8-6-2019 2:30 PM

Characterizing the Familiar-Voice Benefit to Intelligibility

Beatriz Ysabel Domingo, *The University of Western Ontario*

Supervisor: Johnsrude, Ingrid, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Psychology

Abstract

Everyday listening often occurs in the presence of background noise. Listeners with normal hearing can often successfully segregate competing sounds from the signal of interest. To do this, listeners exploit a variety of cues to facilitate the separation of simultaneous sounds into separate sources, and group sequential sounds into intelligible speech streams. One of the cues that has been shown to be an effective facilitator of speech intelligibility is familiarity with a talker's voice. A recent study by Johnsrude *et al*. (2013) measured speech intelligibility of a naturally familiar voice (i.e., that of a long-term spouse) and showed a large improvement in intelligibility when a spouse's voice serves as the target or the masker. This improvement is commensurate with another cue that is well-understood to be a strong facilitator of intelligibility: spatially separating two speech streams. Therefore, the goal of this thesis is to extend the work of Johnsrude *et al*. (2013) by providing a clearer understanding of voice familiarity as a cue for improving intelligibility. Specifically, the aims of this thesis are 1) to measure the magnitude of intelligibility benefit of different types of naturally familiar voices: friends and spouses, (2) to quantify the familiar-voice benefit in terms of degrees of spatial separation, and (3) to compare the neural bases of voice familiarity and spatial release from masking to determine if these cues improve intelligibility by recruiting similar areas of the brain. The primary findings of this thesis were that 1) the familiar-voice benefit of friends and spouses are comparable to each other and that relationship duration does not affect the magnitude of the familiar-voice benefit, (2) that participants gain a similar benefit from a familiar target as when an unfamiliar voice is separated from two symmetrical maskers by approximately 15° azimuth, and (3) that familiar voices and spatial release from masking both activate known temporal voice areas, but attending to an unfamiliar target voice when masked by a familiar voice also recruits attention areas. Taken together, this thesis illustrates the effectiveness of a naturally familiar target voice in improving intelligibility.

Keywords

Speech intelligibility, speech perception, voice familiarity, voice perception, spatial cues, spatial release from masking, selective attention, fMRI, sparse imaging

Lay Summary

Communication typically occurs in noisy environments, where there are competing background sounds such as music, other conversations, and traffic noises. For individuals with normal hearing, it is relatively easy to ignore these background sounds and focus on one person or conversation. However, the processes that make this possible are complex and not completely understood. In this thesis, I aim to gain a deeper understanding of how people understand speech when a competing voice is speaking. Specifically, I want to understand why it is easier to comprehend speech of familiar people compared to speech of strangers. I compared how much intelligibility improved from listening to the voice of a spouse or a friend and found that intelligibility improved by a comparable amount. This means that once a person gains familiarity with a voice, the benefits to intelligibility remain constant over time and does not change depending on the type of relationship. Next, I equated the improvement to intelligibility from a familiar voice in terms of spatial separation and found that the familiar-voice benefit is equal to that of a large spatial separation. This means that familiar voices are highly effective at improving intelligibility. Lastly, I compared the neural mechanisms between speech intelligibility facilitated by voice familiarity and spatial separation and found that brain areas responsible for processing both cues at least partially overlap. Overall, the findings of this thesis highlight the effectiveness of voice familiarity in improving intelligibility and provide preliminary evidence of brain areas responsible for processing intelligibility cues.

Co-Authorship Statement

Chapter 2 was designed and written in collaboration with Dr. Ingrid Johnsrude and Dr. Emma Holmes. This chapter has been submitted to the Journal of Experimental Psychology: Applied and is currently in revision.

Chapters 3 and 4 were designed and written in collaboration with Dr. Ingrid Johnsrude, Dr. Ewan Macpherson, and Dr. Emma Holmes. Chapter 3 is in preparation to be submitted to the Journal of the Acoustical Society of America.

Acknowledgments

This thesis would not have been possible without the help and support of mentors, family, and friends.

First, I am most grateful to Ingrid Johnsrude. Thank you for your endless patience and dedication to helping me grow as a scientist and researcher. It has been a privilege to learn from you and be in your lab for the past five years.

I would also like to thank Emma Holmes, who over the years has become not just a mentor but also a good friend. Thank you for teaching me MATLAB, for helping me analyze and interpret my results, and for being my conference roommate for three years. My PhD would not have been the same without you.

To my advisory committee, Scott MacDougall-Shackleton and Ewan Macpherson, your knowledge and feedback has made my research better.

Thank you to my previous supervisors, Allyson Page and Elizabeth K. Johnson. My experiences in your labs have brought me to where I am today. Thank you for taking a chance on me.

Members of the CoNCH Lab, both past and present, have provided valuable feedback on all posters and presentations in my five years here. Thank you for helping me to become a better researcher. I also want to thank all the friends I have made at Western for sharing in my successes and letting me vent my frustrations. Your friendship and company have made the BMI feel like home, and I am lucky to have you all in my life.

To my family, thank you for your patience and encouragement as I stayed in school until essentially Grade 23! I hope I have made you proud.

Lastly, to Scott. You are my biggest motivator and supporter, and you've helped me grow in so many ways. Thank you for your love.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AAL	Automated Anatomical Labeling
ANOVA	Analysis of Variance
ASA	Auditory Scene Analysis
BA	Brodmann Area
BOLD	Blood-Oxygen-Level Dependent
BU	Both Unfamiliar
BUG	Boston University Gerald
CRM	Coordinate Response Measure
dB	Decibels
EEG	Electroencephalography
EPI	Echo-Planar Imaging
FWE	Family-Wise Error
FFA	Fusiform Face Area
FM	Familiar Masker
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near-Infrared Spectroscopy
FT	Familiar Target
FVA	Frontal Voice Area
GLM	General Linear Model
HL	Hearing Level
HRTF	Head-Related Transfer Function
Hz	Hertz
IC	Inferior Colliculus
IFG	Inferior Frontal Gyrus
ILD	Interaural Level Difference
IPL	Inferior Parietal Lobule
IQR	Interquartile Range
ISSS	Interleaved Silent Steady State
ITD	Interaural Time Difference
KEMAR	Knowles Electronic Mannequin for Acoustics Research
LSO	Lateral Superior Olive
MANOVA	Multivariate Analysis of Variance
MNI	Montreal Neurological Institute (stereotaxic space)
MRI	Magnetic Resonance Imaging
MSO	Medial Superior Olive
MTG	Mid Temporal Gyrus
PET	Positron-Emission Tomography
RAU	Rationalized Arcsin Units
RMS	Root Mean Square
RSA	Representational Similarity Analysis

SCN	Signal-Correlated Noise
SD	Standard Deviation
SE	Standard Error of the Mean
SNR	Speech-to-Noise Ratio
SPL	Sound Pressure Level
SPM	Statistical Parametric Mapping
SRM	Spatial Release from Masking
SRT	Speech Reception Threshold
STG	Superior Temporal Gyrus
STS	Superior Temporal Sulcus
T	Tesla
TE	Time to Echo
TMR	Target-to-Masker Ratio
TR	Repetition Time
TVA	Temporal Voice Area

List of Appendices

Chapter 1

1 General Introduction

Communication in the presence of competing sounds occurs every day with relative ease, yet how listeners accomplish this is not fully understood. It is important to understand how humans with normal hearing are capable of segregating simultaneous sounds to focus on a specific signal of interest. This thesis aims to investigate the role of two cues that have been shown to improve speech intelligibility in noisy environments: voice familiarity and spatial separations between simultaneous speech streams. Specifically, this thesis will characterize and quantify the benefit of a familiar voice (such as that of a friend or spouse) on intelligibility and compare this benefit with that obtained from spatial cues. Lastly, this thesis will identify and compare the neural substrates of familiar-voice processing and spatial release from masking to determine if these two cues activate similar brain areas to facilitate intelligibility.

1.1 Cocktail party listening

In most natural environments, sounds from various sources occur simultaneously. To "hear out" a sound of interest in the presence of competing sounds has been termed the 'cocktail party problem' (Cherry, 1953). Human listeners with normal hearing can typically communicate successfully in noisy environments. Despite the relative ease with which listeners can communicate in these environments, the processes underlying this ability are complex and not completely understood.

Cocktail party listening involves different challenges. First, a listener must segregate simultaneous sounds into separate sources, a process called 'simultaneous grouping'. Next, a listener must segregate sequential sounds over time into separate streams and selectively attend to a specific sound source in the presence of competing sounds. This process is called 'sequential grouping' or 'streaming'. These two processes are discussed in more detail in the next section.

It remains unclear whether segregation happens first to allow attention to a particular speech stream, or if attention to a stream allows it to be segregated. Cusack, Deeks, Aikman, and Carlyon (2004) proposed a hierarchical model explaining the relationship between selective attention and stream formation. In this model, researchers suggest that streaming only occurs on the sound source being attended to, and not on all sounds occurring that the same time. After a stream is attended to, it can be further segregated into small fragments, but this further segregation does not occur in the unattended streams.

There are two types of mechanisms that allow for auditory grouping (Darwin & Carlyon, 1995). The first is *primitive grouping mechanisms*, that involve the use of low-level sound properties, such as harmonicity and onset asynchrony, for segregation. Primitive cues are not primarily relied on when analyzing complex sounds such as musical chords or speech. The second mechanism is *schema-governed mechanisms* that require learned or experience-based information to segregate sounds. An example of this is segregating speech from non-speech sounds.

The cocktail party problem has remained a topic of substantial research since Cherry's (1953) seminal paper. Advances in basic research on this topic has contributed to understanding hearing and speech communication processes and has helped uncover physiological mechanisms and beneficial cues that can be used to facilitated intelligibility of a target in the presence of a competing talker or talkers. A subset of these cues will be discussed in this chapter.

1.2 Auditory scene analysis

The term auditory scene analysis (ASA) refers to the process by which listeners perceptually organize overlapping sounds into distinct sound objects such as speech, music, and environmental sounds into mental representations known as auditory streams (Bregman, 1990). Auditory scene analysis involves two processes: simultaneous grouping and sequential grouping.

Simultaneous grouping is the process of identifying simultaneous sounds as coming from distinct sources. To do this, the auditory system takes advantage of periodic sounds, whose component frequencies are multiples of the fundamental frequency (F_0). When a subset of incoming sound is composed of frequencies that are multiples of the F_0, known as harmonics, those sounds are likely to be grouped together to form an integrated percept of the sound (Darwin & Carlyon, 1995). In addition to harmonicity, another cue that has been shown to indicate that sounds come from the same source is onset and offset synchrony, referring to the tendency for sounds that start and end at the same time to be grouped together (Bee & Micheyl, 2008). Lastly, when the amplitude envelope modulations are correlated across the frequency spectrum, referred to as common amplitude modulation, signal detection in noise is improved.

Sequential grouping refers to the processes by which sound objects that are extended in time, called streams (like voices) are perceptually organized – grouped and segregated – over time. This process is also known as streaming. Auditory objects that are perceptually similar to one another are grouped together as one sound. Perceptual grouping is facilitated by sound features of pure and complex tones. Some examples include acoustic information from the temporal fine structure and envelope, (Fogerty & Humes, 2012; Moon & Sung, 2014), rate of frequency change (Darwin, 1997), F_0 differences between competing streams (Deroche, Culling, Chatterjee, & Limb, 2014), spatial cues (Kidd, Mason, Best, & Marrone, 2010; Marrone, Mason, & Kidd, 2008; Noble & Perrett, 2002; Yost, 2017) and intensity differences (Oxenham, Boucher, & Kreft, 2017). Sequential grouping is useful when processing signals that unfold or evolve over time, such as music or speech.

Features that play a role in ASA can also be classified as bottom-up or top-down. Bottom-up cues refer to the physical properties of a signal. Top-down cues, on the other hand, rely on a listener's knowledge or experience to facilitate sound segregation. Examples of knowledge-based cues are experience with a particular accent, melody, or voice.

Masking refers to the occlusion of auditory objects such that they are not clearly perceived. When competing sounds occur at the same time and at similar frequencies as a target signal and therefore compete with the target sound for cochlear processing, it is known as energetic masking (Brungart, 2001; Carlile, 2014; Scott & McGettigan, 2013). Energetic masking is often produced in situations when a distractor sound has high spectral overlap with the target sound, or is sufficiently loud that the target sound becomes difficult to perceive (Carlile, 2014). Because spectral overlap with the target sound is often the cause of energetic masking, the most effective energetic maskers are wide-spectrum noise or multi-talker babble (Scott & McGettigan, 2013).

Energetic masking occurs at the auditory periphery, where signal and competing sounds create activity at the same areas of the basilar membrane. When there is no spectral or temporal overlap of competing sounds but the target signal is still not clearly perceived, it is referred to as informational masking. Oftentimes, informational maskers make it difficult for a listener to maintain attention to the target stream or object (Carlile, 2014). One example of an informational masker is speech; it is thought that the linguistic information in masking speech competes for processing resources with a speech target. Informational masking is not stimulus-driven and occurs when there is competition between target and distractor sounds at levels higher than the cochlea, creating uncertainty about the target (Scott & McGettigan, 2013). Because there is no interference at the basilar membrane, informational masking is presumed to occur in the central auditory pathway (Kidd & Colburn, 2017). Consistent with this framework, release from energetic masking can be achieved when the target or maskers are physically altered in some way – for example, by moving them further apart. Release from informational masking is thought to occur through higher-level processes such as grouping (Bregman, 1990).

1.3 Speech intelligibility

The term ‘speech intelligibility’ refers to the extent that a target speech signal is understood or correctly identified by a listener. It is not only clear speech, meaning speech produced in the absence of background noise or degradation, that is considered intelligible. Degraded speech, such as noise-vocoded cochlear implant simulations and

bandpass-filtered speech, is often intelligible to a listener because the speech signal contains many redundancies (Plack, 2014), and because listeners can often use context to infer sections of the speech signal that are occluded by competing sounds. Research has shown that when presenting listeners with bandpass-filtered sentences with frequency bands that were 1/3 octave wide, intelligibility reached near-ceiling (over 90%) in conditions with a center frequency of 1100-2100 Hz (Warren, Riener, Bashford, & Brubaker, 1995). Experimenters then presented listeners with bandpass-filtered speech that was only 1/20 octave wide and centered at 1500 Hz, and found that intelligibility was near 80%. Similarly, listeners can identify words with at least 50% accuracy from 3-band or 4-band sine- or noise-vocoded sentences (Dorman, Loizou, & Rainey, 1997; Souza & Rosen, 2009). These experiments suggest that listeners need only a fraction of a full speech signal to report words from it.

How the brain may derive meaning from spoken sentences has been outlined as involving a series of steps that are each influenced by top-down information (Davis & Johnsrude, 2007): (1) grouping of auditory information into a single stream, (2) segmenting speech into meaningful units, and (3) perceptual learning mechanisms to make sense of degraded speech. In the first step of this process, top-down schema-based mechanisms drive perceptual organization of sound into a single stream, often overriding low-level grouping cues. The second step, separating a continuous speech stream into discrete units (words or morphemes), occurs through higher-order processes like word recognition. Similar to the first step, when bottom-up cues are insufficient or unreliable due to signal degradation, these are overridden by top-down information. The third step, perceptual learning to make sense of degraded speech, is supported by top-down processes in which listeners exploit linguistic cues to make sense of acoustically degraded words. When processing a degraded speech signal, a listener can rely on linguistic knowledge and lexical information to help predict upcoming words in a sentence. Furthermore, top-down information is responsible for maintaining categorical perception in the face of the high variability that occurs in speech input both within and between speakers.

The ability of the auditory system to exploit a limited subset of cues to identify words from speech is crucial to communicate successfully in noisy environments or in situations

where listeners hear a degraded speech signal (e.g., over the phone, through a hearing device, etc.). In these situations, the auditory system takes advantage of a variety of features to achieve a release from masking.

Top-down cues have also been shown to improve speech intelligibility by leveraging the listener's knowledge and prior experience to segregate speech streams. Some examples are knowledge of the language of the target (Cooke, Garcia Lecumberri, & Barker, 2008), trained familiarity with a voice (Kreitewolf, Mathias, & von Kriegstein, 2017; Tye-Murray, Spehar, Sommers, & Barcroft, 2016), and natural familiarity with a voice (Johnsrude *et al*., 2013; Newman & Evers, 2007; Souza, Gehani, Wright, & McCloy, 2013).

Johnsrude *et al*. (2013) showed that a familiar voice is more intelligible than an unfamiliar voice when masked by another unfamiliar voice, and that a familiar masker voice can improve the intelligibility of an unfamiliar target voice. In contrast, Newman and Evers (2007) observed an intelligibility benefit from a familiar voice (e.g., the voice of a university professor) when it was the target but not when it was the masker. The differences in findings between these two studies could be due to the different relationship between familiar voice and listener or due to task differences. Nevertheless, an interesting question that stems from the findings of Johnsrude *et al*. (2013) is if the familiar-target and familiar-masker intelligibility benefit could be replicated using more challenging task and different types of naturally familiar voices (e.g., that of friends, roommates, or romantic partners).

1.3.1 Speech intelligibility tasks

Speech intelligibility can be measured using a variety of tasks. Tasks in which participants identify words that they heard from a target sentence are called open-set tasks. Examples of open-set tasks are speech shadowing (Newman & Evers, 2007), verbal target reporting (Freyman, Helfer, McCall, & Clifton, 1999; Huyck & Johnsrude, 2012; Zekveld, Rudner, Johnsrude, Heslenfeld, & Rönnberg, 2012), and word or sentence transcription (Assmann, 1999; Hawley, Litovsky, & Colburn, 1999; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). The main limitation of open-set tasks is that

they are susceptible to response bias in that participants may be more likely to report words in conditions where they feel more confident in their response (e.g., higher SNR, trained/familiar voice, etc). This response pattern may lead to inflated intelligibility scores in these conditions simply because participants are more willing to guess. Therefore, results that use these tasks may be contaminated by responses bias.

In contrast, closed-set tasks are not susceptible to this same response bias. In closed-set tasks, listeners are required to provide a fixed number of responses from a given set of words. A widely used closed-set task is the Coordinate-Response Measure (CRM) (Bolia, Nelson, Ericson, & Simpson, 2000) procedure, in which sentences follow the pattern "Ready <call sign>, go to <colour><number> now." Listeners are instructed to select the colour and number in the sentence that began with the target call sign, which was provided at the start of the experiment. The CRM has been used in speech-on-speech intelligibility tasks (Ericson, Brungart, & Simpson, 2004; Johnsrude *et al*., 2013; Samson & Johnsrude, 2016), as well as speech-in-noise tasks (Brungart, 2001). However, because the response in a CRM tasks is only two words (one colour and one number), it is difficult to determine if a participant is truly streaming the target sentence from the masker or if a participant is simply remembering the colour and number from the target and maskers, and selecting the words that matched the voice that said the target name.

One way to distinguish between these two possibilities is by using a closed-set task in which participants have to recognize more words than the CRM from a greater number of options. An example of a more challenging procedure that fulfills these requirements is the Boston University Gerald (BUG) closed-set task (Kidd, Best, & Mason, 2008), which uses sentences that follow the pattern, "<Name> <verb> <number> <verb> <noun>". The target is identified by the Name word, and participants are required to identify the remaining four words in the sentence out of a possible eight options each.

Table 1. The Boston University Gerald task. Sentences were constructed using one word from each column.

Name	Verb	Number	Adjective	Noun
Bob	bought	two	big	bags
Pat	found	three	blue	cards
	gave	four	cold	gloves
	held	five	hot	hats
	lost	six	old	pens
	saw	eight	new	shoes
	sold	nine	red	socks
	took	ten	small	toys

Using the BUG task, it likely becomes too difficult to hold all words from the target and masker sentences in memory and remember the voices that spoke each Name word to report the target words at the end of the trial. In other words, a task like the BUG has a higher memory load, and so may provide a closer approximation of naturalistic listening than does the CRM task.

1.4 Functional magnetic resonance imaging (fMRI)

fMRI is an imaging technique that is widely used to investigate human brain function and organization, and how these relate to neuroanatomy. fMRI involves the excitation of hydrogen nuclei by a radiofrequency pulse, and a magnetic field gradient to localize the excitation. After this period of excitation, nuclei return to their original state following a time-decay of T1 (for magnetization in the same longitudinal direction as the magnetic field) and T2 (for magnetization transverse to the magnetic field). Different tissues can be seen clearly depending on whether the acquired image is weighted to show contrast based on T1 or T2 signal. For T1-weighted images, tissues that contain fat appear brighter, but for T2-weighted images, tissues that contain water appear brighter. If inhomogeneities are present in the magnetic field, nuclei undergo relaxation following a time-decay of T2* (or ‘effective T2’).

Neural activity is indirectly measured using the blood-oxygen-level-dependent (BOLD) fMRI signal. The BOLD contrast is based on changes in the relative concentrations of oxygenated and deoxygenated blood (Logothetis & Wandell, 2004). Oxygen supply is coupled through a complex process to neural activity, therefore BOLD reflects neural responses to a stimulus (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2002). The vascular system overcompensates for the increased oxygen demand by increasing the amount of oxygenated hemoglobin compared to deoxygenated hemoglobin in a local region. Areas with high levels of oxyhemoglobin produce a higher signal than areas with low oxyhemoglobin (Amaro & Barker, 2006).

Traditional fMRI analyses determine which brain regions are involved in processing various perceptual stimuli or experimental tasks by examining the relationship between cognitive functions and individual voxel activity (Norman, Polyn, Detre, & Haxby, 2006). This type of analysis considers differences in activity of many voxels, but each voxel is considered individually. This approach is an extension of the general linear model and accounts for physiological noise as well as correlations that arise due to temporal smoothing. Common methods of comparing activity between different conditions include subtraction, where there is an assumption that two conditions can be cognitively added and that there are no interaction between the two conditions; or parametric, in which it is assumed that the load on a particular cognitive function (such as working memory) can be increased or decreased without modifying the nature of the function itself; and conjunction, where commonalities between different conditions can be used to identify brain regions involved in a particular cognitive process (Amaro & Barker, 2006).

For auditory research, it is common to use a sparse-sampling fMRI design, in which there are silent periods between volume acquisitions. During the silent period, auditory stimuli is presented to the participant without scanner noise (Hall *et al*., 1999). However, sparse imaging only provides a single measure of the hemodynamic response for each trial. If the volume acquisition time is incorrectly estimated, the peak hemodynamic response will not be captured. A method that overcomes this limitation is interleaved silent steady state (ISSS) imaging, which maintains longitudinal magnetization during volume

acquisition by applying excitation pulses during the silent period, thus avoiding T1-signal decay (Schwarzbauer, Davis, Rodd, & Johnsrude, 2005). ISSS imaging allows for the acquisition of multiple volumes, thus increasing the likelihood of capturing the peak hemodynamic response.

1.4.1 What are the advantages of fMRI?

fMRI is an invaluable tool in cognitive neuroscience research because it overcomes many limitations of other methods. First, fMRI allows researchers to measure neural activity with high spatial resolution, which increases with magnetic field strength (Logothetis, 2008). However, because fMRI measures BOLD signal change, which is quite slow, its temporal resolution is lower than that of other methods such as EEG. For auditory stimuli, BOLD signal peaks about 4-5 seconds after stimulus onset, and returns to baseline about 9-12 seconds after stimulus onset (Hall *et al*., 2000). Second, fMRI poses very little risk to participants (in comparison to PET or X-ray, for example) as participants do not need to ingest contrast agents and does not use radiation. Therefore, the same brains can be studied over several experiments and sessions without harming the participant (Logothetis *et al*., 2002), and fMRI can be used to track changes in brain anatomy and function over time. Third, MRI allows researchers to take both anatomical and functional images of the brain, whereas alternative methods such as EEG can only measure functional activity. Lastly, fMRI allows the researcher to observe activity throughout the entire brain compared to EEG or functional near-infrared spectroscopy (fNIRS), which are most sensitive to activity on the cortical surface.

1.5 How is sound represented in the brain?

Rauschecker and Tian (2000) proposed a framework for higher order auditory pathways in rhesus monkeys that share commonalities with the visual spatial and non-spatial processing models. In this model, the anteroventral “what” stream originates in the anterior lateral temporal lobe and projects to the orbitofrontal cortex and is critical in processing species-specific vocalizations. In contrast, the posterodorsal “where” stream is involved in processing spatial information and originates in the caudal lateral temporal lobe (a similar area to the posterior STG in humans) and projects to the posterior parietal

and dorsolateral prefrontal cortex. More recently, this framework was tested in humans to determine the extent of dissociation between areas involved in sound localization ("where") and sound identification ("what") (Zündorf, Lewald, & Karnath, 2016). Consistent with previous findings of Rauschecker & Tian (2000), researchers found that posterior STG, left and right IPL, posterior parietal cortex, and the superior frontal sulcus are involved in spatial tasks (Arnott, Binns, Grady, & Alain, 2004; Barrett & Hall, 2006; Mathiak *et al*., 2007; Shiell, Hausfeld, & Formisano, 2018; J. D. Warren & Griffiths, 2003). In contrast, the anterior temporal cortex, IFG, dorsolateral prefrontal cortex, and intraparietal sulcus are involved in sound identification (Bethmann, Scheich, & Brechmann, 2012; Maeder *et al*., 2001; Mathiak *et al*., 2007; Relander & Rämä, 2009; Stevens, 2004). Zundorf et al (2016) also identified an area critical to both sound localization and identification: the posterior superior IFG (BA 44).

A meta-analysis of human fMRI research (Arnott *et al*., 2004) supported and extended the initial results of Rauschecker & Tian (2000). Human IPL activity was present in tasks that involve evaluating the location of a sound source (Alain, Arnott, Hevenor, Graham, & Grady, 2001; Griffiths *et al*., 1998; Maeder *et al*., 2001). The superior frontal sulcus and posterior areas of the temporal cortex were also found to be involved in spatial tasks involving sound, but were not reported in every auditory spatial study included in the meta-analysis. In contrast, the anterior temporal lobe and IFG had little involvement with auditory spatial processing, but are active during nonspatial tasks. These findings differ from those of Zundorf *et al*. (2016), who found that IFG is involved in both sound localization and identification, and of other studies that found that the IFG is involved in sound localization (Lewald & Getzmann, 2011; Lewald, Riederer, Lentz, & Meister, 2008) and sound motion perception (Hart, Palmer, & Hall, 2004), indicating that the IFG has some involvement in spatial processing.

Belin and Zatorre (2000) proposed a modified version of the auditory dual-pathway model that is analogous to one proposed for the visual domain (Goodale & Milner, 1992; Mishkin, Ungerleider, & Macko, 1983). In this model, the ventral stream is involved in sound recognition and identification, but the dorsal pathway is involved in analyzing spectral dynamics (analogous to visual spatial motion) to perceive the evolution of sound

over time. This dorsal pathway has been named the ‘how’ stream, and is responsible for processing the verbal content of speech and the melody of music. While the authors did not account for a distinct ‘where’ pathway, areas involved in the ‘how’ pathway may also be responsible for processing auditory spatial information.

Spatial information, specifically ITDs and ILDs, are initially processed in the brainstem. ITD information is carried up the afferent pathway from the auditory nerve fiber and is passed onto the medial superior olive (MSO) (Tollin & Yin, 2009). The MSO contains neurons that identify the time of sound occurrence at each ear. ILD information is also processed in the afferent auditory pathway, primarily by the lateral superior olive (LSO), which receives excitatory input from the ipsilateral anteroventral cochlear nucleus and inhibitory input from the contralateral medial nucleus of the trapezoid body (Tollin & Yin, 2009). The LSO computes differences in ipsilateral and contralateral input and produces action potentials whose firing rate is directly proportional to the sound level. From the LSO, ILD information is passed on to the inferior colliculus (IC) which receives excitatory input from the contralateral ear and inhibitory input from the ipsilateral ear (Brainard, 1994).

1.5.1 Neural correlates of speech processing

Speech processing areas have been found to partially overlap with both the “what” pathway and the “where” pathway (Boldt *et al*., 2013). Pathways for speech and language processing are similar to, but ultimately distinct from, the broader auditory “what” and “where” model. The dual-stream model of speech processing (Hickok & Poeppel, 2000; Hickok & Poeppel, 2007) argues that the ventral stream, involving middle and superior portions of the temporal lobe, processes speech signals for recognition and comprehension. The ventral stream represents different aspects of the speech signal, such as phonemes, syllabic structure, word forms, as well as syntactical and semantic information. In contrast, the dorsal stream, bounded by the posterior frontal lobe and posterior dorsal-most aspect of the temporal lobe and parietal operculum, is responsible for integrating auditory and motor information involved in speech perception and production. The dorsal stream integrates auditory-motor information across two levels: speech segments, and sequences of speech segments. Scott and Johnsrude (2003) provide

support for this account by reporting that the anterior system (involving auditory belt, parabelt, anterior STS, and ventro- and dorso-lateral frontal cortex) may play a role in mapping acoustic-phonetic cues and using those cues in accessing relevant lexical information, whereas the posterior system (involving posterior auditory belt, parabelt, posterior STS, parietal cortex, and ventro- and dorso-lateral frontal cortex) may form articulatory-gestural representations of motor speech.

A newer model by Rauschecker and Scott (2009), based on research on nonhuman primates, builds on the dual-stream model of Hickok and Poeppel (2000, 2007) but extends beyond speech processing. In their model, the antero-ventral stream is responsible for sound identification, perceptual invariance, and speech and voice perception. The postero-dorsal stream is also involved in speech and music perception, including processing of spatial information. Further, this model accounts for articulation and speech production processes differently than Hickok and Poeppel (2000, 2007). During forward-mapping, sound information is thought to be decoded in the antero-ventral stream (i.e., anterior temporal lobe and IFG), and is passed to the premotor cortex where articulatory representations are formed. When information is passed in the reverse direction, called inverse mapping, the IPL is thought to create predictive motor signals that affect articulatory representations in the prefrontal cortex and premotor cortex.

Scott, Blank, Rosen, and Wise (2000) used PET to characterize a neural pathway for intelligible speech involving the left temporal lobe. In this experiment, researchers presented listeners with four types of stimuli: (1) natural unprocessed speech, (2) intelligible noise-vocoded speech, (3) rotated speech which contains similar spectral and temporal properties of natural speech but is unintelligible unless the listener has undergone weeks of extensive training, and (4) rotated noise-vocoded speech which is not speech-like and is completely unintelligible regardless of training. Participants were asked to estimate how much of the sentence they had heard. Results were that areas of the left STG which are lateral and anterior to the primary auditory cortex, and posterior STS were activated by the presence of phonetic cues present in speech, noise-vocoded speech, and rotated speech. However, the anterior left STS was only activated by intelligible signals (e.g., speech and noise-vocoded). This experiment was later followed-up using

fMRI (Narain *et al*., 2003) and intelligible speech was lateralized to the posterior left STS.

Intelligibility of normal speech, noise-segmented speech, noise-vocoded speech, and speech in noise were compared and BOLD response was positively correlated with activation in the left superior and middle temporal gyri (Davis & Johnsrude, 2003; Wild, Davis, & Johnsrude, 2012; Wild, Yusuf, *et al*., 2012). Homologous areas in the right temporal lobe also showed activation to intelligible speech, but to a lesser extent. Further, intelligible speech also correlated with activity in the left hippocampal complex and left IFG.

Instead of using noise-vocoded or spectrally rotated speech, Zekveld, Heslenfeld, Festen, and Schoonhoven (2006) manipulated intelligibility by presenting speech at various signal to noise ratios (SNRs) to identify regions where intelligible and unintelligible speech are processed. Researchers found that bilateral anterior and posterior temporal brain regions and Broca's area in the left IFG are significantly more activated when listening to intelligible speech compared to unintelligible speech. In contrast, the pars opercularis in the left IFG is activated more when listening to unintelligible speech compared to intelligible speech.

A review characterizing neural correlates of masked speech (Scott & McGettigan, 2013) concluded that areas that are recruited in processing masked speech vary according to the masker type and task. Activation was strongest in the left and right dorsolateral temporal lobes for speech-in-speech stimuli compared to speech-in-noise. The opposite contrast (speech-in-noise > speech-in-speech) revealed activation in posterior parietal cortex and left dorsal prefrontal cortex.

One aim of this thesis is to extend this line of research by identifying regions of the brain that are involved in processing intelligible speech using another cue shown to improve intelligibility: voice familiarity.

1.6 Voice familiarity vs. talker normalization

Familiarity with a voice involves knowledge of its acoustic properties. Voice recognition has been suggested to rely on a set of acoustic features like fundamental frequency and the first formant (Baumann & Belin, 2010). Manipulating formant spacing has been shown to adversely affect a listener's ability to recognize a familiar voice, but does not appear to affect intelligibility as much (Holmes, Domingo, & Johnsrude, 2018). Familiar-voice recognition may occur through learning acoustic patterns that are formed from averaging multiple utterances of a single speaker to form a speech prototype (Fontaine, Love, & Latinus, 2017). Therefore, if a listener is exposed to a wide variety of utterances in terms of prosody, affect, and linguistic content, the speech prototype developed will be more flexible than one formed from limited input. When a speech prototype is formed, incoming speech is then compared to it to determine if it was produced by a familiar talker.

Recognition of familiar voices is different to the process of talker normalization, in which speech processing areas recalibrate when listening to speech from a new talker to resolve acoustic-phonetic ambiguities (Wong, Nusbaum, & Small, 2004). In this thesis, I am interested in investigating the intelligibility effects and cortical areas associated with *personally familiar* voices.

Neural responses to familiarity with various stimuli has been studied in vision (Gobbini & Haxby, 2006, 2007; Platek & Kemp, 2009), in audition (Gainotti, 2011; Maguinness, Roswandowitz, & von Kriegstein, 2018; Nakamura et al., 2001; Naoi et al., 2012; von Kriegstein & Giraud, 2006; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005), and in person recognition (Biederman et al., 2018; Blank, Wieland, & Von Kriegstein, 2014; Shah et al., 2001). Despite this, the mental representations of familiar stimuli appear to be poorly understood. Gobbini & Haxby (2007) suggest that familiar face recognition involves the spontaneous retrieval of semantic and personal information related to the individual as well as an emotional response evoked from seeing a familiar person's face. In a familiarity comparison including participants' own faces, Platek & Kemp (2009) compared neural responses to different types of familiar faces (e.g., friends, relatives, and own face) and found that faces of relatives activated areas associated with self-

recognition, suggesting that familiarity and recognition of an individual involves self-referent comparisons. While the current thesis does not investigate the mental representations that are involved in listening to a personally familiar voice, it is possible that familiar voice stimuli may involve representations that are somewhat similar to those for familiar faces.

1.7 Voice perception

1.7.1 Voices and person recognition

Person recognition is cognitively challenging task, involving both physical and semantic information about an individual. When recognizing a familiar person, faces and voices are directly linked to one another, but faces and names are not (O'Mahony & Newell, 2012). In this study, participants were given name-, face-, and voice-information of different actors. They were then asked to give explicit familiarity judgements when presented with faces, voices, or names. The results revealed that when presented with congruent information, participants were significantly faster to recognize face-voice pairs than face-name pairs. These results suggest that faces and voices are integrated with one another for person recognition purposes, but that faces and names are not integrated. Findings from this study may account for how people can become accustomed to the faces or voices of people about whom we have no semantic information.

Human fMRI research provides support for O'Mahony and Newell's account (2012) that supports direct connections between face and voice information before the person recognition step (Blank, Anwander, & von Kriegstein, 2011). Researchers show that face- and voice-sensitive regions of the brain (right fusiform face area (FFA) and right STS, respectively) are structurally connected to each other. Specifically, the FFA has stronger connections with the anterior and middle STS compared to the posterior STS.

1.7.2 Voice discrimination and recognition are separate processes

The ability to recognize a voice as being either familiar or unfamiliar, and the ability to discriminate between unfamiliar voices, are controlled by separate mechanisms (D. Van Lancker & Kreiman, 1987). In this study, participants with left-lateralized lesions, right-

lateralized lesions, and healthy controls were compared in their ability to recognize famous voices and discriminate between unfamiliar voices. Healthy controls demonstrated an ability for both voice recognition and voice discrimination, but performance in both tasks were only weakly correlated. Further, participants with right hemisphere lesions showed impaired recognition abilities, and participants with either right or left hemisphere lesions showed impaired discrimination abilities. Importantly, some participants who showed impaired recognition were not necessarily impaired in discrimination, and vice versa. This was reinforced by a case series (Van Lancker, Cummings, Kreiman, & Dobkin, 1988) of six participants with lesions is temporal or parietal regions. Results of this study suggested that voice recognition and discrimination are tentatively mediated by separate anatomical structures. Specifically, voice recognition deficits appear to be associated with right parietal and temporoparietal regions. Voice discrimination deficits appeared to be associated with right or left temporal lobe lesions. These findings lend further support to the idea that recognition and discrimination are unique processes.

