Specificity effects in spoken word recognition and the nature of lexical representations in memory

Dorina Strori

PhD

University of York

Psychology

March
2016

Abstract

When we hear speech, besides the linguistic content, we may gain a great deal of information about the speaker from their voice, such as their identity, age, gender, or emotional state. No word is uttered the same way, even by the same talker, so one of the main challenges in spoken word recognition research is to understand the cognitive processes that underlie the processing of a complex signal like speech in the presence of high variability. Previous research has shown that listeners encode both linguistic and talker-related voice properties of the speech signal in their memory representations. Speaker variability is not the only variable we encounter; we frequently hear speech in varying auditory contexts as well. Recent evidence suggests that auditory background details, such as non-linguistic sounds co-occurring with spoken words, may be incorporated in our lexical memory. Here, I first test the hypothesis that the acoustic glimpses of words (left-overs) produced by masking from the associated sounds, rather than the sounds per se, are retained in memory. I then identify and examine the role of a novel element in the relationship between a spoken word and its associated sound, perceptual integrality, in the retention of sounds in memory. Last, I investigate the potential impact of the unique pairwise associations between words and sounds on the encoding of sounds in memory. My findings suggest that background sounds can be encoded in memory, but only in certain conditions. Specifically, this can happen when the auditory episode of the word(s) consists of highly contrasted acoustic glimpses of the same word(s), and when the sounds are made integral to, hence more difficult to perceptually segregate from the words, through intensity modulation.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my supervisor, Sven Mattys, for the opportunity to carry out this project within an excellent multidisciplinary research network such as INSPIRE. I am beyond grateful for his mentorship, support, and guidance that he has provided throughout the project, including comments on earlier drafts of the chapters that constitute this thesis. A special thank you goes to my colleague and friend, Johannes Zaar, who contributed to the preparation of stimuli in three experiments. Thanks to him, I have learned more than I would have thought on signal processing and speech intelligibility. I would also like to extend my gratitude to Martin Cooke, who helped with several computational analysis of the stimuli, as well as my second supervisor, Odette Scharenborg, for her support during my academic placement at the Radboud University Nijmegen.

The members of my Thesis Advisory Panel, Gareth Gaskell and Quentin Summerfield, provided useful feedback at regular intervals throughout the project. I greatly appreciate their ideas and constructive feedback on the work that I presented.

My fellow colleagues and friends in the INSPIRE network have provided intellectual support, but above all great friendship, for which I am beyond grateful.
My dear friend, Huarda Valdes-Laribi, has been one of the greatest blessings of this PhD. She is a true sunshine and her precious friendship, a true gift. I am also thankful for the good chats and laughter with my friends at the University of York, they have kept my spirits high.

Finally, I could not have completed this thesis without the constant support of my family. I owe to my parents, Flora and Irakli Strori, the motivation to pursue academic research. Without their selfless dedication and hard work, I would not be where I am now. My sister, Marsida Strori, has always been the best friend, with the tightest hug.

I dedicate this thesis to my son, Emir Arber Bombaci, who never ceases to amaze and inspire me, and my life companion, Ahmet Bombaci, who has made life happy and easy these three years with his love, optimism and support.

Author’s declaration

This thesis was written by myself and represents original work, that has not previously been presented for an award at this, or any other, University. All experiments were designed by the candidate with assistance from the supervisors, Professor Sven Mattys and Dr. Odette Scharenborg. Professor Martin Cooke helped with the computational analysis of the experimental stimuli in Chapters 3 and 4. Johannes Zaar contributed to the preparation of the stimuli used in the experiments presented in Chapters 4 and 5. April Pufahl kindly shared the environmental sounds used in the experiments described in Chapter 5. All testing and statistical analyses were conducted by the candidate. All sources are acknowledged as references.

The research was carried within the FP7 Marie Curie Initial Training Network “Investigating Speech Processing In Realistic Environments (INSPIRE)”, and was funded by the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-PEOPLE-2011-290000.

Conference presentations and proceedings

Strori, D., Zaar, J., & Mattys, S. (2015). Honking is just noise, or just about: effects of energetic masking and speech modulation on spoken word recognition. Paper presented at the 169th Meeting of the Acoustical Society of America, Pittsburgh, PA, USA.

Strori, D., Zaar, J., Scharenborg, O., & Mattys, S. (2015). Honking is just noise, or just about: effects of energetic masking and speech modulation on spoken word recognition. Paper presented at the 3rd International Conference on Cognitive Hearing Science for Communication, Linköping, Sweden.

Strori, D., Zaar, J., Scharenborg, O., & Mattys, S. (2015). The mental lexicon: strictly or leniently lexical? Paper presented at the Experimental Psychology Society’s Meeting, London, UK.

Strori, D., Scharenborg, O., & Mattys, S. (2014). The mental lexicon: strictly or leniently lexical? Paper presented at the Psychonomic Society’s 55th Annual Meeting, Long Island, CA, USA.

Strori, D., Scharenborg, O., & Mattys, S. (2014). Effects of energetic masking on specificity effects in spoken word recognition. Paper presented at the British Society-of Audiology 5th Annual Meeting, Keele, UK.

Chapter 1

Specificity effects and spoken word recognition

1.1. Introduction

A crucial issue regarding speech perception and spoken word recognition is the understanding of the cognitive processes involved in a listener's comprehension of the intended spoken message in the face of the high variability displayed by the speech signal. The investigation of this complicated issue has been organized around several, more specific questions that include: the perceptual analysis stages involved, the processing computations at each stage, and the nature and content of the representations of spoken words in memory. The latter, which is also the focus of this introductory review, has been a matter of debate and controversy for decades, due to the complex nature of speech.

1.2. Speech as a complex and integral stimulus

1.2.1. Complexity

Speech is a complex stimulus, the result of the intertwining between two dimensions: a "linguistic" and an "indexical" one (Abercrombie, 1967). The linguistic component conveys crucial information necessary for the identification of linguistic (phonemic, lexical) contrasts (e.g., Allopenna et al., 1998; Gaskell & Marslen-Wilson, 2002). On the other hand, the indexical properties provide information regarding personal characteristics of the talker, such as: acoustic correlates of their identity (Nygaard et al., 1994; Palmeri et al., 1993; Pisoni, 1997), prosody (Watson et al., 2008), and vocal emotional cues (Morton & Trehub, 2001). These cues can be specific to the point of identifying the talker (Fellowes et al., 1997; Van Lancker et al., 1985), but even when this is not the case, they still convey other talker-related information, such as gender, age, health, and emotional state (Kreiman, 1997; Peterson & Barney, 1952, also cited in Pisoni & Levi, 2007). In Abercrombie's conception, speakers provide a personal "medium" for linguistic messages, implying that certain indexical properties in this medium may be “extra-linguistic".

1.2.2. Integrality

Speech can also be defined as an *integral* stimulus, in which the two components co-exist *simultaneously* and are *integral* to each other, such that it is impossible to segregate one from the other (Vitevitch, 2003). An important characteristic of an integral stimulus is that a change in one of the dimensions (the relevant one) affects the other one as well (the irrelevant one). Vitevitch (2003) notes that although different aspects of the acoustic signal are correlated with linguistic (Zue & Schwartz,1980) and indexical components (Bricker & Pruzansky, 1976; Hecker, 1971), there is evidence from several studies using speeded classification tasks (Garner, 1974) suggesting that spoken language is an integral stimulus with these two dimensions (e.g., Jerger et al., 1995; Jerger et al., 1993). For example, in several studies by Jerger and colleagues, participants selectively attended to either the linguistic (word) or indexical dimension (talker’s gender), while ignoring the other. In both cases, the classification performance of listeners for the relevant dimension was affected by variation in the irrelevant dimension, suggesting that spoken language displays the properties of an integral stimulus.

Given this complex and integral nature of the speech signal, the ensuing variability is also complex. Besides variation in the linguistic dimension, listeners have to deal with the one arising in the indexical dimension as well. Individual talkers differ in their voice properties due to several factors, such as the physical shape and length of the oral, and nasal vocal tract cavities, which in turn affects the acoustic structure of the speech signal (Mullennix et al., 1989). Different talkers display different fundamental frequencies (Van Lancker, Kreiman, & Emmorey, 1985), speaking rates (Van Lancker, Kreiman, & Wickens, 1985), voice onset times (Allen et al., 2003), frication noise (Newman et al., 2001), and realizations of vowels (Bradlow, Torretta, & Pisoni, 1996). A single word may not be uttered the same twice, even by the same talker.

In summary, the acoustic cues to spoken words may vary in several aspects: phonetic, phonological and lexical context, as well as individual talker properties. The way the perceptual system deals with this high degree of variability during the mapping of acoustic signals to lexical representations has been a crucial, as well as controversial issue in speech perception and spoken word recognition research. The present review is only concerned with the variability arising in the indexical dimension of speech; the next section will present an overview of how the traditional view of speech perception has treated it.

1.3. The traditional view of speech perception

The traditional, or “analytical” view of speech perception has its roots in generative linguistic approaches, that engaged a formalist view and focused on explaining two related problems: describing the linguistic competence of native speakers; and finding and explaining systematic regularities and common patterns displayed by all natural languages. To this end, linguists endorsed several assumptions about speech that are of a strong abstractionist nature and rely on symbolic-processing approaches. More specifically, the assumption is that speech is structured in systematic ways and that the linguistic information can be represented economically as a linear sequence of abstract, discrete symbols using an alphabet of conventional phonetic symbols. Further, the regularities and common patterns observed in natural languages could be conveniently explained by sets of formal rules operating on these abstract symbols (Pisoni & Levi, 2007). Since these segmental representations of speech were designed to code only the linguistically significant information, they were assumed to be free of any redundant or incidental information in the speech signal that did not have any linguistic relevance (Licklider, 1952; Twaddell 1952; also mentioned in Pisoni & Levi, 2007 and Pisoni, 1997).

This approach to speech has been embraced across a diverse range of disciplines that study speech processing, such as psycholinguistics, computational linguistics, cognitive and neural sciences, speech and hearing sciences, as well as engineering views of modelling speech intelligibility (see also Jusczyk & Luce, 2002; Pisoni & Levi, 2007). Importantly, it brings along several fundamental theoretical assumptions that directly impact wider theoretical accounts regarding the nature and content of lexical representations. A review of all these assumptions and their influence on theoretical accounts is beyond the scope of the present introduction, where the focus is on the indexical component of the speech signal and how its variability has been treated by different theoretical views of speech perception. Therefore, especially relevant for the present review is that the traditional view of speech relies greatly on a set of psychological processes whose function is to "normalize" the high degree of variation present in the speech signal. More specifically, the general assumption has been that the normalization process is needed during the perceptual stages in order to reduce the acoustic-phonetic variability in the speech signal and make physically different signals perceptually equivalent by bringing them into conformity with some sort of common standard or referent (Pisoni, 1997). Put in alternative terms, *normalization* refers to the filtering of the indexical variation to allow for extracting only the linguistically relevant information for speech recognition. In the normalization phase, representations of stimuli that vary in acoustic detail but are part of the same perceptual category are treated as identical (Jusczyk & Luce, 2002; Lachs et al., 2003).

1.3.1. The abstract view of the lexicon

The abstract view of the lexicon is grounded on the traditional view of speech reviewed above. Namely, listeners normalize the highly variable speech signal and map the acoustic–phonetic input onto abstract phonetic representations, stored in the long-term memory system referred to as the mental lexicon (e.g., Cutler, 2008; Fowler & Smith, 1986; Stevens, 2002). In support of this claim, several studies have specifically looked for invariant, abstract categories of the speech input in the form of acoustic features (Blumstein & Stevens, 1980), or articulatory gestures (Fowler, 1986; Fowler & Rosenblum, 1991).

Abstractionist models postulate that word recognition is subserved by abstract pre-lexical representations. The speech input is mapped onto these abstract phonological representations which, depending on the account, may be features (Gaskell & Marslen-Wilson, 1997), phonemes (Norris, 1994), features and phonemes (McClelland & Elman, 1986), or syllables (Mehler, 1981). Lexical representations are then defined in terms of these sub-lexical prototypes. Importantly, only information relevant for lexical discrimination is retained in the representations. Indexical variability in the speech signal is treated as irrelevant information and discarded at an early stage of encoding.

Several computational models of spoken word recognition have implemented the abstractionst approach (e.g., Distributed Cohort Model: Gaskell & Marslen-Wilson, 1997, 1999, 2002;

Marslen-Wilson & Welsh, 1978; TRACE: McClelland & Elman, 1986; SHORTLIST: Norris, 1994; PARSYN: Luce et al., 2000; see Jusczyk & Luce, 2002 for a detailed review).

Various phenomena have been typically explained in terms of this type of abstraction, such as the interpretation of variant forms of words (e.g., postman versus pos'man) as the same canonical lexical form (Cutler, 2008; Sumner & Samuel, 2005), or the different effects that phoneme transition probability has on the processing of spoken words and non-words (Vitevitch, 2003; Vitevitch & Luce, 1998). The robust nature of speech comprehension in the face of linguistic and indexical variability, as well as the ease with which listeners can comprehend speech from talkers whose voices they are hearing for the first time, have always served as crucial motivating factors for the abstractions view of the lexicon. Indeed, our subjective experience is that understanding an utterance from a new talker— for instance, when a stranger in the street asks for directions, or a shopkeeper names a price—is usually no harder than understanding the same utterance from a speaker whose voice is familiar to us.

The abstract view has been criticised on several grounds, but the one that is crucial to the present review regards the large body of evidence showing that spoken word recognition is sensitive to changes in surface characteristics of the signal, such as talker-related properties. The next section will first introduce an overview of this literature, and will then follow with a detailed review of some relevant studies.

1.4. Indexical Effects in Spoken Word Processing

An extensive number of studies have shown that talker-specific indexical information is not stripped off the speech signal, but is retained in memory and can affect the processing of spoken words (e.g., Church & Schacter, 1994; Luce & Lyons, 1998; Goldinger, 1996; Nygaard et al., 1994; Palmeri et al., 1993). This has been manifested in what is typically referred to in the literature as *indexical effects.*

1.4.1 Definition of the indexical effects

An indexical effect arises when changes in talker-related indexical information affect the identification or memory of spoken words. The change can be within the same talker (change in the speaking rate, emotional tone), or between talkers (change in the talker gender or identity within the same gender). Indexical effects have been usually explained in terms of a *processing cost* to the perceptual system, caused by stimuli that mismatch on the talker-related voice details, and are typically measured in the form of a decrease in accuracy, increase in reaction latencies, or both (e.g., Mullennix & Pisoni, 1990; Luce & Lyons, 1998; McLennan & Luce, 2005). Alternatively, they have been interpreted in terms of a *performance advantage*, induced by the invariable/consistent instances of the stimulus, in this case, the same talker voice (Pufahl & Samuel, 2014).

A typical indexical study consists of an exposure (study), sometimes a short delay, followed by memory test. In the exposure phase, listeners perform a certain task that encourages the stimuli encoding in memory, and then in the test phase, they complete a memory task, with the stimuli repeated either in the same talker voice (the most common manipulation), or in a different talker voice. Besides the most common manipulation (i.e., the voice change) between exposure and test, several other voice properties have also led to the emergence of indexical effects, including gender, emotional intonation, phrasal intonation (statement/question), fundamental frequency (e.g., Church & Schacter, 1994); voice-onset-time (Allen & Miller, 2004); and speaking rate (Bradlow et al., 1999). On the other hand, no effect of the amplitude change has been reported (Bradlow et al., 1999; Church & Schacter, 1994). The effect has also shown to persist over time, with an effect still present (advantage for stimuli repeated in the same voice) after a week delay (Goldinger, 1996).

In summary, indexical effects indicate that listeners seem to encode not only *what* was spoken, but also specific details about *who* said it and *how*. This enriched, specific encoding in turn can improve future understanding of previously encountered speakers. The next section provides an overview of typically used encoding and memory tasks, as well as some controversies regarding their robustness. A more detailed review regarding this issue is provided in **Chapter 2.**

1.4.2. Encoding and memory tasks – Controversies

1.4.2.1. Encoding tasks

Encoding tasks are usually classified in terms of the processing depth they impart on the stimuli. They range from shallow (e.g., categorise words according to the gender of the talker; also mentioned in Pufahl & Samuel, 2014); to moderate (e.g., reporting of the initial phoneme of the word), and (identifying the syntactic class of the words; Goldinger, 1996). In some encoding tasks, attention has been directed to the voice by having participants rate the pitch/clarity of pronunciation (Church & Schacter, 1994; Schacter & Church, 1992), or identify the speaker (Allen & Miller, 2004; Goldinger, 1996; Nygaard & Pisoni, 1998; Nygaard et al., 1994). In contrast, other tasks that have not required any processing of voice properties include, making a lexical decision (e.g., Luce & Lyons, 1998); name the word category (e.g., Schacter & Church, 1992); or counting the number of meanings for the word (e.g., Church & Schacter, 1994; Schacter & Church, 1992). Goldinger (1996) observed indexical effects across several types of encoding tasks that varied in the respective processing depth, thus demonstrating that the type of encoding task did not play a particular role in the emergence or disappearance of an effect. Further, these effects have appeared in both cases when the task requires attention to voice properties, as well as when it does not (Schacter & Church, 1992). Hence, differences in the type of encoding tasks does not seem to play an intrinsic role in the emergence of indexical effects.

1.4.2.2. Memory tasks

Memory tasks have typically been either implicit or explicit in nature, depending on whether they overtly refer to the initial encoding of the stimuli (explicit), or not (implicit). Explicit memory tests have typically tapped into recognition memory for previously heard words via an "old/new" discrimination task (e.g., Church & Schacter, 1994; Goldinger, 1996; Luce & Lyons, 1998; Schacter & Church, 1992), a continuous recognition test (Bradlow et al., 1999; Palmeri et al., 1993), or a cued recall test (Church & Schacter, 1994). In some studies, the recognition tests have included both an "old/new" discrimination on the word, as well as a "same/different" on the talker's voice for the old trials (Palmeri et al., 1993; Bradlow et al., 1999). The "old-same" and "old-different" discrimination tasks have provided more reliable measurements of indexical effects, compared to the only "old/new" task, suggesting that listeners can explicitly access the talker-related information included in the memory episodes of words, and use this information in performing the task.

In general, explicit memory tests have been considered relatively inconsistent in measuring indexical effects. While some studies have reported such effects (Bradlow et al., 1999; Goldinger, 1996; Luce & Lyons,1998; Palmeri et al., 1993), others have not (Church & Schacter, 1994; Pilotti, Bergman, Gallo, Sommers, & Roediger, 2000; Schacter & Church, 1992). Explaining this inconsistency is further complicated by the methodological differences between studies (see Goh, 2005 for a review).

On the other hand, implicit memory tests have displayed a more reliable pattern in revealing indexical effects. Typically, participants have shown performance advantage on stimuli repeated in the same voice, as compared to a different one, on a variety of implicit tests. These tasks include word identification tasks for filtered words, or words presented in noise (Church & Schacter, 1994; Goldinger, 1996; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Pilotti et al., 2000; Schacter & Church, 1992); stem-completion tasks (Church & Schacter, 1994; Pilotti et al., 2000; Schacter & Church, 1992); and speeded classification tasks (Goldinger, 1996). The next section presents a review of some relevant indexical studies from various research areas.

1.4.3. A review of some indexical studies

The evidence from studies investigating the effect of indexical variability on speech processing has emerged since the early fifties and is multi-faceted in terms of the research questions and the methodology used. In one of the early studies, Creelman (1957) compared the intelligibility of words spoken by either single or multiple talkers. He found an inverse relationship between identification performance and the number of talkers. Namely, the increase in the number of talkers led to a decrease in identification performance. Some decades later, Pisoni and colleagues revisited research about the effects of talker variability on spoken word processing and representation. In a well-known example, Mullennix et al (1989) found that participants' identification performance for blocks of familiar English words was both faster and more accurate in single-talker than in multiple-talker conditions (see also, McLennan, 2003; Nusbaum & Morin, 1992; Pisoni, 1993).

Another classical study, Goldinger (1996) exposed listeners to a study list of words spoken from various talkers. Afterwards, listeners were presented with a test list of words from various talkers, and had to identify each word as either *old* (previously heard), or *ne*w. Their recognition accuracy for old words was greater when the voice of the talker at the test phase matched that of the talker at the study phase, compared to when it mismatched the original talker. Using a similar task, Palmeri et al. (1993) had already demonstrated that the observed talker-specificity effect couldn't be simply explained in terms of the semantic encoding of the talker gender with each word. They observed that even when an old word was spoken by a new voice with the same gender as the old voice, listeners' recognition performance decreased, indicating hence that acoustic-matching to a previous voice facilitated word recognition. A possible explanation for this acoustically-specific encoding is related to the fact that the phonemic information is hardly separable from talker-related acoustic factors (Mullennix & Pisoni, 1990). This explanation seems particularly viable given the *integral* nature of spoken words, that makes it impossible to segregate the linguistic dimension from the talker-related acoustical one (Vitevitch, 2003).

There is also evidence for a facilitatory effect of talker familiarity on novel word identification performance. For example, in a study by Nygaard et al. (1994), participants were trained to recognize a set of voices over a 9-day period. A group of them had to identify novel words spoken by the same set of talkers at different signal-to-noise (SNR) ratios. The control group had to identify the same words, spoken by a different set of talkers. The results demonstrated that the ability to identify a talker's voice promoted the intelligibility of novel words produced by that talker. This finding was interpreted as indicating that speech perception involves some talker-dependent processing, such that the perceptual learning of talker-specific characteristics facilitates the subsequent phonetic processing of the acoustic signal.

Consistent evidence comes from a plethora of studies in long-term auditory priming, that have revealed the encoding of indexical information in long-term memory (Craik & Kirsner, 1974; Mullennix et al., 1989; Church & Schacter, 1994; Goldinger, 1996; Goldfinger et al., 1991; Palmeri et al., 1993; Nygaard et al., 1994; Luce & Lyons, 1998; Sheffert, 1998, Sommers, 1999). These studies use the so-called *long-term priming paradigm*, which consists of two blocks of stimuli: a prime or study block and a target or test block. Some items are repeated across blocks, while some in the target block are new. The observed performance (in terms of accuracy and speed) for the repeated items is typically higher than that for the new items. The increase in performance as a function of the item status (repeated or new) has been termed as the *long-term priming effect*. Changes in the surface properties of the stimulus (i.e., talker voice) from prime to target lead to a decrease in the magnitude of the priming effect. This reduction in priming due to changes in the perceptual properties of the stimulus has also been termed as *specificity*. The observed effects indicate that talker-related perceptual details are retained in long-term memory (Ju & Luce, 2006). The notion of *specificity effects* has been synonymous with that of *indexical effects.* It will be particularly useful and more appropriate to use when discussing the presence of these effects beyond the speech domain later in the chapter, since indexical properties are typically confined to the speech domain.

Other evidence for talker-specific effects on spoken word processing comes from on-line spoken word processing. In an eye-tracking study Creet al. (2008) investigated the time course of lexical activation in the presence of talker variation and found that lexical competition was reduced by consistent talker differences between words that were designed to be lexical competitors. In the first experiment, listeners were repeatedly presented with pairs of words that phonologically overlapped at the word onset (e.g., *sheep* and *sheet*). Some of the pairs were consistently spoken by the same talker (e.g., male *sheep,* male *sheet*),while other pairs the words were spoken by different talkers (e.g., male *sheep*, female *sheet*). Upon hearing the target word (e.g., *cow*), participants had to select the corresponding picture out of four pictures displayed on a computer screen. Whenever the targeted word had a cohort, the competitor's picture was also present (*couch*). Participants' eye movements over the picture display were tracked during the word duration. Listeners were significantly more likely to fixate a same-talker competitor picture (e.g., both cohort *sheet* and target *sheep* heard in the male voice) than to fixate a different-talker competitor picture (e.g., cohort *sheet* heard in the female voice and target *sheep* in the male voice). In the second experiment, participants learned to identify black and white shapes from novel labels spoken by one of two talkers. Again, a word and its competitor (cohort or rhyme) were either consistently spoken by the same talker, or by different talkers. Results revealed fewer erroneous selections of competitor pictures for different-talker competitors than for same-talker competitors, indicating a beneficial effect of differentiating competitors by talker. The authors interpreted the results as suggesting that listeners seem to retain talker-specific information when learning new words, and incorporate this information into their word representations.

In a follow-up study, Creel and Tumlin (2011) examined the circumstances under which listeners utilize talker-specific information to inform real-time spoken word processing, with a special focus on the representational levels at which this takes place. The working hypothesis was that listeners might use acoustic cues in the speech signal to access the talker's identity, which would then immediately constrain processing. Alternatively, or simultaneously, listeners might compare the signal to acoustically-detailed representations of words, without awareness of the talker's identity. In a series of eye-tracked word comprehension experiments, participants learned a set of novel words as labels for unfamiliar pictures spoken by several talkers during the study phase and were then tested on their word recognition performance as a function of talker variability. Results revealed talker-specific recognition benefits for newly-learned words both in isolation and with preceding context (embedded in sentences, such that they could give listeners talker-identity information well in advance of the word). Namely, listeners distinguished newly-learned, phonologically overlapping words (such as *boog* and *booj*) faster when the two had been learned from two different talkers (e.g., female *boog*, male *booj*) than when both were learned from the same talker (e.g., female book, female booj). There was little evidence that listeners used the sentential contexts to get talker-information, with talker-specificity effects evident only on the words themselves. The crucial finding was that the demands of the task at test had a significant impact on the way talker-specific information was used, and that when the listeners' attention was fully tuned to talker identification, they could discriminate between two talkers on a similar time scale as between two words. When the task was to learn words, listeners did not necessarily use talker-specific information, but when asked explicitly to relate talkers to novel words, they were able to do so quite easily. The authors argued that at least two processes might be involved, one necessary to store detailed acoustical representations of spoken words, and another involved in associating talkers with this information.

There is also evidence for same-talker benefit on spoken word recognition performance from a neural correlates perspective. For example, Campeanu et al. (2013) measured event-related potentials (ERPs) while participants performed recognition tasks on both the words (*old vs.* new) and the talkers (*same vs. different*), with words spoken in four voices. There were two voice properties (gender and accent) that varied between speakers, such that none, one or two of these parameters was congruent between study and test. Results indicated that talker congruency between study and test facilitated both word and talker recognition, compared to similar or no context congruency at test. These behavioural effects were matched by two ERP modulations. In the word recognition test, the same speaker condition provided the most positive left-parietal deflection of all correctly identified *old* words. In the source recognition test, a right frontal positivity was found for the same speaker condition compared to the different speaker conditions, regardless of response success. Taken together, these results suggest that the benefit of context congruency is reflected behaviourally and in ERP modulations typically associated with recognition memory.

Another area of interest for investigating indexical effects has been that of perceptual learning in speech. Evidence in this area indicates that talker-specific effects at a pre-lexical level of representation play a role in the performance benefits from talker familiarity observed at the word level (Eisner & McQueen, 2005; Kraljic & Samuel, 2007). For example, Kraljic and Samuel (2007) exposed participants to two speakers who differed in their pronunciation of a particular phoneme (/d/ or /t/; /s/ or /S/). Afterwards, participants categorised sounds belonging to a /d/-/t/or /s/-/S/ continuum, in the same two voices they had already heard during the exposure phase. The findings demonstrated that perceptual experience leads to different learning for different types of phonemic contrasts. In the case of fricatives, perceptual learning was found to be talker-specific: listeners were able to maintain multiple different representations simultaneously. On the other hand, for stop consonants, perceptual learning led to more general changes that required the listener to re-adjust their system upon encountering a new pronunciation.

Variability in the speech signal has multiple facets and is not limited to only the talker-related one. In an extensive exploratory study, Bradlow and Pisoni (1999) investigated the combined effects of various talker-, listener-, and item-related characteristics on spoken word recognition by both native and non-native listeners. The study aimed at directly investigating the ways in which multiple sources of variability operate in combination. The main hypothesis was that perceptual difficulties introduced by one factor might be attenuated or amplified by the presence of another factor. To test this hypothesis, two experiments were conducted, each of which examined spoken word recognition under conditions that manipulated talker-, listener-, and item-related factors both separately and in combination, by using a carefully constructed multi-talker, multi- listener speech database. One of the predictions was that listeners might deal better with a high degree of phonetic reduction induced by a fast speaking rate when they become familiar with the speech style of a particular talker. Another prediction was that the recognition oh "hard" words (i.e., words with many phonetically similar neighbours; in contrast, "easy" words have few phonetically similar neighbours) would be impaired for non-native listeners whenever there was a mismatch between the native and target language phoneme inventories. The first experiment investigated the effects of speaking rate (fast versus medium versus slow) and lexical discrimination (easy versus hard) on isolated word intelligibility. Listeners listened to lists of "easy" and "hard" words spoken by several talkers and at different rates and transcribed (typed) the words. Results indicated that as expected, lexical discriminability had an effect on the overall word intelligibility: easy words had higher overall intelligibility than hard words. This effect was categorized as a listener-related factor that results from listener's knowledge of the sound-based structure of the lexicon of the language. An effect of the speaking rate on overall word intelligibility was also observed, such that slow and medium rate words yielded higher overall intelligibility scores than fast rate words. This effect was categorized as a signal-related factor that might have resulted from acoustic-phonetic adjustments on the part of the talker when they were required to consciously adjust the speaking rate. Importantly, the authors observed a relationship between the various factors, such that the difficulties imposed by one factor (fast speaking rate or a difficult lexical item), could be overcome by the advantage gained through the listener's experience with the speech of a particular talker. In the second experiment they investigated some of the characteristics of non-native spoken word recognition as they relate to known characteristics of native spoken word recognition. Interestingly, the results revealed that spoken word recognition by non-native listeners displayed the same overall patterns as for native listeners. Namely, both groups of listeners recognized words more accurately when all the test words were spoken by the same talker relative to a condition where the talker changed from item to item. They also found that both groups of listeners were more accurate in recognizing "easy" words that were distinctive, or easily discriminated in their lexical neighbourhood, than "hard" words that had many phonetically similar neighbours. However, this effect was more robust for the non-native listeners, suggesting that these listeners have particular difficulty in recognizing words that require perception of fine phonetic detail for lexical discrimination. Taken together, their results demonstrated that spoken word recognition depends on a combination of at least three types of factors: 1) signal-related properties, such as speaking rate; 2) lexical factors, such as knowledge of the sound-based structure of the mental lexicon, and 3) instance-specific factors, such as the listener's prior experience with the talker's voice.

In another series of experiments, Bradlow et al. (1999) investigated the encoding of the surface form of spoken words using a continuous recognition memory task, in an attempt to compare three sources of stimulus variability -talker, speaking rate, and overall amplitude- and determine the extent to which each source of variability persisted in episodic memory. In the first experiment, listeners performed an "old/new" discrimination task on each of the spoken words in a list. They were more accurate at recognizing a word as *old* if it was repeated by the same talker and at the same speaking rate; however, no recognition advantage for words repeated at the same overall amplitude was observed. In the second experiment, listeners first decided whether each word was *old* or *new*, and then explicitly decided whether it was repeated by the same talker, at the same rate, or at the same amplitude. Similar to their performance in the first experiment, they showed an advantage in recognition memory for words repeated by the same talker and at same speaking rate, but again, there was no advantage for the amplitude condition. Nevertheless, listeners were able to *explicitly* identify whether an *old* word was repeated by the same talker, at the same rate, or at the same amplitude. This variability was discriminated to a different extent along each of the three dimensions, such that: talker variability was detected better than rate variability, which in turn was detected better than amplitude variability. The results suggested that although information about all three examined properties of spoken words is encoded and retained in memory, variation in each of them affects episodic memory for spoken words to different extents.

1.5. Alternative views of the lexicon

1.5.1. The episodic view

The broad range of evidence in support of indexical effects in spoken word processing in a way "revolutionised" the way speech perception and representation in memory was approached by theoretical accounts. Namely, it became evident that the traditional abstractionist view of speech perception and the respective model of the lexicon could not accommodate the evidence that listeners encode fine-grain talker-related information in their memory representations. This realisation led to the emergence of the exemplar-based, episodic view of the lexicon (Goldinger, 1998; Goldinger & Azuma, 2003). According to this view, detailed unique episodic traces of spoken words are formed during speech perception and they have an effect on subsequent perceptual and memory experiences (Goldinger, 1996; Nygaard & Pisoni, 1998; Goldinger & Azuma, 2003). In what is considered a classical paper, Goldinger (1998) put forward a view of the lexicon as a collection of memory episodes, rather than abstract categories. He provided behavioural evidence from a shadowing task, in which participants first heard several words spoken only in particular voices and then repeated the word lists including those words. It was observed that the more the repetitions of a word in a particular voice during the exposure phase, the faster their shadowing times and the greater the perceived similarity (judged by new listeners) between the imitation and the word imitated. In the other experiments, listeners were first exposed to novel words in a pretraining session that used a talker whom they had not previously heard in the shadowing session. Afterwards, they heard talker-specific presentations and performed a shadowing task as in the previous experiment with real words. The findings pointed in the same direction: fewer exposures at pre-training and limited variability of non-words in the training session led to more pronounced talker-related effects (in terms of both reaction time and imitation similarity) in the shadowing session.

In order to illustrate how episodic perception may work, Goldinger (1998) tested an exemplar computational model (Hintzman, 1988: MINERVA 2) against the behavioural data from the word shadowing tasks. In MINERVA 2, there is a large collection of partially redundant traces in memory for each known word. These traces encode perceptual, conceptual and contextual features of the original encoding event. When a stimulus word is heard, all traces are activated, each in proportion to their mutual similarity. The weighted average of the activated traces forms an echo that long-term memory conveys to ‘‘consciousness”. Echoes contain information not present in the probe (word meanings) by using the information from past traces, hence associating new stimuli to previous knowledge. Goldinger (1998) found that MINERVA 2 successfully predicted the shadowing response time patterns and also a tendency for participants to spontaneously imitate the acoustic patterns of words and non-words. A particularly interesting finding was that the model correctly predicted the strength of shadowing as a function of word frequency. In the face of such evidence, Goldinger (1998) argued that detailed episodes constitute the building blocks of the mental lexicon, as opposed to abstract phonological units.

Further evidence in support of episodic theories comes from studies showing that the lexicon is composed of representations that are detailed in their sub phonemic features, rather than abstract phonemic ones. Research on listeners’ sensitivity to talker differences regarding a phonetically relevant acoustic property, the voice onset time (VOT) has shown that listeners are sensitive to and can learn talker-specific phonetic information (e.g., Allen & Miller, 2004). This information is retained in a way that allows for generalization to novel words and can use the memory for talker-specific details to facilitate later phonetic processing of familiar talkers’ speech. Additionally, eye-tracking (e.g., McMurray et al., 2002) and priming studies (e.g., Andruski et al., 1994) have reported gradient voice onset time (VOT) effects on lexical activation, meaning that the activation of lexical items is dependent upon similarity to a certain VOT prototype and can’t be seen as an all-or-none process. These types of results fit within the framework of a distributional learning mechanism, in which episodic acoustic information is preserved and modulated by frequency (i.e., more frequent VOT exemplars are more likely to have the status of a prototype) (see Maye et al., 2002).

The episodic view has been criticized primarily on the ground that it dispenses with abstract categories. As such, it lacks the power to explain compelling evidence (reviewed in the section below) that demonstrates the need for such categories. The interest in the literature has shifted towards hybrid models that can accommodate both abstract and episodic representations of spoken words. In fact, one of the original proponents of the episodic view, Goldinger, has recently argued for a hybrid model of speech perception that is based on the *complementary systems approach* of memory (McClelland et al., 1995; Goldinger, 2007). This model, as well as other attempts at designing hybrid representational models will be reviewed in the next section.

1.5.2. The hybrid view – Co-existence of episodes and abstractions

As already noted in the previous section, there is considerable evidence in the literature that supports the existence of abstract lexical (and pre-lexical: Cutler, 2008) representations, which cannot be explained by an episodic-only view of the lexicon. This has led to the need for designing hybrid representational models that can accommodate the existing results from both fronts: episodic and abstract. Below I review some relevant examples from this evidence.

Perceptual learning studies have proved particularly useful in testing for the presence of such abstract representations. The results of some of these studies challenge any extreme episodic model in which no abstraction is needed prior to lexical access (Cutler, 2008).

For example, a study by Norris, McQueen, and Cutler (2003) revealed a lexically driven modulation of the category boundary for a consonant contrast, which was introduced in an exposure phase and measured in a subsequent phonetic categorisation task. In the exposure phase, listeners heard naturally produced words, some of which were edited. For one group of listeners, all instances of the fricative sound [s] were replaced by a perceptually ambiguous sound lying midway between [s] and [f]. For another group of listeners, all cases of [f] were replaced by the same ambiguous fricative sound. Results showed that the group that had heard the ambiguous sound in [s]-biased lexical contexts categorised more sounds on an [f]–[s] continuum as [s], whereas the other group categorised most sounds as [f]. Hence, listeners readjusted their fricative categories as a function of the training and the lexical context in which it took place. Overall, this study shows that a perceptual adjustment is made when an idiosyncratic production of a speech sound is placed in an appropriate lexical context. Particularly interesting for the present purposes is the most appropriate explanation behind this phenomenon. According to the authors, it *facilitated* subsequent word recognition. They proposed that readjusting the perception of sounds at an abstract pre-lexical level promoted future recognition of other words in which those sounds were present.

However, McQueen et al. (2006) noted that this evidence did not make a strong argument for the existence of pre-lexical abstraction. Alternatively, it may have been the case that the perceptual learning observed was limited to tasks involving explicit judgments and that these types of judgments benefited from post-lexical phonological representations that are not directly engaged in word recognition. Importantly, both abstractionist and episodic models can accommodate a post-lexical processing stage, hence, it is not accurate to claim that episodic models lack representations of phonemic categories. It is the role that these categories play in word recognition that is arguable. Episodic models may involve phonemic categories, but they serve only as tags for groups of episodic traces and do not serve any abstraction function in lexical access. Therefore, the learning effect in phonemic categorisation found by Norris et al. (2003) is not sufficient for distinguishing between the two types of models. According to McQueen et al. (2006), stronger evidence for differentiating between the two would be testing for lexical generalisation of the re-adjustment effect to words not previously heard in the training phase. The presence of the re-adjustment effect on newly heard words would favour the idea that the readjustment occurs at a pre-lexical level, and also that learning about abstract sub-lexical representations is involved.

Following this line of thought, McQueen et al. (2006) tested whether the perceptual adjustments arising from previous encounters would affect listeners' subsequent interpretation of minimal word pairs that were Dutch words which only differed in their final [f] or [s]. The first part of the experiment was identical to the training phase in Norris et al. (2003). The innovative aspect of the paradigm was in the second part, where cross-modal identity priming with ambiguous primes derived from the minimal pairs was used (e.g., [do?], from *doof* "deaf" versus *doos* "box"). Listeners heard such primes and then performed visual lexical decisions about letter strings presented immediately after the primes. In cross-modal identity priming paradigm, visual lexical decision is facilitated when the same word has just been heard, compared to a new word. However, the mismatch in one phoneme between the prime and the target (e.g., *fake–fate*) reduces this facilitation (Marslen-Wilson et al., 1995; as cited in McQueen et al., 2006). As predicted, the results showed that the perceptual adjustment of fricative categories extended to the recognition of fricative-final words that the listeners were not exposed to during the training phase, hence providing compelling evidence that this adjustment occurred at a pre-lexical stage of processing. McQueen et al. (2006) argued that strictly episodic models cannot account for these results, because they do not take advantage of sub-lexical regularities during word recognition, and hence fail to support generalization of these regularities across words.

Further evidence in support of the view that perceptual readjustment involves phonologically abstract representations comes from studies showing that perceptual learning can generalize to similar sounds (e.g., Kraljic & Samuel, 2006). On the other hand, the evidence that perceptual learning can be talker-specific (Eisner & McQueen, 2005; Kraljic & Samuel, 2007), and that in turn, this is related to the extent to which this talker-specific information is encoded in the manipulated phonemes (Kraljic & Samuel, 2005, 2006), suggests that these representations are also flexible. This dual-faceted evidence supports neither the strictly episodic, nor the strictly abstractionist view of the mental lexicon, but instead calls for the development of hybrid models that would successfully bond both extremes.

However, as Cutler (2008) and McQueen and Cutler (2010) note, reconciling abstractionist and episodic components in the same model might be the greatest challenge for psycholinguistic modelling. Despite promising attempts, there is as yet limited direct evidence. In the following section, I review three recent attempts that seem promising towards building more definite and robust hybrid models.

1.5.2.1. The complementary systems approach

Originally proposed by McClelland et al. (1995) as a *complementary learning systems* (CLS) model of memory, this approach was adapted for spoken word perception by Goldinger (2007). It is based on a *complementary-systems* perspective, wherein reciprocal computational neural networks represent hippocampal and cortical memory systems. The hybrid memory system posited by the model, eliminates the abstract-episodic conundrum. Namely, detailed episodic traces (hippocampal system) and holographic, abstract traces (cortical system) combine to simulate behaviour in real time, thus allowing perceptual or memory data to appear relatively "episodic". Importantly, the two memory systems are inter-connected, such that the traces in each result from the interactivity between the systems. Accordingly, traces in the hippocampus are a result of input from different parts of the cortex. In the case of speech perception, such traces would involve input from the cortical system that segments spoken words and assigns meanings, but could also be extended to include visual and/or emotional information. Hence, although the hippocampal system's primary function is to learn unique traces, its input has already undergone some degree of abstraction; which would presumably occur in the early stages of word perception. On the other hand, cortical representations are formed in a complementary fashion, by the collection of episodes that are stored in the hippocampus and are gradually consolidated back into the cortex. This complementary interaction between the two systems leads to a medium for gradual learning in the cortical system (Goldinger 2007). An instance illustrating this type of process would be the case when extensive exposure to unique episodic traces (e.g., regional accent) slowly affects more abstract knowledge (i.e., adaptation to the accent).

Goldinger (2007) successfully simulated the model with a voice-sensitive priming task on bisyllabic words, and observed that same-voice trials led to fastest settling times (a measure of the network's performance), whereas larger voice changes induced a steady decline in performance. Therefore, the simulation of voice-sensitive priming demonstrated that activity in the hippocampal network reflected changes in the indexical properties of words.

Notably, Davis and Gaskell (2009) proposed a novel theory of the cognitive and neural processes by which adults learn new spoken words, that also relies on the complementary learning systems (CLS) models of memory. This approach seems to have good potential of influencing future psycholinguistic hybrid models of spoken word recognition.

1.5.2.2. The adaptive resonance theory (ART) approach

This attempt was put forward by McLennan and colleagues (2003; 2005) and was motivated by their findings from a series of long-term repetition priming and lexical decision experiments that examined whether surface representations of spoken words are mapped onto underlying, abstract representations. In this case surface variation was allophonic. More specifically, the authors tested the hypothesis that flaps (neutralized allophones of intervocalic /t/s and /d/s), are mapped onto their underlying abstract phonemic counterparts. In casual American English, a /t/ or a /d/ produced between two vowels, as in the word *rater*, is often realized as a flap, a segment that varies between a / t/ and a /d/ (McLennan et al., 2003). In two of the experiments, participants performed a lexical decision task, where they had to decide as quickly and accurately as possible whether the item was a word or non-word. The degree of task difficulty was varied by manipulating the extent to which non-word resembled real words. Namely, in the easy condition, the non-words were quite dissimilar to real words (e.g., *thush-* thudge), whereas in the difficult task condition, the non-words were quite similar to real words (e.g., *bacov*). As expected, overall, participants were quicker to respond in the experiment that involved the easy task compared to the experiment that involved the difficult task. The interesting finding was that allophonic specificity effects were observed only in the easy task condition, not in the difficult one. This pattern of results was interpreted as evidence in support of the hypothesis that the allophonic information that is more frequent dominates initial processing (as in the easy task condition), and that effects of the less frequent underlying information emerge only after some delay (like in the difficult discrimination experiment). Overall, the overall results supported a co-existence of both surface and underlying form-based representations, such that specific allophonic representations dominated processing in the case of rapid responses, and upon slowing of the response time, more abstract underlying representations emerged (i.e., /t/ and /d/ categories). In light of these findings, McLennan et al. (2003) proposed an explanatory account adapted from Grossberg's ARTPHONE neural model (Grossberg et al., 1997).

