the formant structure, while Spanish listeners rely heavily on the durational differences when discriminating the two vowels due to the lack of a similar phonemic contrast in their native language (Escudero et al. 2009). To train Spanish listeners to weight more on the formant structure, these experiments presented listeners with an /ɑ/ - /a:/ continuum that had formant structure in a bimodal distribution and duration fixed. After the training, listeners presented with the bimodal distribution showed better vowel categorisation, and this improvement was larger than in the control condition, where listeners were

exposed to a unimodal distribution of the formant structure. The training effect was also found to last six and twelve months after the training session (Escudero & Williams 2014).

To some extent, CI users face a similar problem as non-native listeners. With the signal degradation and distortion, the acoustic inputs are 'foreign' to listeners with ambiguous and shifted phoneme boundaries. And with the temporal cues intact, listeners might increase the relative weighting to the duration cue and down weight the more informative spectral cue, similar to a non-native listener facing the phonemic pattern of a new language. To facilitate the shift of perceptual attention to the spectral cue for speech processing, listeners need to be trained with stimuli that have an accentuated spectral feature. Therefore, manipulating the distribution of the spectral cue might be effective in drawing listeners' attention to it and encourage listeners to adopt the best listening strategy for a specific language.

1.3 The experiments

This dissertation intends to explore the impact of CI users' acoustic cue weighting strategy on their speech recognition performance and listening effort. It is hypothesised that CI users with a similar cue weighting strategy to NH listeners will have better sentence recognition and less listening effort. The weighting strategy shared by NH listeners within the same language community is developed over years and is robust, suggesting that it might be the optimal strategy to process acoustic cues for speech percep-

tion by allocating more attention to the most informative and reliable cues. Although CI listeners don't have access to all the information in the speech cues, relying on the most informative cue could still bring some benefits to both the accuracy and speed in mapping from acoustic inputs to their phonemic representations. The first two chapters examined this hypothesis. Chapter 2 firstly investigates the impact of different degrees of spectral degradation, distortion and background noise to acoustic cue weighting in a /bɪt/ - /bit/ contrast for NH listeners with CI acoustic simulations. Listeners' spectral and temporal cue weighting strategy was obtained by modelling listeners' vowel categorisation with a logistic regression model, and was calculated as a ratio between the two model coefficients. In this way, listeners' weighting strategy was better quantified and comparable across listeners. CI acoustic simulation only provides a coarse reproduction of CI speech processing, without considering other auditory and cognitive differences between CI users and NH listeners. Therefore, chapter 3 measured both NH and CI listeners sentence recognition and listening effort, and explored how listeners' acoustic cue weighting strategy affect this relation. For CI users, their auditory sensitivity was also measured and taken into account, in order to clarify the relation between listeners' auditory sensitivity and perceptual attention.

The second interest of this dissertation is to investigate whether CI listeners' acoustic cue weighting strategy can be altered by manipulating the distribution of speech cues. If the results of Chapter three suggest that a certain strategy is better for the accuracy and efficiency of speech processing,

then training listeners to adopt this strategy should benefit their sentence recognition and ease their listening effort. Chapter 4 experimented with NH listeners using CI simulations to see whether adult NH listeners still retain the sensitivity to the statistical features of the degraded speech, since previous studies typically used undegraded speech. Chapter 5 trained NH and CI listeners with either the stimulus set that has a distribution of the speech cue that accentuates the spectral aspect, or the stimuli set that accentuates the duration aspect. Before and after the training, listeners' acoustic cue weighting strategy, sentence recognition and listening effort were measured, to investigate the impact of auditory training on these aspects.

In summary, this dissertation will explore the impact and plasticity of listeners' acoustic cue weighting strategy. It will extend the previous studies on this topic, by relating listeners' cue weighting pattern to their general speech perception and cognitive effort, and exploring possible applications for CI users. It will also provide some insights into the methodology of conducting speech and cognitive experiments with a population as variable as CI users.

Part 2

Acoustic cue weighting strategy and its impact

Chapter 2

Listeners' acoustic cue weighting strategy in different spectrally degraded conditions

This chapter investigates normal hearing listeners' (NH) perceptual weighting between temporal and spectral cues using tense/lax vowel categorisation tasks for spectrally shifted noise-vocoded speech. Listeners were tested with different numbers of bands and degree of spectral shifting in Experiment 1, and in different signal-to-noise ratios (SNRs) in Experiment 2. This was to simulate the wide range of signal degradation and distortion perceived by cochlear implant (CI) users and in degraded listening environments.

2.1 Experiment 1

2.1.1 Methods

Participants

Participants were 10 native Southern British English speaking adults recruited via the UCL Psychology Pool, all aged between 18 and 45. Each participant was assessed before the experiment using pure tone audiometry and all had normal hearing, defined as hearing thresholds of 20dB HL or better between 250-8000 Hz at octave frequencies. None of them had extensive exposure to vocoded speech. For their contributions, they were paid at the rate of 7 pounds per hour. All participants read through the information sheet and signed the consent form.

Stimuli

Stimuli were noise-vocoded monosyllabic words, varying orthogonally in vowel duration and formant structure. A male native Southern British English speaker was recorded reading a randomised list of words ('bit', 'beat', 'sit', 'seat', 'pit', 'peat', 'fit', 'feat'), with each word repeated 30 times and all sampled at the rate of 44.1 kHz. The F1 and F2 of both the tense and lax vowels contexts were measured at 50% into the vowel using Praat (Boersma & Weenink 2009) and were converted to the equivalent rectangular bandwidth (ERB) scale to approximate the frequency spacing in the human auditory system (Moore & Glasberg 1983; Glasberg & Moore 1990). The ratio between F2 and F1 was then calculated and the distribution of the ratio was fitted with a custom distribution that was the sum of two Gaussian distri-

Context	$\bar{x}_1$	σ_1	$\bar{x}_2$	σ_2
bVt	2.09	0.01	3.12	0.07
sVt	2.01	0.04	3.10	0.07
pVt	2.02	0.08	3.09	0.11
fVt	2.04	0.04	3.16	0.11

Table 2.1: Means and standard deviations for the bimodal distribution fitted to the F2/F1 ratio of the recorded monosyllable words of each consonant context. The mean corresponds to the F2/F1 ratio of the most typical word in that category.

butions with equal weights. Parameters of each distribution are listed in table 2.1. Based on the F2/F1 ratio distributions, two endpoint values were selected for each continuum that were two standard deviations away from each mean of the bimodal Gaussian distributions. In this way, these two values represented the most typical tense and lax tokens respectively in each context. This process of endpoint selection is illustrated in figure 2.1 for 'bit' - 'beat' continuum. Vowel durations of the recorded words were fitted with a linear regression as a function of F2/F1 ratio. Details of each linear fitting are listed in table 2.2. The duration and fundamental frequency of two endpoint tokens in each context were then matched by PSOLA (Valbret, Moulines, & Tubach 1992) in Praat. 150 tokens of each continuum were synthesised using the Tandem-STRAIGHT algorithm (Kawahara et al. 2008) in Matlab (MATLAB 2013) and the resulting F2/F1 ratios were measured in

Context	Model	Fitting
bVt	$duration = 0.01 \times F2/F1 + 0.07$	$F(1,90) = 185.19$, $p < 0.01$
sVt	$duration = 0.02 \times F2/F1 + 0.07$	$F(1,80) = 70.18$, $p < 0.01$
pVt	$duration = 0.01 \times F2/F1 + 0.07$	$F(1,90) = 117.01$, $p < 0.01$
fVt	$duration = 0.02 \times F2/F1 + 0.06$	$F(1,80) = 97.93$, $p < 0.01$

Table 2.2: Model parameters and fittings for linear regressions with vowel duration as the dependent variable and F2/F1 ratio as the independent variable.

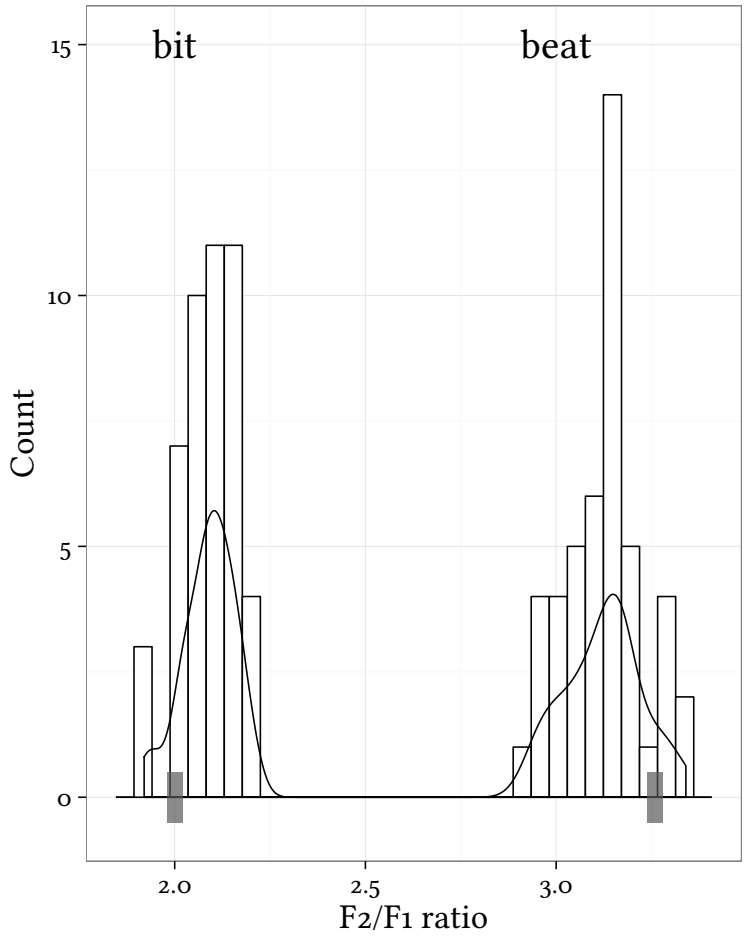

Figure 2.1: An example of endpoint value selection for 'bit' - 'beat' continuum: 1) The F2/F1 ratio was calculated for each recorded words 'beat' and 'bit', and their frequency was plotted; 2) A bimodal Gaussian distribution was fitted to the frequency of the ratio and the density function was plotted; 3) One token on the 'bit' end that was two SDs away from the mean of 'bit' category ($2.09 - 2 \times 0.01 = 2.07$), and one token on the 'beat' end that was two SDs away from the mean of 'beat' category ($3.12 + 2 \times 0.07 = 3.14$) were selected as most typical 'bit' and 'beat' for further processing (shaded in the figure).

the same way as above. Six equal steps in F2/F1 ratio were then selected from these synthesised tokens between the two endpoint values obtained from recorded speech and their corresponding durations were calculated based on the linear regressions in table 2.2. To construct an orthogonal design, each F2/F1 step was then cross-paired with each step in duration using PSOLA, thereby giving altogether $6 \times 6 \times 4 = 144$ stimuli.

Noise-vocoding of stimuli was performed in MATLAB software. Sound files were digitally filtered into six, eight and twelve channels with sixth-order Butterworth infinite impulse response filters. Filter spacing was based on equal basilar membrane distance (Greenwood 1990) across a frequency range of 70-10000 Hz. The output of each band was then half-wave rectified and low-pass filtered at 30 Hz (fourth-order Butterworth) to extract the amplitude envelope. The envelope was then multiplied by a wide band noise, and each filtered by a sixth-order Butterworth output filter. The root mean square (rms) level of the output signal of each channel was adjusted to match the original, and the signals were added together. For unshifted conditions, analysis and output filters had the same centre frequencies. For shifted conditions, cross-over and center frequencies for both the analysis and output filters in shifted conditions were calculated using an equation (and its inverse) relating position on the basilar membrane to characteristic frequency, assuming a basilar membrane length of 35 mm (Greenwood 1990):

$$frequency = 165.4(10^{0.06x} - 1) \tag{2.1}$$

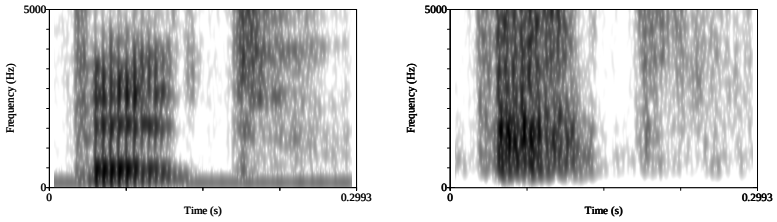

Figure 2.2: Spectrograms for the word 'bit' unprocessed (left) and 6 bands noise-vocoded and 6mm shifted (right).

Output filters then had their centre frequency increased upward by respectively 2mm, 4mm and 6mm on the basilar membrane distance. All stimuli were scaled to equal rms intensity in Praat. In sum, there were thirteen conditions (twelve processed and one unprocessed), each different in the band number and degree of spectral shifting. An example is shown in figure 2.2 to illustrate the impact of noise vocoding and spectral shifting.

Procedure

Experiments were conducted in a quiet room. Auditory materials were presented over Sennheiser HD 25 SP headphones and programs were run on a PC installed with custom MATLAB 2013b software. Participants first received a 15 minutes training session with unprocessed stimuli to familiarise them with the software interface. During the training, they listened to sentences and then were given written and acoustic feedback. No active responses were required. They were then tested with thirteen conditions in a randomised order. For each trial within each condition, participants were presented acoustically with a token selected randomly from all stimuli and visually two words on the computer screen, one containing a tense vowel

and the other a lax vowel in the same context. They were instructed to click on the word they had heard and the program then continued to the next token without feedback until the end of all conditions. Their responses were recorded by the program.

2.1.2 Statistical analysis and results

A measure of cue weighting for each individual was calculated from the binomial response of participants in the word labeling task. In R, logistic regressions were fitted to the proportion of tense vowel responses for each participant, with steps in formant structure and vowel duration as independent variables (no significant interaction was found). Regression coefficients reflected the proportional change in labeling preference with each step change on the continuum, hence could be used as perceptual weighting factors for the spectral and temporal cue. Dividing the coefficient describing sensitivity to changes in formant structure by the coefficient describing sensitivity to changes in duration (after exponentiation) gave a ratio expressing listeners' relative perceptual reliance on the two cues. Therefore, a higher ratio indicates more reliance on the spectral cue relative to the temporal cue; and a lower ratio indicates more reliance on the temporal cue relative to the spectral cue. Previous studies suggested that this relative measurement was more robust and comparable across heterogeneous listeners (Giezen et al. 2010; Nittrouer et al. 2013; Lowenstein & Nittrouer 2015). For instance, listeners might differ in the exact amount of perceptual attention on cues, possibly due to reduced auditory saliency

from a degraded signal input or immature language system, but might still share a similar strategy in distributing attention across cues. Listeners' averaged response functions for each conditions are displayed in figure 2.3. This cue weighting ratio was used in all further studies and analysis as an indication of listeners' temporal and spectral cue weighting strategy.

A boxplot of the cue weighting ratio across different conditions is shown in figure 2.4. To investigate how the number of bands and degree of shifting affects listeners' relative weighting on vowel formant structure and duration cues, a mixed effects model was built using the lme4 package in R (Bates et al. 2014) with cue weighting ratios of the processed conditions as the dependent variable. Mixed effect models allow for controlling the variance associated with random factors without data aggregation. Therefore, by using listener as a random effect in the model, we controlled for the variability in listeners' tense vowel responses (random intercept) and in other fixed factors (random slope) that were associated with them. Factors were entered into the model in the sequence below: the model firstly started with taking *listener* as the random intercept; fixed effect factors *band*, *spectral shifting* and their interaction were then entered; finally, random slopes were entered into the model. Factors were retained in the model only if they significantly improved the model fitting, using Chi-squared tests based on changes in deviance. Details of the best model explaining the variances in cue weighting ratio are shown in table 2.3.

Number of bands (χ^2=23.77, df=2, p<0.01), degree of spectral shifitng (χ^2=34.80, df=3, p<0.01) and their interaction (χ^2=46.079, df=6, p<0.01)

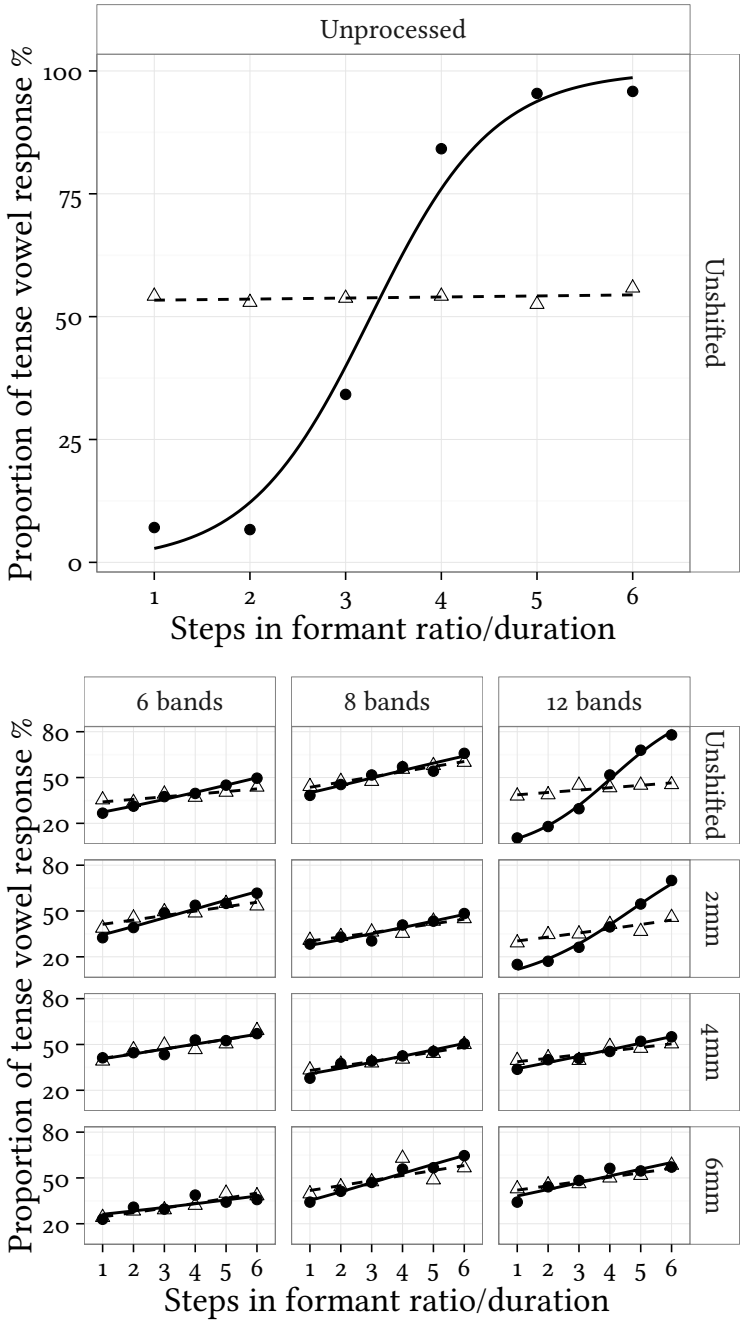

Figure 2.3: The top panel shows the proportion of listeners' tense vowel responses in the unprocessed condition, and the bottom panel shows their proportional responses in the processed conditions. The filled circle (●) is the averaged proportion response for each step in formant structure, and the hollow triangle (△) is the averaged proportion for each step in duration. The filled line (–) is the logistic regression fit to the proportion of listeners' tense vowel responses using steps in formant structure, and the broken line (- -) is the logistic regression using steps in duration.

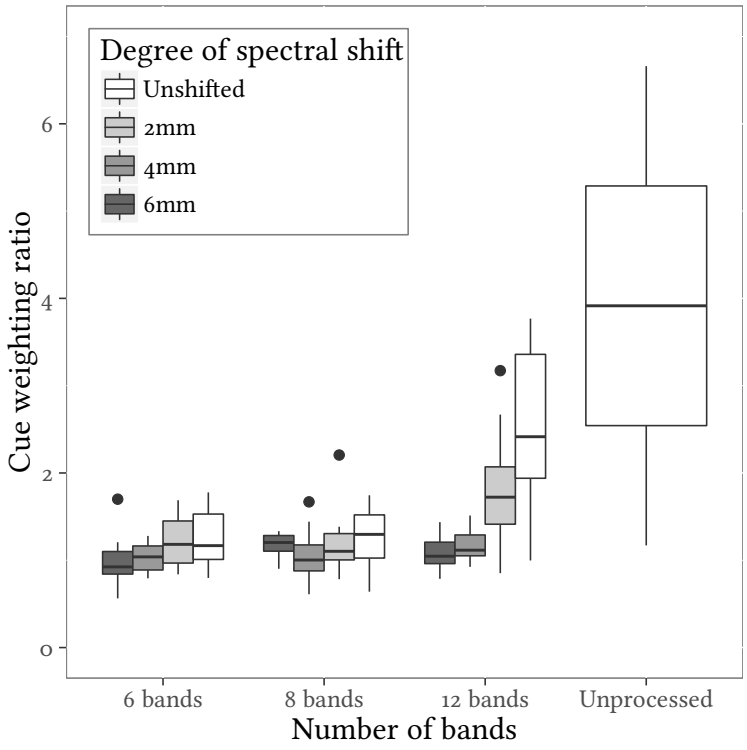

Figure 2.4: Boxplot of listeners' spectral and temporal cue weighting ratio for different noise-vocoding and spectral shifting conditions.

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	1.25	0.13	**<0.05**			
band(8)	-0.01	0.16	>0.05			
band(12)	1.27	0.16	**<0.05**	23.78	2	**<0.05**
shift(2)	-0.03	0.16	>0.05			
shift(4)	-0.22	0.16	>0.05			
shift(6)	-0.27	0.16	>0.05	34.80	3	**<0.05**
band(8):shift(2)	-0.001	0.23	>0.05			
band(12):shift(2)	-0.67	0.23	**<0.05**			
band(8):shift(4)	0.03	0.23	>0.05			
band(12):shift(4)	-1.14	0.23	**<0.05**			
band(8):shift(6)	0.19	0.23	>0.05			
band(12):shift(6)	-1.19	0.23	**<0.05**	46.08	6	**<0.05**
Random effects:	SD	cor		χ^2	df	p
Intercept | listener	0.22			166.96	1	**<0.001**

Table 2.3: Model parameter estimates and model comparison statistics for the best mixed effect model fit to the acoustic cue weighting ratio. The reference level for the categorical factor band is '6 ', and for shift is 'unshifted '.

were found to be significant. In general, listeners' cue weighting ratios were significantly larger in conditions with 12 bands than in 6 bands (β=0.52, SE=0.08, t=6.37, p<0.05) and 8 bands (β=0.48, SE=0.08,t=5.81, p<0.05), but there was no significant difference between 6 bands and 8 bands (β=0.05, SE=0.08, t=0.56, p>0.05). This suggests that listeners allocate more perceptual weighting to the spectral cue than the temporal cue when the spectral resolution is higher, but only when there is sufficient spectral resolution (more than 8 bands). In conditions with spectral shifting, listeners' cue weighting ratios were significantly larger in unshifted and 2mm shifted conditions than in 4mm and 6mm shifted conditions. Cue weighting ratios were also significantly larger in unshifted conditions than in 2mm shifted conditions (β=0.25, SE=0.09, t=2.67, p<0.05), but there was no significant difference between 4mm and 6mm shifted conditions (β=0.01, SE=0.09, t=0.12, p>0.05). This suggests that listeners allocate more perceptual weighting to the spectral cue than the temporal cue when the spectral distortion is smaller, but this effect is insignificant when the spectral distortion is too damaging (over 2mm).

Post-hoc Wald tests showed that for unshifted conditions, there was no significant difference between 6 bands and 8 bands (β=0.001, SE=0.16,t=0.06, p>0.05). But for 12 bands, the cue weighting ratio was significant larger than in 6 bands (β=1.27, SE=0.16, t=7.77, p<0.05) and 8 bands conditions(β=1.28, SE=0.16, t=7.83, p<0.05). The same relation was shown for conditions with 2mm spectral shifting: no significant difference between 6 bands and 8 bands (β=-0.01, SE=0.16,t=-0.07,p>0.05), but a significant larger ratio in 12

bands than in 6 bands (β=0.61, SE=0.16, t=3.72, p<0.05) and 8 bands conditions (β=0.60, SE=0.16, t=3.65, p>0.05). However, for conditions with 4mm and 6mm spectral shifting, there was no difference in acoustic cue weighting ratio among the different number of bands. This seems to suggest that listeners tend to allocate more perceptual weighting to the spectral cue than the temporal cue when the spectral degradation and distortion on the stimuli are smaller. However, once past the threshold (8 bands and 2mm shifting), there are no differences in listeners' weighting strategies.

To investigate whether the change in the weighting ratio is driven mainly by the change in weighting to the spectral or the temporal cue, another two mixed effect models were built with coefficients of the formant structure steps and duration steps as dependent variables respectively. Factors were entered into the model in the same way.

Coefficients of the formant structure showed the same pattern as the weighting ratio across conditions (details of the best fitting model are shown in table 2.4). Coefficients were larger for conditions with smaller spectral degradation and distortion, but there were generally no differences when the number of bands was less than 8 and the shifting was more than 2mm. No fixed effect factors were found significant for duration coefficients. This suggests that listeners maintain the same perceptual weighting to the duration cue, since it is not damaged across different conditions.

To compare listeners' spectral and temporal cue weighting strategy in unprocessed and degraded conditions, a mixed effect linear model was fit to the cue weighting ratio, with *listener* as a random intercept and *unpro-*

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	1.34	0.14	**<0.05**			
band(8)	0.08	0.17	>0.05			
band(12)	1.51	0.17	**<0.05**	25.99	2	**<0.05**
shift(2)	0.04	0.17	>0.05			
shift(4)	-0.15	0.17	>0.05			
shift(6)	-0.15	0.17	>0.05	32.92	3	**<0.05**
band(8):shift(2)	0.01	0.24	>0.05			
band(12):shift(2)	-0.82	0.24	**<0.05**			
band(8):shift(4)	-0.06	0.24	>0.05			
band(12):shift(4)	-1.33	0.24	**<0.05**			
band(8):shift(6)	0.10	0.24	>0.05			
band(12):shift(6)	-1.48	0.24	**<0.05**	55.00	6	**<0.05**
Random effects:	SD	cor		χ^2	df	p
Intercept | listener	0.21			52.40	1	**<0.001**

Table 2.4: Model parameter estimates and model comparison statistics for the best mixed effect model fit to the coefficients of formant structure. The reference level for the categorical factor *band* is 6, and the reference level for *shifting* is 0.

cessed/degraded as the independent factor. The cue weighting ratio in the unprocessed condition is significantly larger than the ratio in degraded conditions (β=61.71, SE=6.01, p<0.01), suggesting that listeners put significantly more weight on the spectral cue when the speech sound is not degraded. Another two identical mixed effect linear models were fit to the coefficients of formant structure and duration, showing significant larger formant structure coefficients in unprocessed conditions than in degraded conditions (χ^2=77.18, df=1, p<0.001, β=72.71, SE=7.10, p<0.001), but no significant difference for duration coefficients (χ^2=0.78, df=1, p>0.05).

Across all conditions, correlation analysis between the coefficients of formant structure and duration was only significant in conditions with 8 bands 4mm shifting (r=-0.67, p<0.05) and 8 bands 6mm shifting (r=0.75, p<0.05).

2.2 Experiment 2

2.2.1 Methods

Participants

Participants were 19 different native Southern British English speaking adults recruited via the UCL Psychology Pool, all aged between 18 and 45. Each participant was assessed before the experiment using pure tone audiometry and all had normal hearing, defined as hearing thresholds of 20dB HL or better between 250-8000 Hz at octave frequencies. None of them had extensive exposure to vocoded speech. For their contributions, they were paid at the rate of 7 pounds per hour. All participants read through the information sheet and signed the consent form.

Stimuli

Sentence stimuli were Basic English Lexicon (BEL) sentences (Calandruccio & Smiljanic 2012), recorded at a sampling frequency of 22.05 kHz from a male native Southern British English speaker. Each of these sentences contained 4 keywords, upon which the scoring was based.