1.7.3 Neural substrates of voice processing

Researchers have identified areas in the right and left STS that show greater brain activity when listening to vocal sounds compared to non-vocal sounds (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000). These 'temporal voice areas' (TVAs) were characterized in greater detail by Pernet *et al*. (2015). Brain areas that show strong voice>non-voice activation have been classified into three 'voice patches' in each temporal lobe: the posterior TVA (right middle/posterior STS), the middle TVA (middle STG/STS), and anterior TVA (anterior STS). The right posterior STS was shown to have the strongest peak activation to vocal sounds compared to non-vocal sounds. The location of TVAs greatly overlaps with identified speech-processing regions, particularly in the left STS (Davis & Johnsrude, 2003; Narain *et al*., 2003; Scott *et al*., 2000; Wild, Davis, *et al*., 2012; Wild, Yusuf, *et al*., 2012).

Three bilateral voice-processing areas in the prefrontal cortex, called the Frontal Voice Areas (FVAs) have also been identified (Aglieri, Chaminade, Takerkart, & Belin, 2018) in the left and right IFG (anterior and middle FVAs), left postcentral gyrus (left posterior

FVA), and right precentral gyrus (right posterior FVA). Aglieri *et al*. (2018) showed that the FVAs are functionally connected to the TVAs. Further, frontal connectivity with anterior and posterior FVAs in the right hemisphere was shown to be correlated with behavioural results of voice recognition. In other words, participants who demonstrate good voice recognition performance have higher functional connectivity between anterior and posterior FVAs.

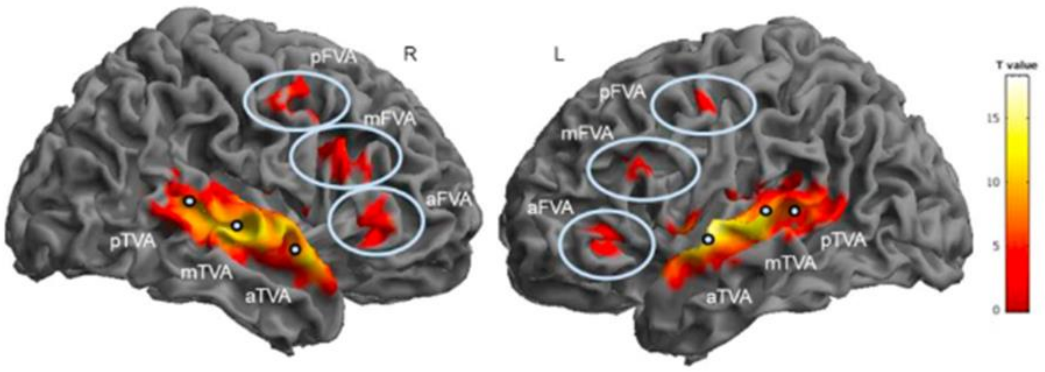


Figure 1. Temporal voice areas (TVAs) and frontal voice areas (FVAs). Figure taken from Aglieri *et al*. (2018).

The IFG has been thought to be part of the extended face perception network (Fox, Iaria, & Barton, 2009), and shows stronger activation to photographs of emotional faces and famous faces, compared to line drawings of faces (Ishai, Schmidt, & Boesiger, 2005). Taken together with the IFG's role in voice processing, it is possible that the IFG is implicated in person recognition processes in general. These findings complement O'Mahony and Newell's (2012) theory that face and voice information are well integrated with one another in person recognition tasks.

Clinical studies have investigated the neural correlates of phonagnosia, or the inability to individuate people on the basis of their voice. In a case study of a 20-year-old female with no known neurological injuries (Herald, Xu, Biederman, Amir, & Shilowich, 2014), significant behavioural deficits in identifying a familiar celebrity from a 7 second voice clip compared to controls were noted. In a sound-imagination task, the subject was able to imagine non-speech sounds, but not able to imagine the voices of familiar individuals.

In the same task, the subject showed decreased precuneus activation and no activation in the ventromedial prefrontal cortex compared to controls.

1.8 Are familiar and unfamiliar voices represented differently?

Studies investigating neural correlates of familiar and unfamiliar voice recognition have produced mixed results. Stronger neural activation in response to familiar compared to unfamiliar voices has been found in the anterior right STS and right temporal poles (Mathiak et al., 2007; Nakamura et al., 2001; von Kriegstein & Giraud, 2004; von Kriegstein et al., 2005), in the anterior temporal lobe and posterior bilateral STS (Bethmann et al., 2012), and in the left MTG (Birkett *et al*., 2007). Further, previous studies have also found decreased activation in response to familiar voices compared to novel voices in the posterior right STG (von Kriegstein & Giraud, 2004; Zäske, Awwad Shiekh Hasan, & Belin, 2017).

The discrepancies may be attributed to differences in type of familiarity. Some of the studies used personally familiar voices (i.e., that of a colleague or friend) (Birkett *et al*., 2007; Nakamura *et al*., 2001; von Kriegstein & Giraud, 2004; von Kriegstein *et al*., 2005). Others familiarized participants with originally novel voices through prior training (Zäske *et al*., 2017), and others used voices of famous people (Bethmann *et al*., 2012). Perhaps these different types of familiarity result in differences in how these voices are encoded and represented and therefore elicit activation in different areas of the temporal lobe.

Furthermore, many of the experiments discussed above also varied in task. Specifically, Zaske *et al.* (2017) trained participants on a specific voice, and then presented participants with either the familiarized voices or novel voices. Participants were instructed to indicate whether the voice was ‘old’ (i.e., familiar) or ‘new’ (i.e,. unfamiliar). Nakamura *et al.* (2001), and Birkett *et al*. (2007) presented listeners with the voice of either a personally familiar voice or an unfamiliar voice. Participants were required to provide a button-press response to indicate whether or not they know the person the voice belonged to. Bethmann *et al.* (2012) also required participants to

indicate familiarity or unfamiliarity with a stimulus voice, but also had participants name the talker (if they were able to) and provide brief descriptions of the talker such as biography or physical attributes.

Although areas involved in familiar voice recognition have been identified, one aspect of familiar-voice processing that remains unknown is how the brain is organized to exploit familiar-voice cues to enhance intelligibility in noisy conditions. The current thesis aims to address this question.

1.9 Spatial release from masking

Spatial release from masking (SRM) is defined as the improvement (decrease) in the speech reception threshold (SRT) when listening to spatially separated target and masker compared to when they are at the same position (collocated). Spatial separations can be created physically by presenting stimuli in free-field or virtually by using head-related transfer functions (HRTFs).

Release from masking can be measured and in a variety of different ways. Many studies have used an adaptive method (Levitt, 1971) to adjust target-to-masker ratio (TMR) in order to obtain a predefined performance threshold (Bronkhorst & Plomp, 1992; Marrone *et al*., 2008). However, the TMRs between participants could differ greatly, and participants may be able to predict the TMR of the next trial. An alternative method is to test participants under the same conditions, and comparing the improvement in intelligibility scores. After obtaining the improvement in intelligibility, performance can then be equated in terms of other measures, such as TMR in dB (Johnsrude *et al*., 2013; Yost, 2017) or spatial separations.

1.9.1 Sound localization cues

When listening to sounds that originate along the horizontal plane, we take advantage of binaural cues such as interaural time differences (ITDs) and interaural level differences (ILDs). ITDs refer to the differences in time of sound arrival in each ear. These differences in time of arrival enable us to detect the direction a sound is coming from. For example, if a sound is coming from the left side of a listener, the sound will arrive at the

left ear before the right ear. Conversely, if a sound is coming directly in front or behind a listener, then the sound will arrive at both ears at the same time, and the ITD in this case is equal to zero. The range of ITDs for a given listener is determined by the diameter of the listener's head. A listener with a larger head will have a larger range of ITDs because the sound has a greater distance to travel to reach the farther ear, making this cue more useful. Listeners weigh ITDs most heavily when presented with low-frequency sounds: ITD is the most important cue in localizing low-frequency sounds (Wightman & Kistler, 1992). ITD cues are most effective at localizing low-frequency sounds because there is less phase ambiguity (Plack, 2014).

ILDs refer to the differences in amplitude of a sound in each ear. Like ITDs, ILDs also depend on the sound's frequency and location. A high-frequency sound will be subject to acoustic shadowing, since high frequency sounds bend less than low frequency sounds. Whereas low frequency sounds from the side bend around the head so that they are nearly as intense at the far ear, high frequency sounds do not, and so have a lower amplitude at the far ear. Typically, sounds of a frequency equal to or less than the diameter of the head will experience more shadowing, making ILDs a useful cue in localizing high-frequency sounds.

To locate sounds in the vertical plane, we can take advantage of spectral shape cues that result from how sound is reflected in the folds of our pinnae. These reflections vary based on location and frequency of the sound.

Spatial separations can be created by playing sounds from different sources in free field or by virtually spatially separating them using HRTFs. HRTFs are a set of measurements that estimate the acoustic filtering of a free field sound by the head, torso, and pinna (Cheng & Wakefield, 1999). Because sound is filtered according to the size and shape of a person’s head, torso, and pinna, HRTFs are unique to each person. However, many studies use HRTFs measured from a Knowles Electronics Mannequin for Acoustic Research (KEMAR) head (Gardner & Martin, 1995; Zhang, Zhang, Kennedy, & Abhayapala, 2009) at various azimuths (e.g., Best, Gallun, Ihlefeld, & Shinn-Cunningham, 2006; Best, Mason, Swaminathan, Roverud, & Kidd, 2017; Bolia, Nelson,

& Morley, 2001; Bronkhorst & Plomp, 1988; Douglas S Brungart & Iyer, 2012; Lorenzi, Gatehouse, & Lever, 1999).

The areas of the brain that are responsible for sound localization and other spatial processes have been identified, yet it remains unknown these same areas are also implicated in improving intelligibility by producing SRM.

1.10 Objectives of the current project

The overarching goal of this thesis is to develop a clearer understanding of the familiar-voice benefit. This thesis aims to replicate and extend the findings of Johnsrude *et al*. (2013), in which researchers showed that the voice of a long-term spouse enhanced intelligibility when the familiar voice is either the target or the masker, and to compare this benefit to that afforded by spatial release from masking.

In the first experiment of this thesis, I will investigate whether the familiar-target and familiar-masker benefits discussed in Johnsrude *et al*. (2013) persists in a more challenging matrix task, and whether pairs of participants who are less naturally familiar than long-term spouses (such as friends, roommates, etc.) demonstrate a comparable familiar-voice benefit. Studies using trained familiar voices have observed considerable familiar-target benefits from four training sessions of 90 minutes each (Kreitewolf *et al*., 2017). Therefore, I hypothesize pairs of friends and roommates who have known each other for at least six months will demonstrate a familiar-target benefit. However, because I will use a more challenging closed-set intelligibility task compared to the CRM, I hypothesize that the familiar-masker benefit observed in Johnsrude *et al.* (2013) may not be replicated.

In the second experiment, I aim to further characterize the familiar-voice benefit to intelligibility in terms of a well-studied auditory cue: spatially separated simultaneous speech streams. The voice of a spouse produces an intelligibility benefit equivalent to 6-9 dB (Johnsrude *et al*., 2013) and spatial separations of ±90° produce an intelligibility benefit of 4 dB (Bronkhorst & Plomp, 1992), 6 dB (Yost, 2017), and 12 dB (Marrone *et al*., 2008). Therefore, I hypothesize that familiar voices will produce an intelligibility

benefit that is comparable to that produced by large spatial separations, in the same group.

Lastly, both familiar voices and spatial separations create large improvements to intelligibility compared to unfamiliar or spatially collocated speech (i.e., two possibly different kinds of release from masking). I will identify and compare brain areas involved in processing each of these cues. I will determine the degree to which the neural pathways involved in producing a familiar-voice benefit to intelligibility are similar to those underlying spatial release from masking. Because both cues improve intelligibility, I hypothesize that both cues will activate intelligibility areas in the superior and middle temporal lobe. I also hypothesize that each cue will have distinct neural substrates. Specifically, I hypothesize that familiar voices will activate frontal regions, specifically the IFG, that has been shown to be part of the 'what' pathway and is one of the frontal voice areas. Spatial separations will activate parietal regions that are part of the 'where' pathway and have been shown to be active in spatial listening tasks, such as the IPL and precuneus.

Taken together, these experiments will advance current knowledge of familiar voices and how they contribute to improving speech intelligibility.

Chapter 2

2 Improvements to intelligibility from voices of spouses and friends do not differ to each other

2.1 Introduction

Verbal communication frequently occurs in listening environments in which multiple sounds occur simultaneously, such as in the presence of competing talkers. To understand speech in these "cocktail party" environments, we must be able to separate these simultaneous sounds and attend to the target speech (Cherry, 1953). In favorable listening conditions, such as those with minimal background noise, listeners with normal hearing can segregate a voice from a mixture of sounds in order to successfully carry on a conversation. In more challenging situations—such as when competing sounds are more intense than target speech, when there are several simultaneous talkers, or when listeners have hearing impairment—intelligibility of target speech is poorer (Brungart, 2001; Dubno, Dirks, & Morgan, 1984; Glyde *et al.*, 2015; Van Engen & Bradlow, 2007), perhaps reflecting difficulty communicating in real-life settings with similar acoustic conditions.

Experience with a talker's voice improves the intelligibility of speech when competing sounds are present (e.g., Gass & Varonis, 1984; Holmes, Domingo, & Johnsrude, 2018; Johnsrude *et al*., 2013; Kreitewolf, Mathias, & von Kriegstein, 2017; Newman & Evers, 2007; Nygaard, Sommers, & Pisoni, 1994; Souza, Gehani, Wright, & McCloy, 2013; Yonan & Sommers, 2000). In the earliest studies that showed this intelligibility benefit for familiar voices (Nygaard & Pisoni, 1998; Nygaard *et al*., 1994) and in a more recent study (Kreitewolf *et al*., 2017), participants were trained in the lab with novel voices. Although these studies demonstrate that experience with a talker's voice improves speech intelligibility, they might underestimate the extent to which a naturally familiar voice can enhance intelligibility: Unlike trained voices, listeners experience naturally familiar voices in a variety of acoustic settings with different masking sounds and hear them over longer periods of time; across several months or years.

Johnsrude *et al.* (2013) examined the speech intelligibility benefit for naturally familiar voices with which listeners had extensive experience: that of a long-term spouse that the listener had been married to for more than 18 years. First, all participants recorded sentences from the Coordinate-Response Measure (CRM; Bolia, Nelson, Ericson, & Simpson, 2000) matrix test, which is a closed-set test often used in multi-talker intelligibility research (e.g., Best, Thompson, Mason, & Kidd, 2013; Brungart, Simpson, Ericson, & Scott, 2001; Kitterick, Bailey, & Summerfield, 2010; Mesgarani & Chang, 2012) and contains sentences in the form "Ready <call sign>, go to <colour> <number> now" (e.g., "Ready Baron go to red two now"). In the listening part of the study, participants heard two CRM sentences simultaneously and reported the colour-number coordinate spoken by the voice that said the callsign "Baron". Intelligibility of the target was better when either the target (familiar-target condition) or masker (familiar-masker condition) were in the spouse's voice than when both voices were unfamiliar (baseline condition).

Since a benefit of familiarity was observed even when the familiar voice was not the focus of attention (i.e. in the familiar-masker condition), Johnsrude *et al.* (2013) concluded that the benefit of a familiar voice probably arises because voice familiarity facilitates stream segregation. The alternative explanation, that voice familiarity merely facilitates extraction of a familiar voice from a mixture, is only possible if the voice to be extracted (i.e., that which matches a mental 'template' generated by previous exposure to the talker) is the focus of attention (Bregman, 1990). Another possibility is that listeners track and remember the color and number from both the target and masker voice, and the familiar voice indicates which pair to report. Interestingly, the intelligibility benefit derived from a familiar masker voice (familiar-masker benefit) in Johnsrude *et al.* (2013) was driven by younger listeners (aged 59 years and below): in general, the majority of errors on this task were words from the masker sentence, but younger listeners were less likely than older ones to mistake the masker voice for the target when the masker was their spouse (familiar-masker condition) than when the masker was also unfamiliar (baseline condition).

In contrast, Newman and Evers (2007) found a speech intelligibility benefit when a naturally familiar voice was the target but *not* when it was the masker. In this experiment, young participants were asked to shadow stories or isolated words spoken by their psychology professor. At the same time, they heard a story spoken by a different person who was unfamiliar to all participants. Participants who had taken classes with the professor made fewer shadowing errors than participants who had taken classes with a different professor. However, in a follow-up experiment in which the professor's voice was presented as the masker, and participants had to shadow the unfamiliar voice, there was no difference in the number of errors between participants who were and those who were not familiar with the professor's voice.

One possible reason why Johnsrude *et al.* (2013) observed a familiar-masker benefit and Newman and Evers (2007) did not is that the professor's voice was not as familiar as the spouses' voices in Johnsrude *et al.* (2013). Perhaps only a highly familiar voice that has personal significance (such as that of a spouse) can aid perceptual organization and improve intelligibility when it is the masker. Perhaps a professor's voice, only encountered in a formal setting during classroom lectures, can be picked out of a mixture when it is attended but is not familiar enough to aid perceptual organization and thereby improve performance when it is the masker.

In addition, the CRM task used in Johnsrude *et al.* (2013) has different psychometric properties to the non-matrix tasks such as those used in Newman and Evers (2007), Levi, Winters, and Pisoni, (2011), Nygaard and Pisoni (1998), and Nygaard *et al*., (1994), in which participants were asked to transcribe the words they heard. If participants were more willing to guess words they were unsure of when the target voice was familiar, they would report more words overall when the target was familiar, leading to a higher score because a subset (even if only a small, semantically predictable, subset) of these guesses would be correct, whereas not reporting any of those words would always be counted as incorrect. One advantage of the CRM task is that listeners select exactly the same number of words from a fixed list on each trial, meaning that differences in performance between trials containing familiar and unfamiliar voices cannot be explained by a difference in bias (i.e., willingness to guess when uncertain).

Nevertheless, a limitation of the CRM task is that listeners only need to report the color and number key words of the target (e.g., “green six”), rather than every word from the target sentence. Typically, the listener reports what they heard by pressing the correctly coloured digit (e.g., the green “6” button) from a matrix of coloured digits presented on the screen. In the Johnsrude *et al.* (2013) experiment, with only a single masking talker, the listener may have been able to attend to the two colour-number pairs, then retrospectively select the correct coloured digit based on the target callsign voice.

One aim of the current experiment was to determine whether the familiar-target and familiar-masker benefits could be replicated using a different closed-set task that requires participants to report every word in an utterance. I used the sentences of the Boston University Gerald (BUG) (Kidd *et al*., 2008, which each contain five words (“<Name> <verb> <number> <adjective> <noun>”). The first (Name) word specifies the target sentence and participants report the remaining four words from that sentence. With a two-talker mixture, if they were to attend to the mixture and select the words that matched the callsign voice, they would have to remember eight items (plus keep track of which voice said the target name), which is much more difficult than remembering two colour-number pairs in the CRM task. Given that Johnsrude *et al*. (2013) found that the magnitude of the familiar-voice benefit depended on the TMR, I presented the stimuli at four different TMRs: -6, -3, 0, and 3 dB.

Another aim of the current study was to examine whether the magnitude of the familiar-voice benefit to intelligibility differs depending on the duration of the relationship. To investigate the length of the relationship, I compared a group of people who heard the voice of their spouse (highly familiar) with a group who heard the voice of a friend (less familiar). In addition, I explored whether within-group differences in relationship duration systematically affect the magnitude of the familiarity benefit. Possibly, the familiar-voice benefit improves gradually with longer durations of knowing someone—and spouses, which are known on average for longer than friends, may provide a greater benefit to intelligibility.

We had a wide age range in the spouse group, so to investigate effects of age, I split the spouse group into older and younger adults. The reason for dividing the spouse group was that older adults have poorer speech comprehension performance than younger adults (Helfer & Freyman, 2008; Tun, O'Kane, & Wingfield, 2002) and this could affect the benefit that listeners get from a familiar voice. Further, I examined whether age affected accuracy differently in each condition. Johnsrude *et al*. (2013) found that younger participants (aged 44-59 years old) were less likely to report words spoken by a familiar masker voice compared to older participants (aged 60+ years old).

2.2 Method

2.2.1 Participants

Participants were 68 individuals, recruited in pairs. I recruited 16 pairs who were married (16 males, 16 females; "Spouses group") and were aged 28–82 years (median = 59.5 years, interquartile range [IQR] = 33.0). I also recruited 18 pairs of friends (11 males, 25 females; "Friends group") who were aged 18–25 years (median = 21 years, IQR = 3.5 years). Of these 18 pairs, 11 pairs were friends or roommates, five pairs were romantic couples, and two pairs were siblings. One couple from the spouse group and three pairs from the friend group (including two romantic couples) did not complete the experiment, which required multiple visits. The data from the remaining 60 individuals were analyzed.

I administered a questionnaire that asked about the length of time participants had known each other or had been married. This questionnaire was completed by 30 spouse participants and 15 friend participants. Spouses reported that they had been married for more than 4 years (range 4.1–51.9 years; median = 27.0 years, IQR = 28.8 years). Friend pairs reported that they had known each other for 1.5–19 years (median = 5.0 years, IQR = 16.0 years). An independent samples Mann-Whitney test indicated that the length of time married pairs had been living together was significantly longer than the length of time friend pairs had known each other [$U = 62.00$, $p < .001$].

I split the Spouses group into two groups of approximately equal size based on age: Older (age ≥ 55 years; N = 16) and Younger (age < 55 years; N = 14). This grouping is similar

to that used in Johnsrude *et al.* (2013) and allowed us to examine age-related differences in the familiar-target benefit. The age range in the Friends group was substantially smaller, and all were younger than the older Spouses group, so the Friends group was not divided. The sample size of the smallest group (N = 14) is estimated to be sensitive to within-subjects effects of size f = 0.41 with 0.95 power (Faul, Erdfelder, Lang, & Buchner, 2007), and therefore should be large enough to detect familiar-voice benefits to intelligibility of the magnitude reported by Johnsrude *et al*. (2013) (f = 0.72). With at least 14 participants in each group, I should be sensitive to group-by-familiarity interactions of size f = 0.23 with 0.95 power.

All participants were self-declared native Canadian English speakers who had no known speech, hearing, or neurological impairments. Participants had hearing levels (measured using pure tone audiometry at four octave frequencies between 500 and 4000 Hz) of 25 dB HL or better averaged across both ears, except for one participant who had an average pure-tone hearing level of 35 dB HL. The same pattern of results obtained whether this individual was included or not, so I reported results including data from this participant.

The study was approved by the University of Western Ontario Non-Medical Research Ethics Board. Informed consent was obtained from all participants.

2.2.2 Materials and Procedure

Participants were tested across two or three sessions. During the first session, each participant was recorded while speaking 480 different sentences, taken from the BUG corpus (Kidd *et al*., 2008). The sentences had the form "<Name> <verb> <number> <adjective> <noun>". In the sub-set used in the experiment, there were two names ('Bob' and 'Pat'), eight verbs, eight numbers, eight adjectives, and eight nouns (see Figure 1). An example is "Bob bought two blue bags". Across the 480 sentences that were recorded, each verb, number, adjective, and noun occurred 60 times. Sentences were recorded at a 44.1 kHz sampling rate using a Sennheiser e845 S microphone connected to a Steinberg UR22 soundcard. Unlike the original BUG corpus, in which each possible word was recorded individually and sentences were later constructed by concatenating individually spoken words, each sentence in this study was recorded in its entirety, thus retaining

natural coarticulation and supra-segmental prosody between words. All sentences were normalized to the same root mean square (RMS) amplitude.

Participants returned for the listening task approximately three months (mean days of separation = 74.4 days, standard deviation [SD] = 73.2 days) after completing the recording session. The listening task was completed in either one session of approximately two hours (N = 36) or two sessions of approximately one hour each, which were separated by less than one month (N = 24; mean days of separation = 14.5, SD = 22.8). Stimuli were presented diotically through Sennheiser HD265 (N = 26) or Grado Labs SR225 (N = 34) headphones. Each participant heard sentences spoken by three different talkers: the participant's partner (familiar talker), and two other participants in the study who the participant did not know but who were from the same group and were the same sex as the participant's partner (unfamiliar talkers). The two unfamiliar voices remained constant for each participant throughout the experiment.

On each trial, participants heard two different sentences spoken simultaneously by different talkers. All of the words of the two sentences were different. The target sentence was identified by one of two names at sentence onset (either Bob or Pat). One name was used as the target for the first half of trials and the other was used for the second half of trials; the order was counterbalanced across participants. Listeners were instructed to identify the remaining four words in the target sentence by clicking on each word on a computer screen. I matched the occurrences of word combinations, so that participants would not know one word in the sentence based on the presence of other words. As illustrated in Figure 2, the words were arranged in four columns, with one column per word type. Participants selected one word from each column, in any order. The target name (Bob or Pat) was displayed at the top of the screen, as a reminder. The response screen remained visible throughout the entire experiment, including during presentation of stimuli, to minimize load on hort-term memory.

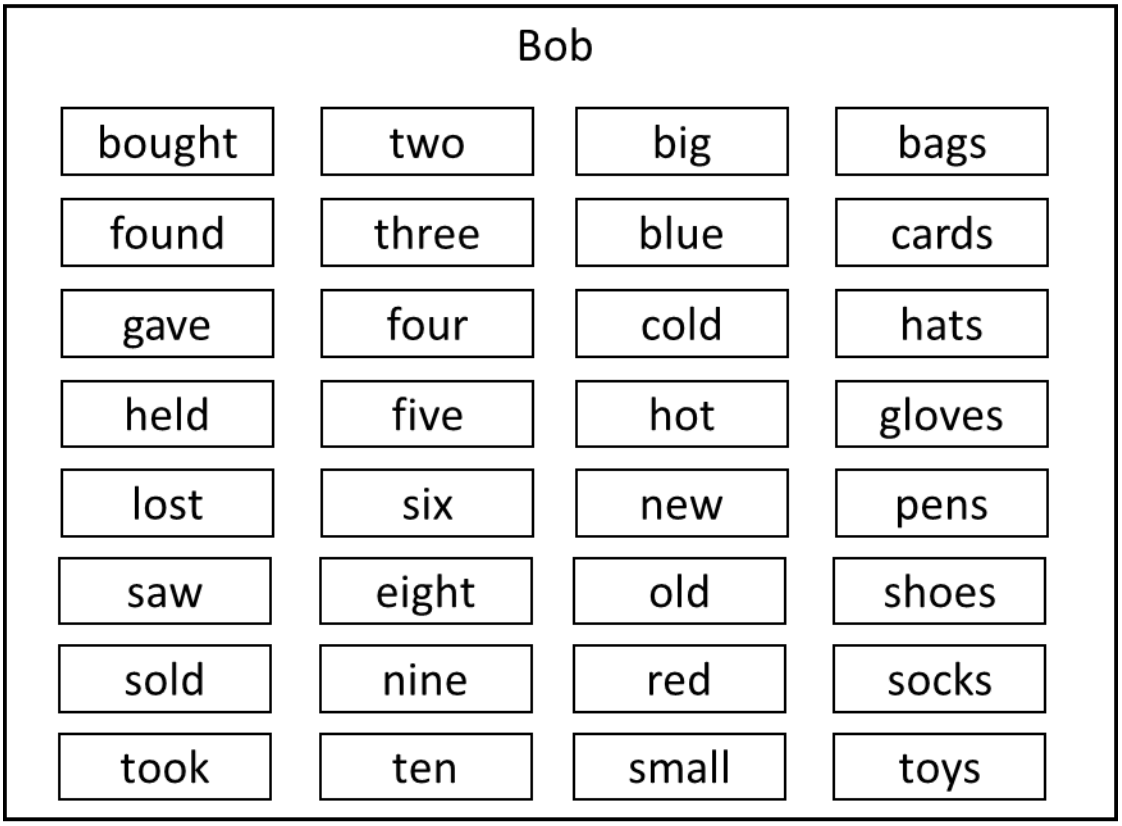


Figure 2: Schematic of the response screen used for the listening task. Participants were asked to choose one word (by a mouse press) from each column according to what they had heard in the target sentence, indicated by the target name (in this example, "Bob").

Intelligibility of the target sentence was tested in three conditions. In the Familiar Target condition, the target sentence was spoken by the participant's partner (i.e., their familiar voice) and the masker sentence was spoken by one of their two unfamiliar talkers (half with each unfamiliar talker). In the Familiar Masker condition, the masker sentence was spoken by the participant's partner and the target sentence was spoken by one of the unfamiliar talkers (half with each unfamiliar talker). In the Both Unfamiliar condition, the target and masker sentences were spoken by the two unfamiliar talkers. In one half of these trials, one unfamiliar voice was the target and the other was the masker; in the other half, the voice roles were reversed.

I varied the target and masker intensities at four target-to-masker ratios (TMRs): -6, -3, 0, and +3 dB. Acoustic stimuli were presented at a comfortable listening level (approximately 67 dB SPL). The overall amplitude of the target and masker sentences in each trial was roved over a range of 3 dB (in 6 equally spaced levels) to ensure that participants could not use the amplitude of either sentence as a cue to identify the target sentence.

Each participant completed 720 trials: 240 trials in each familiarity condition. Across the experiment, participants heard each of the three voices 240 times as the target and 240 times as the masker. Each familiarity condition contained equal numbers of trials at each of the four TMRs and each of the six rove levels. All trial types were randomly interleaved over 30 blocks of 24 trials each. Participants were prompted to rest, if they wished, between blocks. The participant initiated each block of trials by clicking a prompt on the screen when they were ready to begin.

2.2.3 Analyses

2.2.3.1 Accuracy

I calculated the proportion of words (out of a possible 960; 4 words in each of 240 trials) that participants reported correctly in each condition. There were 8 options for each word, so the chance level of performance was 12.5%. I used a 3-way mixed analysis of variance (ANOVA) to compare percent correct across Familiarity Conditions (3 levels: Familiar Target, Familiar Masker, Both Unfamiliar, within-subjects), TMRs (4 levels: -6 dB, -3 dB, 0 dB, 3 dB; within-subjects), and groups (3 levels: Young Friends, Young Spouses, and Older Spouses; between-subjects); see Figure 3.

I always presented unfamiliar voices of the same sex as the participant's familiar voice, but because I used natural voices there was some variability across participants in the degree to which the F_0 of the familiar voice differed from that of each of the unfamiliar voices. At the group level, all three familiarity conditions were acoustically very well matched, because all familiar voices also served as unfamiliar voices, meaning that the voices heard as familiar were acoustically identical to those heard as unfamiliar (with the exceptions noted above). However, given that intelligibility of a target talker in the

presence of a competing talker is known to improve as the difference in F_0 between the two talkers increases (Assmann, 1999; Christopher J Darwin, Brungart, & Simpson, 2003; Summers & Leek, 1998), the F_0 difference has the potential to influence intelligibility at an individual level. I therefore included it as a covariate of no interest in the ANOVA.

We estimated the F_0 of each recorded sentence using an in-house script written in Praat (Boersma & Weenink, 2013), which calculated the median F_0 across each sentence at time steps of 0.01 seconds. To determine each talker's F_0, I averaged the median F_0 values across all of the 480 sentences they recorded. For each participant, I calculated the absolute difference in F_0 between the familiar and the average of the two unfamiliar talkers they heard during the experiment. Fundamental frequencies for each sex in each group are described in Table 2 (median = 12.5 Hz, IQR = 20.6 Hz, which corresponds to 2.06 semitones, IQR = 1.70 semitones).

Table 2. Mean fundamental frequency (F_0) for males and females in each group. Standard deviations are displayed in brackets.

Group	*n*	F_0 (Hz)
Older Spouses		
Male	8	107.69 (16.53)
Female	8	170.76 (10.77)
Younger Spouses		
Male	7	103.73 (12.64)
Female	7	186.95 (22.34)
Young Friends		
Male	7	111.95 (12.84)
Female	23	205.68 (15.84)

Mauchly's tests indicated that the assumption of sphericity was violated for the main effect of Familiarity [$\chi^2(2) = 33.80$, $p < .001$], main effect of TMR [$\chi^2(5) = 89.56$, $p < .001$], and interaction between Familiarity and TMR [$\chi^2(20) = 96.21$, $p < .001$]; these results are reported with Greenhouse-Geisser correction. Pairwise comparisons are reported with Sidak correction for multiple comparisons.

2.2.3.2 Errors

Incorrectly reported words were categorized as one of two types: (1) 'wrong voice' errors, in which the reported word was from the masker sentence; and (2) 'random' errors, in which the reported word was not contained in either of the two sentences spoken on that trial. Percentage of errors was calculated by dividing the number of each type of error by the total number of words in incorrect trials. I used a four-way mixed multivariate analysis of variance (MANOVA), with average F_0 difference as a covariate of no interest, to compare the percentage of Errors (2 levels: Wrong Voice, Random; within-subjects) across familiarity conditions (3 levels: Familiar Target, Familiar Masker, Both Unfamiliar; within-subjects), TMRs (4 levels: -6, -3, 0, 3 dB; within-subjects), and groups (3 levels: Young Friends, Young Spouses, and Older Spouses; between-subjects). I conducted follow-up within-subjects ANOVAs to better understand the effects of Familiarity and TMR on each type of error (Wrong Voice or Random) separately.

2.2.3.3 Age-related differences on intelligibility

Johnsrude *et al*. (2013), using the CRM procedure, observed that task accuracy correlated negatively with age in the Familiar Masker and Both Unfamiliar conditions, but was unrelated to age in the Familiar Target condition. The correlation values differed significantly between the Familiar Target and the other two conditions, and, furthermore, these differences were apparent at TMR values equated across conditions for performance. To examine whether the same relationships obtained in the current matrix-task data, I calculated Spearman correlations between age and accuracy in each of the three familiarity conditions, across the Older and Younger Spouse data. I also statistically compared these correlations (Lee & Preacher, 2013).

2.2.3.4 Influence of relationship duration

To assess whether the magnitude of the intelligibility benefit gained from a familiar voice is related to the length of the relationship, I conducted a partial correlation between Relationship Duration and Familiar-Target Benefit, calculated as the difference in percent correct between the Familiar Target and Both Unfamiliar conditions for each participant, while controlling for the possibly confounding effect of F_0 difference between familiar

and unfamiliar voices. The Relationship Duration was defined for each pair, as the length of time the spouses had been married and the length of time the friends had known each other.

2.3 Results

2.3.1 Accuracy

Data are shown in Figure 3. A three-way mixed ANOVA, controlling for the F_0 difference between familiar and unfamiliar voices, revealed no effect of the covariate (F_0 difference), [$F(1, 56) = 0.28$, $p = .60$, $\omega^2 = -.01$] and no significant interactions involving it (ps > .05).

The main effect of Familiarity was significant [$F(1.37, 76.76) = 8.40$, $p = .002$, $\omega^2 = .11$]. Participants reported more correct words when the target voice was familiar (Familiar Target: mean = 69.28%, standard error [SE] = 2.37) than when the masker voice was familiar (Familiar Masker: mean = 56.97%, SE = 2.23) ($t(59) = 4.81$, $p < .001$), or when both target and masker voice were unfamiliar (Both Unfamiliar: 59.75%, SE = 2.16) ($t(59) = 5.14$, $p < .001$). There was no significant difference in accuracy between the Familiar Masker and Both Unfamiliar conditions ($t(59) = -1.98$, $p = .15$), and the difference trended in the opposite direction to that observed by Johnsrude *et al*. (2013), i.e. towards worse target-word report in the Familiar Masker condition.

As expected, there was a significant main effect of TMR [$F(1.49, 83.36) = 19.60$, $p < .001$, $\omega^2 = .24$]. Participants were more accurate at reporting target words at 3 dB TMR (mean = 68.77%, SE = 1.86) than at 0 dB (mean = 62.00%, SE = 1.91) ($t(59) = 8.95$, $p < .001$), -3 dB (mean = 59.26%, SE = 2.01) ($t(59) = 8.86$, $p < .0001$), and -6 dB (mean = 59.94%, SE = 2.26) ($t(59) = 7.60$, $p < .001$). Accuracy was also better at 0 dB than at -3 dB ($t(59) = 4.32$, $p < .001$) and -6 dB ($t(59) = 4.21$, $p = .001$). The percentage of correctly reported words did not differ between -3 dB and -6 dB TMR ($t(59) = 2.10$, $p = .22$).