In the ART framework, the acoustic–phonetic input, consisting of relatively rich and specific surface representations, activates chunks of features corresponding to more abstract phonological representations, as well as chunks corresponding to less abstract, allophonic representations. A *chunk* represents a learned set of associated features that may vary in size, such that a certain chunk may refer to an individual feature, an allophone, or a word. Crucially, chunks *resonate* with the input, and this resonance between them in turn constitutes the percept and mediates priming and specificity effects. In addition, resonance with the input is determined more quickly by more frequent features and combinations of features (i.e., chunks) in a pattern, than by less frequent ones. When the phonological processing is deep (as in the case of difficult lexical decisions), the restoration of surface representations by underlying abstract representations (chunks) is promoted, which is a process that requires time. In contrast, when the task taps into the recognition process before the underlying abstract form has been activated, robust specificity effects in long-term repetition priming and lexical decision tasks emerge, presumably because the underlying abstract representations may not have had enough time to resonate with a restored surface form.

On a side note, the pattern of results and the proposed theoretical explanation rely on a time-course regarding the emergence of allophonic specificity effects. Namely, allophonic specificity emerges early in processing (in tasks eliciting rapid responses), whilst abstract information takes time to surface in the system and affect perception. Crucially, this time-course is the opposite of the one proposed in McLennan and Luce (2005) for indexical specificity. In the case of indexical effects, immediate processing is dominated by abstract representations, whereas specific information takes time to percolate through the system and affect perception (Luce et al., 2003; McLennan & Luce, 2005). In fact, according to Luce et al. (2003), evidence for episodic theories has emerged mainly from research on indexical variability, while research on allophonic variability has indicated the operation of more abstract representations. The discrepancy in the time-courses of these two types of effects could be used to inform a mixed representational model wherein distinctive effects of abstract and episodic representations can be predicted on the basis of processing time aspects (Luce et al., 2003).

I will elaborate in more detail on the time-course hypothesis regarding indexical effects in section 1.6 below.

1.5.2.3. The socially-weighted dual-route approach

The third and also the most recent attempt is by Sumner et al. (2014). This is a dual-route approach of speech perception that advocates the integration of linguistic and talker-related information from a socio-linguistic perspective. They argue that the perception of spoken words is socially weighted and propose a dual-route approach to speech perception in which listeners map acoustic patterns in speech to linguistic and social representations simultaneously. Accordingly, socially salient tokens are encoded with greater strength (by increased attention to the stimulus), than both typical and atypical non-salient tokens . In this view, a representation derived from one instance of a strongly encoded socially salient token may be as robust as one derived from a high number of less salient, default tokens. An interesting aspect of this approach is that contrary to typical views that try to explain the many-to-one mapping of variable signals to a single linguistic representation, it endorses a one-to-many perspective, in which a single speech string may be mapped to multiple linguistic and social representations concurrently. Speech is considered to be a multi-faceted source of information and its comprehension, broadly, derives from the integration of both linguistic and indexical/social information. A visual illustration of the approach taken from Sumner et al., (2014), is displayed in Figure 1.1.

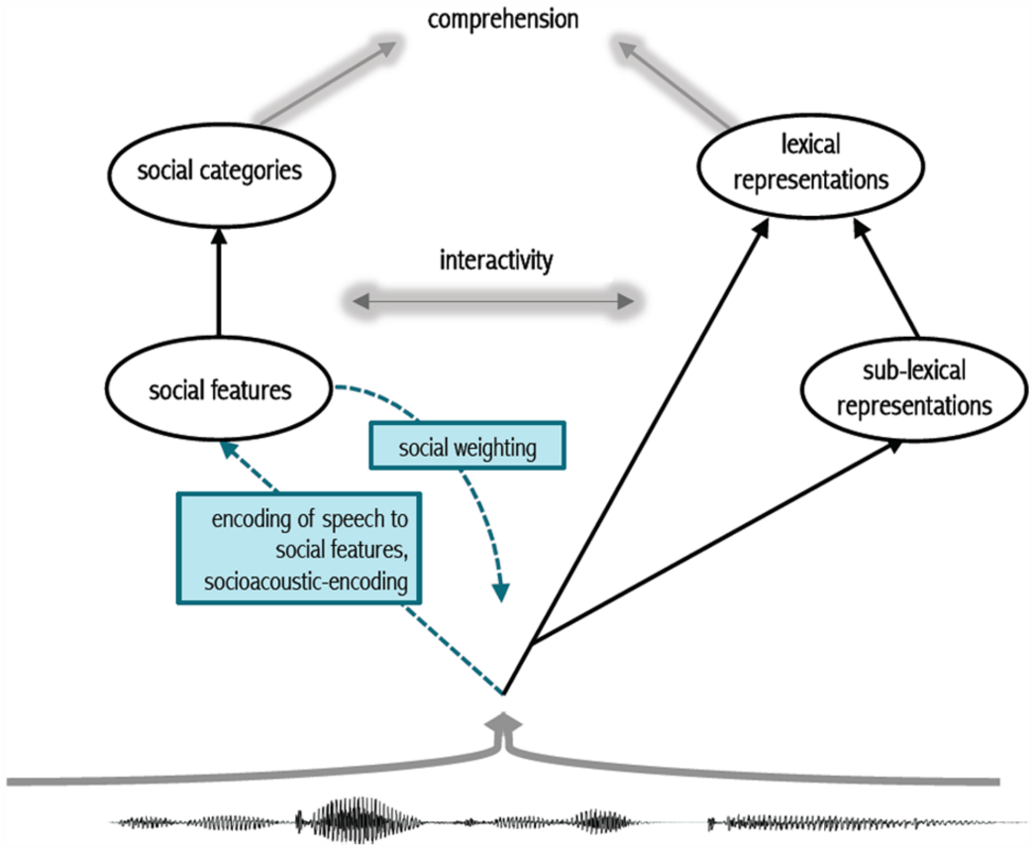

Figure 1.1. (Image and explanation taken from Sumner et al., 2014). In tandem with the encoding of speech to sounds and words (right), acoustic patterns in speech are encoded to social representations (left). Socially weighted encoding results from the heightened activation of social representations that modulates attention to the speech signal. This in turn results in the deep encoding of socially salient acoustic patterns along with linguistic representations, but also independent of them.

It is worth noting that a parallel processing scheme like the one guiding this account, is already a consolidated approach in the speech domain. It constitutes the building block of some of the most prominent dual-pathway neuroanatomical models (e.g., Gow, 2012; Hickok & Poeppel, 2000, 2004, 2007; Rauschecker & Scott, 2009; Scott, 2005; Scott & Wise, 2004), and connectionist models of speech processing (e.g., Gaskell & Marsel-Wilson, 1997; Hinton et al., 1986; Plaut & McClelland, 2010). In the neuroanatomical models, the auditory input is initially processed in the primary auditory cortex, after which higher-level auditory and acoustic–phonetic processing occurs in adjacent superior temporal structures. Similar to the processing type outlined in Gaskell and Marslen-Wilson (1997)'s model, successive mappings are performed in simultaneous parallel processing streams, that comprise a dorsal pathway that provides a mapping between sound and articulation, and a ventral pathway that maps from sound to meaning. The interested reader is referred to Gow (2012) for an extensive review of the evidence from a multitude of sources, as well as the new dual-pathway neural model proposed in the paper.

1.6. The time-course of indexical effects

As mentioned in section 1.3, the emergence of indexical effects has varied in terms of tasks and stimulus quality. Luce, McLennan, and Charles-Luce (2003) posited that the observed inconsistencies in the indexical literature could be best explained by differences in processing-time requirements. According to the *time-course hypothesis*, indexical effects seem to emerge relatively late in processing and the degree of reliance on episodic information depends on the speed with which a response to the task at hand is produced. Namely, slow responses allow retrieval of episodic traces to a greater extent than faster responses. McLennan and Luce (2005) tested this hypothesis in a study consisting of three long-term repetition priming experiments, where the reaction times to targets that were primed by stimuli that matched or mismatched on the indexical variable of interest (talker identity or speaking rate) were examined. In all experiments, the speed with which participants processed the stimuli and hence was manipulated by means of task difficulty. More specifically, in the first two experiments listeners completed either an "easy", or a "difficult" lexical decision task. The easy task involved non-words that were dissimilar to real words and were thus easily and rapidly discriminated, whereas the difficult task comprised non-words that were highly similar to real words and were hence discriminated more slowly and with more effort. As expected, indexical effects were found only in the difficult task, indicating that the emergence of these effects requires processing time. The third experiment intended to tap into the processing system at different times by manipulating the response type: speeded shadowing (a single-word speeded-response shadowing task), or delayed shadowing (cued shadowing).[1] The latter cues the participants on when to respond, providing them with the opportunity to spend additional time processing and rehearsing the stimuli. Again, as predicted, indexical effects emerged only in the delayed-response shadowing task, that necessitated slower processing time. Overall, the results were interpreted as evidence that early perceptual processing is dominated by more abstract, or underlying information, whereas later stages of processing are dominated by more specific, detailed surface information.

Further support for the time-course hypothesis comes from Mattys and Liss (2008), who manipulated processing time in a novel and natural way by using normal vs. impaired (dysarthic) speech. Dysarthric speech provides a convenient and natural medium in which to test this hypothesis, since the degraded quality of speech in this case puts the listener in challenging perceptual conditions, that in turn require more processing time compared to normal speech. Listeners were exposed to either one of three speech conditions: normal (control), mildly dysarthric, and severely dysarthric. The word stimuli were produced by a male and a female talker (different pair in each speech condition). The exposure phase involved passively listening to a series of words, followed by an "old/new" recognition task in the test phase. While there was a voice specificity effect on recognition accuracy in all the speech conditions, it increased with the level of speech impairment. As for response latencies, there was a voice effect only in the impaired speech conditions. Further, analyses performed separately on slow and fast respondents revealed a marked contrast for the two subgroups. Namely, slow respondents showed both a voice effect, whereas fast respondents did not.

In a recent study, the time-course hypothesis was tested in the context of native and foreign-accented speech (McLennan & González, 2012). The lexical decision task used at test, revealed a talker-specificity effect only in the case of the foreign-accented speech, and not for native speech. This result was in accordance with the time-course hypothesis, that would predict slower processing times for the foreign-accented than for the native speech (see also Theodore et al., 2015).

With this section, I conclude the review of the evidence concerning indexical variability and its effect on speech processing and representation in memory. Next, I focus on another type of variability, one that is extrinsic to, but co-occurs with speech. Very recent evidence has suggested that spoken word processing may be sensitive to this type of variability, and that listeners may encode

[1] The term "shadowing" refers to the repetition of the stimuli aloud.

specific details about it in their memory representations, alongside linguistic and indexical information.

1.7. Specificity effects beyond the speech domain

As discussed above, the evidence in favour of indexical effects in speech perception has raised the need for psycholinguistic models of spoken word recognition and accounts of the lexicon to incorporate talker-specific indexical information alongside more abstract lexical representations. However, the variability encountered during speech processing is not confined to the speech domain. We typically process speech in a context, which is also highly variable, an observation that leads to similar questions to the ones concerning indexical variability, Namely, how does the perceptual system cope with the additional external variability from the environment? Do listeners segregate and discard speech-extrinsic variability early in processing, or do they encode it somehow in memory, like they do with speech-intrinsic variability? Is it possible to observe specificity effects in speech processing as a result of speech-extrinsic variability?

Very recent evidence has revealed the presence of speech-extrinsic specificity effects. Specifically, the change of a background sound/noise co-occurring with spoken words seem to impair identification/recognition memory for words previously heard/learned during exposure (Creel et al., 2012; Cooper et al., 2015; Pufahl & Samuel, 2014). Interestingly, this impairment is comparable to the one elicited by the voice change (Pufahl & Samuel, 2014). Below, I review evidence from the three studies that have found speech-extrinsic specificity effects in spoken word processing, using different methodologies.

In a series of experiments, Pufahl and Samuel (2014) investigated the impact of variability in environmental sounds co-occurring with spoken words on word identification memory. Drawing on the analogy to the voice specificity effect, the authors noted that while the variability in the talker voice may have functional relevance for the perceptual system, it is possible to endorse the view that a spoken word always co-exists with the voice. Therefore, a potential explanation for the existence of indexical effects may rely on the co-occurrence element of the word-voice relationship. From this perspective, the indexical properties may not be retained in the lexicon because of their unique indexical status per se, but rather because these properties are co-present with the linguistic information. I will briefly review only the first experiment of this study, due to its particular relevance for the present research.

Specifically, participants heard spoken words paired with environmental sounds. Half of the words represented animate entities (e.g., butterfly) and half of them inanimate ones (e.g., table). Similarly, half of the sounds were from animate sources (e.g., a dog barking), and half were from inanimate sources (e.g., a door bell). All the words were spoken by a male and a female speaker, so there were two versions of each word. Similarly, there were two versions (exemplars) of each sound (e.g., a large dog and a small dog barking; or two different door bells). The experiment consisted of an exposure phase followed by a short delay and then by a test phase. In the exposure phase, participants performed an “animate/inanimate” decision on the word, ignoring the sound. They then completed a word identification task in the test phase, that involved transcribing the word from the highly filtered versions of the word-sound pairs. The degree of match between the stimuli in exposure and test was varied as a function of the change in voice, sound, or both. Accordingly, there were four different combinations of voice-sound manipulations between exposure and test: 1) neither the voice, nor the sound changed; 2) only the voice changed; 3) only the sound changed; and 4) both the voice and the sound changed. Participants were instructed to write down each word they heard, guessing if necessary, and ignore the sound. Results revealed two main things: 1) word identification performance was significantly impaired when the voice of the talker changed from exposure to test phase (the classical voice specificity effect) and, more interestingly, 2) word identification performance was comparably impaired when the paired environmental sound changed from exposure to test phase. The latter finding was interpreted as evidence that a seemingly irrelevant background sound co-occurring with a spoken word was integrated with the word in memory and facilitated subsequent word identification performance. Given that this specificity effect emerged in similar conditions to indexical effects, the authors argued for a further expansion of the mental lexicon to include speech-extrinsic auditory information, alongside linguistic and indexical information. Further, mere co-occurrence between words and sounds was deemed sufficient for their integrated representation to be encoded in the lexicon.

In another recent study, Cooper et al. (2015) examined processing dependencies between background noise and indexical speech features using a speeded classification paradigm (Garner, 1974). In another experiment, they also investigated whether background noise is encoded and represented in memory for spoken words by using a continuous recognition memory paradigm. In both cases, whether or not the noise spectrally overlapped with the speech signal was manipulated. In the first experiment, they investigated perceptual integration versus segregation at an early stage of processing as measured by the Garner (1974) speeded classification paradigm. Results revealed an interdependence of the perceptual processes used to encode information about background noise with indexical information in the speech signal (i.e., gender and talker identity). This suggests that speech and background noise are perceptually integrated at the level of processing tapped into by the speeded classification task. Interestingly, the observed interdependence at the level of perceptual classification appeared to be largely independent of whether the noise and speech are spectrally overlapping or not. However, there was an asymmetry regarding the observed perceptual interference, such that irrelevant indexical feature variation in the speech signal slowed noise classification to a greater extent than irrelevant noise variation slowed speech classification. This asymmetry is not surprising, considering the fact that compared to background noise, speech features are more functionally relevant to listeners, and as such, are more difficult to ignore. In the second experiment, they used the same stimuli to investigate whether listeners' ability to discriminate new (first occurrence) from old (second occurrence) words in a continuous list of spoken words was affected by the variation in the background noise from the first to the second occurrence. An explicit version of this task was implemented, such that participants were explicitly drawn to the background noise by requiring them to decide whether an *old* word (i.e., repeated in the list) had the *same*, or *different* background noise, relative to the first occurrence. Therefore, contrary to the speeded classification task used in the first experiment, where listeners could respond to each trial without referring to a previous trial, in this task participants had to explicitly assess the match between two instances of a spoken word. Results showed that recognition memory for spoken words was affected by the variability in the background noise. However, this effect emerged only when the noise and the speech signal were spectrally overlapping. Taken together, these findings favour an integrated processing of speech and background noise, modulated by the level of processing and the spectral overlap between speech and noise.

Finally, Creel, Aslin, and Tanenhaus (2012) found evidence suggesting that listeners form acoustically-specific representations of newly learned novel words. More specifically, they trained

English listeners on non-words that served as labels for unfamiliar shapes displayed on a computer screen. During the learning phase, the words were heard either in the clear, or embedded in white noise. Similarly, listeners were tested on words in noise or in the clear. The match between the learning and test contexts involved whether the initial exposure to the novel items had been in the clear or in white noise, and whether testing occurred in the clear or in white noise. As a result, there were four between-participants conditions: clear exposure - clear test, clear exposure - noise test, noise exposure - clear test, and noise exposure - noise test. For each word (e.g. *dabo*), there was another word with the same vowels (e.g. *gapo*), as well as one with the same consonants (e.g. *dubei*). Learning was tested via a multiple forced-choice picture-selection task (4 picture candidates). Results revealed that listeners benefited from the match between learning and test contexts, such that those who were exposed to the same context at learning and test (Both clear or both noisy), displayed the highest performance in terms of accuracy and speed. This indicated that listeners' newly formed lexical representations included auditory contextual details pertaining to the speech-extraneous context of the initial exposure.

In summary, there appears to be an increasing interest in the literature recently, that targets the effects of speech-extrinsic auditory variability on the processing and memory representations of spoken words. The initial observation is that this variability seems to affect speech processing, however its novelty brings along the need for further investigation of the conditions in which this takes place. In the next and final section of this chapter, I outline the rationale that motivated the present research and provide a brief overview of the next chapters.

1.8. Sounds in the lexicon? - Exploring speech-extrinsic specificity

To briefly summarise what I have been reviewing so far: previous research spanning several decades has demonstrated that listeners encode talker-specific indexical information in their memory representations, as manifested by what are typically referred to as indexical effects in spoken word processing. This realisation has motivated alternative views of speech perception and the lexicon, wherein the memory representations of spoken words may be richer in content and more flexible than previously thought, comprising both linguistic and indexical information. However, it is yet unclear as to whether the integration of the abstract linguistic information with the more episodic indexical information takes place and is indeed stored within the lexicon. A better understanding of this issue is further complicated by inconsistencies in finding specificity effects across indexical studies, and the different approaches in the literature regarding the nature of the lexicon, with some of them going as far as questioning the need for such a specialised memory structure (e.g., Elman, 2004, 2009). Although the recent trend in the field favours hybrid models of spoken word recognition and the lexicon, developing robust theoretical accounts and models remains a challenge for psycholinguistic theories of spoken word recognition.

Very recent evidence has suggested that spoken word processing seems to be also sensitive to speech-extrinsic auditory variability, manifested in a similar way to effects of indexical variability. To date, effects of such variability have been observed in tasks involving: implicit memory for the identification of highly filtered words, perceptual integration of indexical and speech-extrinsic auditory details (noise), and learning new words in the presence of background noise. The emergence of these effects has led to new claims regarding the nature of the lexicon, positing that besides storing linguistic and indexical information, it could be further expanded to also include speech-extrinsic auditory details (Pufahl & Samuel, 2014).

The discovery of novel effects brings along excitement, but perhaps even more importantly, the need for further investigation. The research covered in this thesis was motivated by the recent speech-extrinsic specificity effects, in particular by the finding of Pufahl and Samuel (2014), to which I will refer to as the *sound specificity effect* henceforth. Before proposing a new view of the lexicon based on this effect, there are several critical questions about it that demand attention.

First, does the emergence of a sound specificity effect really entail the presence of the sounds in memory/lexicon alongside the words? Besides co-occurring with the words, the sounds also act as maskers, leading to degraded versions of them, alternatively known as *acoustic glimpses*. Two different sounds mask the same word differently, creating two distinct degraded versions of it. Accordingly, the specificity effect could have been the result of the mismatch between the two distinct degraded versions of the same word in exposure and test, rather than due to the mismatch between the word-sound associations per se. This scenario would imply that it is not the word-sound associations per se that are encoded in memory, but rather the degraded versions (glimpses) of the words as a result of masking from the sounds. Decoupling these two competing alternatives inspired the first research question explored in this thesis (**Chapter 3)**.

Second, although observed under similar circumstances and methodology used to measure indexical effects, is the sound specificity really similar to the voice specificity effect? A spoken word is intrinsically different in nature from a word-sound pair. As described above, speech has an integral nature, wherein the linguistic and indexical component do not only co-exist, but are also integral to one another, belonging to the same perceptual object. On the other hand, the mere co-existence between words and sounds does not display this integrality element; the sounds can be perceptually segregated from the words with relative ease and the two do not belong to the same object. Does this discrepancy between the two stimuli types (words and word-sound pairs) constrain the emergence of a sound specificity effect? In other words, is mere co-occurrence between words and sounds really sufficient to elicit a specificity effect? These questions were explored in **Chapter 4**.

Third, keeping with the indexical analogy, a spoken word is a unique utterance. No word is spoken in the same way across different speakers, and sometimes even by the same speaker. How would this uniqueness property translate to the co-existence of words and sounds in the investigation of sound specificity effects? For instance, in Pufahl and Samuel (2014), the pairing between a word and a sound was unique, although they do not explicitly explain the reasons behind choosing to implement this association type in their stimuli.[2] Does this element of the co-occurrence between sounds and words play a role in the emergence of a sound specificity effect? Alternatively put, is an effect observed in that context replicable? This question is examined in **Chapter 5**.

These are the main arguments that motivated the present research, spanning several recognition memory experiments and one implicit word identification study. Crucially, this thesis endorses both a comparative and explorative perspective in the investigation of sound specificity effects. The comparative aspect comes from the special attention dedicated to the analogy with indexical ef-

[2] A word was paired with a unique sound exemplar, that was not used in another pairing.

fects, as well as frequent references made to the sound specificity effect found in Pufahl and Samuel (2014), in particular to its co-occurrence element. The explorative aspect lies in its attempts to identify plausible conditions that restrain or promote the emergence of sound specificity effects. In the section below, I briefly outline the organisation of the work in the remaining chapters.

1.8.1 The Present Research – Overview of the next chapters

In line with the rationale outlined above, the present research is organised along the following chapters:

- ♦ Chapter 2 is dedicated to the replication of the classical voice specificity effect, which serves as a comparative basis for the subsequent examination of the sound specificity effect.

- ♦ Chapter 3 consists of three experiments that investigated the role of the acoustic glimpses of the same word(s) in the emergence of a sound specificity effect. In analogy to the two voices in the first experiment, two car horn sounds were used as pair companions of the spoken words.

- ♦ Chapter 4 will present two experiments that explored and identified the *integrality* factor between the word-sound associations as a necessary condition for the emergence of a sound specificity effect. A strict analogy to the voice specificity effect was endorsed, with a particular focus on the intrinsic relationship between words and voices.

- ♦ Chapter 5 involves two experiments (one being an extensive pilot for word intelligibility) that examined the role of pair-wise *association uniqueness* in the emergence of a sound specificity effect. In this case, association uniqueness refers to the unique pairing between a word and a sound. Since this part of the research also aimed at replicating the sound specificity effect originally reported by Pufahl and Samuel (2014), the same environmental sounds (a kind courtesy of April Pufahl) and a similar experimental design were used.

- ♦ Chapter 6 is the final chapter of the thesis and will provide a general discussion and concluding remarks of all the results described in the previous chapters.

Chapter 2

The Voice Specificity Effect

Abstract

The experiment presented in this chapter replicates the *voice specificity* effect. As such, it provides a solid comparative basis in the series of studies that investigated the *sound specificity* effect. While the latter is the primary focus of this thesis, a comparative investigation has the potential to provide better insights into the big picture of specificity effects, as well as into the debate surrounding the representational nature of long term lexical representations. The scope of investigation of these specificity effects in the series of experiments presented in this chapter and the next two lies within the spoken word recognition memory domain. As such, all the studies in the series employed an "old/new" recognition memory task at test. The encoding task used during the exposure phase was a semantic judgement task consisting of an animacy decision on the words, in order to encourage a deep, semantic encoding of the words in memory. As expected, the results revealed a *voice specificity* effect on recognition memory accuracy, such that listeners were overall less accurate in recognising previously heard (old) words when the talker changed from exposure to test. This effect was not present in the response latency, indicating that listeners were not necessarily faster in recognising previously heard words when they were repeated in the same voice compared to a different voice. The results are discussed in the light of previous findings, existing theoretical approaches and their validity as a basis for the subsequent experiments.

2.1. Introduction

Spoken words display a high scale of variability with respect to the way they are conveyed by the talker. Depending on the physical and acoustical properties of the talker voice, a word is uttered differently across talkers and may even never be uttered the same way twice by the same talker. Listeners encounter speakers of different ages, genders, and accents on a regular basis, facing a great amount of variation in the speech signal. How listeners understand spoken words quickly and accurately despite this variation remains an issue essential to psycholinguistic theories. While variation can be often deemed a problem, in daily life people experience relatively few communicative failures. Early models of spoken word processing considered the high degree of variation in the speech signal as detrimental and irrelevant to the perceptual system, positing a normalisation process to discard it in the early stages of processing (e.g., TRACE: McClelland & Elman, 1986; SHORTLIST: Norris, 1994). However, an extensive body of studies have consistently reported talker-specificity effects in spoken word processing and retention in memory. The common finding is that words that are repeated in the same voice in both exposure/study and test phases of an experiment, are recognised/identified/discriminated more accurately and/or faster than words repeated in a different voice. This indicates that listeners retain talker-specific acoustic details in memory, and that this information in turn facilitates the recognition/identification of previously heard words as well as future understanding of previously encountered speakers (e.g., Creel et al., 2008; Nygaard et al., 1994; Palmeri et al., 1993). In the light of such evidence, exclusively abstractionist theories of lexical memory have not been deemed tenable, leaving the spot to theories maintaining that indexical information is not lost during early perceptual processing, but is stored in long-term memory and can affect subsequent recognition. These accounts have typically acknowledged variation as crucial to explaining how listeners understand spoken words uttered at various speaking rates and styles by various speakers, each with their own vocal properties and idiolect (Episodic: Goldinger, 1996, 1998; Roediger, 1990; Hybrid: Goldinger, 2007; McLennan et al., 2003; Sumner et al., 2014).

While the presence of indexical effects in spoken word processing and representation in memory is well-established, a critical question in the literature concerns *how* the indexical and abstract linguistic knowledge about a word are represented in memory. This complex question is further complicated by inconsistencies in detecting indexical effects across implicit and explicit memory tasks. Two main approaches have attempted to explain this discrepancy, as well as shed light into how indexical and linguistic information can co-exist in memory. According to the *memory systems* approach, memory for the voice in which a word is spoken is retained in a memory system that is separate from the system supporting episodic memory (Schacter & Church, 1992; Tulving, 1972). Alternatively, the *transfer-appropriate* approach posits the existence of a single episodic system in which both indexical and linguistic information are represented (Goldinger, 1996; Hintzman, 1986; Roediger, 1990). These approaches will be discussed in more detail in the next section.

Similar to to other indexical studies, the aim of the present study was to replicate the classical voice specificity effect. Crucially, it sets the foundation for a comparative investigation of the sound specificity effect in the next experiments. The next section will present an overview of several methods frequently used to measure talker-specific indexical effects, as well as controversies with respect to their reliability in detecting these effects. It will conclude with the methodological considerations and the rationale behind choosing the paradigm used in the present study.

2.2. Indexical effects and long term memory - Methodological variation

Indexical effects have been investigated via a range of experimental methods, including behavioural and imaging ones (EEG, fMRI). This section will only focus on behavioural methods, since imaging methods are outside the scope of the research conducted in this thesis. Some of the most frequently used behavioural paradigms have been: long-term auditory priming (e.g., Church & Schachter, 1994), continuous word recognition memory (e.g., Bradlow, et al., 1999, Cooper et al., 2015), word recognition memory (Luce&Lyons, 1998; Mattys & Liss, 2008), perceptual identification (e.g., Pufahl & Samuel 2014), eye-tracking (e.g., Creel et al., 2008). Despite differences, these methodologies share a common 'prototype' paradigm that involves two main parts: an exposure/learning/priming phase and a memory test phase. During the first phase participants are exposed to the stimuli, and depending on the paradigm, they either encode the stimuli in memory by performing a task (e.g., word recognition memory paradigms: Goldinger, 1998), or just passively attend to them, without completing any tasks (e.g., Mattys & Liss, 2008). The second phase involves a memory task for the stimuli heard during exposure.

Despite sharing a common prototype paradigm, studies have varied in terms of the encoding and memory tasks, as well as the type of stimuli used. This variability has sometimes been accompanied by discrepancies regarding the emergence of voice specificity effects, with some studies succeeding in measuring it, and others failing to do so.

2.2.1. Encoding tasks

Encoding tasks have differed with respect to the levels of processing they require, from: shallow (e.g., word classification based on the speaker's gender), to moderate (e.g., phoneme identification), and deep (e.g., syntactic classification, semantic judgement). Some types of encoding tasks have focused on the voice by explicitly asking participants to identify the speaker (Allen & Miller, 2004; Goldinger, 1996; Nygaard & Pisoni, 1998; Nygaard et al., 1994), or to rate the clarity/pitch of pronunciation (Church & Schacter, 1994; Schacter & Church, 1992). On the other hand, some tasks have required the processing of the linguistic information, not that of voice, such as: lexical decision (Luce & Lyons, 1998), categorizing a word (Schacter & Church, 1992), or compute the number of meanings for a word (Church & Schacter, 1994; Schacter & Church, 1992).

Although indexical effects have been found across different levels of processing (Goldinger, 1996), and across tasks that require or do not require the participants to pay attention to voice properties, there is some evidence indicating that the type of encoding during study/exposure might play a role in their emergence and/or robustness, especially for studies using a recognition memory task at test. Using a variety of encoding tasks, Goldinger (1996) showed that the voice effects on recognition memory were most robust when participants encoded the target words at shallow levels of processing. In contrast, other studies using shallow encoding tasks like rating the voice pitch and pleasantness (Naveh-Benjamin & Craik, 1995) and voice clarity (Church & Schacter, 1994) did not find voice effects on recognition memory. To make the issue more confusing, in a series of recognition memory experiments, Sheffert (1998a) found voice effects on recognition memory for a range of different encoding conditions—word enunciation (clarity) ratings, semantic ratings (counting the number of word meanings), word identification (transcribing the word) in noise, and word identification in the clear. One of the main findings was that while the depth of encoding affected the overall recognition memory performance - with semantic encoding producing higher overall accuracies - it had no effect on the retrieval of the voice information, since voice effects were found in all the encoding conditions. The finding that overall recognition memory was higher for the semantically encoded items than the non-semantically encoded ones, was in line with the view that explicit memory performance supports a conceptual level of processing. Namely, counting the number of meanings for each word required more conceptual processing compared to rating the clarity of its enunciation, or writing its spelling.

In contrast, a deep encoding task (semantic judgement), did not induce voice effects in an implicit memory task. Jackson and Morton (1984) had participants make semantic (animacy) judgements on the words during the exposure phase. After a delay period, participants completed an implicit memory task, in which they identified the words embedded in white noise. The results revealed no effect of the voice change from exposure to test had on the word identification performance. However, this null voice effect might have also been due to a complicated design, involving both a between- and a within-subjects manipulation of the talker's voice, and a heterogeneous participant sample with a wide age range.

Despite these occasional inconsistencies, the general trend is that indexical effects seem to be relatively robust across encoding tasks, as evidenced by higher performance for same- over different-voice word repetitions across levels of processing (Goldinger, 1996) and across tasks that do and do not require attention to the talker's voice (e.g., Schacter & Church, 1992).

2.2.2. Implicit and explicit memory tasks

Memory tests typically used to measure voice specificity effects have varied in terms of the extent to which they overtly refer to the initial encoding of stimuli. Explicit tasks require a conscious, direct retrieval of a study/exposure episode (e.g., *old/new* decision), whereas implicit tasks do not directly tap into the memory of a previous event (e.g., word identification in noise). Implicit memory tests are typically considered as supporting perceptual, whereas explicit tasks as tapping more into conceptual processing (e.g., Sheffert 1998; see also Pufahl and Samuel (2014) for a review).

Implicit tasks have been frequently used in auditory priming studies, where memory for studied items is tapped by measuring the effects of priming, a facilitation in responding to a repeated test item at test, that is ascribable to information obtained in the study phase. The general finding- the *voice specificity* effect- is that voice changes from study to test lead to a significant reduction in priming. This has been shown for a variety of implicit tasks, such as: perceptual identification in noise (Church & Schacter, 1994; Goldinger, 1996; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Pilotti et al., 2000; Sheffert, 1998b), identification of low-pass-filtered words (Pilotti et al., 2000, Schacter & Church, 1992), word-fragment completion (Schacter & Church, 1992), lexical decision (Luce & Lyons, 1998), and speeded classification task (Goldinger, 1996). These findings suggest that listeners utilise acoustic details of the speaker's voice, in addition to lexical information. However, the presence of the effect has sometimes been varied as a function of tasks and stimulus quality.

For example, Church and Schacter (1992, also Schacter & Church, 1994) found that study-to-test voice changes produced a significant decrease in priming in stem completion and identification of low-pass filtered words, but not in the identification of words embedded in white noise. Alternatively, Sheffert (1998b) found an effect of the voice change in both the above identification tasks, but only when the words were presented in a degraded form (embedded in noise) at study as well. In contrast, Pilotti et al. (2000) did not find a difference in accuracy as a result of the voice change in word-stem completion and word-fragment completion implicit tasks. According to Pilotti et al. (2000), a crucial difference between the two studies is that Schacter and Church (1992) used a relatively small test set of 48 words spoken by six speakers, while Pilotti and colleagues used a large set of 300 words, spoken by one speaker (either male or female). The acoustic homogeneity of using a single talker at test might have made the voice information perceptually less available compared to the other study that involved an acoustically heterogenous talker set at test. In an acoustically heterogeneous test set, the speakers' voices become salient features of the test material, encouraging reliance on voice information to aid in the completion of the task. On the other hand, in an acoustically homogeneous test set, voice information is not a salient feature of the test material, leading participants to rely more on abstract lexical information as the primary means for completing the task. Additionally, another possible explanation for the discrepancy in the findings might be that using a single test voice renders it familiar over the course of the experiment, relative to a paradigm in which six different voices and far fewer stimuli are used.

At the same time, Pilotti et al's finding of a voice effect when the task was identification of words embedded in noise is in contrast with those of Schacter and Church (1992), Jackson and Morton (1984), and Sheffert (1998b), all of which showed no effects of the voice change on priming in this task. According to Pilotti et al, a plausible explanation behind the discrepancy between their results and those of Schacter and Church and of Jackson and Morton is that their signal-to-noise ratio (SNR) was probably higher (- 5dB) than the one used in their studies. The increase in the background noise level might muffle voice-specific information and render it unattainable at test. However this explanation does not hold in the case of the discrepancy with Sheffert's (1998b) results, since the latter also used an SNR of -5 dB. A plausible reason might lie in an important difference between the stimuli used in each of these experiments. Namely, Sheffert (1998b) used high-frequency words, whereas Pilotti et al. used low- and medium-frequency words. It could be the case that high-frequency words need less processing during encoding than do low- and medium-frequency words. If this speculation is true, then high-frequency words might provide a smaller chance for encoding perceptual information and would hence be less likely to reveal voice effects. Therefore, in addition to variability in the encoding and memory tasks, stimulus-related variation that renders a proper comparison and a plausible explanation of the reported discrepancies even more challenging. As the researchers above also speculated, the number of word stimuli (large vs. small set), talkers (2 talkers vs. more), word frequency (high vs. low) and syllabic structure (mono- vs. bisyllabic) might all play a role in the detection of indexical effects.

On the other hand, the pattern of findings from studies implementing explicit tasks to measure voice effects is somewhat more mixed. Frequently used tests have included: recognition memory (Church & Schacter, 1994; Goldinger, 1996; Luce & Lyons, 1998; Mattys & Liss, 2008; Schacter & Church, 1992), continuous recognition (Palmeri et al., 1993; Bradlow et al., 1999), and cued recall (Church & Schacter, 1994). Some recognition tests have required participants to explicitly discriminate both words and the talker's voice ("old same/old different") (Palmeri et al., 1993; Bradlow et al., 1999), providing a higher chance of finding voice effects.

Indexical effects have not always been detectable with a discrete (*old/*new) recognition task.[3] While some studies implementing this type of task have reported significant changes in recognition performance (Mattys & Liss, 2008; Goldinger, 1996; Luce & Lyons, 1998), others have not (Church & Schacter, 1994; Schacter & Church, 1992; Pilotti at al., 2000). For example, Church and Schacter (1994; also Schacter & Church, 1992) found an effect of the voice change only on the implicit memory task, whereas explicit recognition memory was not affected. Alternatively, studies that implemented the continuous recognition memory design have reported more consistent findings (Bradlow et al., 1995, 1999; Craik & Kirsner, 1974; Palmeri et al., 1993; Saldana, Nygaard, & Pisoni, 1995; Sheffert & Fowler, 1995). In the continuous recognition task participants are required to decide whether a word was previously heard (*old* or *new)* after every trial, in contrast with the discrete recognition memory task, in which participants make this decision after the study phase, followed by a short delay. The typical finding is that as the lag increases between the initial presen-

[3] Discrete: consisting of separate experimental phases, usually involving a short break in between.

tation and the repetition of a word, accuracy decreases and response latencies increase. With respect to the voice specificity effect, the evidence from these studies has consistently shown that same-voice repetitions are recognized more accurately and more quickly than different-voice repetitions at all lag intervals (Craik & Kirsner, 1974), regardless of the number of talkers in the list (Bradlow et al., 1995; Palmeri et al., 1993), and at different signal-to-noise ratios (Saldana et al., 1995).

Although implicit memory tasks seem to display a somewhat more consistent pattern of results than their explicit counterparts (especially, recognition tasks), the comparison between the two types of tasks is not straightforward. There are examples of studies in which voice effects were found with an explicit (recognition) task, but not with an implicit (lexical decision) one, despite using the same encoding task during the exposure phase. For example, in Luce and Lyon (1998), participants performed a lexical decision task during exposure, and at test they performed either another lexical decision (Experiment 1: implicit task) or a surprise recognition memory task (Experiment 2: explicit). The results showed a voice effect on the recognition performance, reflected in the response latencies, with the same-voice repetitions being recognised faster than the different-voice ones. However, there was no effect in the lexical decision task, such that same- and different-voice repetitions produced roughly equal amounts of repetition priming. The null result was in contrast with those of Church and Schacter (1994; also Schacter & Church, 1992). According to Luce and Lyon, one crucial difference between the studies involved the type of implicit task used. Namely, Church and Schacter used tasks in which the perceptual identification of the stimuli was made difficult by degrading the stimuli (either via low-pass filtering, or embedding in white noise). This degradation may have led to processing difficulties that in turn may have rendered reliance on the surface details of the stimuli (voice) stronger, either by slowing processing and allowing more time for specificity effects to emerge, or by encouraging the activation of specific previous memory traces to help identification. A second reason for the discrepancy in the results between these studies may be the speed of responses imposed by the lexical decision task used in Luce and Lyon (1998), which may have not allowed a sufficient time window for specificity effects to emerge (the *time-course* hypothesis is explained in the next section**)**.

There is also evidence suggesting that stimulus-related factors may also play a role in the inconsistencies observed with explicit tasks. For example, Sheffert (1998a) points out that one of the possible reasons behind Schacter and Church's failure to obtain voice effects in their recognition memory tasks could be the syllabic structure, number and nature of the word stimuli used. More specifically, the stimuli set consisted of a relatively small number of multisyllabic and low-frequency words. In addition, the words occasionally formed paired associates (e.g., *lemon* and *lemonade*), which in turn may have promoted organisation into semantic categories and/or inter-item elaboration. According to Sheffert (1998 a), all the studies that had reported voice effects on recognition memory, had used monosyllabic, high-frequency words that could not be classified into semantic categories easily (e.g., *shop, case, group, told*).

As a closing remark to the methodological issues reviewed in this section, it is worth pointing out that despite the varied pattern of methods and findings depicted above, the general trend in the literature favors the view that voice information seems to be encoded in long-term memory along with lexical information, and that it is used in both implicit and explicit tasks (see also Goh, 2005).

2.2.3. Theoretical approaches to discrepancies in indexical studies

The dissociation between explicit and implicit memory tasks in measuring voice effects has been a matter of interest and debate since more than four decades. In this section I review three prominent approaches that have attempted to account for it. The first two are approaches to memory: 1) the multiple memory systems and 2) the transfer appropriate processing view. The third account is the time-course hypothesis, which was also reviewed in **Chapter 1**.

The *multiple memory systems* view proposes that implicit and explicit tasks tap anatomically and functionally into separate memory systems (Schacter, 1987; Tulving, 1972). According to Schacter (1992; also Tulving & Schacter, 1990), performance in explicit memory tasks is mediated by explicit, or episodic memory, which stores information about the spatio-temporal relations of events, such as *when* and *where* a word was heard. Furthermore, explicit memory does not retain perceptual information related to the word form and hence, does not play a role in word priming. On the other hand, performance on verbal implicit memory tasks relies on two pre-semantic representational systems (PRSs) that represent only word forms, not meaning or other associative knowledge. During the encoding stage, two representations of a word form are created. One is abstract in nature, stripped of its surface properties and possibly stored predominantly in the left hemisphere. The other representation of the word is a perceptual one, lacking meaning and stored predominantly in the right hemisphere. Each of these memory subsystems serves word priming differently. One of the main predictions of this memory approach is that sensitivity to study-to-test surface changes of the word form depends on the memory system that the task taps into. Accordingly, the effects of perceptual form change would not be observed via a recognition test, because recognition is an explicit task and, as such, it is served by a memory system that has no access to the perceptual form of a word. Therefore, such effects should only be observed on certain implicit memory tasks that tap into the PRS.

However, the results from a set of studies by Goldinger (1992, 1996) challenged this account. In these experiments, Goldinger used a perceptual identification task in order to investigate implicit memory for spoken words. The task in the study phase was the identification of monosyllabic words embedded in white noise. Following a delay period, the participants completed a test phase, in which they heard the same words in a new order, half of them spoken in the same voice as that during the study phase, and the other half in a different voice. The results revealed that the word identification performance was affected by the study-to-test change in voice, such that listeners were faster and more accurate in identifying same-voice word repetitions compared to different-voice ones. Especially relevant for the above account, there was no dissociation between implicit and explicit memory performance, since the voice effect was also present on the recognition memory performance. The latter finding is consistent with a body of other studies that showed this effect on explicit memory tasks (Craik & Kirsner, 1974; Palmeri et al., 1993; Sheffert & Fowler, 1995; Sheffert, 1998 a). In light of these results, the explanation put forward by Schacter and Church that the null findings on tests of word identification in white noise were due to a lack of contribution of a right-hemisphere PRS, does not seem tenable.

