Working Title: Cognitive mechanisms underpinning successful perception of different speech distortions.

Authors:

Dan Kennedy-Higgins[a,b],

Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London, United Kingdom, WC1N 1PF

Joseph T. Devlin,

Department of Experimental Psychology, University College London, 26 Bedford Way, London, United Kingdom, WC1H 0AP

Patti Adank

Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London, United Kingdom, WC1N 1PF

[a] Electronic mail: daniel.kennedy-higgins@kcl.ac.uk.

[b] Current address: *Department of Psychology, King's College London, Guy's Campus, London, United Kingdom, SE1 1UL*

Abstract

Few studies thus far have investigated whether perception of distorted speech is consistent across different types of distortion. We investigated whether participants show a consistent perceptual profile across three speech distortions: time-compressed, noise-vocoded and speech in noise. **Additionally, we investigated whether/how individual differences in performance on a battery of audiological and cognitive tasks links to perception. Eighty-eight** participants completed a speeded sentence-verification task with increases in accuracy and reductions in response times used to indicate performance. **Audiological and cognitive task measures include pure tone audiometry, speech recognition threshold, working memory, vocabulary knowledge, attention switching, and pattern analysis.** Despite previous studies suggesting that temporal and spectral/environmental perception require different lexical or phonological mechanisms, we show significant positive correlations in accuracy and response time performance across all distortions. Results of a principal component analysis and multiple linear regressions suggest that a **component based on vocabulary knowledge and working memory** predicted performance in the speech in quiet, time-compressed and speech in noise conditions. These results suggest that listeners employ a similar cognitive strategy to perceive different temporal and spectral/environmental speech distortions and that this mechanism is supported by vocabulary knowledge and working memory.

Keywords: Time-compressed speech; noise-vocoded speech; speech in noise; individual differences.

Cognitive mechanisms underpinning successful perception of different speech distortions.

Introduction

The ability to perceptually adapt to the countless distortions that are inherent in all speech is essential for successful speech comprehension. Goldstone (1998) refers to perceptual learning as *"relatively long-lasting changes to an organism's perceptual system that improves its ability to respond to its environment and are caused by its environment"* (pg.585). With respect to auditory perceptual learning, this adaptation would represent a change from a situation in which the listener is unable to perceive difficult speech and/or comprehend what is being perceived to a situation in which they can, with effort, recognise what is being said after a period of exposure (Watson, 1980). Such perceptual adaptation occurs after just a few minutes' worth of exposure to time-compressed speech (Fairbanks & Jr., 1957; Mehler et al., 1993; Voor & Miller, 1965), noise-vocoded speech (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Loizou, Dorman, & Tu, 1999; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) and speech embedded in noise (Cainer, James, & Rajan, 2008; Song, Skoe, Banai, & Kraus, 2012). This adaptation is thought to occur as listeners shift their attention from task-irrelevant to task-relevant cues (Adank & Devlin, 2010; Golomb, Peelle, & Wingfield, 2007). In the context of a temporal distortion such as time-compressed speech, changes in attention are believed to occur at phonological levels of processing (Adank & Devlin, 2010; Dupoux & Green, 1997; Golomb et al., 2007; Pallier, Sebastian-Gallés, Dupoux, Christophe, & Mehler, 1998; Sebastián-Gallés, Dupoux, Costa, & Mehler, 2000). The importance of the phonological level of processing for temporal distortions originated from Sebastián-Gallés et al. (2000), who found that adaptation to time-compressed speech is independent from speech comprehension. Participants adapt to an equivalent level after 'training' on time-compressed sentences presented in a language that they cannot speak compared to a language they can comprehend. However, more adaptation occurs if the training language shares an isochrony with the test language. For example, training on a syllable-timed language such as Italian will benefit participants only if the test sentences are also in syllable-timed languages, such as Spanish. Such a benefit will not occur for stress-timed languages such as English. This phenomenon suggests that changes may be occurring at phonological as opposed to lower acoustic levels of processing, as simple exposure to time-compression does not result in equivalent benefits across training languages. Furthermore, as comprehension does not appear to be necessary for adaptation, changes in attention at the phonological level appear most important for temporal distortions.

In comparison for spectral manipulations such as noise-vocoded speech or environmental distortions such as speech in noise, these changes in attention are believed to occur at lexical/semantic levels (Bradlow & Alexander, 2007; Burk, Humes, Amos, & Strauser, 2006; Cainer et al., 2008; Davis et al., 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Hervais-Adelman, Davis, Johnsrude, Taylor, & Carlyon, 2011; Loizou et al., 1999; Mayo, Florentine, & Buus, 1997). The importance of lexical/semantic level changes was demonstrated by Davis et al. (2005). Two groups of participants were trained, via passive listening, on 20 sentences containing words or 20 sentences containing non-word noise-vocoded sentences and then were tested on 20 noise-vocoded English words. Overall, the group trained with words performed significantly better than the group trained on non-words. Additionally, the group trained on non-words performed at a level that was equivalent to subjects that were completely naïve, i.e., subjects that had no prior history of exposure to vocoded speech. This result suggests that adaptation to noise-vocoded speech is dependent on either lexical, semantic and/or syntactic information with phonological information being less important, i.e., the exact opposite of time-compressed speech. For a full review see (Kennedy-Higgins, 2019; Mattys, Davis, Bradlow, & Scott, 2012; Samuel & Kraljic, 2009)

A. Perception of different speech distortions

Most studies thus far have investigated how listeners perceive a single distortion (Bradlow & Bent, 2008; Cainer et al., 2008; Clarke & Garrett, 2004; Davis et al., 2005; Hervais-Adelman et al., 2008; Mehler et al., 1993; Pallier et al., 1998; Zaballos, Plasencia, Gonzalez, de Miguel, & Macias, 2016). Where adaptation to multiple forms of distortion has been investigated, the study either used a between-group design (Davis & Johnsrude, 2003) or a within-group design without reporting any transfer of learning effects (Peelle & Wingfield, 2005). Yet, a between-group design does not provide any insight into whether individuals are capable of perceptually learning equally to multiple distortions or are more skilled at perceiving certain distortions more than others whilst a within-group design without reporting transfer of learning (i.e. order) effects assumes that the adaptation process for the different types of distortion is independent of each other. Thus far only one study has investigated whether the order of presentation influences the degree of subsequent perception to another speech distortion. Adank and Janse (2009) investigated the degree to which exposure and adaptation to one form of compressed speech transferred to the learning of a secondary form of compression. Participants were either presented with a block of artificially time-compressed sentences followed by naturally fast spoken sentences or vice versa. Adapting to the easier, artificially compressed sentences before the more difficult, naturally fast sentences gave participants an advantage over those who adapted to the natural speech before the artificially compressed sentences. This result indicates that learning transferred from the artificial distortion to the natural distortion, but not vice versa suggesting that adaptation to different types of distortion is not independent of each other. Adank and Janse (2009) argue, with reference to Reverse Hierarchy Theory (Ahissar, Nahum, Nelken, & Hochstein, 2009), that this difference in transfer of learning is due to the fact that the artificially time-compressed sentences posed less of a challenge to the perceptual system and therefore participants were able to process this stimuli at a higher level. For artificially time-compressed sentences, adjustments to the timings of expected word boundaries need to occur (cognitive level changes), whereas for **naturally fast** speech, participants need to adapt to the temporal compression as well as additional spectral variability (cognitive and perceptual changes). Consequently, when subsequently faced with the, more difficult, naturally fast speech sentences, participants were better able to process the higher-level features (e.g. temporal compression) and focus on lower-level distortion-specific cues (e.g. spectral variability), as the cognitive temporal adjustments have already been made, but the spectral variability has not been encountered in the previous artificial compression condition. This relationship fits with RHT's prediction that transfer of learning occurs when an easy condition is followed by a more difficult condition. The lack of transfer from naturally fast to artificially fast sentences is believed to be due to the need to immediately focus attention to lower-level properties for the naturally fast stimuli, resulting in learning of stimulus-specific information that does not transfer easily to alternative stimuli.

Only three other studies have specifically investigated whether participants' perceptual profile is consistent across distortions. Bent, Baese-Berk, Borrie, and McKee (2016) investigated recognition of words in phrases across a nonnative accent, a regional dialect and ataxic dysarthric speech within the same group of participants. Results show a significant correlation between performances in the nonnative accented speech condition and both the regional dialect and dysarthric speech conditions, suggesting that individuals who were able to successfully perceive nonnative speech were also more successful in the other conditions. However, no correlation was found between performance in a regional dialect and dysarthric speech conditions. The authors conclude that these results suggest that listeners are not "globally skilled" at perceiving distorted speech. Instead, different individuals can map acoustic-phonetic features found only in certain types of distortions onto words in their mental lexicons. However, in a follow-up study, Borrie, Baese-Berk, Engen, and Bent (2017) investigated the overlap in ability to report words spoken by an individual with dysarthria or words presented in noise. The authors found a significant positive correlation between performances in the two conditions and concluded that similar cognitive-perceptual processes aid comprehension in both conditions, i.e., it appears that participants do possess a global skill that allows them to adapt to a relatively equal level when the speech signal is distorted in an array of forms.