There was no significant main effect of Group [$F(2, 56) = 1.45$, $p = .24$, $\omega^2 = .02$], suggesting that intelligibility does not differ between older spouses, younger spouses, and friends.

There was a significant interaction between Group and TMR [$F(2.98, 83.36) = 9.78$, $p < .001$, $\omega^2 = .23$]. Performance by older spouses was more affected by TMR than was performance in the other two groups (Figure 3). Intelligibility at higher TMRs (-3, 0, and 3 dB) did not differ between Older and Younger Spouses ($0.03 \geq t\text{s}(59) \geq 2.34$, $p\text{s} > .09$), but older spouses reported significantly fewer correct words than Younger Spouses at the lowest TMR, -6 dB ($t(59) = -2.74$, $p = .03$).

The interaction between Group and Familiarity was not significant, [$F(2.74, 76.76) = 1.26$, $p = .29$, $\omega^2 = .01$], neither was the three-way interaction between Group, Familiarity, and TMR [$F(6.68, 190.40) = 0.628$, $p = .73$, $\omega^2 = -.01$], suggesting that the presence of a familiar voice affected intelligibility in a similar way across groups and TMRs. None of the other interactions were significant, either ($0.33 \geq t\text{s}(59) \geq 2.24$, $p\text{s} > .30$).

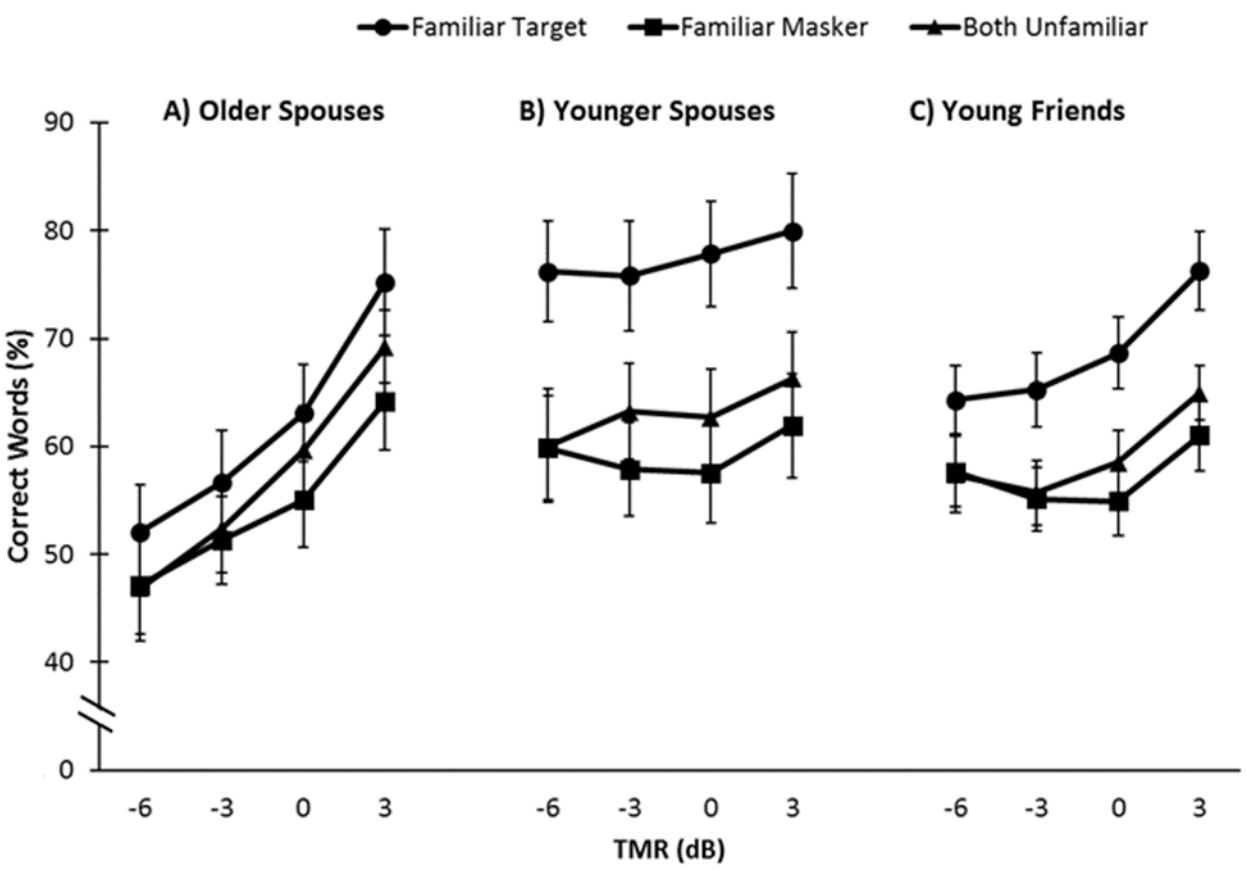


Figure 3. Percentage of correct words in each familiarity condition as a function of target-to-masker ratio (TMR) in Older Spouses (A), Younger Spouses (B), and Friends (C). Data points indicate group means. Error bars show ±1 standard error of the mean.

We repeated the analyses using the percentage of sentences that were reported correctly (rather than correct words), which I defined as trials in which all four words of the target sentence were reported correctly. For this, chance performance is 0.02%. As expected, the percentage of correct sentences was lower than the percentage of correct words across conditions (Familiar Target: mean = 44.74%, SE = 2.58; Familiar Masker: mean = 30.31%, SE = 2.47; and Both Unfamiliar: mean = 32.50%, SE = 2.12). However, the pattern of results did not differ appreciably from the analysis based on words correct.

As a post-hoc analysis, I checked whether the sibling pairs were driving the results in the Friends group; they may have performed differently since their relationship is of a much longer duration, compared to other pairs in the Friends group. I repeated the accuracy analysis but excluded the two sibling pairs, and results did not differ from those reported above. I conducted a separate repeated-measures ANOVA on the Friends group (with siblings excluded) to determine whether there were accuracy differences between friends (n=22) and dating couples (n=10) across Familiarity conditions and TMRs. I did not find any effect of relationship type [$F(1, 23) = 1.84$, $p = .19$, $\omega^2 = .03$].

2.3.2 Errors

In general, people made substantially more 'wrong voice' than 'random' errors. Among the identified words in error trials (those in which at least one word out of a possible four was identified incorrectly), 48.59% (SE = .82) were wrong voice errors, and 10.10% (SE = .56) were random errors. The remaining 41.31% were correctly identified words. The data are presented in Figure 4.

We conducted a four-way mixed MANOVA to compare the proportion of these two types of error across Familiarity Conditions, Groups, and TMR, while controlling for F_0 differences between familiar and unfamiliar voices. The effect of the covariate was not significant, [$F(1, 56) = .559$, $p = .46$, $\omega^2 = -.01$], nor were any of the interactions involving it (ps > .15). I only report the main effect of the Error Type factor and interactions involving it, since the other effects are similar to those reported in the Accuracy analysis (above).

The analysis confirmed that the main effect of Error Type was significant [$F(1, 56) = 374.87$, $p < .001$, $\omega^2 = .87$]. The interaction between Group and Error Type was also significant [$F(2, 56) = 4.43$, $p = .02$, $\omega^2 = .10$]. Whereas the proportion of wrong voice errors did not differ among the three Groups ($1.63 \geq t\text{s}(59) \geq 2.17$, ps > .10), Older Spouses made more random errors than did Younger Spouses ($t(59) \geq 2.51$, $p = .045$) and Friends ($t(59) \geq 3.67$, $p = .01$). The proportion of random errors did not differ between Younger Spouses and Friends ($t(59) \geq 0.67$, $p = .88$).

The interaction between Error Type and Familiarity condition trended towards significance [$F(2, 55) = 2.80$, $p = .07$, $\omega^2 = .06$]. Given that I were expecting to find a difference between the familiar-masker and both-unfamiliar conditions (based on Johnsrude *et al*., 2013), I explored this interaction further. Although the proportion of Random errors did not differ across familiarity conditions ($0.43 \geq ts(59) \geq 1.61$, $ps > .30$), participants made significantly fewer Wrong Voice errors in the familiar-target compared to the familiar-masker and both-unfamiliar conditions ($-4.99 \geq ts(59) \geq -3.58$, $ps < .01$). The percentage of Wrong Voice errors did not differ between the familiar-masker and both-unfamiliar conditions ($t(59) = 0.36$, $p = .98$).

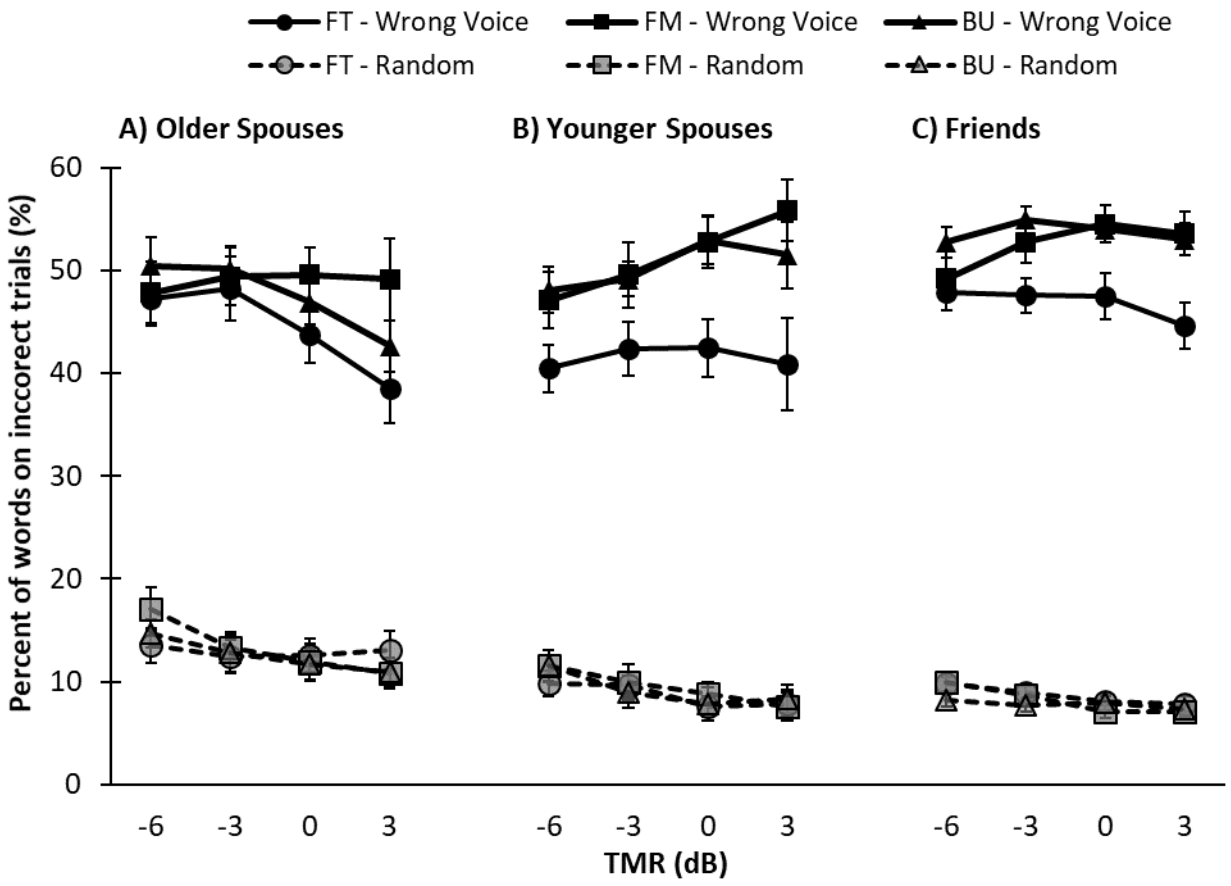


Figure 4. Error analysis. 'Wrong voice' errors (black markers, solid lines), and 'random' errors (grey markers, dashed lines) in incorrect trials as a function of target-to-masker ratio (TMR), expressed as a proportion of all words presented on incorrect trials (trials on which at least one word was reported incorrectly). Left panel (A) shows data from Older Spouses, middle panel (B) shows data from Younger Spouses, and right panel (C) shows data from Friends. Data points show group means. Error bars show ±1 standard error of the mean (SE). FT (circles): Familiar Target; FM (squares): Familiar Masker; BU (triangles): Both Unfamiliar.

2.3.3 Age-related differences on intelligibility

There was a significant negative correlation between age and accuracy in the Familiar Target condition (collapsed across TMRs) [r_s = -.51, p = .004], but not in the Familiar Masker condition [r_s = -.07, p = .70], or in the Both Unfamiliar condition [r_s = -.31, p = .09]. These correlations are shown in Figure 5. I tested for any differences between these

correlations (Lee & Preacher, 2013). Correlations in the familiar-target and familiar-masker conditions differed significantly from each other [Z = -.208, p = .037], whereas correlations in the familiar-target condition did not differ significantly from the correlation in the both-unfamiliar condition [Z = -1.08, p = .28]. These results suggest that Familiar-Target intelligibility decreases more rapidly with age compared to Familiar Masker intelligibility but not more rapidly than Both Unfamiliar intelligibility.

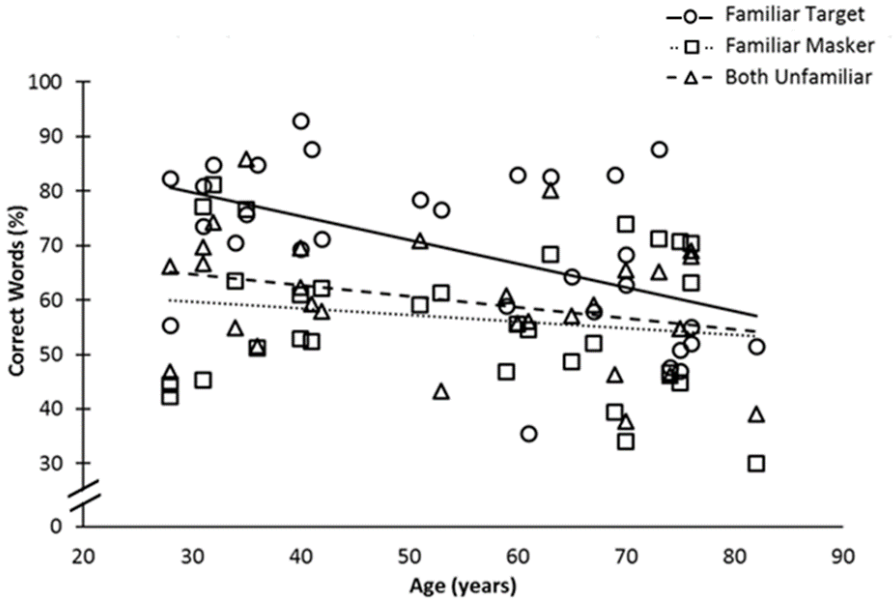


Figure 5. Scatter plot and best-fit regression lines showing the relationship between age and accuracy in the Familiar Target (circles, solid line), Familiar Masker (squares, dotted line), and Both Unfamiliar (triangles, dashed line) conditions.

2.3.4 Influence of relationship duration

Despite finding no significant differences in the familiar-voice benefit between the Spouses and Friends groups (which would have manifest as a significant interaction between Group and Familiarity Condition in the analyses above), I wanted to examine whether variability in length of time participants had known their partner related to the magnitude of the familiar-voice benefit. I tested the partial correlation between Relationship Duration and the familiar-target benefit (difference in intelligibility between the Familiar Target and Both Unfamiliar conditions) across individuals, while controlling for the F_0 difference between the familiar and unfamiliar voices. The correlation was not

significant [r = -.24, p = .12] in the range I had questionnaire data for (1.5-51.9 years). This result suggests that longer relationships do not systematically increase the benefit to intelligibility from a familiar voice.

2.3.5 Influence of talker F_0

In addition to including F_0 as a covariate in my main analysis (above), I also tested post-hoc whether larger F_0 differences were related to bigger apparent familiar-voice intelligibility benefits at an individual level (as we might expect). A Shapiro-Wilk's test indicated that F_0 differences differed significantly from a normal distribution, [$W(60)$ = .917, p = .001]; therefore I conducted a Spearman's correlation between the F_0 difference and the individual intelligibility benefit across participants. For this correlation analysis, I analyzed all participants in one group (spouses and friends combined) so that I had more power to detect a significant relationship.

There was a significant positive correlation between the magnitude of the difference in intelligibility and the magnitude of the F_0 difference between the familiar and the two unfamiliar talkers (averaged together) [r_s = .26, p = .045]. This result demonstrates that the F_0 difference between the familiar and unfamiliar voices explained a significant amount of the individual variability in the magnitude of the familiar-target benefit, as expected.

2.3.6 Influence of sex of familiar voice

Given the unfamiliar voices were sex-matched to the familiar voice, I conducted another post-hoc analysis to determine whether the sex of the familiar voice had an effect of on intelligibility. I conducted a mixed ANOVA with Familiarity (Familiar Target, Familiar Masker, Both Unfamiliar) and TMR (-6, -3, 0, 3 dB) as within-subjects factors and sex of the familiar talker as a between-subjects factor.

There was no effect of the sex of the familiar talker, [$F(1,58)$ = .001, p=.98, ω^2 = -.002], or any significant interactions involving it (ps $\geq$ .38), suggesting that presenting mixtures of male voices or female voices did not affect intelligibility. I therefore collapsed across sex for the remainder of the analyses.

2.3.7 Do unfamiliar voices become ‘familiar’?

Participants heard the two unfamiliar voices many times throughout the experiment, and it is possible that these unfamiliar voices became ‘familiar’ by the end of the experiment. Adaptation to new forms of speech has been shown to occur rapidly, after only 15 trials of exposure (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Huyck & Johnsrude, 2012). Therefore, I conducted a post-hoc analysis to determine whether participants became familiar with the two unfamiliar voices, which would manifest as a greater improvement in intelligibility scores for unfamiliar-target than familiar-target conditions from the beginning of the experiment to the end. Separately for the three Familiarity Conditions, I took a subset of 20 trials from the beginning and end of the experiment, which should be sufficient to get a stable average whilst also being sensitive to perceptual learning of unfamiliar voices of the type described by Huyck and Johnsrude (2012).

I conducted a three-way repeated measures ANOVA to compare the percent of words reported correctly across Groups (3 levels: Older Spouses, Younger Spouses, and Friends), Familiarity conditions (3 levels: Familiar Target, Familiar Masker, Both Unfamiliar) and Trial Positions (2 levels: first and last 20 trials). If, following exposure to the voices, unfamiliar voices became similar to familiar voices, there should be a Familiarity Condition by Trial Position interaction such that accuracy in Both Unfamiliar trials improves to a greater extent between the first and last 20 trials than does accuracy in the Familiar Target condition.

Figure 6 illustrates percent correct in the first and last 20 trials for all three conditions, collapsed across groups. There was a significant effect of Trial Position [$F(1, 57) = 89.05$, $p < .001$, $\omega^2 = .60$]: the last 20 trials (mean: 63.47%, SE = 1.25) were more intelligible than the first 20 trials (mean = 52.68%, SE = 1.29).

Importantly, there was no significant interaction between Familiarity condition and Trial Position [$F(1.35, 76.78) = 1.05$, $p = .33$, $\omega^2 = .00$], but intelligibility of unfamiliar voices did not improve to a greater extent than did intelligibility of familiar voices. Thus, I

found no evidence that intelligibility of the unfamiliar voices was enhanced by learning over the experiment.

There was no significant 2-way interaction between Group and Trial Position [$F(2, 57) = 0.97$, $p = .37$, $\omega^2 = .00$] and no significant 3-way interaction between Group, Trial Position, and Familiarity [$F(2.69, 76.78) = 0.60$, $p = .60$, $\omega^2 = -.01$]. Thus, the magnitude of the improvement in intelligibility between the first and last 20 trials did not differ among groups. Because I assumed that the perceptual learning of the familiar voice has occurred before the experiment and has reached its maximum (as participant pairs are required to know each other for at least six months and speak regularly), I interpret the overall improvement in performance between the first 20 and last 20 trials as attributable to practice effects.

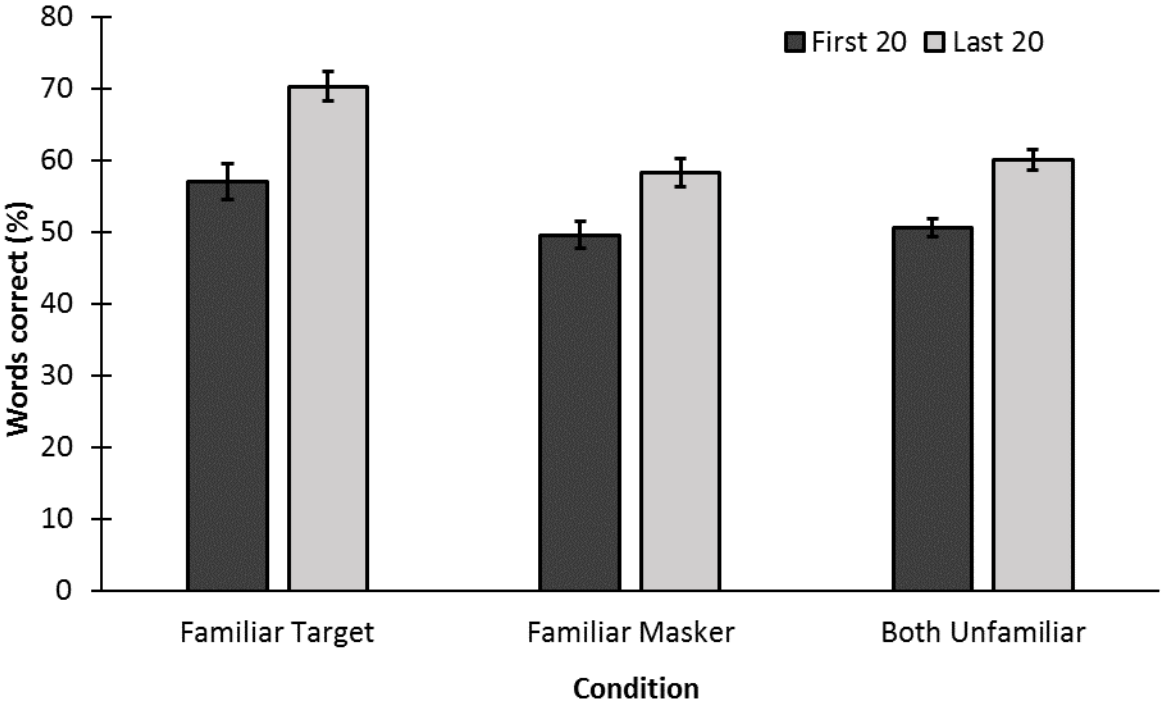


Figure 6. Percent correct of first and last 20 trials, collapsed over Groups and TMRs, for each condition. Error bars represent ±1 standard error of the mean.

2.4 Discussion

2.4.1 Familiar-target benefit is similar for spouses and friends

These results demonstrate that people are better at understanding speech in the presence of a competing talker when the talker they are listening to is a spouse or friend, compared to when it is someone unfamiliar. Words spoken in a familiar voice (Familiar Target condition) were, on average, 10–15% more intelligible than words spoken in an unfamiliar voice (Familiar Masker and Both Unfamiliar conditions; Figure 2). Thus, I replicated the familiar-target benefit found in previous experiments (Gass & Varonis, 1984; Johnsrude *et a*l., 2013; Kreitewolf *et al*., 2017; Newman & Evers, 2007; Nygaard *et al*., 1994; Souza et al., 2013; Yonan & Sommers, 2000) using the closed-set BUG task. Furthermore, I showed that friend's voices and spouse's voices appear to be similarly beneficial for intelligibility when a competing talker is present.

My results extend previous findings by demonstrating a familiar-target benefit for a closed-set test with a high memory load (BUG corpus; Kidd, Best, & Mason, 2008). In contrast, previous experiments used either open-set tests (Newman & Evers, 2007; Yonan & Sommers, 2000) or closed-set tests with a low memory load (i.e., the CRM test in Johnsrude *et al*., 2013). That the familiar-voice benefit is present for closed-set tests indicates that it is not (entirely) due to systematic differences in response bias when people hear speech in familiar and unfamiliar voices: Unlike open-set tasks using naturalistic stimuli, participants reported a fixed number of words on every trial, and the words could never be predicted based on the previous word(s). Therefore, participants must guess if unsure on every trial, regardless of whether they heard a familiar or unfamiliar voice. The high memory load of the BUG task is more similar to everyday conversations than the CRM task used by Johnsrude *et al*. (2013); in most everyday situations, successful communication requires listeners to follow all or most of the words spoken by an interlocutor, whereas the CRM task only requires listeners to extract the colour-number coordinate near the end of each sentence. The current results increase confidence that the familiar-voice benefit improves speech intelligibility in natural communication settings.

The familiar-target benefit I found is similar in magnitude to that reported in previous studies (Johnsrude *et al*., 2013) using closed-set testing that controls for the effect of guessing (bias) on measured intelligibility. Johnsrude *et al*. (2013) found a 10–15% improvement in intelligibility (sentence report) when a target sentence was spoken by the participant's spouse than when it was spoken by an unfamiliar talker. Here, I find a similar benefit for both spouses and friends. Spouses generally knew each other for longer than the friends I tested and presumably have relationships that differ in quality from those of the friend pairs. Nevertheless, the intelligibility benefit was as large for friends as for spouses. Consistent with this result, I found no correlation between the length of time participants had known each other and the magnitude of the intelligibility benefit. Given these results, it is possible that intelligibility due to familiarity with someone's voice manifests rather quickly (within a year and a half of knowing someone) and then remains stable in magnitude as the relationship extends through the years.

The finding that the benefit to intelligibility of friends' voices is as robust as the benefit from a spouse's voice when heard in the presence of a competing talker, has practical significance. To the extent that these results generalize to real-world listening, the intelligibility of casual friends in busy environments should be as high as the intelligibility of a longstanding life partner. People do not need to be exposed to a voice as intensively as they have been exposed to their spouse's voice to improve intelligibility substantially. That familiar voices can improve intelligibility after relatively short exposure is consistent with the results of Newman and Evers (2007), who showed that participants were better at understanding words spoken by a psychology professor by whom they had been taught for one semester than words spoken by a novel voice. In addition, training studies have shown a familiar-voice benefit when participants are exposed to voices for as little as six hours (Kreitewolf *et al*., 2017). However, the benefit of a lab trained voice appears to be of smaller magnitude compared to the benefits I have observed (approximately 10-15%, which I estimate was equivalent to ≥ 3 dB for all groups): 0.52 dB in Kreitewolf *et al*. (2017), approximately 5-10% (Nygaard *et al*., 1994), and approximately 3-15% (Nygaard & Pisoni, 1998). Furthermore, given the impoverished materials that I used, and the lack of natural prosodic and contextual information, I think this measured benefit probably underestimates real-world benefit.

Intelligibility of the unfamiliar voices did not approach the intelligibility of familiar voices by the end of my experiment, demonstrating that more than brief, incidental, exposure to voices is required to produce a familiar-target benefit of the magnitude observed here. A longitudinal study could investigate the time course of the familiar-voice benefit in more detail, and determine what type of experience with a voice is required for an intelligibility benefit to be observed. If a trained talker who the participant has never met could provide an intelligibility benefit as large as that found here, then voice training could have great potential for improving intelligibility in everyday environments—such as public announcements in busy places—and these might help people who find it difficult to listen in noise, including older people who experience declines in hearing with healthy aging. In either case, my results suggest that older people gain as much benefit from a familiar voice as younger people, suggesting that real-world speech intelligibility can be improved by voice familiarity.

2.4.2 No benefit of familiarity with a masker voice

In contrast to Johnsrude *et al*. (2013), I found no benefit of familiarity with the masker voice on the intelligibility of an unfamiliar target voice in any of the three groups. To my knowledge, Johnsrude *et al*. (2013) is the only study to have found a familiar-masker benefit. Johnsrude *et al*. (2013) concluded that the presence of a familiar voice in a mixture (as either the target or masking voice) may aid in perceptual organization. If they had found no familiar-masker benefit, and only a benefit when the familiar voice is the target (and focus of attention), this result could have been accounted for by a template-matching strategy in which participants use a mental representation of the familiar voice to extract it from the mixture (Bregman, 1990). By definition, this strategy is only possible when the speech matching the template is the focus of attention, and therefore cannot explain the familiar-masker benefit they observed.

The absence of a familiar-masker benefit in this study compared to Johnsrude e*t al*. (2013) could be due to the greater cognitive demand of the BUG task compared to the CRM task. The BUG and CRM materials that were used here and in Johnsrude *et al*. (2013) are both closed-set matrix tests, but differ markedly on the number of items to be reported (four words in BUG and one colour-number pair in CRM). To respond correctly

in the BUG task, participants would have to identify the target voice (specified by the ‘name’ word) and correctly report the other four words (‘verb’ ‘number’ ‘adjective’ ‘noun’) in the target sentence. To respond correctly on Familiar Target or Familiar Masker trials in the CRM task, participants need only attend to the call sign at the onset, decide whether their partner’s voice is the target (i.e., said ‘Baron’) or the masker, register both coloured digits, then retrospectively indicate the one spoken by their partner (if target) or the other talker (if their partner is the masker). This strategy is a lot harder to deploy if eight to ten unrelated words have to be held in memory and each assigned to the correct voice. The difference in strategies that could possibly be used for the CRM and BUG tasks could explain why Johnsrude *et al*. (2013) found better overall intelligibility than I found in the current experiment, and why they observed a familiar-masker benefit and I did not.

2.4.3 Older listeners

The pattern of performance in older listeners (55–82 years) was somewhat different to that in younger listeners. Although the groups did not differ in overall intelligibility, performance in the older spouse group was more dependent on TMR (Figure 3) and accuracy in the Familiar Target condition decreased as age increased (Figure 5). Older spouses also made significantly more ‘random’ errors (i.e., words not presented in either the target or masker sentences) than did younger spouses and friends. Both the larger influence of TMR on intelligibility, and increased proportion of ‘random’ errors, is consistent with greater energetic masking in this group, which could result from subclinical hearing loss (i.e., in the absence of shifts in audiometric thresholds) that is related to age—for example, due to broader filter widths (see Badri, Siegel, & Wright, 2011). These results could also be due to age-related attentional decline (Alain & Woods, 1999; Godefroy, Roussel, Despretz, Quaglino, & Boucart, 2010), exacerbated by more challenging listening conditions (i.e., lower TMRs).

Regardless of the mechanism, my results suggest that older people gain as much benefit from a familiar voice as younger people, suggesting that real-world speech intelligibility can be improved by voice familiarity. Familiarity with a voice, which could be gained by speaking to a friend in quiet settings, might help to protect against social isolation in

older adults, which has been linked to increased risk of depression (Carabellese *et al*., 1993) and dementia (Lin *et al*., 2013).

2.4.4 Effect of magnitude of difference in F_0 within listeners

As expected, listeners gained a larger intelligibility benefit from a familiar voice (compared to unfamiliar voices) if the F_0 difference between the familiar voice and unfamiliar voices was larger, demonstrating a well-established effect of acoustics on speech intelligibility (Assmann, 1999; Darwin *et al*., 2003; Summers & Leek, 1998). Counterbalancing of voices ensured that, at the group level, the voices in the Familiar Target condition were the same as those in the Familiar Masker condition and in the Both Unfamiliar condition. The voices in each condition were identical in the spouse groups, but because of the six participants who dropped out of Friends group, the voices in each condition were slightly different. Those six voices only served as unfamiliar, and three other voices only served as familiar. Nevertheless, analyses of the familiar-voice benefit also covaried for the F_0 difference between familiar and unfamiliar voices. The finding of a significant familiar-target benefit, even after factoring out influences of the F_0 difference, indicates that familiarity with a voice (as well as acoustic similiarity between it and a competing unfamiliar voice) contributes to its intelligibility.

2.4.5 Masker words less likely to be mistaken for target words in the familiar-target condition

'Wrong voice' errors—in which the response was from the masker sentence—occurred considerably more frequently than 'random' errors. Whereas 'random' errors probably arise because listeners were not able to hear words from the target (energetic masking; Brungart, 2001; Brungart et al., 2001; Durlach, 2006), 'wrong voice' errors mean that the listener could hear at least part of the target-masker mixture adequately, but they reported a word spoken by the masker voice. This type of error may reflect one of several underlying issues; for example, a difficulty segregating the two speech streams, extracting a target signal from a mixture which becomes more challenging at low TMRs, selectively attending to the target, or potentially some other difficulty that would fall under the umbrella of 'informational masking' (Durlach, Mason, Kidd, *et al*., 2003;

Durlach, Mason, Shinn-Cunningham, *et al*., 2003; Kidd, Mason, Richards, Gallun, & Durlach, 2007). Fewer 'wrong voice' errors were made in the Familiar Target condition than in the other two conditions. This demonstration of less interference by the masker in the familiar-target condition is effectively a 'release from informational masking' and recent work suggests that it may result because speech spoken by a familiar talker is more resistant to interference from maskers that are linguistically similar to the target (Holmes & Johnsrude, 2019). The Familiarity condition and Group factors did not interact, suggesting that familiar voices reduced informational masking—or, more specifically, interference from the masker—to a similar extent for spouses' and friends' voices, and for older and younger people.

2.4.6 No evidence for improved familiarity with previously unfamiliar voices

In all three groups, performance in all of the familiarity conditions improved by a similar magnitude between the start and end of the experiment. I attribute this improvement to task-specific learning (e.g., practice effects). Given that participants were already highly familiar with their friend's or spouse's voice before the experiment began, I expected any learning of unfamiliar voices to manifest as a greater improvement for the unfamiliar than familiar voices between the beginning and end of the experiment. Previous studies have shown that voice training (Kreitewolf *et al*., 2017; Nygaard & Pisoni, 1998; Nygaard *et al*., 1994) or brief prior exposure to a voice (e.g., Brungart *et al*., 2001) can improve intelligibility. However, the incidental exposure provided here did not appear to be sufficient to provide talker-specific learning for the unfamiliar voices.

2.4.7 Conclusions and Implications

Prior experience with a voice leads to a considerable improvement in intelligibility when that voice is heard in the presence of competing sounds. The magnitude of this benefit is similar for friends and spouses, implying that intelligibility of speech spoken by a familiar person plateaus after we have known someone as a friend for 1.5–19 years and stays constant despite increased durations of exposure (up to 52 years of marriage). My work, using a restricted set of words and controlling for variability in speech production,

probably underestimates the benefit derivable in real listening conditions when conversing with a friend or partner. Yet, even under these controlled conditions in which listeners must utilize knowledge of voice acoustics to improve intelligibility, the intelligibility benefit gained from hearing a familiar voice as the target is large (10–20%) and is robust across different tasks (BUG and CRM) and across different types of relationship (friends and spouses). These results highlight the robustness of voice familiarity as a cue to enhance intelligibility.

Chapter 3

3 Using spatial release from masking to estimate the magnitude of the familiar-voice intelligibility benefit

3.1 Introduction

Many everyday conversations occur in the presence of background sounds. The ability to separate simultaneous sounds is essential for successful communication, and recognising what one person is saying in the presence of other talkers (termed 'the cocktail party problem'; Cherry, 1953) is a perceptual challenge that has received considerable attention. Much previous work has focused on how similarity or differences in acoustic features—such as spatial location, frequency, timbre, or onset time—contribute to perceptual grouping/segregation of sounds in mixtures (e.g., Brungart *et al*., 2001; Cusack, Deeks, Aikman, & Carlyon, 2004; Darwin *et al*., 2003; Kitterick *et al*., 2010; Singh & Bregman, 1997).

One feature that robustly improves the ability to segregate speech from competing sounds is prior knowledge of the talker's voice (e.g., Holmes, Domingo, & Johnsrude, 2018; Johnsrude *et al*., 2013; Kreitewolf *et al*., 2017; Newman & Evers, 2007; Souza *et al*., 2013). Benefits of voice familiarity on speech-on-speech listening tasks have been established using training paradigms (Levi *et al*., 2011; Nygaard & Pisoni, 1998; Yonan & Sommers, 2000). A large benefit has also been shown using naturally familiar voices, such as those of the participant's spouse or friend (Domingo, Holmes, & Johnsrude, in revision [Chapter 2]; Holmes et al., 2018; Johnsrude *et al*., 2013). A benefit of 2-9 dB is observed when a familiar voice is masked by a single unfamiliar talker at an SNR of -3 to -6 dB, when using a closed-set matrix task such as the "Boston University Gerald" task (Kidd *et al*., 2008) in which all of the sentences are of the form <Name><past tense verb> < number> <adjective> <noun>, where all the words are monosyllables (e.g. " Pat bought five old gloves".)