The second theoretical approach offers a *processing* angle and posits a single episodic memory system within which various processes take place. More specifically, the *transfer appropriate processing* view posits that every event leaves a unique episodic trace in memory, including those from the processing during the encoding stage (e.g., Jacoby & Brooks, 1984; Kolers & Roediger, 1984, Roediger, 1990). Words are identified or recognized based on their direct similarity to previous episodic traces. Accordingly, memory transfer depends on a *processing match* between study and test. The existence of a processing match between study and test varies as a result of the memory task at test, since different tasks use different processing mechanisms and hence, retrieve different types of information. For instance, an implicit task such as perceptual identification is considered to be "data driven", because it relies on the extent to which the stimulus format at test resembles the stimulus format at study. Therefore, repeating a spoken word in the same rather than a different format (voice) facilitates the processing of the item, resulting in the listener's higher chance of identifying a degraded version of the word (Jacoby, 1983, Roediger, 1990).

A crucial implication of the transfer appropriate processing view is that in implicit tasks that tap into perceptual processing, this type of processing should be encouraged in *both* encoding and test, rather than in only one of the phases. That is, if the implicit task at test is word identification in noise, then the study task should also include these items presented in noise, rather than in clear (Roediger, 1990). Sheffert (1998b) tested this hypothesis in a series of experiments in which memory for words and voices was investigated with two perceptual identification tests (words in noise and low-pass filtered words), after one of two encoding conditions (identification of words in noise and of words in the clear). At test, the talker's voice was manipulated, such that a word was either presented in the same voice as in the study phase, or in a different voice. The results revealed that changes in voice reduced priming and crucially, that voice-specificity effects were greatest when the type of processing required in the study phase matched the one at test. Sheffert interpreted the results as showing that the *goodness of the processing match* between encoding and test was the primary determinant on the attainability of voice-specificity effects on implicit tests that involve perceptual identification. In contrast, most explicit tasks typically require a task that relies either on conceptual processing or on the match between conceptual information at study and at test. Memory transfer relies on the amount of detailed encoding at study and can be relatively insensitive to changes in the perceptual form of a word (Roediger, 1990).

However, another important aspect of the transfer appropriate approach is that it considers the distinction between conceptual and perceptual processing not to be strictly parallel with that between explicit and implicit memory tasks (Sheffert, 1998b). Namely, an explicit task, such as recognition, may involve two distinct processes: one that associates the test item to a previous episode by using conceptually driven search operations, and one that uses perceptual or item-specific representations for initiating a response. A good perceptual match between study and test items facilitates item processing and as a consequence, the perceived familiarity of the item. The increase in perceived item familiarity may then result in a positive recognition decision even in the absence of conscious retrieval (e.g., Gardiner, 1988; Jacoby & Dallas, 1981, Jacoby, 1983). Therefore, the processing view predicts that changes in the perceptual form of a word will have a detectable impact on explicit memory if the explicit task promotes reliance on perceptual fluency, like for example, by encouraging data-driven perceptual processing, or by reducing processing complexity. Such circumstances would minimize the reliance on conceptually based search, providing listeners with the opportunity to recognize the item on the basis of a more perceptually-driven familiarity. This assumption seems to be supported by studies that found indexical effects on explicit memory tasks by showing that recognition is facilitated when the test form of an item matches its study form (e.g., Goldinger, 1996). Further and stronger support comes from studies that implemented a continuous recognition memory paradigm (e.g., Bradlow et al., 1999; Craik & Kirsner, 1974; Palmeri et al., 1993; Sheffert & Fowler, 1995). In these experiments, words are played in a continuous list, where each one of them is either spoken once, or repeated (in the same or in a different voice) after a number of intervening items (lags). Therefore, there is a high degree of similarity between the first and second occurrences of a word, which may also be one of the reasons behind the consistent pattern of voice effects reported by several studies that have used this task.

The third and more recent approach to explaining the discrepancies in indexical studies, is the *time-course hypothesis*, which emphasizes the *processing time* aspect of the stimuli. Luce, McLennan, and Charles-Luce (2003) argued that inconsistencies are best accounted for by differences in processing time requirements (cf. McLennan & Luce, 2005). Namely, voice specificity effects seem to emerge at a relatively late stage in processing. According to the time-course hypothesis, the degree of reliance on instance-specific information is contingent on the speed with which a response is produced, with slow responses allowing retrieval of episodic traces to a greater extent than faster responses. Evidence in support of this claim has revealed specificity effects for stimuli that are processed relatively slowly (e.g., lower frequency bisyllabic words: Luce et al., 1999; McLennan, 2005; or naturally degraded dysarthric speech: Mattys & Liss, 2008), but not for stimuli that are processed more quickly (e.g., higher frequency monosyllabic words: Luce & Lyons, 1998).

This section provided a review of some of the most prominent theoretical approaches put forward to explain the discrepancies in finding a voice effect in the indexical literature. Whilst no single approach can account for the existing inconsistencies on its own, each one of them offers a different angle into the issue. To briefly summarise the major emerging points, it seems like: 1) the processing match between exposure and test may play a role, and 2) allowing more time for the processing of the stimuli may increase the chance of observing an indexical effect. As the literature review on indexical effects suggests, it is quite difficult to point out a certain type of memory task as ideal. They all involve uncertainty and risks with respect to finding an effect, which makes sense, considering the fact that the effects in question are not large in size. In the section below I outline the rationale behind the recognition memory paradigm implemented in the present study and the ones described in the next two chapters .

2.2.4. The present study - Methodological considerations

In this study, a recognition memory paradigm consisting of an exposure phase, a short delay and a memory test phase was implemented. We used the same encoding task in the exposure phase as Pufahl and Samuel (2014) - *animate/inanimate* semantic judgement on the words - since it accesses the deep, semantic lexicon rather than merely lexical representations. For the test phase, contrary to Pufahl and Samuel (2014) and in line with several other indexical studies (e.g., Luce & Lyons, 1998; Mattys & Liss, 2008; Sheffert, 1998 a), we chose a recognition memory task. There are several reasons behind this decision. Although implicit memory tasks have proved relatively reliable in revealing indexical effects, the general pattern is more complex, with occasional inconsistencies. For example, in their study, Pufahl and Samuel (2014) implemented a deep processing encoding task in the exposure phase (an animate/inanimate semantic judgement task on the words), in which the items were heard in the clear. In the test phase, they used an implicit memory task that involved the identification of degraded versions of the words and their accompanying sound. They justified the choice of an implicit instead of an explicit memory task by citing the transfer appropriate approach, that would not suggest using a deep encoding task in conjunction with an explicit memory task. However, if rigorously followed, the transfer appropriate approach would have also predicted a relatively slim chance of finding specificity effects with their paradigm, based on the processing match of the stimuli between exposure and test. Namely, in Pufahl and Samuel (2014), the stimuli format in the exposure and test phases did not match (i.e., clear vs. filtered).

In our paradigm, the stimuli format is consistent between exposure and test, such that the items are always heard in the clear, without any external noise or other form of degradation added to them. In this respect, the transfer appropriate approach would predict a relatively good chance of finding an effect. A crucial considerations guiding our decision was to keep a rigorous methodological consistency among all the experiments in the series that investigated the sound specificity effect. The major reason for starting this quest with the replication of the voice specificity effect was to provide a robust comparative basis for the subsequent investigation of the sound specificity effect. Given that the series of experiments examining the latter would involve the use of word-sound pairs as stimuli - hence a masking component from the sounds - we chose not to increase the level of perceptual uncertainty by further degrading the stimuli. As Luce and Lyon (1998) also point out, perceptual identification tasks that involve stimulus degradation run the risk of inducing processing difficulties that in turn may promote strategic/explicit reliance on contextual cues (i.e., the talker's voice) to aid identification. Additionally, we reasoned that a perceptual identification task may not be the most plausible one for answering questions related to long-term lexical memory. It could be questioned whether identification of degraded stimuli is more indicative of accessing lexical representations in long term memory, or solving a challenging perceptual identification puzzle.

At the same time, we were also aware of the main concerns arising with using an explicit memory task, such as the question regarding whether it was accessing lexical or episodic memory (see also Luce &Lyon, 1998). We assumed that a recognition memory task entails some kind of lexical access and lexical processing during retrieval. Regarding the episodic nature of the task, the indexical literature suggests that lexical representations seem to have an episodic component, consistent with episodic and hybrid models of the lexicon. Further, based on McLennan and Luce (2005)'s time-course hypothesis, it was deemed reasonable to allow for more processing time during retrieval, rather than encouraging speeded responses with a lexical decision task.

The following sections describe the first experiment of this thesis. Aiming at replicating the voice specificity effect, we expected to see it manifested in the listeners' overall recognition memory performance. Specifically, we predicted that recognition accuracy and/or response latencies would be affected by the change in voice from exposure to test, such that the overall accuracy for the words repeated in the "same" voice would be significantly higher than the accuracy for the words repeated in the "different" voice.

2.3. Experiment 1 - Voice specificity effect

2.3.1. Methods

2.3.1.1 Participants

Forty-nine students at the University of York (Mean age = 21.89 years, SD = 3.90) participated in exchange for either course credit or payment. The number of participants was informed by other indexical studies (e.g., Luce and Lyons (1998) tested 60 participants in each of their two experiments; Mattys and Liss (2008) tested 24 participants for each between-subjects conditions;

Schacter and Church (1992) tested 24 participants for each between-subjects conditions in the first experiment and 48 subjects in the second experiment (within-subjects); Sheffert (1998b) tested 15 participants for each between-subjects condition in both experiments). Given the relatively small magnitude of indexical effects, we targeted a relatively large sample size (N > 40).

All participants provided written consent prior to the experiment. They all identified themselves as native speakers of English and none of them reported a history of hearing or speech and language-related problems.

2.3.1.2. Stimulus Recording

The stimuli were recorded in a sound-attenuated booth by a male and a female speaker. Both speakers spoke Standard British English and were instructed to read the stimuli at a normal pace and neutral intonation in front of a microphone (Shure SM58). The words were digitized at a 44.1-kHz sampling rate using a recording software program (Cool Edit Pro, 2000) and stored in individual audio files. All stimuli were filtered to eliminate background noise and 100 milliseconds of silence was appended to the beginning and end of the words to avoid transition artefacts. In addition, all the sound files were normalized to 68 dB intensity by applying a custom-made script in Praat (Boersma & Weenink, 2013), to make the overall amplitude identical across stimuli.

2.3.1.3. Materials and Design

In line with Mattys and Liss (2008), the stimuli set consisted of 80 bisyllabic words, half of which referred to animate entities and the other half to inanimate entities. All the words were of relatively high frequency, as reported in the CELEX database, with the following mean log frequency values per semantic category:

$M_{animate} = 1.22$, $SD = 0.6$; $M_{inanimate} = 1.38$, $SD = 0.45$). The mean frequencies were not different from each other: $F(1,72.34) = 1.67$, $p > .05$. The mean utterance length for each talker was: $M_{female} = 590.41$ ms, $SD_{female} = 93.08$ ms; $M_{male} = 537.28$ ms, $SD_{male} = 80.25$ ms.

Acoustic analyses performed on the stimuli items produced by the two talkers revealed that the mean difference in fundamental frequencies (F0s) between the male and female talkers was 40.5 Hz ($M_{maleF0} = 115.55$ Hz, $M_{femaleF0} = 156.03$ Hz). A list of the word stimuli is provided in Appendix A.

The experiment involved two phases: Exposure and Test, separated by a short delay of 5 -7 minutes. In each phase, participants heard a list of 60 words, each spoken one at a time. None of the words were repeated within a list. Half the stimuli in each list were produced in the female voice and the other half in the male voice. The 60 words in the exposure phase were the same for all participants, although the voice in which they were heard was counterbalanced across participants. In the test phase, 40 out of the 60 words were repeated from the exposure phase, half in the same voice, half in the different voice. Which words in the test phase were in the same or the different voice was counterbalanced across participants. The counterbalancing across talker (male or female) and talker sameness (same or different) resulted in four stimulus lists in total. Each participant was randomly assigned to one of them. The remaining 20 words in test phase had not been heard in the exposure phase. These were the same for all participants, with half of them spoken in the male and half in the female voice.

2.3.1.4. Procedure

Exposure phase

The experiment was run on the DMDX software (Forster & Forster, 2003). Participants sat individually in a sound-attenuated booth and listened to the trials played binaurally over headphones (Sony MDR-V700) at a comfortable listening level of approximately 68 dB SPL. They read instructions on a computer screen and also listened to the experimenter's explanations. They were instructed to make an "animate/inanimate" decision for each word, where animate and inanimate were defined and examples given (e.g., "banana is inanimate", "professor is animate"). The experimenter encouraged them to be as accurate as possible and not to pay attention to the speaker's voice. 500 ms. after the trial, participants saw a message displayed on the screen prompting them to respond by pressing either one of the corresponding 'shift' keys on the computer keyboard. Namely, on the right side of the screen the word 'ANIMATE' (referring them to the right 'shift' key), and on the left side the word 'INANIMATE' (referring them to the left 'shift' key) appeared. Participants were told to wait for the message to appear on the screen before responding and were allowed a maximum of ten seconds to submit a response. The next trial followed immediately after they pressed a response button, or after the maximum allowed time expired, if no response was provided. There were 60 experimental trials in total and their order was randomized for each participant. No feedback was provided and there was no mention of an upcoming recognition task.

Delay

After completing the first experimental phase, participants left the sound-attenuated booth and spent 5-7 minutes on an unrelated distractor task prior to the memory test. This was done in order to ensure that performance in the subsequent test phase was not based on short-term or working memory. The task consisted of playing an online game (Cube Crash 2).

Test phase

In order to assess the effect of voice change on recognition memory, participants completed a surprise word recognition task. They read written instructions on the screen and listened to the experimenter's explanations. The experimenter explained that they would hear spoken words, some of which were words they had already heard in the first part (*old*) and the others were words they would hear for the first time (*new*). They were informed that for every trial, their task was to decide whether the word was *old or new*, by pressing the respective 'shift' key on the keyboard. They were instructed again not to pay attention to the voice change across trials, as the voice of the talker was not relevant for their task. They were encouraged to be as accurate as possible, but to also press the response key as soon as they made their decision. Participants first saw an 'x' symbol appearing at the centre of the screen, which anticipated the playback of a word. After 500 ms., they heard the word and responded by pressing either one of the 'shift' keys on the computer keyboard (right for the 'old' words and left for the 'new' words). Stickers labeled 'OLD' and 'NEW' were put above the corresponding shift keys. The next trial followed immediately after participants response, or after the maximum allocated time of 10 seconds expired and no response was provided. There were 60 experimental trials in total: 20 old words in the same voice as in exposure (old-same), 20 old words in the different voice (old-different), and 20 new words (half in the male voice and half in the female voice). The order of trials was randomized for each participant.

2.3.2. Results

All participants, except one, displayed mean accuracies above 90% on the animacy judgement task in the exposure phase, indicating that they had successfully encoded the words during the task. The participant that failed to show this performance was excluded from further analysis. A one-way repeated measures ANOVA revealed no difference in accuracy in the exposure phase with respect to the semantic class of the words:

$M_{animate}$ = 98.96 %, $SD_{animate}$ = .02; $M_{inanimate}$ = 99.23 % correct, $SD_{inanimate}$ = .02;

F(1,57) = .395, p > .05.

Overall, 48 out of 49 participants were included in the analysis, which included only the critical (*old*) trials. The response times were measured from the onset of the spoken stimulus to the button press. Only the latencies of correct responses were submitted for analysis and latencies longer than 2 SD above the mean on a subject-by-subject basis were omitted. The data were analyzed using linear (LMER) and generalized mixed-effects regression models (GLMER) (Baayen, Davidson, & Bates, 2008), with recognition accuracy (Accuracy) and response times (RT) as dependent variables. Accuracy was coded as a binary variable with values ‘0’ and ‘1 per trial, where ‘1’ meant a correct response , and ‘0’ an incorrect one.

The fixed factors were Voice Sameness (same or different), Semantics (animate or inanimate), and Exposure Voice (male or female). The factors were coded as binary variables as follows: Voice Sameness: 1 (same), -1 (different); Semantics: 1 (animate), -1 (inanimate); Exposure Voice: 1 (female), -1 (male). Prior to adding any fixed factors to the model, for each dependent variable, we tested the maximal random structure, consisting of random slopes and random intercepts for subjects and items, against the basic structure comprising only the random intercepts. This test was done to assess whether adding random slopes for the fixed factors would be necessary. For the main fixed factor of interest, Voice Sameness, random slopes were added for both subjects and items, whereas for the other two fixed factors, only by-subjects random slopes were included. For the Accuracy variable, the maximal random structure comprising all the random slopes failed to converge, but converged when the random slope of Semantics was excluded from the model. However, this new maximal structure was not statistically different from the basic structure: $\chi2(3) = .59$, p = .90. In the case of the RT variable, the maximal random structure comprising all random slopes converged, but it was not different from the basic structure: $\chi2(4) = 6.43$, p = .17. Despite not being mandatory in this case, we decided to use the maximal random structure whenever it converged. This decision was informed by Barr et al (2013)’s analysis, suggesting that Linear Mixed Effects Regression (LMER) models generalise best when they include the maximal structure justified by the design. However, Barr et al (2013) also notes that for categorical variables like the Accuracy variable here, it may be more difficult for the corresponding maximal Generalized Linear Mixed Effects Regression (GLMER) models to converge, especially when mixed logit models are involved. Henceforth in the analysis, the cases wherein the maximal random structure could not be used as a result of the maximal model's failure to converge, will be explicitly mentioned. Otherwise, it means that the maximal random structure with random intercepts and random slopes for subjects and items has been used.

For every dependent variable, the fixed factors, as well as their interactions, were added incrementally to the base model, and improved fit to the model was assessed using the likelihood ratio test. The base model included only the random terms. The main effects of Voice Sameness, Semantics, and Exposure Voice were obtained by testing the improvement in the model fit when each one of these factors was individually added to the base model.

As predicted, there was a main effect of Voice Sameness on recognition accuracy (the *voice specificity effect*), $\beta = .19$, SE = .06, $\chi 2(1) = 9.26$, $p = .002$.[4] Participants were overall more accurate in recognizing previously heard (old) words when they were repeated in the same talker voice (M_{Same} = 81.77 % correct, SD = 12), compared to when the voice was different ($M_{Different}$ = 76.15 % correct, SD = 12.52). The voice specificity effect is graphically illustrated in Figure 2.1.

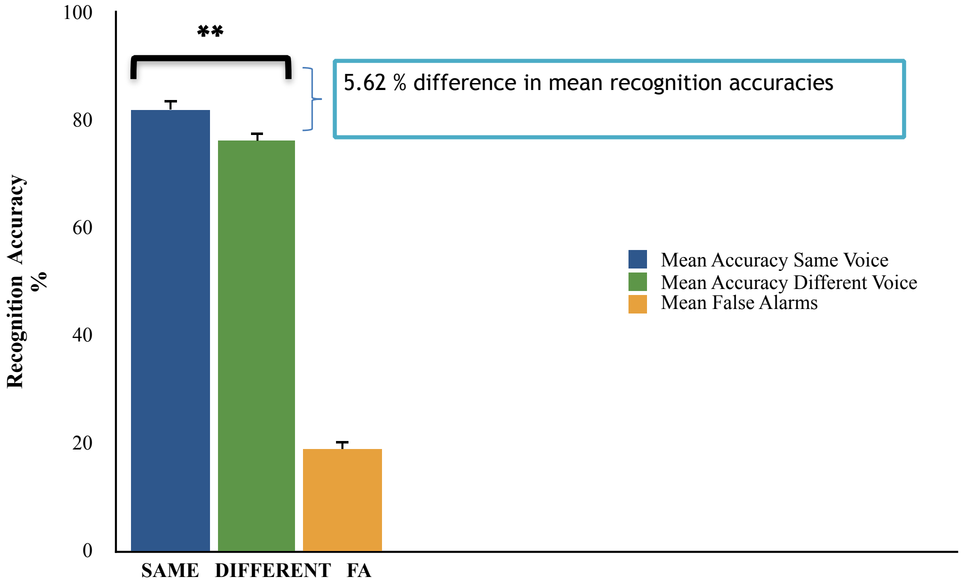

Figure 2.1. Mean recognition accuracy percentages across all participants as a function of the voice change. The mean percentage of false alarms was calculated as (100% - mean accuracy of the responses to the "New" stimuli).

The mean false alarm rate on the new words was $M_{false\ alarms}$ = 19.28 % , SD = 10.62, calculated as: 100% minus the mean hit percentage of the responses to the "New" stimuli (M_{New} = 80.73 % , SD = 10.62). The false alarm rates talker-wise were: M_{male} = 14.59 % , SD = 11.66; M_{female} = 23.96 % , SD = 12.33, and they were significantly different, F(1,47) = 33.71, p < .001, η^2 = .42, indicating a bias towards the female voice for the new stimuli. Since these trials were not critical in

[4] The base and the GLMER models included only the slopes of the fixed factors Voice Sameness (added on a by-subject and by-item basis) and Exposure Voice (added on a by-subject basis), since also adding the slope for Semantics (on a by-subject basis) resulted in these models' failure to converge.

our design and were not eligible for analysis with respect to the sound specificity effect, this bias will not be discussed further.

The voice specificity effect did not manifest in the overall response time (RT). Participants were slightly faster in recognising previously heard words when they were repeated in the same compared to the different voice, M_{Same} = 1184.80 ms., SD = 160; $M_{Different}$ = 1200.99 ms., SD = 173, but this difference was not significant, β = -6.31, SE = 5.48, χ2(1) = 1.32, p = .25.

There was a main effect of Semantics on recognition accuracy, β = .29, SE = .10, χ2(1) = 8.29, p = .004, indicating that overall participants were better at recognizing previously heard words when they were animate compared to when they were inanimate, $M_{animate}$ = 83.33 % correct, $M_{inanimate}$ = 74.58 % correct.[5] However, there was no interaction of Semantics and Voice Sameness, β = .03, SE = .06, χ2(1) = .18 , p = .68, meaning that the robustness of the voice specificity effect on recognition accuracy was not affected by the semantic category of the words.[6] Mean accuracy values for each combination of (Voice Sameness x Semantics) are given in Table 2.1 below:

Table 2.1. Mean recognition accuracy percentage correct values for each combination of Voice Sameness x Semantics.

Voice Sameness x Semantics	Animate	Inanimate
Same	86.04%	77.5%
Different	80.63%	71.67%

There was also a main effect of Semantics on response latencies, β = -23.91, SE = 7.80, χ2(1) = 8.27, p = .004, indicating that participants were faster at recognizing previously heard words when they were animate compared to when they were inanimate, $M_{animate}$ = 1169.05 ms., $M_{inanimate}$ = 1219.62 ms. However, there was no interaction between Semantics and Voice Sameness, β = 7.87, SE = 5.90, χ2(1) = 1.77, p = .18.

No main effect of the Exposure Voice on recognition accuracy was present, β = -0.03, SE = .06, χ2(1) = .19, p = .66, meaning that the voice of the speaker in the exposure phase did not matter for participants accuracy performance in the test phase. Also, no interaction between Exposure Voice and Voice Sameness was found, β = -0.12, SE = .06, χ2(1) = 3.53, p = .06.[7] The mean accuracy values for each combination of (Exposure Voice x Voice Sameness) are displayed in Table 2.2.

[5] Random slopes of the fixed factors Voice Sameness and Semantics were included on a by-subjects basis, since the corresponding base and GLMER models were the only ones that converged.

[6] Random slopes of Semantics (added on a by-subject basis) and Voice Sameness (added both on a by-subject and by-item basis) were included in the base and GLMER models, since they were the only ones that converged.

[7] The base and the GLMER models included only the slopes of the fixed factors Voice Sameness (added on a by-subject and by-item basis) and Exposure Voice (added on a by-subject basis).

Table 2.2. Mean recognition accuracy percentage values for each combination of Exposure Voice x Voice Sameness.

Exposure Voice x Voice Sameness	Different	Same
Female	77.08%	79.58%
Male	75.21%	83.96%

Similarly, there was no main effect of the Exposure Voice on the response times, β = -8.48, SE = 6.30, χ2(1) = 1.77, p = .18, and no interaction of Exposure Voice and Voice Sameness, β = .62, SE = 6.24, χ2(1) = .01, p = .92. Hence, the voice of the speaker in the exposure phase did not matter for participants response speed in the test phase.

In addition to the above analysis, we also performed F1 and F2 analyses to further confirm that the voice specificity effect we found on recognition accuracy was genuine across both subjects and items.

- *By subjects*: A repeated measures ANOVA with Voice Sameness as the within subjects factor revealed a main effect of the vice sameness across subjects, F_1 (1,47) = 9.94, p = .003, η^2 = .175.

- *By items:* A repeated measures ANOVA with Voice Sameness as the within items factor revealed a main effect of the voice sameness across the items, F_2 (1,39) = 8.64, p = .006, η^2 = .181.

In an attempt to test whether the above patterns might be modulated by processing speed (cf McLennan and Luce (2005)'s *time-course hypothesis*), we compared the voice specificity effect on slow and fast respondents. Listeners were categorized as slow or fast on the basis of a median split of their average response latencies. A mixed design measures ANOVA with Voice Sameness (same or different) as the within-subjects factor, and Speed (fast or slow) as the between-subjects factor revealed no main effect of Speed, F(1,46) = .46, p = .50, η^2 = .01. There was a main effect of Voice Sameness (i.e., voice specificity), F(1,46) = 9.8, p = .003, η^2 = .18, but no interaction effect between Speed and Voice Sameness, F(1,46) = .34, p = .57, η^2 = .007. The time-course hypothesis would predict that the slow responders display a stronger indexical effect than their fast counterparts. Our results show no such difference in the effect between fast and slow responders, hence they do not provide support for the time-course hypothesis.

2.4. Discussion and Conclusions

The present study replicated the *voice specificity effect* using a recognition memory paradigm, involving an explicit memory test for previously heard words (e.g., Goldinger, 1998; Luce & Lyons, 1998; Mattys & Liss, 2008). As expected, we found that participants were more accurate in recognizing previously heard words when they were repeated in the same voice, compared to when the voice changed. The effect was not reflected in the reaction times.

Mattys and Liss (2008) reported a similar pattern of results using the same memory task at test. Particularly relevant for the present study is their result regarding the voice specificity effect in the normal speech condition, in which, similar to the present results, the effect was manifested only in the recognition accuracy and not in the response latency. This was not the case for both conditions of dysarthric speech, where the effect was found in both accuracy and response time (Figure 2.2 illustrates this finding).

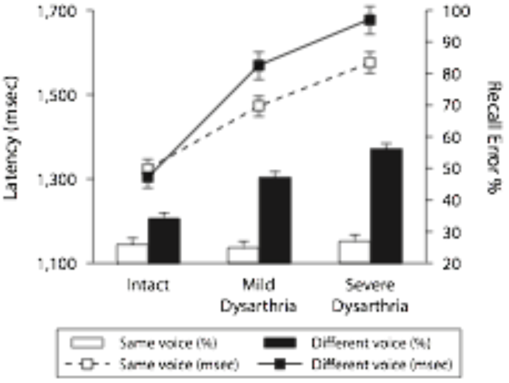

Figure 2.2. From Mattys and Liss (2008). Graph showing the correct recognition latency (lines) and accuracy (bars) as a function of voice similarity and stimulus degradation to to dysarthria. Recall error is calculated as 100 minus the percentage correct of recognised words.

The present findings join the body of indexical studies that found the voice specificity effect using a recognition memory paradigm (e.g., Goldinger, 1996, 1998;Luce & Lyon, 1998; Mattys & Liss, 2008, Sheffert, 1998a). It is worth pointing out that the present pattern of results is slightly different from that in Luce and Lyon (1998). Namely, we found a voice effect in the word recognition accuracy, not in the response latencies, whereas Luce and Lyon found an effect in the response latencies, but not in the recognition accuracy. Potential reasons behind this may be related to methodological differences between our study and Luce and Lyon's. For instance, we used a semantic judgement task in the encoding phase, whereas Luce and Lyon used a lexical decision task. While both tasks presumably access the mental lexicon, the semantic task might provide deeper encoding/access that could in turn promote the appearance of a specificity effect in the recognition accuracy. On the other hand, in line with our results, Mattys and Liss (2008) were able to find a voice effect in recognition accuracy (not response latencies) in the absence of an encoding task. Hence, it seems like the encoding task may not be a key factor in explaining these opposite result patterns.

A second methodological difference regards the type of stimuli used. Namely, like in Mattys and Liss (2008), the stimuli set in the present study consisted of bisyllabic, relatively high frequency words, whereas Luce and Lyon (1998) used monosyllabic words. Perhaps using words of shorter length and duration may have encouraged faster processing and accordingly lower response latencies (which seems to be the case, judging from the mean latency values in Luce and Lyon (1998)), that in turn may have promoted the emergence of the effect in the latency. However, such an argument would be in contrast with the time-course hypothesis, according to which the more processing/response time is allowed, the more likely, and stronger indexical effects are to emerge.

Finally, another methodological difference is the lack of a delay phase in Luce and Lyon (1998). The delay phase could have perhaps helped in consolidating the memory for the words heard during the first phase, which in turn could have led to the effect showing up in the recognition accuracy.[8] However, given that Mattys and Liss (2008) found an effect on accuracy, despite the lack of a delay phase, undermines this possibility as well. Therefore, the methodological differences among the present study and the other two, do not seem to explain the discrepancy in the pattern of results. Nevertheless, the most noteworthy point is that all three studies did find a voice specificity effect, albeit one that manifests itself in different variables of interest (accuracy vs. response latencies).

Contrary to the prediction of the time-course hypothesis, there was no difference with respect to the magnitude of the effect between *fast* and *slow* respondents. Given that response speed was not particularly encouraged, it may be the case that the fast responders were not fast enough for a significant difference between the two groups to show. This result is also consistent with the one that Mattys and Liss (2008) reported for their normal speech condition (very similar to the present experiment) by performing the same analysis (see Figure 2.3 for their results on the matter).[9]

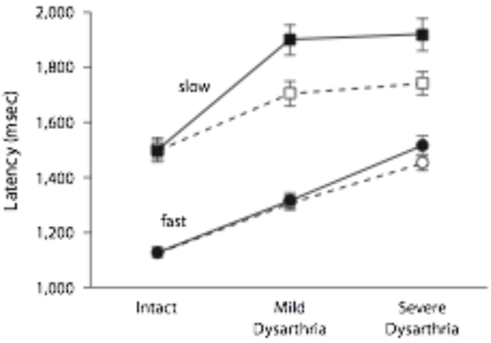

Figure 2.3. From Mattys and Liss (2008). Graph showing the correct recognition latency for slow and fast respondents as a function of voice similarity and stimulus degradation due to dysarthria.

Overall, the present results are consistent with an extensive body of literature indicating that talker-specific details are retained in long-term memory and may play a role in spoken word processing and representation in memory. As such, it can be accommodated by episodic and hybrid models/approaches of spoken word recognition and representation in memory (Goldinger, 1998; Hawkins & Smith, 2001; Hintzman, 1986; McLennan et al., 2003; Sumner, 2014). Abstract models that disregard the information richness that is intrinsic to the variability in the surface properties of the speech signal cannot account for the observed result. It is also important to note that while this study provides support for an episodic component of the mental lexicon, it does not differentiate between the two models. The more recent models of spoken word recognition are steering away from the two extremes of the theoretical spectrum (abstract vs. episodic), positing that the percep-

[8] Although it is arguable that such a moderate time delay could help with consolidation processes.

[9] They observed a difference in the effect between fast and slow respondents only in the dysarthric speech conditions.

tual and memory systems can use and retain either type of information, as dictated by the processing demands imposed by the task.

As Hawkins and Smith (2001) note, the quest for better models of speech processing relies to a good extent on the recognition of the fact that the speech signal is a rich and variable source of information, comprising a high number of different types of linguistic and non-linguistic information. Accordingly, it would not be unreasonable to assume that listeners retain the richness of acoustic information prior to proceeding towards a more abstract representation of speech. Crucially, the speech signal can be considered an *integral* aspect, rather than only a simple carrier of meaning. Obviously, memory encodes experiences, and lexical/linguistic knowledge is attained through continuous experiences. These experiences are ultimately episodic in nature, because the world is episodic. Therefore a view of lexical memory that relies on episodes, at least until a certain level of representation (i.e., until abstraction), seems plausible.

In conclusion, the study described in the present chapter marks a methodological and theoretical starting point for the main aim of this thesis, the exploration of the sound specificity effect in spoken word recognition. The next chapter will begin to investigate this effect, endorsing an energetic masking perspective as a potential explanation behind its emergence.

Chapter 3

The Glimpses Account

Abstract

The study presented in this chapter investigates the role of energetic masking in the emergence of a sound specificity effect. Pufahl and Samuel (2014) found a sound specificity effect on spoken word identification when the paired background sound exemplar changed from exposure to test. It was suggested that mere co-occurrence of the words and background sounds could lead to the word-sound pairs being retained as integrated auditory representations in the lexicon. Here, I argue that mere co-occurrence per se may not be a sufficient, or the only factor in the emergence of a sound specificity effect. The effect could also be accounted for by the presence of different degraded versions (glimpses) of the same word, created by two different masking sounds. It may be the glimpses that are retained in memory, rather than the different word-sound associations per se. The three experiments described here explored this alternative scenario. In line with the previous experiment, the scope of the present investigation also lies within the spoken word recognition memory domain. As such, all the experiments employed the same experimental design as the voice specificity experiment, with identical encoding and recognition memory tasks. The mixed effects regression analyses revealed the anticipated sound specificity effect on recognition memory accuracy, but only when the respective acoustic glimpses of the same word(s) in exposure and test were highly contrasted (Exp. 3). The high contrast was achieved by the combined change of the sound pitch and its temporal overlap with the word. These results provide evidence in favour of an alternative explanation of the sound specificity effect to that of mere co-occurrence. Further, they also point out the susceptibility of such effects to the experimental context in which they are probed.

3.1. Introduction

Spoken words display a great deal of variability that accompanies their acoustical realisations by different talkers. Such variability and its effects on spoken word recognition and representation in long-term lexical memory have been the subject of extensive research and discussion in the past several decades. The previous chapter dealt with this topic and replicated the classical voice specificity effect. This chapter examines another type of variability, that is external to the speech domain. In addition to the speech-intrinsic variability, listeners are also frequently exposed to the external variability of the background in which speech occurs. We rarely hear speech in clear, ideal conditions. On the contrary, it is usually accompanied by background sounds, noises and other speech, all of which also frequently vary. Then, some of the main questions asked in the case of indexical variability become also relevant for this external variability. Namely, what does the perceptual system do with speech-extrinsic variability? Is it discarded immediately and entirely as irrelevant noise, or does it persist at some level in the perceptual and memory systems, impacting the processing, recognition and representation of spoken words? In particular, if the view that regards voices as always co-occurring with spoken words is endorsed, then analogous questions for background sounds co-occurring with spoken words become pertinent. The co-occurrence aspect of the relationship between the two dimensions of speech was critical in Pufahl and Samuel (2014)'s study and the main drive behind the idea that sounds co-occurring with words could also be retained in the mental lexicon in a similar way to voices. More specifically, their main question of interest was whether a change in the co-occurring sound would affect word identification performance in the same way that a change in voice does.

In their first experiment, participants were exposed to words spoken by two talkers, either in isolation, or paired with one of the two exemplars of environmental sounds (e.g., a barking dog, a doorbell sound).[10] In line with the classical indexical paradigm, there was an exposure phase, a delay and a memory test phase. There were four conditions with respect to the voice and sound manipulations from exposure to test: no change, voice change, sound change, both voice and sound change. During the exposure phase participants heard the word-sound pairs in the clear and performed a semantic judgement task on the words only after being instructed to ignore the background sounds as irrelevant. The test phase involved a perceptual identification task, in which participants heard the heavily degraded version of the stimuli (through band-pass filtering) and were asked to transcribe the words. Besides replicating the voice specificity effect, they also found a similar effect of the sound change on word identification accuracy. Namely, listeners were significantly less accurate at identifying the degraded words when the co-occurring background sound changed from exposure to test, compared to when it remained the same. This specificity effect was interpreted as evidence that listeners retain specific acoustic details pertaining to irrelevant, co-occurring background sounds of a spoken word episode in lexical memory. The authors proposed an integrated view of the mental lexicon, in which lexical representations involve a combination of lexical, indexical (voice), as well as speech-extrinsic auditory information (background sounds). In their words: "what co-occurs stays together", which implies that detailed and specific information about the irrelevant co-occurring background sound itself is retained in the mental lexicon, alongside the linguistic information. According to the authors, if indexical effects motivated the expansion of the mental lexicon to include characteristics of the talker's voice alongside the linguistic information, then the new sound specificity effect, obtain by the same indexical paradigm, could require an even further expansion that would additionally incorporate speech-extrinsic auditory information. Furthermore, they posited that the fact that a specificity effect similar to an indexical one is obtained when a speech-extrinsic environmental sound changes between exposure and test suggests that co-occurrence is the critical component, rather than any properties integral to the spoken word.

[10] Every environmental sound had two exemplars (e.g., large barking dog and small barking dog). Every word-sound pairing was unique, i.e., the same sound exemplar was not paired with more than one word.

As a new effect in the indexical literature, the sound specificity effect is exciting and has the potential to advance our understanding of the nature of lexical representations and the structure of the mental lexicon. At the same time, this novelty also necessitates further research for a deeper understanding of its nature and scope. Here, I argue that the effect could also be accounted for by the retention of the degraded versions of the words (acoustic glimpses) produced by the masking sounds, which would make claims about retention of non-linguistic information in memory unnecessary. The next section presents basic background information on masking and the concept of acoustic "glimpses". It concludes with the rationale that motivated the present study.

3.2. Auditory masking and the notion of "glimpses"

When spoken words are paired with background sounds, the sounds are not only co-occurring components in the pairing, but also *masking* ones. Maskers can be anything auditory, from background noises/sounds to competing speech. In the case of background noises, linguistic information may or may not be present in the masker. During masking, the masker signal competes with the target speech signal for processing resources in the auditory system. The properties of masking sounds define the extent to which they compete for the same resources—central or peripheral—as the target speech. As a result, two types of masking have been broadly defined : energetic and informational. Pollack (1975) is credited with coining the terms "energetic/informational masking" and was among the first to propose a distinction between these two types of masking. It is widely accepted that energetic masking usually refers to masking that occurs at the periphery of the auditory system, where the resource competition between the target and masker takes place (i.e., overlapping excitation patterns in the cochlea or auditory nerve). This type of masking is mainly related to the audibility of the target signal and produces partial loss of information at a peripheral level, due to spectral and temporal overlap between the target speech and the noise signal. Since it is typically considered to be directly related to the presence of masker energy in the same frequency region(s) as energy in the target signal, most research in this area has focused on the frequency dimension (Brungart, 2001; Durlach et al., 2003).

Informational masking on the other hand, has proved more challenging to define. It is usually described in terms of what it is not, rather than what it is, and the most frequent definition takes energetic masking as a reference point. Accordingly, informational masking refers to the masking beyond what can be attributed to energetic masking alone. The term has been broadly applied to a wide range of auditory masking processes which may share only the fact that they do not appear to involve energetic masking. In this type of masking, competition for resources seems to be associated with more central auditory processes. The target speech and the noise signals may both be audible, but they may be difficult to segregate, thus hindering the recognition of the target (Brungart, 2001; Rosen et al., 2013). Informational masking therefore depends on factors that inhibit or facilitate stream segregation including linguistic, attentional, and other cognitive factors (Brower et al., 2012; Cooke et al., 2008; Mattys et al., 2009).

Dissociating between energetic and informational masking in auditory processing is not trivial, since for any particular masking signal, the masking effects typically arise from a combination of energetic and informational factors (e.g., Kidd et al., 2007; Lidestam, 2014; Scott et al., 2009; Watson, 2005). For example, if speech is masked with steady-state noise, energetic masking effects will presumably be dominant, whereas masking speech with speech will involve both energetic and informational masking (Scott et al., 2009). The question of how to disentangle the energetic and informational components of the masking effect of noise on speech perception is beyond the aim and scope of this chapter. For our purposes, the important aspect of masking is that if a speech signal is presented together with an acoustic noise signal (such as background sounds), there will be some degree of energetic masking involved.While informational masking cannot be completely discarded in our stimuli, energetic masking is the dominant masking component, therefore it will be the only type of masking we address throughout the rest of the chapter.

The most crucial notion for the issue investigated in this chapter is that of *"glimpsing"* in the presence of background noise/sounds. The successful perception of noisy speech has been often described as "glimpsing" or "dip listening", in which the listener has the possibility to take advantage of the occasional "dips" present in the background noise and perceive relatively undistorted speech portions across the frequency-time domain (Miller & Licklider, 1950; Howard-Jones & Rosen, 1993a; Cooke, 2003; Assman & Summerfield, 2004). These dips occur in two dimensions: temporal and spectral. The temporal dips happen as a result of brief pauses, such as when the overall level of the competing masker signal is low. In these moments, the signal-to-noise ratio (SNR) is relatively high, allowing for 'glimpses' of the target speech to be available to the listener. The spectral dips arise due to the spectrum of the target speech being different from that of the background noise, meaning that there may be certain frequencies of the target speech that are not masked by the background noise signal. This in turn results in a high SNR for those frequencies and enables regions of the target speech spectrum to be glimpsed. The glimpsed acoustic information can then help the listener to infer the complete or near-complete version of the target speech signal (Darwin, 2008). In the case of a fluctuating masker, the process of "glimpsing" ("dip listening") can reduce the effects of energetic masking (Miller & Licklider, 1950; Howard-Jones & Rosen, 1993a; Rosen et al., 2013).

The rest of the chapter will make extensive use of the "*glimpses"* notion, which based on the information outlined above, can be concisely defined as:

- the intelligible left-over of a word after the portions affected by the background masker have been accounted for. Figure 3.1 provides a visual analogy to the concept of masking and glimpses, as well as spectrograms of glimpses of the same word, arising as a result of masking from two different sounds.

Computational models have attempted to quantify the amount of acoustic glimpses that occur during masking, one of the most prominent ones being that proposed by Cooke (2006). This model will be explained in more detail in section 3.7, where the glimpses computation of the stimuli from all the experiments and the respective statistical analysis will be presented.

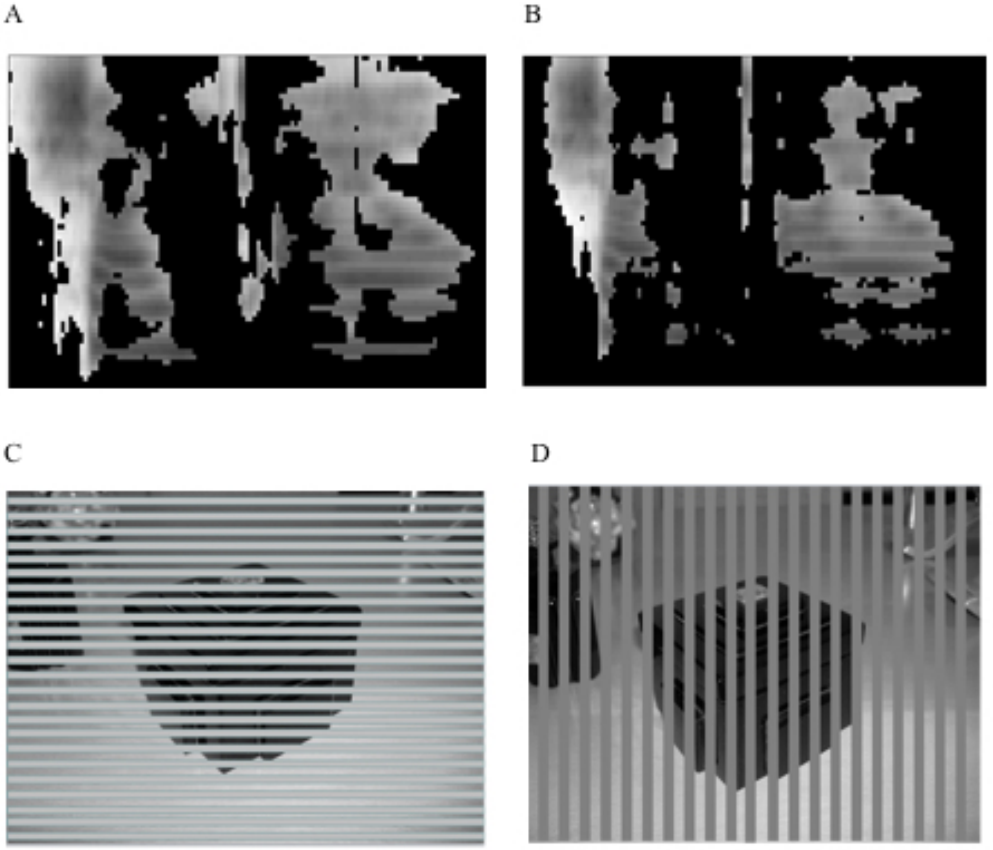

Figure 3.1. Examples of acoustic glimpses and visual masking. The images in A and B represent the spectrograms of the glimpses of the same word after being masked by two different energetic maskers. As a result of different masking, the acoustic left-overs are different. The images in C and D provide an analogy to visual masking and represent the same object being occluded (masked) by two different grid masks. Again, what is left of the object from the occlusion is different in both cases.