Of particular relevance to the current experiment, McLaughlin et al. (2018) investigated if individuals possess a global skill to perceive multiple different distortions. McLaughlin et al. (2018) used four different speech conditions, (1) a native speaker masked in speech shaped noise (energetic masking), (2) a native speaker masked by a single-talker masker (informational masking), (3) a non-native speaker in quiet and (4) a non-native speaker masked by speech-shaped noise. Similarly, to Bent et al. (2016), McLaughlin et al. (2018) found that performance in one condition did not always correlate with performance in all other conditions, but instead performance appeared to correlate for conditions with shared characteristics. For instance, conditions with energetic masking correlated with each other and conditions with a non-native speaker correlated whilst the informational and energetic masker conditions did not correlate. In addition to the previous research however, McLaughlin et al. (2018) also investigated the underlying cognitive mechanism supporting perception in each speech condition and found that greater receptive vocabulary performance was linked to better performance in all conditions. Furthermore, their measure of working memory positively predicted performance for the non-native accented speech. McLaughlin et al. (2018) suggest that vocabulary knowledge may act as a global predictor of individual differences in the perception of any form of difficult speech, while other measures e.g. working memory may be engaged only in certain listening environments.

Thus, only Adank and Janse (2009), Bent et al. (2016) Borrie et al. (2017) and McLaughlin et al. (2018) have so far investigated how individuals adapt to different distortions. The aim of this study was to further elucidate individual differences in the successful perception of different speech distortions and to systematically evaluate links between sensory/cognitive abilities and perception of temporal (time-compressed); spectral (noise-vocoded speech) and environmental (speech in noise) distortions. Based on previous research, where participants have been shown to be capable of adapting fully to time-compressed speech in up to 20 sentences (Adank & Devlin, 2010; Dupoux & Green, 1997; Sebastián-Gallés et al., 2000), while adaptation to noise-vocoded and speech in noise can occur within 30 sentences but can take many hours' worth of training before full adaptation occurs (Cainer et al., 2008; Davis et al., 2005; Zaballos et al., 2016), participants in the current experiment were predicted to adapt rapidly to time-compressed speech and slower and less extensively to the noise-vocoded and speech in noise conditions. However, it is unclear whether performance in one condition will equate to a relatively similar performance in all conditions. Research by Borrie et al. (2017) suggests that participants may possess (or lack) a global skill to perceive speech in any difficult listening environment. Moreover, research by Bent et al. (2016) and McLaughlin et al. (2018) suggest that participants may be better at perceiving speech in some – but not all - listening environments. If participants are better at perceiving speech in specific conditions, performance in the noise-vocoded and speech in noise conditions may correlate as perception of these conditions depends on similar changes at the lexical/semantic level. In contrast, a weaker correlation between noise-vocoded/speech in noise and time-compressed speech perception performance would be expected as this condition is dependent to a greater extent on phonological level changes.

Alternatively, as Borrie et al. (2017) suggest, **participants may possess (or indeed lack) a ‘global skill’ for perceiving speech that deviates from the norm** and thus performance across all three manipulations will be correlated.

B. Individual differences in perceptual adaptation

A key aim of this research is to systematically evaluate links between a battery of audiological and cognitive tests and successful perception of three distortions. The ability to perceive distorted speech has been related to a range of cognitive factors, yet no comprehensive model currently exists that explains which factors are most important and how these factors interact with the type of adverse condition. It is not known, for example, whether different distortions depend on a common cognitive mechanism or whether different distortions require different mechanisms to underpin adaptation. Thus far, associations between four audiological/cognitive abilities and perception of distorted speech have been investigated most: individual hearing thresholds; working memory; selective attention/inhibition and vocabulary knowledge. The impact of individual differences in hearing ability has predominantly been investigated in older populations where difficulty perceiving speech, especially in the presence of background noise, is a common trait. Although overall hearing thresholds have been associated with poorer overall performance on distorted speech tasks (Adank & Janse, 2010; Akeroyd, 2008; Janse & Adank, 2012), research that adjusts for differences in auditory sensitivities suggests that it is not just the decline of the auditory periphery causing the speech in noise deficit; effective listening also relies upon general cognitive processes (Bilodeau-Mercure, Lortie, Sato, Guitton, & Tremblay, 2015; Golomb et al., 2007; Moore, Peters, & Stone, 1999; Tun, 1998; Tun & Wingfield, 1999; Wong et al., 2009).

The Ease of Language Understanding (ELU) model (Rönnberg et al., 2013; Rönnberg, Rudner, Foo, & Lunner, 2008) emphasises the role of working memory capacity in suboptimal conditions where the incoming perceived signal is distorted and does not match any internal phonological representations. Working memory is required to initially keep track of the incoming signal before subsequently assisting in inferring meaning from the incomplete information gained from the distorted signal. In support of this model, working memory has been shown to be an important cognitive mechanism when perceiving speech in noisy environments (Akeroyd, 2008; Rönnberg et al., 2013; Zekveld, Rudner, Johnsrude, & Rönnberg, 2013). Akeroyd (2008) suggests that after hearing thresholds, working memory is the most effective cognitive mechanism in explaining individual differences in performance on tasks requiring perception of speech in noise. This conclusion is in agreement with research investigating perception of an artificial accent (Banks, Gowen, Munro, & Adank, 2015; Janse & Adank, 2012). **<u>However, working memory has not consistently been found to be a significant predictor of distorted speech perception.</u> For instance,** no relationship was found between individual working memory capabilities and performance requiring perception of foreign-accented (Gordon-Salant, Yeni-Komshian, Fitzgibbons, Cohen, & Waldroup, 2013), frequency compressed (Ellis & Munro, 2013), noise-vocoded (Erb, Henry, Eisner, & Obleser, 2012; Neger, Rietveld, & Janse, 2014), or speech in an array of noise backgrounds (Boebinger et al., 2015). Finally, a meta-analysis from Füllgrabe and Rosen (2016) concluded that for young listeners with normal hearing, differences in working memory account for less than two percent of the variance in speech in noise perception.

A similar inconclusive relationship has also been found between individual differences in attention switching/inhibition. For example, Huyck and Johnsrude (2012) simultaneously exposed their participants to noise-vocoded sentences, auditory distractors, and visual distractors. One group of participants were asked to attend to the vocoded sentences, whilst two other groups performed a target detection task for either the visual or auditory distractors. Huyck and Johnsrude (2012) found that in order to effectively learn to perceive the noise-vocoded speech, simple exposure to the stimuli (as occurred in the two distractor groups) is not sufficient, participants must also attend to the noise-vocoded speech. Additionally, attention switching/inhibition has also been linked with greater overall performance for foreign (Tao & Taft, 2017) and novel accented speech (Adank & Janse, 2010; Banks et al., 2015) with a mediating effect in the perception of noise-vocoded speech (Erb et al., 2012). However, Bent et al. (2016) found no relationship with foreign-accented or regionally-accented speech; Ellis and Munro (2013) found no relationship with frequency compressed speech; and Boebinger et al. (2015) found no relationship between attention switching/inhibition and speech in noise. Finally, whilst individual differences in vocabulary knowledge have mainly only been investigated for adaptation to accented speech, the results thus far have been more consistent, with greater vocabulary knowledge associated with greater adaptation to accented speech across numerous studies **(Adank & Janse, 2010; Janse & Adank, 2012; McLaughlin et al., 2018; Neger et al., 2014).** The current experiment aimed to establish how individual differences in a single battery of audiological and cognitive assessments including pure tone audiometry (PTA), speech recognition thresholds (SRT), working memory, vocabulary knowledge, attention-switching and pattern analysis associate with performance across three distortions in a within-subject design, with particular focus on the degree of overlap or divergence in how each cognitive measure relates to each separate speech condition.

C. Summary of research aims

1. Determine the extent to which perceptual performance for one condition correlates with performance on other forms of distortion.

2. Establish whether transfer of learning effects occur between temporal, spectral or environmental distortions.
3. Determine the extent to which individual differences in a battery of audiological and cognitive assessments relate to performance for each distortion and whether the pattern of associations is consistent across distortions.

I. Methods

A. Participants

Ninety participants took part in this experiment (mean age 21.4 ± 2.74 *SD*; range 18-30; 25 males). All participants were native British English speakers, had normal or corrected to normal vision and were highly educated (15.8yrs ± 1.67). No participants reported a history of speech, language, neurological or psychiatric disorder. **The data for two participants were excluded due to their performance on a preliminary test of general cognition falling below a standardised cut-off score (both participants scored below 26 on the Montreal Cognitive Assessment, Nasreddine et al. 2005).** All participants gave informed consent and were compensated with monetary payment or course credit. The study was approved by the UCL research ethics committee (#0599/001).

B. Procedure

Participants underwent audiological and cognitive testing in addition to the main speech adaptation task. All testing was performed in a double-walled soundproof room and lasted up to 90 minutes.