Despite differences in testing paradigms, the considerable improvement in intelligibility from voice familiarity is commensurate with one of the most thoroughly researched cues known to improve speech intelligibility in multitalker situations—spatial release from

masking (Arbogast, Mason, & Kidd, 2005; Best *et al*., 2006; Best, Mason, & Kidd, 2011; Glyde *et al*., 2015; Kidd *et al*., 2010; Singh, Pichora-Fuller, & Schneider, 2008). Spatial release from masking is the improvement in word report when one or more masker talkers are presented at different spatial locations than a target talker, compared to when they are collocated.

Spatial cues include the "better ear effect" due to head shadow, defined as attending to the ear with a more favourable signal-to-noise ratio (Carlile, 2014) and binaural interaction, in which the auditory system leverages interaural time or level differences between target and maskers (Freyman *et al*., 1999).

The magnitude of spatial release from masking depends in part on the spatial relationship between target and masker stimuli. The symmetrical masker paradigm has a stimulus configuration in which two maskers are presented symmetrically (i.e., one on the left and the other the same distance to the right) about a centrally located target (Brungart & Iyer, 2012; Marrone *et al*., 2008). Unlike other designs that have used asymmetrically configured speech signals (Arbogast, Mason, & Kidd, 2002; Freyman *et al*., 1999; Hawley, Litovsky, & Culling, 2004; Johnstone & Litovsky, 2006), this design controls for head-shadow effects because the SNR is the same in the left and right ears (Brungart & Iyer, 2012). Using symmetrical maskers placed at 90˚ about the target, listeners obtained a spatial release from masking of 4 dB in an open-set sentence identification in a modulated-noise task (Bronkhorst & Plomp, 1992), 6 dB in a closed-set word identification in masking speech task (Yost, 2017), and 12 dB in a closed-set Coordinate Response Measure (CRM) speech-in-speech task (Marrone *et al*., 2008).

The current study aimed to more directly compare the benefits to speech intelligibility of spatial separation and voice familiarity. I used an objective measure (spatial separation between target and maskers) to quantify the familiar-voice benefit to intelligibility. I also examined whether, and how, these acoustic (spatial) and cognitive (familiarity) cues interact with one another. I used the symmetric masker paradigm with spatial separations ranging from 0°–90° in order to compare intelligibility of a personally familiar voice to that of an unfamiliar voice in the presence of an unfamiliar masking talker (producing

two different sentences). The target voice was either familiar, such as the listener's friend or romantic partner, or unfamiliar (the friend or partner of another listener). The two maskers were always different sentences and are spoken by an unfamiliar voice different from the target voice. I measured the magnitude of the familiar-voice benefit to intelligibility, and cast this in terms of the degrees of spatial separation required to produce a benefit of equal magnitude (relative to the collocated condition) when the target voice is unfamiliar. I compared the benefits of voice familiarity and spatial separation on intelligibility at three different TMRs (-3, 0, or 6 dB).

3.2 Method

3.2.1 Participants

Participants were nine pairs of friends, siblings, roommates, or romantic couples, who were naturally familiar with each other's voices. Pairs of participants had known each other for longer than six months (median = 4.7 years, interquartile range [IQR] = 5.7) and reported that they spoke to each other between 3 and 90 hours per week (median = 21 hours, IQR = 18.9). The 18 participants (6 male, 12 female) were 18–33 years of age (median = 20.5 years, IQR = 6.8). Participants were native Canadian English speakers with no known history of speech or hearing impairments. Participants were measured using pure-tone audiometry and had 4-frequency (0.5, 1, 2, and 4 kHz) average hearing thresholds of 20 dB HL or better in each ear.

This experiment was approved by the Non-Medical Research Ethics Board at the University of Western Ontario. Informed consent was obtained from all participants prior to testing.

One pair completed the recording sessions but did not return for the listening task and one participant's data was dropped from the analysis due to experimenter error. Data from the remaining 15 participants were analyzed.

3.2.2 Apparatus

The experiment was conducted in a single-walled sound-attenuating booth (Eckel Industries, Model CL-13 LP MR). Participants sat in a chair facing a 24-inch LCD monitor (either ViewSonic VG2433SMH or Dell G2410t).

Speech stimuli were recorded using a Sennheiser e845-S microphone connected to a Steinberg UR22 mkII sound card (Steinberg Media Technologies) and were delivered binaurally through Grado Labs SR224 headphones. Recordings were made and edited using Audacity (version 2.0.3) software.

3.2.3 Stimuli

Stimuli were sentences from the Boston University Gerald corpus (BUG; Kidd, Best, & Mason, 2008). The sentences in this corpus are of the format "<Name> <past-tense verb> <number> <adjective> <noun>". I used a subset of 480 sentences containing two names ("Bob" and "Pat"), eight verbs ("bought," "sold", "found," "lost," "took," "gave," "held," "saw"), eight numbers ("three", "four", "five", "six", "eight", "nine", "ten"), eight adjectives ("blue", "red", "hot" cold", "big", "small", "old', "new"), and eight nouns ("hats", "bags", "shoes", "socks", "pens" "gloves", "toys", "cards"). An example is "Pat held three blue hats".

Unlike the original corpus in which individual words were recorded in citation form, my participants were recorded speaking complete sentences (480 in total, recorded in mono sound; 44.1 kHz sampling rate). Participants were shown a sentence on the screen and a vertical bar moved across the sentence from left to right (Holmes, 2018). Participants were instructed to read the words in the sentence as the bar moved over them, in an effort to maintain a consistent speaking rate throughout the recording session. All sentences were normalized to the same root mean square (RMS) amplitude and each had a duration of approximately two seconds.

Throughout the experiment, each participant heard sentences spoken by three different talkers. These included one familiar voice—that of the participant's partner—and two unfamiliar voices (who were the familiar voices of other participants). All voices were

presented once as familiar and twice as unfamiliar, except for the three participants whose data were not analysed; their voices were only presented as unfamiliar.

The recorded sentences were presented binaurally over headphones using virtual spatial cues in the azimuth plane. Binaural stimuli were created by convolving the speech signal with anechoic head-related transfer functions (HRTFs) measured on a KEMAR mannequin.

Acoustic stimuli were presented at a comfortable listening level (approximately 67 dB SPL). Across trials, the overall amplitude of the stimuli was roved over a range of 3 dB (in 6 equally spaced levels) to ensure that participants could not use the amplitude of either the target or the masker sentences as a cue to identify the target sentence.

3.2.4 Procedure

On each trial, participants were presented with three sentences. The target sentence, which was presented at 0° azimuth (i.e., in front of the participant), was spoken in one voice. The two masker sentences were spoken in a second (always unfamiliar) voice of the same sex as the target. They were either collocated with the target (i.e., also presented at 0° azimuth), or were separated symmetrically about the target at ± 5, 10, 15, 25, 45, or 90° azimuth. A schematic of stimulus configuration is shown in Figure 7A-B. The target sentence always began with a particular name word (“Bob” in one half of the experiment, “Pat” in the other; order counterbalanced across participants). The two masker sentences began with the other name word. The four remaining words were always different in the three sentences. Participants were asked to identify the four words in the target sentence by clicking the words on a screen (Figure 7C). The response screen was visible throughout the entire experiment.

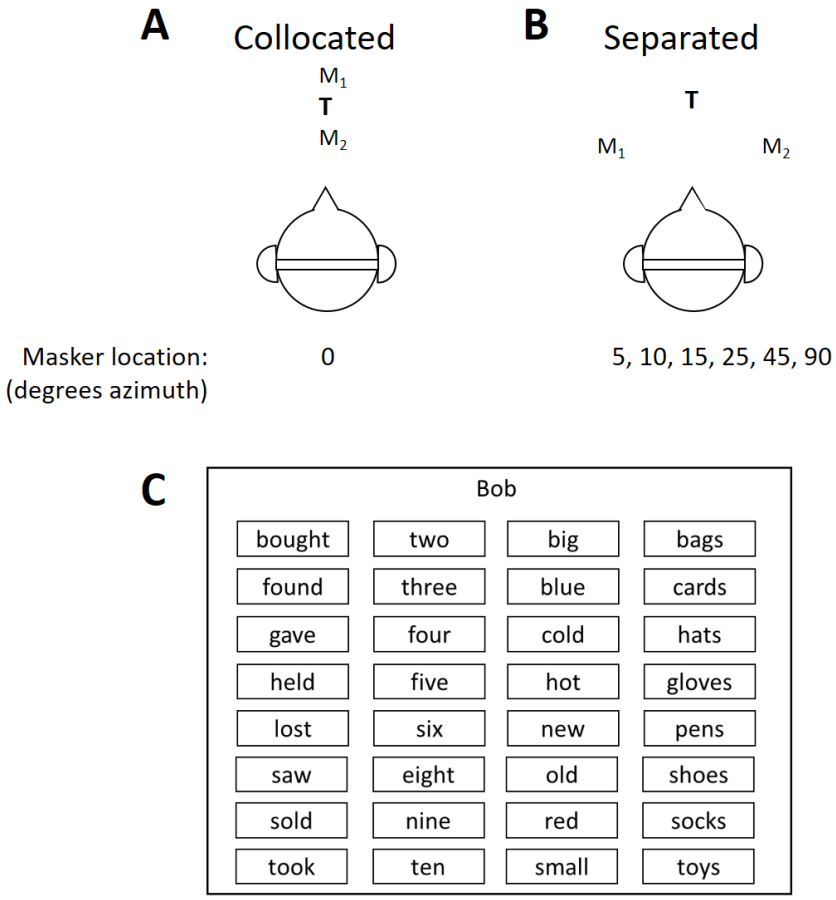


Figure 7. Procedure used in listening sessions. In the collocated condition (A), the target, T, and masker sentences, M_1 and M_2, were played virtually at 0 degrees azimuth. In the spatially separated condition (B), the target was played at 0 degrees azimuth, and the two masker sentences were played symmetrically about the target at 5, 10, 15, 25, 45, and 90 degrees azimuth. Participants tracked the target voice and responded by choosing one word (by a mouse press) from each column on the response screen (C) according to what they had heard in the target sentence, indicated by the target name (in this example, "Bob").

I tested listeners in two familiarity conditions. In the Familiar Target (FT) condition, the target sentence was spoken in the familiar voice, and the two masker sentences were spoken in one of the two unfamiliar voices (half of trials in each of the two unfamiliar

voices). In the Both Unfamiliar (BU) condition, the target was spoken by one of the unfamiliar voices, and the two masker sentences were spoken by the other unfamiliar voice (each unfamiliar voice was the target on half of the trials).

The target and masker sentences were presented at target-to-masker ratios (TMRs) of -3, 0, and 6 dB, defined as the ratio between the target and each individual masker. TMRs were maintained while the overall level of the combined stimuli was roved.

There were 16 trials of each combination of the two familiarity conditions, seven spatial configurations, and three TMRs—producing a total of 672 trials for each participant. Trials were presented in blocks of 48: each condition was presented three times per block in random order. Participants were given the option to take a short break between blocks.

3.2.5 Data analysis

Speech intelligibility was calculated as the proportion of words (out of a possible 64; 4 words in each of the 16 trials) that each participant correctly identified from the target sentence in each condition. Chance performance for each word was 1/8 or 12.5%. These proportions were then normalized into rationalized arcsin units (RAU) (Studebaker, 1985). To determine the effects of voice familiarity and spatial separation on speech intelligibility, I conducted a three-way repeated measures ANOVA on RAU-transformed data, with Familiarity, Spatial Separation, and TMR as within-subjects variables. Mauchly's test indicated that the assumption of sphericity was violated for the main effects of TMR [$\chi^2(2) = 36.4$, $p<.001$] and Spatial Separation [$\chi^2(20) = 51.1$, $p<.001$], and for the interactions between Familiarity and TMR [$\chi^2(2) = 12.9$, $p=.002$], and between Familiarity and Spatial Separation [$\chi^2(20) = 40.71$, $p=.005$]. Thus, these effects are reported with Greenhouse-Geisser correction. Pairwise comparisons are reported with Sidak correction for multiple comparisons.

In order to determine the equivalence point, or the degrees of spatial separation that produces spatial release from masking equivalent to the familiar-voice benefit, I used the lsqcurvefit function on MATLAB R2014b (Mathworks Inc., Natick, MA) to fit the data to the following three-parameter exponential function:

$$y = a(e^{bx}) + c$$

Where a, b, and c are free parameters, and x is spatial separation in degrees.

I then used the function fitted to the Both Unfamiliar data to estimate the degrees of spatial separation that produced an improvement in accuracy equivalent to the average intelligibility in the Familiar Target condition when the maskers were collocated (at 0°). This was done for each TMR separately.

3.3 Results

3.3.1 Familiarity, spatial separation, and TMR affect intelligibility

Figure 8 illustrates the effects of spatial separation and familiarity factors on RAU-transformed proportions of correct words. Intelligibility was significantly better when the target sentence was spoken in the familiar voice (mean = 86.69%, SE = 3.69%) than when it was spoken in the unfamiliar voice (mean = 72.44%, SE = 2.06%) [$F(1, 14) = 23.55$, $p < .001$, $\omega^2 = .58$].

The main effect of Spatial Separation was also significant [$F(2.01, 28.14) = 56.43$, $p < .001$, $\omega^2 = .78$]. Comparing adjacent spatial separation conditions, intelligibility was significantly better for greater spatial separations between 0° and 25° (0–5°: $p = .028$; 5–10°: $p = .04$; 10–15°: $p = .035$; 15–25°: $p = .011$). However, intelligibility did not improve between 25°, 45°, and 90° (ps ≥ .23).

Intelligibility also improved significantly with increasing TMR [$F(1.07, 14.98) = 236.43$, $p < .001$, $\omega^2 = .94$]. Intelligibility was significantly better at 6 dB (mean = 95.77, SE = 2.04) than at 0 dB (mean = 77.44, SE = 2.82) ($p < .001$), and better at 0 dB than at -3 dB (mean = 65.34, SE = 3.22) ($p < .001$).

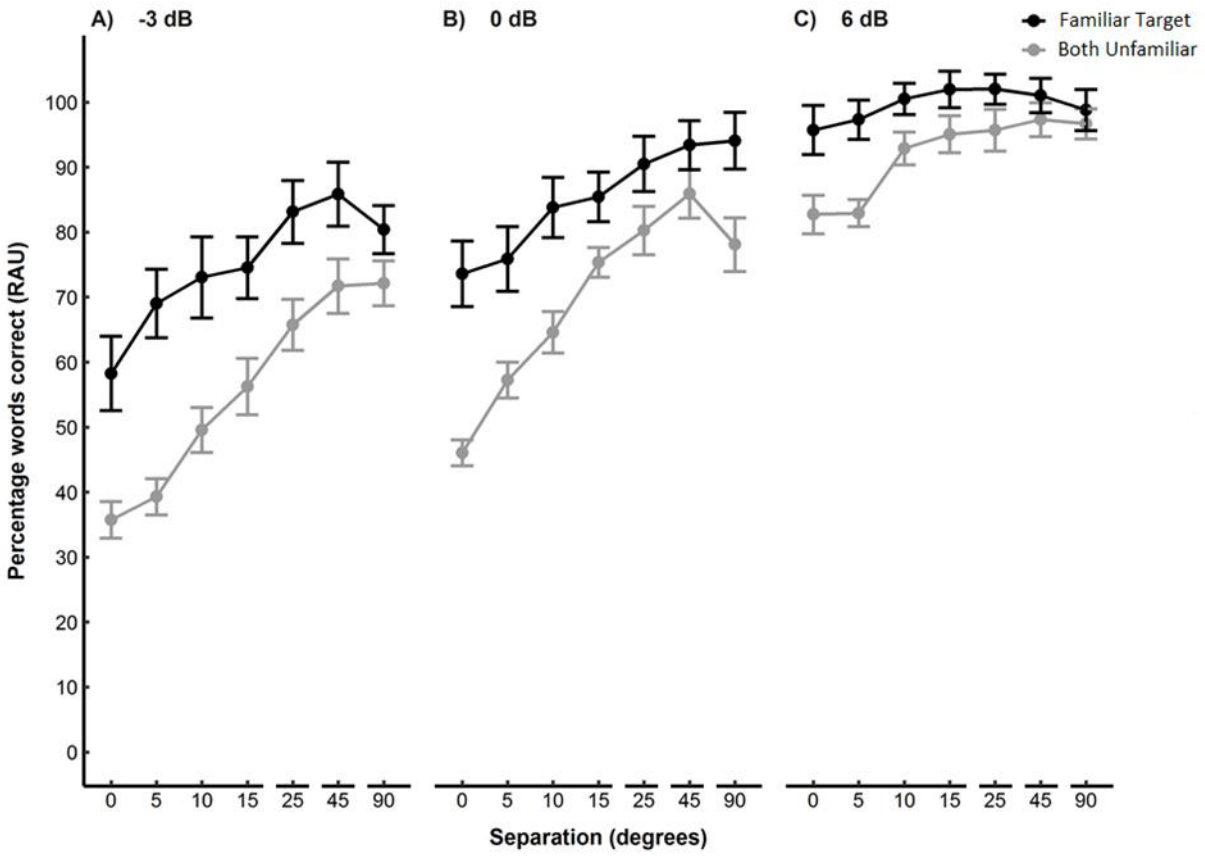


Figure 8. RAU transform of mean percentage of words correct by spatial separation at (A) -3 dB, (B) 0 dB, and (C) 6 dB TMR. The markers represent RAU transformed group mean data for the Familiar Target (black) condition and Both Unfamiliar (grey) condition. Data points show group means. Error bars are ±1 standard error of the mean.

The interaction between TMR and Spatial Separation was significant [$F(12, 168) = 13.05$, $p < .001$, $\omega^2 = .44$], probably due to uniformly high performance in the most favourable TMR condition (+6 dB). At -3 dB and 0 dB TMR, intelligibility at 0° was worse than at all greater separations, intelligibility at 5° and 10° separation was significantly worse than at 45° and 90°), and performance at 15° was worse than at 45° ($4.24 \leq ts(14) \leq 11.25$, $ps \leq .017$. In addition, at -3 dB TMR, intelligibility at 15° was worse than at 90° ($t(14) = 4.81$, $p = .006$). All of these results are consistent with spatial release from masking. Compared to the lower TMRs, at +6 dB, spatial cues had less of an effect on intelligibility. At +6dB TMR, intelligibility at 0° was worse than at 10°, 15°,

and 45° ($3.83 \leq t\text{s}(14) \leq 4.26$, $p\text{s} \leq .038$), intelligibility at 5° was worse than at 45° and 90° ($4.90 \leq t\text{s}(14) \leq 5.28$, $p\text{s} \leq .005$), whereas intelligibility at 15° did not differ from any greater spatial separations ($p\text{s} = 1.00$).

Figure 9 displays the familiar-voice benefit by TMR and spatial separation. The interaction between Familiarity and TMR was significant [$F(2, 28)=8.04$, $p=.008$, $\omega^2 = .31$], again likely because performance was high at +6 dB TMR. The familiar-voice benefit was significantly greater at 0 dB (mean = 15.60%, SE = 3.35) and -3 dB (mean = 19.13%, SE = 4.52) than at 6 dB (mean = 7.71%, SE = 1.54) TMR ($2.95 \leq t\text{s}(14) \leq 3.01$, $p\text{s} \leq .03$).

Familiarity and Spatial Separation also interacted significantly [$F(2.7, 37.85) = 8.37$, $p<.001$, $\omega^2 = .32$], such that talker familiarity was more beneficial at smaller spatial separations. The familiar-voice benefit was larger at 0° than at 15° ($t(14) = 3.71$, $p=.048$) and larger at 5° than at 45° and 90° ($3.74 \leq t\text{s}(14) \leq 3.76$, $p\text{s} \leq .045$).

Lastly, there was a significant three-way interaction between Familiarity, TMR, and Spatial Separation [$F(12, 168)=2.25$, $p=.012$, $\omega^2=.08$]. Again, the generally high performance at 6 dB (see Figure 8C) for even unfamiliar talkers made for weaker effects of familiar voices and spatial separation than in the other two TMR conditions: At 6 dB TMR, the familiar-voice benefit did not differ across spatial separations ($p\text{s} \geq .09$), whereas at -3 and 0 dB TMR, the familiar-voice benefit was greater at small separations than larger separations. This interaction is explained by a significantly greater familiar-voice benefit at 5° than at 45° and 90° at -3 dB TMR ($4.03 \leq t\text{s}(14) \leq 4.30$, $p\text{s} \leq .026$), and at 0° compared to 15° and 45° at 0 dB TMR ($3.85 \leq t\text{s}(14) \leq 4.67$, $p\text{s} \leq .037$).

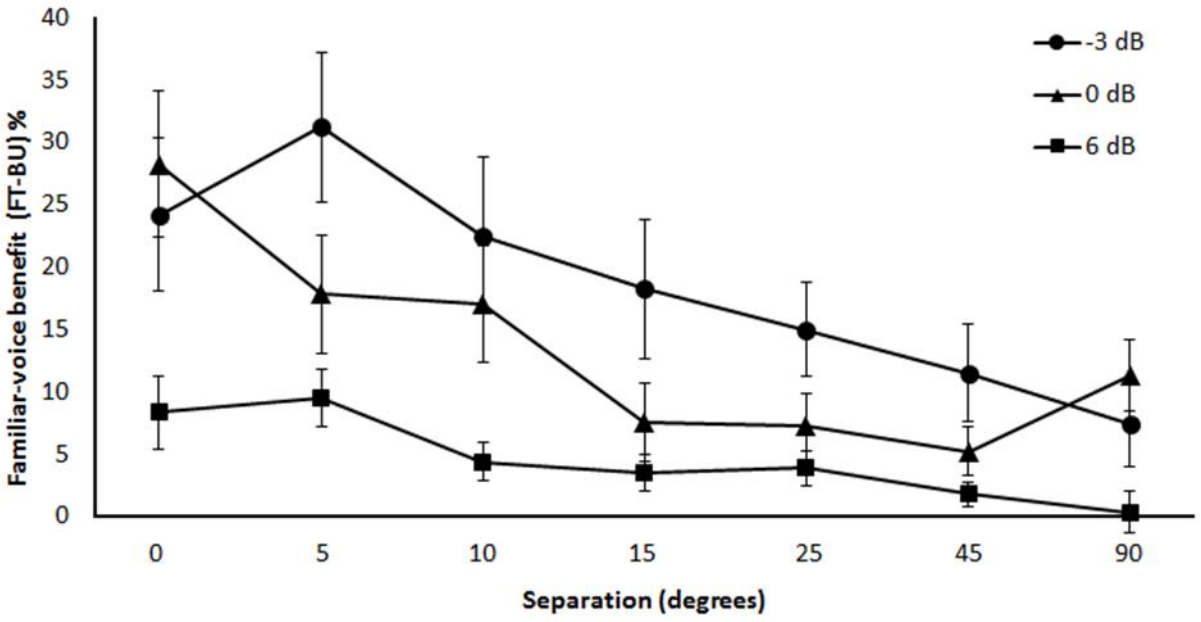


Figure 9. Familiar-voice benefit (difference percentage of correct words identified between the Familiar Target and Both Unfamiliar Condition) at each spatial separation and TMR (-3 dB TMR = circles, 0 dB TMR = triangles, 6 dB TMR = squares). Data points show group means. Error bars are ±1 standard error of the mean. Statistical analyses were based on RAU-transformed data of the familiar-voice benefit.

3.3.2 Sex of listener or his/her familiar voice does not affect intelligibility

I conducted a repeated measures ANOVA as above, and added sex of listener and sex of familiar voice as between-subjects factors and found that neither had a significant effect on intelligibility ($ps \geq .28$), or a significant interaction with familiarity, TMR, or spatial separation ($ps \geq .06$).

3.3.3 Equivalence between familiar-voice benefit and spatial release from masking

Neither benefit from familiarity nor spatial separation is possible when a target that is spoken in an unfamiliar voice, on the midline, is masked by two collocated sentences spoken in another unfamiliar voice. This served as the baseline condition against which to

measure benefits from familiarity and spatial separation. The benefit of a familiar voice was calculated by subtracting intelligibility in the baseline condition from intelligibility in the condition in which the maskers were collocated but the target was familiar.

I then fitted the three-parameter exponential function to averaged Familiar Target (FT) and Both Unfamiliar (BU) data; see Figure 10. The functions provided good fits to the data, with residuals smaller than .045 for each data point. Using the function fitted to the BU data, I then determined the spatial separation that yielded benefit equivalent in magnitude to the familiar-voice benefit (the "equivalence point"), separately at each TMR. At -3 dB TMR, the equivalence point was ±17.1 degrees. At 0 dB TMR, the equivalence point was ±14.6 degrees. At 6 dB TMR, the equivalence point was ±17.0 degrees.

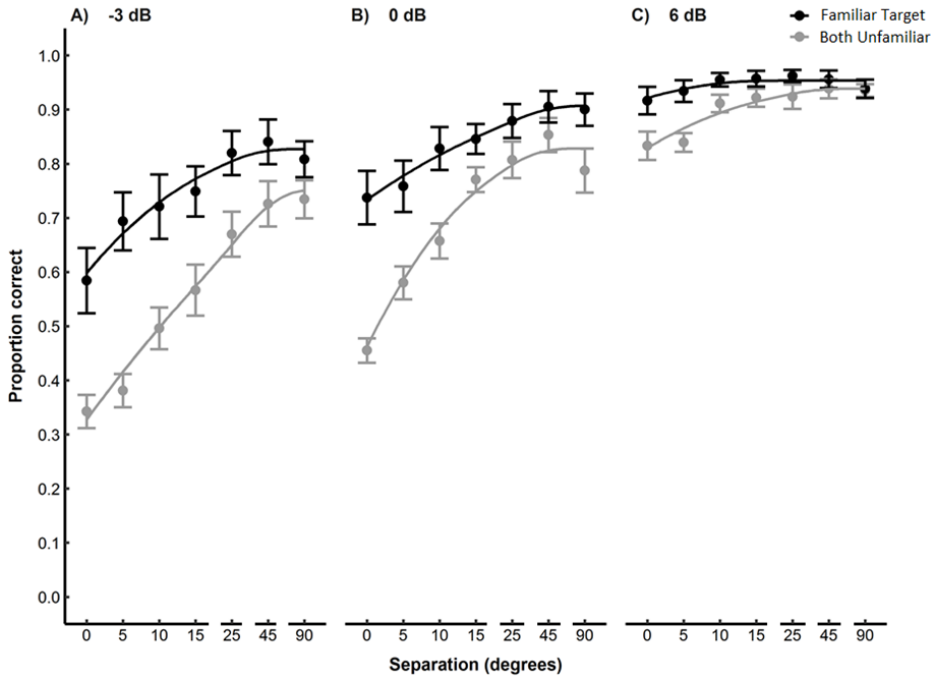


Figure 10. Proportion of correct words as a function of spatial separation at -3 dB (A), 0 dB (B), and 6 dB (C) TMR. The markers represent raw speech intelligibility data in the Familiar Target (black) or Both Unfamiliar (grey) condition. The lines are exponential functions fitted to the raw data. Data points show group means. Error bars are ±1 standard error of the mean.

Next, I quantified the familiar-voice benefit in terms of TMR. When maskers and target were collocated on the midline, participants were 20% more accurate in reporting words spoken by a familiar voice than an unfamiliar voice (averaged across TMRs). In order to quantify this benefit in dB, I fit a linear regression line to the Both Unfamiliar condition when target and masker were collocated at 0° and interpolated the TMR that yields the same accuracy as that in the Familiar Target at -3 dB (collocated). Figure 11 shows the intelligibility in the Familiar Target and Both Unfamiliar conditions at each TMR for collocated (Figure 11A) and ±90° separated (Figure 11B) data. Since I only used 3 TMRs, this is necessarily a rather gross estimate. This is equal to a release from masking of 5.1 dB. When target and maskers were separated by 90°, participants were only 6% more accurate when the target voice was familiar, which is equal to release from masking of 4.4 dB.

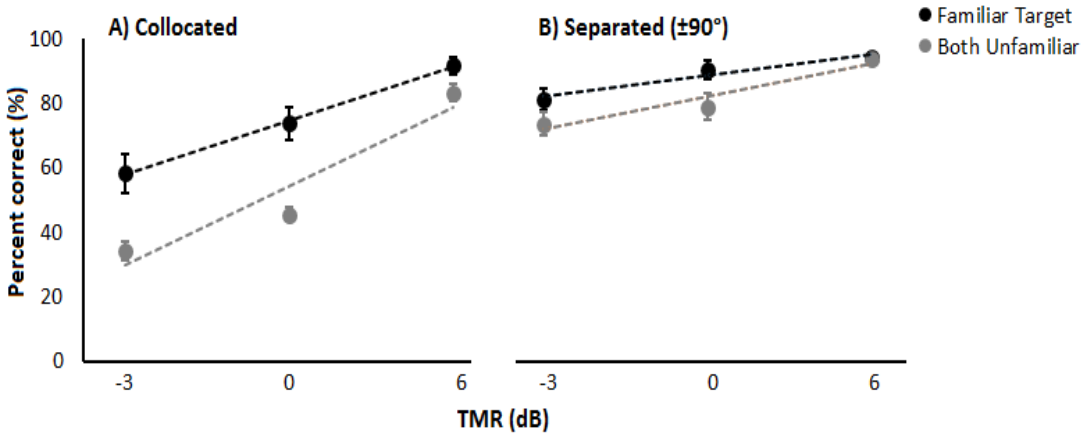


Figure 11. Intelligibility of the Familiar Target (black) and Both Unfamiliar (grey) conditions for (A) Collocated and (B) Spatially Separated data at ±90° as a function of TMR. Dashed lines are the linear regressions for each condition. Data points show group means. Error bars are ±1 standard error of the mean.

3.4 Discussion

My results replicate the familiar-target benefit to intelligibility, consistent with previous studies (Domingo *et al*., in revision [Chapter 2]; Holmes *et al*., 2018; Johnsrude *et al*., 2013; Nygaard & Pisoni, 1998), and extend this by showing a familiar-target benefit in a

three-sentence mixture produced by two voices. When materials were spatially collocated (at 0°) participants correctly reported an average of 20% more words in the Familiar Target than in the Both Unfamiliar condition. These results are highly consistent with previous studies from our laboratory on demographically similar participants, that have found an average improvement in intelligibility of approximately 15% when a familiar, compared to unfamiliar, voice is the target (Domingo *et al*., in revision [Chapter 2]; Holmes *et al*., 2018; Johnsrude *et al*., 2013).

Here, I measured the improvement in intelligibility from a familiar voice to be equivalent to the benefit provided by 14–17°of spatial separation, depending on TMR. In fact, this range of spatial separations produced almost the largest benefit to intelligibility observed in this experiment, because intelligibility plateaued at spatial separations above 15°. Intelligibility at larger separations (25°, 45°, and 90°) were not significantly different from each other (84.4%, 87.0%, and 85.1%, respectively), although they were all significantly better than at 15° (80.2%).

My finding that spatial separation only improves intelligibility up to ±25° is broadly consistent with previous studies showing that spatial release from masking plateaus at large spatial separations. When comparing the intelligibility of a target at 0° in the presence of symmetrically separated speech maskers, the benefit of increasing spatial separation from ±30° to ±90° was only ~0.8 dB (Noble & Perrett, 2002) and 1.5 dB (Yost, 2017). These results are similar to those of Jones & Litovksy (2008) who found that spatial release from masking at 45° accounted for majority of the spatial release from masking observed at 90°, reinforcing the idea that spatial release from masking does not have a linear relationship with spatial separation.

Spatial separations of ±90° have been shown to provide a release from masking up to approximately 4 dB (Bronkhorst & Plomp, 1992), 6 dB (Yost, 2017), and 12 dB (Marrone *et al*., 2008). Findings were influenced by task differences, particularly the number of words participants were required to report. The studies in which listeners reported one word (Yost, 2017) or two words (Marrone *et al*., 2008) showed higher spatial release from masking compared to studies in which listeners were required to

report short sentences (Bronkhorst & Plomp, 1992). The current study also required listeners to report words from a short sentence, with the exception of the first (i.e., Name) word, which was used to identify the target. Using TMRs between -3 and 6 dB, release from masking at ±90° was 4.4 dB, which is highly similar to the findings of Bronkhorst and Plomp (1992).

In a previous study (Marrone *et al*., 2008) that presented symmetric maskers, intelligibility increased with greater spatial separations and reached a maximum at around 45°. Although this is greater than the 25° peak I found in the current study, Marrone *et al*. (2008) did not include any spatial separations between 15° and 45° in their study. It is possible that if a condition at around 25° was included, they may have observed a plateau in intelligibility at that condition. Differences could also be due to task, where Marrone *et al*. (2008) used the CRM corpus (Bolia, Nelson, Ericson, & Simpson, 2000) and I used the BUG task (Kidd *et al*., 2008), but recorded as complete sentences by the participants. The differences could also be due to differences in TMR: Marrone *et al*. (2008) presented stimuli at -5.7 dB and -9.3 dB TMR for 15° and 45° separations, respectively. These TMRs are lower than any used in the current study.

Taken together, Johnsrude *et al*. (2013) and Domingo *et al.* (in revision; [Chapter 2]) found that the release from masking from a collocated familiar target voice ranges from 2 dB to over 9 dB (approximately 10–15% improvement in intelligibility) at TMRs of -3 to -6 dB, suggesting that the release from masking benefit of a familiar voice is commensurate with or even larger than that of a 90° spatial separation reported in previous studies. In the collocated condition of the current study, release from masking benefit of a familiar voice was 5.1 dB (approximately 20% improvement in intelligibility). These results highlight the effectiveness of voice familiarity as a facilitator of intelligibility.

The familiar-voice benefit at smaller spatial separations was significantly larger than at bigger spatial separations (see Figure 8), particularly at low TMRs (-3 dB and 0 dB). This effect cannot be solely attributed to ceiling effects at large spatial separations because I observed the same pattern at the lowest TMR (-3 dB); at this TMR,

intelligibility did not exceed 85%. These results suggest that listeners use voice familiarity to improve intelligibility in challenging listening conditions (i.e., at low spatial separations), but perhaps not as much at higher spatial separations, when acoustic cues are sufficient to identify words in the target sentence.

Voices were counterbalanced so each familiar voice served as the unfamiliar voice for two other participants. At a group level, the acoustics of the voices used as familiar and as unfamiliar voices were therefore identical to each other, and so I focus here exclusively on group level data. Acoustics were not matched at the individual level, therefore investigating individual differences is not possible in the current study. This limitation may be overcome in future research using a training paradigm in which all participants are presented with the same voices, and different subsets of these voices are familiar for different participants.

3.4.1 Conclusion

This paper is the first to directly compare the benefits of voice familiarity and spatial separation on intelligibility. I replicated previous studies showing substantial benefits from both naturally familiar voices and from spatial separation. Moreover, I demonstrated that the familiar-voice benefit is equivalent to spatial release from masking provided by 14-17° of symmetric spatial separation in three-talker listening, and also provided the first data demonstrating a potential trade-off between these cues—my results suggest that individuals rely less on familiar voice information when acoustic cues such as spatial separation and TMR are sufficient to segregate simultaneous speech streams.

Chapter 4

4 Comparing the neural correlates of familiar-voice processing and spatial release from masking

4.1 Introduction

To communicate in noisy environments, listeners must perceptually segregate simultaneous sounds into separate speech streams and attend to a specific stream while tuning out all other competing sounds. Having prior knowledge with a person's voice has been shown to improve intelligibility in noisy environments. Studies from our lab have shown that a familiar target voice is up to 10-20% more intelligible than an unfamiliar target voice in the presence of a single masker (Domingo *et al.*, in revision [Chapter 2]; Holmes *et al.*, 2018; Johnsrude *et al*. 2013). In a symmetric masking paradigm, where two maskers are symmetrically separated by 5-90° from a centrally located target, a familiar voice is 10-30% more intelligible than an unfamiliar voice at -3 dB TMR (Domingo, Holmes, Macpherson, & Johnsrude., in preparation [Chapter 3]). These studies demonstrate the effectiveness of a familiar target voice as a facilitator of intelligibility across different tasks.