In light of the above considerations from the masking perspective, we investigated an alternative hypothesis regarding the sound specificity effect reported by Pufahl and Samuel (2014), as summarised in the following questions:

- Does this effect emerge as a result of the different associations of the same word with two different sounds in exposure and test, or due to the different acoustic glimpses of the same word created as a result of masking by two different sounds?

- The first alternative - the one endorsed by Pufahl and Samuel - implies that the long-term memory episode of the word can retain information about the sound itself, alongside the word. Put in another way, it would mean that the 'word-sound' pairs are integrated in long-term lexical memory as episodic entries (although this is a relatively simplistic conclusion, see Pufahl & Samuel for a more nuanced discussion).

- The second alternative puts emphasis on the masking aspect of the word-sound co-existence. It maintains that the retention of the sounds in memory may not always be necessary for a specificity effect to appear. Rather, it may be the left-over of the word after being masked by the sound(s) that are stored. Therefore, the observed sound specificity effect in Pufahl and Samuel (2014) may not be due to the fact that the word-sound association of the word in the test phase (e.g., word-barking dogB) does not match the one in the exposure phase (word-barking dogA), but because glimpseA of the word at exposure (due to masking from barking dogA) is different from glimpseB at test (due to masking from barking dogB). This alternative scenario would bring along another question of interest:

- Does the degree of glimpse difference matter for the emergence of a sound specificity effect?

It is important to decouple the two alternatives outlined above, especially considering their implications regarding the nature of lexical representations in long-term memory. The set of experiments presented in this chapter aimed at achieving this, by creating a masking context that was favourable to investigate the "glimpses" hypothesis we put forward. Furthermore, different from Pufahl and Samuel's, the present study implements a closer analogy to the typical procedure used to study voice-specificity effects. While they had two talkers and as many sounds as words, we used two sounds, to match the number of talkers used in the voice specificity case (Exp.1, previous chapter).

All the experiments used the same two car horn sounds, with different masking configurations. The second experiment involved two conditions: 2A (Late Masking) and 2B (Early Masking), that differed with respect to the temporal alignment of the sound with the word onset.[11] In experiment 2A, a chance for a moderate word-initial lexical advantage was allowed, such that the sound started later than the word onset. This advantage possibility was later absent in 2B, where the sounds started at the beginning of the words. There has been some debate in the literature of lexical access and spoken word recognition regarding the status of the word-initial information. Whilst some studies have emphasised the importance of an intact word-initial information for successful lexical access and subsequent word identification (e.g., Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987; Marslen-Wilson & Zwitserlood, 1989), others have posited that disrupted word-initial information does not preclude word identification (e.g., Connine et al., 1993; Milberg et al., 1988; Samuel, 1981). However, the possibility of a lexical advantage in the stimuli of 2A was a very minor aspect of the investigation, therefore its investigation does not go beyond comparing the mean recognition accuracies in both the conditions.

Both Experiment 2A and 2B explored the case of a sound specificity effect arising as a result of different acoustic glimpses. Assuming that the change in the sound pitch from exposure to test would create different glimpses of the same word(s), the prediction was the same in both conditions:

- If the acoustic glimpses of the 'word-sound' pairs are retained in memory, then the sound change from exposure to test that also creates different glimpses of the same word(s), should lead to a significant decrease in listeners' overall word recognition accuracy, and/or an increase in their mean response latency. On the other hand, if the glimpses play no significant role in the emergence of the effect, then this pattern of results should not be observed.

3.3. Experiment 2A - Late masking onset

3.3.1. Method

[11] First one in this chapter, second in the entire series.

3.3.1.1 Participants

Fifty-six students at the University of York (Age range: 18 - 23 years) participated in exchange for either course credit or payment. The number of participants in all three experiments described in this chapter was informed by the indexical literature and the very limited number of studies on speech-extrinsic specificity effects. Namely, Cooper et al (2015) tested 44 (36 included in the analysis) and 39 (36 included) participants for each of their two conditions in the first experiment, and 45 (36 included) and 42 (36 included) participants for each of two conditions in the second experiment. Pufahl and Samuel (2014) tested the following number of participants in their experiments: 72 (Exp.1, 64 included), 73 (Exp.2, 64 included), 65 (Exp.3, 64 included), 52 (Exp.4, 48 included), 23 blind adults (Exp.5, 19 included), and 51 (Exp.6) participants.[12] We targeted a relatively large sample size (N > 40), given the reported fragility and the small size of the effects in question. All participants provided written consent prior to the experiment. They all identified themselves as native-speakers of British English and none of them reported a history of hearing or speech and language related problems. None of the participants had done Experiment 1.

3.3.1.2. Materials and Design

The stimuli set consisted of 80 word-car horn sound pairs. The words were the same ones as in the voice specificity experiment, but only in the female voice. Each word was paired with either one of two car horn sounds that had the same intermittent structure, consisting of several 80-ms. sound bursts, separated by silence intervals of varying duration. The intermittent sound structure was selected to create a “glimpsing” context that allowed for a high word intelligibility, and could also be conveniently controlled with respect to the manipulations involved in creating the “Different Sound” condition. The two sounds differed only in their respective pitches (F0s): 190.19 Hz (high pitch), and 131.89 Hz (low pitch). They were both derived from the same car horn sample, from which an 80-ms. portion was edited and manipulated to create a low pitch and a high pitch version. The horn sounds were then customised according to each individual word, such that the word length determined the number of horn bursts and the length of silence intervals between them. In both experimental phases and across all the stimuli, the sound started 80 ms. (the duration of one car horn burst) later than the word , in order to allow for a moderate word beginning lexical advantage. The “Different Sound” condition was realised by the change in the sound pitch. Namely, if a word was paired with the low-pitch car horn sound during Exposure, at Test, it was paired with the high-pitch version of the sound and vice versa. The words and sounds were aligned at the end of the word, mixed at a -3 dB signal-to-noise ratio (SNR) and 100 ms.-silence intervals were appended at the beginning and end of the final mixed stimuli.[13] Visual examples of spectrograms of stimuli in the “Different Sound” condition, prior to mixing are provided in Figure 3.2 (extracted from Praat (Boersma & Weenink, 2013)).

[12] The relevant experiment for the present purposes in Pufahl and Samuel (2014) is Exp.1., the one that reported the sound specificity effect.

[13] The number of horn bursts and the length of the silence intervals between them were calculated by a mathematical formula that used the length of the word, taking into consideration this alignment detail as well.

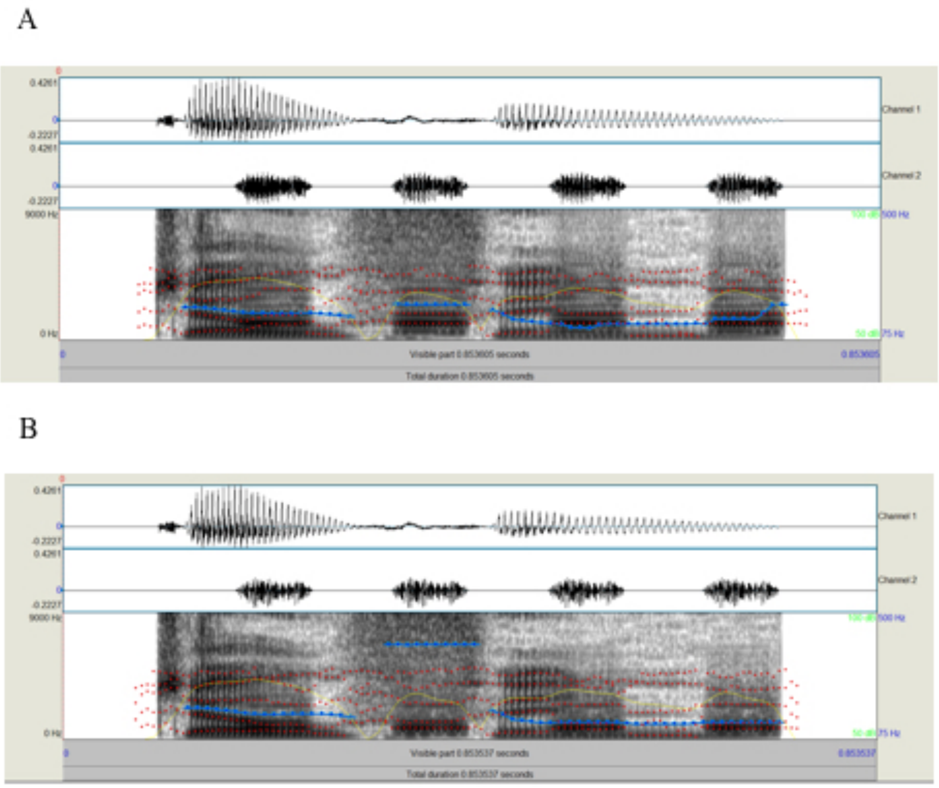

Figure 3.2. Configuration and spectrograms of the stimuli used in Exp.2A, prior to mixing. In the upper part of each image, the upper channel corresponds to the word and the lower channel to the car horn sound. The spectrograms in the lower part of the images represent a mixed version of the two signals, done by the Praat software. The sound starts 80 ms. after the onset of the word. In the example in A, the word is paired with the high pitch version of the car horn, and in B the same word is paired with the low pitch version. In this case, A and B depict the "Different Sound" condition, such that the example in A was played in the Exposure phase and the one in B was played in the Test phase, and vice versa, in a symmetrically counterbalanced way across participants.

3.3.1.4. Procedure

Exposure phase

The experiment was run on the DMDX software (Forster & Forster, 2003). Participants sat individually in a sound-attenuated booth and listened to the trials played binaurally over headphones (Sony MDR-V700) at a comfortable listening level of approximately 68 dB SPL. They read instructions on a computer screen and also listened to the experimenter's explanations. They were instructed to make an "animate/inanimate" decision for each word, where animate and inanimate were defined and examples given (e.g., "banana is inanimate", "professor is animate"). The experimenter encouraged them to be as accurate as possible and ignore the background sound. 500 ms. after the trial, participants saw a message displayed on the screen prompting them to respond by pressing either one of the corresponding 'shift' keys on the computer keyboard. Namely, on the right side of the screen the word 'ANIMATE' (referring them to the right 'shift' key), and on the left side the word 'INANIMATE' (referring them to the left 'shift' key) appeared. Participants were told to wait for the message to appear on the screen before responding and were allowed a maximum of ten seconds to submit a response. The next trial followed immediately after they pressed a response button, or after the maximum allowed time expired, if no response was provided. The order of trials was randomized for each participant. No feedback was provided and there was no mention of an upcoming recognition task.

Delay

After completing the first experimental phase, participants left the sound-attenuated booth and spent 5-7 minutes on an unrelated distractor task prior to the memory test. This was done in order to ensure that performance in the subsequent test phase was not based on short-term or working memory. The task consisted of playing an online game (Cube Crash 2).

Test phase

In order to assess the effect of sound change on recognition memory, participants completed a surprise word recognition task. They read written instructions on the screen and listened to the experimenter's explanations. The experimenter explained that they would hear spoken words, some of which were words they had already heard in the first part (old) and the others were words they would hear for the first time (new). They were informed that for every trial, their task was to decide whether the word was *old or new*, by pressing the respective 'shift' key on the keyboard. They were instructed again to ignore the background sound, as it was not relevant for their task. They were encouraged to be as accurate as possible, but to also press the response key as soon as they made their decision. Participants first saw an 'x' symbol appearing at the centre of the screen, which anticipated the playback of a word. After 500 ms., they heard the word and responded by pressing either one of the 'shift' keys on the computer keyboard (right for the 'old' words and left for the 'new' words). Stickers labeled 'OLD' and 'NEW' were put above the corresponding shift keys. The next trial followed immediately after participants response, or after the maximum allocated time of 10 seconds expired and no response was provided. The order of trials was randomized for each participant.

3.3.2. Results

All participants displayed very high mean accuracies of above 90% correct in the exposure phase, indicating that they had successfully encoded the words during the task. A one-way repeated measures ANOVA revealed no difference in performance in the Exposure phase with respect to the semantic category of the words:
$M_{animate}$ = 98.91 % correct, $SD_{animate}$ = 2; $M_{inanimate}$ = 98.79 % correct, $SD_{inanimate}$ = 3; F(1,54) = .085, p > .05.

One participant displayed very low recognition accuracy (25% correct) for the "old-different" category, and hence was not included in the final analysis. Overall, 55 participants were included in the final analysis, which was performed on the critical (*old*) trials. The response times were measured from the onset of the stimulus to the onset of the button press. Only the latencies of correct responses were submitted for analysis and the ones that were 2*SD above the mean on a subject-by-subject basis were omitted. The two dependent variables were Accuracy (recognition accuracy) and Response Time (RT). Accuracy was coded as a binary variable with values '0' and '1 per trial basis, where '1' meant a correct response to a trial, and '0' an incorrect one. The data were analyzed using linear (LMER) mixed-effects regression models for the continuous RT variable and generalized mixed-effects regression models (GLMER) for the binary Accuracy variable (Baayen, Davidson, & Bates, 2008).

The fixed factors were Sound Sameness (same or different), Semantics (animate or inanimate), and Exposure Sound (high or low pitch sound). The factors were coded as binary variables as follows: Sound Sameness: 1 (same), -1 (different); Semantics: 1 (animate), -1 (inanimate); Exposure Sound: 1 (low pitch horn), -1 (high pitch horn). Like in the previous experiment (**Chapter 2)**, prior to adding any fixed factors to the base model, we tested the maximal random structure of the model for each dependent variable, consisting of random slopes of all the fixed factors and random intercepts for subjects and items. This test was done to see whether adding random slopes for the fixed factors would be necessary. For the main factor of interest, the Sound Sameness, random slopes were added for both subjects and items, whereas for the other two factors, only by-subjects random slopes were added. For the Accuracy variable, the maximal random structure converged and it was statistically different from the structure consisting of only random intercepts, $\chi2(4) = 11$, $p = .03$. For the RT variable the maximal random structure also converged and it was statistically different from the base random structure consisting of only random intercepts for subjects and items, $\chi2(4) = 12.57$, $p = .01$. Based on these results and Barr et al (2013)'s analysis and suggestions, we used the maximal random structure whenever the respective maximal model(s) with the added fixed factors converged. For every dependent variable, the fixed factors (Sound Sameness, Semantics, and Exposure Sound), as well as their interactions, were added incrementally to the base model, and improved fit to the model was assessed using the likelihood ratio test. The base model included only the random terms.

With respect to Accuracy, the change in the paired sound from exposure to test did not lead to a significant decrease in the mean accuracies for the *Same vs. Different* sound conditions, $M_{Acc\_Same\ Sound}$ = 76.91 % correct, SD = 12.74; $M_{Acc\_Different\ Sound}$ = 76.82 % correct, SD = 11.03. Hence, there was no *sound specificity* effect on recognition accuracy, $\beta = .005$, SE = .05 , $\chi2(1) = .007$, $p = .93$, indicating that participants were not more accurate in recognizing previously heard (old) words that were repeated with the same paired sound as in exposure, as compared to the words that were repeated with the different paired sound. There was also no main effect of the sound change on the RT either, $M_{RT\_Same\ Sound}$ = 1405.72 ms., SD = 190.87; $M_{RT\_Different\ Sound}$ = 1391.25 ms., SD = 216.43, $\beta = 10.94$, SE = 10.85, $\chi2(1) = 1.01$, $p = .31$. Hence, participants were not faster in recognizing previously heard (*old*) words that were repeated with the same paired sound as in exposure, compared to the words that were repeated with the different paired sound.

The mean false alarm rate on the new words was: M_{FA} = 20.37 %, SD = 10.71, with 26.18 % , SD = 13.54 for the words accompanied with a LP sound, and 14.55 %, SD = 13.02, for the words accompanied with a HP sound. This difference was, F(1,54) = 30.12, $p < .001$, $\eta_2 = .36$. Since these trials were not critical in our design and were not eligible for analysis with respect to the sound specificity effect, this bias will not be discussed further.

There was a main effect of semantic category (Semantics) on both Accuracy, $\beta = .25$, SE = .09 , $\chi2(1) = 6.98$, $p = .008$, and RT, $\beta = -34.31$, SE = 15.26, $\chi2(1) = 4.82$, $p = .03$. Participants were better and faster at recognizing previously heard words when the words were animate compared to when they were inanimate, $M_{Acc\_Animate}$ = 80.91 % correct, $M_{Acc\_Inanimate}$ = 72.82 % correct,

$M_{RT\_Animate}$ = 1365.2 ms., $M_{RT\_Inanimate}$ = 1435.49 ms.[14] There was no interaction between semantic category and sound sameness on either accuracy, β = -0.01, SE = .05, χ2(1) = .07 , p = . 79, or response times, β = -2.74, SE = 9.35, χ2(1) = .086, p = .77.[15] Hence, participants were more accurate and faster at recognising previously heard animate than inanimate words.

There was no main effect of the exposure sound (high vs. low pitch) on either Accuracy, β = -0.1, SE = .07, χ2(1) = 2.14, p = .14, or RT, β = -7.09, SE = 9.83, χ2(1) = .52, p = .47.

3.3.3. Discussion

The above experiment did not reveal a sound specificity effect on either recognition accuracy, or response latency. Masking from the sounds started late, 80 ms. after the onset of the words (the duration of one car horn sound burst). This was done to allow for a moderate word-initial lexical advantage. The two sounds had the same intermittent structure, differing only in their respective pitches. The glimpses of the same word(s) in each case would be dominated by the unmasked regions (see Figure 3.2), hence would be quite similar. Therefore, the absence of an effect may be due to the fact that the glimpses resulting from this masking configuration were not sufficiently different to elicit the effect.

The next experiment explores the same research question, with the same stimuli, but in a slightly different masking configuration. Namely, in Experiment 2B energetic masking started early, with the word-masker pairs being temporally aligned at the word onset. Like in Experiment 2A, we predicted that the presence of a sound specificity effect on recognition memory would manifest itself in either one of the variables of interest (accuracy/response latency), or both. Namely, the change in the paired sound from exposure to test, leading to different acoustic glimpses of the same word(s), should reveal a significant decrease in listeners' overall word recognition accuracy, and/or an increase in their overall response latency.

3.4. Experiment 2B - Early masking onset

3.4.1. Method

3.4.1.1 Participants

Forty-six students at the University of York (Age range: 18 - 23 years old) participated in exchange for either course credit or payment. All participants provided written consent prior to the experiment. They all identified themselves as native-speakers of British English and none of them reported a history of hearing or speech and language related problems. None of the participants had done any of the previous experiments.

[14] Only random intercepts were included in the respective LMEM and GLMEM, because the maximal models with random slopes for the Semantics fixed factor failed to converge.

[15] For the Accuracy variable, the maximal models did not converge, hence only random intercepts were included.

3.4.1.2. Materials and Design

These were the same as in Experiment 2A, except that now the temporal alignment between words and sounds was different. In the present case, the word and sound were always aligned at the word onset and the sound ended 80 milliseconds (the duration of one car horn burst) earlier than the word. In other words, the honking bursts masked the parts of speech that were left intact in Experiment 2A. Like in experiment 2A, the "Different Sound" condition was realised by the change in the sound pitch (Figure 3.3 provides a visual illustration of this condition). The two signals were mixed at a -3dB signal-to-noise ratio and 100 millisecond-silence intervals were appended at the beginning and end of the final mixed stimuli. The experimental design was identical to the one in experiment 2A, involving the same counterbalancing groups and experimental phases: Exposure and Test, with a short delay in between.

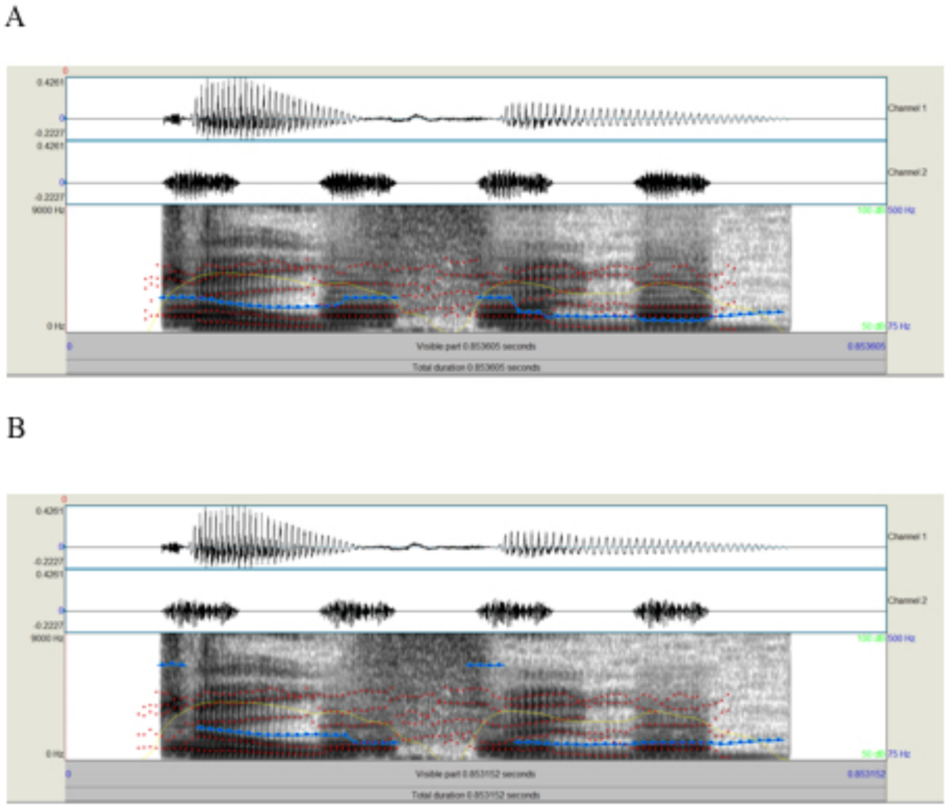
A
Channel 1
Channel 2
Visible part 0.853605 seconds
Total duration 0.853605 seconds
B
Channel 1
Channel 2
Visible part 0.853152 seconds
Total duration 0.853152 seconds

Figure 3.3. Spectrograms of the stimuli configurations used in Exp.2B, prior to mixing. In the upper portion of each image, the upper channel corresponds to the word and the lower channel to the car horn sound. The spectrograms in the lower part of the images represent a mixed version of the two signals in the upper part, done by the Praat software. The word and sound are aligned at the onset of the word. In the example in A, the word is paired with the high pitch version of the car horn, and in B the same word is paired with the low pitch version. In this case, A and B depict the "Different Sound" condition, such that the example in A was played in the Exposure phase and the one in B was played in the Test phase, and vice versa, in a symmetrically counterbalanced way across participants.

3.4.1.3. Procedure

This was the same as in Experiment 2A.

3.4.2. Results

All participants displayed very high mean accuracies of above 90% correct in the exposure phase, indicating that they had successfully encoded the words during the task. A one-way repeated measures ANOVA revealed no difference in performance in the Exposure phase with respect to the semantic category of the words:

$M_{animate}$ = 97.39 % correct, $SD_{animate}$ = .036; $M_{inanimate}$ = 98.04 % correct, $SD_{inanimate}$ = .027; F(1,45) = 1.31, p > .05.

The analysis of performance in the test phase involved the same dependent variables (Accuracy and RT) and fixed factors (Sound Sameness, Semantics, Exposure Sound) as in experiment 2A. The data were coded and analysed in the same way. The maximal random structure for the Accuracy variable, consisting of random slopes for only the Sound Sameness and Semantics converged, but it was not statistically different from the structure consisting of only random intercepts, χ2(3) = 7.24, p = .065. For the RT variable, the maximal random structure consisting of random slopes for all the three factors converged, and it was also statistically different from the base random structure consisting of only random intercepts for subjects and items, χ2(4) = 20.91, p <.001. As before, the maximal random structure was used whenever the respective maximal model(s) with the added fixed factors converged.

With respect to Accuracy, the change in the paired sound from exposure to test did not lead to a significant difference between the mean recognition accuracies for the *Same vs. Different* sound conditions, $M_{Acc\_Same\ Sound}$ = 80.65 % correct, SD = 11.08; $M_{Acc\_Different\ Sound}$ = 81.19 % correct, SD = 12.07. Hence, there was no *sound specificity* effect on recognition accuracy, β = -0.02, SE = .06 , χ2(1) = .083, p = .77, meaning that overall, participants were not more accurate in recognizing previously heard words that were repeated with the same paired sound as in exposure, as compared to the words that were repeated with the different paired sound.

There was also no main effect of the sound change on the mean RTs either, $M_{RT\_Same\ Sound}$= 1216.69 ms., SD = 169.06; $M_{RT\_Different\ Sound}$ = 1225.02 ms., SD = 149.5, β = -6.47, SE = 6.02, χ2(1) = 1.13, p = .29. Hence, participants were not overall faster in recognizing previously heard words that were repeated with the same paired sound as in exposure, compared to the words that were repeated with the different paired sound.

The mean false alarm rate on the new words was: M_{FA} = 18.37 % , SD = .09, with 22.61 % , SD = 11.04 for the words accompanied with a LP sound, and 14.13 %, SD = 11.47, for the words accompanied with a HP sound. This difference was significant, F(1,45) = 19.09, p < .001, η^2 = .30. Since these trials were not critical in our design and were not eligible for analysis with respect to the sound specificity effect, this bias will not be discussed further.

There was a main effect of semantic category (Semantics) on both Accuracy[16], β = .39, SE = .1 , χ2(1) = 13.65, p <.001; and RT, β = -29.48, SE = 10.71, χ2(1) = 6.97, p = .008. Hence, overall participants were better and faster at recognizing previously heard words when they were animate compared to when they were inanimate, $M_{Acc\_Animate} =$ 86.41 % correct, $M_{Acc\_Inanimate} =$ 75.43 % correct, $M_{RT\_Animate} =$ 1193.57 ms., $M_{RT\_Inanimate} =$ 1252.78 ms. There was an interaction of semantic category and sound sameness on Accuracy, β = -0.14, SE = .06, χ2(1) = 4.54 , p = .03, but not on RT, β = .72, SE = 7.1, χ2(1) = .01, p = .92.[17] Participants were overall more accurate and faster at recognising animate words compared to inanimate ones. Further, they were more accurate in recognizing animate words when the paired sound was different and inanimate words when the paired sound was the same as in the exposure phase.

The mean Accuracy values for each combination of sound sameness and semantics are displayed in Table 3.1.

Table 3.1. Mean Accuracy percentage correct values for each combination of Sound Sameness and Semantics.

Sound Sameness x Semantics	Animate	Inanimate
Same	84.35%	76.96%
Different	88.48%	73.91%

There was no main effect of the exposure sound (high vs. low pitch) on either Accuracy, β = -0.05, SE = .07, χ2(1) = .64, p = .42; or RT, β = -6.44 , SE = 8.19, χ2(1) = .61, p = .43.

3.4.3. Discussion

Like Experiment 2A, Experiment 2B failed to find a sound specificity effect on either accuracy or latency. Overall, participants were not more accurate or faster in recognizing previously heard words when they were repeated with the same sound, compared to with the different one. The main conclusion from the above experiments is that neither the contrast between the background sounds nor the contrast in glimpses that these sounds create are sufficient to elicit a sound specificity effect, regardless of whether the masking starts early or late in time. However, perhaps the glimpses of the same word(s) were not sufficiently different to elicit an effect. This possibility may be highly likely considering the fact that the two sounds in question had the same intermittent structure, with only a difference in the frequency domain (pitch) involved. Critically, although in different frequency regions, the sounds masked the same regions of the word(s) in the temporal domain. Hence, the resulting glimpses would have been different, but

[16] Random slopes for all three factors were included for only subjects, as the maximal model containing the by-items random slope of Sound Sameness failed to converge.

[17] For the Accuracy variable, only the by-subject slope of the Semantics factor was included, as the corresponding model was the only one to converge.

perhaps not sufficiently so. More contrasted glimpses of the same word(s) in exposure and test could increase the chance of finding a sound specificity effect.

In Experiment 3, I enhanced the contrast between the glimpses of the same word(s) in exposure and test phase. To this end, a more pronounced masking contrast than the one present in Experiment 2A and 2B was implemented. Crucially, the sound change from exposure to test involved a joint change in both the frequency (pitch) and temporal domain (temporal alignment with the word), rather than just the change in the sound pitch. I reasoned that this manipulation would create more contrasted glimpses, that resulted from different unmasked regions in both the frequency and temporal domains of the word(s). Given that the stimuli in Experiments 2A and 2B differed from each other only in the temporal alignment of the word-sound pairs (onset vs.late onset), their combination provided the necessary stimuli pool for Experiment 3. I predicted that the sound change from exposure to test, leading to more contrasted glimpses of the same words, would elicit a specificity effect in listeners' overall word recognition accuracy and/ or response latency.

3.5. Experiment 3 - Contrasted glimpses

3.5.1. Method

3.5.1.1 Participants

Sixty-eight students at the University of York (Age range: 18 - 27 years) participated in exchange for either course credit or payment. All participants provided written consent prior to the experiment. They identified themselves as native-speakers of British English and none of them reported a history of hearing or speech and language related problems. The information provided in the consent sheets revealed that two participants were not native speakers of English (although highly proficient), and seven other participants were bilinguals that were actively using their other native language. One participant did not provide any answers within the maximum allowed 10 second-time frame per trial, and three other participants had done one of the previous experiments. All these cases were not included in the final analysis, resulting in 55 participants included for analysis.

3.5.1.2. Materials and Design

The stimuli set consisted of 80 word- car horn sound pairs. It was a combination of the stimuli used in Experiment 2A and 2B, where the crucial aspect was the way the "Different Sound" condition was realised. While in the previous two experiments the sound change from Exposure to Test involved only the change in the sound pitch, this time it involved a combined change in two dimensions: the sound pitch and its temporal overlap with the word. There were four word - sound combinations, in terms of the two dimensions of interest: High Pitch Onset Alignment (HPON); Low Pitch Offset Alignment (LPOF); Low Pitch Onset Alignment (LPON); High Pitch Offset Alignment (HPOF). Therefore, there were two ways in which the "Different Sound" condition between Exposure and Test was realized: 1) HPON - LPOF; 2) LPON - HPOF. Every combination was symmetrical with respect to the phase they occurred. For example, if a word was paired in the Exposure phase with the low pitch car horn aligned at the word onset (LPON), in the Test phase it was paired with the high pitch car horn, which started 80 ms. after the word onset (HPOF), and vice versa (Figure 3.4 illustrates this condition). There were no cases in which only the sound pitch, or only the temporal overlap with the word changed from exposure to test. In the "Same Sound" condition both the sound pitch and its temporal overlap with the word were the same in both Exposure and Test. The four word-sound pairing combinations were evenly distributed across the stimuli, resulting in: half of the words being paired with the high pitch car horn sound, and the other half with the low pitch one; half of the words being paired with the sound in the "Onset" temporal alignment, and the other half paired with the sound in the "Offset" temporal alignment (i.e. 80 ms. after the word onset). The counterbalancing in terms of the sound pitch (high or low), its temporal overlap with the word (onset or offset), and sameness (same or different) resulted in a total of 8 stimuli lists (counterbalancing groups). Each participant was randomly assigned to either one of them. The signal-to-noise ratio (SNR) was the same one as in the previous two experiments (-3dB), and 100 ms.-silence intervals were present at the beginning and end of the final mixed stimuli. The experimental design was identical to that in the previous two experiments, consisting of the Exposure and Test phases, with 60 trials played in each.

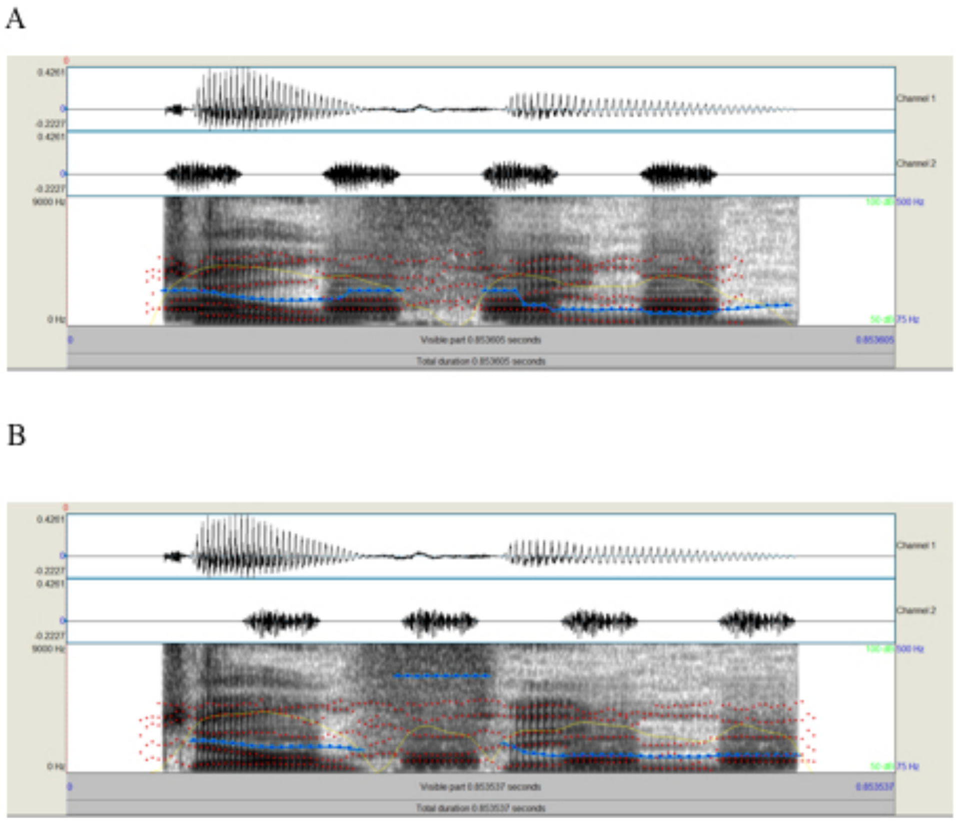

Fig. 3.4. Spectrograms of the stimuli configurations used in Exp.3, prior to mixing. In the upper portion of each image, the upper channel corresponds to the word and the lower channel to the car horn sound. The spectrograms in the lower part of the images represent a mixed version of the two signals in the upper part, done by the Praat software. A and B depict an example of the "Different Sound" condition between Exposure and Test, realised by changing both the sound pitch and its temporal alignment with the word to create highly contrasted glimpses. In the example in A, the word is paired with the high pitch version of the car horn, starting at the word onset, and in B the same word is paired with the low pitch version, starting 80 ms. after the word onset. The critical trials were counterbalanced, hence the opposite configuration to the one shown here was present as well.

3.5.1.3. Procedure

This was the same as in Experiment 2A and 2B.

3.5.2. Results

All participants displayed very high mean accuracies of above 90% correct in the exposure phase, indicating that they had successfully encoded the words during the task. A one-way repeated measures ANOVA revealed no difference in performance in the Exposure phase with respect to the semantic category of the words:
$M_{animate}$ = 98.21 % correct, $SD_{animate}$ = 3.15; $M_{inanimate}$ = 98.89 % correct, $SD_{inanimate}$ = 1.71; F(1,53) = 2.54 , p > .05.

The analysis involved the same variables, Accuracy and RT. There were four fixed factors, the first three of which were the same as in the experiments above: Sound Sameness, Semantics, Exposure Sound Pitch (high vs. low), and Exposure Sound Position (onset vs. offset alignment with the word). The data were coded and analysed in the same way as in the previous two experiments, with the addition of the fourth factor, coded as 1: word onset alignment, and -1: word offset alignment.

For the Accuracy variable, a maximal random structure consisting of random slopes for all factors converged, but it was not statistically different from the structure involving only random intercepts, χ2(5) = 7.66, p = .18. For the RT variable the maximal random structure consisting of random slopes for all factors converged and it was also statistically different from the base random structure consisting of only random intercepts for subjects and items, χ2(5) = 16.54, p = .005. We used the maximal random structure whenever the respective maximal model(s) with the added fixed factors converged.

The analysis revealed the anticipated *sound specificity* effect on word recognition accuracy, such that the mean accuracy for the words repeated with the same paired sound was higher than the mean accuracy for the words repeated with the different paired sound, $M_{Acc\_Same\ Sound}$ = 79.00 % correct, SD = 11.2; $M_{Acc\_Different\ Sound}$ = 75.27 % correct, SD = 14, β = .12, SE = .05, χ2(1) = 4.74, p = .03. Participants were overall more accurate in recognizing previously heard words when the change in the paired sound involved a manipulation of both the sound pitch and its temporal overlap with the word, to create contrasted glimpses of the same word(s) between exposure and test. The mean Accuracy values (percentage correct) for each combination of sound pitch and sound position are depicted in Table 3.2.

Table 3.2. Mean Accuracy percentage values for each combination of Sound Pitch and Sound Position.

Sound Pitch x Position	Word Onset	Word Offset
High Pitch	76%	77.09%
Low Pitch	78.91%	76.55%

The sound specificity effect did not manifest in the mean response latencies, $M_{RT\_Same\ Sound}$ = 1346. 67 ms., SD = 181.14; $M_{RT\_Different\ Sound}$ = 1341.95 ms., SD = 195.96, β = -2.16, SE = 6.59, χ2(1) = .11, p = .74. Overall, participants were not faster in recognizing previously heard words when they were repeated with the same sound configuration (pitch and temporal overlap) as in exposure, compared to when the sound configuration was different.

The overall mean false alarm rate for the Accuracy variable was: $M_{FA} =$ 19.1 % , SD = 8.5.

A two-way repeated measures ANOVA with sound pitch (high vs. low) and position (onset vs. offset) as the within-subjects factors, revealed a main effect of each factor (Pitch and Position) for the responses to the "New" words:

Pitch: F(1,54) = 6.44, p = .01, η^2 = .11; Position: F(1,54) = 10.25, p = .002, η^2 = .16. However, there was no interaction between them: F(1,54) = 1.62, p = .21, η^2 = .03. This result indicates that overall participants were more accurate in identifying the "New" words when the paired sound was temporally aligned with the word at the onset, compared to when it started 80 ms. delayed after the word onset. In the offset alignment configuration, participants were more accurate in correctly identifying the "New" words when the sound was the low pitch car horn compared to when it was the high pitch one. The mean false alarm rates for each combination of sound pitch and sound position are displayed in Table 3.3.

Table 3.3. Mean false alarms percentage values for each combination of Sound Pitch and Sound Position.

Sound Pitch x Position	Word Onset	Word Offset
High Pitch	15%	27.73
Low Pitch	15.76%	19.39%

With respect to the Semantics fixed factor, there was a main effect on both Accuracy, β = .23, SE = .1 , χ2(1) = 5.18, p = .023; and RT, β = -29.35 , SE = 11, χ2(1) = 6.71, p = < .01 Hence, overall participants were more accurate and faster at recognizing previously heard words when they were animate compared to when they were inanimate, $M_{Acc\_Animate} =$ 80.82 % correct, $M_{Acc\_Inanimate}$ = 73.45 % correct, $M_{RT\_Animate} =$ 1318.35 ms., $M_{RT\_Inanimate} =$ 1373.44 ms. However, there was no interaction of Sound Sameness and Semantics for either Accuracy, β = -0.06, SE = .05, χ2(1) = 1.05 , p = .31; or RT, β = -4.29, SE = 6.61, χ2(1) = .42, p = .52.[18] Therefore, the sound specificity effect was not influenced by the semantic category of the words.

There was no main effect of the Exposure Sound Pitch (high vs. low pitch) on either Accuracy: β = .02, SE = .07, χ2(1) = .05, p = .79; or RT, β = 8.99 , SE = 6.72, χ2(1) = 1.78, p = .18.[19] The same was true for the Exposure Sound Position (word onset vs. offset), Accuracy: β = .08, SE = . 05, χ2(1) = 2.25, p = .13; RT: β = -1.59, SE = 6.74, χ2(1) = .06, p = .81. Further, there were no interactions between any of these factors and Sound Sameness for either of the variables. For Accuracy: β = -0.04, SE = .07, χ2(1) = .29, p = .59 (Sound Sameness x Exposure Sound Pitch); β = .09, SE = .06, χ2(1) = 1.79, p = .18 (Sound Sameness x Exposure Sound Position). For RT: β = -15.42, SE = 9.52, χ2(1) = 2.59, p =.11 (Sound Sameness x Exposure Sound Pitch); β = -2.41, SE = 8.85, χ2(1) = .07, p = .79 (Sound Sameness x Exposure Sound Position). Therefore, the sound specificity

[18] For the Accuracy variable, the model containing the interaction term converged when only the by-subject random slope for Sound Sameness was added. Therefore, all the other random slopes were excluded.

[19] For the Accuracy variable, the slope for the Exposure Sound Position was excluded from the model, due to failure to converge.

effect was not affected by either the sound pitch, or its temporal alignment with the word in the exposure phase.

Additionally, F_1 and F_2 analyses were conducted to further confirm that the sound specificity effect we found on recognition accuracy held by subjects and items independently.

- By subjects: A repeated measures ANOVA with Sound Sameness as the within subjects factor revealed a main effect of the sound sameness across subjects, $F_1(1,54) = 5.28$, $p = .025$, $\eta^2 = .09$.

- By items: A repeated measures ANOVA with Sound Sameness as the within items factor revealed a main effect of the sound sameness across the items, $F_2(1,39) = 4.77$, $p = .035$, $\eta^2 = .11$.

3.5.3. Discussion

This study extends the previous two in examining the emergence of a sound specificity effect on recognition memory for spoken words. To this extent, it was designed such that the change in the paired sound between exposure and test would also create highly contrasted glimpses of the same word(s). This contrast was realized by the joint change in two dimensions: the pitch of the sound (frequency domain) and its temporal overlap with the word (temporal domain). Results revealed a small (4%), but significant main effect of the sound change on listeners' overall word recognition accuracy. Namely, listeners were more accurate in recognising previously heard words that were repeated with the same paired sound, compared to the words that were repeated with the different sound. There was no such effect on the overall response latency. The next section offers a comparative analysis between the effects found on the experiments discussed so far (including the voice specificity effect in the previous chapter). It is followed by another section that presents an analysis of the computationally computed glimpse percentages of the stimuli used in the three experiments presented in this chapter.

3.6. Comparative analysis between experiments

A comparative analysis between the experiments described so far was also conducted, in order to better assess the observed s*ound specificity* effect. More specifically, it was compared against the *voice specificity* effect found in the first experiment (**Chapter 1**), as well as against the insignificant effects in Experiment 2A and 2B. If the sound specificity effect is a robust effect, the comparative analysis should reveal that: 1) it is not statistically different from the voice specificity effect, and 2) it is statistically stronger than the insignificant effects in Experiment 2A and 2B. Further, the voice specificity effect was also compared against the insignificant effects in Experiment 2A and 2B, with the expectation that it would be statistically stronger than the insignificant effects.