C. Audiological assessments

Two audiological assessments were performed: (1) Pure Tone Audiometry (PTA) using a clinical audiometer (Maico, MA 41) with each ear tested separately at octave frequencies between 250 and 8000Hz. For each participant, a PTA (average threshold across all measured frequencies) was computed for both ears. (2) Speech Recognition Threshold (SRT) was used to assess the lowest level at which participants could comprehend 50% of an auditorily presented sentence (Plomp & Mimpen, 1979). Each test started at +20dB and varied systematically thereafter. Each sentence had five key words, if participants repeated three or more of the key words then the SNR value would decrease on the subsequent trial, initially in steps of -10dB and subsequently in steps of -2dB, thus making the following trials harder to perceive. The SNR value decreases until participants were only able to comprehend two or fewer of the key words at which point the SNR value would initially increase in steps of +6dB and subsequently in steps of +2dB. The first six lists of the IEEE Harvard Sentences (IEEE, 1969) were used (60 sentences). On average 36 trials/sentences were required to establish each individual speech recognition threshold. Sentences were presented in the same order to all participants.

D. Cognitive assessments

Handedness was assessed using the 10 point Edinburgh Handedness Inventory (Oldfield, 1971). With scores between 50-100 indicative of right-hand dominance and scores from zero-50 indicative of left-hand dominance.

Working Memory was assessed using a forward digit span task. Participants initially heard a set of three numbers and were asked to repeat them back in the same order, for six lists. If participants correctly recalled five or six lists correctly, then the list size increased to four numbers and so on until more than one list was incorrectly recalled for a list size. The last correctly recalled list size was taken as the working memory threshold.

Vocabulary Knowledge was assessed using the auditory version of the spot-the-word section of the Speed and Capacity of Language Processing (SCOLP) test (Baddeley, Emslie, & Nimmo-Smith, 1993; Baddley, Emslie, & Nimmo-Smith, 1992). Participants were presented with 60 pairs of letter strings and had to indicate which one of the letter strings per pair spelled out a real British English word. Reported scores are number of correct identifications out of 60.

Attention-Switching was assessed using the trail-making test (Battery, 1944; Tombaugh, 2004). This task consists of two parts, in part A, participants must draw a line to connect 25 ascending numbers in ascending numerical order (1-2-3-4 etc.) as quickly as possible. In part B, participants have to draw a line to connect 24 circles - 12 of which contain numbers and 12 of which contain letters of the alphabet – in an alternating numerical and alphabetic sequence (1-A-2-B-3-C etc.) again they were required to do this as quickly as possible. We took ratio scores of the two parts as the main outcome statistic (part B/part A).

Pattern/Rule Analysis was assessed using the Wisconsin Card Sorting test (WCST; Grant & Berg, 1948). In this test participants are required to sort a deck of 128 cards into stacks depending on how they correspond to one of four reference cards. Each card (playing and reference) contains a symbol of a certain shape, color and size. The participant must sort the cards depending on one of these features. Critically, participants are initially unaware of how the playing cards and reference cards correspond, with the researcher simply informing them whether each placement is correct or incorrect. After 10 correct placements (for example matching ten playing cards in front of the corresponding color matched reference card) the correspondence rule changes and participants must first notice the rule has changed and then find the new rule. Each of the correspondence rules are repeated twice per test (making six rules), the outcome measure reported here is the number of trials required to complete each rule (i.e. two sets of ten correct placements). Perfect performance would be completing this task in 60 trials.

E. Speech perception task

Task: a computerized version of the SCOLP speed of comprehension test. Participants listened to simple sentences in each of the four conditions outlined below and had to decide whether the sentence was true or false, indicating their response by pressing either the left (true) or right (false) key of a standard PC keyboard. All sentences were clearly true (*'Admirals are people'*) or false (*'Admirals have fins'*). Accuracy and RTs were recorded per trial with adaptation to each condition adjudged via improvements in speed and accuracy of sentence verification. Each set of 48 sentences per condition were retrospectively divided into four sub-blocks of 12 so that the time course of adaptation could be fully assessed.

Stimuli: The auditory sentences were recordings of 192 SCOLP sentences, 96 of which were true and 96 false, with 48 sentences presented per speech condition. All sentences were recorded by four different male speakers of standard Southern British English. At time of recording all speakers were between 30-32 years of age and all were born, raised and educated in South East England (see Adank, Evans, Stuart-Smith, & Scott, 2009 for more details on the recording parameters). Sentences from each speaker were used 12 times per condition with the order of speaker randomized (results related to the effect of multiple speakers will not be discussed here). All sentences were saved to separate files with the beginning and end trimmed to zero crossings as closely as possible to the onset/offset of the initial/final speech sounds, resampled to 22050 Hz, peak normalized to 99% of maximum amplitude and scaled to 70dB SPL using Praat (Boersma & Weenink, 2011). Stimulus presentation was performed using a custom-made MATLAB 2014a program (The MathWorks Inc., Natick, MA, 2000) and Sennheiser headphones, with all stimuli delivered at a comfortable listening level (preset at 74dB SPL but where necessary this was adjusted to fit individual participant preference).

Participants' ability to perceive different types of speech was tested using four different conditions. (1) Time compressed sentences shortened to 40% of their original length using PSOLA implemented in Praat (Charpentier & Stella, 1986). (2) Noise-vocoded sentences were filtered into four logarithmically spaced frequency bands from 50 to 5000Hz (50-528; 528-1248; 1248-2541; 2541-5000Hz). (3) Speech in noise sentences were embedded in a stream of speech-shaped noise at a signal to noise ratio of -4dB. The spectrum of the speech-shaped noise was derived from the 192 sentences used in the adaptation task. (4) Speech-in-quiet sentences were presented without any manipulation (beyond the zero trimmings, peak normalization etc. outlined above). The speech-in-quiet sentences condition was always the first condition that all participants heard. **Presenting the speech-in-quiet sentences consistently as the first condition** ensured that any task practice effects were overcome before the distorted speech stimuli were encountered. Theoretically, therefore any improvement in task performance for the time-compressed, noise-vocoded and speech in noise conditions comes purely from the participants adapting to the specific manipulation and was not due to greater familiarity with the task. Order of presentation of the three non-speech in quiet conditions was fully randomized between participants. Half of the participants were tested on the audiological and cognitive measures first followed by the adaptation task; the other half of participants had the opposite order to ensure results were not due to fatigue. No significant effect of this procedural manipulation was found (*p*>.25) and therefore in all subsequent analyses the data are collapsed across this variable.

II. Results

Differences in performance on the speech perception task across the different speech conditions and sub-blocks of the experiment were assessed using two separate mixed analysis of variance (mixed ANOVA) performed in SPSS 25. One ANOVA was for accuracy and one was for response time (RT) data with condition (speech-in-quiet, time-compressed, noise-vocoded, speech in noise) and sub-block (each set of 12 sentences) as within-subjects factors and order of distortion presentation (speech-in-quiet (Q)-speech in noise (N)-time-compressed (T)-noise-vocoded (V); Q-N-V-T; Q-T-N-V; Q-T-V-N; Q-V-N-T; Q-V-T-N) as a between-subjects factor. For the accuracy descriptive statistics and figures, raw percentage scores are reported, however for the mixed ANOVA the data were transformed to rationalized arcsine units (RAU) to ensure consistent variance over the range of scores obtained (Studebaker, 1985).

A. Accuracy

Accuracy was highest for speech in quiet (*M* = 95.8, *SD* = 4.45), followed by the time-compressed (*M* = 85.3, *SD* = 8.86), speech in noise (*M* = 70.2, *SD* = 10.7) and noise-vocoded speech (*M* = 60.7, *SD* = 12.6). The mixed ANOVA found a significant main effect of condition $F(3, 246)=527$, $p<.001$, $\eta_p^2 = .865$ and post-hoc paired samples *t*-tests revealed significant differences in percentage correct between all four speech conditions at a Bonferroni corrected alpha level ($p = .05/6 = .008$; see Fig. 1 and Supplementary table I[1]). In addition, the mixed ANOVA also revealed a significant main effect of sub-block $F(3,246)=11.3$, $p<.001$, $\eta_p^2 = .121$, reflecting an improvement in performance from the first to the last sub-block (see Fig. 1 and Supplementary table Ii). Post-hoc paired samples *t*-tests revealed significant differences in the average percentage correct for the first 12 sentences in a block of 48 compared to the third set of 12 (sentences 25-36) $t(87) = −5.52$, $p<.001$, Cohen's $d = −0.59$; a significant difference between the first and final sub-block of 12 sentences $t(87) = −4.03$, $p<.001$, Cohen's $d = −0.43$; and a significant difference between the average percent correct for the second and third sub-blocks $t(87) = −3.68$, $p<.001$, Cohen's $d = −0.39$. The comparisons between the first and second sub-blocks $t(87) = −2.1$, $p=.038$, Cohen's $d = −0.22$ and the second and fourth sub-blocks $t(87) = -2$, $p=.049$, Cohen's $d = -0.21$ did not survive Bonferroni correction ($p = .05/6 = .008$).