In general, voices activate the superior temporal gyri (STG) bilaterally, as well areas in the inferior frontal gyrus (IFG) and postcentral gyrus (Aglieri *et al*., 2018; Pernet *et al*., 2015). Studies investigating how the brain is organized to support recognition and discrimination of familiar voices have produced mixed results. A recent review of clinical research examining face, voice, and name recognition (Barton & Corrow, 2016) suggests that right temporal lobe lesions can selectively impair either voice or face familiarity. In neurotypical participants, activation in response to voice identity tasks as been observed in right anterior temporal lobe (Nakamura *et al*., 2001; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003), in left anterior temporal lobe (Latinus, Crabbe, & Belin, 2011; Nakamura *et al*., 2001), in bilateral middle and inferior temporal lobe (Bethmann *et al*., 2012; Birkett *et al*., 2007; von Kriegstein *et al*., 2005), in prefrontal cortex (Latinus *et al*., 2011; Zäske *et al*., 2017), and lastly, in the precuneus (Nakamura *et al*., 2001; von Kriegstein *et al*., 2003; von Kriegstein & Giraud, 2004). The discrepancies in these

results could be due the nature of the familiar voice used. In some cases, the familiar voice was trained (Latinus *et al*., 2011; von Kriegstein *et al*., 2003; Zäske *et al*., 2017), or was personally known to the participant (Birkett *et al*., 2007; Nakamura *et al*., 2001; von Kriegstein & Giraud, 2004; von Kriegstein *et al*., 2005), or belonged to a famous person with whom participants were familiar (Bethmann *et al*., 2012). Notably, the majority fMRI studies that used personally familiar voices (Birkett *et al*., 2007; von Kriegstein & Giraud, 2004; von Kriegstein *et al*., 2005) and one of the studies that used trained familiar voices (von Kriegstein *et al*., 2003) all found activation for familiar voices in Brodmann Area (BA) 21.

Discrepancies in voice recognition areas could also be attributed to task differences. In some tasks, listeners were trained to associate voices with names and identified the name of the voice that produced a word or vowel sound from a set of options (Latinus *et al.*, 2011). In a similar task, listeners were trained on a set of voices and were tested on whether the voices were 'old' (i.e., part of the training phase) or 'new' (i.e., not presented in the training phase). Other studies required listeners to make explicit familiarity decisions about the stimuli. For example, listeners were asked to discriminate between familiar from unfamiliar voices using a button press (Birkett *et al*. 2007). A more challenging version of this task was used by Bethmann *et al*. (2012) and Nakamura *et al*. (2001) in which listeners not only had to identify if the voice was familiar, but also had to indicate if they could identify the talker by name.

Von Kriegstein and Giraud (2004) suggest that there is functional dissociation between anterior and posterior regions of the right superior temporal sulcus (STS) and may explain why different voice recognition tasks activate different areas of the brain. In the 'voice condition' of this study, listeners were played a target sentence, and pressed one button if the subsequent sentences in the block were spoken by the same voice as the target, and another button if the voice was different. In the 'sentence' condition, listeners were asked to press a button if all the words in the sentences in the block were the same as in the target sentence. Results showed that attending to the voice instead of the words activated the right anterior and posterior STS, but only the posterior STS activated in response to an unfamiliar voice. Researchers therefore conclude that the right anterior

STS is active during voice recognition, and the posterior STS is active in processing non-verbal acoustic form. Interestingly, in this study, neither anterior nor posterior STS activated in response to familiar voices. Instead, the areas that selectively responded to familiar voices were not part of the temporal lobe and included the precuneus, amygdala, and parahippocampal gyrus. This is consistent with other work that shows that the precuneus is involved in recognition processes (Dorfel, Werner, Schaefer, von Kummer, & Karl, 2007).

The differences in activation between anterior and posterior STS have been explored more deeply by Schall, Kiebel, Maess, and von Kriegstein (2014). Overall, the posterior STS showed greater activity for voice recognition than speech recognition, whereas the anterior STS showed the opposite pattern. However, only anterior STS activity was correlated with behavioural performance. Participants who were better at recognizing voices had higher anterior STS activity in the voice recognition task compared to the speech recognition task.

Aside from voice familiarity, another cue that has been shown to produce a considerable benefit to speech intelligibility is the spatial separation between simultaneous speech streams. Spatial cues include interaural time differences (ITDs), interaural level differences (ILDs), and spectral cues. Spatial separations of 90° provide an increase in intelligibility of about 4 dB (Bronkhorst & Plomp, 1992), 5 dB (Noble & Perrett, 2002), 6 dB (Yost, 2017), and 12.6 dB (Marrone *et al*., 2008). Some of these studies also show a comparably large benefit at smaller spatial separations. Specifically, spatial separations of ±30° produced a benefit of 4.2 (Noble & Perrett) and 4.5 dB (Yost, 2017) and separations of ±45° produced a benefit of 5.8 (Yost, 2017) and 12.3 dB (Marrone *et al*., 2008). In comparison, a familiar target voice produces an intelligibility benefit of about 2-9 dB compared to an unfamiliar novel voice at TMRs of -3 to -6 dB (Johnsrude *et al*., 2013; Domingo *et al*. in revision [Chapter 2]). In a symmetrical masker paradigm with TMRs between -3 and 6 dB, the benefit of a familiar target voice is 4.4 dB (Domingo *et al*. in preparation [Chapter 3]). Together, these behavioural results show that a familiar voice produces an intelligibility benefit comparable to 30-90° of spatial separation.

Changes in spatial separation in the azimuth plane between concurrent sounds have been shown to be processed in Heschl's gyrus, planum temporale, and surrounding cortical areas (Shiell *et al*., 2018). Further, changes in sound location are a strong activator of the posterior superior temporal gyrus (STG) (Barrett & Hall, 2006) and planum temporale (Warren & Griffiths, 2003), suggesting that posterior auditory cortex areas are sensitive to spatial information (Ahveninen, Kopco, & Jääskeläinen, 2014). The planum temporale appears to play a functional role computing the sound objects and locations that most likely produced the spectrotemporal pattern represented in the primary auditory cortex (Griffiths & Warren, 2002).

Sound localization tasks also recruit the precuneus (Kryklywy, Macpherson, Greening, & Mitchell, 2013) and other inferior and posterior parietal areas (Arnott *et al*., 2004; Coull & Nobre, 1998; Zimmer, Lewald, Erb, & Karnath, 2006; Zündorf *et al*., 2016). The precuneus has also been implicated in visual-spatial tasks (Wolbers, Hegarty, Buchel, & Loomis, 2008), suggesting that the spatial information processed in the precuneus is not modality specific.

Areas in the temporal lobe appear to be sensitive to intelligible speech, particularly in the left hemisphere. For example, when speech intelligibility is manipulated using varying speech-to-noise ratios, areas of the brain that respond to more intelligible speech include bilateral anterior and posterior temporal regions and the left IFG (Zekveld *et al*., 2006). Davis and Johnsrude (2003) used normal speech, speech segmented with signal-correlated noise, noise-vocoded speech, and speech in noise to create different levels of intelligibility and observed bilateral activation in the STG and MTG in response to intelligible speech but with more widespread activation in the left hemisphere. When listeners were presented with normal, noise-vocoded, and spectrally rotated speech, areas associated with processing intelligible speech were identified in the anterior and posterior left temporal lobe (Narain *et al*., 2003; Scott *et al*., 2000).

Voice familiarity and spatial separations appear to provide large improvements to intelligibility but have been studied separately. Although cortical regions associated with familiar voice recognition and integration of spatial information have been identified,

what remains unclear is if these two features (familiarity and spatial distance from masker) improve intelligibility by recruiting similar areas of the brain. The goal of this experiment is to use fMRI to compare the patterns of neural responses to each feature and determine whether they improve intelligibility using similar mechanisms.

I used a symmetrical masking paradigm to measure the degrees of spatial separation between two unfamiliar voices that produced the same level of intelligibility as a familiar voice for each participant. Although the current experiment also aims to investigate intelligibility areas, it is distinct from the previous work described above because voice familiarity conditions are acoustically nearly identical (across the group). Therefore, any activation observed from contrasting familiar voices with unfamiliar voices is not simply due to acoustic differences between conditions. Then, I presented speech with this spatial separation (as well as collocated speech) in a simplified version of the task used in previous familiarity work in our lab (Domingo *et al.*, in revision [Chapter 2]; Domingo *et al.*, in preparation [Chapter 3]; Holmes *et al*., 2018). In this task, listeners were presented with a sentence on the screen and are asked to respond by button press if the sentence matches the target that they heard. The change in task was primarily due to the requirements of fMRI research, in which participants are required to keep as still as possible and have very limited range of movement while inside the bore. Further, participants had to provide a response within a limited window of time before the next trial began. Therefore, a simplified intelligibility task was necessary to identify and compare brain networks recruited for speech perception facilitated by a familiar voice and facilitated by spatial separation.

4.2 Methods

4.2.1 Participants

Participants were recruited in pairs to ensure natural familiarity with one another's voices. Participant pairs could either be friends, roommates, dating couples, or spouses. Thirty-one individuals (15 females) were recruited to be in this study (15 pairs, plus one participant whose partner provided voice recordings in a previous experiment). Participants had a mean age of 25.64 years (SD = 5.53), had known each other an average

of 6.24 years (SD = 4.02), and reported speaking to each other an average of 19.86 hours a week (SD = 15.37). All participants spoke English fluently without an accent (to a native speaker of southwestern Ontario English) and had no known speech or neurological impairments.

One participant was excluded due to abnormal hearing in the left ear. All other participants had normal hearing with a pure-tone average hearing threshold below 25 dB SPL in both ears. Two participants failed to return for the behavioural-only session, three participants did not complete the fMRI session due to light-headedness and claustrophobia, and two participants had incomplete data due to technical errors. Complete datasets were obtained from 23 participants.

This project was approved by the Western University Non-Medical Research Ethics Board. Informed consent was given by all participants before proceeding with the experiment.

4.2.2 Apparatus

The speech recordings and preliminary behavioural session were conducted in a single-wall sound attenuating booth (Eckel Industries, Model CL-13 LP MR). Participants sat in a chair facing a 24-inch LCD monitor (either ViewSonic VG2433SMH or Dell G2410t).

Speech stimuli were recorded using a Sennheiser e845-S microphone connected to a Steinberg UR22 mkII sound card (Steinberg Media Technologies) and were delivered binaurally through Grado Labs SR224 headphones. Recordings were made using Audacity (version 2.0.3).

In the preliminary behavioural-only session, auditory and visual stimuli were presented using an in-house script written in MATLAB R2014b (Mathworks Inc., Natick, MA) over Grado Labs SR25 headphones. For the fMRI session, auditory stimuli was delivered using MRI-compatible Sensimetric Insert Earphones. Stimuli in the fMRI session were generated and presented using a modified version of the same script as was used in the behavioural-only session.

4.2.3 Stimuli

Sentence stimuli for this experiment were taken from the Boston University Gerald corpus (BUG; Kidd, Best, & Mason, 2008). The sentences in this corpus follow the format <Name> <verb><number> <adjective><noun>. A subset of 480 sentences were constructed following this format. Table 1 (in Chapter 1) contains all possible options for each word type. An example of a sentence used in this experiment is “Pat held three blue hats”. Sentences were recorded by each participant in mono sound at a 44 kHz sampling rate. Each sentence had an average duration of 2.5 seconds and was normalized to the same root mean square (RMS) amplitude.

These stimuli were processed to create binaural signals containing virtual spatial cues and presented over headphones. Binaural stimuli were created by convolving the speech signal with anechoic head-related transfer functions (HRTFs) measured on a KEMAR mannequin. The binaural stimuli were then added to simulate speech originating from their assigned locations in space. Spatial locations were manipulated in azimuth only.

The relative target and masker intensities were set to a target-to-masker ratio (TMR) of -3 dB, defined as the ratio between the target and each individual masker. A previous study showed that participants demonstrated a familiar-voice benefit to intelligibility at this TMR without reaching ceiling (Domingo *et al.*, in preparation [Chapter 3]). Acoustic stimuli were presented at a comfortable listening level-- approximately 67 dB SPL – but roved over a range of 3 dB (in 6 equally spaced levels) to ensure that participants could not use the amplitude of either sentence as a cue to identify the target sentence.

4.2.4 Experimental procedure

4.2.4.1 Behavioural-only session: Determining the magnitude of the familiar-voice benefit for each participant

The purpose of this session was to measure the degrees of spatial separation that provided an equivalent intelligibility benefit as a familiar voice for each participant. The spatial separation value determined for each participant from this session was used in their fMRI session.

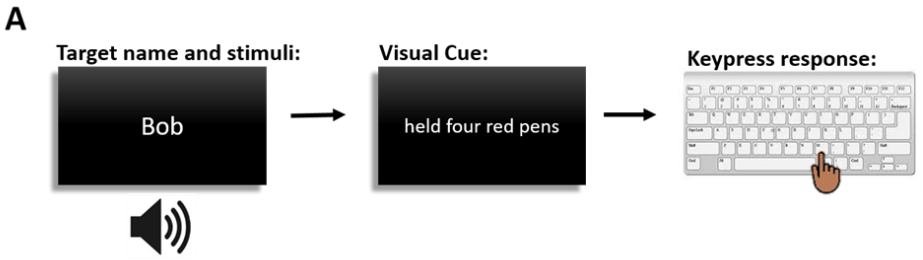
Participants returned approximately two weeks after the recording session (mean days of after recording session = 17 days, SD = 11). The sentences recorded in the first session served as auditory stimuli for this task. Participants were seated in a soundbooth and presented with three sentences played simultaneously. Participants were instructed to attend to the sentence beginning with the name shown on the screen. After the sentences were played, the remainder of the sentence (<verb> <number> <adjective> <noun>) was presented on the screen, and participants decided whether the sentence on the screen matched the sentence in the auditory mixture that began with the word shown on the screen earlier in the trial. The visual cue was a match for 50% of the trials and a mismatch for the remaining 50% of trials, and were presented in random order. The matched visual cue contained all four words from the target sentence. The mismatched visual cue was constructed such that three out of the four words were from the target sentence, and one word was from either masker sentence. Participants indicated that the visual cue sentence was a match or not a match by giving a keypress response. The visual cue remained on the screen until a response was given.

This task was self-paced, so trial lengths and experimental duration varied between participants. The next trial began immediately after the keypress response was received. Figure 12A shows a schematic of the behavioural task.

Speech stimuli were presented in two voice conditions: (1) Familiar Target (FT), in which the target sentence was spoken by the participant's partner, and the two (different) masker sentences were spoken by the same unfamiliar talker (the familiar voice for another participant in the study), and (2) Both Unfamiliar (BU), in which the target sentence, and the two masker sentences, were spoken by two different unfamiliar talkers.

The virtual spatial locations of the auditory stimuli followed a symmetrical masker paradigm, in which the target was always presented at 0° azimuth, and the maskers were presented either at 0° (i.e., collocated with the target), or symmetrically separated about the target at 5°, 10°, 15°, 20°, 25°, or 45° degrees on either side of the target (7 spatial conditions in total). Participants were not informed of the target location and were only cued by the target name. Because the goal of this session was to determine the degrees of

spatial separation that produces the same intelligibility benefit as a familiar voice (without any other cues), stimuli in the Familiar Target condition were always presented at 0° (collocated) only. Figure 12B shows the different conditions used in this session.



B

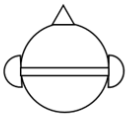
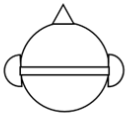
	Collocated	Separated
Stimulus configuration	M_1**T**M_2 	**T** M_1 M_2 
Degrees separation	0	5, 10, 15, 20, 25, 45
Familiarity condition	Familiar Target Both Unfamiliar	Both Unfamiliar

Figure 12. Schematic of task (A) and experimental design (B) for behavioural-only session.

There were a total of eight conditions in this session: Familiar Target at 0° only, and Both Unfamiliar at all seven spatial conditions. Participants completed a total of 30 trials in each condition, totaling 240 trials. Each of the two unfamiliar voices were presented an equal number of times in each condition. For example, in the one Familiar Target (collocated) condition, one unfamiliar voice was used as the masker for 15 trials and a different unfamiliar voice was used as the masker for the remaining 15 trials. Participants were invited to take a short break after the first 120 trials.

From this session, the degrees of spatial separation that provides an equivalent release from masking as a (collocated) familiar voice was determined for each participant. These

individual spatial separation values were used in the 'spatially separated' condition in the fMRI session.

4.2.4.2 fMRI data acquisition

MRI data was acquired using a Siemens 3T Magnetom Prisma Fit at the Center for Metabolic Mapping at the Robarts Research Institute at Western University using a 32-channel head coil. Functional scans consisted of eight multiband echo-planar runs of 89 volumes each. Echo planar imaging (EPI) data were acquired using an interleaved silent steady state (ISSS) acquisition sequence (Schwarzbauer *et al*., 2005) to maintain T1-related signal during volume acquisition and temporal resolution within trials and to avoid further masking sentence stimuli with scanner noise. Each TR consisted of two 1000 ms scans followed by 7000 ms of silence, during which sentence stimuli were presented. During the silent period, seven dummy scans were done (TR = 1000 ms) consisting of relatively silent slice-selective excitation pulses to maintain longitudinal magnetization at a steady state. During each TR, 51 slices were acquired in oblique orientation with a spatial resolution of 2.7 mm^2. For two participants, images were acquired using a 64-channel head coil due to technical difficulties with the 32-channel coil. Anatomical MPRAGE scans covering the whole brain were acquired for image coregistration and normalization (1 mm^3 voxels, 176 slices, TR = 2300 ms, TE = 2.98 ms).

4.2.4.3 fMRI session: Identifying regions of the brain that are sensitive to release from masking from familiar voices and spatial cues

The intelligibility task for the fMRI session was the same as behavioural-only session, except only two spatial conditions were used: collocated and spatially separated. The maskers in the spatially separated conditions were presented at the spatial separation that was determined in the behavioural-only session to produce equal accuracy as the Familiar Target condition. A Familiar Masker condition, in which the two masker sentences were spoken by the participant's partner, and the target voice was unfamiliar to the participant (but was the familiar voice of another participant), was also included in this session, in

order to ensure that participants attend to the target sentence and not simply to a familiar voice when it was present.

The same speech recordings used in the behavioural-only session were used in the fMRI session but were divided into six Speech conditions: (1) collocated Familiar Target (FTcoll), (2) spatially separated Familiar Target (FTsep), (3) collocated Both Unfamiliar (BUcoll), (4) spatially separated Both Unfamiliar (BUsep), (5) collocated Familiar Masker (FMcoll), and (6) spatially separated Familiar Masker (FMsep). Additionally, there were two control conditions: Silence, in which no auditory stimuli were presented, and Noise, in which completely unintelligible signal-correlated noise (SCN), derived from the three-sentence mixture, was presented to the participant. SCN was created with an in-house MATLAB script. First, a noise signal was generated to have the same longterm spectral profile as the three-sentence mixture, and was convolved with the amplitude envelope. Noise trials had the same amplitude envelope as the original three-sentence mixture, but were entirely unintelligible. In both of these conditions, the condition name (e.g., Silence) was presented on the screen instead of target names (Bob or Pat). Instead of a sentence from the BUG task, the visual prompt “Press any button” appeared, and participants pressed either button as a response instead of making a match or no-match decision. Participants were free to choose which of the two buttons to press.

Two volumes were acquired at the end of every trial, involving acoustic noise being generated by the gradient coils. After the volumes were acquired, a delay of seven seconds occurred before the next two volumes. During this delay, stimuli were presented and a silent series of excitation pulses were delivered to ensure constant signal contrast. After stimulus offset, participants were given a window of approximately four seconds to respond before the next trial began. Each trial lasted for 9 seconds, including scanning. Figure 13 shows a schematic the timing of visual cues and auditory stimuli for each trial.

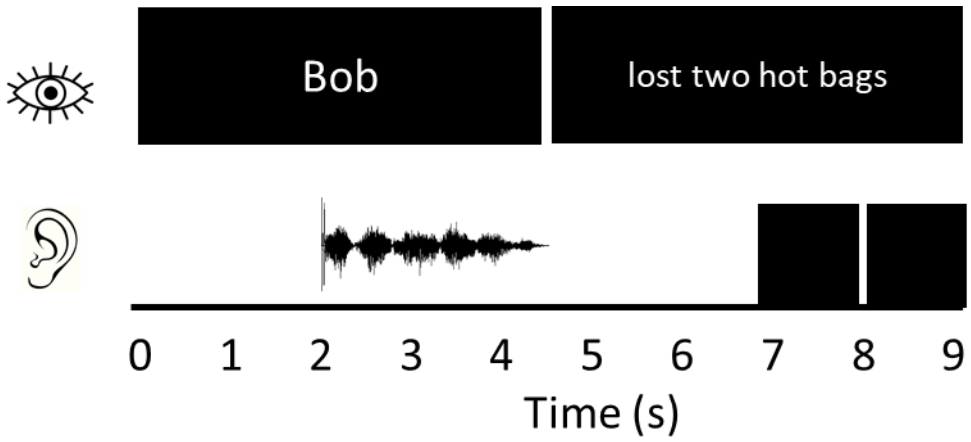


Figure 13. Schematic of trial timing of visual cues (top row) and auditory stimuli (bottom row).

Participants completed eight experimental runs of 42 trials each (336 trials total). There were 21 trials each of Silence and Noise. The Familiar Target and Familiar Masker conditions each had 84 trials (42 were collocated, 42 were spatially separated). In the Both Unfamiliar condition there were 84 collocated trials and 42 spatially separated trials. There were twice as many BUcoll trials (84 trials) in order to use half (42 BUcoll trials) as the baseline condition to compare familiarity effects, and the other half as the baseline condition to compare spatial cue effects.

After participants completed the fMRI task, they were asked to respond to the question "How strongly did you perceive the sentences as coming from different directions?" on a 7-point Likert-type scale (1= not at all, 4=moderately, 7=very strongly).

4.2.5 fMRI preprocessing

Preprocessing and analysis of the fMRI data was conducted using the SPM12 software package (http://www.fil.ion.ucl.ac.uk/spm) at both individual and group levels (modelling subject as a random effect). Preprocessing of the data involved motion correction, coregistration, normalization into standard Montreal Neurological Institute (MNI) template space, and spatial smoothing with a 10 mm full-width half-maximum Gaussian kernel to reduce influence of individual differences in anatomy and to ensure that the data conform to assumptions of Gaussian Random Fields Theory which is used to

apply familywise error correction (FWE) for multiple corrections (Worsley, Evans, Marrett, & Neelin, 1992). Because two volumes were acquired at the end of each trial, I calculated the mean of the two images for each trial and entered those into the first level analysis. I also averaged the realignment parameters for each trial and used the averaged values as regressors.

4.2.6 Behavioural data analysis

4.2.6.1 Behavioural-only session

A single-samples t-test was conducted to determine whether accuracy in each condition was significantly different from chance. To test the effect of a familiar voice on discrimination sensitivity, I conducted a one-tailed paired-samples t-test between collocated FT and BU comparing d' (with loglinear correction to avoid infinite d' values due to extreme false-alarm or hit proportions (Hautus, 1995)) as the dependent measure. Lastly, I conducted a repeated-measures ANOVA on Both Unfamiliar data with spatial separation as the within-subjects factor (7 levels: 0°, 5°, 10°, 15°, 20°, 25°, 45°).

4.2.6.1.1 Magnitude of familiar-voice benefit

The primary goal of the behavioural-only session was to determine the amount of spatial separation in the Both Unfamiliar condition that is associated with the same intelligibility as the collocated Familiar Target condition. This value quantifies the familiar-voice benefit to intelligibility in terms of degrees of spatial separation, as in Domingo *et al*. (in preparation [Chapter 3]). Using the lsqcurvefit function of MATLAB R2014b (Mathworks Inc., Natick, MA), the following three-parameter exponential function was fit to the raw data in the Both Unfamiliar condition and the spatial separation that produced equal intelligibility as the collocated Familiar Target condition was calculated for each participant:

$$y = a(e^{bx}) + c$$

Where a, b, and c are free parameters, and x is spatial separation in degrees.

If the participant did not have a quantifiable spatial separation value because the Familiar Target condition was less intelligible than the collocated Both Unfamiliar condition or because the participant did not show any spatial release from masking, then a spatial separation of 17.1° was assigned to the participant. This value was observed in previous work to be the average spatial separation that yielded an intelligibility benefit for an unfamiliar voice equal to that obtained from a familiar target voice in the collocated condition (Domingo *et al.*, in preparation [Chapter 3]). Because data in the Familiar Target condition was only collected at 0° (i.e., collocated condition), a function was not fit to these data.

4.2.6.2 fMRI session

A within-subjects repeated measures ANOVA was conducted on accuracy data with Familiarity (three levels: Familiar Target, Familiar Masker, Both Unfamiliar) and Separation (two levels: collocated, separated) as within-subjects factors. The same analysis was also conducted for d' values. Lastly, a single-sample t-test was conducted to determine in which conditions d' values were significantly different from chance.

4.2.7 Imaging analysis

Analysis of each participant's data was conducted using a general linear model in which each scan was coded as belonging to one of eight conditions (Silence, Noise, and six Speech conditions). The eight functional runs were modeled as a single session within the design matrix, and eight regressors were entered to remove the mean signal from each run. Six realignment parameters (averaged between the two volumes for each trial) were included to account for motion effects. These models were then fit using the least-mean-squares method to each individual's data and parameter estimates were obtained. Contrast images for each of the eight experimental conditions, as well as the following contrasts were created. Table 3 shows the contrasts that were created:

Table 3. Contrasts and interactions.

Contrast	Conditions
Sound > Silence	(FTcoll + FTsep + FMcoll + FMsep + BUcoll + BUsep + Noise) – Silence
Speech > Noise	(FTcoll + FTsep + FMcoll + FMsep + BUcoll + BUsep) – Noise
FT > BU	(FTcoll + FTsep) – (BUcoll + BUsep)
Sep > Coll	(FTsep + FMsep + BUsep) – (FTcoll + FMcoll + BUcoll)
FM > BU	(FMcoll + FMsep) – (BUcoll + BUsep)
(FT+FM) > BU	(FTcoll + FTsep + FMcoll + FMsep) – (BUcoll + BUsep)
Interaction	
Familiarity x Separation	(FTcoll - FTsep) – (BUcoll - BUsep); (FTsep - FTcoll) - (BUsep - BUcoll)

For the group analysis, the above contrast images were entered into a single-sample t-test. The group analysis was conducted on whole-brain data, with the exception of the FT>BU contrast where I had an a priori hypothesis of where activation would occur. Based on the findings of (Birkett *et al*., 2007; von Kriegstein *et al*., 2003, 2005; von Kriegstein & Giraud, 2004), I conducted a small-volume analysis within BA21. The BA21 mask was created using the PickAtlas tool (Maldjian, Laurienti, & Burdette, 2004; Maldjian, Laurienti, Burdette, & Kraft, 2003) in SPM12. Peaks were localized using the Automated Anatomical Labeling atlas (AAL) (Tzourio-Mazoyer *et al*., 2002). Results of the group analyses were considered significant if they exceeded a threshold of $p <. 05$, family-wise error (FWE) corrected and are shown on the average normalized T1-weighted structural image using the Mango software package (http://ric.uthscsa.edu/mango).

4.3 Results

4.3.1 Behavioural-only session

4.3.1.1 Signal detection analysis

Figure 14A shows d' data in Familiar Target and Both Unfamiliar conditions at each spatial separation. Single-sample t-tests revealed that d' values in all conditions were

significantly above chance (ps<.045) except for in the Both Unfamiliar condition at 0° (p=.22).

d' in the Familiar Target and Both Unfamiliar conditions differed in the predicted direction indicating that familiar voice cues improved sensitivity to the task, [$t(22)$ = 1.85, p=.039 one tailed].

Lastly, the main effect of spatial separation on d' was significant, [$F(6,132)$ = 6.87, p<.0001]. d' at greater spatial separations was larger than at smaller spatial separations. d' at 45° was greater than at 0° and 10°, (ts(22) ≥ 3.52, ps ≤ .047), approached significance at 5°, (p=.06), but was not significantly different from separations between 15-25°, (ps ≥ .99). d' at 0° did not significantly differ from that at 5-10°, (ps ≤ .20).

4.3.1.2 Magnitude of familiar-voice benefit

Magnitude of familiar-voice benefit. I measured the magnitude of the familiar-voice benefit – in terms of degrees of spatial separation – at which an unfamiliar voice was as intelligible as a familiar voice when both are masked by two sentences spoken by an unfamiliar voice. Averaged intelligibility data, expressed as proportion correct, and the fitted function is shown in Figure 14B. Of the 23 participants in this study, 11 did not show an intelligibility benefit from familiar voices. The average familiar-voice benefit for the remaining 12 participants was 12.05°, which is lower than magnitude of 17.1° found in previous work (Domingo *et al.,* in preparation [Chapter 3]). Furthermore, there is a higher proportion of participants who did not gain a benefit to intelligibility, probably due to differences in the behavioural task used. I will explore this further in the Discussion.

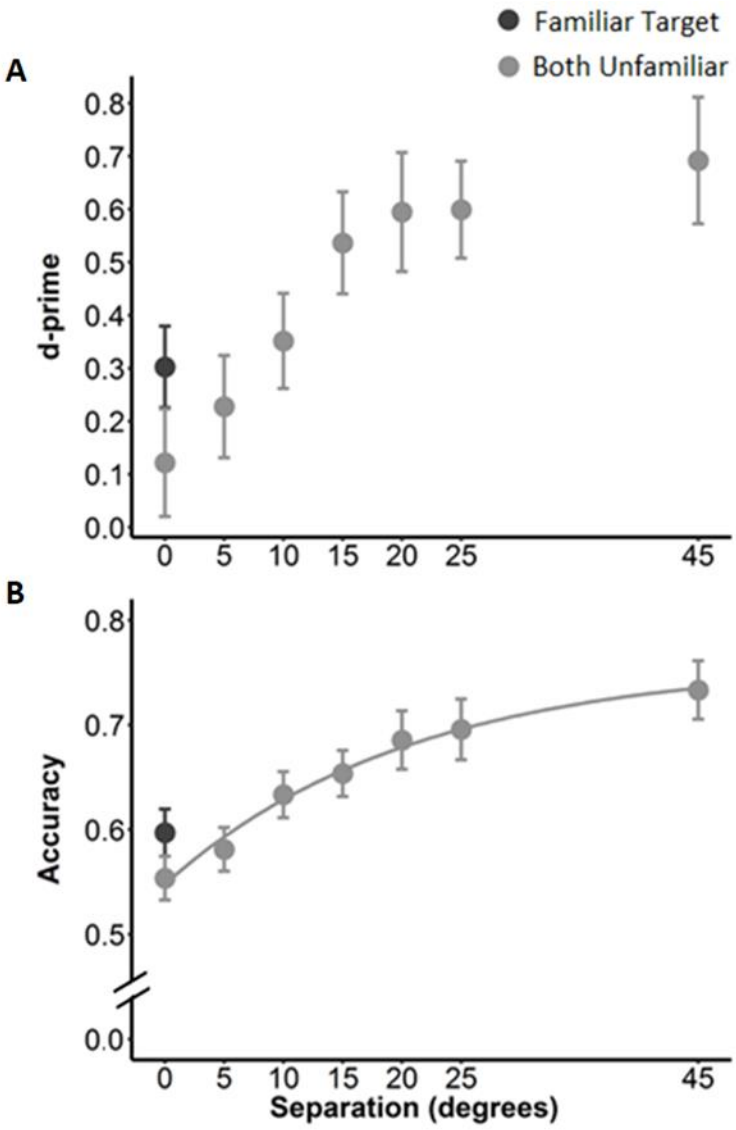


Figure 14. Sensitivity data (A) and accuracy data in proportion correct (B) from the behavioural-only session in each spatial separation for the Familiar Target (FT; dark grey) and Both Unfamiliar (BU; light grey) conditions. The grey line is the fitted function. Data points show group means. Error bars are ±1 standard error of the mean (between subjects). Average accuracy in each condition was calculated for each participant, and the standard error of the mean was calculated from the averaged accuracy across all participants.

4.3.2 fMRI session: Behavioural task

4.3.2.1 Perception of spatialized stimuli

Figure 15 shows each participant's familiar-voice benefit in degrees as a function of their response to the question "How strongly did you perceive the sentence as coming from different directions?" Overall, participants responded that they were able to perceive the fMRI stimuli as coming from different directions (mean = 4.78, SD = 1.38). 19 out of 23 participants responded that the spatialized stimuli were 'moderately perceivable' to 'very strongly' perceivable.

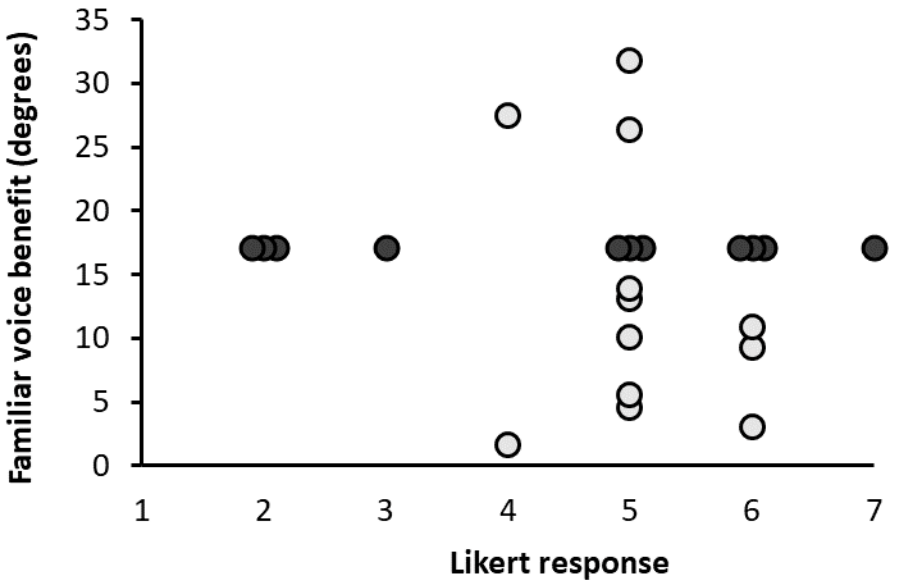


Figure 15. Familiar-voice benefit (measured in degrees) as a function of how strongly each participant perceived the stimuli as coming from different directions (1 = not at all; 4 = moderately; 7 = very strongly). Light grey circles represent participants who showed a familiar-voice benefit during pilot testing. Dark grey circles represent participants who did not show a familiar-voice benefit and were therefore presented with maskers separated by 17.1° during scanning.

4.3.2.2 Accuracy

A 3x2 repeated measures ANOVA was conducted to compare accuracy across familiarity conditions (three levels: Familiar Target, Familiar Masker, Both Unfamiliar; within-subjects) and spatial conditions (two levels: collocated, separated; within-subjects). Figure 16A shows accuracy in each condition for the fMRI task. As expected, accuracy was highest when participants had both familiarity and spatial cues to work with, and was lowest when the familiar voice served as the masker. The main effect of Familiarity was significant, [$F(2, 44) = 7.26$, $p = .002$, $\omega^2 = .21$], indicating that a participant's ability to correctly identify a match or mismatch was affected by the presence of a familiar voice. Participants were significantly more accurate when the target voice was familiar (Familiar Target condition, mean = .57, SE = .02) than when the target was unfamiliar and the masker was familiar (Familiar Masker condition, mean = .51, SE = .01) (p=.020). Participants were also more accurate in the Both Unfamiliar condition (mean = .57, SE = .02) than in the Familiar Masker condition (p=.022). There was no difference in accuracy between the Familiar Target and Both Unfamiliar conditions (p=.961).

Spatial separation had a significant effect on intelligibility, [$F(1, 22) = 9.74$, $p = .005$, $\omega^2 = .27$]. Trials in which the target and masker were spatially separated from one another (Separated condition; mean = .57, SE = .01) were easier to identify as a match or mismatch compared to trials in which the target and masker voices were presented at 0° (Collocated condition; mean = .53, SE = .01).

The interaction between Familiarity and spatial separation was not significant, [$F(1, 22) = 2.50$, $p = .10$, $\omega^2 = .06$].

4.3.2.3 Signal detection analysis

d' values were analyzed for Familiarity and Separation (same levels as above). Results are presented in Figure 16B. As indicated by single-sample t-tests, performance was above chance in the Both Unfamiliar condition when target and maskers were spatially separated and when they were collocated (ps ≤ .015). In the Familiar Target condition, performance was above chance when stimuli were separated (p=.001) but not when they

were collocated (p=.14). Performance in the Familiar masker condition was not above chance either when stimuli were separated or collocated (ps>.49).