- *Sound specificity* vs. v*oice specificity*

The two specificity effects in Experiments 1 and 3 were compared. The data sets of both experiments were collapsed into one joint set, which was then used in the analysis. Similar to the analyses for each individual experiment, a mixed-effects regression analysis was conducted, but with an extra fixed factor added, Experiment (2 levels), coded as: 1 (Exp. 1, previous chapter) and 2 (Exp. 3). The main effects and interaction of the factors Sameness (Voice/Sound) and Experiment were assessed.[20] Since the voice and sound specificity effects were found only for the Accuracy variable, it was the only one included in the analysis.

As expected, there was a main effect of Sameness on Accuracy: β = .15, SE = .04, χ2(1) = 12.09, p < .001.[21] No main effect of Experiment was found, β = -0.12, SE = .14, χ2(1) = .79, p =.38.

There was also no interaction between Sameness and Experiment, β = -0.07, SE = .08, χ2(1) = .76 , p =.38, indicating that the voice and sound specificity effects were not statistically different, hence corroborating the first prediction.[22]

- *Sound specificity* vs. *no sound specificity*

The sound specificity effect found in Experiment 3 was compared against the insignificant effects found in Experiment 2A and 2B. A mixed-effects regression analysis was conducted, with the extra fixed factor, Experiment (3 levels), coded as: 1 (Exp. 2A), 2 (Exp. 2B), and 3 (Exp. 3). The main effects and interaction of Sound Sameness and Experiment were assessed. Since the sound specificity effect was found only for recognition accuracy (Accuracy), the analysis was performed only on this variable.

There was no main effect of Sound Sameness on Accuracy when all three experiments were included in the comparison: β = .04, SE = .03, χ2(1) = 1.45, p = .23. No main effect of Experiment was found either, β = -0.12 , SE = .04, χ2(1) = .04, p =.85. Separate comparisons between the sound specificity effect and each individual insignificant effect also revealed no main effect of Sound Sameness on Accuracy: 1) β = .06, SE = .04, χ2(1) = 2.58, p = .11 (Exp.2A-Exp.3); 2) β = .06, SE = .04, χ2(1) = 2.16, p = .14 (Exp.2B-Exp.3).[23] There was no main effect of Experiment either in each of these comparisons: 1) β = .01, SE = .06, χ2(1) = .03, p = .9 (Exp.2A-Exp.3); 2) β = -0.25, SE = .14, χ2(1) = 3.27, p = .07 (Exp.2B-Exp.3).[24]

There was no interaction between Sound Sameness and Experiment when all three experiments were included in the comparison, β = .06, SE = .04, χ2(1) = 2.31, p =.13.[25] Separate comparisons between the sound specificity effect and each individual insignificant effect also revealed no interactions: 1) β = .06, SE = .04, χ2(1) = 2.27, p = .13 (Exp.2A-Exp.3)[26]; 2) β = .13, SE = .08,

[20] In all the models involved in the comparative analysis, slopes for the Experiment factor were added only on a by-item basis.

[21] Only the random slope of Sound Sameness (added by-subjects and by-items) and the random slope of Experiment (added by-items) were included in the respective models.

[22] Only the random slope of Sound Sameness (added by-subjects) and the random slope of Experiment (added by-items) were included in the models.

[23] Only the random slope of Sound Sameness (added by-subjects and by-items) and the random slope of Semantics (added by-subjects) were included in the respective models.

[24] Only the random slope of Sound Sameness (added by-subjects) and the random slope of Experiment (added by-items) were included in the models.

[25] Only the random slope of Sound Sameness (added by-subjects) and the random slope of Experiment (added by-items) were included in the models.

[26] Only the random slope of Sound Sameness (added by-subjects) was included in the respective models.

$\chi2(1) = 2.69$, $p = .1$ (Exp.2B-Exp.3)[27], suggesting that the sound specificity effect observed in Experiment 3 was not statistically different from its insignificant counterparts in Experiment 2A and 2B. Therefore, the second prediction was not satisfied.

- *Voice specificity vs. no sound specificity*

The voice specificity effect found in Experiment 1 was compared against the insignificant effects found in Experiment 2A and 2B. A mixed-effects regression analysis was performed, with the extra fixed factor, Experiment (3 levels), coded as: 1 (Exp. 1, previous chapter), 2 (Exp. 2A), and 3 (Exp. 2B). The main effects and interaction of Sameness (Voice/Sound) and Experiment were assessed. Since the voice specificity effect was found only for recognition accuracy (Accuracy), the analysis was performed only on this variable.

There was no main effect of Sameness on Accuracy when all three experiments were included in the comparison: $\beta = .06$, $SE = .03$, $\chi2(1) = 2.77$, $p = .1$. No main effect of Experiment was found either, $\beta = .07$, $SE = .07$, $\chi2(1) = .85$, $p = .36$. Separate comparisons between the voice specificity effect and each insignificant effect revealed a main effect of Sameness on Accuracy: 1) $\beta = .08$, $SE = .04$, $\chi2(1) = 4.60$, $p = .03$ (Exp.1-Exp.2A); 2) $\beta = .09$, $SE = .04$, $\chi2(1) = 4.41$, $p = .04$ (Exp.1-Exp.2B).[28] There was no main effect of Experiment in each of these comparisons: 1) $\beta = -0.14$, $SE = .14$, $\chi2(1) = 1.00$, $p = .32$ (Exp.1-Exp.2A); 2) $\beta = .07$, $SE = .07$, $\chi2(1) = .82$, $p = .36$ (Exp.1-Exp.2B).[29]

There was an interaction between Sameness and Experiment when all three experiments were included in the comparison, $\beta = -0.10$, $SE = .04$, $\chi2(1) = 5.90$, $p =.02$. Separate comparisons between the voice specificity effect and each insignificant effect also revealed significant interactions: 1) $\beta = -0.18$, $SE = .08$, $\chi2(1) = 5.11$, $p = .02$ (Exp.1-Exp.2A)[30]; 2) $\beta = -0.10$, $SE = .04$, $\chi2(1) = 5.85$, $p = .02$ (Exp.1-Exp.2B).[31] These results indicate that in line with the prediction, the voice specificity effect was statistically stronger than the insignificant sound specificity effects. The pattern of specificity effects across the experiments discussed so far (this chapter and the previous one) is graphically depicted in Figure 3.5.

[27] Only the random intercepts by-subjects and by-items were included in the respective models, since adding any of the random slopes led to the models' convergence failure.

[28] Only the random slope of Sound Sameness (added by-subjects and by-items) and the random slope of Experiment (added by-items) were included in the respective models.

[29] Only the random slope of Sound Sameness (added by-subjects and by-items) and the random slope of Experiment (added by-items) were included in the respective models.

[30] Only the random slope of Sound Sameness (added by-subjects and by-items) and the random slope of Experiment (added by-items) were included in the respective models.

[31] Only random intercepts for subjects and items were included in the models, since adding any random slope led to the models' convergence failure.

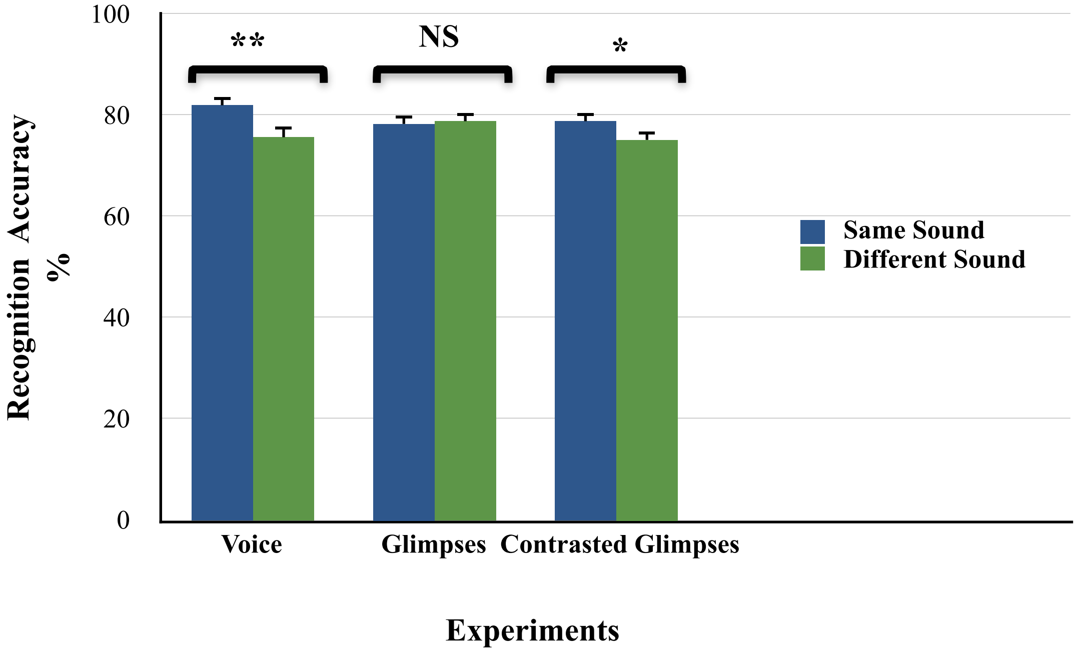

Figure 3.5. The emergence pattern of specificity effects illustrated across experiments, via the mean recognition accuracy percentages as a function of the voice or sound change. The bars in the left represent the *voice specificity* effect; the bars in the middle (Glimpses) the insignificant effect collapsed across Experiments 2A and 2B; and the bars on the right the *sound specificity* effect found in Experiment 3.

3.7. Glimpse computation and analysis

In order to assess whether the acoustic glimpses of the same word(s) at exposure and test were quantitatively different, a computational analysis on all the stimuli used in the above experiments was performed. This analysis produced quantitative measures of the glimpses, expressed in the form of glimpse percentages, that were calculated by using a computational model of speech perception in noise, the *glimpse detection model* (Cooke, 2006). More specifically, the model is based on the use of glimpses of speech in spectro-temporal regions where it is least affected by the background masking. It uses as input simulated spectro-temporal excitation patterns (STEP: Moore, 2003), which are smoothed and compressed representations of the envelope of the basilar membrane response to sound and are typically considered good first-order representations of auditory stimuli at an early stage of processing. Based on the assumption that listeners may be unable to detect very brief regions of speech target dominance, or regions that occupy a very narrow portion of the spectrum, the glimpse detection model includes a *minimum glimpse area* criterion. Namely, all connected regions of spectro-temporal elements that satisfy a given local signal-to-noise (SNR) criterion also have to possess an "area" (i.e., glimpse extent) greater than a specified amount. In this context, "area" is defined as the number of time–frequency elements making up the glimpsed region. Cooke (2006) draws attention to the fact that this is not an area in the traditional sense, since the time and frequency units are not identical. Additionally, different choices of time and frequency resolution in the STEP would produce slight differences in the calculated "areas". This model uses choices based on those commonly employed in studies involving STEPs. For the present glimpses calculations, the spectro-temporal excitation pattern used as input to the model was initially processed by a bank of 55 gamma- tone filters (Patterson et al., 1988), between 100 and 8000 Hz. The SNR criterion was 3 dB, meaning that speech had to be 3 dB stronger than the masker to be counted as a glimpse. This threshold value constitutes a relatively conservative approach to glimpse counting, increasing the confidence in thinking that they are glimpses that listeners make use of. The calculation of glimpses was based on 5 ms frames, rather than time sample by time sample, due to the high level of excitation-related fine structure in the stimuli, for which the latter approach would not work very well. The glimpse percentage value that the computational analysis produces for a masked speech input[32], corresponds to the percentage of all the individual glimpses in the input that meet the criteria mentioned above. We ran statistical tests on the calculated glimpse percentages of the stimuli in each experiment, to assess the difference between the glimpses of the same word(s), corresponding to the two different masking sounds. Only the critical trials (i.e., those corresponding to the "Old" words) were analysed.

- *Experiment 2A*:

The mean glimpse percentages for each car horn sound were: $M_{Glimpses\_HP\ Sound} =$ 42.95 % SD = 3.75; $M_{Glimpses\_LP\ Sound} =$ 45.35 % , SD = 3.35, F(1,39) = 174.14, p < .0001, η^2 = .82. On average, there were more glimpses resulting from the low pitch masker compared to the high pitch one. This probably reflects the fact that the words were spoken by a female talker, whose voice was high pitch, hence more masking occurred in the higher frequencies.

Hence, the sound change from exposure to test led to overall statistically different glimpse percentages of the same word(s). Visual examples of glimpses for a word-masker pair are provided in Figure 3.6.

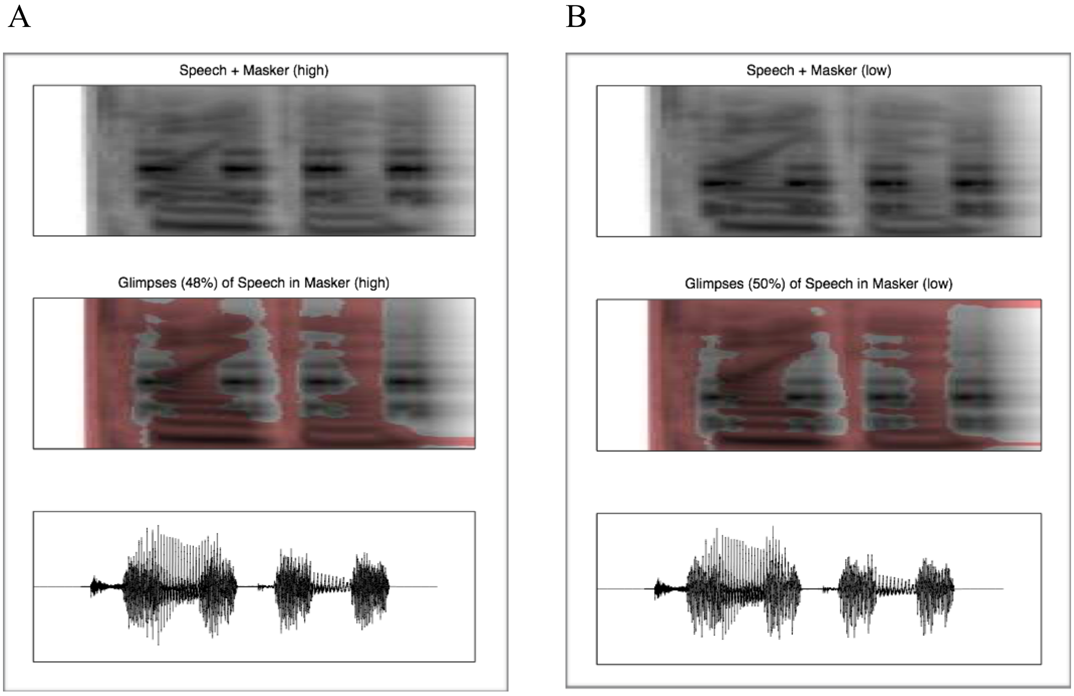

Figure 3.6. Spectrograms of the mixtures and resulting glimpses of the same word, paired with the two car horn sound in Exp. 2A. The masked regions are shown in black and the glimpses in red, with masking starting 80 ms. later than the word onset. Both the glimpse areas and the respective overall glimpse percentages are shown. The two images represent an example of the *Different Sound* condition. The image in A depicts the case when the word is paired with the high pitch masker and the image in B the case when the same word is paired with the other masker, the low pitch one. Although the sound change leads to quantitatively different glimpses, they are qualitatively similar, with the masking happening in very similar regions.

[32] In the present case, word(s) masked by sound(s).

- *Experiment 2B*:

The mean glimpse percentages for each car horn sound were: $M_{Glimpses\_HP\ Sound}$ = 51.95 % SD = 2.98; $M_{Glimpses\_LP\ Sound}$ = 54.35 % , SD = 3.09, F(1,39) = 145.87, p < .0001, η^2 = .79. Like in Experiment 2A and as expected, there were overall more glimpses resulting from the low pitch masker compared to the high pitch one. Visual examples of a word-masker pair in both experimental phases, and the respective glimpses are provided in Figure 3.7.

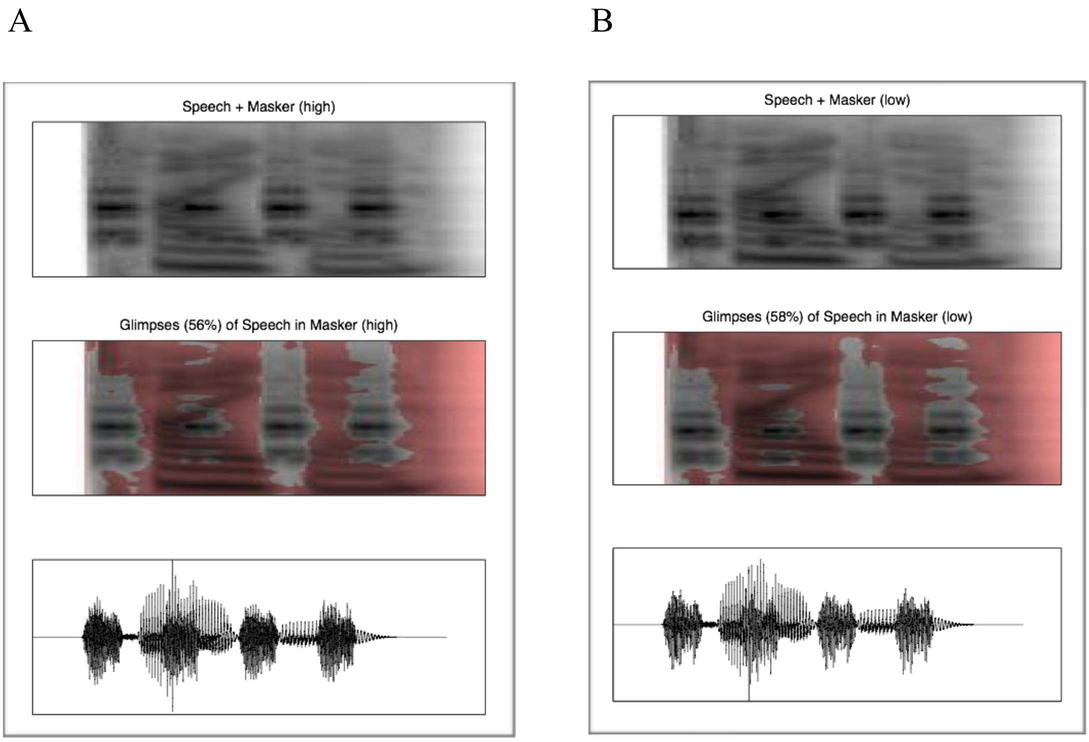

Figure 3.7. Spectrograms of the mixtures and resulting glimpses of the same word, paired with the two car horn sound in Exp. 2B. The masked regions are shown in black and the glimpses in red, with masking starting at the word onset. The two images represent an example of the *Different Sound* condition. The image in A depicts the case when the word is paired with the high pitch masker and the image in B the case when the same word is paired with the other masker, the low pitch one. Although the sound change leads to quantitatively different glimpses, they are qualitatively similar, with the masking happening in very similar regions.

- *Experiment 3*:

The stimuli used in this experiment were a cross-over combination of the stimuli in Experiments 2A and 2B. There were four different word-masker combination in terms of the two dimensions of interest: sound pitch and its temporal alignment with the word. Therefore, there were two *"*Different Sound*"* combinations: 1) High Pitch Onset Alignment (HPON) - Low Pitch Offset Alignment (LPOF); 2) Low Pitch Onset Alignment (LPON) - High Pitch Offset Alignment (HPOF). The mean glimpse percentages for each pairing combination were (repeated here for the reader's convenience from the two prior sections): $M_{Glimpses\_HPON\ Sound}$ = 51.95 % SD = 2.98 ; $M_{Glimpses\_LPOF\ Sound}$ = 45.35 % , SD = 3.35; $M_{Glimpses\_LPON\ Sound}$ = 54.35 %

SD = 3.09; $M_{Glimpses\_HPOF\ Sound}$ = 42.95 % , SD = 3.75. A two-way repeated measures ANOVA with Combination (2 levels) and Masker (2 levels:high vs. low pitch) as the within items factors revealed a robust main effect of the Masker on the glimpse values, F(1,39) = 396.73, p < .0001, η^2 = .91, indicating that the two masking sounds, contrasted in both pitch and temporal alignment with the word, led to significantly different glimpses of the same word(s). There was no main effect of Combination, F(1,39) = 0, p = 1, but there was an interaction of Combination and Masker, F(1,39) = 298.52, p < .0001, η^2 = .88, indicating that one of the "Different Sound" combinations (combination 2: LPON-HPOF) led to a greater glimpse difference than the other combination. However, the main result of interest for the present analysis is that the sound change from exposure to test led to significantly different glimpses of the same word(s). Figure 3.8 shows an example of a word-masker pair , and the respective glimpses. As it can be seen, in this case, the glimpses are both quantitatively and qualitatively different, with masking occurring in different spectro-temporal regions. This enhanced masking contrast seems to be the critical factor in the emergence of a sound specificity effect in Experiment 3.

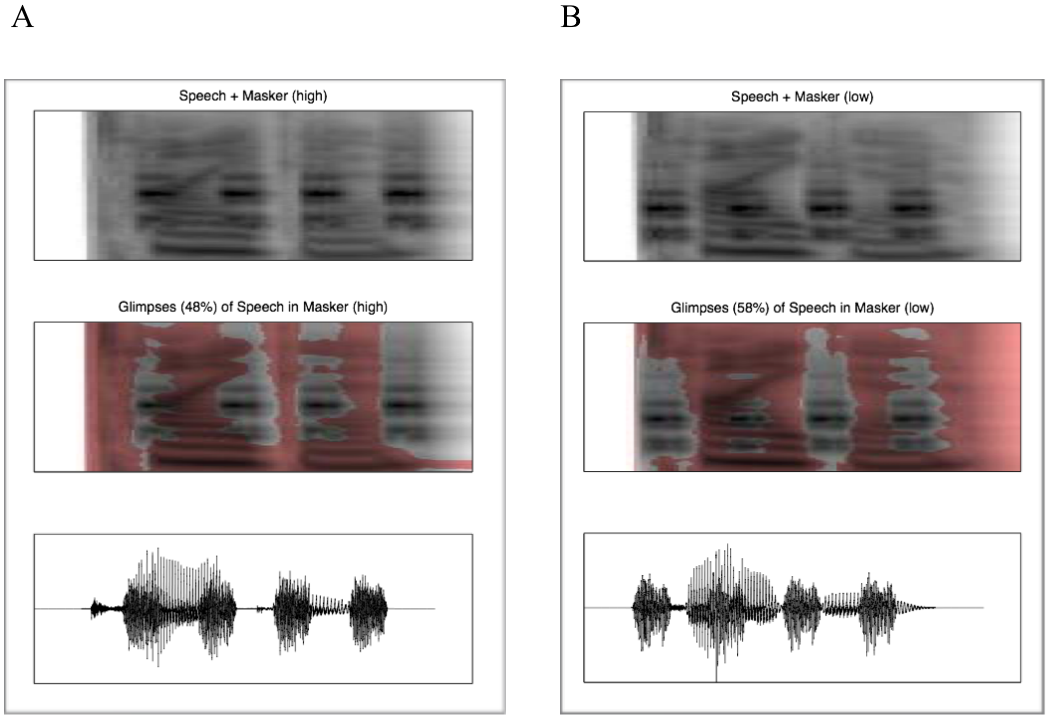

Figure 3.8. Spectrograms of the mixtures and resulting contrasted glimpses of the same word, paired with the two car horn sound at different temporal positions in Exp.2. The masked regions are shown in black and the glimpses in red, with masking starting either at the word onset, or 80 ms. delayed. The two images represent an example of the *Different Sound* condition, with each one belonging to either the Exposure or Test phase. The image in A depicts the case when the word is paired with the high pitch masker that starts 80 ms. after the word onset, and the image in B illustrates the case when the same word is paired with the other masker, the low pitch one, that starts at the word onset. The change in sound in this case leads to highly contrasted glimpses, as a result of the joint change in both the sound pitch and its temporal overlap with the word. The glimpses are both quantitatively and qualitatively different.

3.8. General discussion and conclusions

This chapter explored an account of the sound specificity effect in terms of the acoustic glimpses (intelligible left-overs) of the same word(s), created by the distinct masking of two co-occurring sounds. To this end, the present experiments created favourable and controlled "glimpsing" contexts, where spoken words were paired with one of two car horn sounds that had the same intermittent structure and different pitches. Experiments 2A and 2B explored the hypothesis that different glimpses of the same word(s), created by the change in the paired sound from exposure to test, would elicit a sound specificity effect. The glimpse difference was realised by the change in the pitch of the paired car horn sounds. As the statistical analysis of the computationally computed glimpse percentages also demonstrated, the glimpses of the same word(s) corresponding to the two masking sounds were indeed significantly different from each other, with the low pitch car horn sound creating a higher overall glimpses percentage. Nevertheless, there was no sound specificity effect in either 2A, or 2B. The failure to find an effect in these experiments encouraged the idea to implement more contrasted glimpses of the same word(s) between exposure and test in Experiment 3. We reasoned that although different, the masking from the two background sounds in Experiment 2A and 2B occurred in the same word regions temporally, therefore it may have led to glimpses that were not sufficiently contrasted in eliciting the targeted effect. The high glimpse contrast in Experiment 3 was achieved by a combined change in both the sound pitch and its temporal alignment with the word(s). The intermittent structure of the car horn sounds was particularly convenient for both the creation of glimpsing opportunities and the manipulation of the temporal alignment between the words and their paired sounds. As anticipated, a sound specificity effect was present in the word recognition accuracy. Similar to Experiment 2A and 2B, the computed glimpses of the same words, corresponding to the two different masking sound configurations, were quantitatively different. I argue that what seems to be crucial for the emergence of a sound specificity effect from a glimpses perspective, is the combined quantitative and qualitative difference in the glimpses. More specifically, in Experiment 3, the two sounds masked different regions in both the frequency and temporal domains of the words, creating both quantitatively and qualitatively different glimpses of the same word(s) in exposure and test.

The comparative analysis among the experiments in this chapter revealed that the sound specificity effect was not statistically different from the voice specificity effect. However, no interactions between the sound specificity effect in Experiment 3 and the insignificant effects in Experiment 2A and 2B were found. This is not entirely unexpected, considering the relatively small magnitude of specificity effects in general and that of the sound specificity in particular. On the other hand, there was an interaction between the voice specificity effect (Experiment 1, Chapter 2) and the insignificant effects in Experiment 2A and 2B. This suggests that the voice specificity effect seems more robust than the sound specificity effect, which is also consistent with the overall findings of Pufahl and Samuel (2014).[33] It is not very surprising that indexical information pertaining to human voices seems to persist better in memory than the information associated with external sounds co-existing with spoken words. The auditory system is tuned to detecting and interpreting changes in human voices as meaningful events with functional value. Tracking these changes has

[33] Across experiments, Pufahl and Samuel (2014) found an average of about 6% decrease in the overall word identification accuracy as a result of the voice change from exposure to test, and an average of about 3% decrease in identification accuracy from the sound change. Similarly, we found a higher decrease in accuracy for the voice change (5.63% decrease in the overall word recognition accuracy) compared to the sound change (3.73% decrease in recognition accuracy).

useful practical and adaptive functions for successful speech understanding and communication. The functional value of tracking changes in background sounds co-occurring with spoken words, on the other hand, is arguably less relevant for the auditory system. Consistent with existing literature (e.g., Pufahl and Samuel, 2014), the comparative analysis indicates that the sound specificity effect observed in Experiment 3 is fragile and less stable than its voice counterpart.

In the experiments discussed in this chapter, the fact that a specificity effect only appeared when the co-occurring sounds created highly contrasted glimpses between conditions, undermines to a certain degree the idea that mere co-occurrence between speech and sounds leads to episodic traces of the sounds in memory. Taken together, the present results indicate that energetic masking may play a role in the emergence of a sound specificity effect. As such, they support the possibility that the acoustic glimpse of a word, rather than the co-occurring sound may be retained in memory and affect subsequent word recognition accuracy. However, they do not provide enough evidence to reject the alternative that the sounds are encoded in memory alongside the words. Another plausible explanation of the present results could be that increasing the difference between the items in exposure and test in Experiment 3 (compared to Experiment 2A and 2B), may have led to the emergence of the sound specificity effect. Further, although these experiments present a "glimpses scenario" as a potential explanation for a sound specificity effect, they do not directly test this scenario. In all the experiments, the sounds were still present, i.e., co-occurring with the words, albeit in different masking configurations. A more direct test of the presence of glimpses in memory would have been to have another experiment with only the glimpses of the words as stimuli, instead of word-sound pairs. Without such an experiment, the glimpse hypothesis remains relatively weak.

Speech-extrinsic specificity effects are a recent phenomenon in the indexical and spoken word recognition literature, with very few studies reporting them (Cooper et al., 2015; Creel et al., 2012; Pufahl & Samuel, 2014). This recent development has the potential to add new insights into the representational nature of lexical entries and the organization of the mental lexicon. However, consistent with other studies that investigated speech-extrinsic specificity effects (e.g., Cooper et al., 2015; Pufahl and Samuel, 2014), the findings discussed in this chapter suggest that these effects are fragile and highly sensitive to the context in which they are tested. In the rest of the thesis, my goal is to further explore contexts in which such effects can emerge. The experiments presented here established one such context: a "glimpses" scenario. In the next chapter, I continue this investigation by analyzing the analogy (or lack thereof) between voice specificity and sound specificity, focusing on the notion of integrality (Vitevitch, 2003), that is, the degree to which spoken words are necessarily undissociable from the accompanying voice as opposed to the accompanying sound.

Chapter 4

The Integrality Account

Abstract

The two experiments presented in this chapter investigate the role of a novel element in the relationship between words and sounds, namely, integrality, as a potential factor in the emergence of a sound specificity effect. Pufahl and Samuel (2014) posited that mere co-occurrence between words and background sounds was sufficient to reveal a sound specificity effect and that even auditory information that is non-integral to speech is retained in the lexicon alongside words. Keeping in focus a close analogy to the relationship between words and voices, I argue that co-occurrence per se may not be always sufficient to reliably reveal such an effect. The experiments described here identify another crucial factor that seems to play a role in the emergence of the effect, the integrality between the words and sounds. I define this new factor as the perceptual blending of the two signals into a relatively unified auditory entity. Like in the previous chapters, the present experiments involved the same design, with identical encoding and recognition memory tasks. The results revealed a sound specificity effect on both word recognition accuracy and response latency, but only when the sounds were rendered integral to the words by intensity envelope modulation (Exp. 4). Importantly, this specificity effect disappeared when the integrality factor was removed from the stimuli (Exp.5). Thus, listeners were less accurate and slower in recognising previously heard words when the sound changed from exposure to phase only when the sounds were difficult to segregate from the words. These results provide evidence in favour of an alternative explanation of speech-extrinsic specificity effects to that of mere co-occurrence. Further, while supporting an episodic view of the mental lexicon, they also draw attention to the vulnerability of such effects to the auditory context in which they emerge.

4.1. Introduction

There has been growing interest in the literature recently towards speech-extrinsic specificity effects that occur as a result of auditory variability external to the speech signal. Pufahl and Samuel (2014) were the first to report what we referred to as a “sound specificity effect”. The experiment of interest in the set of experiments they conducted was explained in detail in the previous chapter, hence only a brief outline is provided here, for the reader’s convenience. Namely, words spoken by either a male or a female talker were paired with one of two sound exemplars of a certain sound category, such that each sound exemplar was unique to a word. Besides the classical voice specificity effect, the authors also found a sound specificity effect, that was manifested in a similar decrease in the word identification accuracy when the paired sound exemplar changed from exposure to test. This novel effect was interpreted as evidence that listeners retain specific acoustic details of irrelevant background sounds that co-occur with spoken words in long term lexical memory. Therefore, the authors proposed an integrated view of the mental lexicon, in which lexical representations consist of a combination of lexical, indexical (voice), as well as speech-extrinsic auditory detail. The crucial aspect of their argument is that the *co-occurrence* between the words and sounds rather than any properties *integral* to the spoken word, seems to play a role in the emergence of a sound specificity effect.

The previous chapter started the investigation of the sound specificity effect from a masking perspective. However, it still revolved around the co-occurrence element between the words and the background sounds. Importantly, while providing a plausible context for the emergence of the sound specificity effect, the glimpse account also highlighted the fragility of this effect with respect to the context in which it occurs. It demonstrated that a sound specificity effect seems to emerge in some contexts, while failing to do so in others. A way of approaching the vulnerability of this speech-extrinsic specificity effect is by comparing it to its speech-intrinsic counterpart, namely the voice–specificity effect. The latter shows a relatively reliable emergence pattern. It has been replicated many times, using a variety of encoding and memory tasks. Therefore, an understanding of what makes voices special and why they persist in long-term memory more reliably than background sounds when the same experimental tasks are used, might contribute to a better understanding of the sound specificity's context-dependency.

Obviously, the auditory system is tuned towards voices, with them being so frequently encountered on a regular basis. Hence, voices have an inherent advantage from both an occurrence frequency and a pragmatic perspective. Successful communication in spoken language necessitates paying attention to the talker's voice, be it for identifying them, inferring the nature of their message/their emotional state, or adjusting one's own speech to accommodate understanding that of the talker. Thus, it is not surprising that voice-related details are retained in memory and used during spoken word processing.

On the other hand, background sounds that co-occur with spoken words usually do not serve any pragmatic purpose. The auditory system is used to treat them as noise and ignore them through selective attention in order to enhance speech intelligibility. Furthermore, if the experimental task at hand does not require paying attention to them, the chance that they are reliably retained in memory and not discarded seems slim. As our previous experiments show, mere co-occurrence with the words does not guarantee the sound's incorporation into long-term memory, since additional factors, such as spectrally and temporally contrasted acoustic glimpses, seem to play role in the emergence of an effect. It becomes interesting then to identify other factor(s) that might play an intrinsic role in the emergence of a more reliable and robust sound specificity effect. To examine this issue, we deemed it useful to keep a closer analogy with the voice effect and revisit the relationship between words and voices.

A crucial property of a spoken word is its integral nature as a stimulus, consisting of two components: a linguistic (the word) and an indexical one (the talker's voice). The linguistic component conveys prepositional information about objects and events in the world. Indexical information refers to acoustic correlates in the speech signal that provide information about the talker, including identity, age, gender, dialect, and emotional state (Pisoni, 1997; Vitevitch, 2003). Importantly, these two components necessarily co-exist at any time and are also integrated with one another to form a single entity (Vitevitch, 2003). In Gestalt terms, they share a “common fate”. As such, it is impossible to perceptually segregate them, rendering the task of ignoring of the talker’s voice by the listeners very difficult. Since the two components are perceived as one common entity, it makes sense for the voices to be encoded in memory with the words.

On the other hand, the word-sound pairs in our previous experiments, as well as in Pufahl and Samuel’s study, consisted of two dimensions that co-occurred, but the sounds were not integral to the words and segregating them from one another was relatively easy. Therefore, the listeners most probably perceived the pairs as two separate objects, rather than a single one, as in the case of spoken words. This perceptual segregation may be one of the crucial factors behind the fragility of the sound specificity effect. This realisation in turn evokes the question whether making the sounds more integrated to the words may play an intrinsic role in the emergence of a sound specificity effect. Specifically, we were interested to see whether emulating the relationship between words and voices more accurately would yield a sound specificity effect, possibly one that is more comparable to the voice specificity effect in terms of robustness. Can listeners’ retention of sound details in memory be promoted by making them integral to the words, like voices are integral to speech? We reasoned that if besides co-existing, the words and sounds are also integral to one another, such that it is difficult to perceptually segregate them, then they might be perceived as relatively unified objects, rather than two separate signals. This perceptual blending might in turn lead to a more robust sound specificity effect.

Given the strong reference to the voice effect and similar to the previous studies, the same number of background sounds as that of voices (Experiment 1) was used, namely two. The first experiment of the study (Exp.4) investigates the impact of rendering the sounds integral to the words, by modulating them according to each individual word’s intensity envelope, on the appearance of a sound specificity effect. By revealing such an effect in both the word recognition accuracy and response latency, it posits another plausible context, besides the glimpses one, for its emergence. This context combines two elements of the relationship between the words and sounds, integrality and co-occurrence, similar to the case of words and voices. The second experiment (Exp. 5) further confirms that integrality is indeed the crucial factor behind the observed sound specificity effect in the first experiment, because it demonstrates that removing the integrality element from the stimuli, leads to the disappearance of the effect. Furthermore, it also consolidates the claim that mere co-occurrence is not always sufficient for the appearance of this specificity effect. The next section provides a definition of the concept of integrality as used in the present study, and explains the rationale behind the method used to implement it in the stimuli of Experiment 4.

4.2. Integrality - Definition and implementation

The concept of integrality endorsed here refers to a degree of acoustical integration between the words and sounds, with the aim to make their segregation challenging and promote their perceptual blending into single entities. As mentioned, given the close comparison to the voice specificity effect and the intrinsic relationship between words and voices, the goal was to imitate this relationship as accurately as possible in the word-sound pairs. More specifically, we wanted the sounds to be bound to the words in such a way that every pair would be acoustically and perceptually blended into one unique item, similar to a spoken word. Nevertheless, we also wanted the sounds to retain their identity across the different pairings, like a voice preserving its identity across different utterances.

With these requirements in mind, we decided to implement the integrality element by modulation of the sounds according to the intensity envelope of each individual word. It is well-established in the literature that speech intelligibility strongly depends on the intensity fluctuations over time. While a detailed review of this literature is beyond the present scope, a brief overview of some major relevant arguments is provided for the reader's convenience. For instance, noise-vocoded speech is perfectly intelligible given that enough sub-band envelopes are used (Shannon et al., 1995). In their seminal study, Shannon et al. (1995) demonstrated that using only the speech envelopes and replacing the fine structure with noise yields perfect speech intelligibility, provided that at least 3 sub-bands are used. Also, several powerful speech intelligibility prediction models use only modulation information (e.g., Jørgensen & Dau, 2011; Jørgensen et al., 2013). Therefore, we chose the intensity envelope of the word(s) as the link between the words and the sounds. To preserve the identity of the sounds, we selected sounds whose identity is mainly conveyed by their temporal fine structure, rather than their intensity (amplitude) modulation. This quality makes them suitable candidates for amplitude modulation by another sound, the spoken word. The integral maskers were then created by preserving the fine structure of the sounds and replacing their intensity envelopes with those of the words. Such a method gives rise to "tailored" maskers that are fitted and integrated into each word uniquely, yet also retain their own identity as speech extrinsic sounds.

The next section describes Experiment 4, which investigated the role of integrality in the emergence of a sound specificity effect. We anticipated to see such an effect would manifest itself in the overall word recognition accuracy, such that the recognition accuracy for the words repeated with the same integral sound as in exposure would be higher than the accuracy for the words repeated with the different sound.

4.3. Experiment 4 - Integral maskers

4.3.1. Method

4.3.1.1 Participants

Fifty-four undergraduate students at the University of York (Age range: 18 - 27) participated in exchange for either course credit or payment. The number of participants in the two experiments presented in this chapter was informed by the indexical literature and the very limited number of studies on speech-extrinsic specificity effects. Namely, Cooper et al (2015) tested 44 (36 included in the analysis) and 39 (36 included) participants per condition in their first experiment, and 45 (36 included) and 42 (36 included) participants per condition in their second experiment. Pufahl and Samuel (2014) tested the following number of participants in their experiments: 72 (Exp.1, 64 included), 73 (Exp.2, 64 included), 65 (Exp.3, 64 included), 52 (Exp.4, 48 included), 23 blind adults (Exp.5, 19 included), and 51 (Exp.6) participants.[34] In line with our previous experiments, a relatively large sample size (N > 40) was targeted, given the reported fragility and the small size of the effects in question. All participants provided written consent prior to the experiment. They all iden-

[34] The relevant experiment for the present purpose in Pufahl and Samuel (2014) is Exp.1, the one that reported the sound specificity effect.

tified themselves as native-speakers of British English and none of them reported a history of hearing or speech and language related problems. None of the participants had done any of the previous experiments.

4.3.1.2. Materials and Design

The stimuli set consisted of 80 word-sound pairs. The words were the same ones as in the previous experiments. Like in the case of two talkers (Exp. 1) and as in the other previous experiments, two sounds were used and each word was paired with either one of them. The sounds were rendered integral to speech via intensity modulation along the envelope of each individual word. Hence, although the same two maskers were used, the modulation process led to relatively unique pairs. In both the present and the next experiment, all the stimuli files were generated with a sampling rate of 44.1 KHz and a resolution of 16 bit. Every stage of the stimuli preparation process was implemented using the Matlab software (version R2014b). The sections below explain in detail the selection criteria for the sounds, as well as the implementation of the intensity modulation process and the creation of the final stimuli.

I. Sound Selection for the Integral Maskers

Two environmental sounds were chosen as maskers. The main criteria for the sound selection were that: i) the sounds were continuous (i.e., not with an intermittent structure), non-fluctuating over time and ii) their identity was conveyed mainly by their pitch and timbre information (related to the temporal fine structure of the sound). Hence, their overall intensity envelope should not be important for their identity. To this end, a cat sound and a violin sound (playing one sustained tone) were selected as the best candidates. The temporal waveform of the sounds and their spectrograms are depicted in Figures 4.1 and 4.2, respectively.

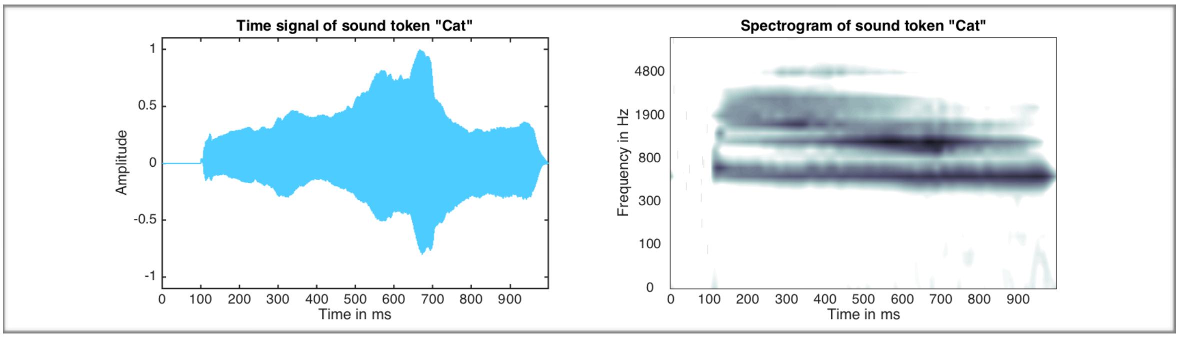

Figure 4.1: Time signal (left) and spectrogram (right) of the cat sound.