The mixed ANOVA revealed a significant three-way interaction of condition, sub-block and order $F(45, 738)=372$, $p<.001$, $\eta_p^2 = .108$, reflecting differential rates of adaptation between conditions depending on the order in which the participants were exposed to the different distortions. This interaction was investigated with separate two-way repeated measures ANOVAs for each order. This follow-up analysis found a significant condition by sub-block interaction for the Q-N-T-V order $F(9, 117)=4.3$, $p<.001$, $\eta_p^2 = .249$ with post-hoc paired samples *t*-tests showing a significant difference between the first and third sub-blocks in the time-compressed condition $t(13)=-4.28$, $p=.001$, Cohen's $d =1.14$ and between the first and last sub-blocks for the noise condition $t(13)=-4.17$, $p=.001$, Cohen's $d =1.12$. All remaining comparisons had a significance level greater than the Bonferroni-corrected alpha level (.05/24 = .002). In addition, a significant condition×sub-block interaction was found in the Q-T-N-V order $F(9, 117)=1.99$, $p=.046$, $\eta_p^2 = .13$, but no follow-up comparison survived Bonferroni correction (Supplementary Fig. 1a2).

The mixed ANOVA also revealed a significant condition×order interaction. When investigating the main effect of condition separately for each order, the speech-in-quiet, time-compressed and speech in noise conditions all differed significantly from each other irrespective of their position in the order. Whilst performance in the noise-vocoded condition always differed significantly from the speech-in-quiet and time-compressed condition (always significantly poorer), differences between the noise-vocoded and speech in noise conditions only occurred when the noise-vocoded condition was second in the order (Q-V-N-T and Q-V-T-N). In both cases, performance was significantly poorer in the noise-vocoded relative to speech in noise condition. This interaction suggests that participants found the noise-vocoded condition especially difficult when they had not yet encountered any of the other distortions (as speech in quiet always came first in the order). The interaction between speech condition and sub-block approached significance $F(9, 738)=1.87$, $p=.054$, $\eta_{p}^{2} = .022$. All other main effects and interactions were non-significant (p's >.05).

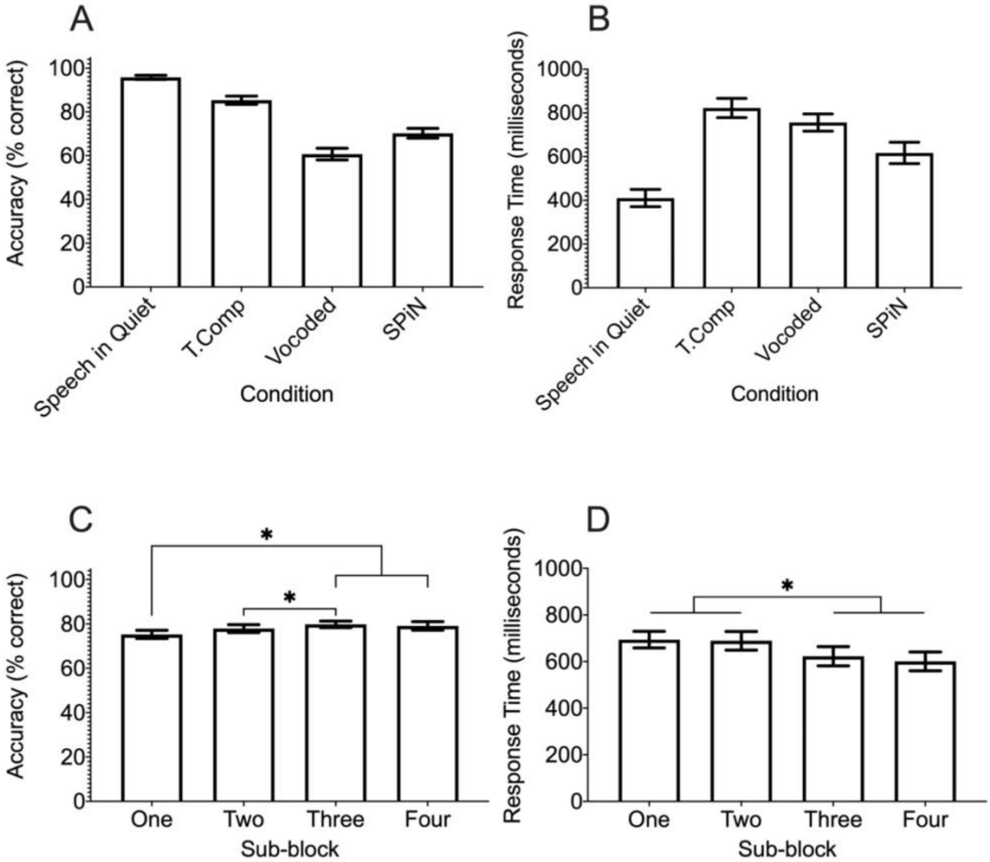

FIG. 1. Average accuracy and response time data for each condition (A, B) and sub-block (C, D). Errors bars represent 95% confidence intervals of the mean. For A and C all comparisons are significantly different at Bonferroni corrected alpha level of .008. For C and D a * represents all comparisons that are significantly different at the Bonferroni corrected alpha level of .008.

B. Response times

Response times (RTs) in milliseconds were analysed for correct responses only. RTs were measured relative to the end of each sentence, therefore allowing for negative RTs. Overall participants were quickest to respond in the speech in quiet condition (*M* = 410, *SD* = 186) followed by speech in noise (*M* = 617, *SD* = 230), then noise-vocoded (*M* = 756, *SD* = 185) and finally the slowest overall RTs were in response to time-compressed speech (*M* = 823, *SD* = 206). The mixed ANOVA revealed a significant main effect of condition $F(2.69, 220)=200$, $p<.001$, $\eta_{p}^{2}=.709$, post-hoc comparisons revealed significant differences in RTs between all four conditions at the corrected alpha level (p =.008; Fig. 1, Supplementary table I_1). Additionally, the mixed ANOVA revealed a significant effect of sub-block $F(2.49, 246)=31.5$, $p<.001$, $\eta_{p}^{2}=.278$, post-hoc paired samples *t*-tests revealed significant RT differences between the first two and final two sub-blocks. The comparison between the third and fourth sub-blocks $t(87)=-2.03$, $p=.046$, Cohen's $d=-0.22$ did not survive Bonferroni correction ($p = .05/6 = .008$), no difference was found between the first and second sub-blocks (Fig. 1, Supplementary table I_1).

The mixed ANOVA result revealed a significant condition×sub-block×order interaction $F(36.3, 596)=1.48$, $p=.037$, $\eta_{p}^{2}=.083$. When investigating the condition by sub-block interaction separately for each order a significant two-way interaction was only found for the Q-N-V-T order $F(9,126)=3.15$, $p=.002$, $\eta_{p}^{2}=.184$, within this order, the only post-hoc paired samples *t*-test to survive Bonferroni correction was the comparison between the first and third sub-block for the speech in quiet condition $t(14)=4.8$, $p<.001$, Cohen's $d = 1.24$, reflecting the quicker response of participants in the third set of 12 sentences relative to the first set of 12 sentences in the speech in quiet condition (Supplementary Fig. $1b_3$).

The mixed ANOVA also revealed a significant condition×sub-block interaction, $F(7.27, 596)=3.61$, $p=.001$, $\eta_{p}^{2} = .042$. When analysing the four conditions in separate RM ANOVAs, a significant effect of sub-block was found for all conditions: speech in quiet $F(2.48, 216)=17.8$, $p<.001$, $\eta_{p}^{2} = .17$; time-compressed speech $F(2.65, 230)=17.8$, $p<.001$, $\eta_{p}^{2} =.170$; noise-vocoded $F(3, 261)=12.5$, $p=.001$, $\eta_{p}^{2} =.125$; and speech in noise $F(3, 261)=3.14$, $p=.026$, $\eta_{p}^{2} =.035$. In the speech in quiet condition, a significant difference was found in RTs between all sub-blocks except the third and fourth where reductions in RTs levelled off. In the time compressed condition, the first two sub-blocks differed significantly from the final two sub-blocks and for the speech in noise condition a significant difference was found between the second and fourth sub-block. In all condition's participants became quicker to make a correct response as the number of trials increased. For the noise-vocoded speech, however, a significant difference was observed between the second sub-block and all other sub-blocks, this difference was due to participants on average becoming slower to respond in the second sub-block of 12 sentences compared to the other sub-blocks (see Fig. 2). All other main effects and interactions were non-significant (*p*'s >.05).

In summary, overall adaptation was most evident in the response time data where participants became quicker to make a correct response as the number of trials increased for the speech-in-quiet, time-compressed and speech in noise conditions. Whilst, transfer of learning effects were most noticeable in the accuracy data where participants found the noise-vocoded condition especially difficult when they had not yet encountered any of the other distortions.