Sensitivity (d') was affected by Familiarity, [F(2, 42)=5.99, p=.005, ω^2 = .18]. Sensitivity was higher in the Familiar Target condition (mean=0.24, SE=0.07) and Both Unfamiliar condition (mean = 0.23, SE = 0.06) than in the Familiar Masker condition (mean = 0.03, SE = 0.04) (ps≤.036). There were no significant differences between the Familiar Target and Both Unfamiliar conditions (p=.99).

Spatial separation did not have an effect on discrimination sensitivity, [F(1, 21) = 1.25, p = .28, ω^2 = .01]. There was also no interaction between Familiarity and Separation, [F(2, 46) = 1.62, p = .213, ω^2 = .02].

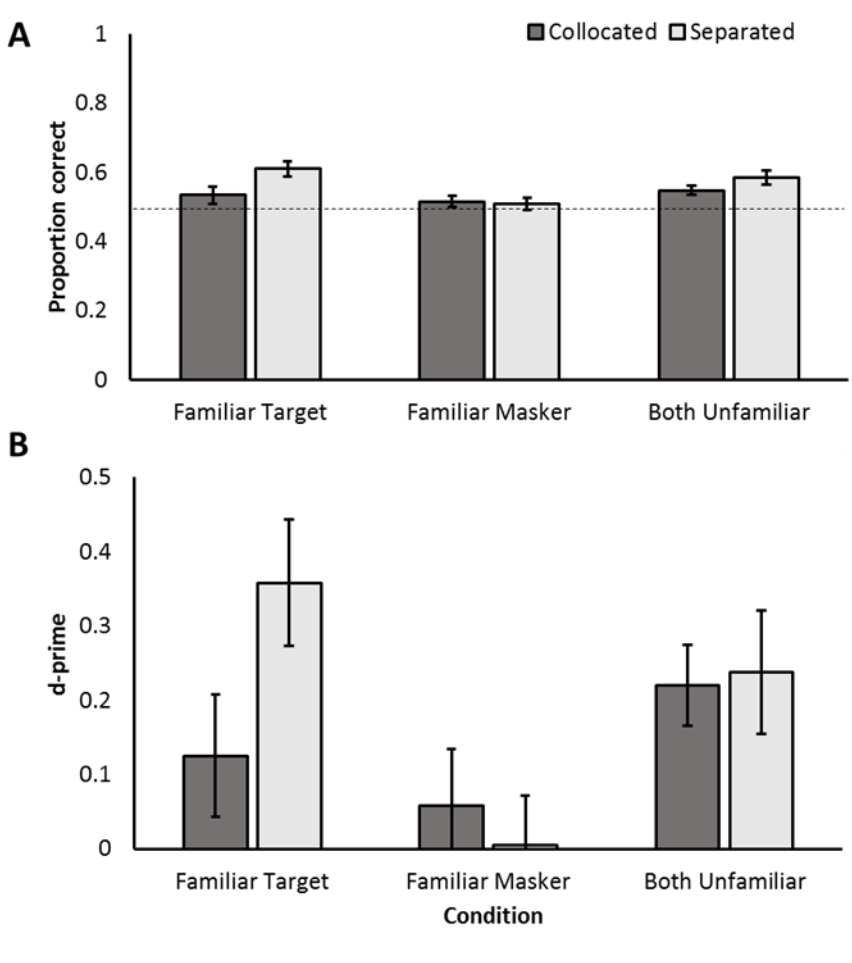


Figure 16. (A) Accuracy expressed as proportion correct and (B) Sensitivity for the behavioural task in the fMRI session by condition for Collocated (dark grey) and Separated (light grey) trials. Error bars are ±1 standard error of the mean. Dotted line in Panel A indicates chance performance.

4.3.3 Functional imaging results

4.3.3.1 Auditory and speech perception

When speech-in-speech stimuli were presented, participants had to decide whether the visually presented sentence was an identical match to the target they had heard. For Silence and Noise trials, participants had to press any button when prompted. The conditions in which sounds were presented (speech and noise trials) activated the bilateral middle and superior temporal areas as well as premotor and prefrontal areas when

contrasted against Silence (contrast: Sound>Silence; Table 4, Figure 17A). When contrasted against noise, speech trials activated the left middle temporal region, premotor areas, and prefrontal areas (contrast: Speech>Noise; Table 4, Figure 17B).

4.3.3.2 Familiar-voice sensitive areas

A paired t-test revealed that the activation between (FT>BU) and (FM>BU) did not differ, therefore the results from all trials containing a familiar voice were collapsed together. I contrasted all trials with a familiar voice present against trials where the target and masker voices were unfamiliar (contrast: (FT+FM)>BU; Table 5, Figure 18) and observed significant clusters in the right posterior STG, left supramarginal gyrus, precuneus, and right IFG.

4.3.3.3 Spatial-cue sensitive areas

Spatially separated trials activated areas in posterior superior and middle temporal areas and precuneus when contrasted against collocated trials (contrast: Separated>Collocated; Table 4, Figure 18).

The interaction between familiarity and spatial cues was not significant.

Table 4. Local response maxima in statistical parametric maps for the second-level analyses, probing sound perception (Sound > Silence, *p*<.05, FWE), and speech perception (Speech > Noise, *p*<.05, FWE), familiar vs. unfamiliar voice (FT > BU. *p*<.05, FWE, masked by BA21), and the effect of spatial separation on target speech perception, (Separated > collocated, *p*<.05, FWE).

Contrast	*x*	*y*	*z*	*T*	N voxels	Anatomical location
Sound > Silence	40	-22	6	15.7	1906	Heschl's gyrus, *r.*
	-54	-34	8	15.3	2129	Posterior STS, *l.*
	-8	16	50	9.9	117	Suppl. Motor Area, *l.*
	-44	-4	56	8.71	63	Precentral gyrus, *l.*
	-32	20	-2	7.89	25	Insula
	30	-66	-54	12.9	108	Cerebellum, Lobule VIII, *r. hemis.*
	6	-74	-22	9.2	48	Cerebellum, Lobule VII, vermis
	32	-64	-24	9.06	49	Cerebellum, Lobule VI, *r. hemis.*
Speech > Noise	-62	-28	2	11.4	525	Posterior STS, *l.*
	58	-18	2	7.59	40	Middle STG, *r.*
	-6	16	48	10.1	115	Suppl. Motor Area, *l.*
	-50	-2	48	7.92	21	Precentral gyrus, *l.*
	28	-66	-50	7.42	8	Cerebellum, Lobule VIII, *r. hemis.*
Separated > Collocated	48	-26	10	7.64	9	Posterior STS, *r.*
	-62	-44	6	7.51	7	Posterior MTG, *l.*
	4	-58	56	7.87	17	Precuneus, *r.*

x, y, z are MNI coordinates of local maxima (in mm). *T*, level of significance

Table 5. Significant clusters in statistical parametric maps in the second-level analysis. Clusters were considered significant if they reached a threshold of *p*<.05, FEW.

Contrast	*x*	*y*	*z*	T	N voxels	Anatomical location
(FT + FM) > BU	50	-42	22	7.15	1736	Posterior STG, *r.*
	-58	-46	24	6.33	1310	Supramarginal gyrus, *l.*
	48	22	20	5.53	773	IFG triangularis, *r.*
	-6	-62	44	5.14	890	Precuneus, *l.*

x, y, z are MNI coordinates of local maxima (in mm). T, level of significance

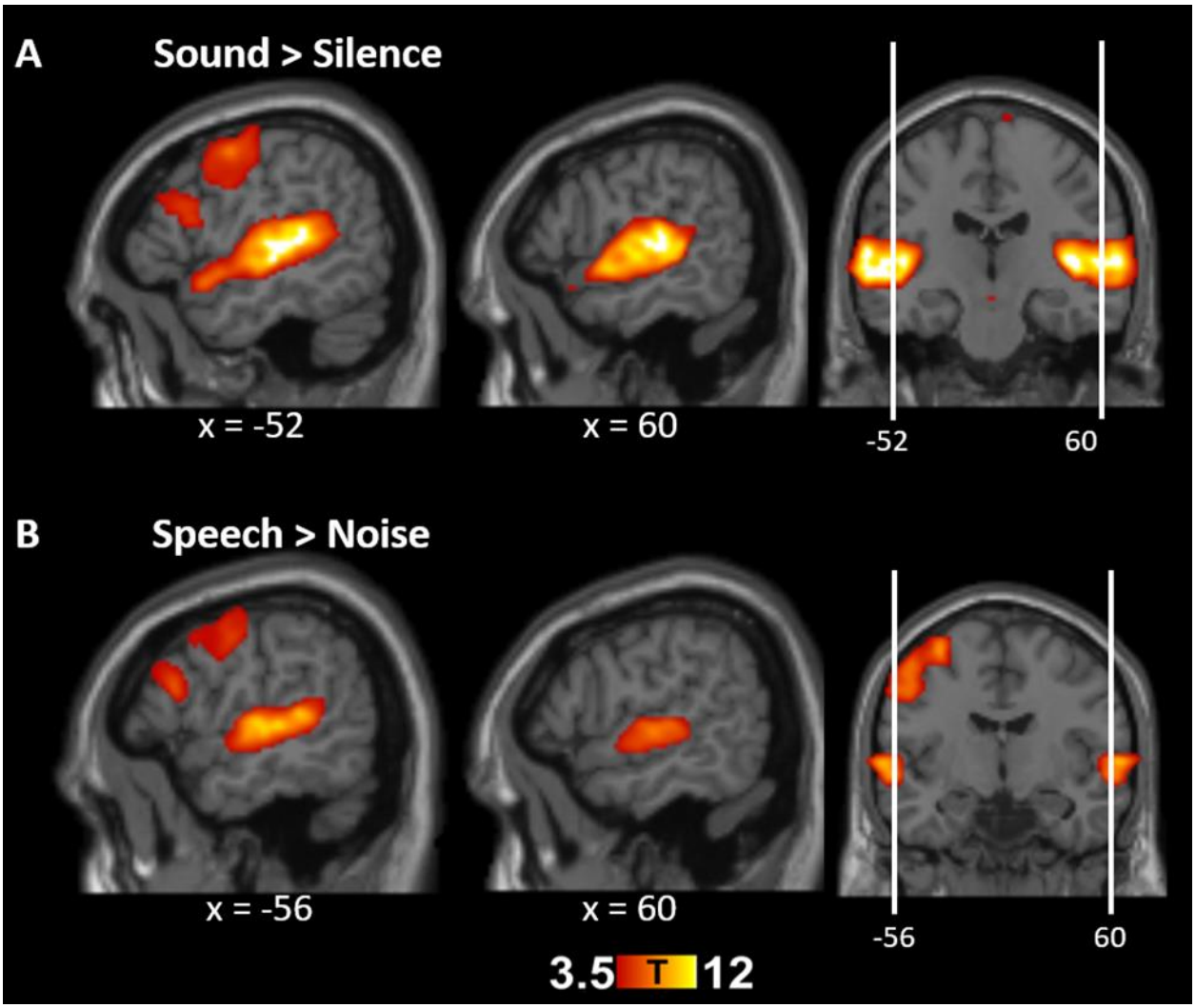


Figure 17. Regions activated when (A) listening to sounds versus silence, and (B) speech versus signal-correlated noise, *p*<.001 uncorrected. The coronal slice on the right shows the location of the sagittal slices for each contrast.

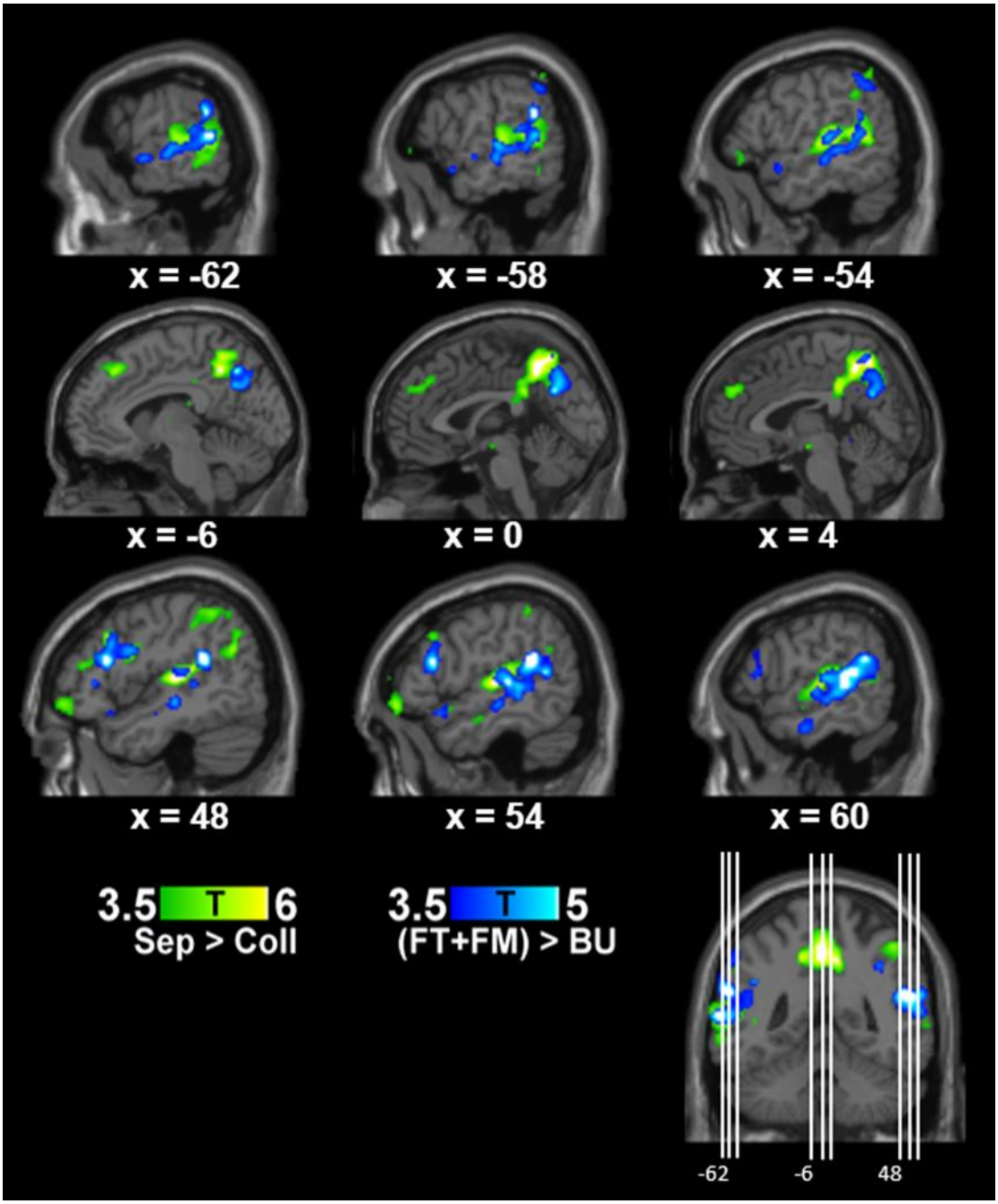


Figure 18. Regions activated when a familiar voice was present versus when both target and maskers were unfamiliar ((FT+FM)>BU; blue-light blue colour scale) and spatially separated versus collocated stimuli (Sep>Coll; green-yellow colour scale) at *p*<.001 uncorrected. The coronal slice on the bottom-right shows the location of the sagittal slices for each contrast.

4.3.3.4 Region by Condition Interaction

To verify that different regions respond differently when intelligibility is improved by acoustic (spatial separation) and cognitive (voice familiarity) factors, I conducted 2x2 repeated measures ANOVA comparing the activity in peak voxel in the (FT+FM)>BU and in the Sep>Coll contrasts and specifically examined the region by contrast (condition) interaction.

As expected, the interaction was significant [$F(1, 22) = 284.54$, $p<.0001$, $\omega^2 = .92$], indicating that activity in the peak voxels significantly differed based on the intelligibility cue used. Activity of the peak voxel in the (FT+FM)>BU contrast was significantly greater in response to familiar voices compared to spatial separations ($p<.0001$). Similarly, activity of the peak voxel in the Sep>Coll contrast was significantly greater in response to spatial cues compared to voice familiarity ($p<.0001$). This interaction is illustrated in Figure 19.

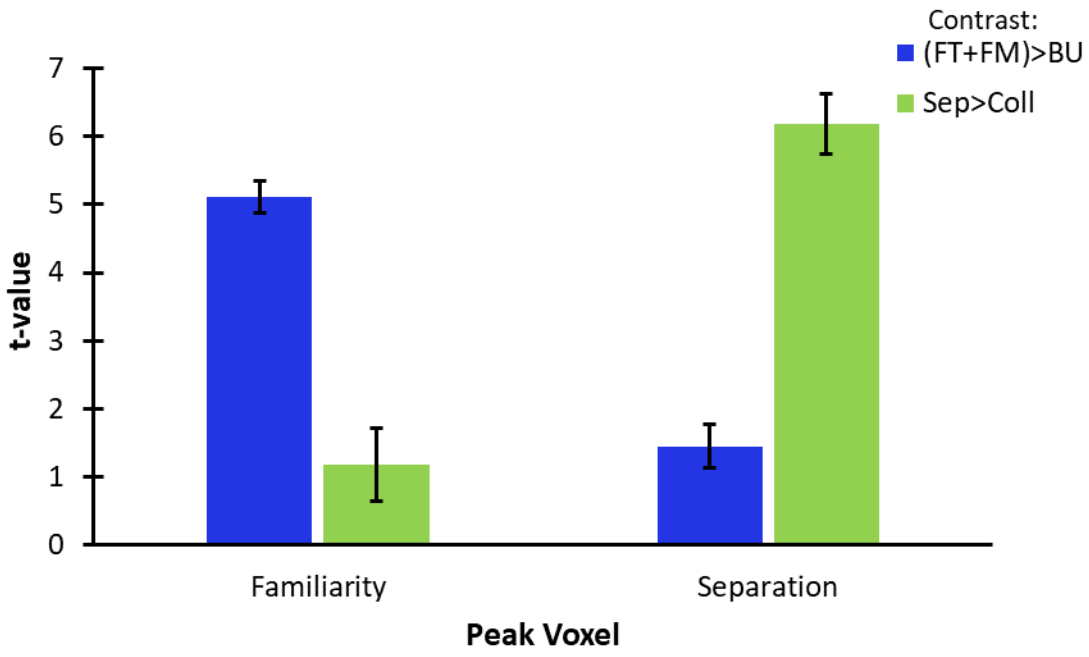


Figure 19. Differences in peak voxel activity in the (FT+FM)>BU contrast (blue) and Sep>Coll contrast (green) measured as t-values. Error bars are ±1 standard error of the mean.

4.4 Post-hoc data collection

The low overall performance in the behavioural task suggests that perhaps the task was too difficult. Responses in the current task were scored as correct or incorrect only and chance level was 50%, whereas the task used in Chapters 2 and 3 were scored based on every correct word instead of correct sentences and had a chance level of 12.5%. Therefore, the task used in the current study has less resolution to detect differences in accuracy. The current task was piloted in a previous voice familiarity study in our lab (Holmes, unpublished results) involving a two-voice mixture (one target voice and one masker voice), and not only was the familiar-voice benefit to intelligibility replicated, but it also significantly correlated with traditional matrix tasks in which participants are instructed to click on each word on the screen from a list of options. However, the current study differed from pilot work in terms of the number of maskers and TMR. The current study presented all stimuli at a TMR of -3 dB, whereas the pilot experiment used an adaptive threshold task to determine each participant's 40% threshold. The mean TMR across participants was -0.24 dB (SD = 3.77, range = -13 – 5 dB) in the Holmes *et al.* study. These differences made the task more challenging and could explain the absence of a behavioural familiar-voice benefit in the current study that was observed in pilot work.

To determine if the lower TMR and change in number of maskers were reasons for the observed low performance, the current study was followed up using a TMR of 0 dB. For the behavioural-only session, a traditional matrix task (the same as in Chapter 3) was used to measure the magnitude of the familiar-voice benefit for each participant. In the fMRI session, participants completed the same target-matching task used in the current chapter.

4.4.1 Participants

Two pairs of participants were tested (mean age = 23.7 years, SD = 0.6) who have known their partner an average of 3.6 years (SD = 2.8 years) and reported that they speak an average of 18.3 hours a week (SD = 6.5) were tested.

4.4.2 Results

4.4.2.1 Behavioural-only session

Three out of four participants demonstrated a familiar-voice benefit. On average, the collocated Familiar Target condition (mean = .67, SE = .07) was more intelligible than the collocated Both Unfamiliar condition (mean = .46, SE = .04) by a proportion of .21. The average magnitude of the familiar-voice benefit in the collocated condition was equivalent to a spatial separation of two unfamiliar talkers of 29.7°. The individual familiar-voice benefits ranged from 11.6° to 22.8°. These results are more consistent with the results of previous work (Domingo *et al.*, in preparation [Chapter 3]).

4.4.2.2 fMRI session

Accuracy in this session were similarly low compared to the original 23 participants. Proportion of correct responses was highest in the Familiar Target condition (mean = .57, SE = .03). Accuracy was slightly lower in Both Unfamiliar (mean = .53, SE = .04) and Familiar Masker (mean = .54, SE = .02).

The results of the behavioural-only session and fMRI session for these four participants suggest that the results of the target-matching task used in the current study did not approximate the results obtained from a traditional matrix intelligibility task involving one target and two symmetrical masker voices.

4.5 Discussion

Previous research has shown that familiar voices and spatial cues are both robust facilitators of speech intelligibility. In the current study, I used fMRI and virtually spatialized auditory stimuli to determine the neural regions that are associated with each intelligibility cue, and to examine whether these two cues improve intelligibility through mechanisms that depend on similar brain networks. Results from this study revealed a set of brain regions that support familiar-voice perception and spatial cue processing. These areas overlap in the posterior temporal regions and precuneus. Further, results suggest differential patterns of neural activation when a familiar voice is present or not, involving activation in the right STG, right IFG, left supramarginal gyrus, and precuneus. Despite

the absence of a behavioural familiar-voice intelligibility benefit in the fMRI session, which may be due to the insensitivity of this task, given the previous literature on this topic it seems likely that this is a real effect, and that the neural substrates of the processing of familiar voices differ somewhat from those that support processing of unfamiliar voices.

4.5.1 Familiar voices activate voice, person recognition, and attention areas

Familiar voices appear to activate areas in known 'voice' areas in the temporal and frontal lobes. In the current study, the right posterior STG was the only peak in the temporal lobe that activated in response to a familiar voice. The right posterior STG is part of the 'temporal voice areas' (Pernet *et al*., 2015) and has also been implicated in processing speaker information (Chandrasekaran, Chan, & Wong, 2011). The current results are also consistent with clinical research that suggests that the posterior right temporal lobe is critical for voice recognition (Ellis, Young, & Critchley, 1989).

The right posterior STG activation observed in response to familiar voices may also be reflective of an intelligibility benefit from a familiar voice. Intelligibility and voice areas both involve the length of the left and right STS/STG (Davis & Johnsrude, 2003; Pernet *et al*., 2015; Wild, Davis, *et al*., 2012; Wild, Yusuf, *et al*., 2012). Therefore, the right posterior STG is a small subset of established intelligibility areas. While intelligibility has been shown to be left-lateralized (Davis & Johnsrude, 2003; Narain *et al*., 2003; Wild, Davis, *et al*., 2012; Wild, Yusuf, *et al*., 2012; Zekveld *et al*., 2006), areas of the right temporal lobe have also been shown to respond to intelligible speech (Davis & Johnsrude, 2003). The limited activation in intelligibility areas may be reflective of the lack of familiar-voice benefit in the behavioural results and may also simply be responding to the familiar voice in some way.

The regions of temporal lobe that I observed to be active have been reported to be sensitive both to voices (Pernet *et al*., 2015; Schall *et al*., 2014) and to the intelligibility of the content (Davis & Johnsrude, 2003; Wild, Davis, *et al*., 2012; Wild, Yusuf, *et al*., 2012). This appears to contradict the findings of von Kriegstein *et al*. (2003) who suggest

that voice identification activates anterior temporal areas and verbal information activates the left posterior middle temporal region. A possible explanation for this discrepancy is that in the current study, participants were never explicitly instructed to attend to the target voice as they were in the von Kriegstein *et al.* (2003). However, Schall *et al*. (2014) also required participants to attend to a voice and found that posterior right STS activity was associated with voice identity recognition. Perhaps these differences in findings are task-related, in which von Kriegstein *et al*. (2003) instructed participants to respond via button press if the voice presented on a trial was the same was the target voice presented at the beginning of the block. In comparison, Schall *et al*. (2014) asked participants to indicate if the voice presented on a trial matched the name assigned to that voice in a preceding training phase. It may be the case that the task that von Kriegstein *et al*. (2003) used may also be involving voice *discrimination* processes, and the task in Schall *et al.* (2014) was primarily about voice identity *recognition*.

Two areas in the parietal lobe activated in response to familiar voices. The first is the precuneus, which has been shown in other studies to be involved in familiar-voice recognition and discrimination (Nakamura *et al*., 2001; von Kriegstein *et al*., 2003; von Kriegstein & Giraud, 2004). Precuneus activation in the current study can perhaps be associated with processes of word recognition, which has also been observed by (Dorfel *et al*., 2007). The precuneus is considered to be a higher-order area involved in processes such as visuo-spatial imagery, episodic memory retrieval, self-processing, and consciousness (for a review on anatomy of the precuneus and behavioural correlates of activation in the region, see Cavanna & Trimble, 2006). The other parietal region that responded to familiar voices is the left supramarginal gyrus. This area has been shown to be active in speech recognition (Benson *et al.*, 2001) and word recognition (Relander & Rämä, 2009). Perhaps the presence of a familiar voice enabled participants to better perceptually segregate the target and masker voices in order to recognize the target sentence, leading to activation in this area.

Another region activated by familiar voices was the right IFG. The right IFG comprises the anterior and middle frontal voice areas (FVAs; Aglieri *et al*. 2018) and is part of the auditory ‘what’ pathway in humans (Zündorf *et al.*, 2016). In line with this, the IFG has

been demonstrated to be involved in discriminating between trained familiar and novel voices (Zäske *et al*., 2017). Therefore, activation in this region might indicate that participants noticed or recognized their familiar voice. In addition to voice discrimination processes, the IFG has also been implicated in language processing (Buckner, Raichle, & Petersen, 1995), attentional switching (Lee, Larson, Maddox, & Shinn-Cunningham, 2014), and attending to speech masked with competing speech spoken by the same talker (Nakai, Kato, & Matsuo, 2005). Perhaps the observed IFG activation is not only due to recognizing a familiar voice in the three-sentence mixture, but also due to increased attention involved in processing personally salient stimuli.

Taken together, the observed activation in the right and left temporal regions, parietal lobe, and right IFG, in response to the presence of a personally familiar voice, is consistent with observations made in other neuroimaging studies investigating voice perception studies (Nakamura *et al*., 2001; Pernet *et al*., 2015; von Kriegstein *et al*., 2003; von Kriegstein & Giraud, 2004), person-recognition (Barton & Corrow, 2016; Ellis *et al*., 1989; Zäske *et al*., 2017), speech and word recognition (Benson *et al*., 2001; Relander & Rämä, 2009), and selective attention and inhibition (Lee *et al*., 2014; Nakai *et al*., 2005).

4.5.2 Spatialized voices activate temporal regions and precuneus

When the maskers were spatially separated from the target, I observed activation in the precuneus and in regions of the posterior temporal lobe. Activated temporal regions are consistent with the temporal voice areas (Pernet *et al.*, 2015), and may reflect the improved perception of the target voice due to spatial separation. In line with this, the observed temporal-lobe activation comprises a subset of the intelligibility areas defined above and could reflect spatial release from masking observed in the behavioural results of the current study. Further, the majority of participants (19 out of 23) indicated that they at least moderately perceived the stimuli as coming from different directions. Therefore, another possibility is that posterior temporal lobe activation could be representing the spatial separation between concurrent sounds, and not necessarily their specific locations in space, as demonstrated by Shiell, Haufeld, & Formisano (2018).

The fact that posterior temporal regions appear sensitive both to the presence of familiar voices and to spatial cues supports the notion that this region acts as a computational hub that is involved in a variety of spectotemporal analyses (Griffiths & Warren, 2002). Based on the model proposed by Warren and Griffiths (2002) planum temporale activation could be indicative of perceptually segregating the three voices in the stimuli, or perceiving changes in spatial location between the collocated and spatially separated conditions, or matching the familiar-target speech with stored templates. All three of these processes may play a role in improving intelligibility of target speech.

Lastly, the precuneus activated in response to spatialized sounds. This region has also been involved in processing spatial location change in non-speech sounds (Kryklywy *et al*., 2013; Maeder *et al*., 2001). More generally, the precuneus has been involved in processing object location changes and visuo-spatial imagery (Cavanna & Trimble, 2006; Wolbers *et al*., 2008), suggesting that the spatial information processed in the precuneus is not modality specific.

4.5.3 Limitations of this work

Accuracy and d' scores of the current study (Figure 16) were low, although generally better than chance. The low levels of performance suggest that the task was perhaps too difficult. Further, the weak or absent familiar-voice benefit in this study, despite having observed a strong benefit using different tasks numerous times (Domingo *et al.*, in revision [Chapter 2], Domingo *et al*., in preparation [Chapter 3], Holmes *et al.*, 2018; Johnsrude *et al*., 2013) may be due to a floor effect. In previous work, participants were instructed to select each word that they thought was spoken by the target voice, and their overall accuracy was scored using correct words instead of correct sentences. Therefore, if a participant were to identify three out of four words correct on a given trial, they would have been scored as 75% accurate. In contrast, the trials in the current study were scored as either correct or incorrect, which reduced resolution, increased the chance rate to 50% and may have increased variability. Furthermore, the foils were highly confusable with the target: three out of four words were from the target sentence and one word was from one of the two masker sentences.

The results of the post-hoc data collection supports the interpretation that the lack of familiar-voice benefit is task related. The four post-hoc participants did both a traditional matrix task (in the behavioural-only session) as well as the match/no-match identification task (in the fMRI session). In the behavioural-only session, there appeared to be a strong familiar-voice benefit. However, in the fMRI session, the Familiar Target condition was only marginally more intelligible than the other conditions, and overall accuracy was still near chance level (.50). These results appear to suggest that the match/no-match task used in the current study did not approximate the results of the matrix task.

4.6 Conclusion

In this study, I presented participants with a target sentence with symmetrical maskers spatially separated to the degree that produced equal intelligibility as a familiar voice. When the familiar voice was present in the mixture, there was activation in the right posterior STG, left supramarginal gyrus, precuneus, and right IFG. The specific processes driving this activation are difficult to ascertain because of the weak familiar-target benefit that was observed in the behavioural-only session but not in the fMRI session.

When attending to spatially separated speech, participants showed a significant improvement in target-matching accuracy compared to when they attended to collocated speech. Further, spatially separated speech activated posterior temporal areas and the precuneus, which also activated in response to familiar voices. The activation of the precuneus and temporal areas for both familiar voices and spatial separation may suggest that the neural mechanisms that support intelligibility from these two cues at least partially overlap and are therefore not entirely distinct from one another.

Chapter 5

5 General Discussion

Voice familiarity has been shown to be an effective facilitator of speech intelligibility. Johnsrude *et al*. (2013) found that the voice of a long-term spouse can improve intelligibility by 10-15% (equivalent to 9 dB) when it serves as either the target or the masker voice. This benefit is commensurate with another cue that has also been shown to greatly improve intelligibility: spatial separation between simultaneous speech streams. The aim of this thesis was to extend the findings of Johnsrude *et al*. (2013) and provide a deeper understanding of voice familiarity as an intelligibility cue. Specifically, I (1) measured the magnitude of the familiar-voice benefit from different types of naturally familiar voices (friends and spouses); (2) quantified the familiar-voice benefit in terms of degrees of spatial separation; and (3) compared the neural bases of voice familiarity and spatial release from masking to determine if these cues improve intelligibility by recruiting similar areas of the brain. The main findings of the three experiments I conducted are discussed below. In this chapter, I also identify limitations of these experiments, as well as make recommendations for future research in this area.

5.1 Summary of key findings from Chapter 2

In Chapter 2, I measured the familiar-target and familiar-masker benefits in older spouses (age ≥ 55 years), younger spouses (age < 55 years), and friends, in a matrix task using the BUG corpus (Kidd *et al*., 2008) in which sentences followed the format “<Name> <verb> <number> <adjective> <noun>. Participants identified the target sentence by the Name word (e.g., Bob or Pat) and responded with the remaining four words from the sentence by clicking from a set of options on a screen.

Overall, intelligibility was highest when participants attended to a familiar compared to an unfamiliar voice. Interestingly, the intelligibility benefit from a familiar voice did not differ between older spouses, younger spouses, and friends. Further, the familiar-voice benefit did not correlate with relationship duration, suggesting that longer relations do not systematically increase the benefit to intelligibility from a familiar voice. This suggests

that once the familiar-voice benefit has developed (which, based on our data, probably takes less than 1.5 years), it remains constant over time. This work is currently under revision for the Journal of Experimental Psychology: Applied; the requested revisions are minor.

5.2 Summary of key findings from Chapter 3

Chapter 2 showed that the familiar-voice benefit to intelligibility previously demonstrated by Johnsrude *et al*. (2013) is replicable using a more challenging task, such as the BUG (Kidd *et al*., 2008). In Chapter 3, I measured intelligibility in terms of proportion correct and equated each participant's performance to degrees of spatial separation to quantify the familiar-voice benefit to intelligibility. Spatial release from masking is a well-known and well characterized benefit, and it is helpful to know whether the benefit realized from a familiar voice is commensurate with it. Further, in this experiment, I examined whether, and how, these two cues interact with one another.

Between TMRs of -3 to 6 dB, participants reported 10-30% more words correctly when the target sentence was spoken in a familiar voice compared to an unfamiliar voice. In terms of degrees of spatial separation and TMR, the magnitude of the familiar-voice benefit to intelligibility was 14-17° and 5.1 dB, respectively. The improvement to intelligibility from attending to a familiar target voice is commensurate with or even larger than that of a 90° spatial separation reported in previous studies.

At 6 dB TMR, the familiar-voice benefit did not differ across spatial separations, likely because the target sentence was sufficiently intelligible from TMR cues alone so that participants did not need to rely voice familiarity or spatial separations. However, at lower TMRS (-3 dB and 0 dB) the familiar-voice benefit was greater at smaller separations than at larger separations. These results suggest that voice familiarity is most beneficial for listeners when acoustic cues, such as those providing information about spatial separation, alone are insufficient to perceptually separate speech streams. This paper is nearly ready for submission to the Journal of the Acoustical Society of America.

5.3 Summary of key findings from Chapter 4

Chapter 3 showed that attending to a familiar voice led to an improvement in intelligibility of about 20%, which is equal to 17.1° at -3 dB TMR. The magnitude of intelligibility improvement from a familiar voice in Chapter 3 is comparable to that of a 90° spatial separation reported in other studies. Because voice familiarity and spatial separation both provide considerable improvement to intelligibility, the goal of Chapter 4 was to using fMRI to determine whether the neural substrates supporting these two cues are similar to or different from one another.

Due to the requirements of fMRI experiments, I used a simplified version of the task used in Chapters 2 and 3. In this simplified task, participants responded by button press if the sentence presented on the screen matched the target sentence they heard. The task was likely too difficult for participants, as overall accuracy and sensitivity was low and the familiar-voice benefit was substantially weaker than in other studies. Nevertheless, when a familiar masker was present in the mixture there was activation in right posterior STG, left supramarginal gyrus, right IFG, and precuneus. These activations are consistent with previous voice research and may be associated with person-recognition, speech recognition, voice processing, or attention involving the familiar voice. When spatially separated speech was contrasted against collocated speech, there was activation right posterior STS, left posterior MTG, and precuneus, again, similar to what has been observed previously during spatial perception. Although the results are not conclusive, primarily due to the absence of a familiar-voice benefit in behavioural data, it appears that the mechanisms that support intelligibility from familiar voices and spatial separation partially overlap with one another but are partially distinct, as predicted. A paper based on this work is currently being prepared for publication.