A 'beat' - 'bit' word continuum was created with a Klatt synthesiser (Klatt 1980), with a sampling rate of 22.05 kHz. Formant values of the two endpoints were based on best exemplars from previous perceptual studies (Evans & Iverson 2004; Iverson et al. 2006). Formant steps were then interpolated using the ERB scale to approximate the bandwidths of the auditory filters in human hearing. Vowel durations were modelled from a report by House (1961) and duration steps were interpolated linearly. Altogether,

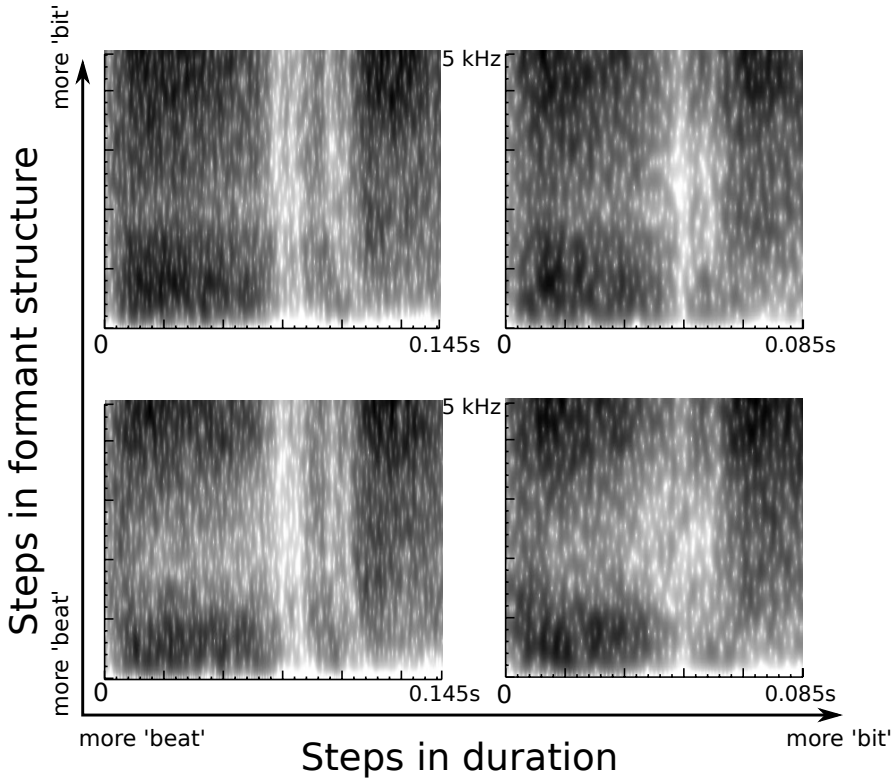

Figure 2.5: Spectrograms of the continuum endpoints of the word stimuli in the labelling task. Tokens varied in F1 F2 formant structure and vowel duration orthogonally. Each step in formant structure was paired with each step in vowel duration. The left bottom token had the most typical /i/ formant structure and duration; right bottom had the most typical /i/ formant structure and /ɪ/ vowel duration; left top had the most typical /ɪ/ formant structure and /i/ duration; right top had the most typical /ɪ/ formant structure and duration. All other stimuli varied in equal steps in between.

there were six steps in formant structure (with the first two formants varying simultaneously) and six steps in vowel duration.

Each step in formant structure was paired with each step in vowel duration, making $6 \times 6 = 36$ tokens. Exact values of all tokens are listed in table 2.5 and spectrograms of example tokens are shown in figure 2.5 . Training materials consisted of two short stories: Aesop's *The north wind and the sun* and *The wolf and goat*, recorded from a male native Southern British English speaker, ten BEL sentences in quiet and ten BEL sentences

	Step Number					
	1	2	3	4	5	6
F1 (Hz)	210	239	269	302	336	372
F2 (Hz)	2707	2604	2505	2410	2318	2230
Duration (s)	0.145	0.133	0.121	0.109	0.097	0.085

Table 2.5: Levels of vowel formant structure and duration for the 'beat' to 'bit' continua. Fo of all stimuli is 140 Hz, and F3 is 3200 Hz.

masked by speech-shaped noise at an SNR level of 10dB.

All stimuli were then 8-band noise-vocoded and 4mm upward shifted in MATLAB 2013b. This was to allow for enough spectral resolution for word recognition and to match the performance of successful CI listeners while reducing the auditory saliency of spectral cues by systematic distortion with over 3mm spectral mismatch, as shown in Dorman et al. (1997); Friesen et al. (2001); Faulkner et al. (2003) and the previous experiment. Each sound file was digitally filtered into 8 bands, using sixth-order Butterworth infinite impulse response filters. Filter spacing was based on equal basilar membrane distance (Greenwood 1990) across a frequency range of 70-6000 Hz. The output of each band was full-wave rectified and low-pass filtered (fourth-order Butterworth) at 30 Hz, using a zero-phase forwards-backwards technique to extract the amplitude envelope. The envelope was then multiplied by a wide band noise carrier. The resulting signal was passed through 8 output filters with their cut-off frequencies shifted upwards from the analysis filters by 4mm on the basilar membrane distance according to the Greenwood map. The rms level of the output signal intensity was adjusted to match the original analysis band and the signal was summed across all bands.

Procedure

Participants were seated in a quiet room. All audio stimuli were presented binaurally through Sennheiser HD25 headphones at a comfortable level. Experiments were run in Matlab 2013b using custom software.

Before the testing, participants were familiarised to the degraded speech through a 15 minutes training. During the training, they listened to sentences and then were given written and acoustic feedback. No active responses were required.

Considering the great variability in listeners' degraded speech recognition performance, an adaptive procedure was used to find a suitable SNR level that was of similar difficulty to each listener. This was intended to control for the confound of intelligibility. Listeners were firstly tested with 30 randomly chosen BEL sentences in quiet. In each trial, they listened to a sentence and were scored by the experimenter based on the number of keywords correctly reported. Their percentage correct was averaged across trials to obtain their speech recognition score. This was followed by an adaptive speech perception threshold tests (Plomp & Mimpen 1979) using speech-shaped noise to track 50% of each individuals' speech recognition score. Each SRT test consisted of 15 randomly selected sentences.

Finally, participants' cue weighting strategies were measured with a word labeling task. After hearing a word token, they were instructed to choose what they heard on the screen from either 'beat' or 'bit'. Their response was recorded by the computer and the next trial started. Their weighting

strategies were tested in four different conditions: in quiet (*quiet*), at the SRT level obtained from each individual (*SNR50%*), and 3dB over and below that level (*SNR50%+3, SNR50%-3*). The order of testing was randomised for each listener.

2.2.2 Statistical analysis and results

The mean SNR for condition SNR50% is 9.68dB (σ=5.78dB).

In each SNR condition (SNR50%, SNR50%+3, SNR50%-3, noNoise), the cue weighting ratio for each listener was calculated from the binomial response of participants in the word labelling task, identical to the method in Experiment 1. A boxplot of the cue weighting ratio across different SNR conditions is shown in figure 2.6

To investigate the impact of noise on the relative perceptual weighting between the spectral and temporal cues, a mixed effect model was built with the cue weighting ratio as the dependent variable, condition as independent factor and listener as the random intercept. No significant difference was found across SNR conditions (χ^2=1.10, df=3, p>0.05). Another two mixed effects models were built with the coefficients of the vowel formant structure and duration as dependent variables respectively, in order to look at whether noise had an impact on listeners' reliance on each cue. No significant difference was found for the spectral (χ^2=2.69, df=3, p>0.05) and the temporal cue (χ^2=3.44, df=3, p>0.05).

A significant correlation between the coefficients of the vowel formant structure and duration was found in quiet (r=0.75, p<0.001) and SNR50%+3

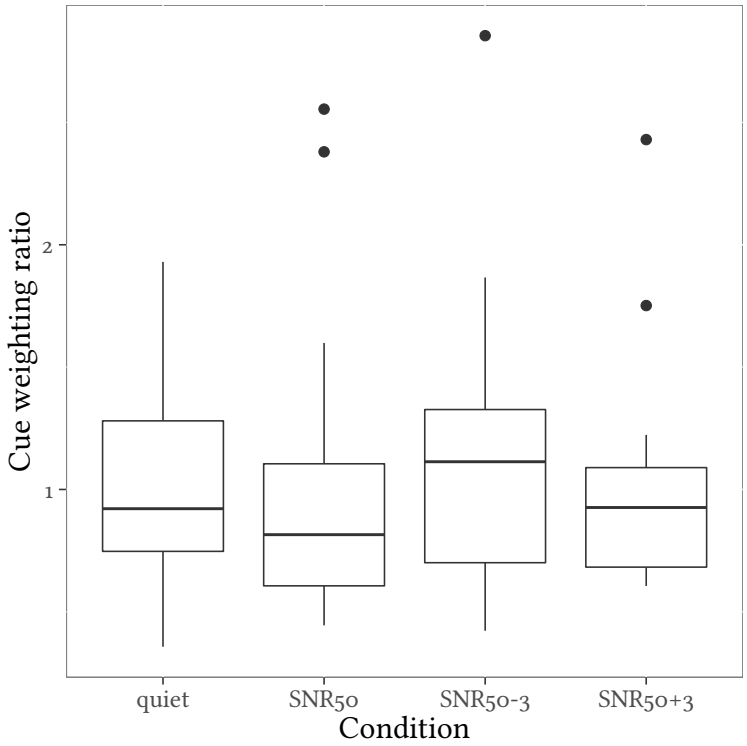

Figure 2.6: Boxplot of listeners' spectral and temporal cue weighting ratio in 'beat' - 'bit' word categorisation task in different SNR levels.

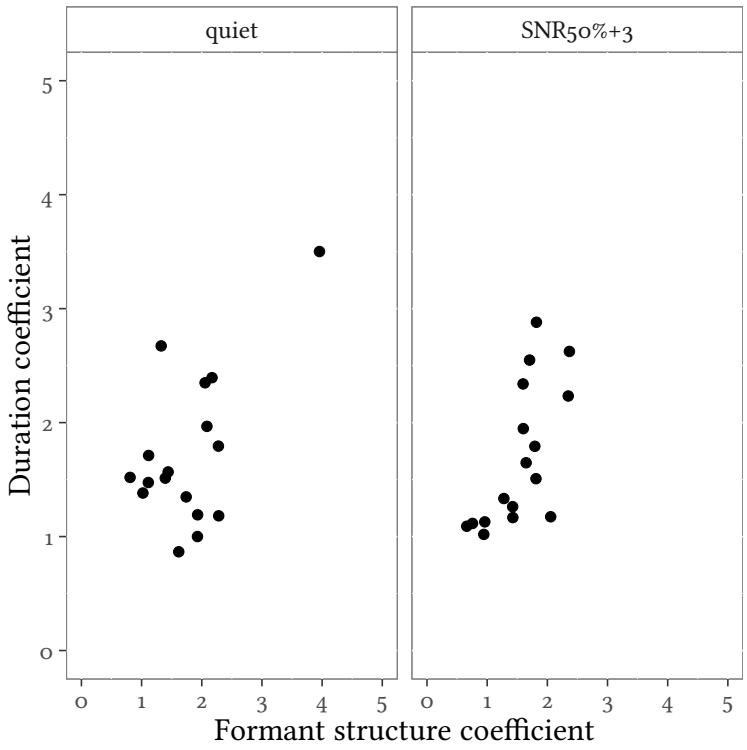

Figure 2.7: The left panel shows the scatterplot between formant structure and duration coefficients in condition quiet, and the right panel shows the scatterplot in condition SNR50%+3.

conditions (r=0.90, p<0.001), demonstrated in figure 2.7

2.3 Discussion

This chapter investigates the effect of spectral resolution, spectral distortion and noise masking on NH listeners' spectral and temporal cue weighting strategies for spectrally-shifted and noise-vocoded words. Word stimuli contained the tense - lax vowels varying along two acoustic dimensions, and listeners' categorical labeling responses were recorded and characterised with logistic regressions. Coefficients of the two acoustic cues, vowel formant structure and duration, were used as an indication of listen-

ers' perceptual weighting on the spectral and temporal aspects, and their ratio was used as an indication of the listeners' weighting strategy.

Testing stimuli in Experiment 1 were synthesised using the Tandem-STRAIGHT algorithm based on natural recordings, and stimuli in Experiment 2 were based on the best exemplars from the earlier studies. Listeners' recognition of synethsised vowels was sensitive to the synthesis fidelity, and the STRAIGHT algorithm has been found to preserve well the spectral features of vowels (Assmann & Katz 2005). The best exemplars were obtained from native Southern British English speakers who were asked to fine tune vowel acoustic dimensions to find the perceptually best vowel representatives (Evans & Iverson 2004; Iverson et al. 2006). Therefore, both methods should preserve well the acoustic profile of the tense and lax vowels in Southern British English, the same dialect as all CI and NH participants.

Spectral resolution was altered here by reducing the spectral information to 6, 8 or 12 bands wish a noise-band vocoder, and the degree of spectral distortion was altered by shifting the noise carrier bands relative to the analysis band by 2mm, 4mm or 6mm. NH listeners relied most heavily on the spectral cue when words were undegraded. Similar strategies are shown in other studies using different phonemic contrasts (Souza et al. 2015; Nittrouer et al. 2013; Winn et al. 2013a; Iverson et al. 2006), suggesting that when spectral information is intact NH listeners almost wholly rely on it for word categorisation. This strategy might be developed over years of exposure to the specific language, finding the cues that are most

informative for categorisation and least variable across different speaking styles or speakers (Holt & Lotto 2006; Nittrouer et al. 2014).

Once spectral degradation and distortion were applied to the stimuli, listeners decreased their perceptual weighting of the formant structure cue. The decreased reliance on the formant structure cue could be due to not having a sufficient number of bands to transmit the formant movement from the tense to the lax end of the continuum through cross-channel amplitude modulations (Prendergast & Green 2012). Although unavailability of sufficient spectral resolution might explain the significant decrease in the use of the spectral cue for the 6 bands conditions, there was enough spectral resolution to perform the word categorisation task here in 8 and 12 bands conditions. Therefore, the degradation also changed listeners' perceptual attention on the spectral cue, so that it was treated as less important in categorising words, although it might still be useful. Note that the impact of immediate spectral degradation and distortion was not independent, which is indicated by the significant interaction between the two factors. The higher spectral resolution did not always guarantee more reliance on the spectral cue, since increasing the number of bands didn't change the cue weighting pattern for spectral shifting larger than 3mm. This might be due to the considerable mismatch from the representation of formants in natural speech for shifts greater than 3mm, making it impossible for listeners to utilise this damaged cue to restore the phonemic structure from acoustic inputs immediately. In this experiment, the average F1 value is between 465 Hz and 268 Hz, and the average F2 value is between 1670 Hz and 2243 Hz

from the lax end to the tense end of the vowel continuum. After a 2mm upward shift, the average F1 value increases to between 666 Hz and 406 Hz, and the average F2 value to between 2253 Hz and 3010 Hz. This might not be unusual in a normal listeners' hearing experiences since female speakers can have formants as much as 23% higher than male speakers, and children can be 39% higher(Kwon 2010; Peterson & Barney 1952). After 4mm shifting, however, F1 ranges from 929.9 Hz to 587.2 Hz and F2 3023.2 Hz to 4020.4 Hz, which might extend the normal experience of listeners. This is consistent with other studies using similar types of spectral degradation and distortion. Fu and Shannon (1999) reported a significant drop in vowel categorisation for tonotopic shifts either apically or basally of over 3mm in 4, 8 and 16 bands noise-vocoded speech. But no interaction between the number of bands and shifts was found. This might be due to the task difference: the vowel categorisation task involved using both durational and spectral cues, while the current study investigated independently two cues with them varying orthogonally. Therefore, the resistance to spectral degradation could be stronger in the first case, where intact durational cues vary consistently with spectral cues to assist the recognition. Rosen et al. (1999) reported near chance sentence recognition performance for 4-band noise vocoded and 6.46mm shifted speech compared to the unshifted speech. Only after 3 hours of training, performance with a single talker increased to half of that observed with the original unshifted condition. Arguably, a bigger mismatch makes the mapping from the acoustic inputs to the phonemic representations more difficult and requires more inputs

before listeners could re-establish the representation.

Although a wide range of SNR levels was used in Experiment 2, no significant changes in the reliance on the spectral and temporal cue were found. Previous studies have shown that listeners rely more on the acoustic cues that are less damaged by the masking noise (Winn et al. 2013b; Wardrip-Fruin 1985). For instance, the VOT cue is less weighted than the Fo cue in the perception of word-initial voicing in stop consonants when words are masked by speech-shaped noise. The aspiration noise that characterises the VOT cue is aperiodic and concentrated in the high-frequency region, so its intensity will be more affected by the masking noise than the Fo cue in the vowel section (Winn et al. 2013b). The lack of significance in this study might be due to the higher SNR level used than previous studies (6.7dB in this study, odB in Wardrip-Fruin (1985) and Winn et al. (2013b)). The SNR level pushing individual sentence recognition performance down by more than 50% could still not be enough to mask the spectral shape of the vowel to bring out the advantage of shifting the weighting to the duration cue. Therefore, the performance level might not be a good indication of listeners' weighting strategy.

The current study indicates that listeners are able to adjust their acoustic cue weighting strategy dynamically, depending on the level of acoustic degradation and distortion. Their perceptual weighting of the speech cues is constrained by the amount of information left for mapping to the phoneme, and also related to listeners' allocation of perceptual attention based on how far the speech cues are from the stored representations. Furthermore,

the large between-individual variances in the cue weighting strategy are unaccounted for. It is reasonable to hypothesise that how well listeners can utilise their auditory sensitivity and apply the appropriate language-specific weighting strategies for reaching phonemic decisions should be related to how well and how easily they can recover phonemic structure in the speech signal. The next chapter explores the possibility that listeners' acoustic weighting strategies for degraded speech might be related to their general speech performance and listening effort, for both NH listeners and CI users.

Chapter 3

Listeners' acoustic cue weighting strategy, speech performance and listening effort

It is hypothesised that, despite great variability, listeners employing the strategy predominately used in non-degraded conditions by NH listeners, should have better speech recognition performance achieved with less listening effort. This hypothesis is supported by previous studies reviewed in chapter 1 that have shown that CI listeners' acoustic cue weighting strategy is related to their word recognition performance. Typically, CI listeners who put more weighting to the cues that are weighted more by NH listeners have better word recognition score. Considering that the weighting strategy shared by NH listeners within the same language community is developed over years and highly specific, it should allocate more importance on acoustic cues that are most informative and reliable in that language. Therefore, applying this weighting strategy should restore the phonemic

structure from the acoustic inputs with better accuracy as well as with less cognitive effort, since listeners will not 'waste' cognitive resources on acoustic cues that are not useful for that language.

To test this hypothesis, NH listeners with CI acoustic simulations (Experiment 1) and CI users (Experiment 2) performed a series of sentence recognition tasks with pupil response simultaneously recorded. It was followed by a word labelling task to find their relative weighting of temporal and spectral cues, and an auditory discrimination task using non-word stimuli to find their temporal and spectral auditory sensitivity. The pupillary response is a sensitive and reliable index for cognitive processing load in listening tasks and reveals an aspect of speech comprehension not necessarily reflected in recognition scores. If the task becomes more difficult, for instance, with lower signal-to-noise ratios (SNRs), divided attention and spectral degradation, pupil size will increase accordingly (Zekveld et al. 2011; Koelewijn et al. 2012; Zekveld & Kramer 2014; Koelewijn et al. 2015). Therefore, the task difficulty level needs to be controlled across listeners in order to observe the dynamic relationship between speech comprehension and listening effort without other confounds affecting pupil size. This is especially necessary when testing CI or NH listeners with CI acoustic simulation since their speech recognition performance varies greatly under conditions of auditory or acoustic degradation. Without fixing the perceptual difficulty of the speech task to a level similar for each listener, the variability observed in listening effort would be entirely due to the interindividual variability in speech recognition, leaving little to be explained

by other factors. In this chapter, perceptual difficulty levels were controlled by masking sentences with a speech-shaped noise at individually set SNRs. The exact SNRs were obtained from each listener beforehand from adaptive speech reception threshold (SRT) tests tracking 40%, 50% or 80% of their degraded sentence recognition performance in quiet. Therefore, each trial in each condition would be similarly easy or difficult across listeners.

3.1 Experiment 1

3.1.1 Methods

Participants

14 normal-hearing native standard Southern British English speaking adults were recruited via the University College London (UCL) Psychology Pool. All participants were aged between 18 and 45 and had normal hearing (defined as hearing thresholds of 20dB HL or better between 250 - 8000 Hz tested at octave frequencies). None of them had prior experience with vocoded speech. For their contributions, they were paid at the rate of 7 pounds per hour. All participants consented to take part by reading and signing a consent form, as approved by the UCL Research Ethics Committee.

Stimuli

Sentence stimuli were Basic English Lexicon (BEL) sentences (Calandruccio & Smiljanic 2012), recorded at a sampling frequency of 22.05 kHz from a

male native Southern British English speaker. Each of these sentences contained 4 keywords, upon which the scoring was based. Sentences were manipulated using PSOLA (Moulines & Charpentier 1990) in Praat (Boersma 2002) to be of the same duration (mean = 2.02s, standard deviation = 0.24s). The same synthesised 'beat' - 'bit' word continnum as in chapter 2 Experiment 2 was used. Each step in formant structure was cross-paired with each step in vowel duration, making $6 \times 6 = 36$ tokens.

Training materials consisted of two short stories (Aesop's *The north wind and the sun* and *The wolf and goat*, recorded from a male native Southern British English speaker), ten BEL sentences in quiet and ten BEL sentences masked by speech-shaped noise at an SNR level of 10 dB.

All stimuli were then 8-band noise-vocoded and 4mm upward shifted in MATLAB 2013b, using a similar procedure in chapter 2. This was to allow for enough spectral resolution for word discrimination and to match the performance of successful CI listeners while reducing the auditory saliency of spectral cues by systematic distortion with over 3mm spectral mismatch shown in Dorman et al. (1997); Friesen et al. (2001); Faulkner et al. (2003) and the previous experiment. Each sound file was digitally filtered into 8 bands, using sixth-order Butterworth infinite impulse response filters. Filter spacing was based on equal basilar membrane distance (Greenwood 1990) across a frequency range of 70-6000 Hz. The output of each band was full-wave rectified and low-pass filtered (fourth-order Butterworth) at 30 Hz, using a zero-phase forwards-backwards technique to extract the amplitude envelope. The envelope was then multiplied by a wide band noise

carrier. The resulting signal was passed through 8 output filters with their cut-off frequencies shifted upwards from the analysis filters by 4mm on the basilar membrane distance according to Greenwood map. The root mean square (rms) level of the output signal was adjusted to match the original analysis band and the signal was summed across all bands.

Procedure

Participants were seated in a quiet room, 70 cm from a 17-inch white screen monitor and 55 cm from an infrared monocular eye-tracker (Eyelink 1000, SR Research, 500 Hz sampling rate). All audio stimuli were presented binaurally through Sennheiser HD25 headphones at a comfortable level. Experiments were run in Matlab 2015b, using Psychtoolbox and custom software.

The illuminance of the room was adjusted for each participant (mean illumination = 101 lx), such that the pupil diameter was midway between maximum and minimum size (elicited by turning off and on the room lighting consecutively).

Firstly, participants were familiarised with the simulated speech through training for 15 minutes. During the training, they listened to sentences and then were given written and acoustic feedback. No active responses were required.

To find each individuals' speech recognition score, participants were then tested with 30 randomly chosen BEL sentences in quiet. In each trial, they were scored by the experimenter based on the number of keywords

correctly reported. Their percentage correct was averaged across trials to obtain their speech recognition score. This was followed by two adaptive speech perception threshold tests (Plomp & Mimpen 1979) using speech-shaped noise to track either 40% or 80% of each individuals' speech recognition score. Each SRT test consisted of 15 randomly selected sentences and the order of the two tests was randomised.

At the 2 SRT levels obtained from each individual, 2 fixed SNR speech recognition tests (condition SNR40% and SNR80%) were then performed and participants' pupil responses were recorded simultaneously. Due to the experimental design, the condition using the SNR from tracking 40% of participants' speech recognition (SNR40%) was more difficult compared to the condition with 80% (SNR80%) for each participant, but each condition was similarly difficult or easy across participants. In both conditions, the masking noise was set to the same level, in order to prevent listeners from predicting the test difficulty based on noise level. For each trial, the presentation of the speech shaped noise masker started 3s before sentence onset and finished 2s after sentence offset. Participants were instructed to fixate the black fixation cross on the white monitor and avoid excessive blinks. After the masker offset, they were prompted by the colour change of the fixation cross to repeat back the sentence. Their responses were scored by the experimenter and the program proceeded to the next trial. An example trial is displayed in figure 3.1. Each fixed SNR test condition consisted of 30 randomly selected sentences with the order of conditions randomised.

Finally, participants' cue weighting strategies were measured with a word

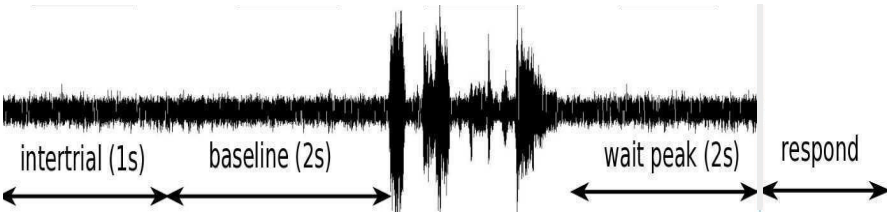

Figure 3.1: The trial starts with acoustic presentation of 1s speech-shaped noise and visual presentation of a black fixation cross against the white monitor screen ('intertrial'). This is to allow for pupil size to recover from the previous trial. Another 2s of baseline measurement follows, with the same acoustic and visual presentation ('baseline'). The sentence starts 3s into the trial and finishes after 2.02s (the duration of the sentence), followed by noise presentation for 2s ('wait peak'), with the same visual presentation. The black fixation cross then changes to yellow to prompt listeners to repeat back the sentence ('respond'). Pupil measurements during 'baseline', sentence presentation and 'wait peak' are included for processing and analysis.

labeling task. After hearing a word token, they were instructed to choose what they heard on the screen from either 'beat' or 'bit'. Their response was recorded by the computer and next trial started.

Data processing

Baseline pupil diameter in each trial was calculated as averaged pupil traces 2s before the start of the sentence. The rest of the pupil diameter measurements were subtracted by that baseline level to obtain pupil size change elicited by sentence recognition. Pupil diameter values below 3 standard deviations (SD) of the mean of the trace were coded as blinks. Traces between 50 data points before the start and after the end of blink were cubically interpolated in Matlab, to further decrease the impact of the obscured pupil from blinks. Trials that had over 20% of the data points coded as blinks from the start of baseline to the end of masker presentation were excluded. Trials containing blinks longer than 0.4s were also excluded, since they

were more likely to be artefacts than normal blinks (Bristow et al. 2005). Altogether, 748 trials of pupil response recordings were included, with 53 trials on average for each participant (SD = 9). All valid traces were then low-pass filtered at 10 Hz with a first order Butterworth filter to preserve only cognitively related pupil size modulation (Klingner et al. 2008) and downsampled to 50 Hz. Three indices of pupil response (mean pupil dilation, peak pupil dilation and peak latency) were obtained from processed traces, consistent with the method in Zekveld et al. (2010, 2011). Mean pupil dilation and peak pupil dilation were the average and maximum diameter of pupil measurements from sentence onset to response prompt, relative to the baseline pupil size. Peak latency response was the time between onset of the sentence to the peak dilation. These indices were calculated for each trial.

The cue weighting ratio for each individual was calculated from the binomial response of participants in the word labeling task, using the same method as in chapter 2. Similarly, a larger ratio indicates more reliance on the spectral cue relative to the temporal cue; and a smaller ratio indicates more reliance on the temporal cue relative to the spectral cue. Listeners' response functions and cue weighting ratio are shown in figure 3.2.