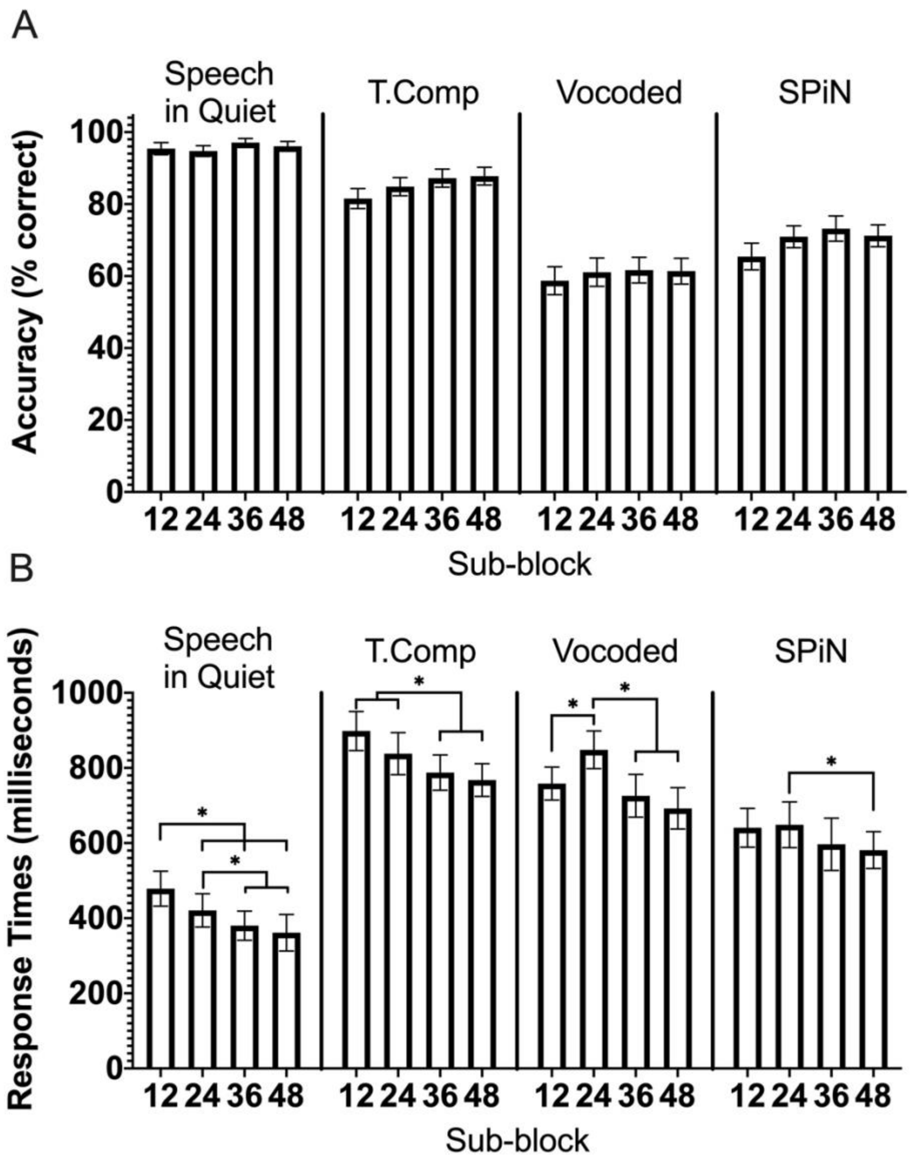

FIG. 2. A. Accuracy of responses for each condition across the four sub-blocks per condition. **B**. Average RTs for each condition across the four sub-blocks per condition. Error bars represent 95% confidence intervals of the mean. All significant differences are represented by a *. T.Comp = Time compressed, Vocoded = Noise Vocoded, SPiN = Speech in Noise.

C. Correlation in performance between conditions

Results for both accuracy and RT data reveal significant positive Pearson correlations between all conditions (p < .008; Fig. 3, Supplementary table II[4]). This result suggests that participants appear to possess (or lack) a general ability to perform relatively equally in different adverse listening conditions, irrespective of the type of distortion (spectral, temporal or environmental).

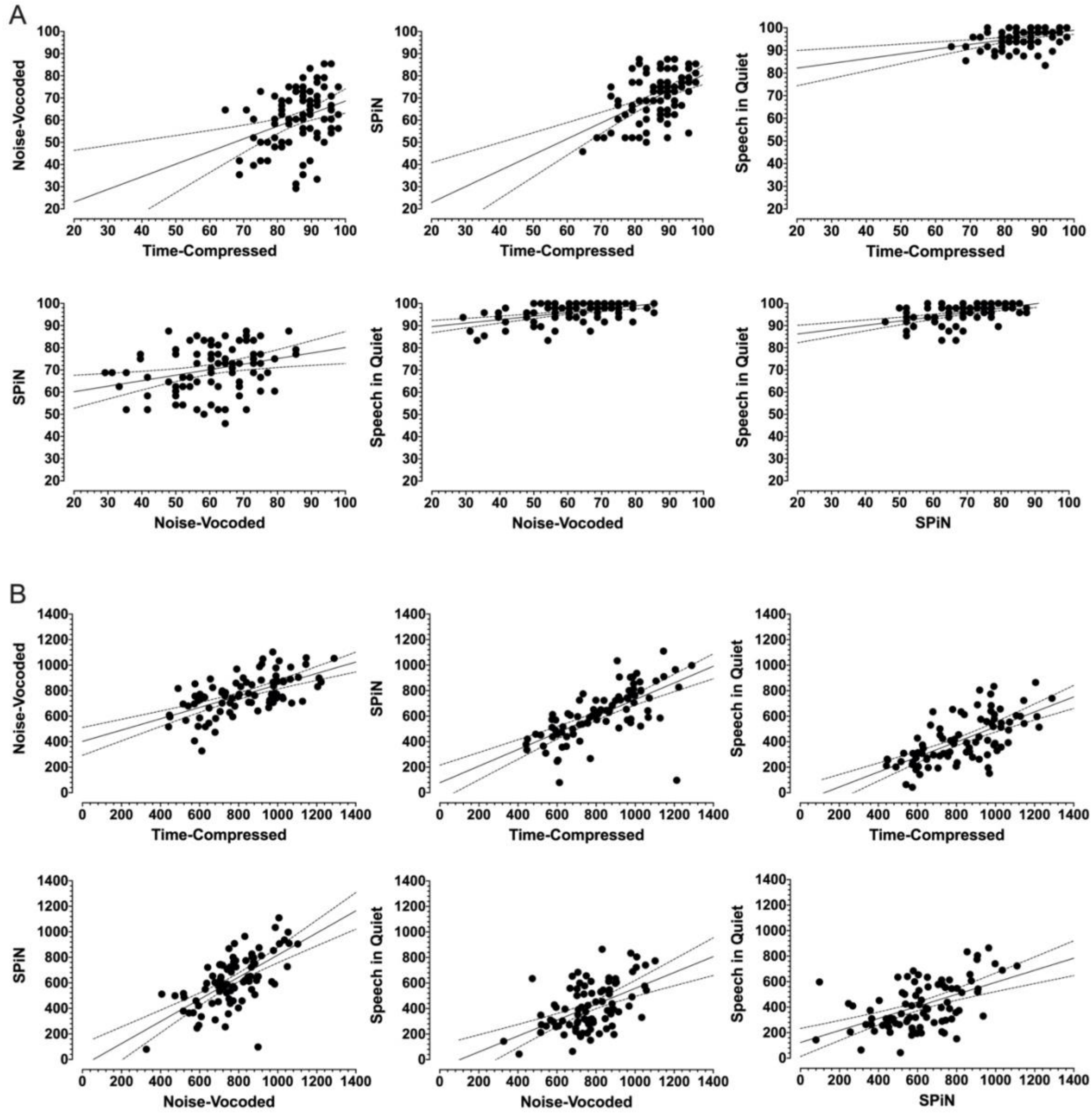

FIG. 3. Scatterplot matrices representing the correlations between overall accuracy (A) and RTs (B) for each participant across the 4 speech conditions. Dotted lines represent 95% confidence intervals of the regression line. All subplots represent significant positive correlations.

D. Relationship between cognitive assessments and performance in different speech conditions

Overall performance on each of the audiological and cognitive assessments was high, as would be expected from a homogenous young, highly educated population of participants (Table I). Prior to establishing the relationship between individual differences in audiological/cognitive ability and performance on each of the speech distortions a Principal Component Analysis (PCA) was run on the audiological and cognitive assessment scores to establish correlations between variables. Initially, average PTA thresholds for both ears, SRTs, working memory, vocabulary knowledge, attention-switching (trail-making test ratio) and pattern analysis (number of trials required to complete all rules on WCST) were included in the analysis. Scores on the SRT, attention-switching or pattern analysis tests showed no correlation coefficients >.3 and were therefore excluded from the subsequent PCA. Bartlett's test of sphericity was found to be significant indicating sufficient between variables correlations in the remaining measures to be suitable for a PCA ($X_2(21) = 54.6$, $p<.001$).