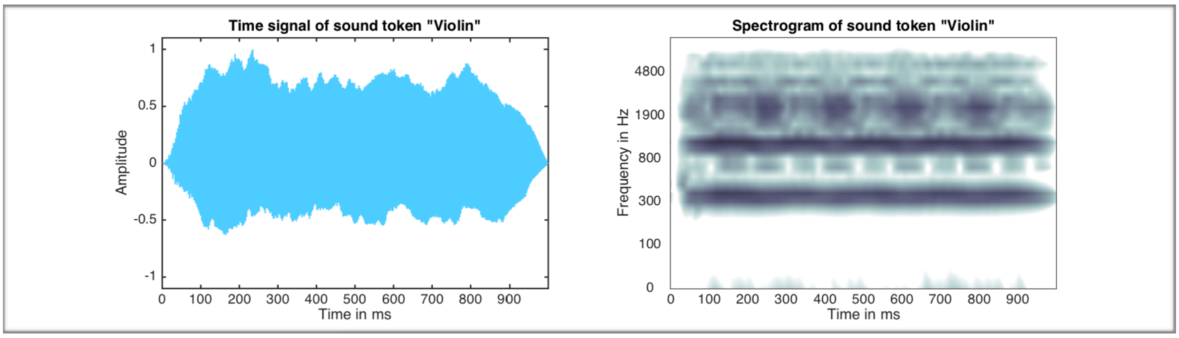

Figure 4.2: Time signal (left) and spectrogram (right) of the violin sound.

Acoustic analyses performed on the cat and violin sounds revealed that the mean difference in fundamental frequencies (F0s) between the cat and the violin sounds was 203.3 Hz (Cat_{F0} = 551.94 Hz, $\text{Violin}_{\text{F0}}$ = 348.64 Hz). The pitch contours (fundamental frequency over time) for both sounds are displayed in Figure 4.3.

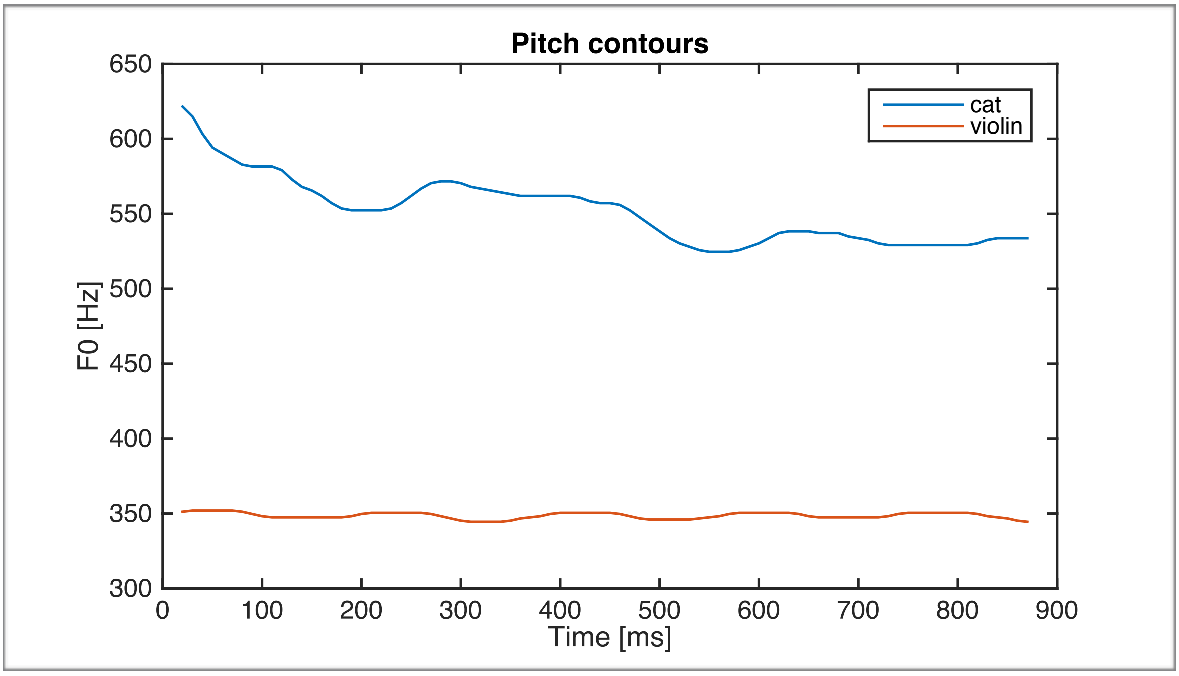

Figure 4.3: Pitch contours of the cat (blue) and violin (red) sounds.

II. Preparation of the Integral Maskers

To generate a certain level of perceptual integrality between the two chosen sounds and each of the spoken words, the sounds' intensity envelopes were shaped according the intensity envelopes of the individual words. The intensity envelopes were extracted by filtering the words to the frequency band between 0.3 and 6 kHz, extracting their Hilbert envelopes, and low-pass filtering the envelopes with a third-order low-pass filter at a cut-off frequency of 30 Hz. To generate the cat and violin maskers for a given word, the sounds were limited to the same frequency band (0.3-6 kHz) and then either lengthened by adding silence at the end or shortened by cropping the end to match the duration of the speech token and its intensity envelope. The sounds were then multiplied by the intensity envelope, such that they followed the intensity envelope of the word, which defines the "rhythm" of the token. It is important to note that the integral maskers created in this way were specific and unique to each word, since their resulting envelope followed the envelope of the individual words they were later mixed with. However, the integral maskers did not contain any intelligible/identifiable speech information but rather sounded like amplitude-modulated versions of the original sounds (with the type of amplitude modulation determined by the word's intensity envelope). Examples of the processing scheme for the two sounds, in their envelope-shaped versions and then mixed with a word ("tiger"), are shown in the upper three panels of Figures 4.4 and 4.5, respectively.

III. Mixing Words and Maskers

Each word was mixed with the corresponding two integral maskers (obtained from the cat and violin sounds) to obtain the final experimental stimuli. For the majority of the stimuli, the signal-to-noise ratio(SNR) used was -3dB, but other SNR values were also used where deemed necessary for maximum intelligibility of the individual mixtures. The additional SNR values were: -1, 0, +1 and +3 dB. The SNR values were piloted prior to the experiment, and the ones that yielded the maximum word identification accuracy (100 % correct) were used. Examples for the final stimuli can be seen in the bottom panel of Figures 4.4 and 4.5, respectively.

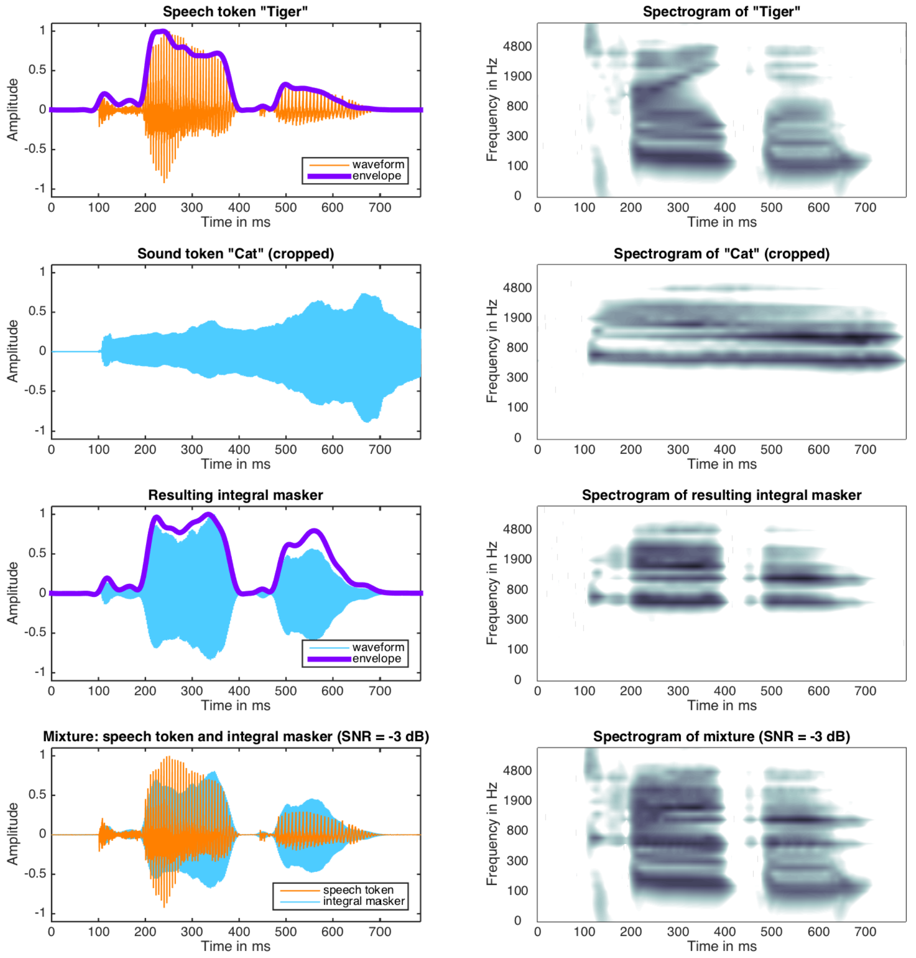

Figure 4.4: Processing scheme for generating the integral maskers applied to speech token “Tiger” and sound token “Cat”. Left panel, from top to bottom: speech token “Tiger” (orange) and its envelope (purple); sound token “Cat” cropped to length of speech token; integral masker and its envelope; mixture of speech token (orange) and integral masker. Right panel: corresponding spectrograms.

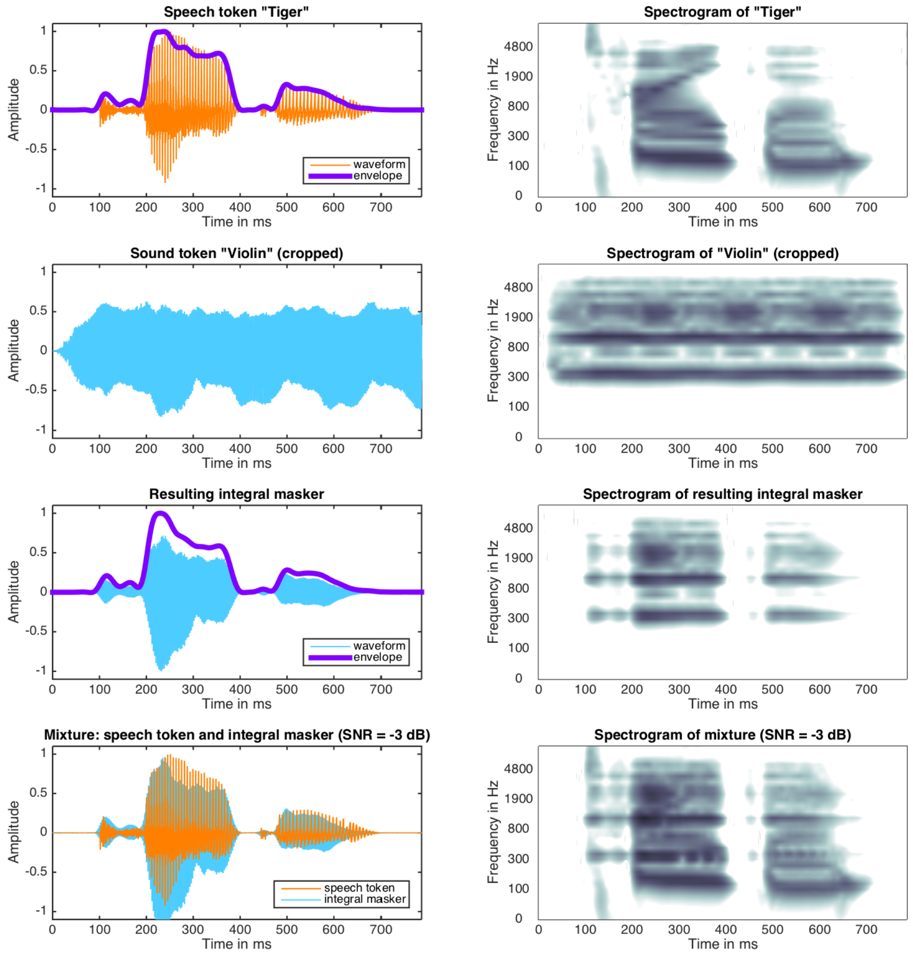

Figure 4.5: Processing scheme for generating the integral maskers applied to speech token "Tiger" and sound token "Violin". Left panel, from top to bottom: speech token "Tiger" (orange) and its envelope (purple); sound token "Violin" cropped to length of speech token; integral masker and its envelope; mixture of speech token (orange) and integral masker. Right panel: corresponding spectrograms.

The same paradigm as in the previous experiments was used in both Experiment 4 and 5, consisting of two phases: Exposure and Test, and a short delay in between. In each phase, participants heard a block of 60 trials, each played one at a time. None of the trials were repeated within a block. Half the words in each block were paired with the cat sound and the other half with the violin sound. While the words in the exposure trials (Block 1) were the same for all participants, what sound they were paired with was counterbalanced across participants. In the test trials (Block 2), 40 of the 60 words were repeated from the exposure phase and constituted the "OLD" words. Half of the "OLD" were paired with the same sound as in exposure, and the other half with the different sound. Which words in the test phase were paired with the same or the different sound was counterbalanced across participants. The counterbalancing in terms of the paired sound (cat or violin) and sound sameness (same or different from Exposure to Test) resulted in 4 stimulus lists (counterbalancing groups) in total. Each participant was randomly assigned to either one of them. The words in the remaining 20 trials in Block 2 had not been heard in the exposure phase (Block 1). Hence, these were the same for all participants, with half of them paired with the cat sound and half with the violin sound.

4.3.1.4. Procedure

Exposure phase

The experiment was run on the DMDX software (Forster & Forster, 2003). Participants sat individually in a sound-attenuated booth and listened to the trials played binaurally over headphones (Sony MDR-V700) at a comfortable listening level of approximately 68 dB SPL. They read instructions on the computer screen and also listened to the experimenter's explanations. They were instructed to make an "animate/inanimate" decision for the word in each trial and ignore the background sound. The 'animate' and 'inanimate' concepts were defined and examples for each of the categories were provided (e.g., "banana is inanimate", "professor is animate"). The experimenter encouraged them to be as accurate as possible. After 500 milliseconds, a message was displayed on the screen prompting them to respond by pressing either one of the corresponding 'shift' keys on the computer keyboard. Namely, on the right side of the screen the word 'ANIMATE' (referring them to the right 'shift' key), and on the left side the word 'INANIMATE' (referring them to the left 'shift' key) appeared. Participants were told to wait for the message to appear on the screen before responding and were allowed a maximum of 10 seconds to submit a response. The next trial followed immediately after they hit a response button, or after the maximum allowed time expired, if no response was provided. Prior to the experimental trials, participants completed 4 practice trials, where different words from the experimental ones were spoken by a male talker, with half of them were paired with the high-pitch car horn sound and the other half with the low-pitch one. There were 60 experimental trials (Block 1) in total and their order was randomized for each participant. No feedback was provided after each trial and there was no mention of an upcoming recognition task. The task lasted approximately 10 minutes.

Delay

After completing the first experimental phase, participants left the sound-attenuated booth and spent 5-7 minutes on an unrelated distractor task prior to the memory test. This was done in order to ensure that performance in the subsequent test phase was not based on short-term or working memory. The task consisted of playing an online game (Cube Crash 2).

Test phase

In order to assess the effect of sound change on recognition memory, participants completed a surprise word recognition task. They read written instructions on the screen and listened to the experimenter's explanations. The experimenter explained that they would again hear word-sound pairs, in which some of the words had already been heard in the first part of the experiment (old), and the rest would be heard for the first time (new). They were also told that the background sounds were the same ones as in the first part of the experiment, but that they should again ignore them, as they were not relevant for the task at hand. The experiment explained that for every trial, their task was to decide whether the word was "old" or "new", by pressing the respective 'shift' key on the keyboard. They were encouraged to be as accurate as possible, but to also hit the response key as soon as they made their decision. Participants first saw an 'x' symbol appearing at the centre of the screen, which anticipated the coming of a word. After 500 milliseconds, they heard the word and responded by pressing either one of the 'shift' keys on the computer keyboard (right for the "old" words and left for the "new" words). Labels written with either 'OLD' or 'NEW' were put above the corresponding shift keys for participants' convenience. The next trial followed immediately after participants response, or after the maximum allocated time of 10 seconds expired and no response was provided. There were 60 experimental trials in total: 20 old words-same sound (old-same), 20 old words-different sound (old-different) and 20 new words. The order of trials was randomized for each participant. The task lasted approximately 10-15 minutes.

4.3.2. Results

All participants displayed very high mean accuracies of above 90% correct in the exposure phase, indicating that they had successfully encoded the words during the task. A one-way repeated measures ANOVA revealed no difference in performance in the Exposure phase with respect to the semantic category of the words:
$M_{animate} = 98.26$ % correct, $SD_{animate} = 2.48$; $M_{inanimate} = 98.75$ % correct, $SD_{inanimate} = 2.44$; F(1,47) = .68, $\eta^2 = .01$, p =.41.

Six participants were excluded from analysis for the following reasons: 1) three participants due to technical failure of the experimental software, 2) two participants judged the sounds, instead of the words in the exposure phase, and 3) one participant judged all the "inanimate" words in the exposure phase incorrectly. Overall, 48 participants were included in the final analysis, which consisted of only the critical (o*ld*) trials. The response times were measured from the onset of the stimulus to the onset of the button press. Only the latencies of correct responses were submitted for analysis and the ones that were 2*SD above the mean on a subject-by-subject basis were omitted. The two dependent variables were Accuracy (recognition accuracy) and Response Time (RT). Accuracy was coded as a binary variable with values '0' and '1 per trial basis, where '1' meant a correct response to a trial, and '0' an incorrect one. Like in the previous experiments, the data were analyzed using linear (LMER) mixed-effects regression models for the continuous RT variable, and generalized mixed-effects regression models (GLMER) for the binary Accuracy variable (Baayen, Davidson, & Bates, 2008).

The fixed factors were the same ones as in the previous experiments, namely: Sound Sameness (same or different), Semantics (animate or inanimate word), and Exposure Sound (cat or violin). The factors were coded as binary variables as follows: Sound Sameness: 1 (same), -1 (different); Semantics: 1 (animate), -1 (inanimate); Exposure Sound: 1 (violin sound), -1 (cat sound). Prior to adding any fixed factors to the base model, we tested the maximal random structure of the model for each dependent variable, consisting of random slopes of all the fixed factors and random intercepts for subjects and items. For the main factor of interest, the Sound Sameness, random slopes were added for both subjects and items, whereas for the other two factors, only by-subjects random slopes were added. For the Accuracy variable, the maximal random structure converged, but it was not statistically different from the structure consisting of only random intercepts, $\chi2(4) = 5.03$, p = .28. For the RT variable, the maximal random structure also converged, but it was not statistically different from the base random structure consisting of only random intercepts for subjects and items, $\chi2(4) = 0.60$, p = .96. Nevertheless, based on Barr et al (2013)'s analysis and suggestions, we used the maximal random structure whenever the respective maximal model(s) with the added fixed factors converged.

For every dependent variable, the fixed factors, as well as their interactions, were added incrementally to the base model, and improved fit to the model was assessed using the likelihood ratio test. The base model included only the random terms. The main effects of Sound Sameness, Semantics, and Exposure Sound were obtained by testing the improvement in the model fit when each one of these factors was individually added to the base model.

As expected, there was a main effect of the sound change from exposure to test on Accuracy, $M_{Acc\_Same\ Sound} =$ 80.42 % correct, SD = 12.54; $M_{Acc\_Different\ Sound} =$ 76.04 % correct, SD = 9.56; β = .14, SE = .06 , χ2(1) = 5.95, p = .01, meaning that overall, participants were more accurate in recognizing previously heard words that were repeated with the same integral sound as in exposure, compared to the words that were repeated with the different sound.[35]

Interestingly, the sound specificity effect was also present in the response latency, $M_{RT\_Same\ Sound} =$ 1424.98 ms., SD = 229.68; $M_{RT\_Different\ Sound} =$ 1470.67 ms., SD = 264.06; β = -19.62, SE = 8.25, χ2(1) = 5.42, p = .02. Hence, there was a small, but significant increase in the mean RT as a result of the sound change from exposure to test. Participants were faster in recognising previously heard words that were repeated with the same integral sound as in exposure, compared to the words that were repeated with the different sound. Contrasted with the voice specificity effect (Experiment 1, **Chapter 2**), that emerged only for recognition accuracy, in the present case, the sound specificity effect is manifested in both variables of interest.

The mean false alarm rate on the new words was: M_{FA} = 19.58 % , SD = 11.34, with 17.08 % , SD = 13.52 for the words accompanied with a violin sound, and 22.08 %, SD = 12.37 for the words accompanied with a cat sound. This difference was, F(1,47) = 7.62, p = .008, η^2 = .14. Since these trials were not critical in our design and were not eligible for analysis with respect to the sound specificity effect, this bias will not be discussed further.

There was a main effect of semantic category (Semantics) on both Accuracy: β = .28, SE = . 1 , χ2(1) = 7.39, p = .007, and RT: β = -39.38, SE = 13.82, χ2(1) = 7.45, p = .006. Overall participants were better and faster at recognizing previously heard words when they were animate compared to when they were inanimate, $M_{Acc\_Animate} =$ 82.60 % correct, $M_{Acc\_Inanimate} =$ 73.85 % correct; $M_{RT\_Animate} =$ 1414.65 ms., $M_{RT\_Inanimate} =$ 1484.11 ms. However, there was no interaction between the semantic category (Semantics) and the sound specificity effect (Sound Sameness) on either Accuracy, β = .09, SE = .06, χ2(1) = 2.2, p = .14; or RT, β = 6.43, SE = .8.28, χ2(1) = .6, p = .44.[36] Hence, participants were more accurate and faster at recognising animate words compared to inanimate ones, regardless of whether they were repeated with the same background sound or not.

There was no main effect of the Exposure Sound (cat vs. violin) on either Accuracy, β = -0.02 , SE = .06, χ2(1) = .13, p = .71; or RT, β = .03, SE = 8.61, χ2(1) = 0, p = 1. Further, there was

[35] Random slopes of the Sound Sameness and Exposure Sound were added only for subjects, as the maximal models with all the random slopes, added to both subjects and items, did not converge.

[36] For the Accuracy variable, the maximal models did not converge, hence only random intercepts were included.

no interaction between the sound specificity effect and the exposure sound on either Accuracy, $\beta = -0.04$, SE = .06, $\chi2(1) = .57$, p = .45; or RT, $\beta = 4.05$, SE = 8.47, $\chi2(1) = .23$, p = .63. Therefore, the sound with which the words were first heard during exposure did not affect either the recognition memory performance of participants at test, or the sound specificity effect.

Additionally, F_1 and F_2 analyses were conducted to further confirm that the *sound specificity effect* found on Accuracy and RT was genuine, hence present across both subjects and items.

Accuracy:

- *By subjects*: A repeated measures ANOVA with Sound Sameness as the within subjects factor revealed a main effect of the sound sameness across subjects, $F_1(1,47) = 8.67$, p = .005, η^2 = . 16.
- *By items*: A repeated measures ANOVA with Sound Sameness as the within items factor revealed a main effect of the sound sameness across the items, $F_2(1,39) = 4.35$, p = .04, η^2 = .10.

Response Time (RT)

- *By subjects*: A repeated measures ANOVA with Sound Sameness as the within subjects factor revealed a main effect of the sound sameness across subjects, $F_1(1,47) = 8.41$, p = .006, η^2 = . 15.
- *By items*: A repeated measures ANOVA with Sound Sameness as the within items factor revealed a main effect of the sound sameness across the items, $F_2(1,39) = 6.59$, p = .01, η^2 = .15.

4.3.3. Discussion

The experiment described above investigated the role of a novel dimension in the relationship between words and sounds, integrality, on the emergence of a sound specificity effect. The rationale behind it was inspired by a close analogy to the case of the voice specificity effect. The integrality element was implemented by modulating the sounds according to the intensity envelope of each word, such that their own envelopes were replaced by those of the words, while their fine, spectral structure was kept intact. The aim was to render the perceptual segregation between words and sounds difficult, as well as to mirror the relationship between words and voices as closely as possible. As expected, the mixed effects regression analysis revealed a sound specificity effect for word recognition accuracy, such that the words repeated with the same paired sound as in exposure were recognised more accurately than the words repeated with the different sound. Interestingly, there was also a main effect of the sound change on the overall response latency, such that listeners were faster in recognising the words repeated with the same paired sound as in exposure, compared to the words repeated with the different sound. The F_1 and F_2 analyses showed that the effect was present across subjects and items for both variables of interest, hence further conforming its genuineness. Given these positive results, it is tempting to postulate that the integrality element added to the word-sound pairs is the factor behind the emergence of the sound specificity effect. However, a strong argument about this claim is not possible, unless the alternative explanation for the appearance of the effect has been ruled out. As mentioned, although they were integral to the words, the sounds retained their identity throughout the different pairings. It was relatively easy to identify them, and a cat sound is clearly quite different from a violin one. Therefore, it might be the case that the sound specificity effect observed is not due to integrality per se, but rather a result of the acoustical and semantic difference between the two sounds. To address this possibility, we ran a modified version of the previous experiment, in which the integrality element was removed from the stimuli.

The same experimental design, tasks and word-sound pairs were used in experiment 5, but the sounds were not modulated by the intensity envelope of the words. Hence, the words and sounds only co-occurred, without being integral to one another. This meant that the masking sounds where easily segregable from the words and also not unique to each one of them, like in the case of the integral sounds in Experiment 4. If the acoustical and/or semantic difference between the cat and the violin sounds plays any role in the emergence of a sound specificity effect, then we should observe such an effect in either of the dependent variables, or both. Namely, listeners should be more accurate and/or faster in recognising the "old" words that are repeated with the same sound as in exposure, compare to the "old" words repeated with the different sound. On the other hand, if integrality is indeed the crucial factor in the appearance of the sound specificity effect, then we should not be able to observe a specificity effect in either of the variables in Experiment 5, where integrality is eliminated from the stimuli.

4.4. Experiment 5 - Non-integral maskers

4.4.1. Method

4.4.1.1 Participants

Forty-six undergraduate students at the University of York (age range: 18 - 23 years) participated in exchange for either course credit or payment. All participants provided written consent prior to the experiment. They all identified themselves as native-speakers of British English and none of them reported a history of hearing or speech and language related problems. None of the participants had done Experiment 4 and any of the previous experiments.

4.4.1.2. Materials and Design

The same two sounds (cat and violin) as in Experiment 4 were used, but the envelope shaping described in Experiment 4 was omitted, such that the masking sounds were not integral to the words. Since all the word tokens started with a 100-ms silence phase and the violin sound started almost immediately (cf. Figure 4.2 above), the violin sound was delayed such that it started together with the word tokens (compare Figures 4.2 and 4.7). The different start times of word tokens and violin sound were not an issue in the integral masker experiment as the sound was shaped by the word's envelope and therefore the integral masker could not start earlier than the word. As it can be observed in the spectrogram of the violin sound (right panel of Figure 4.7), there is virtually no variation over time. Therefore, the time shift in the violin sound can be assumed to have had no relevant effect in terms of masking. The time shift was introduced to prevent the sound from dominating the percept due to its earlier start as compared to the word token. This early perceptual domination was not the case in Experiment 4, hence it was avoided in the present experiment as well.

The cat sound (cf. Figure 4.4) started at 100 ms (as the speech tokens) and was therefore not changed. In order to ensure a fair comparison across experiments in terms of the spectral content, the sounds were filtered to the same band (0.3 – 6 kHz) they had been filtered to when used as integral maskers. Then, 100 ms of silence was appended to the sound tokens (the word tokens had 100 ms of silence at the end). The word tokens were mixed with the sound tokens at -3dB SNR for the majority of the stimuli. However, other SNR values: -1, 0, +1 and +3 dB, were also used in certain cases to ensure maximum intelligibility of the respective mixtures. The SNR values were piloted prior to the experiment, and the ones that yielded maximum word identification intelligibility (100 % correct) were used. The SNR specifies the ratio of the word token's level and the sound token's level (or the difference between the levels given in decibels (dB)). In our case the levels were defined as the levels of the energy-containing portions of the sound files, such that appended silence did not affect the SNR calculation. The longer sound file (word or sound) determined the overall duration of the mix, such that neither the word nor sound token was cropped. As the sound tokens were longer than the word tokens (also illustrated by the examples in Figures 4.8 and 4.9), the effective duration of the mix was the duration of the sound (including the appended silence).

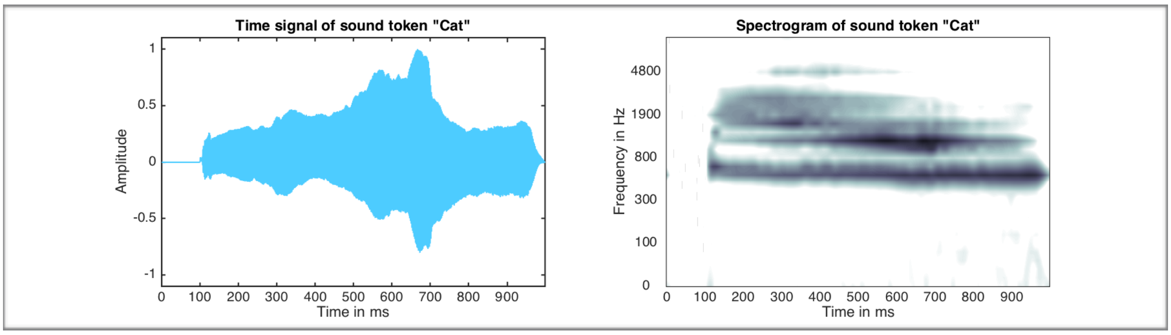

Figure 4.6: Time signal (left) and spectrogram (right) of cat sound (same as in Figure 1).

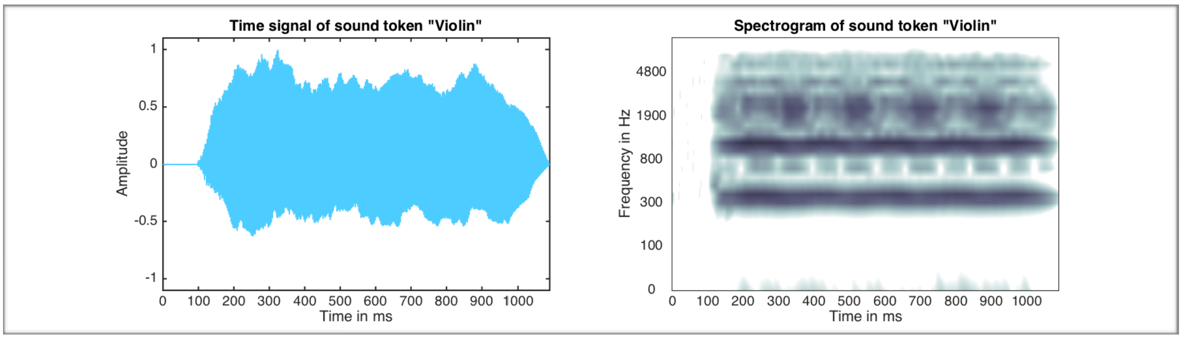

Figure 4.7: Time signal (left) and spectrogram (right) of violin sound. The only difference to Figure 2 is that the onset is shifted such that the sound starts at 100 ms.

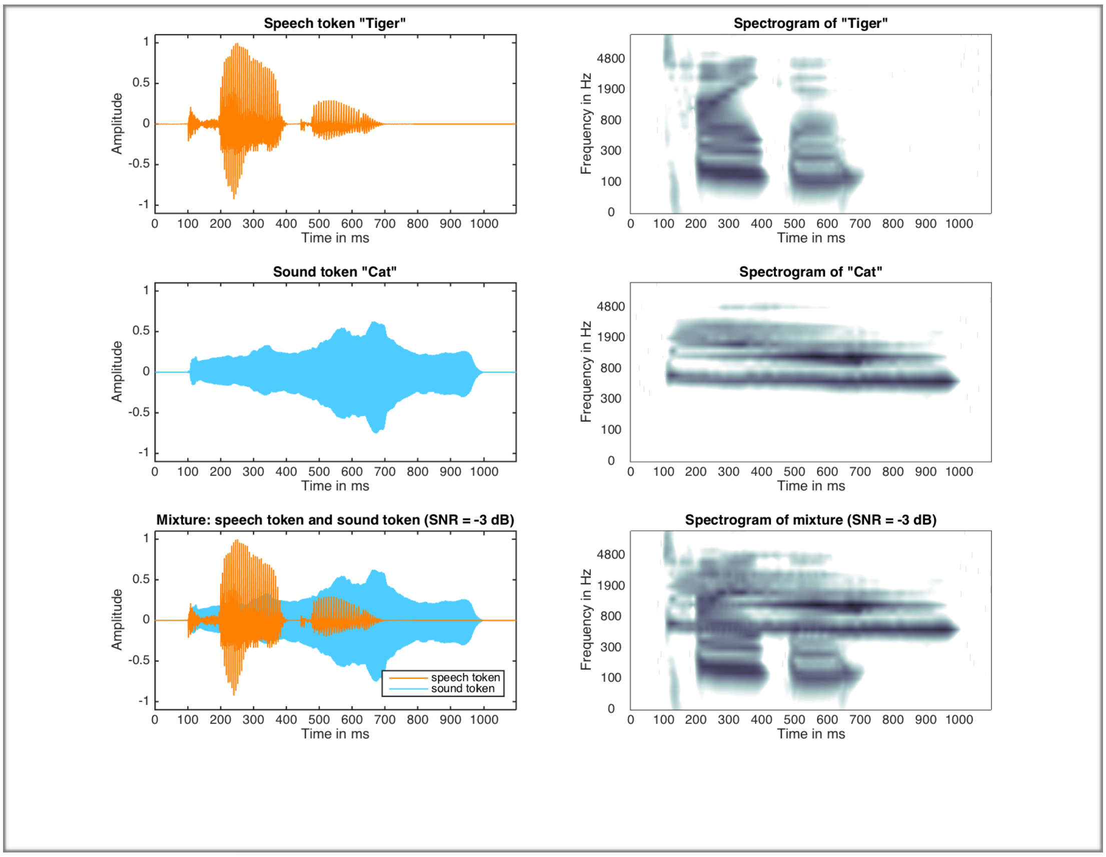

Figure 4.8: Processing scheme for mixing speech tokens and non-integral maskers applied to the word "Tiger" and the sound "Cat". Left panel, from top to bottom: word "Tiger" (orange), sound "Cat" (light blue) with 100 ms silence appended, the corresponding mixture. Right panel: corresponding spectrograms.

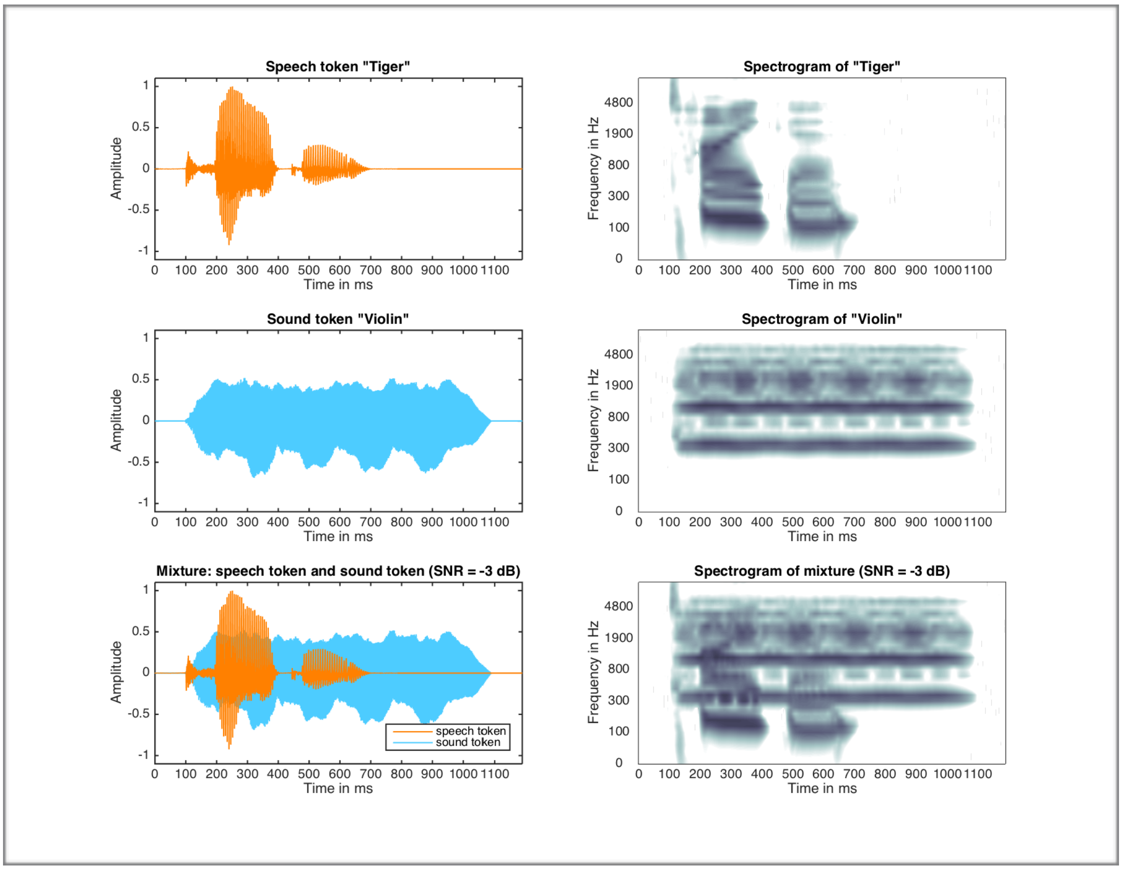

Figure 4.9: Processing scheme for mixing speech tokens and non-integral maskers applied to the word "Tiger" and the sound "Violin". Left panel, from top to bottom: word "Tiger" (orange), sound "Violin" (light blue) with 100 ms silence appended, the corresponding mixture. Right panel: corresponding spectrograms.

The experimental design was identical to the one in Experiment 4, involving the same counterbalancing groups and experimental phases.

4.4.1.3. Procedure

This was the same as in Experiment 4.

4.4.2. Results

All participants displayed very high mean accuracies of above 90% correct in the Exposure phase, indicating that they had successfully encoded the words during the task. A one-way repeated measures ANOVA revealed a very small, yet significant difference in performance with respect to the semantic category of the words:

$M_{Animate}$ = 99.20 % correct, $SD_{Animate}$ = 2; $M_{Inanimate}$ = 99.93 % correct, $SD_{Inanimate}$ = .5; F(1,45) = 5.28, η^2 = .11, p =. 026.

Like in Experiment 4, the response latencies analysed in the Test phase were measured from the onset of the stimulus to the onset of the button press. Only the latencies of correct responses were submitted for analysis and the ones that were 2*SD above the mean on a subject-by-subject basis were omitted. The data were analysed in the same way as in Experiment 4. The same two dependent variables, Accuracy (recognition accuracy) and Response Time (RT), as well as the same fixed factors were involved and coded in an identical fashion.

Prior to adding any fixed factors to the base model, the maximal random structure of the model was tested for each dependent variable. For the main factor of interest, the Sound Sameness, random slopes were added for both subjects and items, whereas for the other two factors, only by-subjects random slopes were added. For the Accuracy variable, the maximal random structure converged, but it was not statistically different from the structure consisting of only the random intercepts for subjects and items, χ2(4) = 7.57, p = .11. Similarly, with respect to the RT variable, the maximal random structure also converged, but it was not statistically different from the base random structure consisting of only the random intercepts, χ2(4) = .67, p = .95. Nevertheless, based on Barr et al (2013)'s analysis and suggestions, we used the maximal random structure whenever the respective maximal model(s) with the added fixed factors converged.

For every dependent variable, the fixed factors, as well as their interactions, were added incrementally to the base model, and improved fit to the model was assessed using the likelihood ratio test. The base model included only the random terms. The main effects of Sound Sameness, Semantics, and Exposure Sound were obtained by testing the improvement in the model fit when each one of these factors was individually added to the base model.

With respect to a sound specificity effect on recognition accuracy (Accuracy), there was no significant decrease in the mean accuracy as a result of the sound change from exposure to test, $M_{Acc\_Same\ Sound}$ = 77.07 % correct, SD = 13.64; $M_{Acc\_Different\ Sound}$ = 76.74 % correct, SD = 10.39; β = .007, SE = .07, χ2(1) = .009, p = .92.[37] Overall, participants were not more accurate in recognizing previously heard words when they were repeated with the same non-integral sound as in exposure, compared to when the sound was different.

[37] For the Accuracy variable, random slopes of only the Sound Sameness factor were added to both subjects and items, as the maximal model containing the slopes of the other factors as well, did not converge.

There was also no main effect of the sound change on response latency (RT), $M_{RT\_Same\ Sound}$ = 1507.31 ms., SD = ; $M_{RT\_Different\ Sound}$ = 1529 ms., SD = ; β = -10.75, SE = 10.84, χ2(1) = .98, p = .32. Hence, participants were not faster in recognising previously heard words when they were repeated with the same non-integral sound as in exposure, compared to when the sound was different.

The mean false alarm rate on the new words was: M_{FA} = 20.22 % , SD = 10.95, with 15.87 %, SD = 11.85 for the words accompanied with a violin sound, and 24.57 %, SD = 14.71 for the words accompanied with a cat sound. This difference was, F(1,47) = 7.62, p = .008, η^2 = .14. Since these trials were not critical in our design and were not eligible for analysis with respect to the sound specificity effect, this bias will not be discussed further.

There was a main effect of semantic category (Semantics) on Accuracy: β = .27, SE = .12 , χ2(1) = 4.99, p = .03, but not on RT: β = -24.15 , SE = 12.99, χ2(1) = 3.29, p = .07. Overall, participants were more accurate in recognising inanimate words and faster with animate words, although not significantly so: $M_{Acc\_Animate}$ = 77.61 % correct, $M_{Acc\_Inanimate}$ = 78.12 % correct; $M_{RT\_Animate}$ = 1512.83 ms., $M_{RT\_Inanimate}$ = 1531.73 ms. Additionally, there was no interaction between the semantic category and sound specificity effect on either Accuracy, β = -0.03, SE = .06, χ2(1) = .19, p = .67; or RT, β = 5.09, SE = 10.87, χ2(1) = .22, p = .64.[38] Participants were more accurate at recognising inanimate words compared to animate ones, regardless of whether they were repeated with the same background sound or not.

There was no main effect of the Exposure Sound (cat vs. violin) on either Accuracy, β = .03 , SE = .06, χ2(1) = .18, p = .67; or RT, β = -6.75, SE = 10.41, χ2(1) = 0.42, p = .52. Further, there was no interaction between the sound specificity effect and the exposure sound on either Accuracy, β = -0.04, SE = .06, χ2(1) = .37, p = .54; or RT, β = -11.24 , SE = 10.25, χ2(1) = 1.2, p = .27. Therefore, the sound with which the words were first heard during exposure did not affect either the recognition memory performance of participants at test, or the sound specificity effect.

4.4.3. Discussion

The experiment above examined the emergence of a sound specificity effect in the presence of the same two background sounds used in Experiment 4, but with the integrality component removed from the stimuli. The aim was to decouple the two alternative explanations for the appearance of the sound specificity effect in Experiment 4: integrality vs. acoustical/semantic difference between the two sounds. We wanted to see whether the sound specificity effect found in Experiment 4 would persist in the absence of integrality between the words and sounds. To this end, the sounds were not modulated by the intensity envelope of the words, hence they only co-occurred with the words when paired, without being integral. As expected, the data revealed no main effect of the change in the paired sound from exposure to test on either the overall word recognition accuracy, or response latency. Therefore, the sound specificity effect found in Experiment 4 disappeared

[38] For the Accuracy variable, the maximal models did not converge, hence only random intercepts were included.

with the removal of the integrality element from the stimuli. This result consolidates the argument that the integrality between the words and sounds seems to play a critical role in the emergence of a relatively robust sound specificity effect, which in our study, was present in both variables of interest.