5.4 Limitations

5.4.1 Familiar-voice benefit was not present in all participants

All experiments in this thesis used the psychophysical method of constant stimuli, in which a fixed number of trials were presented to participant in each condition. This method ensures that data can be easily aggregated and compared because all participants

are tested under identical conditions. This procedure also limits bias because participants cannot predict the stimuli or condition.

An adaptive staircase procedure (Levitt, 1971) that allows researchers to modify the TMR based on each participant's performance ensures that all participants demonstrate the same level of accuracy. Therefore, this method can be used adapt TMR to ensure that all participants show a familiar-voice benefit. Further, it can also be used to adjust difficulty to avoid ceiling effects such as those observed in Chapter 3 because the detection threshold can be defined in advance.

5.4.2 Unable to investigate individual differences

The voices in Chapters 2-4 were counterbalanced such that every participant served as a familiar voice to their partner in the study and also served as an unfamiliar voice for two other participants in this study. Counterbalancing in this away allows for the familiar and unfamiliar voices to be acoustically matched *across all participants*, meaning that comparisons between conditions are not biased by acoustic differences. However, this design does not allow me to study individual differences in familiar-voice benefit. At the individual level, intelligibility is affected by F_0 differences, as well as by other less well-defined perceived similarities between the familiar and unfamiliar voice. No attempts were made to keep the F_0 differences between familiar and unfamiliar voices similar between participants. Instead, assigning unfamiliar voices to each participant was done with consideration of which other participant a participant was least likely to know, but also with the constraint that each voice served as an unfamiliar voice for two other participants.

Examining individual differences in the magnitude of the familiar-voice benefit would allow researchers to determine the extent of the familiar-voice benefit and identify underlying characteristics common to people with strong familiar-voice benefits. One way to examine individual differences is by using a training paradigm in which all participants are presented with the same familiar and unfamiliar voices. Performance in training paradigms could also be correlated with other behavioural recognition measures

such as face or object recognition. Results of those correlations could suggest that different types of person recognition process are not independent of one another.

5.4.3 HRTF measurements were not personalized

Chapters 3 and 4 presented virtually spatialized stimuli to measure spatial release from masking. These stimuli were created using the HRTFs of a KEMAR mannequin. Using HRTFs measured from KEMAR is fairly common in auditory research; however, a participant's ability to perceive the stimuli as coming from different directs strongly depends on how similar their pinna and head shape and size is to the KEMAR. In Chapter 4, many participants reported that they could perceive the sounds coming from different directions, but some could not. Not only could this affect behavioural results, but this could also influence the cortical activations observed. This limitation could be overcome by measuring each participant's own HRTF and applying those to stimuli. Doing this would ensure that every participant would be able to perceive spatialized stimuli, leading perhaps to more robust results, with a clearer interpretation.

5.4.4 Closed-set tasks are not generalizable

Chapters 2-4 all used closed-set tasks in which participants could select from a defined set of possible responses. In Chapters 2 and 3, there were eight options for each of the four words participants had to select. In Chapter 4, participants determined whether the visual cue sentence was a match or not with the target sentence they heard. I used these tasks because they control for potential response biases. Further, similar studies (Kreitewolf *et al*., 2017; Johnsrude *et al*., 2013) also used closed-set tasks making our results more easily comparable to theirs.

The primary limitation to closed-set tasks is that they may not generalize to natural communication because real-world communication is not limited to a small word set and because closed-set tasks involve word identification, not recognition. The tasks used in this thesis do not take suprasegmental elements such as intonation and prosody into account. The way a familiar voice is encoded or represented in memory may include these characteristics so the tasks used may not be exploiting all the possible cues

available from a familiar voice. Therefore, the current experiments may underestimate the true magnitude of the familiar-voice benefit.

Using an open-set task, with more naturalistic sentences and in which participants identify the words from a target sentence instead of merely recognizing them from a set of options will provide a more ecologically valid measurement of the familiar-voice benefit to intelligibility. Observing a robust familiar-voice benefit from a naturally familiar voice in an open-set task would provide more conclusive evidence that familiar voices are more intelligible in the presence of competing speech.

5.5 Recommendations and directions for future research

Findings from Chapter 2 suggest that the familiar-voice benefit is robust by about 1.5 years of knowing someone, and did not increase with increased exposure. However, it cannot build up very quickly since I observed no improvement in the intelligibility of (at the outset) unfamiliar voices throughout the experiment (approximately two hours of exposure). These results raise an important question: *How long does it take for the familiar-voice benefit to develop*?

Nygaard *et al*. (1994) show a familiar-voice benefit for 10 familiarized voices after nine days of training. Similarly, Kreitewolf *et al.* (2017) observed a familiar-voice benefit for one previously novel voice after a total of six hours of training, spread across four consecutive days. It is possible that the timeline for learning a familiar voice artificially via a lab procedure may differ from that for learning a voice in real life. To investigate the development of the familiar-voice benefit from *naturally* familiar voices, a longitudinal design tracking intelligibility of new familiar voice over time would be ideal. For example, recruiting first year university residence roommates (who did not previously know each other) when they had just moved in with one another, and again after each month of the first year of university would allow intelligibility of the roommate's voice to be compared across different timepoints to identify how long it takes to develop a familiar-voice benefit.

In the imaging results of Chapter 4, there was no difference in regions of activation between the Familiar Target and Familiar Masker conditions, when contrasted against Both Unfamiliar. Although it is possible that familiar voices are represented the same way in the brain when they are presented as targets or maskers, it may also be the case that *patterns* of activation may be different between the two conditions containing a familiar voice, but the univariate analysis used in Chapter 4 was not sensitive enough to detect those differences. To address this, a representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008) could be used to compute similarity between activation patterns between experimental conditions. RSA could also be used to compare the activation patterns and interactions for familiar and unfamiliar voices. The activation pattern at different TMRs may be more similar for familiar target voices than unfamiliar target voices. The similarities that may occur between participants may also correlate with a behavioural familiar-voice benefit. Participants who show a strong familiar-voice benefit may have different neural representations of a familiar voice compared to participants who do not show this benefit.

In natural listening environments, it is impossible that multiple voices originate from exactly the same point (as in the collocated conditions of Chapters 3 and 4). It is similarly implausible that two talkers are located at the same distance from the listener and are 5-10° apart. In order to increase ecological validity and accurately measure the practical benefit of familiar voices to intelligibility, future research must take spatial separations between target and maskers into account.

5.6 Implications

The experiments in the current thesis were intended to contribute to the larger body of literature on speech intelligibility in noisy environments. Compared to other cues such as spatial separation and TMR, voice familiarity is not as well investigated and therefore not as well understood. However, the current thesis highlights the magnitude and robustness of voice familiarity as a facilitator of speech intelligibility in noisy environments and demonstrates that its effectiveness is similar to that of larger spatial separations. These findings suggest that further research into voice familiarity is warranted to gain a deeper understanding of how familiarity improves intelligibility.

This research may also have health-related implications. The World Health Organization International Classification of Functioning, Disability, and Health (WHO ICF; World Health Organization, 2001) includes a concept of *participation*, defined as "involvement in life situations". One construct that is based on this concept is *communicative participation*, defined as participating in life situations involving verbal communication of ideas or information (Eadie et al., 2006). Communicative participation is considered a critical indicator of intervention success to ensure that these interventions are contributing a meaningful improvement in lives of clients, including older adults who suffer from age-related hearing loss and often experience difficulty communicating in noisy or reverberant environments (Huang & Tang, 2010). Chapter 2 in the current thesis suggests that the intelligibility benefit gained from a familiar voice is as strong in older listeners as it is in younger listeners. Although all participants in Chapter 2 were audiometrically tested to have normal pure-tone hearing thresholds, these findings raise the possibility that repeatedly exposing a listener to a voice may improve their ability to understand that person's voice in noisy environments. Auditory training on voices of frequent communication partners has recently begun to be explored and has demonstrated positive outcomes (Tye-Murray et al., 2016).

5.7 Conclusions

Using a closed-set matrix intelligibility task in which participants have to select each target word from a list of options, the intelligibility benefit from a familiar voice is about 10-30%, which is comparable to the benefit gained from large spatial separations in other studies. In terms of degrees of spatial separation, the magnitude of the familiar-voice benefit to intelligibility is 15-17° using a symmetrical masker paradigm. Therefore, the experiments in this thesis highlight the potential effectiveness of familiar voices as a cue to improve intelligibility.

Further, this thesis provides preliminary evidence that the neural mechanisms that underlie intelligibility from a familiar voice and spatial separations are partially distinct and partially overlap, particularly in areas including the right posterior superior temporal lobe and precuneus. Although the results from Chapter 4 are inconclusive, they suggest that these two robust cues to intelligibility may converge.

None of the experiments in this thesis replicated the familiar masker benefit demonstrated by Johnsrude et al. (2013), suggesting that the familiar masker benefit is task-related or that the closed-set task used in this thesis are not sensitive enough to observe it. In this thesis, intelligibility from the familiar masker condition was no different from that when both target and maskers were unfamiliar to the listener. At the higher TMRs in Chapter 2 and in Chapter 4, in fact, the familiar masker condition was *less* intelligible than the control Both Unfamiliar condition, suggesting that the presence of a familiar masker voice may actually be distracting to participants and make them less able to hear out an unfamiliar talker. At least within the designs used here, this thesis shows that a familiar voice only improves intelligibility when it serves as the target voice.

References

Aglieri, V., Chaminade, T., Takerkart, S., & Belin, P. (2018). Functional connectivity within the voice perception network and its behavioural relevance. *NeuroImage*, *183*, 356–365. https://doi.org/10.1016/j.neuroimage.2018.08.011

Ahveninen, J., Kopco, N., & Jääskeläinen, I. P. (2014). Psychophysics and neuronal bases of sound localization in humans. *Hearing Research*, *307*, 86–97. https://doi.org/10.1016/j.heares.2013.07.008

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., & Grady, C. L. (2001). "'What'" and " 'where'" in the human auditory system. *Proceedings of the National Academy of Sciences*, *98*, 12301–12306. Retrieved from www.pnas.orgcgidoi10.1073pnas.211209098

Alain, C., & Woods, D. L. (1999). Age-related changes in processing auditory stimuli during visual attention: Evidence for deficits in inhibitory control and sensory memory. *Psychology and Aging*, *14*, 507–519. https://doi.org/10.1037/0882-7974.14.3.507

Amaro, E., & Barker, G. J. (2006). Study design in fMRI: Basic principles. *Brain and Cognition*, *60*, 220–232. https://doi.org/10.1016/j.bandc.2005.11.009

Arbogast, T. L., Mason, C. R., & Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, *112*(5), 2086–2098. https://doi.org/10.1121/1.1510141

Arbogast, T. L., Mason, C. R., & Kidd, G. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *117*(4), 2169–2180. https://doi.org/10.1121/1.1861598

Arnott, S. R., Binns, M. A., Grady, C. L., & Alain, C. (2004). Assessing the auditory dual-pathway model in humans. *NeuroImage*, *22*, 401–408. https://doi.org/10.1016/j.neuroimage.2004.01.014

Assmann, P. F. (1999). Fundamental frequency and the intelligibility of competing voices. In *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 179–182). San Francisco. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0179.pdf

Badri, R., Siegel, J. H., & Wright, B. A. (2011). Auditory filter shapes and high-frequency hearing in adults who have impaired speech in noise performance despite clinically normal audiograms. *The Journal of the Acoustical Society of America*, *129*, 852–863. https://doi.org/10.1121/1.3523476

Barrett, D. J. K., & Hall, D. A. (2006). Response preferences for "'what'" and " 'where'" in human non-primary auditory cortex. *NeuroImage*, *32*, 986–977. https://doi.org/10.1016/j.neuroimage.2006.03.050

Barton, J. J. S., & Corrow, S. L. (2016). Recognizing and identifying people: A

neuropsychological review. *Cortex, 75*, 132–150. https://doi.org/10.1016/j.cortex.2015.11.023

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research, 74*, 110–120. https://doi.org/10.1007/s00426-008-0185-z

Bee, M. A., & Micheyl, C. (2008). The "Cocktail Party Problem": What is it? How can it be solved? And why should animal behaviorists study it? *Journal of Comparitive Psychology, 122*, 235–251. https://doi.org/10.1037/0735-7036.122.3.235

Belin, P., & Zatorre, R. J. (2000). "What", "where", and "how" in auditory cortex. *Nature Neuroscience, 3*(10), 965–966. Retrieved from https://www.nature.com/articles/nn1000_965.pdf

Belin, Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309–312. https://doi.org/10.1038/35002078

Benson, R. R., Whalen, D. H., Richardson, M., Swainson, B., Clark, V. P., Lai, S., & Liberman, A. M. (2001). Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain and Language, 78*, 364–396. https://doi.org/10.1006/brln.2001.2484

Best, V., Gallun, F. J., Ihlefeld, A., & Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *Journal of the Acoustic Society of America, 120*(3), 1506–1516. https://doi.org/10.1121/1.2234849

Best, V., Mason, C. R., & Kidd, G. (2011). Spatial release from masking in normally hearing and hearing-impaired listeners as a function of the temporal overlap of competing talkers. *The Journal of the Acoustical Society of America, 129*(3), 1616–1625. https://doi.org/10.1121/1.3533733

Best, V., Mason, C. R., Swaminathan, J., Roverud, E., & Kidd, G. (2017). Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. *The Journal of the Acoustical Society of America*, *141*(1), 81–91. Retrieved from https://doi.org/10.1121/1.4973620

Best, V., Thompson, E. R., Mason, C. R., & Kidd, G. (2013). An Energetic Limit on Spatial Release from Masking. *Journal of the Association for Research on Otolaryngology, 14*(4), 603–610. https://doi.org/10.1007/s10162-013-0392-1

Bethmann, A., Scheich, H., & Brechmann, A. (2012). The temporal lobes differentiate between the voices of famous and unknown people: An event-related fMRI study on speaker recognition. *PLoS One, 7*(10), 1–15. https://doi.org/10.1371/journal.pone.0047626

Biederman, I., Shilowich, B. E., Herald, S. B., Margalit, E., Maarek, R., Meschke, E. X., & Hacker, C. M. (2018). The cognitive neuroscience of person identification. *Neuropsychologia, 116*, 205–214. https://doi.org/10.1016/j.neuropsychologia.2018.01.036

Birkett, P. B., Hunter, M. D., Parks, R. W., Farrow, T. F., Lowe, H., Wilkinson, I. D., &

Woodruff, P. W. (2007). Voice familiarity engages auditory cortex. *Neuroreport, 18*, 1375–1378. https://doi.org/10.1097/WNR.0b013e3282aa43a3

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice- and face-recognition areas. *The Journal of Neuroscience*, *31*(36), 12906–12915. https://doi.org/10.1523/JNEUROSCI.2091-11.2011

Blank, H., Wieland, N., & Von Kriegstein, K. (2014). Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neuroscience and Biobehavioral Reviews*, *47*, 717–734. https://doi.org/10.1016/j.neubiorev.2014.10.022

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer.

Boldt, R., Malinen, S., Seppä, M., Tikka, P., Savolainen, P., Hari, R., & Carlson, S. (2013). Listening to an audio drama activates two processing networks, one for all sounds, another exclusively for speech. *PLoS ONE*, *8*(5), 1–10. https://doi.org/10.1371/journal.pone.0064489

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065–1066. https://doi.org/10.1121/1.428288

Bolia, R. S., Nelson, W. T., & Morley, R. M. (2001). Assymetric performance in the cocktail party effect: Implications for the design of spatial audio displays. *Human Factors*, *43*, 208–216. https://doi.org/10.1518/001872001775900887

Brainard, M. S. (1994). Neural substrates of sound localization. *Current Opinion in Neurobiology*, *4*(4), 557–562. https://doi.org/10.1016/0959-4388(94)90057-4

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press. Retrieved from http://webpages.mcgill.ca/staff/Group2/abregm1/web/pdf/2004_ Encyclopedia-Soc-Behav-Sci.pdf

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *83*, 1508–1516. https://doi.org/10.1121/1.395906

Bronkhorst, A. W., & Plomp, R. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America*, *92*(6), 3132–3139. https://doi.org/10.1121/1.2202888

Brungart, D S. (2001). Evaluation of speech intelligibility with the coordinate response measure. *The Journal of the Acoustical Society of America*, *109*(January 2001), 2276–2279. https://doi.org/10.1121/1.1357812

Brungart, Douglas S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Brungart, Douglas S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *103*(3), 1101–1109. https://doi.org/10.1121/1.1408946

Brungart, Douglas S, & Iyer, N. (2012). Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America*, *132*, 2545–2556. https://doi.org/10.1121/1.4747005

Buckner, R. L., Raichle, M. E., & Petersen, S. E. (1995). Dissociation of human prefrontal cortical areas across different speech production tasks and gender groups. *Journal of Neurophysiology*, *74*, 2163–2173. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.477.2092&rep=rep1&type=pdf

Carabellese, C., Appollonio, I., Rozzini, R., Bianchetti, A., Frisoni, G. B., Frattola, L., & Trabucchi, M. (1993). Sensory impairment and quality of life in a community elderly population. *Journal of the American Geriatrics Society*, *41*, 401–407. https://doi.org/10.1111/j.1532-5415.1993.tb06948.x

Carlile, S. (2014). Active Listening: Speech intelligibility in noisy environments. *Acoustics Australia*, *42*(2), 90–96. Retrieved from http://www.acoustics.asn.au/journal/2014/Vol42No2_CARLILE.pdf

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain*, *129*, 564–583. https://doi.org/10.1093/brain/awl004

Chandrasekaran, B., Chan, A. H. D., & Wong, P. C. M. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience*, *23*(10), 2690–2700. https://doi.org/10.1162/jocn.2011.21631

Cheng, C. I., & Wakefield, G. H. (1999). Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. *IEEE International Conference on Acoustics, Speech, & Signal Processing*, *2*, 961–964. Retrieved from http://www.eecs.umich.edu/~coreyc

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustic Society of America*, *25*(5), 1262–2527. https://doi.org/10.1121/1.1408946

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, *123*(1), 414–427. https://doi.org/10.1121/1.2804952

Coull, J. T., & Nobre, A. C. (1998). Where and when to pay attention: The neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *The Journal of Neuroscience*, *18*, 7426–7435. Retrieved from http://www.jneurosci.org/content/jneuro/18/18/7426.full.pdf

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 643–656. https://doi.org/10.1037/0096-1523.30.4.643

Darwin, Chris J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, *1*, 327–333. Retrieved from http://web.mit.edu/hst.723/www/ThemePapers/ASA/Darwin97.pdf

Darwin, Christopher J, Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*(5), 2913–2922. https://doi.org/10.1121/1.1616924

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. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147. https://doi.org/10.1016/j.heares.2007.01.014

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*, 222–241. https://doi.org/10.1037/0096

Deroche, M. L. D., Culling, J. F., Chatterjee, M., & Limb, C. J. (2014). Roles of the target and masker fundamental frequencies in voice segregation. *The Journal of the Acoustical Society of America*, *136*(3), 1225–1236. https://doi.org/10.1121/1.4890649

Dorfel, D., Werner, A., Schaefer, M., von Kummer, R., & Karl, A. (2007). Distinct brain networks in recognition memory share a defined region in the precuneus. *European Journal of Neuroscience*, *30*, 1947–1959. https://doi.org/10.1111/j.1460-9568.2009.06973.x

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, *102*, 2403–2411. https://doi.org/10.1121/1.419603

Dubno, J. R., Dirks, D. D., & Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, *76*(1), 87–96. https://doi.org/10.1121/1.391011

Durlach, N. (2006). Auditory masking: Need for improved conceptual structure. *The Journal of the Acoustical Society of America*, *120*(4), 1787–1790. https://doi.org/10.1121/1.2335426

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., & Shinn-Cunningham, B. G. (2003). Note on informational masking (L). *The Journal of the Acoustical Society of America*, *113*(6), 2984–2987. https://doi.org/10.1121/1.1570435

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd, G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America*, *1141*(10), 2112–2984. https://doi.org/10.1121/1.1577562

Ellis, A. W., Young, A. W., & Critchley, E. M. R. (1989). Loss of memory for people following temporal lobe damage. *Brain*, *112*, 1469–1483. Retrieved from https://academic.oup.com/brain/article-abstract/112/6/1469/343529

Ericson, M. A., Brungart, D. S., & Simpson, B. D. (2004). Factors that influence intelligibility in multitalker speech displays. *The International Journal of Aviation Psychology*, *14*(3), 313–334. https://doi.org/10.1207/s15327108ijap1403_6

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. Retrieved from https://link.springer.com/content/pdf/10.3758/BF03193146.pdf

Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1490–1501. https://doi.org/10.1121/1.3676696

Fontaine, M., Love, S. A., & Latinus, M. (2017). Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology*, *8*(JUL), 1–9. https://doi.org/10.3389/fpsyg.2017.01180

Fox, C. J., Iaria, G., & Barton, J. J. S. (2009). Defining the face processing network: Optimization of the functional localizer in fMRI. *Human Brain Mapping*, *30*, 1637–1651. https://doi.org/10.1002/hbm.20630

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America*, *106*, 3578–3588. Retrieved from http://www.erin.utoronto.ca/~w3psy385/FreymanETAL(1999)SpeechMask_spatial.pdf

Gainotti, G. (2011). What the study of voice recognition in normal subjects and brain-damaged patients tells us about models of familiar people recognition. *Neuropsychologia*, *49*(9), 2273–2282. https://doi.org/10.1016/j.neuropsychologia.2011.04.027

Gardner, W. G., & Martin, K. D. (1995). HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, *97*, 3907–3908. https://doi.org/10.1121/1.412407

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*(1), 65–87. Retrieved from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/98153/j.1467-1770.1984.tb00996.x.pdf?sequence=1

Glyde, H., Buchholz, J. M., Nielsen, L., Best, V., Dillon, H., Cameron, S., & Hickson, L. (2015). Effect of audibility on spatial release from speech-on-speech masking. *The Journal of the Acoustical Society of America*, *138*(5), 3311–3319. https://doi.org/10.1121/1.4934732

Gobbini, M. I., & Haxby, J. V. (2006). Neural response to the visual familiarity of faces. *Brain Research Bulletin*, *71*(1–3), 76–82. https://doi.org/10.1016/j.brainresbull.2006.08.003

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*, 32–41.

https://doi.org/10.1016/j.neuropsychologia.2006.04.015

Godefroy, O., Roussel, M., Despretz, P., Quaglino, V., & Boucart, M. (2010). Age-related slowing: Perceptuomotor, decision, or attention decline? *Experimental Aging Research*, *36*, 169–189. https://doi.org/10.1080/03610731003613615

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. Retrieved from https://pdf.sciencedirectassets.com/271059/1-s2.0-S0166223600X02020/1-s2.0-0166223692903448/main.pdf?x-amz-security-token=AgoJb3JpZ2luX2VjEI3%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCIQCCdE7Ue%2Fhr9RAprz5NEelNFQDvSLSqHVYbhqa6HEiZxAIgD6PB47PtKUS

Griffiths, T. D., Rees, G., Rees, A., Green, G. G. R., Witton, C., Rowe, D., … Frackowiak, R. S. J. (1998). Right parietal cortex is involved in the perception of sound movement in humans. *Nature Neuroscience*, *1*(1), 74–79. https://doi.org/10.1038/276

Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *TRENDS in Neurosciences*, *25*(7), 348–353. Retrieved from http://tins.trends.com0166-2236/02/$-seefrontmatter

Hall, D A, Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., … Bowtell, R. W. (1999). “Sparse” temporal sampling in auditory fMRI. *Human Brain Mapping*, *7*(3), 213–223. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10194620

Hall, Deborah A., Summerfield, A. Q., Gonçalves, M. S., Foster, J. R., Palmer, A. R., & Bowtell, R. W. (2000). Time-course of the auditory BOLD response to scanner noise. *Magnetic Resonance in Medicine*, *43*(4), 601–606. https://doi.org/10.1002/(SICI)1522-2594(200004)43:4<601::AID-MRM16>3.0.CO;2-R

Hart, H. C., Palmer, A. R., & Hall, D. A. (2004). Different areas of human non-primary auditory cortex are activated by sounds with spatial and nonspatial properties. *Human Brain Mapping*, *21*, 178–190. https://doi.org/10.1002/hbm.10156

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d’. *Behavior Research Methods. Instruments. & Computers*, *27*(1), 46–51. Retrieved from https://link.springer.com/content/pdf/10.3758/BF03203619.pdf

Hawley, M. L., Litovsky, R. Y., & Colburn, H. S. (1999). Speech intelligibility and localization in a multi-source environment. *The Journal of the Acoustical Society of America*, *105*, 3436–3448. https://doi.org/10.1121/1.424670

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, *115*(2), 833. https://doi.org/10.1121/1.1639908

Helfer, K., & Freyman, R. (2008). Aging and speech-on-speech masking. *Ear and Hearing*, *29*(1), 87–98. https://doi.org/10.1097/AUD.0b013e31815d638b.Aging

Herald, S. B., Xu, X., Biederman, I., Amir, O., & Shilowich, B. E. (2014). Phonagnosia: A voice homologue to prosopagnosia. *Visual Cognition*, *22*, 1031–1033. https://doi.org/10.1080/13506285.2014.960670

Hickok, G, & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131–138. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10740277

Hickok, Gregory, & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(May), 393–402.

Holmes, E. (2018). Speech recording videos. Zenodo. https://doi.org/10.1177/0956797618779083

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science*, *29*(10), 1575–1583. https://doi.org/10.1177/0956797618779083

Holmes, E., & Johnsrude, I. (2019). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *PsyArXiv*. https://doi.org/10.31234/OSF.IO/2EBRS

Huang, Q., & Tang, J. (2010). Age-related hearing loss or presbycusis. *Eur Arch Otorhinolaryngol*, *267*, 1179–1191. https://doi.org/10.1007/s00405-010-1270-7

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*, EL236–EL242. https://doi.org/10.1121/1.3685511

Ishai, A., Schmidt, C. F., & Boesiger, P. (2005). Face perception is mediated by a distributed cortical network. *Brain Research Bulletin*, *67*, 87–93. https://doi.org/10.1016/j.brainresbull.2005.05.027

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467

Johnstone, P. M., & Litovsky, R. Y. (2006). Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults. *The Journal of the Acoustic Society of America*, *120*(4), 2177–2189. https://doi.org/10.1121/1.2225416

Kidd, G., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, *124*(6), 3793–3802. https://doi.org/10.1121/1.2998980

Kidd, G. J., & Colburn, H. S. (2017). The auditory system at the cocktail party. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.) (pp. 75–109). New York: Springer.

Kidd, G. J., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2007). Informational Masking. In W. Yost (Ed.), *Auditory Perception of Sound Sources*

(pp. 143–190). New York.

Kidd, G., Mason, C. R., Best, V., & Marrone, N. (2010). Stimulus factors influencing spatial release from speech-on-speech masking. *The Journal of the Acoustical Society of America*, *128*(4), 1965–1978. https://doi.org/10.1121/1.3478781

Kitterick, P. T., Bailey, P. J., & Summerfield, a Q. (2010). Benefits of knowing who, where, and when in multi-talker listening. *The Journal of the Acoustical Society of America*, *127*(4), 2498–2508. https://doi.org/10.1121/1.3327507

Kreitewolf, J., Mathias, S. R., & von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Frontiers in Psychology*, *8*, 1–8. https://doi.org/10.3389/fpsyg.2017.01584

Kryklywy, J. H., Macpherson, E. A., Greening, S. G., & Mitchell, D. G. V. (2013). Emotion modulates activity in the "what" but not "where" auditory processing pathway. *NeuroImage*, *82*, 295–305. https://doi.org/10.1016/j.neuroimage.2013.05.051

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex*, *21*(December), 2820–2828. https://doi.org/10.1093/cercor/bhr077

Lee, A. K. C., Larson, E., Maddox, R. K., & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, *307*, 111–120. https://doi.org/10.1016/j.heares.2013.06.010

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer Software]. *Computer Software*. https://doi.org/http://quantpsy.org/corrtest/corrtest2.htm

Levi, S. V, Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: whose familiar voices are more intelligible? *The Journal of the Acoustical Society of America*, *130*(6), 4053–4062. https://doi.org/10.1121/1.3651816

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*, 467–477. Retrieved from https://engineering.purdue.edu/~ece511/LectureNotes/reading05.pdf

Lewald, J, & Getzmann, S. (2011). When and where of auditory spatial processing in cortex: A novel approach using electrotomography. *PLoS ONE*, *6*(9), e25146-. https://doi.org/10.1371/journal.pone.0025146

Lewald, Jörg, Riederer, K. A. J., Lentz, T., & Meister, I. G. (2008). Processing of sound location in human cortex. *European Journal of Neuroscience*, *27*, 1261–1270. https://doi.org/10.1111/j.1460-9568.2008.06094.x

Lewis, M. B., Sherwood, S., Moselhy, H., & Ellis, H. D. (2001). Autonomic responses for familiar faces without autonomic reponses to familiar voices: Evidence for voice-specific Capgras delusion. *Cognitive Neuropsychiatry*, *6*(3), 217–228. https://doi.org/10.1080/13546800143000041

Lin, F. R., Yaffe, K., Xia, J., Xue, Q.-L., Harris, T. B., Purchase-Helzner, E., …

Simonsick, E. M. (2013). Hearing loss and cognitive decline in older adults. *Journal of the American Medical Association Internal Medicine*, *173*, 293–299. https://doi.org/10.1001/jamainternmed.2013.1868

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–878. https://doi.org/10.1038/nature06976

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2002). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*, 150–157. https://doi.org/10.1038/35084005

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD signal. *Annual Review of Physiology*, *66*, 735–769. https://doi.org/10.1146/annurev.physiol.66.082602.092845

Lorenzi, C., Gatehouse, S., & Lever, C. (1999). Sound localization in noise in normal-hearing listeners. *The Journal of the Acoustical Society of America*, *105*, 1810–1820. https://doi.org/10.1121/1.426719

Maeder, P. P., Meuli, R. A., Adriani, M., Bellmann, A., Fornari, E., Thiran, J.-P., … Clarke, S. (2001). Distinct pathways involved in sound recognition and localization: A human fMRI study. *NeuroImage*, *14*, 802–816. https://doi.org/10.1006/nimg.2001.0888

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179–193. https://doi.org/10.1016/j.neuropsychologia.2018.03.039

Maldjian, J. A., Laurienti, P. J., & Burdette, J. B. (2004). Precentral gyrus discrepancy in electronic versions of the Talairach atlas. *NeuroImage*, *21*(1), 450–455. https://doi.org/10.1016/j.neuroimage.2003.09.032

Maldjian, J. A., Laurienti, P. J., Burdette, J. B., & Kraft, R. A. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*, 1233–1239. https://doi.org/10.1016/S1053-8119(03)00169-1

Marrone, N., Mason, C. R., & Kidd, G. (2008). Tuning in the spatial dimension: Evidence from a masked speech identification task. *Journal of the Acoustical Society of America*, *124*, 1146–1158. https://doi.org/10.1121/1.2945710

Mathiak, K., Menning, H., Hertrich, I., Mathiak, K. a, Zvyagintsev, M., & Ackermann, H. (2007). Who is telling what from where? A functional magnetic resonance imaging study. *Neuroreport*, *18*(5), 405–409. https://doi.org/10.1097/WNR.0b013e328013cec4

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*, 233–236. https://doi.org/10.1038/nature11020

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. Retrieved

from http://apps.usd.edu/coglab/schieber/psyc707/pdf/Mishkin1983.pdf

Moon, I. J., & Sung, H. H. (2014). What is temporal fine structure and why is it important? *Korean Journal of Audiology, 18*(1), 1–7. https://doi.org/10.7874/kja.2014.18.1.1

Nakai, T., Kato, C., & Matsuo, K. (2005). An fMRI study to investigate auditory attention: A model of the cocktail party phenomenon. *Magnetic Resonance in Medical Sciences, 4*(2), 75–82. https://doi.org/10.2463/mrms.4.75

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., … Kojima, S. (2001). Neural substrates for recognition of familiar voices: A PET study. *Neuropsychologia*, *39*, 1047–1054. Retrieved from www.elsevier.com/locate/neuropsychologia

Naoi, N., Minagawa-Kawai, Y., Kobayashi, A., Takeuchi, K., Nakamura, K., & Yamamoto, J. (2012). Cerebral responses to infant-directed speech and the effect of talker familiarity. *NeuroImage*, *59*, 1735–1744. https://doi.org/10.1016/j.neuroimage.2011.07.093

Narain, C., Scott, S. K., Wise, R. J. S., Rosen, S., Leff, A., Iversen, S. D., & Matthews, P. M. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex, 13*, 1362–1368. https://doi.org/10.1093/cercor/bhg083

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*(1), 85–103. https://doi.org/10.1016/j.wocn.2005.10.004

Noble, W., & Perrett, S. (2002). Hearing speech against spatial separate competing speech versus competing noise. *Perception and Psychophysics*, *64*, 1325–1336. https://doi.org/10.3758/BF03194775

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Nygaard, L., & Pisoni, D. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*, 355–376. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9599989

Nygaard, L., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

O'Mahony, C., & Newell, F. N. (2012). Integration of faces and voices, but not faces and names, in person recognition. *British Journal of Psychology*, *103*(1), 73–82. https://doi.org/10.1111/j.2044-8295.2011.02044.x

Oxenham, A. J., Boucher, J. E., & Kreft, H. A. (2017). Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy. *The Journal of the Acoustical Society of America*, *142*(3), EL264–EL269. https://doi.org/10.1121/1.5002149

Pernet, C. R., Mcaleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G., … Belin, P. (2015). The human voice areas: Spatial organization and inter-

individual variability in temporal and extra-temporal cortices. *NeuroImage*, *119*, 164–174. https://doi.org/10.1016/j.neuroimage.2015.06.050

Plack, C. J. (2014). *The Sense of Hearing* (2nd ed.). New York: Psychology Press.