3.1.2 Statistical analysis and results

The first aim of the statistical analysis was to examine the effect of an individual's cue weighting strategy on sentence recognition. To this purpose, a logistic regression model was fitted to listeners' sentence recogni-

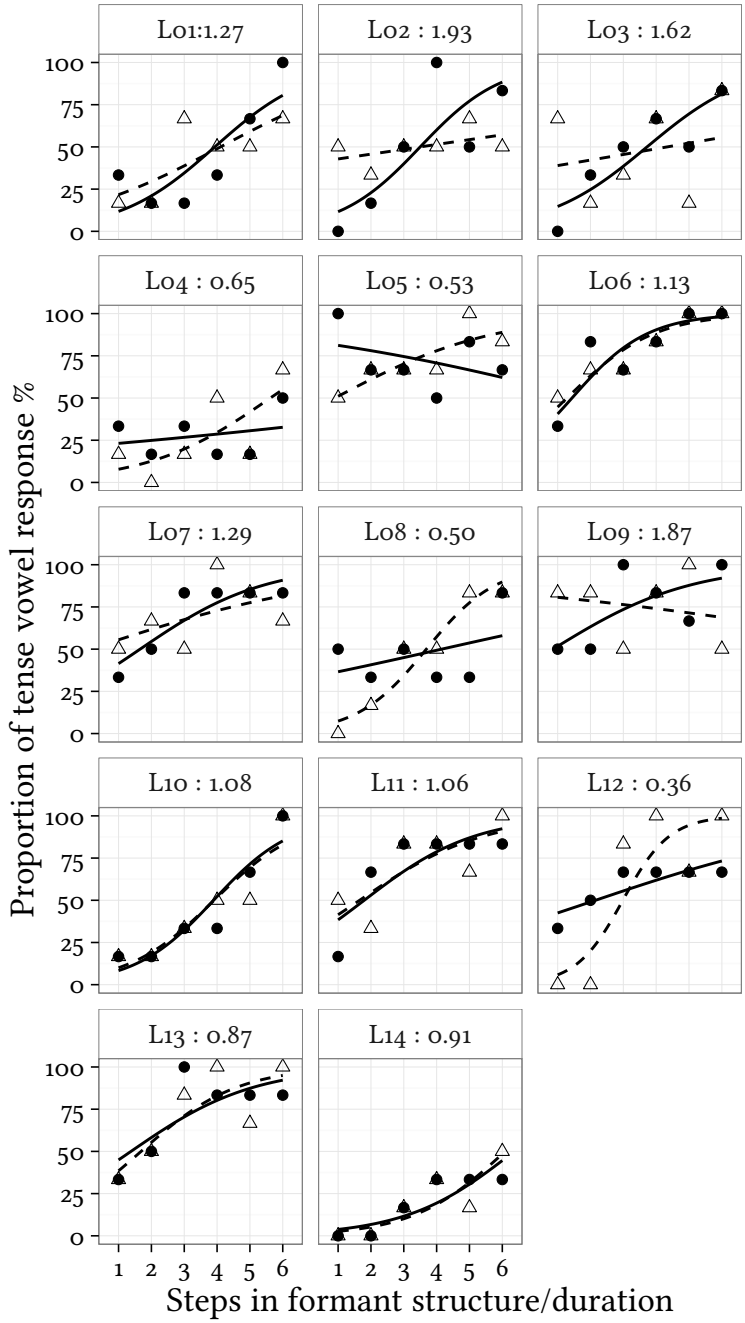

Figure 3.2: NH listeners' proportion of tense vowel responses for 8 band noise-vocoded and 4mm shifted word stimuli. The filled circle (●) is the averaged proportion response for each step in formant structure, and the hollow triangle (△) is the averaged proportion for each step in duration. The filled line (–) is the logistic regression fit to the proportion of tense vowel responses using steps in formant structure, and the broken line (- -) is the logistic regression using steps in duration. The top panel shows listeners' code and their cue weighting ratio.

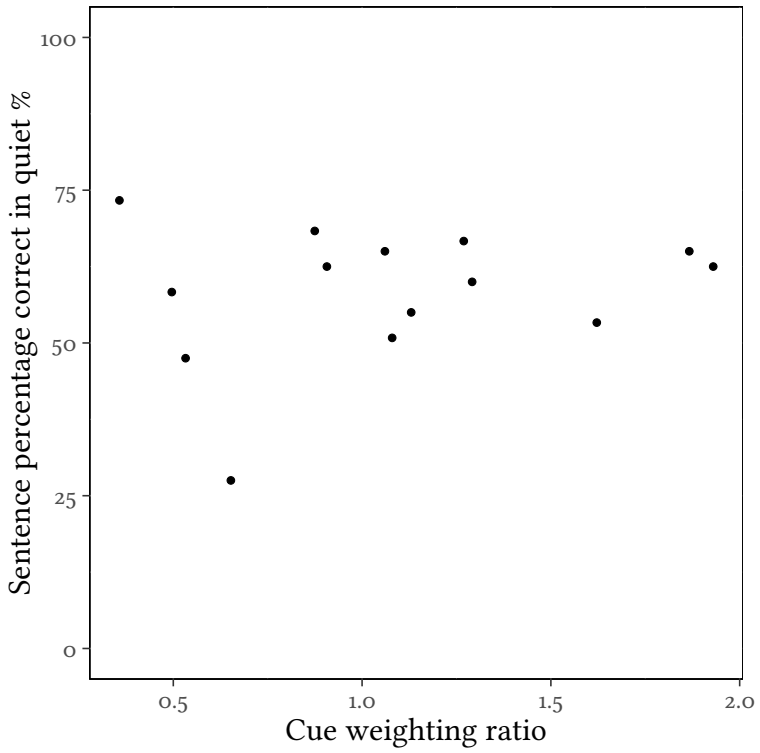

Figure 3.3: Scatterplot of 14 listeners for their cue weighting ratio and degraded sentence recognition scores in quiet.

tion performance in quiet using the cue weighting ratio as the independent variable.A scatterplot of participants' sentence recognition performance in quiet and their cue weighting ratio is displayed in figure 3.3. Over all listeners, no significant effect of cue weighting ratio was found (F=0.22, df=1, p>0.05), suggesting that listeners' acoustic cue weighting strategy was not related to their sentence recognition performance in quiet.

The second aim of the analysis was to explore the impact of an individual's cue weighting strategy on listening effort during speech perception. Therefore, trials in SNR40% and SNR80% conditions were analysed. Logistic mixed effect models were fitted to the proportion correct levels of each trial, predicted by the fixed effect factors *pupil response*, *condition* (SNR40%

and SNR80%) and *cue weighting ratio*. By using *listener* and *sentence* (the exact BEL sentence tested) as random effect factors in the model, we controlled for the variance in correct levels (random intercept) and other fixed factors (random slope) that were associated with them. Three models were constructed using the lme4 package in R (Bates et al. 2014) with each pupil response index (baseline corrected pupil mean dilation, peak dilation and latency response) as an independent variable. Factors were entered into the model in the sequence below: the model firstly started with taking *listener* and *sentence* as random intercepts; fixed effect factors *pupil response*, *condition* and *cue weighting ratio* and their interactions were then entered; finally, random slopes were entered into the model. Factors were retained in the model only if they significantly improved the model fitting, using Chi-squared tests based on changes in deviance.

Details of the best models predicting proportion correct using mean pupil size, peak pupil size and latency response are shown in table 3.1. For the model containing mean pupil size, significant fixed effect factors are: *condition* (χ^2 = 8.89, df = 1, p < 0.01), *condition* × *mean pupil size* (χ^2 = 6.81, df = 1, p < 0.01), *mean pupil size* × *cue weighting ratio* (χ^2 = 9.53, df = 1, p < 0.01). Unsurprisingly, post hoc Wald tests showed that the difficult condition SNR40% had significantly lower averaged sentence recognition score than the easy condition SNR80% (β = 0.74, SE = 0.26, p < 0.01). In both conditions, the bigger the mean pupil response in a trial, the higher proportion of correct responses (SNR40%: β = 3.90, SE = 1.60, p < 0.05; SNR80%: β = 6.64, SE = 1.50, p < 0.001). But in the easy condition SNR80%, this effect is

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	-0.65	0.51	0.20			
Condition(SNR80%)	0.74	0.26	**<0.05**	8.90	1	**<0.05**
Cue Weighting Ratio	0.25	0.41	0.53	<0.05	1	0.97
Mean Pupil Size(SNR40%)	3.90	1.60	**0.01**			
Mean Pupil Size(SNR80%)	6.64	1.50	**<0.001**	1.79	1	0.18
Mean Pupil Size × Condition(SNR80%)	2.74	1.14	**0.02**	6.81	1	**<0.001**
Mean Pupil Size × Cue Weighting Ratio	3.58	1.10	**<0.05**	9.53	1	**<0.05**
Random effects:	**SD**	**cor**		χ^2	df	p
Intercept | Listener	0.61			489.80	1	**<0.001**
Intercept | Listener × Condition	0.55			69.89	1	**<0.001**
Intercept | Sentence	1.52			488.00	1	**<0.001**
Mean Pupil Size | Sentence	5.74	-0.74		25.16	3	**<0.001**

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	-1.11	0.57	0.05			
Condition(SNR80%)	0.87	0.27	**<0.05**	8.89	1	**<0.05**
Cue Weighting Ratio	0.36	0.46	0.44	0.04	1	0.83
Peak Pupil Size	3.43	0.93	**<0.001**	13.58	1	**<0.001**
Peak Pupil Size × Cue Weighting Ratio	1.75	0.73	**0.02**	4.35	1	**<0.05**
Random effects:	**SD**	**cor**		χ^2	df	p
Intercept | Listener	0.60			489.80	1	**<0.001**
Intercept | Listener × Condition	0.60			69.89	1	**<0.001**
Intercept | Sentence	1.49			124.61	1	**<0.001**

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	-0.57	0.25	**0.02**			
Condition(SNR80%)	0.82	0.10	**<0.001**	8.89	1	**<0.05**
Latency Response	<0.05	<0.05	0.24	3.99	1	**<0.05**
Random effects:	**SD**	**cor**		χ^2	df	p
Intercept | Listener	0.60			489.80	1	**<0.001**
Intercept | Sentence	1.38			488.00	1	**<0.001**

Table 3.1: Model parameter estimates and model comparison statistics for the best logistic mixed effect models fit to proportion correct. For the categorical factor Condition, the reference level is Condition SNR40%.

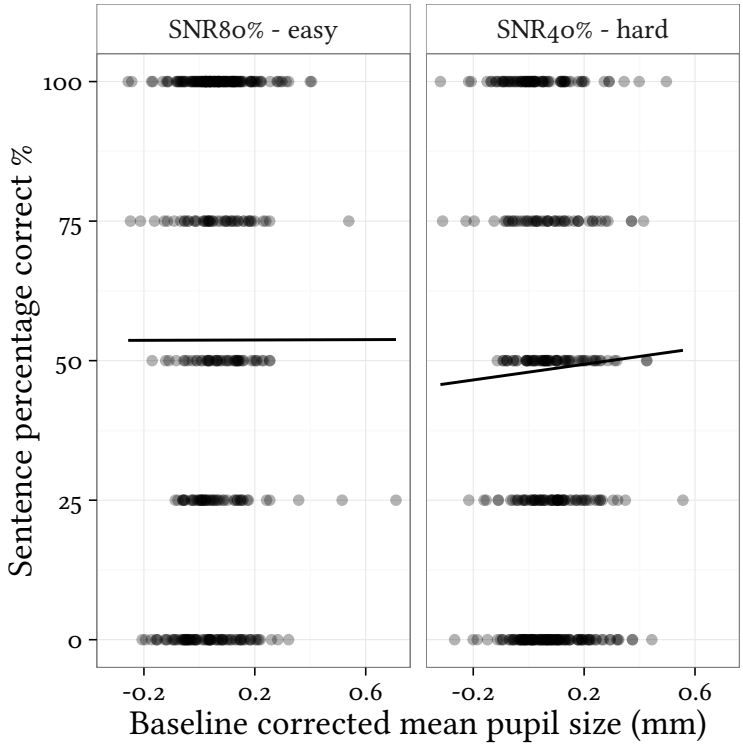

Figure 3.4: Figure illustrating the interaction between condition and mean pupil size. Black dots are listeners' performances in each valid trial. Black lines are logistic regressions fitted to listeners' sentence recognition scores.

bigger than in the difficult condition SNR40% (β = 2.74,SE = 1.14, p < 0.05). This interaction is illustrated in figure 3.4.

The interaction between mean pupil response and cue weighting ratio suggested that cue weighting strategy explained a significant amount of variance in the relation between listening effort and sentence recognition performance. To examine whether cue weighting ratio has any systematic effect on this relation, a trend analysis was performed. It showed a significant linear trend of cue weighting ratio for its interaction with mean pupil response (F = 14.03, df = 1, p < 0.01), suggesting that for participants with smaller ratios (more weighting of the duration cue), bigger mean pupil responses were associated with poorer sentence recognition; and for listeners

with bigger ratios (more weighting of the spectral cue), bigger mean pupil responses were associated with better sentence recognition (β = 7.96, SE = 2.18, p< 0.001).

For the model containing peak pupil size, significant fixed effect factors are: *condition* (χ^2 = 8.89, df = 1, p < 0.01), *peak pupil size* (χ^2 = 13.58, df = 1, p < 0.001), *cue weighting ratio* × *peak pupil size* (χ^2 = 4.35, df = 1, p < 0.05). For the interaction between peak pupil response and cue weighting ratio, a trend analysis also showed a significant linear trend (F = 7.37, df = 1, p < 0.01) of cue weighting ratio, suggesting that for participants with a smaller ratio (more weighting of the duration cue), bigger peak responses were associated with lower sentence recognition performance; and for participants with bigger ratio (more weighting of the spectral cue), bigger peak responses were associated with better performance (β = 3.91, SE = 1.47, p < 0.01). The interactions are shown in figure 3.5 and figure 3.6.

For the model using latency response, the only significant factor is: *condition* (χ^2 = 8.89, df = 1, p < 0.01). The easy condition SNR80% had a shorter average latency than the hard condition SNR40% (β = -0.82, SE < 0.001, p < 0.001).

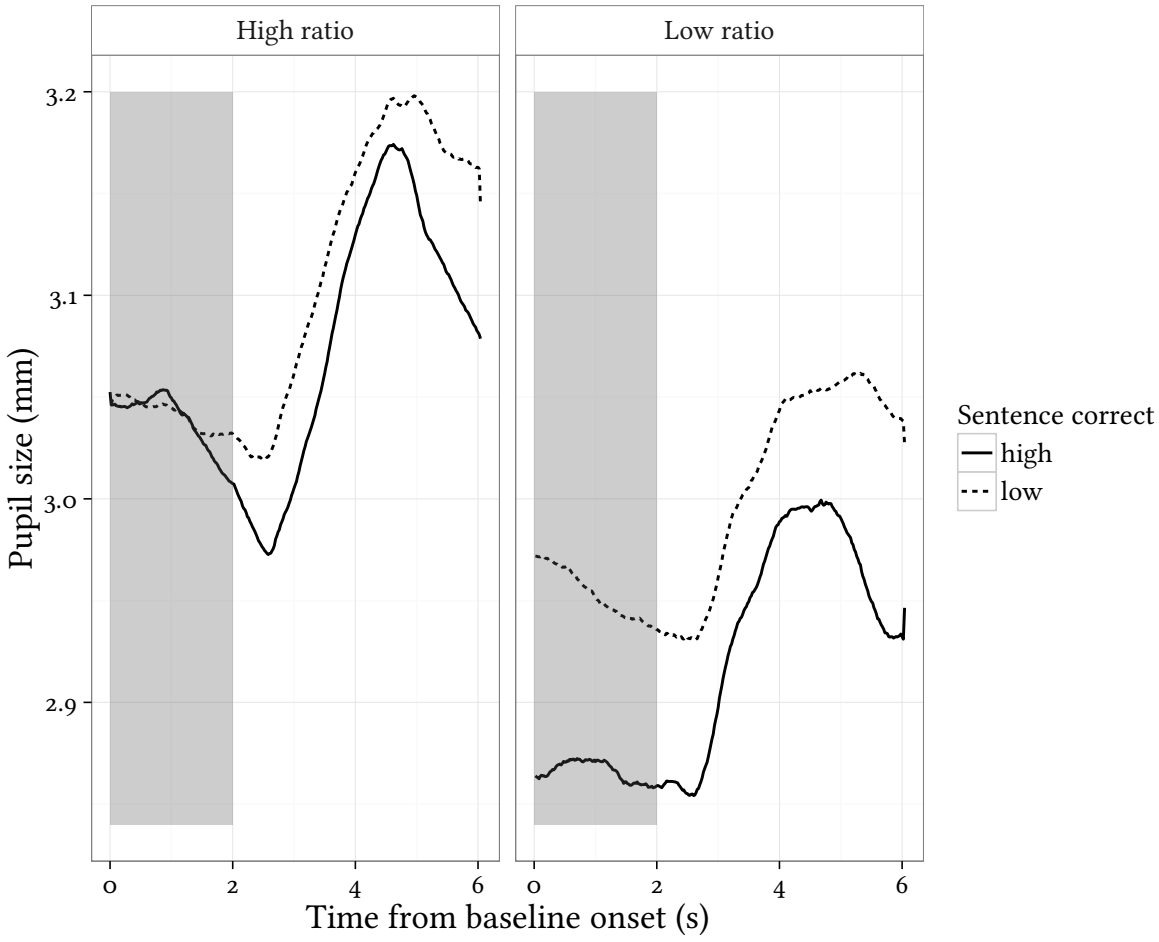

Figure 3.5: The shaded region is for baseline measurement, and the rest of the pupil trace is within the analysis window starting from the offset of the baseline to the response prompt. The left panel shows the aggregated pupil traces for 7 participants with higher cue weighting ratio (greater weighting of the formant structure cue), and the right panel is for 7 participants with lower cue weighting ratio (greater weighting of the vowel duration cue). The back lines show the aggregated pupil traces for trials with >50% correct, and the dashed lines show the aggregated pupil traces for trials with $\leq$50% correct.

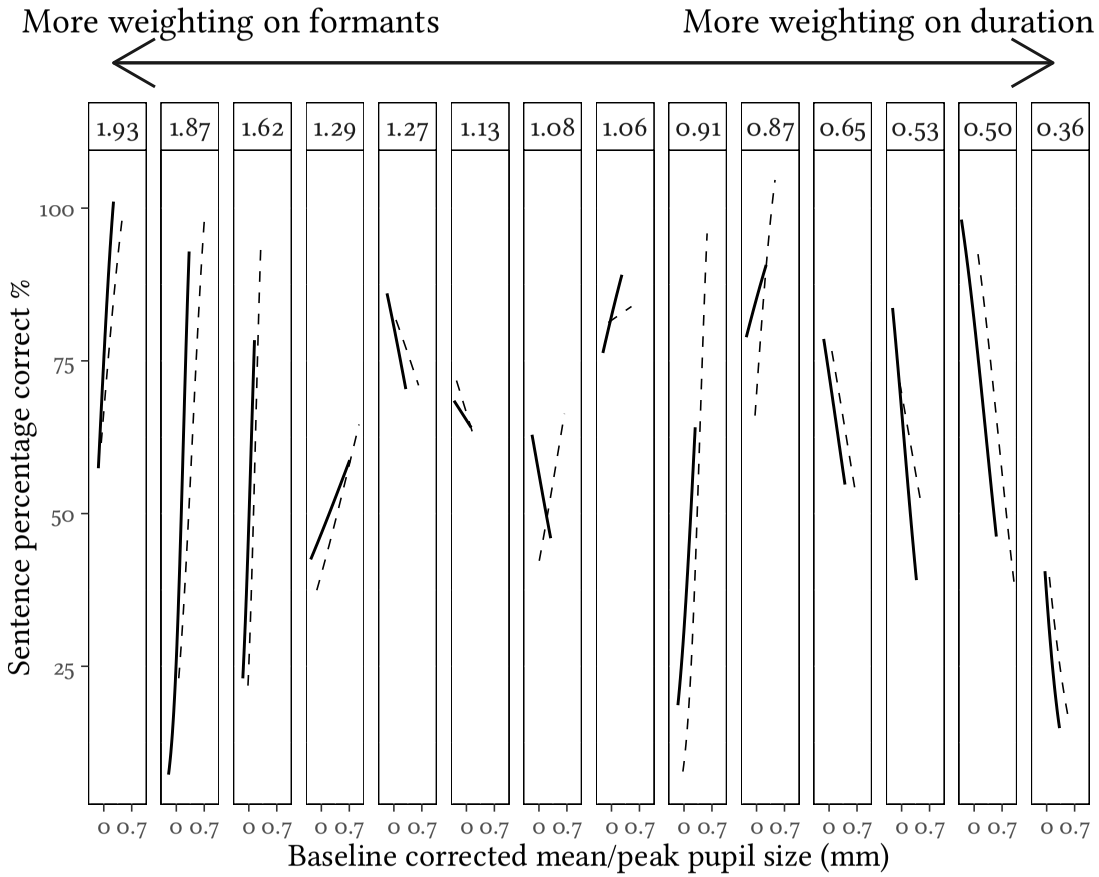

Figure 3.6: Plots illustrating the interaction between the cue weighting ratio and mean/peak pupil size. Each panel displays performance of a single listener, with their cue weighting ratios on the panel top. The larger the ratio, the greater the relative weighting of the spectral cue; and the smaller the ratio, the greater the weighting of the temporal cue. Black lines are logistic regressions with mean pupil size as the independent variable; dashed lines are logistic regressions with peak pupil sizes as the independent variable.

3.2 Experiment 2

3.2.1 Methods

Participants

7 native Southern British English speaking and post-lingually deaf cochlear implant users were recruited and 1 dropped out during the experiment (participant Cf). Summary information is shown in table 3.2. All participants consented to take part by reading and signing a consent form, as approved by the UCL Research Ethics Committee.

Stimuli

Sentence stimuli were Basic English Lexicon (BEL) sentences (Calandruccio & Smiljanic 2012) recorded from a male native Southern British English speaker. Sentences were manipulated using PSOLA in Praat to be of the same duration (mean = 2.02s, standard deviation = 0.24s).

Training materials consisted of ten BEL sentences in quiet and ten BEL sentences masked by speech-shaped noise at an SNR level of 10 dB.

To measure listeners' acoustic cue weighting strategy, the same synthesised 'beat' - 'bit' word continuum as in Experiment 1 was used. Each step in formant structure was cross-paired with each step in vowel duration, making $6 \times 6 = 36$ tokens.

To measure listeners' auditory sensitivity to the spectral shape and duration along the 'beat' - 'bit' continuum, the stable vowel section of the most typical 'beat' and 'bit' according to table 2.5 were extracted. Two continuua

Subject	Age	Gender	Cause of deafness	Age at onset	Age at implantation	Model
Ca	48	F	Viral infection	35	46	Nucleus 6
Cb	61	F	Meningitis	42	57	Nucleus 5
Cc	72	F	Ushers Syndrome	16	66	Advanced Bionic
Cd	34	F	Genetic	3	27	Nucleus 6
Ce	40	M	Premature birth	2	27	Advanced Bionic
Cf	73	F	Genetic	2	70	Nucleus 6
Cg	36	M	Viral infection	27	28	Nucleus 6

Table 3.2: Demographic information of cochlear implant users.

were then synthesised. For measuring listeners' sensitivity to the spectral shape, 60 tokens were synthesised using Tandem-STRAIGHT algorithm in Matlab, with F1 F2 values varying linearly from /i/ (F1 = 210 Hz, F2 = 2707 Hz) to /ɪ/ (F1 = 372 Hz, F2 = 2230 Hz) stable formants section, and durations fixed to 0.12s. For measuring listeners' sensitivity to the duration, 60 tokens were synthesised, with durations varying linearly from /i/ (0.145s) to /ɪ/ (0.085s), and F1 F2 values fixed to 269 Hz and 2505 Hz. The spectrum of some example tokens for the auditory discrimination task are shown in figure 3.7. All stimuli were then scaled to the same rms intensity level.

Procedure

Participants were seated in a quiet room, 70 cm from a 17-inch white screen monitor and 55 cm from an infrared monocular eye-tracker (Eyelink 1000, SR Research, 500 Hz sampling rate). The illuminance of the room was adjusted for each participant, such that the pupil diameter was midway between maximum and minimum size (elicited by turning off and on the room lighting consecutively). All audio stimuli were presented through a Yamaha MS101 loudspeaker, calibrated at 72 dB SPL. Experiments were run in Matlab 2015b Psychtoolbox and custom software.

Firstly, participants were familiarised to the speech material through training for 15 minutes. During the training, they listened to sentences and then were given written and acoustic feedback. No active responses were required.

To find each individuals' speech recognition score, CI users were then tested

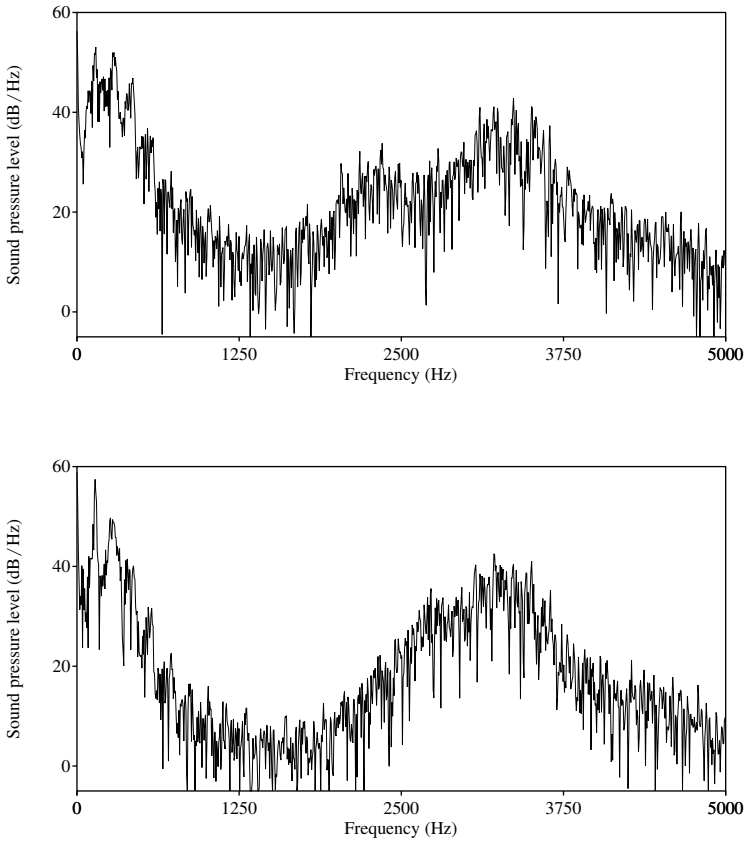

Figure 3.7: The top panel shows the spectrum of step 10 in the spectral shape discrimination continuum, and the bottom panel shows the spectrum of step 50 in the spectral shape discrimination continuum.

with 30 randomly chosen BEL sentences in quiet. In each trial, they were scored by the experimenter based on the number of keywords correctly reported. Their percentage correct was averaged across trials to obtain their speech recognition score. This was followed by an adaptive speech perception threshold test, using the updated maximum-likelihood (UML) package in Matlab, to track 50% of each individuals' speech recognition score with speech-shaped noise. UML function estimates the psychometric function by using Bayesian procedures that minimise the expected entropy of the posterior parameter distribution (Shen & Richards 2012; Shen et al. 2015). Listeners' sentence recognition score was set in UML function as the upper bound of the logistic psychometric function. The prior distribution range of the psychometric function slope was set between 0.1 to 10, and the range of the threshold was set between -10 dB to 30 dB. The SNR level of the first trial was set as 5 dB. Then the number of keywords reported by the participants were entered into the function, and the SNR level for the next trial was estimated online based on the sweet point for threshold estimation using the Bayesian minimum-variance procedure. The SRT test was set to converge either when the 90% confidence interval of the threshold estimation was within 3 dB or reaching the maximum trial number of 30.

At the SRT level obtained from each individual, a fixed SNR speech recognition test with 20 sentences was then performed and participants' pupil responses were recorded simultaneously. Due to this control on the intelligibility across participants, each trial in the fixed SNR test was similarly difficult for each participant, regardless of their speech recognition per-

formance. The presentation of the speech-shaped noise masker started 2s before sentence onset and finished 2s after sentence offset. Participants were instructed to fixate the black fixation cross on the white monitor and avoid excessive blinks. After the masker offset, they were prompted by the colour change of the fixation cross to repeat back the sentence. Their responses were scored by the experimenter and the program proceeded to the next trial.