However, the absence of a sound specificity effect in Experiment 5 could also be explained by a glimpses account. The previous chapter demonstrated that highly contrasted glimpses of the same word(s), resulting from the masking of two different sounds, can lead to the emergence of a sound specificity effect. In the present case, it is possible that there was a greater difference between the acoustic glimpses resulting from the two sounds in Experiment 4 than in Experiment 5. If so, a glimpse contrast, rather than an integrality contrast would be responsible for the different outcomes in Experiments 4 and 5. To decouple these possibilities, we computed the acoustic glimpses of the stimuli in both experiments and conducted a comparative statistical analysis.

4.5. Glimpse computation and analysis

We obtained quantitative measures of the acoustic glimpses resulting from the masking of the two sounds for all the stimuli used in the two experiments described above. The glimpses were computed by using a glimpse detection model (Cooke, 2003; 2006), which was described in the previous chapter. We ran statistical tests on the calculated glimpses for each experiment, in order to assess the difference in the glimpses corresponding to the two masking sounds. Only the critical trials (i.e., those comprising the *old* words) were analysed.

Experiment 4

The mean glimpse percentages for each sound were: $M_{Glimpses\_Violin} = 34.09\ \%$ SD = 4.93; $M_{Glimpses\_Cat} = 46.10\ \%$, SD = 4.61. Hence, on average, there were more glimpses resulting from the cat sound compared to the violin sound. This means that the violin sound seems to be more masking than the cat sound, which considering their spectrograms (Figures 4.1 and 4.2), is not surprising. A repeated measures ANOVA with Masker (2 levels: cat vs. violin) as the within-items factor revealed a robust main effect of the masker difference on the glimpse value, F(1,39) = 483.33 , $p < .0001$, $\eta^2 = .93$. Thus, the glimpses of the same word(s) resulting from the two masking sounds were different. Examples of a word paired with each of the two maskers and the respective glimpses are provided in Figure 4.10.

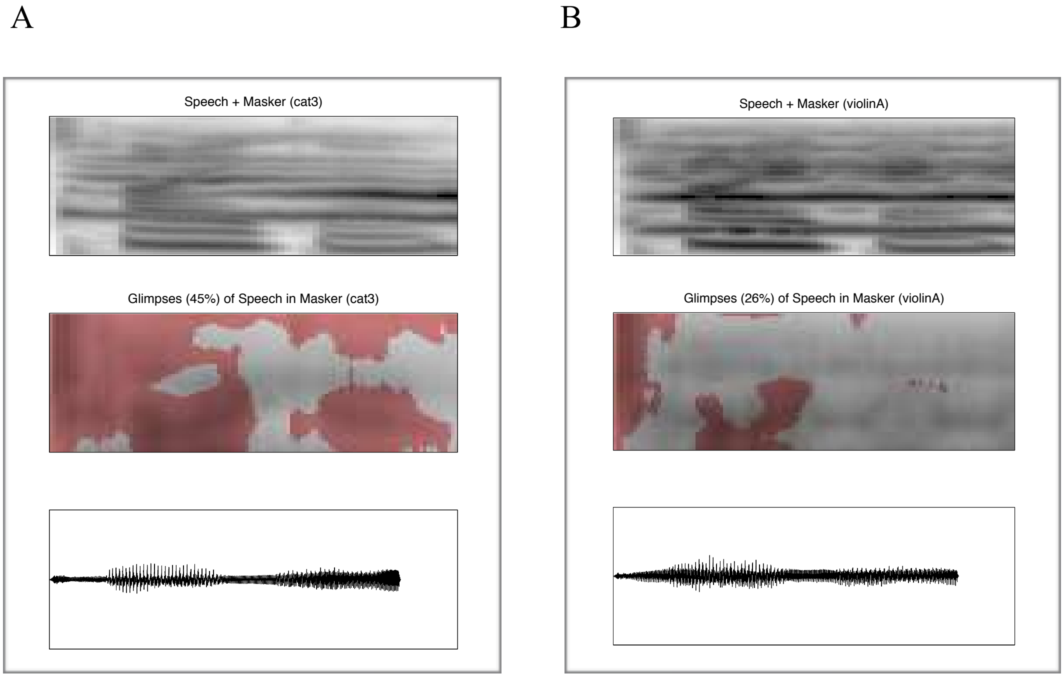

Figure 4.10. Spectrograms of the mixtures and the resulting glimpses of the same word paired with the two sounds in experiment 4. The masked regions are shown in black and the glimpses in red. The image in A depicts the case when the word is paired with the cat sound and the image in B the case when the same word is paired with the violin sound.

Experiment 5

The mean glimpse percentages for each car horn sound were: $M_{Glimpses\_Violin}$ = 22.40 % SD = 3.84; $M_{Glimpses\_Cat}$ = 44.07 % , SD = 6.06. Like in experiment 4, there were overall more glimpses resulting from the cat sound compared to the violin sound. This difference in glimpses was significant, as revealed by a repeated measures ANOVA with Masker (2 levels: cat vs. violin) as the within-items factor. There was a robust main effect of the masker difference on the glimpse value, F(1,39) = 405.10, p < .0001, η^2 = .91. Visual examples of a word paired with each of the two maskers and the respective glimpses are provided in Figure 4.11.

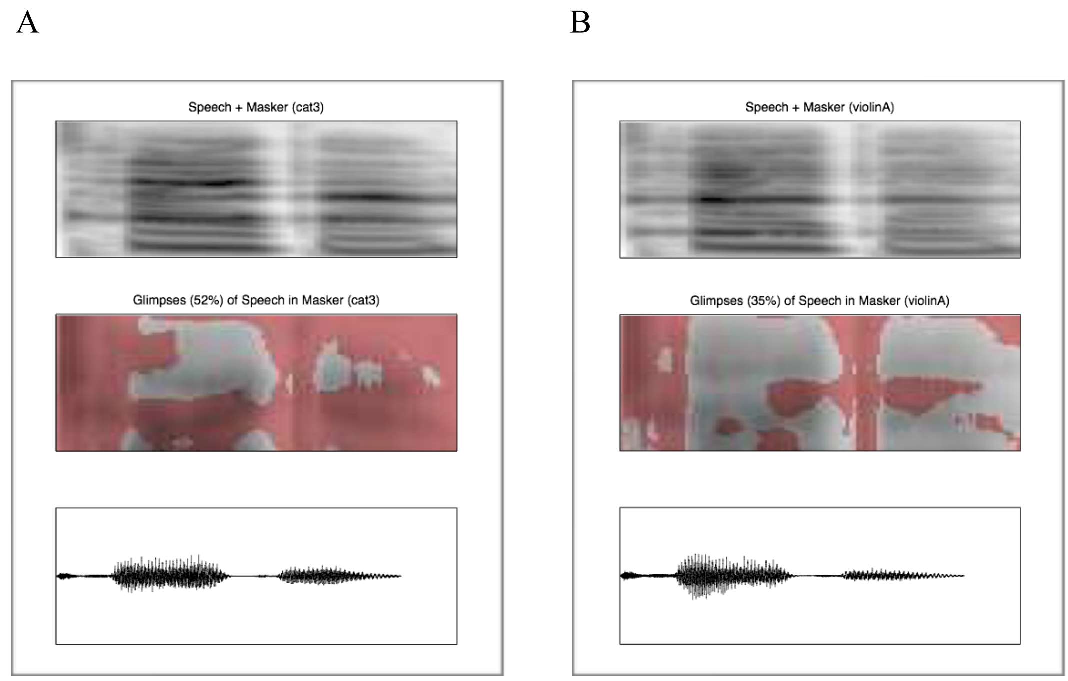

Figure 4.11. Spectrograms of the mixtures and the resulting glimpses of the same word paired with the two sounds in experiment 5. The sounds are not modulated according to the word's intensity envelope. The masked regions are shown in black and the glimpses in red. The image in A depicts the case when the word is paired with the cat sound and the image in B, the case when the same word is paired with the violin sound.

<u>*Comparison between experiments*</u>

The glimpse differences in both experiments were compared via a two-way repeated measures ANOVA with Experiment (2 levels) and Masker (2 levels) as factors. As anticipated, there was a robust main effect of Masker, $F(1,39) = 704.56$, $p < .0001$, $\eta^2 = .95$. A main effect of Experiment was also present, $F(1,39) = 86.20$, $p < .0001$, $\eta^2 = .69$.

Importantly, there was an interaction between Experiment and Masker, $F(1,39) = 71.50$, $p < 0001$, $\eta^2 = .65$, showing that the glimpse difference in Experiment 5 (Diff.: 44.07 % - 22.40 % = 21.67 %) was significantly greater than the glimpse difference in Experiment 4 (Diff.: 46.10 % - 34.09 % = 12.01 %).

This analysis undermines the possibility of a glimpses account for the specificity effect found in Experiment 4, since the expectation in this account is that a greater glimpse difference has a greater chance of leading to a sound specificity effect. Accordingly, the glimpses account would have favoured the chance of an effect in Experiment 5, that displays a significantly greater glimpse difference than that in Experiment 4. Contrary to such a prediction, the opposite pattern of results was observed, further consolidating the integrality account as the explanation behind the sound specificity effect observed in Experiment 4.

4.6. Comparative analysis between experiments

A comparative analysis between all the experiments discussed so far was also conducted, in order to compare the sound specificity effect found in Experiment 4 against the other specificity effects found in the previous chapters (voice and sound), as well as the insignificant effect in Experiment 5. Ideally, we would expect the voice specificity and the sound specificity effect found in Experiment 4 to be statistically comparable (no interaction). Further, we would expect the sound specificity effect in Experiment 4 to be statistically stronger than the insignificant effect in Experiment 5.

- *Integral vs. non-integral maskers*

The sound specificity effect found in Experiment 4 was compared with the insignificant effect in Experiment 5. A mixed-effects regression analysis was conducted, with an extra fixed factor added, Experiment (2 levels), coded as: 1 (Exp. 4) and 2 (Exp. 5). While random slopes were attempted for all the fixed factors, the main and interaction effects of only Sound Sameness and Experiment were assessed.[39] Since the sound specificity effect was found on both Accuracy and RT, both variables were included in the analysis.

- Accuracy

There was no main effect of Sound Sameness, $M_{Acc\_Same\ Sound}$ = 78.78 % correct, $M_{Acc\_Different\ Sound}$ = 76.38 % correct; β = .08, SE = .05, χ2(1) = 2.26, p = .13. Similarly, no main effect of Experiment was found, β = -0.06, SE = .0.14, χ2(1) = .2, p =.66.[40] There was also no interaction between Sound Sameness and Experiment, β = -0.13, SE = .08, χ2(1) = 2.54, p =.11.[41]

- Response Time (RT)

There was a main effect of Sound Sameness, $M_{RT\_Same\ Sound}$ = 1464.12 ms., $M_{RT\_Different\ Sound}$ = 1499.52 ms.; β = -15.29, SE = 6.45, χ2(1) = 5.31, p = .02. No main effect of Experiment was found, β = 77.68, SE = 50.18, χ2(1) = 2.37, p = .12. Further, there was no interaction between Sound Sameness and Experiment, β = 8.90, SE = 12.91, χ2(1) = .48, p = .49.

The presence of an interaction for either of the variables or both, would have put the sound specificity effect in a stronger position, but the lack of it is not very surprising, given that speech-extrinsic specificity effects tend to be relatively small in magnitude. Apparently, the sound effect in question is not strong enough to elicit a significant interaction.

- *Voice vs. integral maskers*

[39] In all the models involved in the comparative analysis, slopes for the Experiment factor were added only on a by-item basis.

[40] Only the random slope of Sound Sameness (added both by-subjects and by-items) and the random slope of Experiment (added by-items) were included in the respective models.

[41] Only the random slope of Sound Sameness was added by-subjects and the random slope of Experiment by-items, since the respective models were the only ones including random slopes to converge.

The sound specificity effect found in Experiment 4 was compared against the voice specificity found in Experiment 1. The extra fixed factor Experiment (2 levels) was coded as: 1 (Exp. 1) and 2 (Exp. 4). Since the sound specificity effect was found on both Accuracy and RT, both variables were included in the analysis.

- Accuracy

There was a main effect of Sameness (specificity effect), M_{Acc_Same} = 81.09 % correct, $M_{Acc\_Different}$ = 76.09 % correct; β = .16, SE = .04, χ2(1) = 15.51, p < .0001.[42] No main effect of Experiment was found, β = -0.06, SE = .14, χ2(1) = .18, p = .67. As expected, there was no interaction between the specificity effect (Sameness) and Experiment, β = -0.04, SE = .08, χ2(1) = .26, p = .61.[43] The lack of an interaction shows that the voice and sound specificity effects are statistically comparable, as anticipated.

- RT

There was a main effect of Sameness (specificity effect), $M_{RT\_Same\ Sound}$ = 1303.60 ms., $M_{RT\_Different\ Sound}$ = 1335.25 ms.; β = -13.09, SE = 4.97, χ2(1) = 6.47, p = .01. A main effect of Experiment was also found, β = 254.67, SE = 41.99, χ2(1) = 31.27, p < .0001. However, there was no interaction between the specificity effect and Experiment, β = -12.61, SE = 9.86, χ2(1) = 1.64, p = .20, indicating that the specificity effect on response latency persists between experiments, but is not strong enough to elicit an interaction.

- *Voice vs. non-integral maskers*

The voice specificity effect (Experiment 1) was compared against the insignificant effect found in Experiment 5. The extra fixed factor Experiment (2 levels) was coded as: 1 (Exp. 1), 2 (Exp.5). Given that the voice specificity effect was found only for Accuracy, only this variable was included in the analysis.

There was a main effect of Sameness on Accuracy: M_{Acc_Same} = 79.47 % correct, $M_{Acc\_Different}$ = 76.44 % correct; β = .10, SE = .04, χ2(1) = 5.56, p = .02.

No main effect of Experiment was found, β = -0.12, SE = .15, χ2(1) = .63, p = .43.

However, there was an interaction between Sameness and Experiment, β = -0.18, SE = .08, χ2(1) = 4.46, p = .03, indicating the robustness of the voice specificity effect compared to the insignificant effect of Experiment 5.[44]

- *Integral maskers vs. contrasted glimpses*

[42] The models included a random slope of Sameness (added only by-subjects) and the random slope of Experiment (added by-items).

[43] The models included a random slope of Sameness (added only by- subjects), and the random slope of Experiment (added by-items), since these models were the only maximal ones to converge.

[44] All the models involved in this comparison included the random slopes of Sameness (added both by-subjects and by-items) and Experiment (added by-items).

The sound specificity effect due to contrasted glimpses (Experiment 3) was compared against the sound specificity effect due to integral maskers (Experiment 4). The extra fixed factor Experiment (2 levels) was coded as: 1 (Exp. 3), 2 (Exp.4). The analysis was performed on both Accuracy and RT, since a sound specificity effect was found on both these variables in Experiment 4.

- Accuracy

There was a main effect of Sameness: M_{Acc_Same} = 79.67 % correct, $M_{Acc\_Different}$ = 75.63 % correct; β = .13, SE = .04, χ2(1) = 10.66, p = .001.[45]

No main effect of Experiment was found, β = .07, SE = .13, χ2(1) = .29, p = .59.[46]

Further, there was no interaction between Sameness and Experiment, β = .02, SE = .08, χ2(1) = .09, p = .76, showing that the two sound specificity effects are not statistically different from each-other.[47]

- Response Time (RT)

There was a marginal main effect of Sameness (specificity effect), $M_{RT\_Same\ Sound}$ = 1383.59 ms., $M_{RT\_Different\ Sound}$ = 1402.20 ms.; β = -10.02, SE = 5.22, χ2(1) = 3.68, p = .055.

A main effect of Experiment was also found, β = 102.76, SE = 42.25, χ2(1) = 5.75, p = .02. However, there was no interaction between the specificity effect and Experiment, β = -17.22, SE = 10.45, χ2(1) = 2.71, p = .10. The pattern of specificity effects and insignificant effects across all the experiments is graphically depicted in Figure 4.12.

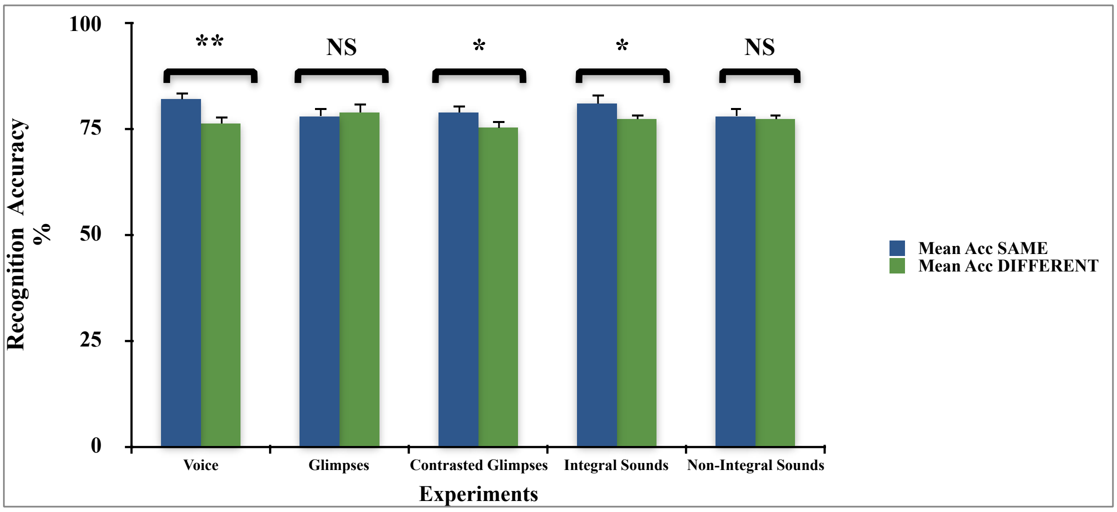

Figure 4.12. The pattern of specificity effects across experiments is illustrated via the mean recognition accuracy percentages at test, as a function of the voice or sound change from exposure to test.

[45] The models included a random slope of Sameness (added only by-subjects) and the random slope of Experiment (added by-items).

[46] The models included only the random slopes of Sameness (added both by-subjects and by-items) and Experiment (by-items), since the addition of the other random slopes did not allow the models to converge.

[47] The models included only a random slope of Sameness (added only by-subjects), since the addition of the other slopes lead to a failure to converge.

4.7. General discussion and conclusions

This chapter explored another plausible constraining factor for the appearance of a sound specificity effect, one that adds another dimension to the co-occurrence of words and sounds, namely, integrality. The rationale behind the studies was motivated by the analogy to the voice specificity effect. More specifically, I concentrated on the fact that words and voices necessarily co-exist, but are also integrated to form a single, common acoustic signal. As such, they cannot be perceptually segregated. I reasoned that the interplay between these properties of spoken words may be what makes the voice specificity effect more robust and more resilient to the experimental context in which it is tested, compared to the sound specificity effect.[48] Therefore, I sought to implement both these properties in the word-sound pairs in Experiment 4: co-occurrence and integrality, as closely as possible to the case of spoken words. I was interested to see whether a sound specificity effect that was relatively comparable to the voice specificity effect would appear. Speech modulation was used to implement integrality between the words and sounds, such that the intensity envelope of the sounds was replaced by the envelope of each word. As expected, there was a sound specificity effect when the stimuli were made integral, in Experiment 4. Interestingly the effect manifested itself in both the word recognition accuracy and response latency, marking the first time a specificity effect was observed in response latency in the series of experiments described so far.

To account for the possibility that the acoustic and/or semantic difference between the two sounds elicited the sound specificity effect, instead of integrality, a second experiment was conducted. It was identical to the first one, except that the integrality element was removed from the stimuli. As expected, the sound specificity effect disappeared. Further, a comparative glimpse analysis confirmed that the observed effect could not be explained by a difference in glimpse contrast. The results contradicted the prediction of the glimpses account, since there was a greater overall glimpse difference in Experiment 5 than in Experiment 4.

The comparative analysis across experiments did not reveal all the anticipated interactions, indicating the fragility of the sound specificity effect. When comparing the specificity effects obtained so far, the following concept may be useful. Consider the integrality between words and voices/sounds to be placed along a continuum as illustratively depicted in Figure 4.13. On one end there is maximum integrality, represented by spoken words and the respective voice specificity effect. On the other end, there is no integrality, represented by the word-sound pairs used in Experiment 3 (contrasted glimpses) and the respective sound specificity effect. Somewhere in between, closer to the integrality end, there is the integrality implemented in the stimuli of Experiment 4 and the respective sound specificity effect. An interaction between the integrality effect in Experiment 4 and the insignificant effect in Experiment 5, as well as an interaction between the former and the contrasted glimpses effect in Experiment 3, would have strengthened the position of the integrality effect in the continuum of specificity effects. However, considering the fragility of speech-extrinsic specificity effects in general and the fact that they appear relatively small in magnitude, the lack of interactions is not very surprising.

At this point, it is also worth noting that technically speaking, it is possible to further increase the level of integrality/difficulty in segregation between the words and sounds via the modulation technique. From a perspective of a “time-frequency” representation (spectrogram) of a

[48] By context, I refer to the experimental conditions that elicit the effect.

signal, there are two dimensions across which envelopes can be extracted: time and frequency. In the present integral maskers, the intensity envelope of the words across time was used as the modulator for deriving the intensity envelope of the integral sound maskers, while the envelope across frequency, that yields the spectral properties of the signal, was essentially derived from the original sounds.[49] There are ways to make the sound harder to segregate from speech via amplitude modulation by increasing the number of bands in which the sound signals are filtered, which in the present case was only one band, between 300-6000 Hz. The presence of only one wide filtering band kept the spectral information of the sounds relatively intact. Alternatively, higher integrality could be achieved by combining amplitude and frequency modulation, such that both the frequency and intensity envelopes of the words would be used as modulators for the sound maskers. However, in both these cases and especially in the second one, the sound would loose its peculiar identity and there would hardly be any contribution from it in the overall mixture. Instead, depending on how further integrality is enhanced in the stimuli, the sound would sound more like a weird distortion related to the word, than like a distinct co-occurring masker. Similarly, the overall mixture would sound like an awkward and unnatural version of the spoken word. In other words, the stimuli would be more integral with the sound blending even more with the word, and perhaps a stronger sound specificity effect would have emerged. However, there would be little-to-no space left for discussing the contribution of a co-occurring speech-extrinsic sound to word recognition and encoding in memory. Given that the aim was that the sounds were integral to the words, but also preserve their distinct identities, the chosen modulation technique worked particularly well in satisfying both these conditions. It is thus quite interesting to observe a relatively robust sound specificity effect at a relatively moderate level of integrality.

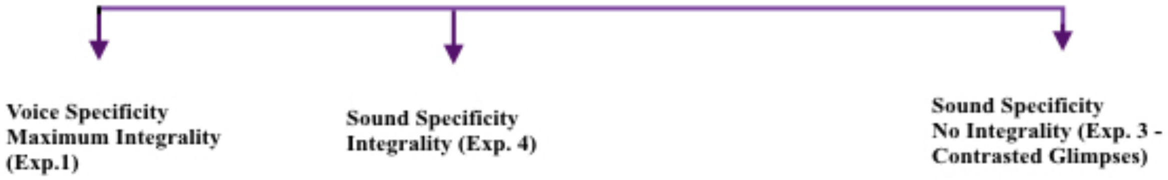

Figure 4.13. A simplistic visualisation of the integrality continuum across specificity effects.

Taken together, the present results provide evidence for an integrality account, in which the acoustical and perceptual blending of spoken words with their paired background sounds seems to play a crucial role in the emergence of a sound specificity effect. They extend prior work on speech-extrinsic specificity effects by revealing a novel condition on the stimuli: the difficulty in perceptually segregating the sounds from the words, which suggests a perceived functional/causal link between words and sounds. The latter is reminiscent of the functional/causal link between words and voices, which was also the main motivation behind the rationale of the studies described above. The focus on the strong analogy with the case of the voice specificity effect led to a closer inspection of the intrinsic relationship between words and voices, which in turn paved the way for

[49] More specifically, the words' intensity envelopes were multiplied with the sounds. Therefore, the resulting sound maskers got a new intensity envelope that followed that of the individual words. The information across the frequency envelope of the words was not used at all, solely the intensity envelope, or in alternative terms, the overall intensity contour.

the integrality account. This approach of investigating speech-extrinsic specificity effects through a solid analogy to speech-intrinsic ones, is novel on it own, and I argue, important for a better assessment and understanding of these effects.

The present findings also contribute to consolidating the view that sound specificity effects are fragile, conditional and constrained by the context in which they are probed. Another study recently reported a similar finding with respect to the conditional nature of such effects, were perceptual segregation was identified as a potential constraint. In one of their experiments, using an explicit continuous recognition memory task at test, Cooper et al (2015) observed that the recognition of spoken words was affected by a change in the background noise. However, this effect was constrained, such that it only appeared under the condition of spectral overlap between the speech and the background noise. Spectrally overlapping signals are more difficult to segregate than non-spectrally overlapping ones, which is in line with our argument. However, it is worth noting that the difficulty of perceptual segregation in the present integral stimuli is greater that the one involved in Cooper et al (2015)'s stimuli. Spectrally-overlapping does not necessarily entail integrality in the way that the concept was defined and implemented in the present account. Further, all our non-integral stimuli across the experiments described so far have been spectrally overlapping, but a sound specificity effect appeared only when the two masking sounds created contrasted acoustic glimpses of the same word(s) (Experiment 3).

Another important implication of the current findings is that they further reinforce the argument that mere co-occurrence between words and sounds may not always be sufficient for the emergence of a sound specificity effect. If co-occurrence per se was sufficient, then there should have been a sound specificity effect in both experiments discussed here, regardless of the integrality factor.

In conclusion, the results obtained in this chapter support episodic views of the mental lexicon, where the memory episodes of the words can include speech-extrinsic auditory information that co-occurs with spoken words, provided additional conditions are also satisfied. In the present case, memory episodes are stronger when the sounds are made integral to the words, such that perceptual segregation becomes challenging. The previous chapter put forward a "contrasted glimpses" account and the present chapter proposes an integrality account, as conditions in the stimuli that seem to play a role in the appearance of a sound specificity effect. The next chapter will deal with another potential condition, uniqueness in the pairwise associations between words and sounds. In doing so, it will also attempt at a more truthful replication of the original sound specificity effect reported by Pufahl and Samuel (2014).

Chapter 5

Effect of pairwise speech-sound association on the sound specificity effect

Abstract

The present chapter investigates the emergence of a sound specificity effect in the presence of a one-to-one, unique pairing between a spoken word and a background sound, which I refer to as *association uniqueness.* As such, this chapter explores another plausible context in which a speech-extrinsic specific specificity effect may potentially emerge. At the same time, it also attempts at a replication of the original sound specificity effect found in Pufahl and Samuel (2014), where a similar pairing method was used. To ensure as truthful a replication as possible, the same encoding, test tasks and stimuli filtering technique as in the original study were used. Listeners heard the filtered versions of the stimuli at test and unlike in the previous experiments, performed an implicit memory task, namely perceptual identification of words. The filtering technique implemented was multiple band-pass filtering and the first experiment of the study served as a pilot step to determine the optimal multiple band-bass filter-banks to be used in the main experiment. The main experiment shared design and stimuli details with the respective experiment in Pufahl and Samuel (2014). However, unlike their main result, no sound specificity effect on word identification accuracy was observed. Namely, the change in the paired sound exemplar from exposure to test did not lead to a decrease in the overall accuracy, which may be further evidence that speech-extrinsic specificity effects are relatively fragile and susceptible to the context in which they are probed. The result is discussed in light of the Pufahl and Samuel (2014)'s finding, as well as in light of the main results from the previous chapters.

5.1. Introduction

Pufahl and Samuel (2014) were the first to report a sound specificity effect on spoken word identification. The relevant experiment in the study that revealed the effect was explained in detail in the previous chapters, hence only a brief summary is provided here, for the reader's convenience. Words spoken by either a male or a female talker were paired with either one of two sound exemplars belonging to the same sound category, such that each sound category was unique to a word. For example, the word "butterfly" was paired with either the exemplar A or B of a harmonica sound. Hence, the sound change from exposure to test involved only a change in the exemplar, while preserving the sound category. During exposure, listeners performed an "animate vs. inanimate word" semantic judgement task on stimuli heard in the clear. At test, they transcribed the words from the filtered versions of the stimuli. The change in the paired sound exemplar from exposure to test evoked a sound specificity effect, reflected in a decrease in the overall word identification accuracy.

Throughout the work described in the previous chapters, we have argued that mere co-occurrence is not sufficient, or the only factor in the emergence of speech-extrinsic specificity effects. The vulnerability of such effects to the context they are being probed in supports this argument, as well as triggers a quest to identify plausible contexts that reveal their appearance. A context can consist of anything related to the nature of the stimuli, the encoding and test tasks, the listeners, or a combination of all of them. Each chapter described so far in this thesis has explored one potential context that has added another necessary element to the co-occurrence between words and sounds. These additional elements so far have been related to the nature of stimuli, while the same experimental tasks have been consistently used across the studies.

The present chapter investigates another plausible context, but this time with a focus on the nature of the stimuli, as well as the memory task in the test phase. In doing so, it diverts from the previous studies in both these methodological aspects, but gets considerably closer to the original study by Pufahl and Samuel (2014), targeting a closer replication of the original effect. Before delving into further details, I define the factor of interest, whose potential role in the appearance of a sound specificity effect inspired the present study. To this end, the relevant study by Pufahl and Samuel (2014) will serve as a reference point.

I will refer to this factor as "association uniqueness", which stands for the unique pairing between a word and a sound category. On a comparative basis to the case of speech-intrinsic (indexical) effects, "uniqueness" seems to be an inherent characteristic of spoken words. More specifically, every spoken word is a unique utterance, with peculiar linguistic and indexical properties that are hardly matched by another utterance. No two words are realised in the same way acoustically, even when spoken by the same talker. Perhaps the consistency and robustness of indexical effects, especially compared to speech-extrinsic specificity effects, is due to this "uniqueness" element of spoken words. For example, when the word "table" is spoken by a male talker in exposure, this unique utterance is encoded with the peculiar properties that characterise it. Afterwards, when the same word is heard at test spoken by a different talker, it constitutes another unique utterance that is different from the one heard at test. Since the two different acoustical realisations of the same linguistic item do not match, the recognition/identification performance for this item decreases and/or is slowed down, leading to an indexical effect.

Similarly, in the case of speech-extrinsic specificity effects, a unique, 'one-to-one' association between a word and a sound may be more salient and encoded better in memory than a non-unique association. As such, it could be more sensitive to a change in the context of the stimulus, namely the paired sound. For example, if the word "window" is paired with an harmonica sound during exposure and it is the only word to be paired with this sound, the sound might be encoded as an external feature associated with this particular word. In the test phase, when the same word is repeated with a different harmonica sound (assuming the sound change occurs within the category, like in Pufahl and Samuel, 2014), the listeners may remember the unique association between "window" and an harmonica sound, but also realise that the sound is not the same as in exposure, and fail to recognize/identify the word. Therefore, the association uniqueness may increase the chance that background sounds become encoded in memory alongside the respective words.

In addition, increasing the number of background sounds contributes to higher uncertainty and unpredictability levels in the overall experimental context, promoting listeners' reliance on contextual cues, such as the paired sounds. In the case of only two sounds, the listener learns quickly within the course of the experiment that a word is paired with either one of them. Hence, there is no surprise unpredictability issue regarding the next stimulus: the word will be paired with one of the already heard sounds. This enhanced predictability may in turn facilitate the ignoring of the background sound. However, in the case of many sounds, especially when the association between the words and sounds is 'one-to-one', it is impossible for the listener to anticipate the sound that the next word will be paired with. Such unpredictability may then lead to the listeners' paying more attention to the stimuli as pairs, rather than easily ignoring the sound. In contrast, if the word is paired with one of two sounds, and half of the words in the stimuli set also are, there is a 'many-to-one' relationship between the words and the sounds (as it was the case in our previous studies). The listeners may not easily realise at test that a word is paired with the other sound, for two reasons. First, the sound was paired with half of the other words in exposure, making it highly familiar and a relatively stable element within the overall experimental context. This could in turn contribute towards listeners' habituation to the sound and consequently, their improvement at ignoring it. Second, the different sound that the word is paired with at test, has also been heard multiple times during exposure with other words, thus being a familiar sound to which the listeners have grown habituated to as well. Adding to these points the relative ease of segregation between words and sounds (unless they are integral to one-another, as in the previous chapter), ignoring the sounds becomes quite easy.

Following this line of argument, perhaps a possible explanation for the lack of a sound specificity effect in two previous experiments (Experiment 2A and 2B, **Chapter 3**) may be the fact that the associations between words and sounds were not unique.[50] Similar to Pufahl and Samuel's study, the sound change from exposure to test also took place within the same sound category (car horn) and between different exemplars (high pitch vs. low pitch) in these experiments. However, unlike in the former case, such a change was not unique. Therefore, given Pufahl and Samuel's finding and the lack of an effect in the two aforementioned experiments, I was interested to see whether adding the uniqueness element to the stimuli would evoke a sound specificity effect.

A plausible way to implement "association uniqueness" as a primary factor in the quest for speech-extrinsic specificity effects, would be to enrich the stimuli context by increasing the number of background sounds. This approach brings along another question of interest: how many sounds would be needed for the emergence of an effect? While this question is interesting and deserves investigation on its own, it is beyond the scope of the present work. For the present purposes, a maximum "association uniqueness" was target, which involves a 'one-to-one' association type between the words and sounds. Being the first and only study to report a sound specificity effect using multiple background sounds with this type of unique association, Pufahl and Samuel's study served as a useful reference regarding the plausible number and type of sounds. In order to increase

[50] Note that the studies that revealed an effect also used two sounds, however those studies involved an additional element in the paired stimuli, namely contrasted masking and integrality.

the likelihood of a sound specificity effect, the same sound set used in the reference study was used in the present work.

Regarding the tasks in exposure and test, the same encoding task as in the previous studies was used. In contrast, the memory task was different from the one previously used. Instead of the recognition memory task, an implicit memory task, like the one in Pufahl and Samuel (2014) was delivered. This served two purposes: 1) a closer experimental design to the original study that first reported the effect, and 2) the possibility to try a different memory task, that among other things, has been considered more reliable in measuring specificity effects, compared to its explicit counterpart.

In summary, the present study's purpose was twofold: 1) to investigate the role of association uniqueness in the emergence of a sound specificity effect, and 2) to obtain a relatively close replication of the original effect by Pufahl and Samuel (2014). The first experiment (Experiment 6) was designed to pilot the optimal filtering level for the stimuli in the implicit memory task at test. The main experiment (Experiment 7) investigated the role of "association uniqueness" in the emergence of a sound specificity effect. Based on the potential impact of this factor and the main result of Pufahl and Samuel (2014), I anticipated to find a sound specificity effect reflected in the overall word identification performance.

5.2. Experiment 6 - Piloting the intelligibility of the filtered words

5.2.1. Method

5.2.1.1 Participants

Fifty-seven students at the University of York (age range: 18-23) participated in exchange for either course credit or payment. Forty-five participants participated in the first phase of the pilot study (Group 1) and the other twelve participants participated in the second phase of the pilot study (Group 2). The relatively large number of participants for a pilot study was due to the several filtering conditions piloted (details in the next section), with the aim of testing at least five participants per condition. All participants provided written consent prior to the experiment. They all identified themselves as native-speakers of British English and none of them reported a history of hearing or speech and language related problems. None of the participants had done any of the previous experiments.

5.2.1.2. Materials and Design

The stimuli consisted of 80 word-sound pairs. The words were the same as in all the previous experiments. The same background sounds as in Pufahl and Samuel (2014) were used.[51] They consisted of 80 different sound categories (e.g., barking dog, guitar sound, bird chirping, doorbell ringing, etc.) and for each of these categories there were two different exemplars, labeled as A and B. Half of the sound categories belonged to animate sources and half of them to inanimate sources. Each word was paired with one of the exemplars belonging to the same semantic category. Hence, there was a unique, one-to-one pairing between a word and a sound category, that once randomly

[51] The first author, April Pufahl, kindly provided their set of sounds.

assigned, remained fixed (e.g., the word “peanut” was always paired with the barking dog sound category, either exemplar A, or B). The sound change from exposure to test involved only the exemplar change (A to B, or vice versa), while preserving the semantic category of the sound. This pairing method satisfied our “association uniqueness” notion and was also identical to the one implemented in Pufahl and Samuel (2014). A list of all the sound categories used is provided in Appendix B.

Further, in line with the filtering technique used in Pufahl and Samuel (2014), we also implemented multiple band-pass filtering to degrade our stimuli for the test phase. In both the present and the next experiment, all the stimuli files were generated with a sampling rate of 44.1 KHz and a resolution of 16 bit. Every stage of the stimulus preparation process was implemented using the Matlab software (version R2014b). The following sections explain the details behind the technique, the selection of an optimal filter-bank for our purposes and the preparation of the final stimuli.

Multiple bandpass filtering

A) Retrieving the filter transfer function from Pufahl and Samuel (2014)

We performed reverse-engineering on the original stimuli used in Pufahl and Samuel (2014)’s first experiment. The overall frequency transfer function of the filtering used was obtained by calculating the frequency spectra of the unfiltered and filtered stimuli (spoken words as well as environmental sounds), and dividing the spectra of the filtered stimuli by the spectra of the unfiltered stimuli. The following figure shows the retrieved frequency transfer function form the filtering used in the aforementioned study:

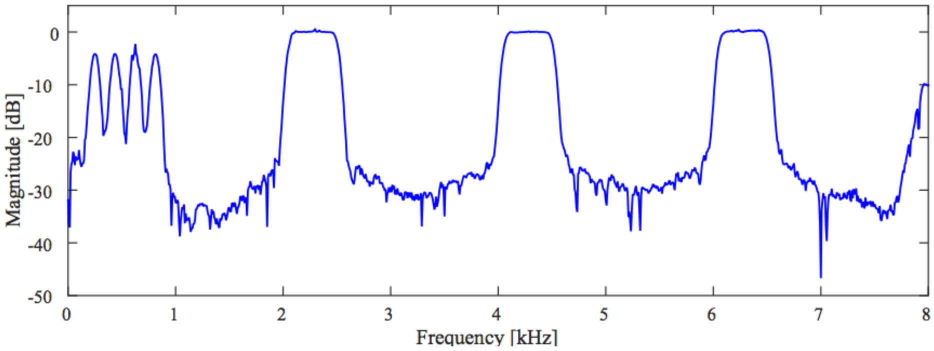

Figure 5.1. Frequency transfer function retrieved from Pufahl and Samuel's stimuli

The lower and upper cut-off frequencies (f_{low} and f_{high}) of the individual bandpass filters used in Pufahl and Samuel (2014) are listed in Table 5.1:

Table 5.1. The cut-off frequencies as reported in Pufahl and Samuel (2014).

Band No.	1	2	3	4	5	6	7	8
f_{low} [Hz]	200	400	600	800	2000	4000	6000	8000
f_{high} [Hz]	250	450	650	850	2500	4500	6500	8500

It is worth noting that Pufahl’s filtered stimuli had a sampling rate of 16 kHz, which means that the maximum frequency that can actually be represented as upper limit is half of that, i.e., 8 kHz. Therefore, the highest filter band (8 – 8.5 kHz) is practically non-existent, as it is outside the representable frequency region. For this reason, this band was not considered during the preparation of the filter-bank for our stimuli.

B) Defining a filter-bank with similar characteristics to that used in Pufahl and Samuel (2014)

The next step was to define a bank of bandpass filters that showed a similar frequency transfer function as the one obtained from Pufahl and Samuel's stimuli (Figure 5.1). The bandpass filter-bank was implemented in Matlab (version R 2014). For the low-frequency bandpass filters (band no. 1-4, Table 5.1), four 4th-order Butterworth bandpass filters were defined using Matlab's "butter" filter function. For the high-frequency bandpass filters (band no. 5-7, Table 5.1), three 4th-order elliptic filters were defined using Matlab's "ellip" filter function (peak-to-peak ripple: 0.1 dB, minimum stop-band attenuation: 30 dB). The cut-off frequencies for the individual bandpass filters were adjusted such that the overall frequency transfer function of the entire filter-bank matched the frequency transfer function retrieved from the Pufahl and Samuel's stimuli. The final values of these cut-off frequencies differed slightly from those displayed in Table 5.2:

Table 5.2. Cut-off frequencies that yielded the best fit with the transfer function derived from Pufahl and Samuel's stimuli

Band No.	1	2	3	4	5	6	7
f_{low} [Hz]	222	410	600	785	2080	4080	6080
f_{high} [Hz]	272	460	650	835	2480	4500	6500

Using these cut-off frequencies, we obtained a strong fit with the transfer function estimated from Pufahl and Samuel's stimuli. Both transfer functions are shown in the figure below:

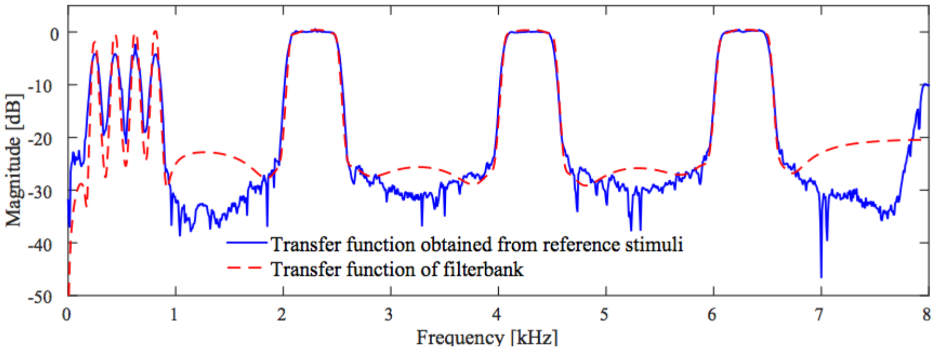

Figure 5.2. Frequency transfer function retrieved from Pufahl and Samuel's stimuli (blue) along with transfer function of our designed filter-bank.

C) Modifying the bandwidth of the filters to create additional filters

The initial stage of the pilot experiment, in which the original filter-bank (*d* = *1*) was piloted for both the words and sounds, revealed a higher average word intelligibility level (Mean Acc. = 84.17 % correct) than the one targeted (70-75% correct). The targeted mean intelligibility level was meant to be neither too high, nor too low, and indicate that the task was challenging, but not overly so. This high average word identification accuracy may have been partly due to the fact that our female talker had a very clear and highly intelligible overall pronunciation. Given that the aim of the implicit memory task in the test phase was to render perceptual identification of the words challenging, we reasoned that a high average word intelligibility level would not serve this aim very well. At the same time, we did not want to make the task at test so challenging that the participants would struggle too much and mostly engage in a transcription task, rather than in an implicit memory task. Therefore, the targeted average intelligibility value of 70-75% correct was deemed optimal.

Importantly, the initial piloting also revealed that the words were not evenly intelligible when filtered by the same filter-bank, with some of them being highly intelligible and others much less so. This indicated that using one filter-bank would not yield an even filter for all words. Therefore, we piloted several filter banks. As a result, we were able to choose the optimal filter-bank for each word (or group of words) individually, such that intelligibility for each individual word was as close as possible to the targeted level.

To achieve even filtering, we made some changes to the bandwidth of the initial filters (displayed in Table 5.2) and created several additional filtering conditions. Filters can either be described by their lower and upper cut-off frequencies (f_{low} and f_{high}) or by their center frequency f_c and bandwidth B. The center frequency is the frequency in the middle of the filter band: $f_c = (f_{low} + f_{high})/2$. The bandwidth is the width of the filter band and is calculated as the difference between the upper and lower cut-off frequencies: $B = f_{high} - f_{low}$. We decided to preserve the center frequencies and modify the bandwidth of the filters to create additional filter-banks.