Platek, S. M., & Kemp, S. M. (2009). Is family special to the brain? An event-related fMRI study of familiar, familial, and self-face recognition. *Neuropsychologia*, *47*(3), 849–858. https://doi.org/10.1016/j.neuropsychologia.2008.12.027

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. https://doi.org/10.1038/nn.2331

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "'what'" and " 'where'" in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11800–11806. https://doi.org/https://doi.org/10.1073/pnas.97.22.11800

Relander, K., & Rämä, P. (2009). Separate neural processes for retrieval of voice identity and word content in working memory. *Brain Research*, *1252*, 143–151. https://doi.org/10.1016/j.brainres.2008.11.050

Samson, F., & Johnsrude, I. S. (2016). Effects of a consistent target or masker voice on target speech intelligibility in two-and three-talker mixtures. *The Journal of the Acoustical Society of America*, *139*(139), 1037–1046. https://doi.org/10.1121/1.4942628

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2014). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience*, *27*(2), 280–291. https://doi.org/10.1162/jocn_a_00707

Schwarzbauer, C., Davis, M. H., Rodd, J. M., & Johnsrude, I. (2005). Interleaved silent steady state (ISSS) imaging: A new sparse imaging method applied to auditory fMRI. *NeuroImage*, *29*, 774–782. https://doi.org/10.1016/j.neuroimage.2005.08.025

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*, 2400–2406. Retrieved from https://watermark.silverchair.com/1232400.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAcswggHHBgkqhkiG9w0BBwagggG4MIIBtAIBADCCAa0GCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMkKSiF1l22OGMvz15AgEQgIIBftCsC2E7Inrj2o3GVKSQSxscvpnuaYUAPbuuavhzb3NWRJy

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*(2), 100–107. https://doi.org/10.1016/S0166-2236(02)00037-1

Scott, S. K., & McGettigan, C. (2013). The neural processing of masked speech. *Hearing Research*, *303*, 58–66. https://doi.org/10.1016/j.heares.2013.05.001

Shah, N. J., Marshall, J. C., Zafiris, O., Schwab, A., Zilles, K., Markowitsch, H. J., …

Vogt-, O. (2001). The neural correlates of person familiarity A functional magnetic resonance imaging study with clinical implications. *Brain*, *124*, 804–815. Retrieved from http://www.ece.uvic.ca/~bctill/papers/facerec/Shah_2001_Brain.pdf

Shiell, M. M., Hausfeld, L., & Formisano, E. (2018). Activity in human auditory cortex represents spatial separation between concurrent sounds. *The Journal of Neuroscience*, *38*, 4977–4984. https://doi.org/10.1523/JNEUROSCI.3323-17.2018

Singh, G., Pichora-Fuller, M. K., & Schneider, B. a. (2008). The effect of age on auditory spatial attention in conditions of real and simulated spatial separation. *The Journal of the Acoustical Society of America*, *124*(2), 1294–1305. https://doi.org/10.1121/1.2949399

Singh, P. G., & Bregman, A. S. (1997). The influence of different timbre attributes on the perceptual segregation of complex-tone sequences a),b). *Journal of the Acoustic Society of America*, *102*(4), 1943–1952. Retrieved from http://webpages.mcgill.ca/staff/Group2/abregm1/web/pdf/1997_Singh_Bregman.pdf

Skuk, V. G., & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing Research*, *296*. https://doi.org/10.1016/j.heares.2012.11.004

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, *24*(8), 689–700. https://doi.org/10.3766/jaaa.24.8.6

Souza, P., & Rosen, S. (2009). Effects of envelope bandwidth on the intelligibility of sine-and noise-vocoded speech. *The Journal of the Acoustical Society of America*, *126*, 792–805. https://doi.org/10.1121/1.3158835

Stevens, A. A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research*, *18*, 162–171. https://doi.org/10.1016/j.cogbrainres.2003.10.008

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research*, *28*, 455–462. https://doi.org/10.1017/CBO9781139644549

Summers, V., & Leek, M. (1998). F0 processing and the seperation of competing speech signals by listeners with normal hearing and with hearing loss. *Journal of Speech, Language, and Hearing Research*, *41*(6), 1294–1306. https://doi.org/10.1044/jslhr.4106.1294

Tollin, D. J., & Yin, T. C. T. (2009). Sound localization: Neural mechanisms. In *Encyclopedia of Neuroscience* (pp. 137–144). Elsevier. https://doi.org/10.1016/B978-008045046-9.00267-9

Tun, P. A., O'Kane, G., & Wingfield, A. (2002). Distraction by competing speech in young and older adult listeners. *Psychology and Aging*, *17*, 453–467. https://doi.org/10.1037//0882-7974.17.3.453

Tye-Murray, N., Spehar, B., Sommers, M., & Barcroft, J. (2016). Auditory training with frequent communication partners. *Journal of Speech, Language, and Hearing Research*, *59*, 871–875. https://doi.org/10.1044/2016_JSLHR-H-15-0171

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix,

N., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*, 273–289. https://doi.org/10.1006/nimg.2001.0978

Van Engen, K., & Bradlow, A. (2007). Sentence recognition in native-and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*, *121*(1), 519–526. Retrieved from http://scitation.aip.org/content/asa/journal/jasa/121/1/10.1121/1.2400666

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829–834. https://doi.org/10.1016/0028-3932(87)90120-5

Van Lancker, Diana, Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). *Phonagnosia: A dissociation between familiar and unfamiliar voices*. *Cortex* (Vol. 24). https://doi.org/10.1016/S0010-9452(88)80029-7

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*, 48–55. https://doi.org/10.1016/S0926-6410(03)00079-X

von Kriegstein, K., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*, 948–955. https://doi.org/10.1016/j.neuroimage.2004.02.020

von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), 1809–1820. https://doi.org/10.1371/journal.pbio.0040326

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*, 367–376. https://doi.org/10.1162/0898929053279577

Warren, J. D., & Griffiths, T. D. (2003). Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain. *The Journal of Neuroscience*, *23*, 5799–5804. Retrieved from www.jneurosci.org

Warren, R. M., Riener, K. R., Bashford, J. A., & Brubaker, B. S. (1995). Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Perception & Psychophysics*, *57*(2), 175–182. Retrieved from https://link.springer.com/content/pdf/10.3758/BF03206503.pdf

Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, *91*(3), 1648–1661. https://doi.org/10.1121/1.402445

Wild, C. J., Davis, M. H., & Johnsrude, I. S. (2012). Human auditory cortex is sensitive to the perceived clarity of speech. *NeuroImage*, *60*, 1490–1502. https://doi.org/10.1016/j.neuroimage.2012.01.035

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on

attention. *The Journal of Neuroscience*, *32*(40), 14010–14021. https://doi.org/10.1523/JNEUROSCI.1528-12.2012

Wolbers, T., Hegarty, M., Buchel, C., & Loomis, J. M. (2008). Spatial updating: How the brain keeps track of changing object locations during observer motion. *Nature Neuroscience*, *11*, 1223–1230. https://doi.org/10.1038/nn.2189

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural Bases of Talker Normalization. *Journal of Cognitive Neuroscience*, *16*, 1173–1184. Retrieved from https://www.mitpressjournals.org/doi/pdfplus/10.1162/0898929041920522

Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, *12*, 900–918. Retrieved from https://journals.sagepub.com/doi/pdf/10.1038/jcbfm.1992.127

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, *15*(1), 88–99. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10755292

Yost, W. A. (2017). Spatial release from masking based on binaural processing for up to six maskers. *The Journal of the Acoustical Society of America*, *141*(3), 2093–2106. Retrieved from https://doi.org/10.1121/1.4978614

Zäske, R., Awwad Shiekh Hasan, B., & Belin, P. (2017). It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. *Cortex*, *94*, 100–112. https://doi.org/10.1016/j.cortex.2017.06.005

Zekveld, A. A., Rudner, M., Johnsrude, I. S., Heslenfeld, D. J., & Rönnberg, J. (2012). Behavioral and fMRI evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. *Brain & Language*, *122*, 103–113. https://doi.org/10.1016/j.bandl.2012.05.006

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. https://doi.org/10.1016/j.neuroimage.2006.04.199

Zhang, M., Zhang, W., Kennedy, R. A., & Abhayapala, T. D. (2009). HRTF measurement on KEMAR manikin. In *Proceedings of ACOUSTICS*. Australian Acoustical Society. Retrieved from https://www.acoustics.asn.au/conference_proceedings/AAS2009/papers/p8.pdf

Zimmer, U., Lewald, J., Erb, M., & Karnath, H.-O. (2006). Processing of auditory spatial cues in human cortex: An fMRI study. *Neuropsychologia*, *44*(3), 454–461. https://doi.org/10.1016/j.neuropsychologia.2005.05.021

Zündorf, I. C., Lewald, J., & Karnath, H.-O. (2016). Testing the dual-pathway model for auditory processing in human cortex. *NeuroImage*, *124*, 672–681. https://doi.org/10.1016/j.neuroimage.2015.09.026

Appendices

Appendix A: Ethics Approval for Chapters 2 and 3

Western Research

Research Ethics

Western University Health Science Research Ethics Board
NMREB Delegated Initial Approval Notice

Principal Investigator: Dr. Ingrid Johnsrude
Department & Institution: Social Science\Psychology,Western University

NMREB File Number: 106492
Study Title: The role of voice familiarity on speech perception in multi-talker backgrounds
Sponsor: Canadian Institutes of Health Research

NMREB Initial Approval Date: April 17, 2015
NMREB Expiry Date: April 17, 2016

Documents Approved and/or Received for Information:

Document Name	Comments	Version Date
Letter of Information & Consent	LOI and Consent Form	2015/03/05
Recruitment Items	Phone questionnaire for participant recruitment	2015/03/06
Other	Debriefing form for Participants	2015/03/05
Recruitment Items	Revised flyer	2015/04/02
Recruitment Items	Revised flyer	2015/04/02
Recruitment Items	Revised flyer	2015/04/02
Revised Western University Protocol		2015/04/02

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the above named study, as of the NMREB Initial Approval Date noted above.

NMREB approval for this study remains valid until the NMREB Expiry Date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario.

Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB.

The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 0000941.

[redacted]

Ethics Officer, on behalf of Riley Hinson, NMREB Chair or delegated board member

Ethics Officer to Contact for Further Information

Erika Basile [redacted]	Grace Kelly [redacted]	Mina Mekhail [redacted]	Vikki Tran [redacted]

This is an official document. Please retain the original in your files.

Appendix B: Ethics Approval for Chapter 4

Western Research

Western University Non-Medical Research Ethics Board
NMREB Delegated Initial Approval Notice

Principal Investigator: Dr. Ingrid Johnsrude
Department & Institution: Social Science\Psychology, Western University

NMREB File Number: 109098
Study Title: Functional Magnetic Resonance Studies of Speech Intelligibility of Familiar Voices
Sponsor: Natural Sciences and Engineering Research Council

NMREB Initial Approval Date: July 18, 2017
NMREB Expiry Date: July 18, 2018

Documents Approved and/or Received for Information:

Document Name	Comments	Version Date
Letter of Information & Consent		2017/01/17
Other	Screening questionnaire for participants to determine safety for fMRI scanner	2017/01/20
Recruitment Items	Recruitment flyer/advertisement	2017/01/20
Other	Demographic questionnaire for participants	2017/01/20
Other	Debriefing Form for SONA-recruited participants	2017/01/20
Western University Protocol	Received March 1, 2017	

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the above named study, as of the NMREB Initial Approval Date noted above.

NMREB approval for this study remains valid until the NMREB Expiry Date noted above, and Pending timely submission and acceptance of NMREB Continuing Ethics Review.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario.

Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB.

The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

[redacted]

Ethics Officer: [redacted] rn, NMREB Chair or delegated board member

EO: Erika Basile___ Grace Kelly ___ Katelyn Harris___ Nicola Morphet___ Karen Gopaul Patricia Sargeant___

Appendix C: Letter of Information and Consent Form for Chapters 2 and 3

Western SocialScience

Psychology

Project Title: The Role of Voice Familiarity on Speech Intelligibility in Multi-talker Backgrounds

Research Investigators:

Ysabel Domingo
PhD Student
The Brain & Mind Institute
Department of Psychology
e-mail: [redacted]

Emma Holmes, PhD
Postdoctoral Fellow
The Brain & Mind Institute
e-mail: [redacted]

Ingrid Johnsrude, PhD
Western Research Chair
School of Communication Sciences & Disorders and Department of Psychology
The Brain and Mind Institute
ph: [redacted]
e-mail: [redacted]

Letter of Information

You are being invited to participate in this research study that examines the effect of familiar voices on speech intelligibility because you and your spouse/friend had responded to our advertisements and expressed interest in participating in this study.

The purpose of this letter is to provide you with information required for you to make an informed decision regarding participation in this research. This letter describes the current research study and what you may expect if you decide to participate. Please read the letter carefully and ask the researcher any questions that you may have before making a decision about whether or not to participate. This form contains important information and telephone numbers, so you should keep this copy for future reference. If you decide not to participate in this study, the decision will not be held against you and will not affect your future care, employment, or eligibility for future research in any way.

Purpose of this Study:

The purpose of this study is to investigate the effect of familiar voices on understanding words in noisy environments. We are interested in determining if familiarity with a talker affects how accurately they are understood in comparison with novel voices.

Inclusion Criteria:

Individuals who are between 18-80 years old, who speak English fluently as a first language, who have no discernable accent, speech impediment, or neurological impairment, who are participating in this study as a pair (must be married for at least 5 years, or friends for at least 6 months), are eligible to participate in this study. In addition, individuals between 18-30 years old must not have any known hearing difficulties.

Exclusion Criteria:

Individuals below 18 years old or above 80 years old, or who are not fluent in English, or have a discernible accent, speech impediment, or neurological impairment; or who are not participating in this study as part of a pair, are not eligible to participate. Individuals between 18-30 years old with known hearing difficulties are also not eligible to participate. Additionally, individuals who are familiar with other participants in this study aside from their pair, or who do not consent to have their voices recorded are not eligible to participate in this study.

If you suspect you have hearing difficulties, it is recommended that you do not participate in this study. However, if you are concerned about your hearing, you may obtain a clinical assessment at the UWO audiology clinic in Elborn College. More information about the services provided at Elborn College can be found online at [redacted]

Study Procedures:

This study takes place over a maximum of five sessions. All five sessions will be the same for you and your pair, although you will participate at different times. During the first session, you will first be asked to detect quiet sounds and to record sentences into a microphone inside a sound booth. The second to fifth sessions will be completed up to two months after the first. In the second to fifth sessions, you will be asked to complete a listening task wherein you will hear sentences and will be asked to indicate the words heard in the sentence by selecting from a list of options on a screen. The sentences that you record will be used in the listening task for other participants. The sentences you record may also be digitally altered before being presented to other participants. Additionally, an excerpt of your voice recording may be played to other participants to determine if they recognize your voice. Your name will not be linked to any of

your recorded sentences, and no information about you will be released to any of the other participants.

After the listening tasks, you will be asked to fill out a demographics questionnaire about your language background, which is of critical importance to a speech intelligibility study.

It is anticipated that the entire experiment will take 120-450 minutes in total (five sessions of 30-90 minutes each). The experiment will be conducted in the Brain & Mind Institute at UWO.

Possible Risks and Benefits

There are no known or anticipated risks or discomforts associated with participating in this study.

You may not directly benefit from participating in this study. However, the results of this project may improve understanding factors that enhance speech intelligibility, which is of benefit to society in the long term.

Compensation:

You will be compensated $10 per hour for your participation the first three sessions of this study. If you do not complete the first three sessions, you will still be compensated at a pro-rated amount of $5 per half hour.

You will be compensated $12 per hour for your participation in the fourth and fifth sessions of this study. If you do not complete the fourth and fifth sessions, you will still be compensated at a pro-rated amount of $6 per half hour.

Voluntary Participation:

Participation in this study is voluntary. You may refuse to participate, refuse to answer any questions or withdraw from the study at any time with no effect on your eligibility for future research or promise of a course credit. If you choose to withdraw from the study, your data will be removed and destroyed from our database.

Confidentiality:

All data collected will remain confidential and accessible only to the investigators of this study. If the results are published, your name will not be used. Representatives of The University of Western Ontario Non-Medical Research Ethics Board may contact you or require access to

your study-related records to monitor the conduct of the research. You do not waive any legal rights by agreeing to participate in this research.

Contacts for Further Information:

If you require any further information regarding this research project or your participation in the study you may contact Ysabel Domingo, [redacted], Emma Holmes, [redacted] or Ingrid Johnsrude, [redacted]

If you have any questions about your rights as a research participant or the conduct of this study, you may contact The Office of Research Ethics [redacted], email: [redacted].

Publication:

If the results of the study are published, your name will not be used and no information that discloses your identity will be released or published. If you would like to receive a copy of any potential study results, please provide your name and contact number on a piece of paper separate from the Consent Form.

Consent:

If you agree to participate in this study, please sign the consent form on the next page. If you decide that you would *not* like to take part, please notify the experimenter and do not complete the consent form.

This letter is yours to keep for future reference.

Western SocialScience

Psychology

Consent Form

Project Title: The Role of Voice Familiarity on Speech Intelligibility in Multi-talker Backgrounds

Study Investigator's Name: Ysabel Domingo

I have read the Letter of Information, have had the nature of the study explained to me, and I agree to participate. I understand the requirements of this study and all questions have been answered to my satisfaction. I have been given sufficient time to consider the information and to seek advice if I choose to do so. I understand that I can contact the researcher(s) at any time if I have further questions. I understand that I am allowed to withdraw from the experiment at any time without giving a reason.

I understand that participation in this study requires recordings of my voice. I understand that the sentences I record will be presented to other participants in this study, but all information about me will be kept confidential.

Please check the following boxes as they apply to you. Leaving boxes unchecked does not affect your eligibility to participate in the current study.

☐I allow the researchers to use my voice recordings in future studies. I understand that all of my information will remain anonymous.

☐I agree to be contacted for invitation to participate in future research. I understand that receiving an invitation does not require/obligate me to participate in any future studies.

Participant's Name (please print): _____

Participant's Signature: _____

Date: _____

Person Obtaining Informed Consent (please print): _____

Signature: _____

Date: _____

Appendix D: Letter of Information and Consent Form for Chapter 4

Western
The Brain and Mind Institute

INFORMATION/CONSENT FORM

TITLE OF PROJECT: Functional Magnetic Resonance Studies of Intelligibility of Familiar Voices

Principal Investigator Dr. Ingrid Johnsrude [redacted]
The Brain and Mind Institute ([redacted])

1. Invitation to Participate

You are being invited to participate in a research study investigating how the brain is organized to perceive speech and voices. Previous research has identified regions implicated in speech and voice processing. The current study will investigate how speech spoken by familiar and unfamiliar voices is processed in these regions. Up to 100 people will participate in this study. This study is funded by the Natural Sciences and Engineering Research Council (NSERC).

2. Purpose of the Letter

The purpose of this letter is to provide you with information required for you to make an informed decision regarding participation in this research.

3. Purpose of this Study

The purpose of this study is to investigate how the brain is organized for the perception and comprehension of speech spoken by familiar and unfamiliar voices. Your brain's activity will be measured using a magnetic resonance imaging scanner. This scanner uses a strong magnetic field to create detailed images of brain structure and function. By taking a series of images while you perform a task, we can build up a picture of the brain areas that are active when you listen to speech.

4. Inclusion Criteria

You will be considered for the study if your first language is English and if you have spent most of your life living in Canada. In addition you must be between the ages of 18 and 80, and have no known hearing or speech disorders, no chronic neurological disorders, no prior accidents involving prolonged loss of consciousness, and no long-term psychoactive medication use.

5. Exclusion Criteria

You may be considered ineligible for the study if you are below 18 or over 80 years old, if you have elevated hearing thresholds, or if you have a foreign accent or a speech disorder. You will be considered ineligible for this study if you cannot complete the task or if it is deemed unsafe for you to be in an fMRI scanner (i.e., if you do not meet MRI safety requirements).

Western
The Brain and
Mind Institute

6. Study Procedures

If you agree to participate, your brain will be imaged while you are lying in a magnetic resonance imaging (MRI) scanner in the UWO Robarts Imaging Facility. Your heart rate and breathing may also be monitored.

Prior to your scanning session, you may be asked to complete a recording session where we will record you reading sentences aloud into a microphone. During this session, we will also ask that you complete a questionnaire about the languages you know and how long you and your pair have known each other. This session will take 60-90 minutes. If you have participated in a previous voice study in our lab before and we already have your voice recordings and completed questionnaire, you will not be asked to complete another set of recordings.

Additionally, before the scan session you may also be asked to complete a training session in an audiobooth (sound attenuated chamber) at the Brain and Mind Institute (BMI). This session may take 30 to 60 minutes and will involve sitting at a computer while audio stimuli are presented over headphones and visual instructions presented on a monitor. No MRI will be used in this part. You may be asked to make responses using a keyboard in this training session, which will familiarize you with the types of tasks and stimuli you will be asked to respond to within the MRI scanner. This audiobooth session may take place on a different day to, or earlier on the same day as, your scanning session. If you cannot complete the training task, you may not be invited to complete the imaging session in the scanner.

For your scanning session at the UWO Robarts Imaging Facility, the entire session may last up to 2 hours, including getting ready for the study, being positioned in the magnet, etc. This part will be completed in a single session.

a) You will begin by filling out a checklist and questionnaire to make sure you are eligible. This will be completed first, and will take up to 10 minutes. If you are pregnant or are trying to conceive you will not be eligible. If there is any uncertainty regarding whether or not you are pregnant and you want to participate in the study, a pregnancy test must be done prior to the experiment.

b) Your informed consent to participate in the study will be obtained. If you agree that your data may be anonymously shared with other researchers, we will ask you to answer some questions about yourself (again, this information will be kept anonymous) so that your imaging

Western
The Brain and
Mind Institute

data will be more informative.

c) You must wear clothing containing no metal or bring a change of clothing that does not contain metal. Metal in zippers, snaps, and the wire and metal clasps in some bras can interfere with the imaging. Many shoes contain metal as well. For your own safety, you will be asked to remove or change out of any clothes that contain metal that will be near the area being imaged, and you will be asked to remove your shoes. For imaging the brain, the snaps and zippers in jeans or other trousers are far enough from the area being imaged that they do not cause a problem.

d) For most studies, you will be given a pair of earplugs to block out the loud noise of the scanner. This makes the process more comfortable for you. You will also be wearing headphones with ear muffs, which block external sound but enable you to hear the experimental sounds clearly. Although the operator/researcher can talk to you via an intercom, during the scan it will be too noisy, so you will be given an alarm bulb so that you can call us during the scan if something is wrong or you want to come out.

e) You will be asked to lie on your back on the bed of the magnet. Pillows will be placed under your legs for comfort and a blanket will be placed over your legs if you wish. A head coil will be placed over your head. This is fitted with a mirror so that you can see out of the magnet towards your head or feet. You and the bed will then slide into a long tube (the magnet).

f) You will need to keep very still while the images are taken. To help you, we will make you as comfortable as possible and we will pack soft foam around your head if needed.

g) During the scan you will be listening to sounds, like beeps and noise bursts, speech sounds, spoken words, or sentences. You may be asked to press a button or speak in response to sounds of a particular type, and the occurrence of this response will be recorded. You will be familiarized with the task before the experiment begins so that you know exactly what to expect. Your brain will be imaged repeatedly during the experiment.

h) The scanning session will take about one hour, although functional scanning will not last more than about 45 minutes of this time.

i) At the end of the session, additional images will be taken of the anatomy (or structure) of your brain (about 15 minutes).

Western
The Brain and
Mind Institute

7. An explanation of the special research techniques that will be used:

Part of your participation in this study will involve a research test with Magnetic Resonance Imaging (MRI) system, a common medical diagnostic tool that uses a strong magnetic field, a low frequency magnetic field, and a radio frequency field. No X-rays are used. As with any technology there is a risk of death or injury. For MRI the risk of death is less than 1 in 10 million and the risk of injury is less than 1 in 100,000. These risks do not arise from the MRI process itself, but from a failure to disclose or detect MRI incompatible objects in or around the body of the study participant or the scanner room. It is therefore very important that you answer all the questions honestly and fully on the MRI screening questionnaire.

Almost all the deaths and injuries related to MRI scans have occurred because the MRI operator did not know that surgically implanted metal hardware (such as a cardiac pacemaker) was present inside the study participant during the MRI scan. Other remote risks involve temporary hearing loss from the loud noise inside the magnet. This can be avoided with ear headphone protection that also allows continuous communication between the study participant and staff during the scan.

For comparison, the risk of death in an MRI is similar to travelling 10 miles by car, while the risk of injury during an MRI is much less than the risks associated with normal daily activities for 1 hour.

You may not be allowed to continue in this research study if you are unable to have a MRI scan because, for example, you have some MRI incompatible metal in your body, you may be pregnant or attempting to become pregnant, or you may have a drug patch on your skin that contains a metal foil. Should you require a medical MRI scan in the future, the final decision as to whether you can be scanned will be made by a qualified physician considering all the risks and benefits.

This MRI machine uses a strong magnet and radio waves to take images of the body interior. You will be asked to lie on a long narrow couch for a certain amount of time [usually 1 hour, but in some cases it may take 2 hours] while the machine gathers data. During this time you will be exposed to magnetic fields and radio waves. You will not feel either. You will, however, hear repetitive tapping noises that arise from the magnets that surround you. You will be provided with earplugs or headphones that you will be required to wear to minimize the sound and protect your hearing. The space within the large magnet in which you lie is somewhat confined, and participants with claustrophobia may find this experience uncomfortable, although we have taken many steps to relieve the "claustrophobic" feeling. There are no known significant risks with this procedure at this time because the radio waves and magnetic fields, at the strengths used, are thought to be without harm.

The exception is if you have a cardiac pacemaker, or a metallic clip in your body (e.g., an aneurysm clip in your brain), have severe heart disease, body piercings, tattoos containing metallic ink or slow release pharmaceutical skin patches – you will not be able to take part in the study if you have any of these items at the time of scanning.

There is a possibility that you will experience a localized twitching sensation due to the magnetic field changes during the scan. This is not unexpected and should not be painful. However, you can stop the scan at any time. The magnetism and radio waves do not cause harmful effects at the levels used in the MRI machine. However, because the MR scanner uses a very strong magnet that will attract metal, all metallic objects must be removed from your person before you approach the scanner. In addition, watches and credit cards should also be removed as these could be damaged. (These items will be watched for you while you are in the scanner).

8. Risks/Side-Effects:

There are no known risks involved with magnetic resonance imaging. However, the MR scanner uses a very strong magnet that will attract metal. Therefore ALL metallic or magnetic objects MUST be removed from your person before you approach the scanner.

9. Benefits

There are no known benefits to you as a result of participating in this study.

10. MRI Exclusions:

If you have any history of head or eye injury involving metal fragments, if you have some type of implanted electrical device (such as a cardiac pacemaker), if you have severe heart disease (including susceptibility to heart rhythm abnormalities), you should NOT have an MRI scan unless supervised by a physician. Additionally you should not have a MRI scan if you have conductive implants or devices such as skin patches, body piercing or tattoos containing metallic inks because there is a risk of heating or induction of electrical currents within the metal element causing burns to adjacent tissue.

11. Confidentiality

The findings of this study will be reported in scientific journals but your name will remain confidential. Data from your images will be stored on a secure computer system and identified only with the date and a study participant code. Only the researchers directly related to this study will have access to the data files and the study participant codes. You will not be identified in any publication or reports. The paper copy of your demographic questionnaire will be stored for 5 years, following which it will be destroyed. Questionnaire responses and data collected in the MRI scanner

Western
The Brain and Mind Institute

will be stored electronically and will be accessible by the principal investigator (Dr. Ingrid Johnsrude), the investigator's research team, and representatives from the Health Sciences Research Ethics Board (to ensure that study is following the proper laws and regulations). If you agree on the consent form to share your data with other researchers, this will be accomplished via your study participant code (not your name).

Like faces, brains come in all shapes and sizes, and there are many normal variations, but some variations are less usual. Although this is not a diagnostic scan and any images obtained are for research purposes only, it is possible that the MR scan may disclose an unknown abnormality. In this event, a medical imaging specialist will be asked to review the images and we would send a report to your physician. The researchers directly involved with this procedure do not have the credentials to diagnose medical conditions.

12. Voluntary Participation/Freedom to withdraw or participate:

Your participation in this study is voluntary. You may withdraw from this study at any time and your withdrawal will not affect your future medical care, academic standing, or career. You are not obliged to answer any questions that you find objectionable or which make you feel uncomfortable. If you decide to withdraw for any reason, all information collected will be destroyed.

13. Liability:

In the event that you are injured as a result of the study procedures, medical care will be provided to you until resolution of the medical problem has been identified.

By signing this consent form, you do not waive your legal rights nor release the investigator(s) and sponsors from their legal and professional responsibilities.

14. Compensation: Some studies compensate for participant's expenses and inconvenience.

You will receive $25 per hour (or $12.50 per half hour) to reimburse time and travel costs. In addition, if this is your first time participating in an MRI study at the UWO Robarts Imaging Facility, you will receive a black and white picture of your brain structure, which will be emailed to you. If you were recruited through Western's Psychology Research Participation Pool (SONA), you will receive participation credits for your time. You will have the pleasure of knowing that you have made a contribution to our understanding of the relationship between brain and behaviour.

Western
The Brain and
Mind Institute

STUDY PARTICIPANT STATEMENT AND SIGNATURE SECTION:

I have read and understand the consent form for this study. I have had the purposes, procedures and technical language of this study explained to me. I have been given sufficient time to consider the above information and to seek advice if I chose to do so. I have had the opportunity to ask questions which have been answered to my satisfaction.

I have named Dr. _____at_____as the physician to be contacted for follow-up purposes. I am voluntarily signing this form. I will receive a copy of this consent form for my information. If at any time I have further questions, problems or adverse events, I can contact

Dr. Ingrid Johnsrude (P.I.)
The Brain and Mind Institute, UWO
[redacted]
e-mail: [redacted]

Janet Wallace
Robarts Imaging Research Manager
Phone: [redacted]
e-mail: [redacted]

If I have questions regarding my rights as a research study participant I can contact The Office of Research Ethics [redacted] mail: [redacted]

Circle as appropriate:

I **agree/do not agree** to have data from this study shared with other researchers at UWO. I understand that my data will be identified only by number, not by my name.

I **agree/do not agree** to permit researchers to use my voice recordings in future studies. I understand that my voice recordings will be identified only by number, not by my name.

I **am interested/am not interested** in being contacted so that I may participate in other studies in the UWO Robarts Imaging Facility and at the Brain and Mind Institute. I understand that being contacted for research opportunities does not obligate me to participate in any future projects.

By signing this consent form, I am indicating that I agree to participate in this study.

_____ _____
Signature of Volunteer Date

_____ _____
Signature of Witness Date

STATEMENT OF INVESTIGATOR:

I, or one of my colleagues, have carefully explained to the study participant the nature of the above research study. I certify that, to the best of my knowledge, the study participant understands clearly the nature of the study and demands, benefits, and risks involved to participants in this study.

_____ _____
Signature of Principal Investigator Date

Version 1 (January 17, 2017)

Appendix E: Demographics Questionnaire for Chapters 2-4

FOLLOW-UP QUESTIONNAIRE
The role of voice familiarity on speech perception in multi-talker backgrounds

1. Participant ID (to be filled by experimenter): _____

2. Month and Year of birth: _____

3. Are you male or female? _____

4. Describe the nature of your relationship with your pair (select all that apply):

spouse friend

dating roommate

relative:_____ other: _____

4. How many months or years have you and your spouse been married/ have you and your friend known each other? _____ years _____ months

5. How many hours a week do you say you talk to each other (phone, Skype, FaceTime, face-to-face)? _____

6. What is your first language? _____

7. Where did you grow up? _____

8. Have you ever lived abroad or outside of Ontario? _____

   a. If yes, where? _____

   b. If yes, for how long? _____

9. What other languages do you speak? _____

   a. Are you fluent? _____

   b. When was this language acquired? _____

10. Do you have any hearing difficulties or do you wear a hearing aid? _____

11. Have you been diagnosed with a neurological impairment? _____

12. Are you at all color-blind? Or are there any colors that you know you cannot see?_____

Version 2015/10/28

Appendix F: Spatialized Speech Perception Questionnaire

PID: _____ Date:_____

How strongly did you perceive the sentences as coming from different directions?

Please circle your response.

Not at all			Moderately			Very strongly
1	2	3	4	5	6	7

Curriculum Vitae

Name: Ysabel Domingo

Post-secondary Education and Degrees:
University of Toronto
Mississauga, Ontario, Canada
2008-2012 H.B.Sc.

The University of Western Ontario
London, Ontario, Canada
2012-2014 M.Sc.

The University of Western Ontario
London, Ontario, Canada
2014-2019 Ph.D. (degree expected August 2019)

Honours and Awards:
Western Graduate Research Scholarship
2014-2018

Association for Research in Otolaryngology Travel Award
2016

Best Oral Presentation, HRS Graduate Research Forum
2014

Recognition of Exceptional Academic Excellence
2012

Related Work Experience:
Teaching Assistant
The University of Western Ontario
2014-2018

Research Assistant, Communicative Participation Lab
The University of Western Ontario
2012-2013

Lab Manager, Infant Language and Speech Lab
University of Toronto Mississauga
2010-2012

Publications:

Domingo, Y., Page, A., Adams, S., & Jog, M. (accepted). Examining the Speech Intelligibility of Individuals With Oromandibular Dystonia Receiving

Botulinum Toxin: A Series of Cases. *Canadian Journal of Speech-Language Pathology and Audiology.*

Domingo, Y., Holmes, E., & Johnsrude, I. (in revision). The benefit to speech intelligibility of hearing a familiar voice. *Journal of Experimental Psychology: Applied.*

Holmes, E., **Domingo, Y.,** & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science, 29,* 1575-1583.

Dykstra, A., **Domingo, Y**., Adams, S., & Jog, M. (2015). Examining speech intelligibility and self-ratings of communicative effectiveness in speakers with oromandibular dystonia receiving botulinum toxin therapy. *Canadian Journal of Speech-Language Pathology and Audiology, 39,* 334-345.

Posters and Presentations:

Domingo Y., Holmes, E., & Johsrude, I. "Using spatial release from masking to estimate the magnitude of the familiar-voice intelligibility benefit", Association for Research in Otolaryngology, San Diego, CA, February 2018.

Holmes, E., **Domingo, Y.,** & Johnsrude, I. "Talker familiarity improves speech understanding in the presence of simultaneous talkers", Association for Research in Otolaryngology, San Diego, CA, February 2018.

Holmes, E., **Domingo, Y.,** & Johnsrude, I. "Improvement to speech intelligibility is for familiar voices is robust to manipulations of fundamental frequency and vocal tract length", Association for Research in Otolaryngology, Baltimore MA, February 2017.

Domingo, Y., Holmes, E., Johnsrude, I. "How voice familiarity facilitates speech intelligibility in multi-talker situations", Association for Research in Otolaryngology, San Diego CA, February 2016.

Domingo, Y. "Using botulinum toxin type A as speech intelligibility intervention in oromandibular dystonia," Speech & Language Sciences Seminar, Western University, November 4, 2015.

Dykstra, A. D., Mancinelli, C., **Domingo, Y.,** Dworschak-Stokan, A., & Husein, M. "Examining communicative effectiveness in adult speakers with velopharyngeal insufficiency", American Speech and Hearing Association Annual Convention, November 20, 2014, Orlando, FLA.

Dykstra, A.D., **Domingo. Y**., Adams, S., & Jog, M. The effect of botulinum toxin type A on speech intelligibility and self-ratings of communicative effectiveness by

speakers with oromandibular dystonia. Conference on Motor Speech, Sarasota, FLA., February 2014

Domingo, Y., Dykstra, A.D., Adams, S.G., Johnson, A., & Jog, M. "The effect of botulinum toxin type A (Botox) on speech intelligibility in oromandibular dystonia", Health & Rehabilitation Sciences Graduate Research Forum, February 2014

Domingo, Y., Dykstra, A.D., Adams, S.G., Johnson, A., & Jog, M. "Evaluating the impact of botulinum toxin A injections on speech intelligibility in oromandibular dystonia", Aging, Rehabilitation and Geriatric Care & Faculty of Health Science Symposium, "Partnerships and Possibilities in Health Research", Western University, February 7, 2014.

Mancinelli, C., **Domingo, Y.**, Dykstra, A.D., Dworschak-Stokan, M.S., & Husein, M. "An examination of speech intelligibility, hypernasality, and self-ratings of communicative effectiveness in adults with velopharyngeal insufficiency", Health & Rehabilitation Sciences Graduate Research Forum 2014.

Mancinelli, C., **Domingo, Y.**, Dykstra, A.D., Dworschak-Stokan, M.S., & Husein, M. "An exploration of the relationships between speech intelligibility, hypernasality, and self-ratings of communicative effectiveness in adults with velopharyngeal insufficiency", Aging, Rehabilitation and Geriatric Care & Faculty of Health Science Symposium, "Partnerships and Possibilities in Health Research", Western University, February 7, 2014.

Domingo, Y., Dykstra, A.D., Jablecki, D., Adams, S.G., Johnson, A., & Jog, M. "An evaluation of speech intelligibility based on technique in Oromandibular Dystonia", Health & Aging Graduate Research Conference "Urban Health and Well-being" McMaster University, March 1, 2013

Jablecki, D., Dykstra, A.D., **Domingo, Y.,** & Jog, M. "Examining levels of speech intelligibility in an individual with Oromandibular Dystonia", Health & Aging Graduate Research Conference "Urban Health and Well-being" McMaster University, March 1, 2013

Domingo, Y., Dykstra, A.D., Jablecki, D., Adams, S.G., Johnson, A., & Jog, M. "A comparison of speech intelligibility measures obtained from three measurement techniques in Oromandibular Dystonia", Health & Rehabilitation Sciences Graduate Research Forum 2013: "Sowing Seeds of Ideas for Fruitful Trees", Western University, February 6, 2013

Domingo, Y., Dykstra, A.D., Jablecki, D., Adams, S.G., Johnson, A., & Jog, M. "Evaluating speech intelligibility based on technique in Oromandibular Dystonia", Aging, Rehabilitation & Geriatric Care Research Centre & Faculty of

Health Science Symposium “Research to Action: Technology, Innovation & Health”, Western University, February 1, 2013

Jablecki, D., Dykstra, A.D., **Domingo, Y.,** Adams, S.G., & Jog, M. "The effect of task on speech intelligibility in Oromandibular Dystonia: A case report.", Aging, Rehabilitation & Geriatric Care Research Centre & Faculty of Health Science Symposium “Research to Action: Technology, Innovation & Health”, Western University, February 1, 2013

Cheema, S., **Domingo, Y.,** Erwood, A., Loong, M., Singh, G., & Pichora-Fuller, K. “Teleaudiology: Factors related to future applications”, Research Opportunity Program Fair, University of Toronto Mississauga, February 2011

Leadership and Community Involvement

Psychology Colloquium Committee, Member
2015-2017

Industry Fair Head Organizer, Western WiNS
2017-2018

Co-Chair, SOGS Academic Committee
2016-2017

HRS Student Field Mentor
2013-2014

HRS Research Forum Committee, Member
2013-2014