Participants' cue weighting strategies were measured with a word labeling task using the 'beat' - 'bit' word continuum. After hearing a word token, they were instructed to choose what they heard on the screen from either 'beat' or 'bit'. Their response was recorded by the computer and next trial started.

Then, a three-alternative forced-choice (3AFC) test procedure was used to measure CI users' auditory discrimination to the spectral shape and duration of vowels along the 'beat' - 'bit' continuum. Three frogs appeared on the screen, with each 'saying' one of the stimuli from the continuum. Participants were instructed to click on the frog that uttered a sound different from the other two. The interstimulus interval was set at 500 ms. Two fixed reference discrimination tasks were used to for each just noticeable difference (jnd) measurement for the formant structure and duration. In each task, the standard stimulus was either the 'bit' or 'beat' endpoint of the continuum. The test started with the token from the other endpoint as the comparison stimulus, which was an easy task. A three-down/one-up adaptive procedure was the used to choose the comparison stimulus based

on participants' correct responses so that the stimulus could be discriminated from the standard 79.4% of the time (Levitt 1971). Step size varied throughout the test, from eight steps at the start and decreasing linearly over the first three reversals to four steps. The task ended after six reversals on each track or maximum of 30 trials. The jnd was then calculated by taking the mean of the final three reversals. A jnd of 10 steps would typically indicate that the listener was able to discriminate the difference between 2 steps in the acoustic cue weighting test.

Data processing

Baseline pupil diameter in each trial was calculated as averaged pupil traces 1s before the start of the sentence. The rest of the pupil diameter measurements were divided by that baseline level to obtain the proportional pupil size change elicited by sentence recognition. Pupil diameter values below 3 SD of the mean of the trace were coded as blinks. Traces between 50 data points before the start and after the end of blink were cubically interpolated in Matlab, to further decrease the impact of the obscured pupil from blinks. Trials that had over 20% of the data points coded as blinks from the start of baseline to the end of masker presentation were excluded. Trials containing blinks longer than 0.4s were also excluded since they were more likely to be artefacts than normal blinks (Bristow et al. 2005). Altogether, 132 trials of pupil response recordings were included, with on average 19 trials for each participant (SD = 2). All valid traces were then low-pass filtered at 10 Hz with a first order Butterworth filter to preserve only cognitively related

pupil size modulation (Klingner et al. 2008), and downsampled to 50 Hz. The processed pupil traces are shown in figure 3.8, aggregated by participants.

Three indices of pupil response (mean pupil dilation, peak pupil dilation and peak latency) were obtained from processed traces, consistent with the method in Zekveld et al. (2010, 2011). Mean proportional pupil size change and peak proportional pupil size were the average and maximum of proportional pupil changes from sentence onset to response prompt, relative to the baseline pupil size. Peak latency response was the time between onset of the sentence to the peak dilation. A principal component analysis (PCA) was then performed in R, with three variables scaled to the same unit of standard deviation and centering at zero. PCA is a mathematical algorithm that reduces large numbers of variables and data points to smaller numbers of independent factors, without much loss of information (Jolliffe 1986; Ringnér 2008). This reduction is accomplished by identifying directions, called principal components, along which the variation in the data is maximal. Therefore, these components are a good summary of the variation in the data. Details of the principal components are displayed in table 3.3 The first principal component (PC1) explained the most amount of variances in the three indices (67%), with large loadings on the mean (0.68) and peak pupil size changes (0.69). This suggests that PC1 is most representative of the trends in the three pupillary indices. Therefore, it was selected as an indication for sentence-evoked pupil response.

The cue weighting ratio for each individual was calculated from the bino-

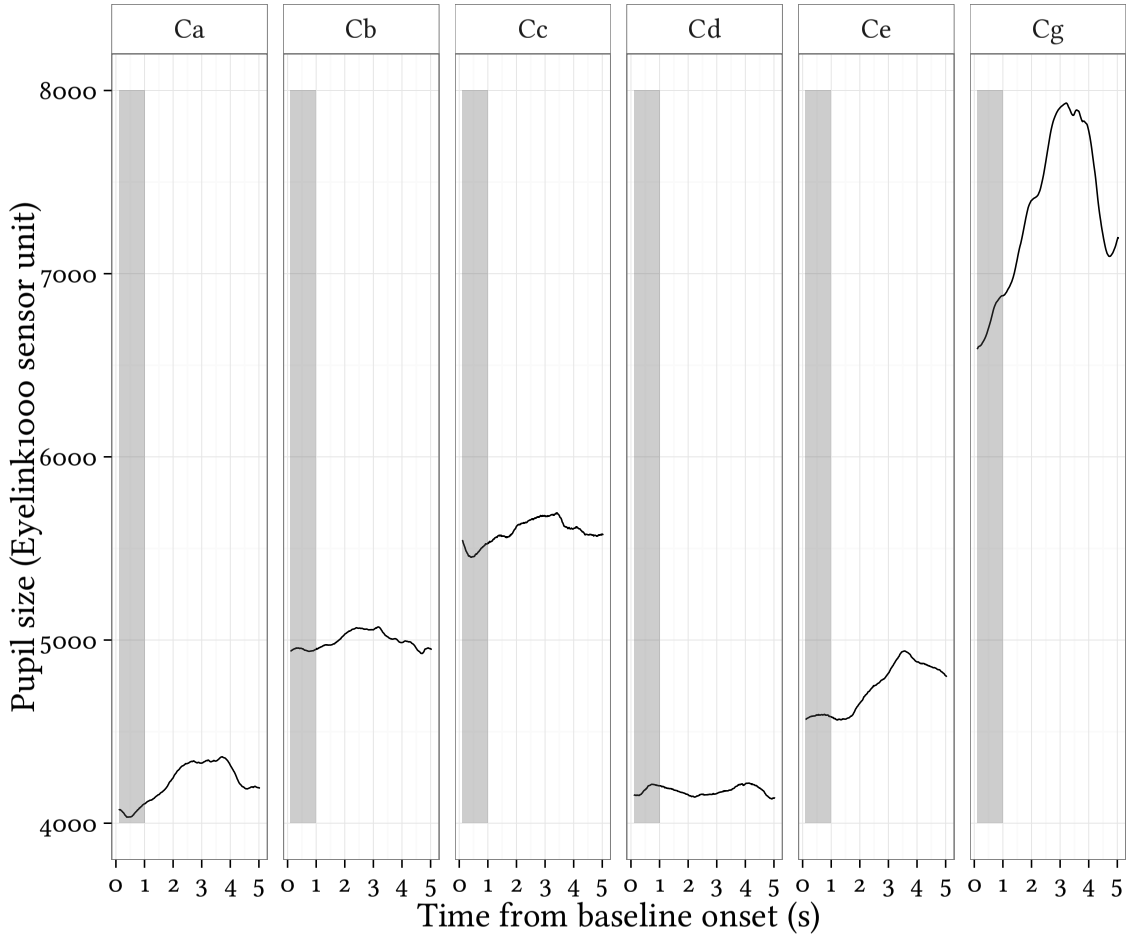

Figure 3.8: The y-axis shows the arbitrary Eyelink 1000 camera sensor units, with reference to the number of threshold camera pixels during each recording session. The x-axis shows the time from baseline onset. The shaded region is for baseline measurement, and the rest of the pupil trace is within the analysis window starting from the offset of the baseline to the response prompt. Each panel displays performance of a single listener, with their participant code on the panel top.

Importance of principal components:			
	PC1	PC2	PC3
Standard deviation	1.42	0.97	0.24
Proportion of variance	0.67	0.31	0.02
Rotation:			
	PC1	PC2	PC3
Latency	0.24	-0.97	0.04
Mean proportional change	0.68	0.20	0.70
Peak proportional change	0.69	0.15	-0.71

Table 3.3: The importance of principal components and variable rotation of the principal component analysis on CI users' pupil dilation measurements.

mial response of participants in the word labelling task, using the same method in Experiment 1. Similarly, a higher ratio indicates more reliance on the spectral cue relative to the temporal cue; and a lower ratio indicates more reliance on the temporal cue relative to the spectral cue. CI listeners' response functions and cue weighting ratio are shown in figure 3.9.

Since there was no significant difference between the jnd measurements in spectral shape and duration for the first and second discrimination tasks (spectral shape discrimination: t=-1.48, p>0.05; duration discrimination: t=-1.29, p>0.05), the smaller jnd step (better discrimination) was chosen for each measurement. The mean jnd step for spectral shape discrimination is 24.44, with the standard deviation of 5.96; the mean jnd step for duration discrimination is 32.60, with the standard deviation of 4.47. The jnd step of the spectral shape was divided by the jnd step of duration, giving a ratio indicating listeners' relative auditory sensitivity to the spectral shape and duration along the tense to lax vowel continuum. A higher ratio indicates better discrimination of the duration relative to the spectral shape, and a

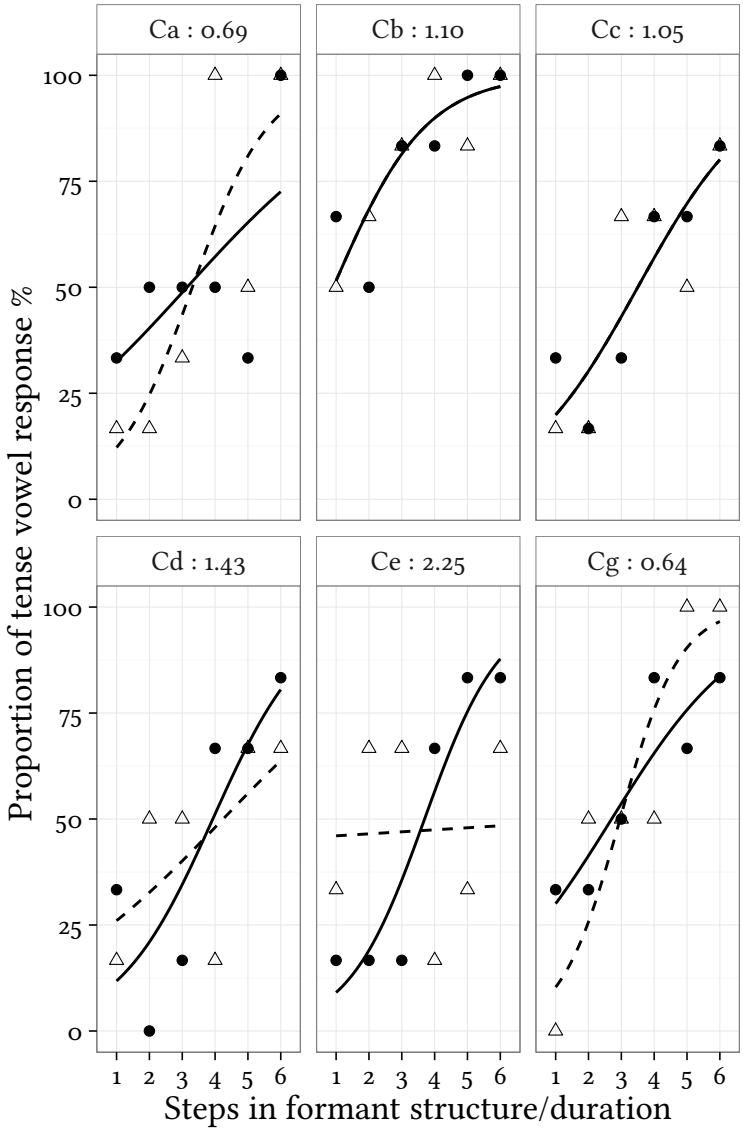

Figure 3.9: CI listeners' proportion of tense vowel responses word stimuli. The filled circle (●) is the averaged proportion response for each step in formant structure, and the hollow triangle (△) is the averaged proportion for each step in duration. The filled line (–) is the logistic regression fit to the proportion of tense vowel responses using steps in formant structure, and the broken line (- -) is the logistic regression using steps in duration. The top panel shows listeners' code and their cue weighting ratio.

lower ratio indicates better discrimination of the spectral shape relative to the duration.

3.2.2 Statistical analysis and results

The first aim of the analysis is to compare the cue weighting strategy between CI users and NH listeners with CI simulations (in Experiment 1). An independent two sample t-test showed no significant difference between the cue weighting ratio of CI users and NH listeners (t=-1.07, p>0.05; CI: $\bar{x}$=1.47,σ=0.92; NH: $\bar{x}$=1.08, σ=0.49), suggesting that both groups have relatively more weighting on the formant structure cue than the duration cue, but no significant difference between two groups. To compare their perceptual weightings on the two cues, another two t-tests were performed on the coefficients of the two cues. No significant difference was found between two groups of listeners for the formant structure cue (t=-0.24, p>0.05) and duration cue (t=0.70, p>0.05).

The second aim of the analysis is to investigate whether CI users sentence recognition performance in quiet is related to their relative perceptual weighting on the spectral and temporal cue, and their auditory sensitivity to the spectral shape and duration of the vowels tested. A scatterplot matrix is shown in figure 3.10, showing the relation among sentence recognition, jnd steps for spectral shape and duration, and coefficients for formant structure cue and duration cue. A logistic regression model was fitted to listeners' sentence recognition performance in quiet using the *cue weighting ratio* and *auditory discrimination ratio* as the independent vari-

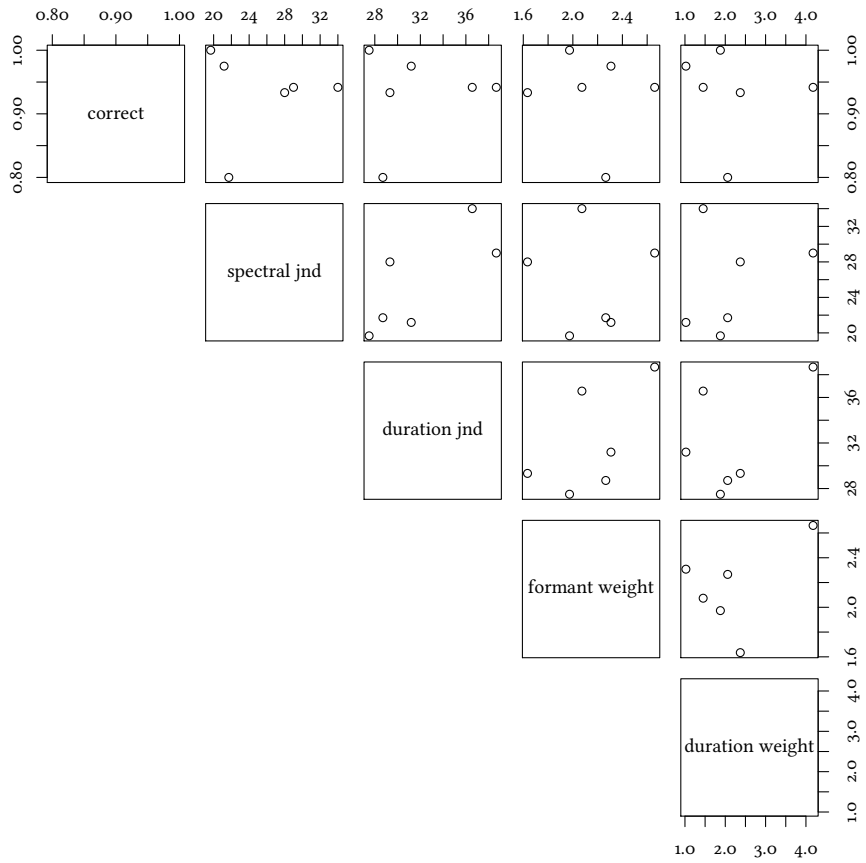

Figure 3.10: Scatterplot matrix displaying the relation among sentence recognition, jnd steps for spectral shape and duration, and coefficients for formant structure cue and duration cue.

ables. Acoustic cue weighting ratio (χ^2=0.98, df=1, p>0.05), relative discrimination ratio (χ^2=0.78, df=1, p>0.05) and their interaction (χ^2=0.81, df=1, p>0.05) were not found significant in explaining the variances in listeners' sentence recognition in quiet. Also, no significant correlation was found 1) between listeners' acoustic cue weighting ratio and relative auditory discrimination ratio (r=-2.55, p=0.05); 2) between listeners coefficients on formant structure and jnd step on formant structure discrimination (r=-1.22, p>0.05); 3) between coefficients on duration and jnd step on duration discrimination (r=0.76, p>0.05). Note that the correlation analysis might be prone to outliers and not reliable, due to the small sample size.

Thirdly, the analysis intends to explore the impact of an individual's cue weighting strategy and relative discrimination ability on listening effort during speech perception. Therefore, trials in the fixed SNR test were analysed. One logistic mixed effect model was fitted to the proportion correct levels of each trial, predicted by the fixed effect factors *pupil response (PC1)*, *auditory discrimination ratio* and *cue weighting ratio*. By using *listener* and *sentence* as random effect factors in the model, we controlled for the variability in correct levels (random intercept) and other fixed factors (random slope) that were associated with them. Factors were entered into the model in the sequence below: the model firstly started with taking *listener* and *sentence* as random intercepts; fixed effect factors and their interactions were then entered; finally, random slopes were entered into the model. Factors were retained in the model only if they significantly improved the model fitting, using Chi-squared tests based on changes in deviance. Details of the

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	-2.07	3.11	0.51			
PC1	1.93	1.16	0.10	0.41	1	0.52
Cue Weighting Ratio	-0.23	0.65	0.73	0.01	1	0.96
Discrimination Ratio	2.76	3.39	0.42	0.07	1	0.78
PC1:Cue Weighting Ratio	0.22	0.24	0.37	3.55	1	0.05
PC1:Discrimination Ratio	-2.88	1.29	**<0.05**	9.19	1	**<0.05**
Random effects:	SD	cor		χ^2	df	p
Intercept | Listener	0.64			300.80	1	**<0.001**

Table 3.4: Model parameter estimates and model comparison statistics for the best logistic mixed effect models fit to proportion correct in fixed SNR test for CI users.

best fitting logistic mixed effect model for explaining the variances in proportional correct levels in the fixed SNR test are shown in table 3.4.

There is a significant interaction between correct level and relative auditory discrimination abilities (χ^2=11.68, df=1, p<0.001). A trend analysis showed a significant linear trend of auditory discrimination ratio in its interaction with pupil response (F=8.44, d=1, p<0.01), suggesting that for CI listeners with smaller ratio (better spectral shape discrimination than duration discrimination on the tense - lax vowel continuum), bigger pupil responses were associated with better sentence recognition; and for listeners with bigger ratios (better duration discrimination than spectral shape discrimination), bigger pupil responses were associated with worse sentence recognition(β=-1.52, SE=0.56, p<0.01). This trend is illustrated in figure 3.11. There is also a borderline insignificant interaction between correct level and acoustic cue weighting ratio (χ^2=3.55, df=1, p=0.05), suggesting that as acoustic cue weighting ratio increases (more weighting on the formant structure cue), more cognitive effort might associate with better sentence correct (β=0.22, p>0.05). This borderline effect is illustrated in figure 3.12.

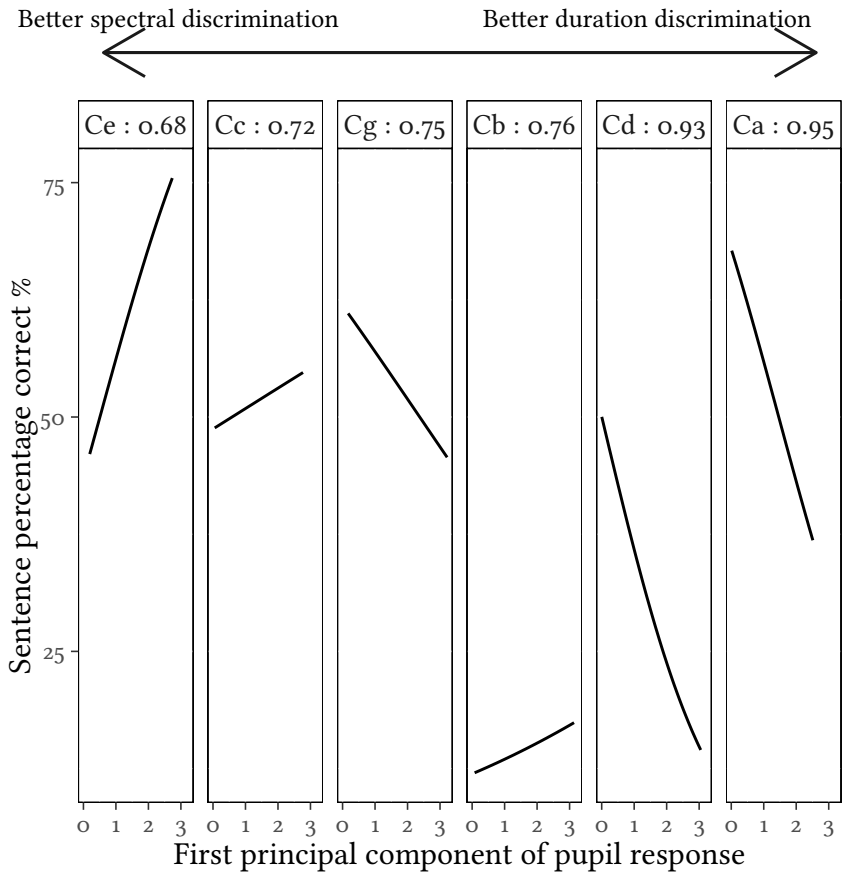

Figure 3.11: The interaction between pupil response and relative auditory discrimination ratio. Each panel displays performance of a single listener, with their relative discrimination ratio on the panel top. A bigger ratio indicates better duration discrimination, and a smaller ratio indicates better spectral shape discrimination. Black lines are the logistic regression fitted to sentence proportion correct, with pupil response as the independent variable.

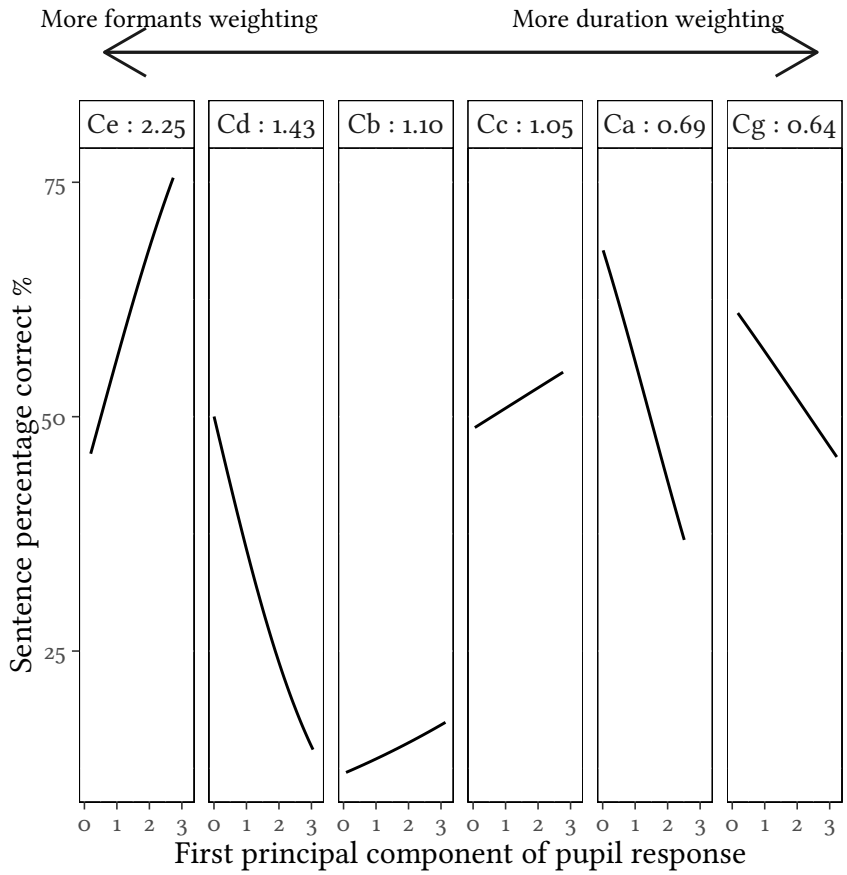

Figure 3.12: The interaction between pupil response and acoustic cue weighting ratio. Each panel displays performance of a single listener, with their acoustic cue weighting ratio on the panel top. A bigger ratio indicates more perceptual weighting on the spectral cue, and a smaller ratio indicates more perceptual weighting on the duration cue. Black lines are the fitted logistic regression on sentence proportion correct, with pupil response as the independent variable.

3.3 Discussion

This chapter investigates the impact of acoustic cue weighting strategy for NH listeners with CI acoustic simulation and CI users. Specifically, it examined how much between-individual variance in sentence recognition and listening effort could be explained by the relative perceptual attention on

the temporal and spectral cues (the vowel duration and formant structure in 'beat' - 'bit' discrimination). Listening effort was always measured at the SNR level obtained from SRT tests tracking a certain percentage of individual sentence recognition performance in quiet. This procedure intends to control for the inherent differences among listeners in their abilities and efforts to comprehend degraded sentences, in order to focus on how their acoustic cue weighting strategies affect listening effort at a level similarly difficult or easy for each participant.

3.3.1 Acoustic cue weighting strategy, auditory sensitivity and sentence recognition

This study expands findings from previous studies of cue weighting strategy by CI users and NH listeners with CI acoustic simulation. By applying greater distortion to the acoustic simulations and using the same method for calculating acoustic cue weighting ratio, this study is able to compare more closely acoustic cue weighting strategies of two groups. There is no significant difference in the perceptual weighting on the temporal and spectral cue between CI and NH listeners in word labeling task, and both groups allocate more perceptual weighting on the spectral cue. However, Hedrick and Carney (1997),Winn et al. (2012),Donaldson et al. (2015) and Moberly et al. (2016) all showed a bigger reliance on amplitude/duration cues for CI user compared to NH listeners, and a smaller reliance on dynamic/static spectral cues for CI users compared to NH listeners. This might be due to the different speech recognition levels of CI users recruited in these stud-

ies. Here, CI users' sentence recognition in quiet reaches a mean of 92.9% and standard deviation of 6.4%, suggesting that they land on the good performers' side on the continuum of CI users' speech outcome. Therefore, it might be a possibility that good CI performers also have a larger perceptual weighting on the spectral cue. However, this possibility is not testable since previous studies didn't report listeners' sentence recognition performance. Also, different acoustic cues and contrasts were used in those studies. While typically duration and amplitude cues are similarly intact, different types of spectral cues are affected differently via the CI speech processor. Although modern CI systems provide a better representation of the spectral envelope using faster stimulation rates and wider dynamic ranges, dynamic spectral cues are still poorly transmitted compared to static spectral cues. Dynamic spectral cues are the transitions in formant frequencies that arise from movement between consonant constrictions and vowel postures, and are typically shorter than the static spectral cues. Therefore, they require CI users to have better spectral acuity and integration of spectral and temporal processing. Typically CI users perform worse in discrimination or labeling tasks with dynamic spectral cues (Nittrouer et al. 2014; Donaldson et al. 2015; Moberly et al. 2016). The static spectral cue (formant structure) used in this study is more accessible and requires less auditory sensitivity, and may receive more perceptual weighting compared to the dynamic cues used in other studies (i.e., formant transition) (Winn et al. 2012; Nittrouer et al. 2014).

No significant relation between acoustic cue weighting strategy and sen-

tence recognition was found for either group of listeners (even when auditory discrimination abilities were taken into consideration for CI users). However, much stronger correlations were found in other studies between word recognition and cue weighting (for instance, r = 0.55 in Lowenstein and Nittrouer (2015), r = 0.77 in Moberly et al. (2014), r = 0.54 in Moberly et al. (2016)). This could be due to several reasons. Firstly, the current study used relative weighting between temporal and spectral cues as independent variables, instead of weighting on one cue. Temporal cues in previous studies generally didn't explain significantly word recognition, therefore, including it in the model might reduce the amount of variability explained in word recognition performance. To investigate this discrepancy, two separate logistic regression models were fitted to NH and CI listeners' sentence recognition performance in quiet using the weighting on coefficients of spectral and temporal cues as independent variables. No significant effect was found for either group of listeners. Secondly, we used open-set sentences, instead of words, for speech recognition, which includes more top-down processing. Variability among listeners at those stages was unaccounted for, but could modulate the relationship between cue weighting pattern and speech recognition significantly.