Table I. Descriptive statistics for battery of audiological and cognitive tests.

Assessment	N	*M*	*SD*	Min.	Max.
PTA Left Ear	88	3.90	4.74	-5	25
PTA Right Ear	88	5.30	5.58	-2.50	35
SRT	87	-3.78	1.11	-7	0
Working Memory	88	6.31	1.22	4	9
Vocabulary Knowledge	88	50.1	3.77	42	58
Attention Switching	88	2.14	0.72	1.01	5.65
Pattern Analysis	88	77	11.7	63	129

Two PCA components showed eigenvalues >1 and together explained 70.5% of all variance. A Varimax orthogonal rotation was employed to aid interpretability. Component 1 loads most strongly onto vocabulary knowledge and working memory. Component 2 reflects general hearing ability loading most strongly onto the PTA thresholds of the two ears.

Table II. Summary of Principal component analysis loadings. Bolded font represents the strongest loadings for each item.

	Rotated Component Coefficients	
Items	1	2
Vocabulary Knowledge	**.834**	-.109
Working Memory	**.839**	.077
PTA Left Ear	.077	**.844**
PTA Right Ear	-.180	**.811**

Multiple linear regressions were conducted to establish if performance in each speech distortion condition could be predicted based on individual differences in both of the two components and the SRT, attention switching and pattern analysis tasks (separate analyses were conducted for accuracy and response time data). The regression model significantly predicted overall accuracy performance in the speech-in-quiet, time-compressed and speech in noise conditions (Table III) but not in the noise-vocoded condition (p=.541). Across all three significant models, component 1 significantly predicted performance, suggesting a link between vocabulary knowledge, working memory and performance in the different distortions.

Table III. Summary of multiple regression analyses for RAU accuracy data across conditions.

Clear: $F(5,79)=3.42$, $p=.008$, adj. $R^2=.126$

Variable	*B*	*SE_B*	*β*	*p*
Intercept	108	7.75		<.001
Component One	2.01	0.96	.226	**.038**
Component Two	-0.6	0.93	-.066	.525
SRT	-1.86	0.88	-.232	**.037**
Attention Switching	-0.52	1.3	-.042	.69
Pattern Analysis	-0.09	0.08	-.112	.304

Time-Compressed: $F(5,79)=4.13$, $p=.002$, adj. $R^2=.157$

Variable	*B*	SE_B	*B*	*p*
Intercept	85.4	9.18		<.001
Component One	3.51	1.13	.327	**.003**
Component Two	-0.3	1.1	-.027	.789
SRT	-1.73	1.04	-.179	.1
Attention Switching	1.78	1.54	.119	.25
Pattern Analysis	-0.08	0.1	-.085	.426

Noise-Vocoded: $F(5,79)=0.82$, $p=.541$, adj. $R^2=-.011$

Variable	*B*	SE_B	*B*	*p*
Intercept	52.7	11.8		<.001
Component One	0.97	1.46	.077	.509
Component Two	-0.02	1.42	-.002	.988
SRT	-2.07	1.34	-.181	.127
Attention Switching	0.5	1.99	.028	.802
Pattern Analysis	-0.02	0.13	-.014	.904

SPiN: $F(5,79)=2.64$, $p=.029$, adj. $R^2=.089$

Variable	*B*	SE_B	*B*	*p*
Intercept	61.3	10.3		<.001
Component One	3.59	1.27	.311	**.006**
Component Two	-0.29	1.24	-.025	.813
SRT	-1.09	1.16	-.105	.351
Attention Switching	2.38	1.72	.148	.172
Pattern Analysis	-.001	0.11	-.001	.994

Notes: *B* = standardised regression coefficient; SE_B = Standard error of the coefficient; *β* = standardised coefficient. All *p*-values highlighted in bold are significant at an alpha level of <0.05

For the RT data, the multiple linear regression model was non-significant for all of the distorted speech conditions (all *p*'s>.05; Table IV).

Table IV. Summary of multiple regression analyses for response time data across conditions.

Clear: $F(5,79)=0.73$, $p=.607$, adj. $R^2=-.017$

Variable	B	SE_B	β	p
Intercept	389	172		<.001
Component One	-10.4	21.3	-.056	.626
Component Two	17.8	20.5	.098	.388
SRT	13.03	19.4	.079	.503
Attention Switching	-17	29.1	-.066	.561
Pattern Analysis	1.46	1.82	.093	.424

Time-Compressed: $F(5,79)=2.25$, $p=.057$, adj. $R^2=.069$

Variable	B	SE_B	β	p
Intercept	903	182		<.001
Component One	-22.4	22.5	-.11	.323
Component Two	53.7	21.7	.266	**.016**
SRT	30.1	20.5	.165	.146
Attention Switching	13.3	30.8	.047	.667
Pattern Analysis	0.15	1.92	.008	.939

Noise-Vocoded: $F(5,79)=0.9$, $p=.486$, adj. $R^2= -.006$

Variable	B	SE_B	β	p
Intercept	827	141		<.001
Component One	17.4	17.4	.142	.22
Component Two	18.5	16.8	.123	.274
SRT	19.6	15.9	.144	.22
Attention Switching	27.8	23.9	.131	.247
Pattern Analysis	-0.55	1.49	-.042	.715

SPiN: $F(5,79)=.868$, $p=.507$, adj. $R^2=-.008$

Variable	B	SE_B	β	p
Intercept	871	190		<.001
Component One	5.65	23.4	.028	.81
Component Two	32.4	22.6	.161	.156
SRT	27.9	21.3	.153	.194
Attention Switching	15.2	32	.053	.637
Pattern Analysis	-2.33	2	-.135	.248

Notes: B = standardised regression coefficient; SE_B = Standard error of the coefficient; β = standardised coefficient. All p-values highlighted in bold are significant at an alpha level of <0.05

III. Discussion

The present study investigated whether participants show a consistent perceptual profile across three speech distortions: time-compressed, noise-vocoded and speech in noise.

Additionally, we investigated whether/how individual differences in performance on a battery of audiological and cognitive tasks links to perceptual performance to uncover the underlying cognitive mechanisms that underpin the perceptual process.

A. <u>**Speech perception in different listening conditions**</u>

Speech perception performance was highest in the speech in quiet, followed by time-compressed, speech in noise, and noise-vocoded speech condition. Second, individual participants' performance in one condition was highly predictable from their performance in the three other conditions. This supports the notion that participants possess (or lack) a general ability to successfully perceive multiple forms of speech, even when the distortions differ in the degree of spectral, temporal or environmental manipulation. It should be noted that the significant correlations in both the current experiment and those of Bent et al. (2016) and Borrie et al. (2017) do not imply a causal link and thus do not definitively illustrate a common cognitive mechanism underpinning general auditory perceptual adaptation to any form of distorted speech. It is possible, even likely, that listeners can reach a similar level of perception for *different* distortions using *different* cognitive strategies with individuals who are skilled on one strategy also being more skilled on other strategies. As highlighted previously, we think the perceptual system is stressed differently by the speech distortions used in this experiment. The temporal distortion likely resulted in changes at phonological processing levels (Sebastián-Gallés et al., 2000), whereas the spectral and environmental distortions likely resulted in changes at lexical/semantic levels (Davis et al., 2005). Future research could investigate this further by using functional imaging to establish whether the neural patterns that occur during the perceptual/adaptation processes are similar or different across distortions, <u>**this research**</u> may elucidate whether one common mechanism underpins this process or whether multiple mechanisms are required/responsible (Adank, Davis, & Hagoort, 2012; Adank & Devlin, 2010). Such work could consequently inform research using neuromodulatory techniques such as Transcranial Magnetic Stimulation, which could be used to demonstrate causality between the observed neural activation and the associated cognitive mechanisms during tasks requiring perceptual adaptation to distorted speech.

In the current experiment, individual differences in accuracy performance in the speech in quiet, time-compressed and speech in noise conditions were all associated with individual differences in performance on tests of working memory and vocabulary knowledge. These results are strikingly similar to the results of McLaughlin et al. (2018) who also found a link between (receptive) vocabulary knowledge, working memory, and performance on a speech perception task, using accented speech. The association between greater vocabulary knowledge and perception of accented speech has been shown previously (Adank & Janse, 2010; Banks et al., 2015; Bent et al., 2016). The findings from the present experiment, extend those from the accented speech literature to show that individual differences in vocabulary knowledge and working memory are also associated with individual differences in the ability to perceive speech in quiet and artificially time-compressed speech. McLaughlin et al. (2018) suggest that individuals with greater receptive vocabularies are able to perceive distorted speech to a greater extent "*because they have stronger lexical mappings that allow them to access semantic representations from input even when it is environmentally degraded*" (McLaughlin et al, 2018, page 1567). It is expected that the increased working memory capability can be used to retain larger chunks of the degraded speech signal for longer. This retention of larger chunks of information in turn allows the listener to analyse and compare the stimulus with pre-existing speech templates. Thus, a larger vocabulary may increase the chance of correctly identifying words and/or phrases in the distorted stimuli, with improved speech perception as a result.