For CI listeners, their auditory sensitivity to the nonword stimuli's spectral shape and duration underlie tense to lax word labeling and sentence recognition. The size of the jnd's for spectral shape and duration suggests that on average CI users could successfully discriminate two steps along the 'beat' - 'bit' word continuum using the spectral shape difference,

and three steps using the duration difference. Although a better spectral shape discrimination than duration discrimination score for CI users seems counter-intuitive, this is mainly due to using the step size in between the 'beat' - 'bit' continuum instead of the physical stimulus values as the unit. No evidence was found to suggest that CI users' auditory sensitivity to the acoustic cues was linked to their general speech perception performance. Also, listeners' acoustic cue weighting strategy was not explained by their auditory sensitivity to these acoustic cues. Two other studies with larger sample sizes (Nittrouer et al. (2014) tested 51 CI children and Moberly et al. (2016) tested 34 CI adults) also showed no significant relation between non-speech auditory discrimination on duration/spectral cues and acoustic cue weighting strategy, or between auditory sensitivity and word recognition. But for listeners who were able to access the dynamic spectral cue, better sensitivity to the spectral cue is linked to better word recognition. This seems to suggest that having access to certain cues doesn't always guarantee that CI listeners will allocate perceptual attention on them for word labeling and speech recognition. Factors other than auditory saliency contribute to CI users' attention allocation strategies. For instance, whether or not listeners have enough language exposure to develop the optimal perceptual weighting strategy for that language seem to affect both their perceptual weighting on speech cues and speech recognition. It was reported that CI users who have developed hearing loss later in life are more likely to weight spectral cues heavily and have better word recognition, probably because they are more likely to have developed an efficient per-

ceptual weighting strategies similar to NH listeners (Moberly et al. 2014).

3.3.2 Perceptual difficulty and listening effort

Different perceptual difficulties were shown in Experiment 1 to affect the efficiency of using cognitive effort in speech recognition. Listening effort didn't differ significantly between easy (SNR80%) and hard conditions (SNR40%) in this study (for mean pupil dilation β=-0.003, SE=0.01, p>0.05; for peak pupil dilation β<0.03, SE=0.03, p>0.05; for latency response β=0.28, SE=3.67, p>0.05). Instead, it was the gain in the behavioural outcome that was affected: in the easy condition, listeners need relatively less cognitive effort to improve sentence recognition scores compared to the difficult condition. This suggests that higher perceptual difficulty didn't load listeners directly with greater listening effort, but made the resources allocated for speech recognition less efficient, possibly a sacrifice for coping with the extra noise. Intuitively, this seems to be in contradiction to previous studies on listening effort. Typically, more difficult conditions (for instance with lower SNRs, poorer auditory spectral resolution, or more competitive lexical competition) were reported to incur greater listening effort (Koelewijn et al. 2012; Kuchinsky et al. 2013; Zekveld & Kramer 2014; Winn et al. 2015). This was not found in this study, although the SNR difference between the two conditions was big enough according to previous literature (SNR40%: mean = 9.7 dB, SD = 5.3 dB; SNR80%: mean = 16.1 dB, SD = 6.7 dB). This discrepancy could be attributed to the method of obtaining the appropriate level for testing. SRT tests were set in this study

to find the threshold with reference to listeners' individual performance in quiet, instead of assuming that all listeners had similar slope and ceiling in their psychometric functions for degraded speech perception. Therefore, the threshold level would match more closely their real degraded speech perception proficiency, compared to experiments tracking an identical performance level across all participants (Zhang et al. 2014). Considering the great variability in both speech perception and listening effort for both NH listeners with CI simulation and CI users, this method should be more sensitive in unmasking the modulating effect of individual differences.

3.3.3 Acoustic cue weighting strategy and listening effort

The results showed a complex relation between sentence recognition and listening effort. For both groups of listeners, pupil response as the main factor was not found to be significant in explaining the variance in sentence recognition score, even after controlling for inherent individual differences in intelligibility and auditory discrimination abilities during experimental design and statistical analysis. This suggests that sentence recognition and listening effort are not related directly, therefore, potentially for some listeners, a low speech perception performance might be accompanied by high listening effort. Previous studies have suggested some individual differences at auditory, linguistic or cognitive levels that could contribute to the variation in performance, for instance, auditory spectral resolution, hearing loss, working memory and attention allocation (Zekveld et al. 2011; Koelewijn et al. 2012; Kuchinsky et al. 2013; Zekveld & Kramer 2014; Winn

et al. 2015). In this study, the perceptual cue weighting strategy was shown to affect the relation between sentence recognition and listening effort for each listener.

This was demonstrated by a significant linear trend of cue weighting ratio in its interaction with mean and peak pupil response in predicting sentence recognition score in Experiment 1 for NH listeners with CI acoustic simulation. For listeners with more weighting on the temporal cue, bigger pupil responses (indicated by bigger mean pupil dilation and peak pupil dilation while listening to sentences) were associated with poorer sentence recognition; and for listeners with more weighting on the spectral cue, bigger pupil responses were associated with better sentence recognition. It seemed that listeners weighting more on the spectral cue than the temporal cue had an advantage in the efficiency of using listening effort for degraded speech perception. In other words, they were able to enhance their behavioural performance by investing more effort. In comparison, listeners weighting more on the temporal cue failed to gain the same benefit from increasing listening effort. It was as if that they were 'wasting' their cognitive resources by directing them to less useful information. Therefore, using the same amount of cognitive effort, listeners weighting more on the spectral cue would score relatively higher in sentence recognition, because they are able to focus cognitive resources on more informative and reliable information for the task.

Even with the small number of CI participants, CI users' auditory sensitivity is shown to affect significantly how efficiently they use cognitive

resources for speech recognition, indicated by the significant linear trend of auditory discrimination ratio in its interaction with pupil response. It shows that CI listeners who have better spectral shape discrimination are able to increase their speech performance by investing in more cognitive resources. This is not surprising, since spectral information is important for speech perception but seriously diminished in the face of hearing loss and subsequent implantation. CI users' access and auditory sensitivity to the spectral information would be strongly associated with the success of implants and the remaining integrity of listeners' auditory system. Previous studies have shown that having access to dynamic spectral cues is linked with better word recognition (Nittrouer et al. 2014; Moberly et al. 2016). Therefore, it might be the case that good CI performers are also those who can effectively use their cognitive resources. There is also a borderline significant interaction between acoustic cue weighting strategy and pupil response, suggesting that CI users' acoustic cue weighting strategy might affect listeners' efficiency of using cognitive resources for speech recognition. This insignificant effect shares the same trend with NH listeners, that is listeners who weight more on the formant structure cue benefit in sentence recognition from investing cognitive resources. This suggests that how CI users' allocate perceptual attention to different acoustic cues could still be an important individual feature explaining variances in speech recognition and listening effort, especially considering that this strategy is independent of auditory sensitivity to the acoustic cues involved.

This finding supports what has only been suggested in past studies. The

common perceptual weighting strategies observed from mature language users for their native language are acquired through years of language exposure (Nittrouer et al. 2014; Lowenstein & Nittrouer 2015). They should, in principle, utilise the most linguistically informative and reliable acoustic cues for a certain language. Indeed, some studies have shown that this strategy is predictable through unsupervised learning based on a weighting-by-reliability principle (McMurray et al. 2009; Toscano & McMurray 2010). These studies support, indirectly, that acoustic cues favoured by NH listeners are ideal for speech perception. It is possible that this strategy could also be beneficial cognitively. Potentially, it could be the easiest route to successful speech recognition, requiring less processing time and effort. One cognitive benefit illustrated by this study could be that a better strategy allows cognitive resources to be used more efficiently for speech perception. By allocating cognitive resources on informative and reliable acoustic cues, listeners are prioritising information that could lead to fewer ambiguities and variability in restoring phonemic structure from acoustic inputs. Therefore, they should need less cognitive effort, compared to listeners who don't employ this strategy, to support a certain level of speech perception. This difference in the efficiency of using cognitive effort, rather than just the total amount of effort expended, could lead to high fatigue level for some CI users. Those users who fail to employ an optimal speech perception strategy might try constantly and actively to allocate more cognitive resources for a conversation, intending to support a better speech communication. However, little benefit could be gained from this investment. The cognitive

resources wasted in speech communication would then negatively impact one's ability to perform other mental operations, causing fatigue generally in life (McGarrigle et al. 2014). This might explain the lack of correlation between quality of life measurements and behavioural or cognitive factors in some studies, since they might overlook this complex relation during speech communication (Capretta & Moberly 2016b). Therefore, how efficiently listeners use their limited amount of cognitive resources might be more relevant when explaining individual differences in listening effort and speech perception performance. Essentially, expending listening effort for speech recognition is no bad thing, but expending more listening effort without much improvement in speech recognition is potentially exhausting for listeners.

The results of this study suggest there is a good reason to look at CI users' listening strategy, and the possibility of retuning perceptual attention to acoustic features that best facilitate language processing, even though these acoustic cues might only be coarsely transmitted through CIs and of low auditory saliency to listeners. An optimal weighting strategy might provide listeners with better speech recognition and more manageable listening effort. In this study using the English tense to lax vowel continuum, listeners (using CI simulation) weighting more on the spectral cue were more efficient in using their cognitive resources for sentence recognition, with proportionate listening effort input and speech recognition gain. However, listeners weighting more on the temporal cue seemed to expend effort, but without much benefit to sentence recognition. Although results are less

conclusive for CI users, previous studies with larger sample sizes suggest better word recognition for CI users with a similar weighting strategy as NH listeners. It is very likely that this benefit could extend to their efficiency in using cognitive resources for speech recognition. Therefore, work with more CI participants and both cognitive and behavioural assessments are needed. Furthermore, although ongoing work to enhance the accuracy of spectral information transmission and rehabilitate listeners' auditory abilities for CI users is important, it is still not sufficient to restore a good listening strategy. The availability of acoustic cues doesn't automatically translate to listeners' reliance on them, as suggested by this study and previous ones (Nittrouer et al. 2014; Lowenstein & Nittrouer 2015; Kong et al. 2016; Moberly et al. 2016). Therefore, more active and targeted training on listeners' perceptual attention should be considered and they should focus on helping listeners to acquire efficient listening strategies (specific to their language).

The next two chapters investigate the malleability of listeners' acoustic cue weighting strategy through training. Both infants and adults have been shown to harness statistical regularities of speech cues to acquire new visual and auditory knowledge (Fiser & Aslin 2002; Wanrooij et al. 2013). Specifically, distributional training, which exposes listeners to a series of stimuli varying in a speech cue that has its frequency distribution following that of a new language, helped listeners to learn non-native speech sound categories (Ingvalson et al. 2012; Escudero & Williams 2014). This might also be useful in CI rehabilitation. By constructing training stimuli that

have statistically-accentuated spectral features, listeners might be able to gain more reliance and attention on them. Ultimately, the ideal outcome of rehabilitation is to make CI users better in both speech communication as well as managing listening effort in everyday life. With training schemes targeting different aspects of language processing, CI users should be able to benefit more from the prosthesis and lead a better life.

Part 3

The impact of distributional training

Chapter 4

NH listeners' cue weighting plasticity in CI acoustic simulation

NH adults have shown to be able to harness statistical regularities of speech cues to acquire new visual and auditory knowledge (Fiser & Aslin 2002; Wanrooij et al. 2013). For instance, NH listeners can learn non-native speech sound categories through exposure to stimuli varying in a speech cue that has its frequency distribution following that of a new language (Ingvalson et al. 2012; Escudero & Williams 2014). NH listeners can also adjust reliance on a speech cue depending on its relative probability distribution for a word or speech category (Clayards et al. 2008b; Toscano & McMurray 2012). However, it is unclear whether listeners can still utilise the statistical information of degraded and distorted speech cues. This experiment intended to investigate the plasticity of cue weighting strategies of NH listeners with spectrally degraded and distorted stimuli (12-band noise-vocoded and 4mm shifted) and the effect of distributional training on the plasticity. It was hypothesised that listeners who were exposed to stimuli

with speech cues resembling the statistical regularity in the undegraded speech would establish the representation of speech categories from the degraded and distorted acoustic inputs much more easily.

4.1 Methods

4.1.1 Participants

20 normal-hearing native standard Southern British English speaking adults were recruited via the UCL Psychology Pool. All participants were aged between 18 and 45 and had normal hearing (defined as hearing thresholds of 20dB HL or better between 250 - 8000 Hz tested at octave frequencies). None of them had prior experience with vocoded speech. For their contributions, they were paid at the rate of 7.5 pounds per hour. All participants consented to take part by reading and signing a consent form, as approved by the UCL Research Ethics Committee.

They were then randomly assigned to either the experimental group or the control group, with ten participants in each.

4.1.2 Stimuli

All stimuli were 12-band noise-vocoded and 4mm upward shifted, using the same method and parameters in chapter 2.

Testing sentences were pre-recorded BKB sentences, sampled at the rate of 44.1 kHz and spoken by a male native speaker of British English (Bench, Kowal, & Bamford 1979).

Context	Uniform distribution range
bit - beat	1.92 - 3.34
fit - feat	1.96 - 3.36

Table 4.1: Model parameters of the uniform distribution fitted to the F2/F1 ratio of the recorded monosyllable words of each consonant context.

For testing listeners' acoustic cue weighting strategy, the same four continua of stimuli as in Experiment 1 chapter 2 were used. They were monosyllabic words containing lax to tense vowels /ɪ/ and /i/ ('bit' - 'beat', 'sit' - 'seat', 'pit' - 'peat', 'fit' - 'feat'), varying orthogonally in vowel duration and formant structure. Each step in F2/F1 ratio was paired with each step in vowel duration using PSOLA, thereby giving altogether $6 \times 6 \times 4 = 144$ stimuli.

Training materials were two continua of words ('bit'- 'beat' and 'fit' - 'feat'), selected from the same set of synthesised words in Experiment 1 chapter 2. For the experimental group, 60 word tokens in each word continuum were selected based on the bimodal Gaussian distribution fitted to the F2/F1 ratios of the recorded words in Experiment 1 (see details in table 2.1). Each of these words has F2/F1 ratio with equal intervals in the probability density of the bimodal distribution, a similar procedure as in Wanrooij and Boersma (2013). Therefore, all training words have different F2/F1 ratios, but still preserve the same statistical feature of the original recorded word continuum. For the control group, 60 word tokens in each continuum were selected based on the uniform distribution fitted to the F2/F1 ratios of the recorded words in Experiment 1 chapter 2. Parameters of the uniform distributions are shown in table 4.1 Each of these word tokens have F2/F1

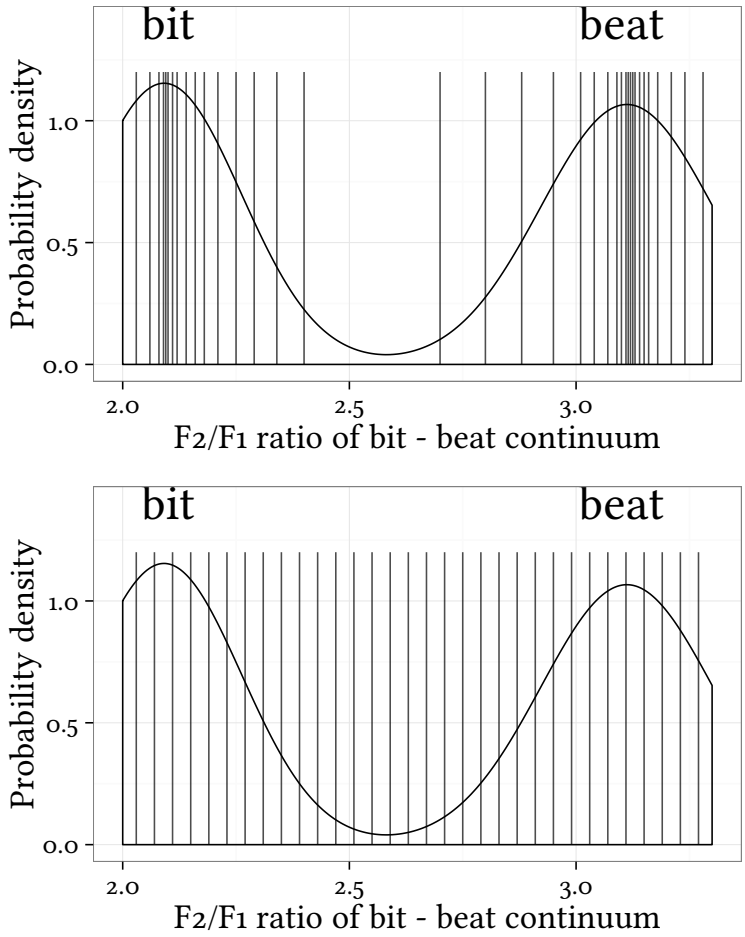

Figure 4.1: The curve is the bimodal distribution fit to the F2/F1 ratios of the recorded 'bit' - 'beat' words. The grey lines correspond to the F2/F1 values of the selected words for training. The top panel illustrates the selection for the experimental group. The under-curve area between consecutive lines is identical, therefore, values close to the mean are selected more often and values far from the mean are selected less often. The bottom panel illustrates the selection for the control group. Values are selected equally along the continuum, therefore, values close to the mean and far from the mean are selected with the same frequency.

ratio with equal intervals in the probability density of the uniform distribution. An illustration of selecting training words for both training groups is shown in figure 4.1. Then, the durations of all word tokens were manipulated with PSOLA, according to the predictions from linear regressions in table 2.2. With this design, in the experimental group, the sampling of the formant structure cue in the word stimuli was manipulated in a

way that it followed a bimodal distribution that resembled natural speech. This was to increase the statistical saliency of the spectral cue, in order to encourage listeners to re-tune their attention towards it after spectral degradation. For the control group, the sampling of the spectral cue did not allow the construction of reliable speech categories. A 4mm upward shift significantly impacted listeners' categorisation and the use of the spectral cue, as indicated in chapter 2. Therefore, the difference in the acoustic cue weighting strategy after the training between the two groups would indicate how much listeners have utilised the statistical regularities of the training material for recovering from the signal degradation and distortion.

4.1.3 Procedure

Experiments were conducted in a quiet room. Auditory materials were presented over Sennheiser HD 25 SP headphones and programs were run on a PC installed with custom MATLAB 2013b software.

Before the first session, participants were introduced to the testing and training software with unprocessed speech materials.

They were firstly tested with four randomly selected lists of BKB sentences. During each trial the number of keywords correctly reported by participants was noted down. Then, participants' cue weighting strategies were measured with a word labeling task. After hearing a word token, they were instructed to choose what they heard on the screen from either 'beat' or 'bit'. Their response was recorded by the computer and the next trial started.

Training sessions were conducted on two consecutive days, each lasting for approximately one hour. Participants were presented acoustically one word randomly selected from the stimulus set and visually four words on the computer screen, two foils and two containing targeted tense/lax vowels. They were then instructed to select from those four choices the one they had heard. After registering the response, the program proceeded to the next trial without feedback. All together, participants were exposed to two sets of stimuli ('beat' - 'bit' and 'feat' - 'fit'), repeated nine times each (randomised), totalling to $2 \times 60 \times 9 = 1080$ words. The second session on the next day was conducted with the same number of stimuli but in a different sequence.

After finishing two training sessions, all participants were tested with another four different lists of BKB sentences. Finally, they were tested with the word labeling task using the same set of stimuli as in the pre-training tests, but in a different sequence.

All participants finished the two training sessions and two testing sessions.

4.2 Statistical analysis and results

To investigate whether different types of word training had an impact on listeners' open-set sentence recognition, a mixed effect logistic regression model was built with *sentence correct* as the dependent variable in R with lme4. Random effect factor *listener*, and fixed effect factors *training session* (before or after the training), *training group* (experimental or con-

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	1.79	0.37	**<0.05**			
Type(untrained)	0.08	0.47	>0.05	7.22	1	**<0.001**
Session(post)	0.66	0.47	>0.05	13.07	1	**<0.001**
Type(untrained):Session(post)	1.93	0.68	**<0.05**	7.92	1	**<0.001**
Random effects:	SD	cor		χ^2	df	p
Intercept | listener	0.57			170	1	**<0.001**

Table 4.2: Model parameter estimates and model comparison statistics for the best mixed effect model fit to acoustic cue weighting ratio. The reference level for the categorical factor training session is *pre*, and for training type is *trained*.

trol group) were entered into the model in the same way as in the previous two chapters. Only *training session* was found significant (χ^2 = 158.82, df=1, p<0.01), suggesting that regardless of the training materials, listeners' sentence recognition performance increased after the training (β=0.19, SE=0.02, p<0.01).

Secondly, to investigate whether exposure to the different sampling of words had an impact on listeners' acoustic cue weighting strategy, a mixed effect linear regression model was fit to the *cue weighting ratio*. Listeners' acoustic cue weighting ratio was calculated in the same way as in the previous two chapters. Therefore, a bigger ratio suggests more perceptual weighting of the formant structure cue, and a smaller ratio suggests more perceptual weighting of the duration cue. Random effect factor *listener*, and fixed effect effect factors *training session* (before or after the training), *training group* (experimental or control group) and *training type* (whether the continuum was trained or not) were entered into the model in the same way. Details of the best fitting model are displayed in table 4.2. Post-hoc Wald tests showed that after training, listeners increased significantly their

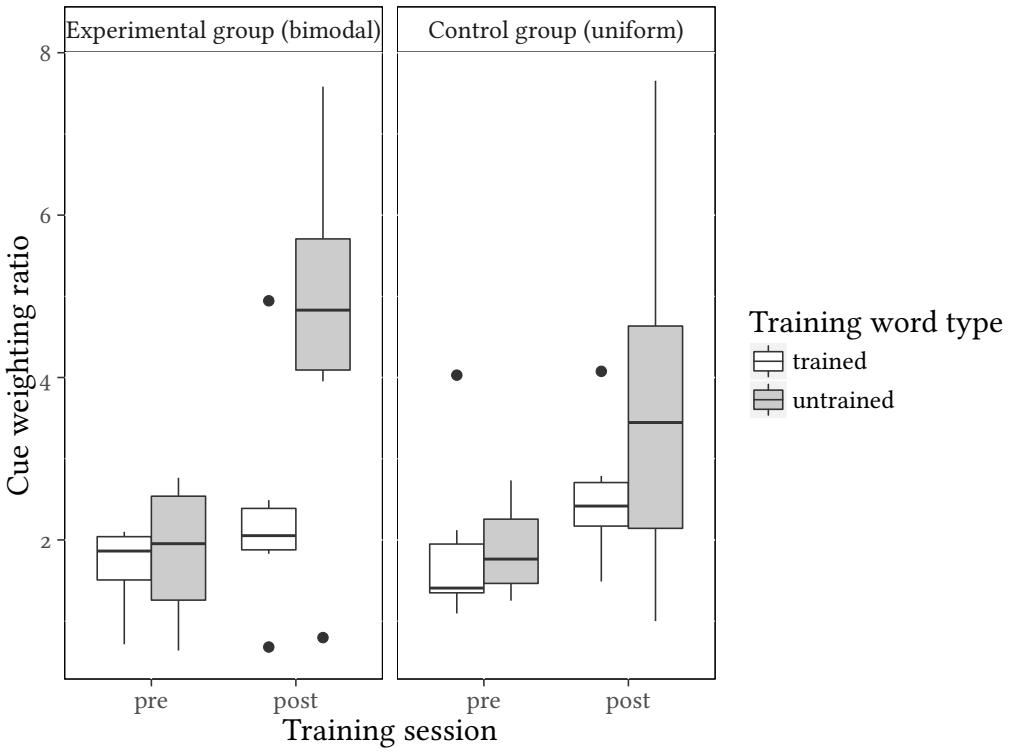

Figure 4.2: Boxplot showing the effect of training session, training group and training type on listeners' acoustic cue weighting ratio.

cue weighting ratio (β=1.63, SE=0.29, p<0.05), suggesting that they put relative more perceptual weighting to the formant structure cue after more exposure to the degraded words. Word continua that were not in the training stimuli had higher ratios than the continua used in the training (β=1.04, SE=0.34, p<0.05), but only after the training (β=2.01, SE=0.49, p<0.05). This interaction is illustrated in figure 4.2.

Finally, to investigate how the sampling of speech cues affects listeners' perceptual weighting of the formant structure and duration cues, another two mixed effects linear regression models were fit to the *formant cue coefficients* and *duration cue coefficients*, using the same fixed effect and random effect factors. Details of the best fitting model on the formant structure cue

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	5.46	0.63	**<0.05**			
Type(untrained)	-0.11	0.57	>0.05	3.56	1	>0.05
Session(post)	1.52	0.70	**<0.05**	11.09	1	**<0.001**
Group(control)	0.33	0.79	>0.05	0.57	1	>0.05
Session(post):Type(untrained)	1.91	0.82	**<0.05**	5.31	1	**<0.05**
Session(post):Group(control)	-1.60	0.82	>0.05	4.06	1	**<0.05**
Random effects:	SD	cor		χ^2	df	p
Intercept | listener	0.94			150	1	**<0.001**

Fixed effects:	β	SE	p	χ^2	df	p
Intercept	1.22	0.07	**<0.05**			
Session(post)	0.12	0.70	>0.05	11.09	1	**<0.001**
Group(control)	-0.01	0.09	>0.05	0.10	1	>0.05
Session(post):Group(control)	-0.30	0.13	**<0.05**	5.88	1	**<0.05**
Random effects:	SD	cor		χ^2	df	p
Intercept | listener	0.05			121	1	**<0.001**

Table 4.3: Model parameter estimates and model comparison statistics for the best mixed effect models fit to the formant structure (top table) and duration coefficients (bottom table). The reference level for the categorical factor training session is *pre*, for the reference level for training type is *trained*, and for training group is *experimental*.

and duration cue coefficients are in table 4.3. Post-hoc Wald tests showed that the formant structure coefficients increased significantly after training (β=1.67, SE=0.41, p<0.05), suggesting that listeners put more reliance on the spectral cue generally after the training. The two significant two-way interaction indicated that untrained word continua had bigger coefficients than trained word continua after the training (β=1.80, SE=0.59, p<0.05), and in the experimental group coefficients are bigger after the training (β=2.47, SE=0.58,p<0.05). This interaction is shown in figure 4.3. Coefficients of the duration cue were significantly bigger after the training in the experimental group than the control group (β=0.30, SE=0.10, p<0.05). This interaction is

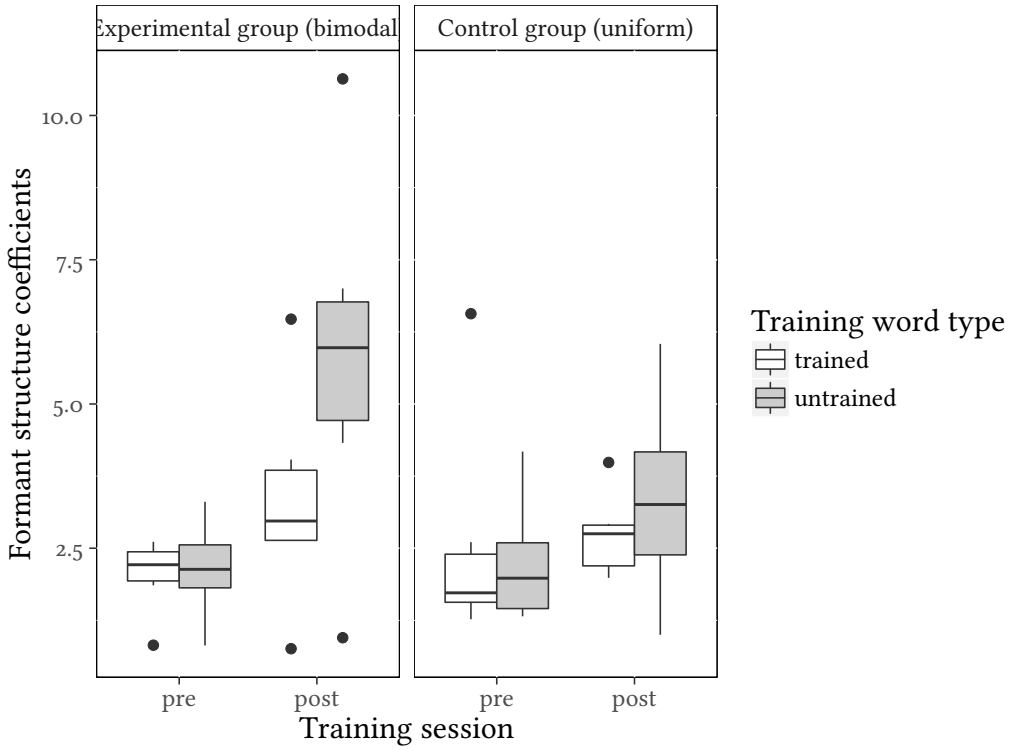

Figure 4.3: Boxplot showing the effect of training session, training group and training type on the coefficients of the formant structure cue.

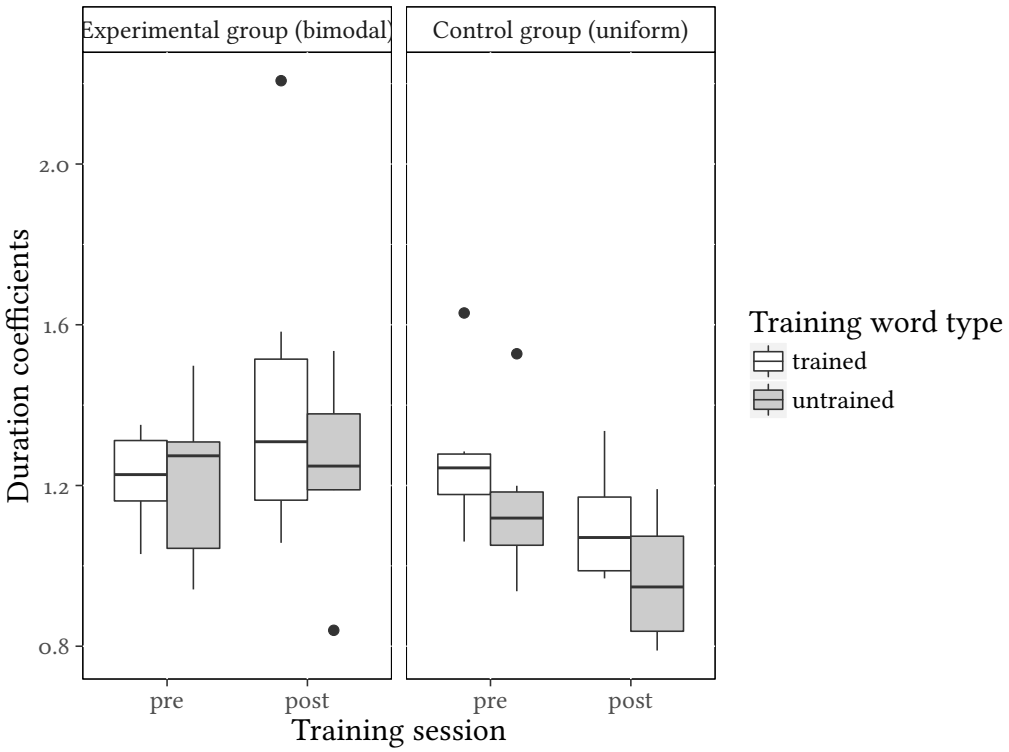

Figure 4.4: Boxplot showing the effect of the training session, training group and training type on the coefficients of the duration cue.

shown in figure 4.4.

4.3 Discussion

This experiment intends to investigate whether NH listeners could use the statistical regularity of degraded speech cues to restore the mismatch between the acoustic inputs and the phonemic representations. Two groups of listeners were trained with the same set of word continua, but differing in the sampling of the speech cues. The experimental group listened to words that have their spectral and duration cues closely matching the originally recorded words, while the control group listened to words that

have speech cues equally sampled over the continuum.

Although listeners' open-set sentence recognition increased after training, there was no significant difference between two training groups. Therefore, the improvement is probably due to more exposure to the degraded acoustic inputs, instead of the training.

Along with the improvement in sentence recognition, listeners also allocated more weighting to the formant structure cue than the duration cue after the training. It seems to suggest that as listeners recover from the initial signal degradation, they also regain the acoustic cue weighting strategy used for undegraded speech (more perceptual weighting to the spectral than the temporal cue). This expands the previous literature in the investigation of the audio training effect that mainly focused on the change in speech recognition performance. For instance, Rosen et al. (1999) trained NH participants with 4-band 6.46mm shifted noise-vocoded speech for around three hours, and showed that all test scores (BKB sentences, vowel and consonant recognition) improved significantly, but only the performance of consonants reached a similar level with the unshifted NV speech at the end of the training. Stacey and Summerfield (2008) trained NH listeners with 6mm shifted 8-band noise-vocoded sentences, words and consonant-vowel syllable pairs. They showed no significant effects of phonetic training on sentence tests, and none of both sentence and phonetic training on consonant and vowel tests. Listeners' strategies are typically not investigated in training studies with degraded speech, but it could provide important information on listeners' individual differences from training.

For instance, Wanrooij et al. (2013) showed that although both pre-training high performers and low performers improved their word categorisation after the training, they relied on different cues to achieve the improvement. However, it is unclear from the current study whether this change in acoustic cue weighting strategy is due to the manipulation of training stimuli. There was no significant interaction of training group effect on the change in acoustic cue weighting ratio before and after the training, suggesting that the sampling of speech cues in the training materials didn't change listeners' reliance on them. This might be due to an increase in the coefficients of both speech cues in the experimental group after the training, making the change in the ratio of these two cues insignificant. Listeners might be assisted by the 'peaks' of both formant structure and duration cues in the experimental group. Typically, when one acoustic property is measured across many tokens of a speech sound category, most values are likely to cluster around a central value (as illustrated in figure 2.1) (Lisker & Abramson 1964; Newman et al. 2001; Lotto et al. 2004). The number of clusters would be indicative of the number of speech sound categories. Past studies suggest that both adult and infants are able to observe the number of clusters of speech cues and utilise this information to construct novel speech categories. For instance, the distribution of phoneme /a/ F1 values in Spanish centers at 12.9 ERB, with only one peak representing one phoneme category. Around similar values, Dutch contains two peaks, centering at 12.2 ERB and 13.6 ERB, representing the phonemes /a/ and /ɑ/. After exposure to words with a bimodal sampling of F1 cues similar to that

of Dutch, Spanish listeners improved significantly in the discrimination and categorisation of words with /a/ and /ɑ/, compared to the listeners who were exposed to a unimodal sampling of F1 cues (Wanrooij & Boersma 2013; Wanrooij et al. 2013, 2015). The 4mm spectral shifting used here induced a significant shift of F1 F2 values out of listeners' experience as indicated in chapter 2, making it difficult for listeners to use this spectral cue for delimiting the tense and lax vowel boundary. With exposure to the bimodal sampling of the shifted spectral cues (Experimental group), listeners could be quicker to adjust the vowel boundary by observing the clustering in the acoustic inputs and recover from the distortion. For the group trained with no peaks in the spectral cue distribution (Control group), there is no clustering of speech cues to direct listeners' attention to the new vowel boundary. This is consistent with the findings of the current experiment. Listeners increased their perceptual attention on both the formant structure and duration cues after the training in the Experimental group, and the increases were significantly bigger compared to the Control group, since both cues clustered around the typical values of tense and lax vowels in the Experimental group but there were no cluterings of speech cues in the control group.

However, the increase in the coefficients of the formant structure cue was bigger for the untrained word context than the trained word context. This is inconsistent with other training studies, which typically observe similar or bigger improvements for the trained stimuli (Fu et al. 2004; Fu, Nogaki, & Galvin III 2005; Shafiro et al. 2012). Although the 'sit' - 'seat' and 'pit'

- 'peat' continua are not in the training materials, their vowel F2/F1 ratios are within the same range as the trained 'bit' - 'beat' and 'fit' - 'feat' continua (table 2.1). It might be possible for listeners to apply the new vowel boundaries onto the novel consonant contexts. Meanwhile, training only contains two continua and the procedure is repetitive. It is likely that listeners could develop fatigue over the trained continua, making their labeling performance in the post-training testing less reliable.

This study suggests that NH listeners' use of speech cues was not affected by the distributional training here that intended to allocate relative more perceptual attention on the spectral cue. Therefore, the next chapter will reduce the statistical saliency of the secondary duration cue in the training, in order to focus listeners attention to the formant structure cue. Ideally, the training would encourage listeners to employ the optimal cue weighting strategy specific for their language, and benefit listeners both behaviourally and cognitively as observed in chapter 3. CI listeners will also be tested, to investigate whether they could still utilise the statistical regularities of speech cues and change their acoustic cue weighting strategy accordingly.

Chapter 5

CI listeners' cue weighting plasticity

Chapter 4 showed that NH listeners allocated more perceptual weight to the spectral cue in the English tense - lax vowel contrast after the auditory training. It is unclear whether CI listeners would also be sensitive to the statistical regularities of speech cues and utilise this information to adapt their listening strategies to benefit speech recognition. It is also hypothesised that the benefits of auditory training not only lie in the improvement of speech recognition, but also in the improvement of the efficiency of using cognitive resources for speech recognition. Currently, typical intervention outcomes are measured as the change in speech recognition performance between follow-up sessions relative to the baseline session. However, listeners with hearing loss often report increased effort and fatigue, even when speech performance is equivalent. It would be likely that after the intervention, listeners might still suffer from increased effort despite an improvement in speech performance. Therefore, it is necessary to

compare listeners' use of cognitive resources for speech recognition before and after the intervention, using both subjective rankings and objective physiological measurements. Essentially, this provides an insight into how hard listeners work to achieve a certain level of speech performance and how the level of effort changes over training.

In this experiment, both NH (with 8-band noise-vocoded and 4mm upward shifted simulation) and CI listeners were trained with multiple word continua spoken by different talkers, which contained tense - lax vowels with formant structure cues sampled either in a pattern similar to natural speech (Experimental group) or evenly across the vowel categories (Control group). The difference in the perceptual weighting of speech cues, word and sentence recognition was compared pre- and post- training. This was to investigate whether the manipulation of the statistical regularity of speech cues in the training materials had an impact on how much perceptual attention listeners allocate to the cues for word labelling and open-set speech recognition. Meanwhile, their listening effort to achieve 50% of their sentence recognition performance in speech-shaped noise was recorded using pupillometry pre- and post- training. Similar to chapter 3, this adaptive design aimed to control for the confound of inherent intelligibility differences across listeners, so that the listening effort was measured at a level similarly difficult or easy for all participants. There was also within-individual change in sentence recognition score after the training. Therefore, this design also made sure that listening effort was measured independently from speech recognition performance.

5.1 Methods

5.1.1 Participants

22 normal-hearing native standard Southern British English speaking adults were recruited via the UCL Psychology Pool. All participants were aged between 18 and 45 and had normal hearing (defined as hearing thresholds of 20dB HL or better between 250 - 8000 Hz tested at octave frequencies). None of them had prior experience with vocoded speech.

The same 6 post-lingually deaf cochlear implant users in Experiment 2 chapter 3 were recruited. Summary information is shown in table 3.2.

For their contributions, they were paid at the rate of 7.5 pounds per hour. All participants consented to take part by reading and signing a consent form, as approved by the UCL Research Ethics Committee. They were then randomly assigned to either the Experimental or the Control group, with 11 NH and 3 CI participants in each.

5.1.2 Stimuli

Testing sentences were pre-recorded Basic English Lexicon (BEL) sentences (Calandruccio & Smiljanic 2012), sampled at the rate of 44.1 kHz and spoken by a male native speaker of British English. Sentences were manipulated using PSOLA in Praat to be of the same duration (mean = 2.02s, standard deviation = 0.24s).

To measure listeners' acoustic cue weighting strategy, the same synthesised

Speaker	Recorded words	For testing	For training
M1	pool, pull	Yes	Yes
	fool, full	Yes	No
	feet, fit	Yes	Yes
	seat, sit	Yes	No
	cart, cat	Yes	Yes
	park, pack	Yes	No
M2	pool, pull	Yes	Yes
	feet, fit	Yes	Yes
	cart, cat	Yes	Yes
W1	pool, pull	Yes	Yes
	fool, full	Yes	No
	feet, fit	Yes	Yes
	seat, sit	Yes	No
	cart, cat	Yes	Yes
	park, pack	Yes	No
W2	pool, pull	Yes	Yes
	feet, fit	Yes	Yes
	cart, cat	Yes	Yes

Table 5.1: The words recorded from different speakers (2 female W1 W2 and 2 male M1 M2) used in either the testing (as indicated under the heading 'For testing ') or the training (as indicated under the heading 'For training ').

'beat' - 'bit' word continuum as in chapter 3 was used. Each step in formant structure was paired with each step in vowel duration, making $6 \times 6 = 36$ tokens.

To construct word stimuli for testing and training, 2 male and 2 female native Southern British English speakers were recorded reading a randomised list of words containing 3 tense - lax vowel contrasts in British English: /i/ - /ɪ/, /ɑ/ - /ae/, /u/ - /ʊ/. Each word was repeated 30 times and the exact words uttered by each speaker are listed in table 5.1. For words to be used in the testing (as indicated by the appropriate column in table 5.1), 10 words in each continuum (5 with tense vowels and 5 with lax vowels) were randomly selected from the recordings. Altogether, there were

$(6 \times 2 + 3 \times 2) \times 10 = 180$ tokens in the word test.

To synthesise words for training, only pool - pull (/u/ - /ʊ/), cart - cat (/ɑ/ - /ae/) and feet - fit (/i/ - /ɪ/) word continua from the 4 speakers were used, as indicated by the appropriate column in table 5.1. The synthesising procedure was similar to chapter 4. The only difference was that this experiment used the distribution of both F1 and F2 instead of the distribution of the F2/F1 ratio when selecting endpoint tokens. Firstly, the F1 and F2 of each tense and lax vowels from each speaker were measured at 50% into the vowel using Praat and were transformed into an ERB scale. The distribution of F1 and F2 along each tense - lax vowel continuum was then separately fitted with a custom distribution that was the sum of two Gaussian distributions with equal weights. Parameters of each F1 and F2 distribution are listed in table 5.2. Based on the bivariate bimodal distribution of F1 and F2, two endpoint values were selected for each continuum that had the highest and lowest values on the cumulative density function. These two values represented the most typical tense and lax tokens respectively in each continuum. Vowel durations of the recorded words were also measured for each continuum and fitted with a linear regression as a function of F2/F1 ratio. Details of each linear fitting are listed in table 5.3. Then, 100 tokens of each continuum were synthesised using the two tokens (matched in duration and fundamental frequency) in the Tandem-STRAIGHT algorithm, and the resulting F1 and F2 were measured in the same way as above.

The selection of word tokens for training was also similar to chapter 4. For the Experimental group, 40 word tokens in each word continuum were

Speaker	Continuum	Formant	$\bar{x}_1$	σ_1	$\bar{x}_2$	σ_2
W1	feet - fit	F1	11.52	0.33	9.32	0.08
		F2	23.11	0.20	21.20	0.17
	cart - cat	F1	15.15	0.79	13.36	0.47
		F2	19.74	0.33	17.17	0.60
	pool - pull	F1	10.59	0.24	9.30	0.63
		F2	15.62	0.22	15.32	0.70
W2	feet - fit	F1	10.89	0.46	8.63	0.41
		F2	22.82	0.73	19.10	0.17
	cart - cat	F1	16.27	0.24	14.33	0.38
		F2	19.32	0.24	16.56	0.28
	pool - pull	F1	11.41	0.35	8.90	0.41
		F2	18.71	0.63	14.28	0.41
M1	feet - fit	F1	9.74	0.20	6.97	0.21
		F2	21.58	0.17	19.61	0.27
	cart - cat	F1	12.94	0.26	12.29	0.47
		F2	18.99	0.23	14.87	0.48
	pool - pull	F1	10.09	0.47	8.08	0.35
		F2	22.86	0.28	12.40	0.33
M2	feet - fit	F1	9.17	0.34	6.48	0.11
		F2	20.99	0.36	19.21	0.23
	cart - cat	F1	12.08	0.10	11.38	0.24
		F2	18.03	0.23	15.51	0.31
	pool - pull	F1	8.59	0.47	7.43	0.30
		F2	21.24	0.12	12.58	0.87

Table 5.2: Model parameters of the bimodal distribution fitted to the F2 and F1 of the recorded monosyllable words of each speaker.

Speaker	Continuum	Model	Model fit
W1	feet - fit	*duration* = 0.05 × *F2/F1* + 0.05	F(1,60)= 185.19, $p< 0.01$
	cart - cat	*duration* = -0.03 × *F2/F1*+ 0.14	F(1,60)= 102.19, $p< 0.01$
	pool - pull	*duration* = 0.16 × *F2/F1* -0.17	F(1,60)= 170.10, $p< 0.01$
W2	feet - fit	*duration* = 0.04 × *F2/F1* - 0.02	F(1,60)= 162.22, $p< 0.01$
	cart - cat	*duration* = -0.73 × *F2/F1* + 0.98	F(1,60)= 185.19, $p< 0.01$
	pool - pull	*duration* = -0.07 × *F2/F1* + 0.24	F(1,60)= 100.31, $p< 0.01$
M1	feet - fit	*duration* = 0.01 × *F2/F1* + 0.07	F(1,60)= 151.48, $p< 0.01$
	cart - cat	*duration* = -0.25 × *F2/F1* + 0.47	F(1,60)= 190.88, $p< 0.01$
	pool - pull	*duration* = -0.01 × *F2/F1* + 0.16	F(1,60)= 117.19, $p< 0.01$
M2	feet - fit	*duration* = 0.03 × *F2/F1* + 0.04	F(1,60)= 147.39, $p< 0.01$
	cart - cat	*duration* = -0.14 × *F2/F1* + 0.33	F(1,60)= 135.92, $p< 0.01$
	pool - pull	*duration* = 0.05 × *F2/F1* + 0.03	F(1,60)= 192.19, $p< 0.01$

Table 5.3: Model parameters and fittings for the linear regression fitted to the vowel duration with the F2/F1 ratio as the predictor for each recorded word and speaker.

selected based on the bivariate bimodal Gaussian distribution fitted to the recorded words (see details in table 5.2). Each of these words has F1 and F2 values with equal intervals in the probability density of the bimodal distribution. Therefore, all training words have different F2 and F1 values, but still preserve the same statistical features of the original recorded word continuum. For the control group, 40 word tokens in each continuum were selected based on the uniform distribution fitted to the F1 and F2 of the recorded words. Vowel durations were then calculated for each token based on the linear regressions fitted to the recorded words (table 5.3). For the Experimental group, the duration of the vowel in each training word was with PSOLA in Praat to the average of each tense - lax vowel continuum from each speaker, so that the vowel duration of the tense vowel was the same as that of the lax vowel for each speaker. For the control group, the duration of the vowel in each training word was manipulated to be the calculated values from table 5.3.

With this design, the sampling of the formant structure cue in the Experimental training group for the word stimuli was manipulated in a way that it followed a bimodal distribution that resembled the natural speech. Different from chapter 4, vowel durations were fixed, making the vowel duration cue invariant across the tense - lax vowel categories. This was to increase the statistical saliency of the spectral cue and decrease the saliency of the duration cue, in order to encourage listeners to re-tune their attention towards the spectral aspect. For the control group, the sampling of the formant structure cue did not reliably relate to speech categories, but

the duration cue could be used since its sampling was closer to the natural speech. In addition to the synthesised monosyllabic words, 20 common and distinctively different multi-syllabic words recorded from the 4 speakers were also included in both training stimulus sets ('chocolate', 'university', 'computer', etc.).

For NH listeners, all testing and training stimuli were 8-band noise-vocoded and 4mm upward shifted, using the same method and parameters in chapter 2.

5.1.3 Procedure

Participants were seated in a quiet room, 70 cm from a 17-inch white screen monitor and 55 cm from an infrared monocular eye-tracker (Eyelink 1000, SR Research, 500 Hz sampling rate). All audio stimuli were presented through Yamaha MS101 loudspeaker, calibrated at 72 dB SPL. The illuminance of the room was adjusted for each participant, such that the pupil diameter was midway between maximum and minimum size (elicited by turning off and on the room lighting consecutively). Experiments were run in Matlab using Psychtoolbox and custom software.

Before the first session, all participants were introduced to the testing and training software. They listened to different sets of BEL sentences (simulated speech for NH listeners and unprocessed speech for CI listeners), and then were given written and acoustic feedback. No active responses were required.

To find listeners' baseline performance before the training, they were tested

with a series of auditory tasks. Firstly they were tested with 30 randomly selected BEL sentences in quiet to obtain each individuals' speech recognition score. In each trial they listened to a sentence and the number of keywords correctly reported by participants was noted down. Their percentage correct was averaged across trials to obtain their speech recognition score. Then their listening effort was measured in the same way as in Experiment 2 chapter 3. An adaptive speech perception threshold test was performed for each listener, using speech-shaped noise to track 50% of each individuals' speech recognition score in quiet with the UML package in Matlab (Shen & Richards 2012). Their averaged recognition score was set as the upper bound of the logistic psychometric function. The prior distribution range of the psychometric function slope was set between 0.1 to 10, and the range of the threshold was set between -10 dB to 30 dB. The SNR level of the first trial was set as 5 dB. Then the number of keywords reported by the participants were entered into the function, and the SNR for the next trial was estimated online based on the sweet point for threshold estimation using the Bayesian minimum-variance procedure. The SRT test was set to converge either when the 90% confidence interval of the threshold estimation was within 3 dB or reaching the maximum trial number of 30. At the SRT level obtained from each individual, a fixed SNR speech recognition test with 20 sentences was then performed and participants' pupil responses were recorded simultaneously. Due to this control on the intelligibility across participants, each trial in the fixed SNR test was of a similar difficulty for each participant, regardless of their speech recognition

performance. The presentation of the speech-shaped noise masker started 2s before sentence onset and finished 2s after sentence offset. Participants were instructed to fixate the black fixation cross on the white monitor and avoid excessive blinks. After the masker offset, they were prompted by the colour change of the fixation cross to repeat back the sentence (see a similar example in figure 3.1). Their sentence recognition responses were scored by the experimenter and the program proceeded to the next trial. Then, participants' cue weighting strategies were measured with the same word labeling task as in chapter 3, with the same synthesised 'beat' - 'bit' word continuum. Each step in formant structure was paired with each step in vowel duration, making $6 \times 6 = 36$ tokens. Exact values of each token are illustrated in table 2.5. After hearing a word token, listeners were instructed to choose what they heard on the screen from either 'beat' or 'bit'. Their response was recorded by the computer and the next trial started. Finally, listeners' word recognition scores were obtained. After hearing a word, they were instructed to click on the word they heard on the screen, from all possible 12 words in the testing stimuli. Their response was recorded by the computer and the next trial started.

For NH listeners, training was conducted in the same room on two consecutive days, each lasting for approximately one and half hours. In each training session, participants were exposed to 3 monosyllabic word continua recorded from 4 speakers and 20 random multi-syllable words, repeated 2 times each and all randomised, totalling to $3 \times 4 \times 40 \times 2 + 20 \times 2$ = 1000 words. The manipulated monosyllable words serve as training stim-

uli to adjust listeners' perceptual cue weighting strategy, and the multi-syllable words were to check whether listeners were paying attention. Participants were presented acoustically one randomly selected word, and visually four words on the computer screen, two foils and two containing targeted tense/lax vowels. They were then instructed to select from those four choices the one they heard, or click the button 'none ' to indicate that the word was not a monosyllable. If a multi-syllable word was mistakenly recognised as one of the four choices or vice versa, listeners were prompted by a window to keep paying attention. After registering the response, the program proceeded to the next trial without feedback. The second session on the next day was conducted with the same number of stimuli but in a different sequence.

For CI listeners, the same training program was compiled in Matlab into a standalone application. It was then installed on a Windows computer tablet, which was issued to each CI participant. Altogether, there were three training sessions, each lasting for approximately one and half hours. Each session contained the same number of word stimuli, and the sequence was randomised for each participant. CI listeners were instructed to finish the three sessions over 5 days at home, and return to the lab for the post-training testing within 1 week from the baseline testing.

After finishing their training sessions, all participants were tested with the same series of auditory tasks.

All participants finished their training and testing sessions.

5.1.4 Data processing

The processing of pupillometry data was identical to Experiment 2 in chapter 3. Baseline pupil diameter in each trial was calculated as averaged pupil traces 1s before the start of the sentence. The rest of the pupil diameter measurements were divided by that baseline level to obtain the proportional pupil size change elicited by sentence recognition. Blinks and problematic trials were excluded using the same rules as in chapter 3. Altogether, 981 trials of pupil response recordings were included for analysis, with 34 trials on average for each NH participant (SD = 4) and 37 trials for each CI participant (SD = 2). All valid traces were then low-pass filtered at 10 Hz and downsampled to 50 Hz. Three indices of pupil response (mean pupil dilation, peak pupil dilation and peak latency) were obtained from processed traces, consistent with the method in Zekveld et al. (2010, 2011). Mean proportional pupil size change and peak proportional pupil size were the average and maximum of proportional pupil changes from sentence onset to response prompt, relative to the baseline pupil size. Peak latency response was the time between onset of the sentence to the peak dilation. A principal component analysis was then performed in R, with three variables scaled to the same unit of standard deviation and centering at zero. Details of the principal components are reported in table 5.4. The first principal component was mostly related to the proportional mean and peak pupil size changes, indicated by their large coefficients (0.70 and 0.70). The first principal component was selected as an index for sentence-evoked pupil

Importance of principal components:			
	PC1	PC2	PC3
Standard deviation	1.40	0.99	0.22
Proportion of variance	0.65	0.33	0.02
Rotation:			
	PC1	PC2	PC3
Latency	0.14	-0.99	-0.08
Mean proportional change	0.70	0.04	0.71
Peak proportional change	0.70	0.15	-0.70

Table 5.4: The importance of principal components and variable rotation of the principal component analysis on all valid trials.

response, since it explained the most variance in the three indices (65%).

A cue weighting ratio for each individual was calculated from the binomial response of participants in the word labelling task, using the same method as in chapter 3. Similarly, a higher ratio indicates more reliance on the spectral cue relative to the temporal cue; and a lower ratio indicates more reliance on the temporal cue relative to the spectral cue.

5.2 Statistical analysis and results

A summary of results is displayed in table 5.5. Firstly, to investigate whether training had an impact on listeners sentence recognition performance, a mixed effect linear regression model was built for NH listeners, with *sentence correct* as the dependent variable. Random effect factor *listener*, and fixed effect factors *training session* (before or after the training), *training group* (Experimental or Control group) were entered into the model in the same way as previous chapters. Only *training session* was found significant (χ^2=19.72, df=1, p<0.01) and a post-hoc Wald test indicated that after

Sentence recognition:	NH		CI	
	pre	post	pre	post
Experimental:	0.65 (±0.15)	0.72 (±0.13)	0.96 (±0.03)	0.96 (±0.03)
Control:	0.55 (±0.08)	0.68 (±0.10)	0.89 (±0.08)	0.90 (±0.02)
Word recognition:	**NH**		**CI**	
	pre	post	pre	post
Experimental:	0.45 (±0.14)	0.51 (±0.15)	0.80 (±0.19)	0.85 (±0.10)
Control:	0.39 (±0.09)	0.48 (±0.15)	0.67 (±0.13)	0.79 (±0.05)
PC1 score:	**NH**		**CI**	
	pre	post	pre	post
Experimental:	1.33 (±0.97)	0.54 (±0.51)	0.66 (±0.72)	0.36 (±0.30)
Control:	0.93 (±0.93)	0.43 (±0.32)	2.10 (±0.28)	1.15 (±1.06)
Cue weighting ratio:	**NH**		**CI**	
	pre	post	pre	post
Experimental:	1.04 (±0.29)	2.07 (±2.10)	1.33 (±0.82)	0.87 (±0.29)
Control:	1.04 (±0.52)	1.18 (±0.92)	1.06 (±0.40)	1.75 (±1.07)
Formant cue coefficient:	**NH**		**CI**	
	pre	post	pre	post
Experimental:	2.00 (±0.81)	3.15 (±2.61)	1.97 (±0.34)	2.57 (±1.23)
Control:	1.88 (±0.68)	2.28 (±0.91)	2.33 (±0.30)	3.03 (±1.14)
Duration cue coefficient:	**NH**		**CI**	
	pre	post	pre	post
Experimental:	1.89 (±0.55)	2.06 (±1.11)	1.76 (±0.68)	2.87 (±0.45)
Control:	2.11 (±0.80)	2.21 (±0.76)	2.56 (±1.43)	2.14 (±1.10)

Table 5.5: The mean and standard deviation of the testing results for NH and CI listeners.