Future research could build on the current research to investigate which other mechanisms, if any, enable successful perception of speech in different adverse listening conditions in an attempt to build a more comprehensive model of the supporting cognitive and neural mechanisms. One potential mechanism to explore is statistical learning. Research in infant language learning suggests that statistical learning is critical for early language acquisition and development (Romberg & Saffran, 2010; Saffran, 2003). Similar statistical learning abilities may be utilised in older children and adults when faced with the challenge of adapting to distorted speech. For example, Neger et al. (2014) found that perceptual learning of distorted speech was modified by statistical learning ability, with participants who showed better performance on a statistical learning task also showing greater perceptual learning of noise-vocoded speech (in their younger group of participants, but not in the older group). Additionally, Neger et al. (2014) found a significant effect of vocabulary knowledge on perceptual learning of noise-vocoded speech. The authors argue that perceptual learning abilities may rely directly on sensitivity to the probabilistic information inherent in all speech (i.e., statistical learning). It is believed that individuals who are more capable, and faster, to identify subunits of the distorted speech signal are able to transfer this information to higher level processors thus facilitating faster access to their larger store of lexical representations and greater overall perception and potential adaptation to the distortion. Future research should extend this research and investigate the role of statistical learning in perception of distorted speech, along with other potential cognitive mechanisms and investigate how each of the different mechanisms interact to underpin successful perception of speech in adverse listening conditions.

We found that vocabulary knowledge and working memory did not predict accuracy of performance in the noise-vocoded condition. It is possible that this result is indicative of an alternative mechanism sub serving perception of vocoded speech. Alternatively, the parameters we used to vocode the speech were more stringent than previous studies (Davis et al., 2005; Hervais-Adelman et al., 2012; Hervais-Adelman et al., 2008; Huyck & Johnsrude, 2012). Indeed Loizou et al. (1999) noted that adaptation to noise-vocoded speech drops rapidly below five channels. Future research could establish whether perception of noise-vocoded speech is underpinned by a separate cognitive mechanism compared to other speech distortions or whether the lack of a relationship found in the current experiment is the result of the parameters used during vocoding.

B. Transfer of learning effects

With reference to Reverse Hierarchy Theory (Ahissar et al., 2009), Adank and Janse (2009) explain their transfer of learning result on the basis of the easier to comprehend artificial condition providing the listener with a training signal (Hervais-Adelman et al., 2008) for the harder naturally fast spoken sentences. In the current experiment, overall performance for the noise-vocoded condition was poorest of all four conditions, however performance improved for the vocoded condition when listeners had encountered either or both time-compressed and speech in noise conditions before the noise-vocoded condition. When encountered third or fourth in the order, participants would have experience of either 96 or 144 occurrences of the noun plus predicate structure of the target sentences. The superior knowledge of the sentence structure may have helped the listeners to perform better on the noise-vocoded condition when it occurred later in the order. These advantages would be absent when noise-vocoding occurred immediately after the speech in quiet condition and thus this knowledge could not be transferred to assist in performance on this distortion resulting in poorer performance when noise-vocoding was encountered early in the order. Furthermore, of the three distorted speech conditions, the noise-vocoded condition represents the condition that was designed to sound most different from anything experienced regularly by individuals with hearing in the normal range. Listeners can be expected to hear speech in quiet and fast speech on a daily basis but cannot be expected to hear the specific changes to the spectral composition of the sentences in the noise-vocoded condition. Therefore, it is possible that the poorer performance in the noise-vocoded condition, when presented first, may in part be attributable to attentional effects. Indeed, Floccia, Butler, Goslin, and Ellis (2009) found that the initial perturbation in performance (accuracy and RTs) that is often observed when a stimulus changes from one condition to another e.g., from speech in quiet to noise-vocoded speech, can be dependent on the task instructions. Floccia et al. (2009) found that individuals who were aware that the accent changed during testing - and who were exposed to this accent during training – showed less perturbation when the accent changed. **The reduced perturbation** was the case for those who were not told to specifically focus on the differences between accents only. Future studies could explore whether the effect of providing the participants with 1 or 2 “training” sentences before the test phase of the experiment would reduce any initial attentional effects (Peelle & Wingfield, 2005). In turn, this reduction could then change the nature of any transfer of learning effects, as found in the current experiment.

C. Limitations and directions for future research

There are a few limitations of the current experiment that could be addressed in future research. For example, the working memory task that was chosen required participants to listen as a list of numbers were read to them, internally rehearse the list and subsequently repeat the list in the exact (forward) order as the list was presented to them. While this design fits within the definition of working memory as the ‘temporary storage of information in connection with performance of other cognitive tasks’ (Baddeley, 1983), it might be useful for future research to use a working memory task that has been more closely related to the more traditional definition of working memory as an online memory manipulation process, such as an operation span or reading span task (Banks et al., 2015).

Future studies could consider task-related effects on performance. We used a speeded sentence-verification task with accuracy and RTs of responses as a measure of perception differences between conditions. While accuracy performance was above chance, sentence verification is a somewhat crude measure of successful perception of speech, especially in difficult listening environments. A task that requires word identification and recall, e.g., a transcription task, might have been better. First, a transcription task requires participants to explicitly recognise, recall, and reproduce each word of the sentence, which is arguably associated with more focused and extensive linguistic processing. Second, this task could also provide a more fine-grained accuracy measure. However, the downside of such a task would be that it would not be feasible to measure RTs, so the speed of processing could not be assessed. Nevertheless, as it is still possible that the type of task interacts with processing of different types of distortions of the speech signal, different tasks should be considered in future research.

IV. Conclusions

In conclusion, results from the current experiment suggest that individual listeners possess a general skill to adapt to various speech distortions. This skill allows participants to perform to a relatively equivalent level across different conditions irrespective of whether the distortion is temporal, spectral or environmental in nature. The results of this experiment suggest that measures of vocabulary knowledge and working memory could underpin the perceptual learning process. **Overall, the current research adds to and extends the work of Adank and Janse (2009), Bent et al. (2016), Borrie et al. (2017) and McLaughlin et al. (2018).**

The present work supports previous work by providing a more comprehensive overview of the potential underlying cognitive mechanism(s) that underpins perception across a range of different speech distortions, an area of the field that is currently under researched. Future research should try to establish whether equivalent perception across distortions is dependent on the use of a singular neural-cognitive-perceptual mechanism in all adverse listening conditions as suggested here or whether different mechanisms are employed dependent on the distortion.

1 See supplementary material at [URL will be inserted by AIP] for a summary of the pairwise comparison statistics for accuracy and response times.

2 See supplementary material at [URL will be inserted by AIP] for a figure showing the average accuracy data for each condition across the four sub-blocks and each order of presentation.

3 See supplementary material at [URL will be inserted by AIP] for a figure showing the average response time for each condition across the four sub-blocks and each order of presentation.

4 See supplementary material at [URL will be inserted by AIP] for a table of the Pearson correlation coefficients for average performance between conditions for RAU accuracy and response time data.

.

References

Adank, P., Davis, M. H., & Hagoort, P. (2012). Neural dissociation in processing noise and accent in spoken language comprehension. *Neuropsychologia*, *50*(1), 77-84. doi:https://doi.org/10.1016/j.neuropsychologia.2011.10.024

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage*, *49*(1), 1124-1132. doi:https://doi.org/10.1016/j.neuroimage.2009.07.032

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J Exp Psychol Hum Percept Perform, 35*(2), 520-529. doi:10.1037/a0013552

Adank, P., & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *J Acoust Soc Am, 126*(5), 2649-2659. doi:10.1121/1.3216914

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychol Aging, 25*(3), 736-740. doi:10.1037/a0020054

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1515), 285-299. doi:10.1098/rstb.2008.0253

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int J Audiol, 47 Suppl 2,* S53-71. doi:10.1080/14992020802301142

Baddeley, A. (1983). Working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences, 302*(1110), 311-324. doi:doi:10.1098/rstb.1983.0057

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The Spot-the-Word test: a robust estimate of verbal intelligence based on lexical decision. *Br J Clin Psychol, 32 (Pt 1),* 55-65.

Baddley, A., Emslie, H., & Nimmo-Smith, I. (1992). *The Speech and Capacity of Language Processing Test manual*. Suffolk, UK: Thames Valley Test Company.

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *J Acoust Soc Am, 137*(4), 2015-2024. doi:10.1121/1.4916265

Battery, A. I. T. (1944). *Manual of Directions and Scoring*. Washington D.C., USA: War Department, Adjuntant General's Office.

Bent, T., Baese-Berk, M., Borrie, S. A., & McKee, M. (2016). Individual differences in the perception of regional, nonnative, and disordered speech varieties. *The Journal of the Acoustical Society of America, 140*(5), 3775-3786. doi:10.1121/1.4966677

Bilodeau-Mercure, M., Lortie, C. L., Sato, M., Guitton, M. J., & Tremblay, P. (2015). The neurobiology of speech perception decline in aging. *Brain Struct Funct, 220*(2), 979-997. doi:10.1007/s00429-013-0695-3

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., & Scott, S. K. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *J Acoust Soc Am, 137*(1), 378-387. doi:10.1121/1.4904537

Boersma, P., & Weenink, D. (2011). Praat: doing phonetics by computer (Version 5.4.02).