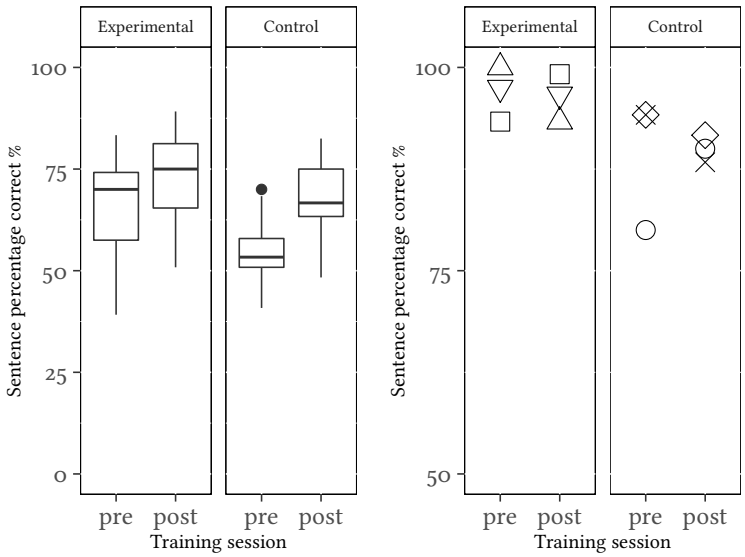

Figure 5.1: The left pair of panels shows the boxplots for NH listeners' sentence recognition scores in quiet, and the right pair of panels shows the scores of each CI participant.

training, NH listeners' sentence recognition performance in quiet was significantly better (β=0.1, SE=0.02, p<0.05). There were no improvement in sentence recognition scores after the training for CI users (pre: 0.93, post: 0.93), and between training group (Experimental pre: 0.97, Experimental post: 0.96; Control pre: 0.90, Control post:0.9). Boxplots for NH listeners and point plots for CI listeners are displayed in figure 5.1.

To investigate whether training had an impact on listeners word recognition performance, a mixed effects linear regression model was built for NH listeners, with *word correct* as the dependent variable. Random effect factor *listener*, and fixed effect factors *training session* (before or after the training), *training group* (Experimental or Control group), *word type* (whether the word was in the training or not) were entered into the model. Only *training session* (χ^2=18.61, df=1, p<0.001) was found significant. A post-hoc

Wald test showed that NH listeners' word recognition performance was significant better after the training (β=0.07, SE=0.02, p<0.05). A similar increase in the word recognition score was also observed for CI users (pre: 0.74, post:0.82).

To compare listeners' cognitive effort for sentence recognition between training groups and sessions, a mixed effect linear regression model was built for NH listeners, with the *first principal component score (PC1)* of pupil responses as the dependent variable. Random effect factor *listener*, and fixed effect factors *training session* (before or after the training), *training group* (Experimental or Control group) were entered into the model in the same way. Similarly, only *training session* was significant (χ^2=8.71, df=1, p<0.01), and the pupil dilation *PC1* was significantly smaller after the training (β=-0.64, SE=0.20, p<0.05), suggesting less listening effort after the training. Another three mixed effect regression models with the same independent variables were built, using the *mean pupil size*, *peak pupil size* and *latency response* as the dependent variable respectively. No significant effect of *training session* was found for *mean pupil size* (χ^2=0.91, df=1, p=0.35), *peak pupil size* (χ^2=2.19, df=1, p=0.14) and *latency response* (χ^2=1, df=1, p=0.32). A similar decrease in the mean PC1 score after the training was also observed for CI users (pre: 1.38, post:0.76). The pupil size variation of CI listeners before and after the training is shown in figure 5.2. A boxplot of PC1 score for NH listeners and a point plot of PC1 score for CI listeners are displayed in figure 5.3.

Finally, to examine whether the exposure to a different sampling of speech

Figure 5.2: The y-axis shows the arbitrary Eyelink 1000 camera sensor units, with reference to the number of threshold camera pixels during each recording session. The x-axis shows the time from baseline onset. The shaded region is for baseline measurement, and the rest of the pupil trace is within the analysis window starting from the offset of the baseline to the response prompt. Each panel displays the performance of a single C1 listener. The top panel shows the pupil size change before the training, and the bottom panel shows the pupil size change after the training.

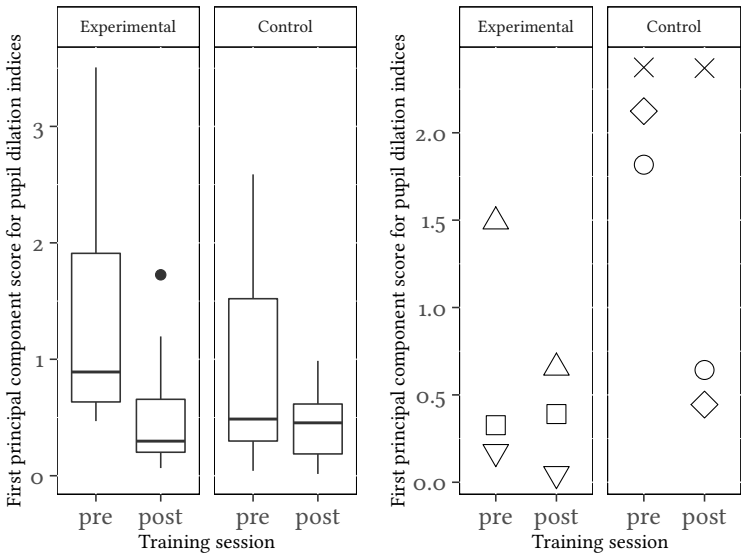

Figure 5.3: The left pair of panels shows the boxplot for NH listeners' first principal component scores for the three pupil dilation indices (mean dilation, peak dilation and latency), and the right pair of panels shows the scores of each CI participant.

cues affected listeners' acoustic cue weighting strategy, a mixed effects linear regression model was built for NH listeners, using *cue weighting ratio* as the dependent variable, with random effect factor *listener* and fixed effect factors *training session* (before or after the training), *training group* (Experimental or Control group). No factor was found significant. Generally, NH listeners in both Experimental and Control group showed an increase in the cue weighting ratio after the training (suggesting more perceptual weighting to the formant structure cue than the duration cue). However, two out of three CI listeners in the Experimental group ('Cc' and 'Ce') had an decrease in the cue weighting ratio, with an increase in the weighting of the duration cue and a decrease in the formant structures cue. This is illustrated in figure 5.4 for NH listeners and figure 5.5 for CI listeners.

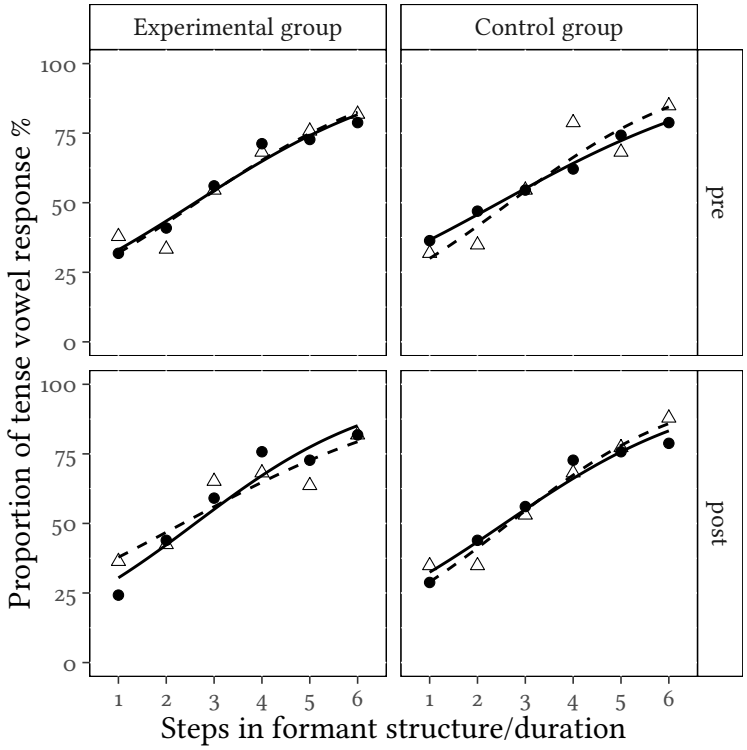

Figure 5.4: NH listeners' proportion of tense vowel responses in the word categorisation task. The filled circle (●) is the averaged proportion response for each step in formant structure, and the hollow triangle (△) is the averaged proportion for each step in duration. The filled line (–) is the logistic regression fit to the proportion of tense vowel responses using steps in formant structure, and the broken line (- -) is the logistic regression using steps in duration.

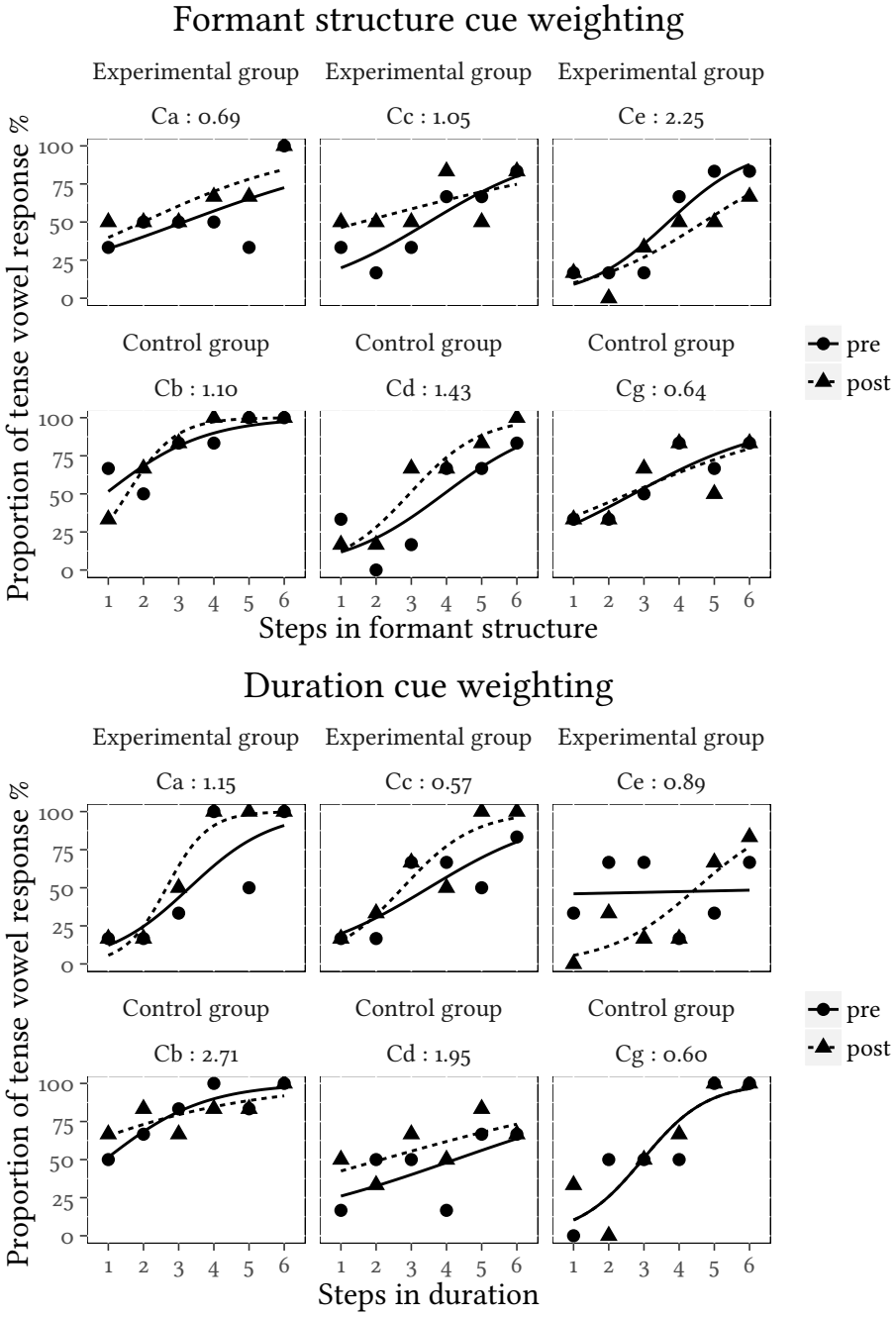

Figure 5.5: The top panel shows CI listeners' proportion of tense vowel responses with steps in formant structure, and the bottom panel shows the responses with steps in duration. The circle is the averaged proportion response for each step before the training, and the triangle is the averaged response after the training. The filled line is the logistic regression fit to the proportion of tense vowel responses before the training, and the broken line is the logistic regression fit after the training. The panel annotation indicates each listener's training group and acoustic cue weighting ratio.

5.3 Discussion

This experiment extends the findings in chapter 4 by investigating CI users' sensitivity to the statistical regularities of speech cues and comparing the amount of listening effort used for sentence recognition before and after the auditory training. While chapter 4 allowed for the variation of vowel duration in the training stimuli, this experiment fixed vowel duration in the Experimental group so that the spectral cue was the only speech cue varying across tense and lax vowel categories. It was hypothesised that by reducing the utility of the duration cues, listeners would be more likely to allocate more attention on the spectral cue, which is the cue NH listeners weighted perceptually more on. As suggested by previous literature and chapters, this change in the acoustic cue weighting strategy should benefit listeners' speech recognition and the efficiency of using cognitive resources for speech recognition.

No significant impact of training on listeners' perceptual weighting of the speech cues was found. Nevertheless, NH listeners showed a similar pattern of change as in chapter 4. There was an increase in the mean of acoustic cue weighting ratio after the training, suggesting that listeners allocated more perceptual weighting to the formant structure cue than the duration cue. And the increase was bigger in the Experimental group than in the Control group. This was due to a bigger increase in the mean of the formant structure coefficient for the Experimental group than the Control group,

while the increase in the mean of the duration coefficients was similar between two groups. This difference could be attributed to the manipulations on the statistical features of the speech cues in the training stimuli. In the Experimental group, the formant structure cue had a bimodal distribution pattern along the tense - lax vowel continuum and the duration cue was invariant; in the control group, the formant structure cue had a uniform distribution and the duration cue was bimodal. It is likely that the exposure to a bimodal distribution facilitated listeners' restoration of perceptual weighting of the damaged spectral cue in the Experimental group, since it provided clusterings along the vowel continuum that help to adjust listeners' speech category boundary. The new vowel and tense vowel boundaries were shifted beyond listeners' normal experience by the 4mm upward shifting, and listeners' sensitivity to the spectral shape was diminished by the 8 band noise-vocoding (as indicated by chapter 2). These might make the spectral cue less useful for listeners in categorising vowels. Although no past studies have introduced a similar type of distortion, some have shown that both top-down lexical tuning and statistical regularity in the acoustic cues improve the mapping of perceptually mismatched stimuli induced by foreign accent or experimental manipulations (Clayards et al. 2008a; Escudero et al. 2011; Idemaru & Holt 2011; Wanrooij & Boersma 2013; Idemaru & Holt 2014; Escudero & Williams 2014; Reinisch & Holt 2014; Ong et al. 2015). Typically in these studies, NH listeners are exposed to speech cues with a different frequency distribution than their native language, for instance a different number of clusters (the number of speech categories) and clus-

ter boundaries as in Escudero et al. (2011); Wanrooij and Boersma (2013); Escudero and Williams (2014), or a different correlation with other cues as in Idemaru and Holt (2011, 2014); Liu and Holt (2015). Before the training, listeners' use of the speech cues was dominated by that of their native language, making their recognition and discrimination of the new accent poor. After the training, their performance improved, suggesting that they accommodated to the difference and adopted the new pattern. Similarly, NH listeners in this experiment might also have observed the new distribution pattern of the spectral cue during the training and adapted it to optimise vowel categorisation and discrimination. This procedure would draw more attention to the usefulness of the spectral cue that would otherwise be down-weighted due to the spectral degradation and distortion. Meanwhile, since the duration cue is intact through the CI simulation, there is little mismatch between the immediate acoustic input and the long-term representation, so the manipulation of its statistical features might not overcome listeners' original perceptual attention to it.

Due to the small sample size, no reliable statistical test with three factors could be performed on the data of CI users. Nevertheless, they showed some changes in the reliance on the formant structure and duration cues after the training. However, different from the trend for NH listeners, two out of three CI participants in the Experimental group showed a decrease in the cue weighting ratio. Specifically, the coefficient of duration cue increased after the training, although the durations of vowels in the Experimental group training words were fixed and not reliably associated with

any speech category. It is likely that although the formant structure cue in the Experimental group formed two clusters to facilitate the construction of two speech categories, this pattern is not unusual in natural speech. But fixed duration is uncommon in naturally uttered speech, and this might attract listeners' attention to the duration cue instead. Also, the increase in the mean of the formant structure coefficient was similar between the Experimental and Control group, suggesting that the experimental manipulation on the statistical distribution of the spectral cue didn't have an impact on CI listeners' perceptual reliance. Past studies have suggested that listeners could utilise the probabilistic speech cues that contribute to the phonetic identity, and tune their weighting of these cues in order to adjust for the specific distribution of the cues in different styles of speech (Holt & Lotto 2006; Clayards et al. 2008b; Lau et al. 2016). However, there could be a number of sensory and cognitive constraints that might prevent listeners from achieving this optimised performance. For instance, a secondary task during speech perception could impact listeners' speech encoding (Mitterer & Mattys 2017). Specifically for CI listeners in this experiment, they might not have enough spectral resolution to be sensitive to the distribution of the formant structure cues in the training stimuli. Although all CI participants could perform the auditory discrimination tests in chapter 3, the tests were constrained within the frequency range of the testing stimuli. It is unclear whether they have the similar auditory sensitivity at the vowel-specific and speaker-specific frequency region in the training stimuli to detect the variation in the formant structure. CI listeners' cognitive abilities

could also affect their performance. Moberly et al. (2017) showed poorer working memory accuracy and highly impaired phonological sensitivity for CI listeners group compared to the age-matched NH listeners group. Both cognitive functions are important for processing phonological information, and the impairment could damage listeners' observation and use of the statistical properties of speech cues. All these confounds could introduce a significant amount of stress on CI listeners' phonological processing, making them unable to be as flexible as NH listeners in adjusting their perceptual weighting of the speech cues to optimise their performance.

Note, however, that these trends were not statistically significant for either NH or CI listeners. This might be due to the difference in the training and testing stimuli. While chapter 4 used the same speaker and vowel in the training and testing, this experiment used multiple vowels and talkers in training and synthesised words for testing. Similar studies typically used the same vowel or consonant from the same speaker in the training and testing (Idemaru & Holt 2011; Liu & Holt 2015; Schertz et al. 2016). Those investigating the generalisation of the training effect found that the learning effect was constrained by the perceptual space and category (Escudero & Williams 2014; Idemaru & Holt 2014; Reinisch & Holt 2014; Ong et al. 2015). Therefore, using synthesised words for testing might not fully reflect the impact of the training that contained different vowels and frequency regions. This experiment also introduced a bigger degradation in CI acoustic simulation that significantly changed listeners' use of the spectral cue (see chapter 2). Longer training might be needed before listeners could recover

the phonemic structure from the distorted acoustic inputs. Furthermore, it is possible that adults are generally not as sensitive to the statistical features of the speech cues as infants, since this low-level mechanism could be dominated by the top-down influence from the higher-level linguistic representations that are acquired and matured over years of exposure to a native language (Wanrooij et al. 2014). Passive exposure might also not be the ideal training format for adults. Adding explicit procedures along with the exposure to the statistical distribution of speech cues has been found to impact listeners' training outcome significantly. For instance, requiring active responses from listeners improved their discrimination of lexical tones, and presenting listeners with exemplars prior to the training improved their response time of syllable detection (Batterink et al. 2015; Ong et al. 2015). Explicit procedures might help to increase listeners attention to the statistical distribution of speech cues, which is a low-level feature of speech that would otherwise be outweighed by other higher-level features (lexical level, semantic context, etc.).

While it has long been recognised that listeners with hearing loss report increased effort and fatigue, interventions that aim to help listeners to benefit maximally from hearing aids seldom investigates its impact on this aspect. Speech recognition measurements are not enough to characterise listeners' speech performance. In some cases, a high level of listener effort was still reported even when performance was equivalent (McCoy et al. 2005; Bologna et al. 2013); while in other cases, listeners reported more effortless listening with no significant change in speech recognition (Sarampalis et al.

2009; Ahlstrom et al. 2014). An independent and unbiased measurement of listening effort should be applied, in order to better characterise the intervention outcomes. For instance, Kuchinsky et al. (2014) showed improved word recognition scores as well as more rapid and bigger pupil response after a word training for older adults with hearing loss. Cochlear implantation has been proven to improve speech recognition abilities and life quality for post-lingually deaf listeners (Sladen et al. 2017). Specifically, training has been found to improve listeners speech recognition performance, and this study is the first to investigate the change in listening effort induced by the auditory training. CI listeners were found to have significantly smaller pupillary response, accompanied by a significant improvement in sentence recognition after training. Although this result might look inconsistent with the finding of Kuchinsky et al. (2014), it could be due to the different interpretation of pupillary response in the two studies caused by different experimental manipulations. In Kuchinsky et al. (2014), no adaptive procedure was applied, therefore, the pupillary response before and after the training reflects the absolute amount of attention and cognitive resources allocated for word recognition. After the training, the same task used before the training became easier, so listeners have extra cognitive 'space' to invest in more resources to enhance their speech performance. Therefore, the increase in the pupillary response includes both release from the speech understanding difficulty and listeners' active allocation of more cognitive resources. In comparison, this experiment measured pupillary responses at 50% of each listeners' performance before and after the training, making

sure that the measurements reflected the amount of effort required to reach the same level of performance. Any change in speech processing abilities due to the training will not affect the adaptive measurement of cognitive effort. Therefore, a decrease in pupillary response would suggest that after training, listeners need significantly less cognitive resources to cope with a similarly difficult speech task as before the training. Therefore, they experience less cognitive load in general for speech understanding. This also illustrates the importance of using experimental designs that take into consideration between- and within-individual variances, especially for HI listeners considering their high variability in speech and cognitive abilities. However, whether this decrease in pupillary response is due to the auditory training used here is unclear, since there is no significant interaction of the training group. The decrease in listening effort might just be the result of CI listeners getting used to the sentence materials and testing procedures.

This study suggests that CI listeners' sensitivity to the statistical features of speech cues might be diminished by their auditory and cognitive constraints. Nevertheless, exposure to multi-talker words materials, regardless of the distribution of speech cues, helps to improve CI listeners' sentence and word recognition, as well as decrease their listening effort.

Part 4

General Discussion

Chapter 6

General Discussion and Future Direction

This dissertation investigates the impact and plasticity of NH (with CI acoustic simulations) and CI listeners' acoustic cue weighting strategies. It was hypothesised that how listeners allocate perceptual attention to different speech cues is related to how accurately and effectively listeners can restore the phonemic structures from the acoustic inputs. Therefore, it would be beneficial to use auditory training to guide listeners' attention to the more reliable and informative cues for their specific language, in order to improve their speech recognition and ease listening effort. The first hypothesis was examined in Chapter 3, and the second hypothesis was examined in Chapters 4 and 5.

Chapter 2 firstly explored the impact of different degree of spectral degradation, distortion and background noise on the relative spectral and duration cue weighting in a /bɪt/ - /bit/ contrast for NH listeners with noise-vocoded and spectrally-shifted speech. This was to simulate the wide range of signal

degradation and distortion perceived by CI users. Listeners were found to decrease their perceptual weighting on the formant structure cue but retained the same weighting on the duration cue once spectral degradation and distortion were applied to the stimuli. This suggests that the CI simulation not only reduced the amount of spectral information available to listeners, but also changed their perceptual attention to the spectral cue. Also, the impact of immediate spectral degradation and distortion was not independent, since increasing spectral resolution by increasing the number of vocoding bands didn't change the cue weighting pattern for spectral shifting larger than 3mm. Therefore, both spectral degradation and distortion had a significant impact on listeners' acoustic cue weighting. Past studies typically didn't use spectral shifting in CI simulations, which might underestimate the degree of signal mismatch experienced by CI users hence couldn't characterise their cue weighting pattern.

Chapter 3 investigated whether listeners' acoustic cue weighting strategies were related to how they use cognitive resources for sentence recognition. It was shown that for both NH and CI listeners, there was a trend that more weighting of the formant structure cue than the duration cue was associated with more efficient use of cognitive resources for sentence recognition. Even for CI users, when their auditory sensitivity was taken into account, this relation bordered on significance. This seems to suggest that using the acoustic cue weighting strategy similar to a native language speaker, even when the speech signals are compromised, benefits the mapping from acoustic inputs to their phonemic representations. The weighting strategy

shared by NH listeners within the same language community is robust and developed over years. Therefore, this strategy is likely to be the most effective one for that particular language or dialect in processing acoustic cues for speech perception, by allocating more attention to the most informative and reliable cues. Post-lingually deaf CI listeners who keep this strategy after hearing loss and apply it to the speech signals perceived through their implants are essentially using the most useful strategy for that language, so they should be better in the speech recognition performance. To further support this hypothesis, firstly more CI listeners need to be recruited. Meanwhile, CI listeners from other languages/dialects with different types of cue weighting strategy should be tested to see whether this effect is replicable. According to this hypothesis, whatever relative cue weighting pattern a certain group of adult NH listeners uses, if post-lingually deaf CI users in the same language community used a similar strategy, they should be better speech performers.

There could also be an alternative explanation. The relation between acoustic cue weighting and speech performance might not be a causal one. In another word, the better CI performers might be more efficient in speech processing *not* because they use a certain cue weighting strategy, but rather, that these better performers have more intact internal phonological representations and phonological processing abilities for their native language, which were developed prior to their hearing loss. Hence, they have a perceptual cue weighting strategy similar to that of their NH counterparts. Previous studies showed that adults with hearing impairment had poor perfor-

mance on even visually presented rhyme-matching tasks, suggesting that their phonological representations had deteriorated (Lyxell et al. 1998; Classon et al. 2013). In a more recent study, Moberly et al. (2017) showed even after controlling for age and working memory, that CI users performed significantly worse in nonword repetition and lexical decision tasks, which both require listeners to recognise detailed phonological structure in the speech inputs. Therefore, CI listeners who have better phonological representations and abilities, possibly due to a shorter period of deafness and better peripheral and central auditory system integrity, are most likely to be good performers. This might explain the concurrent relation between listeners' cue weighting strategy and speech perception efficiency shown in this experiment.

Past studies and Chapter 3 reported no significant correlation between listeners' acoustic cue weighting strategy and their auditory sensitivity (Moberly et al. 2014, 2016). This suggests that some CI listeners don't use all the spectral differences in the perceived auditory signals for speech categorisation. Also, previous studies and Chapter 4 have shown that NH adult listeners are still sensitive to the statistical regularities in the visual and auditory inputs, although not at a similar level as infants (Wanrooij & Boersma 2013; Wanrooij et al. 2014). These present the possibility that CI listeners could be trained to allocate more perceptual attention on the spectral cue by manipulating the distribution of speech cues. However, results in Chapter 5 suggested that CI listeners were not sensitive to this manipulation of the training materials. This could be due to many reasons. Firstly, CI listen-

ers might not have access to this information. As mentioned above, their phonological processing ability has been compromised due to prolonged hearing loss and distorted acoustic inputs. This might include the ability to track occurrences of certain speech categories, observe the differences in acoustic cues and generate an updated representations of the speech categories. Secondly, due to the low resolution and saliency of the input signals, CI listeners' speech perception might be mainly driven by top-down processing. During speech communication, a listener needs to access to the acoustic signals, employ attention and intention, interpret the linguistic and pragmatic contexts, etc. However, this continuous cooperation between the top-down and bottom-up information processing is disturbed when the acoustic inputs are degraded and distorted. To compensate for that, listeners are shown to rely more on the contextual information and cognitive processes (Davis et al. 2005; Zekveld et al. 2006; Obleser et al. 2007). Therefore, changes in low-level signal properties, for instance, the statistical features of acoustic cues, might not have a great impact on CI listeners' speech perception.

The results in Chapter 5 also suggest that, after training, both groups of listeners tend to have decreased pupil responses. This suggests that listeners had less listening effort understanding sentences after the training, although the improvement was not likely due to the training implemented. However, this shows that it is possible for listeners to increase their sentence recognition performance and decrease their listening effort at the same time. Therefore, rehabilitations for CI users should have an updated

goal: training should not only improve CI listeners' speech recognition performance, but also decrease their listening effort. Both aspects are important in assessing the efficacy of a CI training program, since speech perception should both be more accurate and less costly in terms of cognitive resources to support other tasks of daily life.

In summary, this dissertation presents studies that are among the first to investigate the cognitive impact of individual variability and auditory training for CI users. Although the specific auditory training method in the current study didn't have an impact on CI listeners sentence recognition and listening effort, the line of research to find individual auditory or linguistic features that explains the great variability in listening effort remains important. Many studies have investigated the variability in CI users' accuracy of speech perception; therefore, how efficiently they understand speech should receive the same amount of attention. Only by drawing more attention to their listening effort and tailor auditory training to resolve the problem of high listening fatigue, can CI users benefit more from the use of their device and enjoy a higher quality of life.

References

Ahlstrom, J. B., Horwitz, A. R., & Dubno, J. R. (2014). Spatial separation benefit for unaided and aided listening. *Ear and Hearing*, *35*(1).

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear and Hearing*, *38*(1), e39–e48.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552.

Arai, T., & Greenberg, S. (1998). Speech intelligibility in the presence of cross-channel spectral asynchrony. In *Acoustics, speech and signal processing, 1998. proceedings of the 1998 IEEE international conference* (Vol. 2, pp. 933–936).

Assmann, P. F., & Katz, W. F. (2005). Synthesis fidelity and time-varying spectral change in vowels. *The Journal of the Acoustical Society of America*, *117*(2), 886–895.

Babel, M., & Munson, B. (2014). 19 producing socially meaningful linguistic variation. *The oxford handbook of language production*, 308.

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *J Exp Psychol Hum Percept Perform*, *6*(3), 536.

Baskent, D., & Shannon, R. V. (2003). Speech recognition under conditions of frequency-place compression and expansion. *The Journal of the*

Acoustical Society of America, *113*(4), 2064–2076.

Bates, D., Maechler, M., Bolker, B., et al. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version*, *1*(7).

Batterink, L. J., Reber, P. J., & Paller, K. A. (2015). Functional differences between statistical learning with and without explicit training. *Learning & Memory*, *22*(11), 544–556.

Bench, J., Kowal, Å., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology*, *13*(3), 108–112.

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, *29*(3), 191–211.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9/10), 341–345.

Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer (version 5.1. 05)[computer program]. retrieved may 1, 2009.*

Bologna, W. J., Chatterjee, M., & Dubno, J. R. (2013). Perceived listening effort for a tonal task with contralateral competing signals. *The Journal of the Acoustical Society of America*, *134*(4), EL352–EL358.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training japanese listeners to identify English /r/ and/l/: Long-term retention of learning in perception and production. *Attention, Perception, & Psychophysics*, *61*(5), 977–985.

Bradshaw, J. (1968). Pupil size and problem solving. *The Quarterly Journal*

of Experimental Psychology, *20*(2), 116–122.

Bristow, D., Haynes, J.-D., Sylvester, R., et al. (2005). Blinking suppresses the neural response to unchanging retinal stimulation. *Current Biology*, *15*(14), 1296–1300.

Budenz, C. L., Cosetti, M. K., Coelho, D. H., Birenbaum, B., Babb, J., Waltzman, S. B., & Roehm, P. C. (2011). The effects of cochlear implantation on speech perception in older adults. *Journal of the American Geriatrics Society*, *59*(3), 446–453.

Calandruccio, L., & Smiljanic, R. (2012). New sentence recognition materials developed using a basic non-native English lexicon. *Journal of Speech Language and Hearing Research*, *55*(5), 1342–1355.

Capretta, N. R., & Moberly, A. C. (2016a). Does quality of life depend on speech recognition performance for adult cochlear implant users? *The Laryngoscope*, *126*(3), 699–706.

Capretta, N. R., & Moberly, A. C. (2016b). Does quality of life depend on speech recognition performance for adult cochlear implant users? *The Laryngoscope*, *126*(3), 699–706.

Chatterjee, M., & Shannon, R. V. (1998). Forward masked excitation patterns in multielectrode electrical stimulation. *The Journal of the Acoustical Society of America*, *103*(5), 2565–2572.

Clark, G. M., Clark, J. C., & Furness, J. B. (2013). The evolving science of cochlear implants. *JAMA*, *310*(12), 1225–1226.

Classon, E., Rudner, M., & Rönnberg, J. (2013). Working memory compensates for hearing related phonological processing deficit. *Journal of*

Communication Disorders, *46*(1), 17–29.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008a). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008b). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Dahan, D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 704.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222.

De Filippo, C. L., & Scott, B. L. (1978). A method for training and evaluating the reception of ongoing speech. *The Journal of the Acoustical Society of America*, *63*(4), 1186–1192.

Dillon, M. T., Buss, E., Adunka, M. C., King, E. R., Pillsbury, H. C., Adunka, O. F., & Buchman, C. A. (2013). Long-term speech perception in elderly cochlear implant users. *JAMA Otolaryngology–Head & Neck Surgery*, *139*(3), 279–283.

Di Nardo, W., Scorpecci, A., Giannantonio, S., Cianfrone, F., Parrilla, C., & Paludetti, G. (2010). Cochlear implant patients' speech understanding in background noise: effect of mismatch between electrode assigned

frequencies and perceived pitch. *The Journal of Laryngology & Otology*, *124*(08), 828–834.

Donaldson, G. S., Rogers, C. L., Cardenas, E. S., Russell, B. A., & Hanna, N. H. (2013). Vowel identification by cochlear implant users: Contributions of static and dynamic spectral cues. *The Journal of the Acoustical Society of America*, *134*(4), 3021–3028.

Donaldson, G. S., Rogers, C. L., Johnson, L. B., & Oh, S. H. (2015). Vowel identification by cochlear implant users: Contributions of duration cues and dynamic spectral cues. *The Journal of the Acoustical Society of America*, *138*(1), 65–73.

Dorman, M. F., Dankowski, K., McCandless, G., Parkin, J. L., & Smith, L. (1991). Vowel and consonant recognition with the aid of a multichannel cochlear implant. *The Quarterly Journal of Experimental Psychology*, *43*(3), 585–601.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding. *The Journal of the Acoustical Society of America*, *102*(5), 2993–2996.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *The Journal of the Acoustical Society of America*, *97*(1), 585–592.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, *95*(5), 2670–2680.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2

perceptual cue weighting for dutch vowels: The case of dutch, german, and spanish listeners. *Journal of Phonetics*, *37*(4), 452–465.

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America*, *130*(4), EL206–EL212.

Escudero, P., & Williams, D. (2014). Distributional learning has immediate and long-lasting effects. *Cognition*, *133*(2), 408–413.

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *The Journal of the Acoustical Society of America*, *115*(1), 352–361.

Faulkner, A., Rosen, S., & Green, T. (2012). Comparing live to recorded speech in training the perception of spectrally shifted noise-vocoded speech. *The Journal of the Acoustical Society of America*, *132*(4), EL336–EL342.

Faulkner, A., Rosen, S., & Stanton, D. (2003). Simulations of tonotopically mapped speech processors for cochlear implant electrodes varying in insertion depth. *The Journal of the Acoustical Society of America*, *113*(2), 1073–1080.

Finley, C. C., & Skinner, M. W. (2008). Role of electrode placement as a contributor to variability in cochlear implant outcomes. *Otology & neurotology: official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology*, *29*(7), 920.

Firszt, J. B., Holden, L. K., Skinner, M. W., Tobey, E. A., Peterson, A., Gaggl, W., … Wackym, P. A. (2004). Recognition of speech presented at soft to loud levels by adult cochlear implant recipients of three cochlear implant systems. *Ear and Hearing*, *25*(4), 375–387.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826.

Friesen, L. M., Shannon, R. V., Baskent, D., et al. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America*, *110*(2), 1150–1163.

Fu, Q.-J., Galvin, J., Wang, X., & Nogaki, G. (2004). Effects of auditory training on adult cochlear implant patients: a preliminary report. *Cochlear Implants International*, *5*(S1), 84–90.

Fu, Q.-J., Galvin, J., Wang, X., & Nogaki, G. (2005). Moderate auditory training can improve speech performance of adult cochlear implant patients. *Acoustics Research Letters Online*, *6*(3), 106–111.

Fu, Q.-J., & Galvin, J. J. (2008). Maximizing cochlear implant patients' performance with advanced speech training procedures. *Hearing Research*, *242*(1), 198–208.

Fu, Q.-J., & Galvin III, J. J. (2001). Recognition of spectrally asynchronous speech by normal-hearing listeners and nucleus-22 cochlear implant users. *The Journal of the Acoustical Society of America*, *109*(3), 1166–1172.

Fu, Q.-J., Nogaki, G., & Galvin III, J. 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.-J., & Shannon, R. (2000). Effects of stimulation rate on phoneme recognition in cochlear implant users. *The Journal of the Acoustical Society of America*, *107*, 589–597.

Fu, Q.-J., & Shannon, R. V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing. *The Journal of the Acoustical Society of America*, *105*(3), 1889–1900.

Giezen, M. R., Escudero, P., & Baker, A. (2010). Use of acoustic cues by children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *53*(6), 1440–1457.

Gifford, R. H., Shallop, J. K., & Peterson, A. M. (2008). Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. *Audiology and Neurotology*, *13*(3), 193–205.

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

Gopal, H. (1990). Effects of speaking rate on the behavior of tense and lax vowel durations. *Journal of Phonetics*, *18*(4), 497–518.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species - 29 years later. *The Journal of the Acoustical Society of America*, *87*(6), 2592–2605.

Gulian, M., Escudero, P., Boersma, P., et al. (2007). Supervision hampers

distributional learning of vowel contrasts. *Proceedings of the 16th International Congress of Phonetic Sciences*.

Hakerem, G., & Sutton, S. (1966). Pupillary response at visual threshold. *Nature*, *212*(5061), 485–486.

Harris, M. S., Capretta, N. R., Henning, S. C., Feeney, L., Pitt, M. A., & Moberly, A. C. (2016). Postoperative rehabilitation strategies used by adults with cochlear implants: a pilot study. *Laryngoscope Investigative Otolaryngology*, *1*(3), 42–48.

Hazan, V., & Rosen, S. (1991). Individual variability in the perception of cues to place contrasts in initial stops. *Attention, Perception, & Psychophysics*, *49*(2), 187–200.

Hedrick, M. S., & Carney, A. E. (1997). Effect of relative amplitude and formant transitions on perception of place of articulation by adult listeners with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1445–1457.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192.

Hillenbrand, J., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013–3022.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099–3111.

Hillenbrand, J., & Nearey, T. M. (1999). Identification of resynthesized /hvd/ utterances: Effects of formant contour. *The Journal of the Acoustical Society of America*, *105*(6), 3509–3523.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*(4108), 177–180.

Hirschfelder, A., Gräbel, S., & Olze, H. (2008). The impact of cochlear implantation on quality of life: the role of audiologic performance and variables. *Otolaryngology—Head and Neck Surgery*, *138*(3), 357–362.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisitions). *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071.

Holt, L. L., Lotto, A. J., & Diehl, R. L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *The Journal of the Acoustical Society of America*, *116*(3), 1763–1773.

Hoonhorst, I., Colin, C., Markessis, E., Radeau, M., Deltenre, P., & Serniclaes, W. (2009). French native speakers in the making: From language-general to language-specific voicing boundaries. *Journal of Experimental Child Psychology*, *104*(4), 353–366.

House, A. S. (1961). On vowel duration in English. *The Journal of the Acoustical Society of America*, *33*(9), 1174–1178.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception & Performance*, *37*(6), 1939.

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009.

Ingvalson, E. M., Holt, L. L., & McClelland, J. L. (2012). Can native Japanese listeners learn to differentiate /r–l/ on the basis of F3 onset frequency? *Bilingualism: Language and Cognition*, *15*(02), 255–274.

Ingvalson, E. M., Lee, B., Fiebig, P., & Wong, P. C. (2013). The effects of short-term computerized speech-in-noise training on postlingually deafened adult cochlear implant recipients. *Journal of Speech, Language, and Hearing Research*, *56*(1), 81–88.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57.

Iverson, P., Smith, C. A., & Evans, B. G. (2006). Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration. *The Journal of the Acoustical Society of America*, *120*(6), 3998–4006.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in "vowelless" syllables. *Attention, Perception, & Psychophysics*, *34*(5), 441–450.

Jolliffe, I. T. (1986). Principal component analysis and factor analysis. In *Principal component analysis* (pp. 115–128). Springer.

Jusczyk, P. W., Pisoni, D. B., Walley, A., & Murray, J. (1980). Discrimination of relative onset time of two-component tones by infants. *The Journal of the Acoustical Society of America*, *67*(1), 262–270.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Prentice-Hall Englewood Cliffs, NJ.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, Fo, and aperiodicity estimation. In *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference* (pp. 3933–3936).

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, 286–319.

Kirk, K. I., Tye-Murray, N., & Hurtig, R. R. (1992). The use of static and dynamic vowel cues by multichannel cochlear implant users. *The Journal of the Acoustical Society of America*, *91*(6), 3487–3498.

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

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the*

2008 symposium on eye tracking research & applications (pp. 69–72).

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., et al. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, *323*, 81–90.

Koelewijn, T., Zekveld, A. A., Festen, J. M., et al. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*(2), 291–300.

Kohler, K. J. (1982). Fo in the production of lenis and fortis plosives. *Phonetica*, *39*(4-5), 199–218.

Kohler, K. J. (1984). Phonetic explanation in phonology: the feature fortis/lenis. *Phonetica*, *41*(3), 150–174.

Kong, Y.-Y., Winn, M. B., Poellmann, K., et al. (2016). Discriminability and perceptual saliency of temporal and spectral cues for final fricative consonant voicing in simulated cochlear-implant and bimodal hearing. *Trends in Hearing*, *20*, 2331216516652145.

Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2014). Speech-perception training for older adults with hearing loss impacts word recognition and effort. *Psychophysiology*, *51*(10), 1046–1057.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., et al. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*(1), 23–34.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do

not. *Attention, Perception, & Psychophysics*, *50*(2), 93–107.

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

Kwon, H.-B. (2010). Gender difference in speech intelligibility using speech intelligibility tests and acoustic analyses. *The Journal of Advanced Prosthodontics*, *2*(3), 71–76.

Lau, J. H., Clark, A., & Lappin, S. (2016). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*, *62*, 91–123.

Lazard, D., Lee, H., Gaebler, M., Kell, C., Truy, E., & Giraud, A.-L. (2010). Phonological processing in post-lingual deafness and cochlear implant outcome. *Neuroimage*, *49*(4), 3443–3451.

Lehet, M., & Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cognitive science*, *41*(S4), 885–912.

Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. *The Journal of the Acoustical Society of America*, *33*(3), 268–277.

Leung, K. K., Jongman, A., Wang, Y., & Sereno, J. A. (2016). Acoustic characteristics of clearly spoken english tense and lax vowels. *The Journal of the Acoustical Society of America*, *140*(1), 45–58.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America*, *49*(2B), 467–477.

Li, T., & Fu, Q.-J. (2007). Perceptual adaptation to spectrally shifted vowels:

training with nonlexical labels. *Journal of the Association for Research in Otolaryngology*, *8*(1), 32–41.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., et al. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431.

Lisker, L. (1978). Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research*, *54*, 127–132.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783.

Loizou, P. C., Dorman, M., Poroy, O., & Spahr, T. (2000). Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution. *The Journal of the Acoustical Society of America*, *108*(5), 2377–2387.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of japanese acquisition of /r/ and /l/. *From sound to sense*, *50*(2004), C381–C386.

Lowenstein, J. H., & Nittrouer, S. (2015). All cues are not created equal: The case for facilitating the acquisition of typical weighting strategies in children with hearing loss. *Journal of Speech, Language, and Hearing Research*, *58*(2), 466–480.

Lyxell, B., Andersson, J., Andersson, U., Arlinger, S., Bredberg, G., & Harder,

H. (1998). Phonological representation and speech understanding with cochlear implants in deafened adults. *Scandinavian Journal of Psychology*, *39*(3), 175–179.

Manrique, M., Espinosa, J., Huarte, A., Molina, M., Garcia-Tapia, R., & Artieda, J. (1997). Cochlear implants in post-lingual persons: results during the first five years of the clinical course. *Acta otorrinolaringologica espanola*, *49*(1), 19–24.

MATLAB. (2013). *version 8.1.0 (r2013a)*. Natick, Massachusetts: The MathWorks Inc.

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *The Quarterly Journal of Experimental Psychology Section A*, *58*(1), 22–33.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper '. *International Journal of Audiology*, *53*(7), 433–440.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Re-*

view, *118*(2), 219.

Miller, S. E., Zhang, Y., & Nelson, P. B. (2016). Efficacy of multiple-talker phonetic identification training in postlingually deafened cochlear implant listeners. *Journal of Speech, Language, and Hearing Research*, *59*(1), 90–98.

Mitterer, H., & Mattys, S. L. (2017). How does cognitive load influence speech perception? An encoding hypothesis. *Attention, Perception, & Psychophysics*, *79*(1), 344–351.

Moberly, A. C., Harris, M. S., Boyce, L., & Nittrouer, S. (2017). Speech recognition in adults with cochlear implants: The effects of working memory, phonological sensitivity, and aging. *Journal of Speech, Language, and Hearing Research*, *60*(4), 1046–1061.

Moberly, A. C., Lowenstein, J. H., & Nittrouer, S. (2016). Word recognition variability with cochlear implants: Perceptual attention versus auditory sensitivity? *Ear and Hearing*, *37*(1), 14–26.

Moberly, A. C., Lowenstein, J. H., Tarr, E., Caldwell-Tarr, A., Welling, D. B., Shahin, A. J., & Nittrouer, S. (2014). Do adults with cochlear implants rely on different acoustic cues for phoneme perception than adults with normal hearing? *Journal of Speech, Language, and Hearing Research*, *57*(2), 566–582.

Moore, B. C. (2004). Dead regions in the cochlea: conceptual foundations, diagnosis, and clinical applications. *Ear and Hearing*, *25*(2), 98–116.

Moore, B. C., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the*

Acoustical Society of America, *74*(3), 750–753.

Mosnier, I., Bebear, J.-P., Marx, M., Fraysse, B., Truy, E., Lina-Granade, G., … others (2015). Improvement of cognitive function after cochlear implantation in elderly patients. *JAMA Otolaryngology–Head & Neck Surgery*, *141*(5), 442–450.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453–467.

Nadol, J. B., & Eddington, D. K. (2006). Histopathology of the inner ear relevant to cochlear implantation. In *Cochlear and brainstem implants* (Vol. 64, pp. 31–49). Karger Publishers.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America*, *80*(5), 1297–1308.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196.

Nittrouer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *The Journal of the Acoustical Society of America*, *112*(2), 711–719.

Nittrouer, S., Caldwell-Tarr, A., Moberly, A. C., et al. (2014). Perceptual weighting strategies of children with cochlear implants and normal hearing. *Journal of Communication Disorders*, *52*, 111–133.

Nittrouer, S., Lowenstein, J. H., & Tarr, E. (2013). Amplitude rise time

does not cue the /bɑ/–/wɑ/ contrast for adults or children. *Journal of Speech, Language, and Hearing Research*, *56*(2), 427–440.

Nittrouer, S., & Studdert-Kennedy, M. (1986). The stop–glide distinction: Acoustic analysis and perceptual effect of variation in syllable amplitude envelope for initial /b/ and /w/. *The Journal of the Acoustical Society of America*, *80*(4), 1026–1029.

Oba, S. I., Fu, Q.-J., & Galvin III, J. J. (2011). Digit training in noise can improve cochlear implant users' speech understanding in noise. *Ear and Hearing*, *32*(5), 573.

Obleser, J., Wise, R. J., Dresner, M. A., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, *27*(9), 2283–2289.

Ong, J. H., Burnham, D., & Escudero, P. (2015). Distributional learning of lexical tones: a comparison of attended vs. unattended listening. *PloS one*, *10*(7), e0133446.

Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science*, *205*(4409), 924–927.

Peavler, W. S. (1974). Individual differences in pupil size and performance. In *Pupillary dynamics and behavior* (pp. 159–175). Springer.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., … others (2016). Hearing impairment and cog-

nitive energy: The framework for understanding effortful listening (fuel). *Ear and Hearing*, *37*, 5S–27S.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. *The Journal of the Acoustical Society of America*, *61*(5), 1352–1361.

Plomp, R., & Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, *18*(1), 43–52.

Port, R. F. (1981). Linguistic timing factors in combination. *The Journal of the Acoustical Society of America*, *69*(1), 262–274.

Prendergast, G., & Green, G. G. (2012). Cross-channel amplitude sweeps are crucial to speech intelligibility. *Brain and Language*, *120*(3), 406–411.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 539.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychology Bulletin*, *92*(1), 81.

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304.

Roberts, B., Summers, R. J., & Bailey, P. J. (2010). The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20101554.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and

linguistic aspects. *Philosophical Transactions of the Royal Society of London 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. *The Journal of the Acoustical Society of America*, *106*(6), 3629–3636.

Rudner, M. (2016). Cognitive spare capacity as an index of listening effort. *Ear and Hearing*, *37*, 69S–76S.

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

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception & Psychophysics*, *78*(1), 355–367.

Schumann, A., Serman, M., Gefeller, O., & Hoppe, U. (2015). Computer-based auditory phoneme discrimination training improves speech recognition in noise in experienced adult cochlear implant listeners. *International Journal of Audiology*, *54*(3), 190–198.

Shafiro, V., Sheft, S., Gygi, B., & Ho, K. T. N. (2012). The influence of environmental sound training on the perception of spectrally degraded speech and environmental sounds. *Trends in Amplification*, 1084713812454225.

Shannon, R. V., Galvin III, J. J., & Baskent, D. (2002). Holes in hearing.

JARO-Journal of the Association for Research in Otolaryngology, *3*(2), 185–199.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303.

Shen, Y., Dai, W., & Richards, V. M. (2015). A MATLAB toolbox for the efficient estimation of the psychometric function using the updated maximum-likelihood adaptive procedure. *Behavioral Research Methods*, *47*(1), 13–26.

Shen, Y., & Richards, V. M. (2012). A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention. *The Journal of the Acoustical Society of America*, *132*(2), 957–967.

Simos, P. G., & Molfese, D. L. (1997). Electrophysiological responses from a temporal order continuum in the newborn infant. *Neuropsychologia*, *35*(1), 89–98.

Sinex, D. G., & McDonald, L. P. (1989). Synchronized discharge rate representation of voice-onset time in the chinchilla auditory nerve. *The Journal of the Acoustical Society of America*, *85*(5), 1995–2004.

Sinex, D. G., McDonald, L. P., & Mott, J. B. (1991). Neural correlates of nonmonotonic temporal acuity for voice onset time. *The Journal of the Acoustical Society of America*, *90*(5), 2441–2449.

Skinner, M. W., Ketten, D. R., Holden, L. K., Harding, G. W., Smith, P. G., Gates, G. A., … Blocker, B. (2002). Ct-derived estimation of cochlear

morphology and electrode array position in relation to word recognition in nucleus-22 recipients. *Journal of the Association for Research in Otolaryngology*, *3*(3), 332–350.

Sladen, D. P., Peterson, A., Schmitt, M., Olund, A., Teece, K., Dowling, B., … others (2017). Health-related quality of life outcomes following adult cochlear implantation: A prospective cohort study. *Cochlear Implants International*, *18*(3), 130–135.

Sladen, D. P., & Zappler, A. (2015). Older and younger adult cochlear implant users: speech recognition in quiet and noise, quality of life, and music perception. *American Journal of Audiology*, *24*(1), 31–39.

Souza, P. E., Wright, R. A., Blackburn, M. C., Tatman, R., & Gallun, F. J. (2015). Individual sensitivity to spectral and temporal cues in listeners with hearing impairment. *Journal of Speech, Language, and Hearing Research*, *58*(2), 520-534.

Stacey, P. C., Raine, C. H., O'Donoghue, G. M., Tapper, L., Twomey, T., & Summerfield, A. Q. (2010). Effectiveness of computer-based auditory training for adult users of cochlear implants. *International Journal of Audiology*, *49*(5), 347–356.

Stacey, P. C., & Summerfield, A. Q. (2008). Comparison of word-, sentence-, and phoneme-based training strategies in improving the perception of spectrally distorted speech. *Journal of Speech, Language, and Hearing Research*, *51*(2), 526–538.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of phonetics*, *17*(1), 3–45.

Tillman, G., Benders, T., Brown, S. D., & van Ravenzwaaij, D. (2017). An evidence accumulation model of acoustic cue weighting in vowel perception. *Journal of Phonetics*, *61*, 1–12.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434–464.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301.

Valbret, H., Moulines, E., & Tubach, J.-P. (1992). Voice transformation using psola technique. *Speech Communication*, *11*(2), 175–187.

Wanrooij, K., & Boersma, P. (2013). Distributional training of speech sounds can be done with continuous distributions. *The Journal of the Acoustical Society of America*, *133*(5), EL398–EL404.

Wanrooij, K., Boersma, P., & Benders, T. (2015). Observed effects of "distributional learning" may not relate to the number of peaks. A test of "dispersion" as a confounding factor. *Frontiers in Psychology*, *6*.

Wanrooij, K., Boersma, P., & van Zuijen, T. L. (2014). Distributional vowel training is less effective for adults than for infants. A study using the mismatch response. *PloS one*, *9*(10), e109806.

Wanrooij, K., Escudero, P., & Raijmakers, M. E. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. *Journal of Phonetics*, *41*(5), 307–319.

Wardrip-Fruin, C. (1985). The effect of signal degradation on the status of cues to voicing in utterance-final stop consonants. *The Journal of the Acoustical Society of America*, *77*(5), 1907–1912.

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in australian english vowels. *The Journal of the Acoustical Society of America*, *106*(1), 458–468.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2012). The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearinga). *The Journal of the Acoustical Society of America*, *131*(2), 1465–1479.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2013b). Roles of voice onset time and fo in stop consonant voicing perception: effects of masking noise and low-pass filtering. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1097–1107.

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, *36*(4), e153–e165.

Winn, M. B., & Litovsky, R. Y. (2015). Using speech sounds to test functional spectral resolution in listeners with cochlear implants. *The Journal of the Acoustical Society of America*, *137*(3), 1430–1442.

Winn, M. B., Rhone, A. E., Chatterjee, M., & Idsardi, W. J. (2013a). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in Psychology*, *4*.

Wu, J.-L., Yang, H.-M., Lin, Y.-H., & Fu, Q.-J. (2007). Effects of computer-assisted speech training on mandarin-speaking hearing-impaired children. *Audiology and Neurotology*, *12*(5), 307–312.

Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2013). Task difficulty differentially affects two measures of processing load: The pupil response during sentence processing and delayed cued recall of the sentences. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1156–1165.

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *NeuroImage*, *32*(4), 1826–1836.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*(3), 277–284.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510.

Zeng, F.-G., & Galvin III, J. J. (1999). Amplitude mapping and phoneme recognition in cochlear implant listeners. *Ear and Hearing*, *20*(1), 60–74.

Zénon, A., Sidibé, M., & Olivier, E. (2014). Pupil size variations correlate with physical effort perception. *Frontiers in Behavioral Neuroscience*,

8, 91–123.

Zhang, J., Xie, L., Li, Y., et al. (2014). How noise and language proficiency influence speech recognition by individual non-native listeners. *PloS one*, *9*(11), e113386.

Zhou, N., Xu, L., & Lee, C.-Y. (2010). The effects of frequency-place shift on consonant confusion in cochlear implant simulations. *The Journal of the Acoustical Society of America*, *128*(1), 401–409.