Borrie, S. A., Baese-Berk, M., Engen, K. V., & Bent, T. (2017). A relationship between processing speech in noise and dysarthric speech. *The Journal of the Acoustical Society of America, 141*(6), 4660-4667. doi:10.1121/1.4986746

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America, 121*(4), 2339-2349. doi:10.1121/1.2642103

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729. doi:10.1016/j.cognition.2007.04.005

Burk, M. H., Humes, L. E., Amos, N. E., & Strauser, L. E. (2006). Effect of training on word-recognition performance in noise for young normal-hearing and older hearing-impaired listeners. *Ear Hear, 27*(3), 263-278. doi:10.1097/01.aud.0000215980.21158.a2

Cainer, K. E., James, C., & Rajan, R. (2008). Learning speech-in-noise discrimination in adult humans. *Hearing Research, 238*(1), 155-164. doi:https://doi.org/10.1016/j.heares.2007.10.001

Charpentier, F., & Stella, M. (1986, Apr 1986). *Diphone synthesis using an overlap-add technique for speech waveforms concatenation.* Paper presented at the ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *J Acoust Soc Am, 116*(6), 3647-3658.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical Processing in Spoken Language Comprehension. *The Journal of Neuroscience, 23*(8), 3423-3431.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A. G., 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), 222-241. doi:10.1037/0096-3445.134.2.222

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *J Exp Psychol Hum Percept Perform, 23*(3), 914-927.

Ellis, R. J., & Munro, K. J. (2013). Does cognitive function predict frequency compressed speech recognition in listeners with normal hearing and normal cognition? *Int J Audiol, 52*(1), 14-22. doi:10.3109/14992027.2012.721013

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2012). Auditory skills and brain morphology predict individual differences in adaptation to degraded speech. *Neuropsychologia, 50*(9), 2154-2164. doi:http://dx.doi.org/10.1016/j.neuropsychologia.2012.05.013

Fairbanks, G., & Jr., F. K. (1957). Word Intelligibility as a Function of Time Compression. *The Journal of the Acoustical Society of America, 29*(5), 636-641. doi:10.1121/1.1908992

Füllgrabe, C., & Rosen, S. (2016). On The (Un)importance of Working Memory in Speech-in-Noise Processing for Listeners with Normal Hearing Thresholds. *Frontiers in Psychology, 7*(1268). doi:10.3389/fpsyg.2016.01268

Goldstone, R. L. (1998). Perceptual learning. *Annu Rev Psychol, 49*, 585-612. doi:10.1146/annurev.psych.49.1.585

Golomb, J. D., Peelle, J. E., & Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *J Acoust Soc Am, 121*(3), 1701-1708.

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., Cohen, J. I., & Waldroup, C. (2013). Recognition of accented and unaccented speech in different maskers by younger and older listeners. *The Journal of the Acoustical Society of America, 134*(1), 618-627. doi:10.1121/1.4807817

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology, 38*(4), 404-411. doi:10.1037/h0059831

Hervais-Adelman, A. G., 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), 460-474. doi:10.1037/0096-1523.34.2.460

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *J Exp Psychol Hum Percept Perform, 37*(1), 283-295. doi:10.1037/a0020772

Huyck, J. J., & Johnsrude, I. S. (2012). Rapid perceptual learning of noise-vocoded speech requires attention. *The Journal of the Acoustical Society of America, 131*(3), EL236-EL242. doi:10.1121/1.3685511

IEEE. (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE No 297-1969*, 1-24. doi:10.1109/IEEESTD.1969.7405210

Janse, E., & Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *The Quarterly Journal of Experimental Psychology, 65*(8), 1563-1585. doi:10.1080/17470218.2012.658822

Kennedy-Higgins, D. (2019). *Neural and cognitive mechanisms affecting perceptual adaptation to distorted speech.* UCL (University College London),

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *J Acoust Soc Am, 106*(4 Pt 1), 2097-2103.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7-8), 953-978. doi:10.1080/01690965.2012.705006

Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of Second-Language Acquisition and Perception of Speech in Noise. *Journal of Speech, Language, and Hearing Research, 40*(3), 686-693. doi:10.1044/jslhr.4003.686

McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., Van Engen, K. J. J. A., Perception,, & Psychophysics. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. *80*(6), 1559-1570. doi:10.3758/s13414-018-1537-4

Mehler, J., Sebastian, N., Altmann, G., Dupoux, E., Christophe, A., & Pallier, C. (1993). Understanding Compressed Sentences: The Role of Rhythm and Meaning a. *Ann N Y Acad Sci, 682*(1), 272-282. doi:10.1111/j.1749-6632.1993.tb22975.x

Moore, B. C., Peters, R. W., & Stone, M. A. (1999). Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips. *J Acoust Soc Am, 105*(1), 400-411.

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., . . . Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc, 53*(4), 695-699. doi:10.1111/j.1532-5415.2005.53221.x

Neger, T. M., Rietveld, T., & Janse, E. (2014). Relationship between perceptual learning in speech and statistical learning in younger and older adults. *Frontiers in Human Neuroscience, 8*(628). doi:10.3389/fnhum.2014.00628

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia, 9*, 97-112.

Pallier, C., Sebastian-Gallés, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition, 26*(4), 844-851. doi:10.3758/bf03211403

Peelle, J. E., & Wingfield, A. (2005). Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. *J Exp Psychol Hum Percept Perform, 31*(6), 1315-1330. doi:10.1037/0096-1523.31.6.1315

Plomp, R., & Mimpen, A. M. (1979). Speech-reception threshold for sentences as a function of age and noise level. *The Journal of the Acoustical Society of America, 66*(5), 1333-1342. doi:10.1121/1.383554

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *1*(6), 906-914. doi:10.1002/wcs.78

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., . . . Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience, 7*(31). doi:10.3389/fnsys.2013.00031

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *Int J Audiol, 47 Suppl 2*, S99-105. doi:10.1080/14992020802301167

Saffran, J. R. (2003). Statistical Language Learning:Mechanisms and Constraints. *12*(4), 110-114. doi:10.1111/1467-8721.01243

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Atten Percept Psychophys, 71*(6), 1207-1218. doi:10.3758/app.71.6.1207

Sebastián-Gallés, N., Dupoux, E., Costa, A., & Mehler, J. (2000). Adaptation to time-compressed speech: Phonological determinants. *Perception & Psychophysics, 62*(4), 834-842. doi:10.3758/bf03206926

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304.

Song, J. H., Skoe, E., Banai, K., & Kraus, N. (2012). Training to Improve Hearing Speech in Noise: Biological Mechanisms. *Cerebral Cortex, 22*(5), 1180-1190. doi:10.1093/cercor/bhr196

Studebaker, G. A. (1985). A "Rationalized" Arcsine Transform. *Journal of Speech, Language, and Hearing Research, 28*(3), 455-462. doi:doi:10.1044/jshr.2803.455

Tao, L., & Taft, M. (2017). Influences of Cognitive Processing Capacities on Speech Perception in Young Adults. *Frontiers in Psychology, 8*(266). doi:10.3389/fpsyg.2017.00266

Tombaugh, T. N. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology, 19*(2), 203-214. doi:https://doi.org/10.1016/S0887-6177(03)00039-8

Tun, P. A. (1998). Fast noisy speech: Age differences in processing rapid speech with background noise. *Psychol Aging, 13*(3), 424-434. doi:10.1037/0882-7974.13.3.424

Tun, P. A., & Wingfield, A. (1999). One voice too many: adult age differences in language processing with different types of distracting sounds. *J Gerontol B Psychol Sci Soc Sci, 54*(5), P317-327.

Voor, J. B., & Miller, J. M. (1965). The effect of practice upon the comprehension of time-compressed speech. *Communications Monographs, 32*(4), 452-454.

Watson, C. S. (1980). Time Course of Auditory Perceptual Learning. *Annals of Otology, Rhinology & Laryngology, 89*(5_suppl), 96-102. doi:10.1177/00034894800890s525

Wong, P. C., Jin, J. X., Gunasekera, G. M., Abel, R., Lee, E. R., & Dhar, S. (2009). Aging and cortical mechanisms of speech perception in noise. *Neuropsychologia, 47*(3), 693-703. doi:10.1016/j.neuropsychologia.2008.11.032

Zaballos, M. T., Plasencia, D. P., Gonzalez, M. L., de Miguel, A. R., & Macias, A. R. (2016). Air traffic controllers' long-term speech-in-noise training effects: A control group study. *Noise Health, 18*(85), 376-381. doi:10.4103/1463-1741.195804

Zekveld, A. A., Rudner, M., Johnsrude, I. S., & Rönnberg, J. (2013). The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. *134*(3), 2225-2234. doi:10.1121/1.4817926