Table 5.3 provides an overview of the cut-off frequencies, center frequencies and bandwidths in the filter-bank used in Pufahl and Samuel's study:

Table 5.3. Cut-off frequencies, center frequencies, and bandwidths used in Pufahl and Samuel (2014).

Band No.	1	2	3	4	5	6	7	8
f_{low} [Hz]	200	400	600	800	2000	4000	6000	8000
f_{high} [Hz]	250	450	650	850	2500	4500	6500	8500
f_c [Hz]	225	425	625	825	2250	4250	6250	8250
B [Hz]	50	50	50	50	500	500	500	500

The values of the cut-off frequencies, centre frequencies and bandwidths that were used in filtering our stimuli are displayed in Table 5.4. These cut-off frequencies yielded the best fit with the transfer function retrieved from Pufahl and Samuel's stimuli.

Table 5.4. Cut-off frequencies, center frequencies, and bandwidths used for filtering our stimuli.

Band No.	1	2	3	4	5	6	7
f_{low} [Hz]	222	410	600	785	2080	4080	6080
f_{high} [Hz]	272	460	650	835	2480	4500	6500
f_c [Hz]	247	435	625	810	2280	4290	6290
B [Hz]	50	50	50	50	400	420	420

The bandwidths B_d for the additional filtering conditions were defined by the divisor d: $B_d = B/d$. The values of divisor d chosen for the pilot experiment were: 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, 4. The center frequencies were unchanged throughout all the filtering conditions (values displayed in Table 5.4). The bandwidths B_d corresponding to the different d-values are specified in the following table, calculated according to the formula above ($B_d = B/d$) and rounded to integer values:

Table 5.5. Center frequencies and the bandwidths used for the several filtering conditions.

Band No.	1	2	3	4	5	6	7
f_c [Hz]	**247**	**435**	**625**	**810**	**2280**	**4290**	**6290**
B [Hz]	**50**	**50**	**50**	**50**	**400**	**420**	**420**
$B_{0.25}$ [Hz]	200	200	200	200	1600	1680	1680
$B_{0.5}$ [Hz]	100	100	100	100	800	840	840
$B_{0.75}$ [Hz]	67	67	67	67	533	560	560
B_1 [Hz]	50	50	50	50	400	420	420
$B_{1.25}$ [Hz]	40	40	40	40	320	336	336
$B_{1.5}$ [Hz]	33	33	33	33	267	280	280
$B_{1.75}$ [Hz]	29	29	29	29	229	240	240
B_2 [Hz]	25	25	25	25	200	210	210
B_3 [Hz]	17	17	17	17	133	140	140
B_4 [Hz]	13	13	13	13	100	105	105

As it can be seen from Table 5.5, the new bandwidths (B_d) are inversely proportional to the old ones (B), i.e., if $d < 1$, the bandwidth increases, and, if $d > 1$, the bandwidth decreases (relative to the original B used in the calculation). For a visual illustration of this, consider Figure 5.3, which shows six different transfer functions obtained for different bandwidths B_d (only a subset of the d-values was used to generate the image):

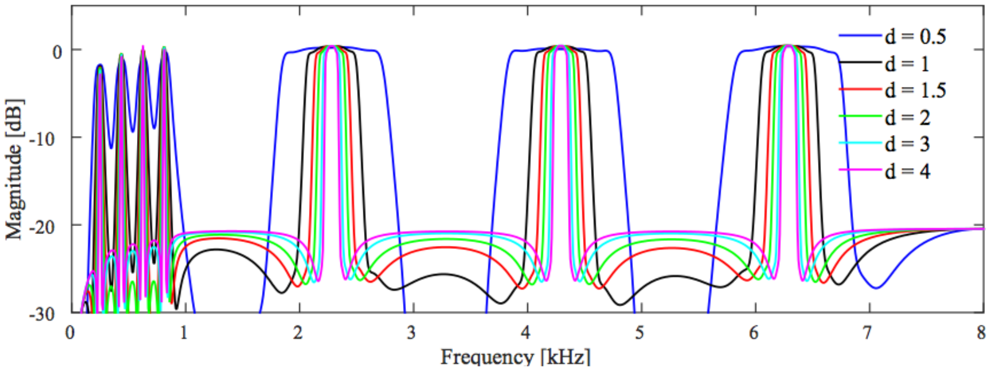

Figure 5.3. The filterbank transfer functions for different selected d-values

The environmental sound were filtered with the original filter-bank (*d* = 1), that was very similar to the one used in Pufahl and Samuel (2014).

II. Creating the final stimuli

The final stimuli consisted of word-sound pairs in both the unfiltered and the filtered versions. The unfiltered stimuli were used in the exposure phase, and their filtered version in the test phase. In the case of the filtered stimuli, the words and sounds were filtered separately prior to mixing. As explained above, while the sounds were filtered with the original (Pufahl) filter-bank (*d = 1),* the words were filtered with the following filter-banks: *d* = 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, and 4. Hence, there was only one filtered version for the sounds and several filtered versions for the words. The several filtering conditions that were eventually piloted were a result of the different filter-banks used to filter the words, with the filtered sounds being the same in each of them. There were six main filtering conditions (groups), where all the words were filtered with the same respective filter in each one of them: *d* = 1, 1.25, 1.5, 1.75, 2, 3. These conditions constituted the first phase of the piloting study (Phase 1), completed by the participants in Group 1. The other filters (*d* = 0.25, 0.5, 0.75, 4) were applied only to a few words that were on the two extreme ends of the intelligibility spectrum when filtered. These words were identified once Phase 1 was complete and the filter-banks for the majority of the words had been finalised. It was noticed that the mean word identification accuracy was slightly lower than the targeted level and that the aforementioned words constituted “edges” among the other filtered words and as such, needed to be “smoothed” further. Therefore, the additional filter-banks mentioned above were used for these words and their filtered versions were then added to the set of the other filtered words that had resulted from Phase 1, thus creating an almost-final set of filtered stimuli. This set was then further piloted in the second phase of the study (Phase 2), which was completed by the participants in Group 2. A list of all the words and their respective filter-banks chosen at the end of the pilot experiment is provided in Appendix C.

In both the unfiltered and filtered versions, prior to mixing, the words and sounds were normalised to 68 dB by using a customised script in the Praat software . They were then mixed at 0 dB SNR . During mixing (implemented in Matlab), the centres of the two signals (words and sounds) were aligned and any difference in duration was filled with silence. Namely, for the first half of the shorter signal, silence was added at the beginning and, for the second half, at the end. After mixing, silence frames were added to form a 100-ms silence interval at each end of the mixed signal.

5.2.1.4. Procedure

The experimental paradigm was slightly different from the one used in the previous experiments, as well as from the design involved in Pufahl and Samuel (2014).[52] Namely, in the present experiment, there were 80 trials played in each phase, compared to the 60 trials per phase played in the previous experiments. Pufahl and Samuel (2014) had 64 trials per phase in their study. Further, unlike the previous experiments that involved an explicit memory task in the test phase (word

[52] Only the first experiment in Pufahl and Samuel (2014) is relevant for the present purposes, hence it will be taken as reference.

recognition), an implicit memory task was used in the present experiment, that involved identifying the word(s) from heavily filtered word-sound pairs. The encoding task in Exposure (animate/inanimate judgement) was identical to that in the previous experiments. Both the encoding task in Exposure and the memory task in Test were identical to those used in Pufahl and Samuel (2014)'s study. Another difference between the present experiment and its counterpart in Pufahl and Samuel (2014), is the number of experimental conditions at test that describe how well the items at test match the corresponding ones in exposure. Specifically, in the present case, there was only one independent variable (the paired sound) and two possible conditions at test related to it: 1) no change (same paired sound as in exposure), and 2) sound change (different paired sound from the one in exposure). On the other hand, Pufahl and Samuel (2014) had two independent variables (the paired sound and the voice of the talker), resulting in four conditions at test: 1) no change (neither the paired sound, nor the voice changed from exposure), 2) sound change (only the paired sound changed from exposure), 3) voice change (only the voice of the talker changed from exposure), and 4) sound and voice change (both the paired sound and the talker's voice changed from exposure).

Exposure phase

Participants heard a block of 80 unfiltered word-sound mixtures (Block 1), each played one at a time. None of the trials were repeated within the block. Each word was paired with a unique exemplar sound. Half the animate words were paired with an animate sound and the other half with an inanimate sound. The same was true for the inanimate words. Hence, there were 20 of each of the following pairing combinations: Animate word - Animate sound (A-A); Animate word - Inanimate sound (A-I); Inanimate word - Animate sound (I-A), and Inanimate word - Inanimate sound (I-I). While the word-sound category pairings were the same for all participants, which sound exemplar (either A or B) of the sound category was used in the pairing was counterbalanced across participants.

The experiment was run on the same software and using the same equipment as in the previous experiments. Participants sat individually in a sound-attenuated booth and listened to the trials played binaurally over headphones at a comfortable listening level. They read instructions on the computer screen and also listened to the experimenter's explanations. While their task was to make an "animate/inanimate" decision for the word in each trial, they were also instructed to pay attention to the accompanying sound. This was done with the main experiment (Experiment 7) in mind. Since the participants in the main experiment would be encouraged to pay attention to the background sounds, in order to further promote the emergence of a sound specificity effect, we included such an encouragement in the pilot experiment as well.

The 'animate' and 'inanimate' concepts were defined and examples for each of the categories were provided (e.g., "banana is inanimate", "professor is animate"). The experimenter encouraged them to be as accurate as possible. After 500 milliseconds, a message was displayed on the screen prompting them to respond by pressing either one of the corresponding 'shift' keys on the computer keyboard (INANIMATE left; ANIMATE right). Participants were told to wait for the message to appear on the screen before responding and were allowed a maximum of 10 seconds to submit a response. The next trial followed immediately after they hit a response button, or after the maximum allowed time expired, if no response was provided. Prior to the experimental trials, participants completed 8 practice trials, involving words spoken by a male talker that were different from the experimental words and paired with different environmental sounds from those used in the experimental trials. There were 80 experimental trials (Block 1) in total and their order was randomised for each participant. No feedback was provided after each trial and there was no mention of an upcoming word identification task. The task duration was around 15 minutes on average.

Delay

After completing the first experimental phase, participants left the sound-attenuated booth and spent 5-7 minutes on an unrelated distractor task prior to the memory test. This was done in order to ensure that performance in the subsequent test phase was not based on short-term or working memory. The task consisted of playing an online game (Cube Crash 2). It was different from the distractor task used in Pufahl and Samuel (2014), which involved answering a number of semantic illusions questions.[53]

Test phase

Participants heard the same 80 word-sound mixtures, but in their heavily filtered version (Block 2). Like in the exposure phase, none of the trials were repeated within the block. The same four animacy pairing combinations mentioned above were present. In each of these combinations, half of the words were paired with the same sound exemplar as in exposure and the other half with the different exemplar. Which words in the test phase were paired with the same or the different sound exemplar was counterbalanced across participants. The counterbalancing in terms of the sound exemplar (A or B) and sound sameness (same or different) resulted in four stimulus lists (counterbalancing groups) in total. Every participant was randomly assigned to either one of them. This counterbalancing scheme was present in every filtering condition in Phase 1, as well as in Phase 2, that involved only one condition, with the assorted set of filter-banks chosen for individual words or groups of words at the end of Phase 1. Every participant in Phase 1 was randomly assigned to one of the six main filter conditions and to one of the four counterbalancing groups within the condition, and as such, was exposed to only one filtered version of all the words. On the other hand, every participant in Phase 2 was randomly assigned to one of the four counterbalancing groups and was exposed to the various filter-banks involved in the almost-final set of filtered stimuli.

Participants performed an implicit memory task that involved transcribing the word(s) in the heavily filtered pairs. Participants read written instructions on the screen and also listened to the experimenter's explanations. The experimenter explained that they would hear again the same word-sound pairs, but they would sound distorted and would be challenging to understand. Their task was to type in the word they heard for every trial. They were encouraged to be as accurate as possible, but also to guess whenever necessary. Similar to the instructions in the first part of the experiment, they were instructed to pay attention to all the auditory information in a trial. No time limit was imposed. The next trial followed immediately after participants response. Prior to the experimental trials, participants completed the same 8 practice trials as in the exposure phase, but this

[53] Description taken from Pufahl and Samuel (2014): "Participants were given a sheet with 24 semantic illusions, like the Moses Illusion. In this illusion, when people are asked the question ''How many animals of each kind did Moses take on the ark?'' they generally respond ''two'' even though they know it was Noah, not Moses, who built the ark (Erickson & Mattson, 1981). Participants wrote their answers on the sheet below each question and circled ''yes'' or ''no'' to indicate if they had ever heard the question before."

time in their filtered version. There were 80 experimental trials in total and their order was randomised for each participant. In half of the trials, the words were paired with a different sound exemplar from the one in the exposure phase (i.e., exemplar B instead of the exposure exemplar A , or vice versa), and in the other half, they were paired with the same exemplar. The task duration varied from 20 to 30 minutes.

5.2.2. Results

A. Phase 1

Only the transcribed responses in the Test part of the experiment were analysed. Response accuracy was coded as a binary variable with values '1' and '0' per trial, where '1' corresponded to a correct response in a trial, and '0' to an incorrect one. During the rating of the responses, occasional spelling mistakes were judged as correct responses, whereas words that appeared similar in form to the target word(s), but belonged to a different grammatical and/or semantic category, were considered incorrect responses. For example, a response like "pantha" for "panther" was rated as correct and responses like "work" or "acting" for "worker" and "actor", respectively, were rated as incorrect. When in doubt, the experimenter sought the rating of a second person.

The average intelligibility accuracies for each piloted filter-bank were calculated on an individual word basis. The filter-bank that corresponded to the accuracy that was closer to the targeted overall intelligibility accuracy (70 - 75 % correct) was selected for each word (see the full list in Appendix C).

B. Phase 2

The filtered word-sound pairs, in which the filtered versions of the words were selected at the end of Phase 1, constituted the near-final stimuli set for the Test phase of the main experiment. They were further piloted and some more filters were attempted for those few words that were still highly intelligible or unintelligible. The additional filter-banks mentioned above were applied to these words and every new change was piloted again, with a few participants at each intermediate piloting step. The set of the filtered words was eventually finalised and was ready to be used in the main experiment. It is displayed in Appendix C as a list containing the word set and their respective final filter-banks. The set of the filtered environmental sounds was already final, without requiring any piloting.

5.2.3. Discussion

The aim of the above experiment was to determine the optimal multiple bandpass filtering level for the words and the sounds. This task proved relatively easy for the sounds, as they were filtered by a multiple bandpass filter-bank very close to the one used in Pufahl and Samuel, whose relevant filtering function was extracted and served as a basis for developing our filter-bank. However, ensuring a relatively even filter for all the words proved more difficult. It became evident that a single filter-bank would not be effective for all the words. Therefore, several filter-banks were developed and piloted. The pilot experiment simulated the main experiment, hence involved the same experimental phases and tasks. The resulting filtered versions of the words and sounds constituted the final stimuli set for the Test phase of the main experiment (Experiment 7), described in the next sections.

5.3. Experiment 7 - Association uniqueness

5.3.1. Method

5.3.1.1 Participants

Sixty-six students at the University of York (age range: 18-23) participated in exchange for either course credit or payment. The number of participants was informed by the study of interest, Pufahl and Samuel (2014), that tested the following number of participants in their experiments: 72 (Exp.1, 64 included), 73 (Exp.2, 64 included), 65 (Exp.3, 64 included), 52 (Exp.4, 48 included), 23 blind adults (Exp.5, 19 included), and 51 (Exp.6) participants.[54] All participants provided written consent prior to the experiment. They all identified themselves as native-speakers of British English and none of them reported a history of hearing or speech and language related problems. None of the participants had done Experiment 6 and any of the other previous experiments.

5.3.1.2. Materials and Design

The stimuli consisted of : 1) the unfiltered word-sound pairs for the exposure phase that were identical to the ones used in Experiment 6, and 2) the filtered word-sound pairs, with the assorted filtered versions of the words selected from the piloting in Experiment 6. The experimental design was identical to the one in experiment 6, involving the same counterbalancing groups and experimental phases. Like in Experiment 6, participants were encouraged to pay attention to all the auditory information in a trial, but in addition, they were told that there would be some questions regarding the background sounds after the experiment. Hence, the test phase was followed by a brief questionnaire (Appendix D). The answers to the questionnaire were not considered in the analyses.

5.3.1.3. Procedure

The same procedure and tasks as in Experiment 6 were involved in the exposure and test phases, and their respective durations were similar to those in Experiment 6. The additional questionnaire part lasted between 10 - 15 minutes.

5.3.2. Results

Participants displayed accuracy above 90% in the Exposure phase, indicating that they had successfully encoded the words during the task. A one-way repeated measures ANOVA with Animacy (2 levels) as the within-subjects factor, revealed a small, yet significant difference in performance with respect to the semantic category of the words:

$M_{Animate}$ = 97.20 % correct, $SD_{Animate}$ = 2.97; $M_{Inanimate}$ = 98.98 % correct, $SD_{Inanimate}$ = 1.59; $F(1,65) = 17.55$, $\eta^2 = .21$, $p < .001$.

Only the transcribed responses in the Test phase of the experiment were included in the main analysis. The dependent variables, Accuracy (word identification accuracy), was coded as a binary

[54] The relevant experiment for the present purpose in Pufahl and Samuel (2014) is Exp.1, the one that reported the sound specificity effect.

variable with 1 for correct and 0 for incorrect. As in Experiment 6, spelling mistakes were judged as correct, whereas words that appeared similar in form to the target word(s), but belonged to a different grammatical and/or semantic category, were considered incorrect. Like in the previous experiments, the data were analyzed using generalized mixed-effects regression models (GLMER) for the binary Accuracy variable (Baayen, Davidson, & Bates, 2008). The fixed factors consisted of the same ones as in the previous experiment(s): Sound Sameness (same or different sound at test), Word Semantics (animate or inanimate word at test), Exposure Sound Exemplar (exemplar A or B at exposure), plus an additional one: Exposure Sound Semantics (i.e., whether the paired sound in exposure was animate or inanimate). Note that since the sound change from exposure to test involved only the change in exemplar and as such, occurred within the same sound category, the semantic class of the sound remained the same in both exposure and test phases (i.e., animate/inanimate in both phases). Therefore, there was no need to have a separate factor for the semantic class of the sound at test. Similarly, given that there was a fixed factor for the sound exemplar at exposure (Exposure Sound Exemplar) and one for whether this exemplar was the same/different one at test (Sound Sameness), there was no need for including a separate factor for the paired sound exemplar at test.

The factors were coded as binary variables in the following way: Sound Sameness: 1 (same), -1 (different); Word Semantics: 1 (animate), -1 (inanimate); Exposure Sound Exemplar: 1 (exemplar A), -1 (exemplar B); Exposure Sound Semantics: 1 (animate), -1 (inanimate).

Prior to adding any fixed factors to the base model, we tested the maximal random structure of the model for each dependent variable, consisting of random slopes of all the fixed factors and random intercepts for subjects and items. This was done to see whether adding random slopes for the fixed factors was necessary. For the main factor of interest, Sound Sameness, slopes were added for both subjects and items, whereas for the other three factors, only by-subjects slopes were added. The maximal random structure for the Accuracy variable converged, but it was not statistically different from the basic structure consisting of only the random intercepts for subjects and items, χ2(5) = 1.39, p = .93, indicating that adding the random slopes to the model(s) was not necessary. Nevertheless, following Barr et al (2013)'s suggestions, we used the maximal random structure whenever it converged with the added fixed factors.

The fixed factors, as well as their interactions, were added incrementally to the base model, and improved fit to the model was assessed using the likelihood ratio test. The base model included only the random terms. The main effects of the factors were obtained by testing the improvement in the model fit when each one of these factors was individually added to the base model.

With respect to the main factor of interest, Sound Sameness, there was no significant decrease in the mean word identification accuracy as a result of the change in the paired sound exemplar from exposure to test, $M_{Acc\_Same\ Sound\ Exemplar}$ = 74.92 % correct, SD = 8.53; $M_{Acc\_Different\ Sound\ Exemplar}$ = 74.09 % correct, SD = 9.65; β = .02, SE = .03, χ2(1) = .34, p = .56. Therefore, contrary to the initial prediction, we did not replicate the respective finding in Pufahl and Samuel (2014).

No main effect of the semantic category of the word at test (Word Semantics) on Accuracy was found: β = -0.02, SE = .1 , χ2(1) = .05, p = .82. While there was a slight advantage in identifying inanimate words, the difference was not significant: $M_{Acc\_Animate}$ = 74.02 % correct, SD = 10.04; $M_{Acc\_Inanimate}$ = 75 % correct, SD = 9.29. Additionally, there was no interaction between the semantic category of the word and the sound sameness factor: β = .004, SE = .03, χ2(1) = .01, p = .91.[55] Hence, participants were not more accurate at identifying inanimate words compared to animate ones, regardless of whether they were repeated with the same background sound exemplar or not. The mean accuracy values for each combination of the factors are displayed in Table 5.6.

Table 5.6. Mean word identification accuracy values (% correct) for each combination of Word Semantics x Sound Sameness.

Word Semantics x Sound Sameness	Same Sound	Different Sound
Animate Word	74.54%	73.48%
Inanimate Word	75.30%	74.70%

With respect to the Exposure Sound Exemplar (A vs. B) factor, no main effect on Accuracy was observed: β = -0.02, SE = .06, χ2(1) = .13, p = .72. The mean word identification accuracies for each of the exposure sound exemplars were: $M_{Acc\_Exposure\ Exemplar\ A}$ = 74.13 % correct, SD = 8.76; $M_{Acc\_Exposure\ Exemplar\ B}$ = 74.89 % correct, SD = 7.72. Further, there was no interaction between the exposure sound exemplar and sound sameness, β = .02, SE = .03, χ2(1) = .49, p = .48.[56] Therefore, what sound exemplar the words were first heard with during exposure did not matter for either the word identification performance at test, or (the lack of) a sound specificity effect. The mean accuracy values for each combination of the factors are displayed in Table 5.7.

Table 5.7. Mean word identification accuracy values (% correct) for each combination of Exposure Sound Exemplar x Sound Sameness.

Exposure Sound Exemplar x Sound Sameness	Same Sound	Different Sound
Exemplar A	74.85%	73.41%
Exemplar B	75%	74.77%

[55] Only by-subject random slopes for Sound Sameness and Word Semantics factors were included in the respective models, because the addition of the other slopes led to the models' failure to converge.

[56] Only the random slope of Sound Sameness was added to Subjects, since the models involving the other random slopes failed to converge.

Similarly, no interaction was found between the exposure sound exemplar and the semantic category of the word at test (Word Semantics), β = .005, SE = .04, χ2(1) = .02, p = .89.[57] The slight advantage in identifying inanimate words more correctly than their animate counterparts at test, was not affected by what sound exemplar the words were paired with in the exposure phase. The mean accuracy values for each combination of the factors are shown in Table 5.8.

Table 5.8. Mean word identification accuracy values (% correct) for each combination of Exposure Sound Exemplar x Word Semantics.

Exposure Sound Exemplar x Word Semantics	Animate Word	Inanimate Word
Exemplar A	73.71%	74.54%
Exemplar B	74.32%	75.54%

Last, there was also no main effect of the Exposure Sound Semantics (animate vs. inanimate) on Accuracy: β = .03, SE = .1, χ2(1) = .1, p = .75. The mean word identification accuracies for each of the exposure sound semantic category were: $M_{Acc\_Exposure\ Animate\ Sound}$ = 74.81 % correct, SD = 8.95; $M_{Acc\_Exposure\ Inanimate\ Sound}$ = 74.20 % correct, SD = 9.66. Additionally, no interaction between this factor and the sound sameness (Sound Sameness) was observed, β = .008, SE = .03, χ2(1) = .05, p = .82.[58] More specifically, the animacy of the sound the words were first heard with during exposure did not matter for either the overall word identification performance at test, or (the lack of) a sound specificity effect. The mean accuracy values for each combination of the factors are displayed in Table 5.9.

Table 5.9. Mean word identification accuracy values (% correct) for each combination of ExposureSound Semantics x Sound Sameness.

Exposure Sound Semantics x Sound Sameness	Same Sound	Different Sound
Animate Sound	75.23%	74.39%
Inanimate Sound	74.62%	73.79%

Similarly, no interaction was found between the semantic category of the exposure sound (Exposure Sound Semantics) and the semantic category of the word at test (Word Semantics), β = -0.08, SE = .09, χ2(1) = .77, p = .38. Hence, the slight advantage in identifying inanimate words more

[57] Only random slopes of Word Semantics and Exposure Sound Exemplar were added to Subjects only, as the models involving the rest of the slopes as well, failed to converge.

[58] Random slopes were added only for the following variables: Sound Sameness (to both Subjects and Items), and Exposure Sound Semantics (only Subjects).

correctly than their animate counterparts at test, was not affected by the animacy of the sounds the words were paired in the exposure phase. The mean accuracy values for each combination of the factors are shown in Table 5.10.

Table 5.10. Mean word identification accuracy values (% correct) for each combination of Exposure Sound Semantics x Word Semantics.

Exposure Sound Semantics x Word Semantics	Animate Word	Inanimate Word
Animate Sound	72.73%	76.89%
Inanimate Sound	75.30%	73.11%

5.3.3. Discussion

This experiment examined the emergence of a sound specificity effect in the presence of a unique word-sound pairing context. The context uniqueness element was realised by pairing every word with one of two exemplars belonging to a unique environmental sound category. The environmental sounds were from the same set used in Pufahl and Samuel (2014), that also implemented the same unique pairing method in their stimuli. The sound change from exposure to test occurred between exemplars of the same sound category (e.g., dog A vs. dog B; guitar A vs. guitar B). Despite using the same sound set, a similar design with the same encoding task and implicit memory test as Pufahl and Samuel (2014), we did not replicate their sound specificity effect. Possible reasons for this failure to replicate are discussed in the next section.

5.4. Discussion and conclusions

This chapter explored another potential context for the emergence of a sound specificity effect, one incorporating an “association uniqueness” factor in the word-sound pairs. We were interested to see whether unique pairings between words and sounds would promote better encoding in memory than non-unique pairs. The latter type of association was present in the stimuli of our previous experiments that involved only two background sounds. These studies demonstrated that when only two background sounds merely co-occur with the words (Exp. 2A and 2B), no sound specificity effect is observed. It was only when another element was added to the stimuli, such as a high masking contrast on the same word (Exp. 3), or integrality between the words and sounds (Exp. 4), that an effect appeared.

To increase the likelihood of observing an effect, I adopted a similar experimental design to Pufahl and Samuel (2014)’s first experiment, and also used their set of environmental sounds. In this respect, the present study served two main purposes: investigating the role of association uniqueness in the appearance of a sound specificity effect, and attempting to replicate the sound specificity effect observed in Pufahl and Samuel (2014).

However, contrary to my prediction, I did not find a sound specificity effect. Below are a few possible explanations as to why this may have occurred:

I. The 'one-to-one' association in the word-sound pairs, together with the within-category change may have led to too many background sounds overall, perhaps increasing the uncertainty in the experimental context to a higher-than-tolerated level. Part of the original reasoning was that high stimulus uncertainty and unpredictability could potentially lead to an increased reliance on contextual cues, and hence promote the appearance of a sound specificity effect. However, perhaps if the uncertainty is too high, as it might have been the case here, it may increase the overall noise level in the experimental system, which could in turn lead to the lack of an effect.

II. The diversity in the nature of the sounds, while promoting association uniqueness, may have also served as a distracting element for the participants. The uneven familiarity of the sounds (e.g., dog barking vs gorilla roaring) may have contributed towards participants paying more attention to some sounds than others. Thus, instead of integrating the word and sound as a pair, directed attention towards the unfamiliar and/or alerting sounds may have contributed to the perceptual segregation of the respective pairs.

III. The task at test involved heavy filtering of the stimuli, perhaps making the task more like an word intelligibility puzzle-solving, than a memory task per se. Instead of implicitly accessing the episodes encoded during exposure, participants may have focused primarily on figuring out the word and purely transcribing it. Furthermore, the episodes the listeners encoded in the exposure phase involved word-sound pairs in clear listening conditions, and as such did not match in format with the episodes heard in the test phase. This may be problematic from a theoretical point of view. The *transfer appropriate* approach to memory (reviewed in **Chapter 2**) emphasises the " processing match" between encoding and test in implicit tasks. More specifically, it maintains that the same type of processing should be encouraged in both encoding and test, rather than in only one of the phases. For example, if the implicit task at test is word identification in noise, then the study task should also include these items presented in noise, rather than in clear (Roediger, 1990). Sheffert (1998b) demonstrated that the "goodness of the processing match" between encoding and test was the primary determinant on the attainability of indexical effects on implicit tests that involve perceptual identification (study reviewed in detail in **Chapter 2**). From this perspective, using an implicit memory task that does not require degrading the stimuli, or an explicit recognition task could have increased the likelihood of the appearance of a sound specificity effect. In both cases, the episodes heard during exposure would match in the presentation format (i.e., both heard in the clear) with those heard at test.

Although the above factors may have contributed the lack of a sound specificity effect, they do not explain the difference between our results and Pufahl and Samuel's. However, it is worth considering some small, but potentially consequential differences between the two designs:

I. Although similar, the stimuli were not exactly the same. The same sound set was used, but the words were different. Since the words were the primary focus of the respective tasks in exposure and test, different word sets may have played a role in this inconsistency in observing a sound specificity effect.

II. The word-sound associations were different. In both experiments, a word was randomly paired with a unique sound category. Different sound pairings meant that the two experiments could have contained different idiosyncrasies. .

III. The experimental design was similar, but not exactly the same. In the Pufahl and Samuel study, there were two independent variables (talker's voice and the paired sound) and 4 conditions in which these variables were manipulated from exposure to test: 1) no change, 2) voice change only, 3) sound change only, and 4) both voice and sound change. In contrast, there was only one independent variable in our study, the paired sound, and hence only two conditions in which it was manipulated from exposure to test: 1) no change, and 2) sound change. Perhaps the additional change in an another dimension of the stimuli (indexical) made the listeners more prone to paying attention to and/or detecting contextual change.

IV. Our explicit instruction that asked participants should pay attention to the sounds as well was not present in Pufahl and Samuel's study. While this instruction aimed at promoting the integration of the words and sounds as pairs, by expecting participants to pay attention to the association (e.g., the word "tiger" is paired with an harmonica sound), it may have had the opposite effect since, as explained above, directed attention to the sounds may have facilitated their perceptual segregation from the words.

In conclusion, the insignificant result observed in this chapter is in sharp contrast with the main finding of Pufahl and Samuel (2014)'s first experiment. It is nevertheless an informative result, in that it confirms that speech-extrinsic specificity effects are fragile and are conditional on the context in which they are probed. While the existence of sound specificity effects (Creel,Aslin, & Tanenhaus, 2012; Pufahl and Samuel, 2014; Cooper, Brower, & Bradlow, 2015) is constraining current models of spoken word recognition, the fragility of these effects makes their status in long-term memory far from clear. In the same way that the inclusion of indexical properties in the lexicon is still being debated, further research is necessary to determine whether speech-extrinsic auditory properties can be incorporated within the lexicon. Before more compelling and conclusive evidence emerges on this issue, models of the lexicon will have to be cautious in incorporating such information in lexical representations. I discuss this topic, along with others, in detail, in the next and final chapter.

Chapter 6

General Discussion

6.1. Introduction

In the last chapter of this thesis, I will engage in discussing its major findings, their implications for theoretical approaches, as well as potential limitations and directions for future research. Similar to the comparative approach between sound and voice specificity effects undertaken in the previous chapters, the discussion will continue to draw parallels between the two, in an attempt to better understand the implications of the present work for the broader debate about specificity effects. The next section provides an overview of the main findings reported across the previous chapters.

6.2. Summary of findings

I. Although the focus of the thesis was the sound specificity effect, a comparative perspective to the well-established voice specificity effect was presented. Therefore, I started the investigation into the sound specificity effect with the successful replication of this effect in **Chapter 2**. An interesting aspect of this replication was that it occurred with a recognition memory task, which being typically classified as explicit in nature, is thought to be less efficient in revealing indexical effects compared to implicit tasks (a detailed review of the debate surrounding this issue was provided in **Chapter 2**, but see also Pufahl and Samuel, 2014).

II. I then proceeded with an examination of the sound specificity effect from several perspectives. In **Chapter 3**, I explored an account focused on the masking component that is inherent in the pairing of a word with a background sound. More specifically, I was interested to see whether it was the word-sound associations that were retained in memory, or the degraded version (acoustic glimpses) of the words as a result of masking by the paired sounds. To this end, I used two sounds with the same intermittent pattern, but different fundamental frequencies (high vs. low pitch) as maskers and manipulated the energetic masking on the words from exposure to test. I found that when the change in the paired sound involved only the frequency dimension of the sounds (i.e., pitch), such that the resulting acoustic glimpses were broadly similar, there was no sound specificity effect on the overall word recognition accuracy. However, adding a temporal contrast to the change in the sound pitch led to highly contrasted glimpses of the same word(s), which in turn elicited a sound specificity effect on word recognition memory.

III. In **Chapter 4**, I took a closer look at the analogy with the voice specificity effect and the intrinsic relationship that exists between the words and voices in spoken utterances. More specifically, I focused on the fact that this relationship has two crucial components: *co-occurrence* and *integrality* between the words and voices. In the existing studies regarding the sound specificity effect, only the “co-occurrence” element had been explored. I reasoned that if the sounds were integral to the words, hence making the relationship between the word-sound pairs resemble more the intrinsic relationship between words and voices, a sound specificity effect would appear. To create integrality, the sounds were modulated according to the temporal intensity envelope of each individual word and these modulated maskers were then paired with the corresponding words. As expected, there was a relatively robust sound specificity effect. Crucially, this effect disappeared when the integrality element was removed from the stimuli and the words and sounds only co-occurred

IV. In **Chapter 5,** I explored the role of the *uniqueness* factor in the association between words and sounds regarding the emergence of a sound specificity effect. Contrary to the previous studies where only two background sounds were involved, there were as many sound categories as there were words. Hence, there was a “one-to one” association between a word and a sound, such that a word was paired with only one sound category and that sound category was not paired with another word. The study was also in part motivated by the sound specificity effect originally reported by Pufahl and Samuel (2014), that involved the same type of association between the words and sounds Surprisingly, a sound specificity effect was not found. This finding is indicative of two things: 1) the pairwise association uniqueness between words and sounds does not always seem to play a role in the appearance of sound specificity effects, and 2) these effects are fragile and may not be replicated easily, despite similarities in the methodology used.

The results across experiments, their respective effect sizes and the observed power in each experiment are summarised in Table 6.1.[59]

Table 6.1. The effect sizes and the observed power in each of the experiments described in the thesis

Experiment	Sample Size	Independent Variable	Effect Targeted	Effect Found?	Effect Size *r*	Effect Size Cohen’s *d*	Partial Eta Squared (η^2)	Observed Power
1	48	Voice change	Voice specificity	Found	0.22	0.46	0.18	0.87
2A	55	Sound pitch change-Late masking	Sound specificity	Not found	0.004	0.08	0.004	0.08
2B	46	Sound pitch change-Early masking	Sound specificity	Not found	-0.02	-0.05	0.003	0.07
3	55	Sound pitch and position change	Sound specificity	Found	0.15	0.29	0.09	0.62
4	48	Sound change-Integrality	Sound specificity	Found	0.19	0.39	0.16	0.82
5	46	Sound change-No integrality	Sound specificity	Not found	0.01	0.03	0.001	0.06
7	66	Sound exemplar change-Unique pairs	Sound specificity	Not found	0.05	0.09	0.01	0.13

[59] The observed power and partial eta squared values are from the repeated measures analysis of variance (ANOVA) conducted across subjects, with the respective IVs (voice/sound change) as the within-subjects factor. The *r* and *d* values for the effect sizes estimation were obtained via an online calculator (http://www.uccs.edu/~lbecker/), provided by Dr. Lee A. Becker, University of Colorado Colorado Springs (UCCS), with the respective means and standard deviations entered as input.

Based on Cohen (1988)'s heuristics for interpreting effect sizes[60], the specificity effects observed (sound/voice) can be placed on a small-to-medium size range. Namely, the "voice specificity" effect (Exp.1) qualifies as a small-to-medium effect ($|.1| < r < |.3|$; $|.2| < d < |.5|$). The sound specificity effect obtained in the case of contrasted glimpses of the same word(s) (Exp.3) represents a small effect ($|r < |.3|$; $d < |.5|$). Finally, the sound specificity effect obtained in the presence of integral sounds (Exp.4) represents a small effect (approaching medium size) ($|.1| < r < |.3|$; $|.2| < d < |.5|$). Regarding the power that each experiment had to detect the targeted effect, from the observed power values, it seems that Experiment 1 and Experiment 4 had enough statistical power (observed power > .8) to detect their respective effects. The contrasted glimpses experiment (Exp.3) revealed an effect, despite the observed power being lower than the typical threshold (.8). In all the other experiments that did not reveal an effect, the observed power seems quite low. However, this is not surprising, given that the observed power of a study is based on the significance level and the effect size observed in that study. As several researchers have pointed out, observed power values in the case of insignificant effects can be uninformative and misleading, since these effects always correspond to low observed power estimates (e.g., Goodman & Berlin, 1994; Hoenig & Heisey, 2001; O'Keefe, 2007). Further, the sample size was comparable across experiments (significant and insignificant results) and the targeted sound specificity effect was consistently small in size whenever present. Hence, there seems to be a relatively slim chance that the failure to reveal this targeted effect in the experiments with insignificant results (that tested comparable numbers of people with the experiments that revealed the effect in question) is due to insufficient power, rather than due to the manipulation associated with the independent variable. The pattern of appearance/lack of this effect across the experiments in question is an indication of its fragile and unstable nature. Nevertheless, it remains true that the experiments with insignificant results lack the power to distinguish between the possibility that the null hypothesis is true (the targeted effect is not present) and the possibility of a Type-II error (targeted effect present, insufficient power to detect it).

Taken together, the results of this thesis highlight the general observation that the sound specificity effect is contingent on the experimental context in which it is being probed, a statement that can be decomposed into the following main themes:

- The sound itself may not necessarily need to be encoded in memory alongside the word for a specificity effect to emerge.

- Mere co-occurrence between words and sounds may not be always sufficient for the appearance of a sound specificity effect.

- The uniqueness factor in the pairing between words and sounds does not seem to play a critical role in the emergence of a sound specificity effect.

I will discuss each of these themes in more detail in the following section.

6.3. Sound specificity effects and spoken word recognition - Main themes

The discovery of a novel effect, like the sound specificity effect, brings along several questions, the most important one being: how genuine is it? The most convincing way to test whether an

[60] Cohen (1988) suggested that an r of $|.1|$ represents a 'small' effect size, $|.3|$ represents a 'medium' effect size and $|.5|$ represents a 'large' effect size. Similarly, $d = 0.2$ represents a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size.

effect is real or not is by replicating it. Hence, the crucial question regarding the novel sound specificity effect is then: can it be replicated? I examined this question from several perspectives and what I found suggests in a nutshell, a conditional "yes" answer. The longer answer depicts a more complex picture, of course, and revolves around the themes outlined in the previous section.

6.3.1.Theme 1: Sounds may not be necessarily encoded in memory - The case for highly contrasted acoustic glimpses

This theme emerged from the results in **Chapter 3**, which highlighted the idea that the sound itself may not always need to be encoded alongside the word in memory for a specificity effect to emerge. The masking component that the pairing between words and sounds brings along is an important one and needs to be factored in when investigating specificity effects that arise as a result of changes in the paired sound(s). The series of three experiments demonstrated that given the proper energetic masking configuration, it is the acoustic left-overs (glimpses) of the word(s) and not necessarily the word-sound associations per se, that could be encoded in memory upon first encounter during exposure. However, this type of encoding is conditional on the type and amount of masking exerted by the sound. Namely, a change in the frequency domain of the sound (pitch) was not sufficient in eliciting an effect, despite the co-occurrence factor between the words and sounds. As the computational analysis revealed, the resulting glimpses of the same word(s) were quantitatively different, but qualitatively speaking, still quite similar, given that the masking occurred in the same temporal regions of the word. It was only when a temporal contrast in the overlap between the word and sound between exposure and phase was added to the change in the sound pitch that a small sound specificity effect emerged. In this case, the resulting acoustic glimpses of the same word were both quantitatively and qualitatively different, with masking from the two sounds taking place in relatively different areas of the spectro-temporal domain of the word.

- How reliable is the effect of contrasted glimpses?

The evidence at hand points to a week effect. The glimpse account makes an interesting case for the role of energetic masking in sound specificity effects and reveals it as a plausible alternative factor to mere co-occurrence. By highlighting the possibility that the degraded versions of the words may be retained in memory, it challenges to a certain degree the notion that the emergence of a sound specificity effect necessitates the presence of the word-sound association, and hence, of the sound itself, in memory. In the right energetic masking context, it may be the degraded version of the word from the masking of the sound that might persist in the memory episode of the word, rather than the word-sound pair itself. However, the context-selective nature of this effect restrains the scope of claims that can be made with regard to its implications. The glimpses of the same word had to be highly contrasted in both dimensions of interest (frequency and time) for a relatively small effect to emerge. Even in this context, it is possible to argue that the observed sound specificity effect may not be due to the contrasted glimpses, but rather due to the presence of sufficient cues attached to a given word from the paired sound, that could have in turn strengthened the memory episode of the word. Further, the "glimpses in memory" hypothesis was not directly tested, since in all the experiments involved, the sound(s) still co-occurred with the word(s). A more direct and stronger case for the presence of the degraded versions of the words in long-term memory would have been possible by having an additional experiment in which only the degraded versions (glimpses) of the words were used as stimuli, without the explicit word-sound associations. If the effect persisted in that context as well, then it would strengthen the case for the " glimpses in memory" hypothesis, as visualized in Figures 6.1 A and B. Therefore, as it currently stands, the evidence in favour of the "glimpses in memory" hypothesis is not stronger than the evidence in favour of the "sounds in memory" hypothesis that previous studies have put forward (e.g., Pufahl and Samuel, 2014).

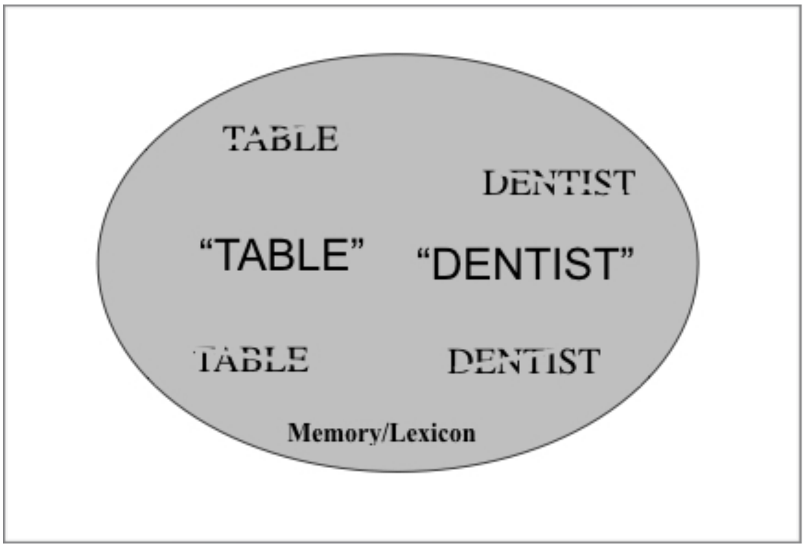

Figure 6.1A. A visualisation of the presence of only the degraded versions of words (glimpses) in the memory system, as a result of masking by sounds, without including the sounds per se. No specialized memory structures or boundaries are depicted in this case, indicating a relatively unrestrained co-existence between purely lexical and episodic information in long-term memory.

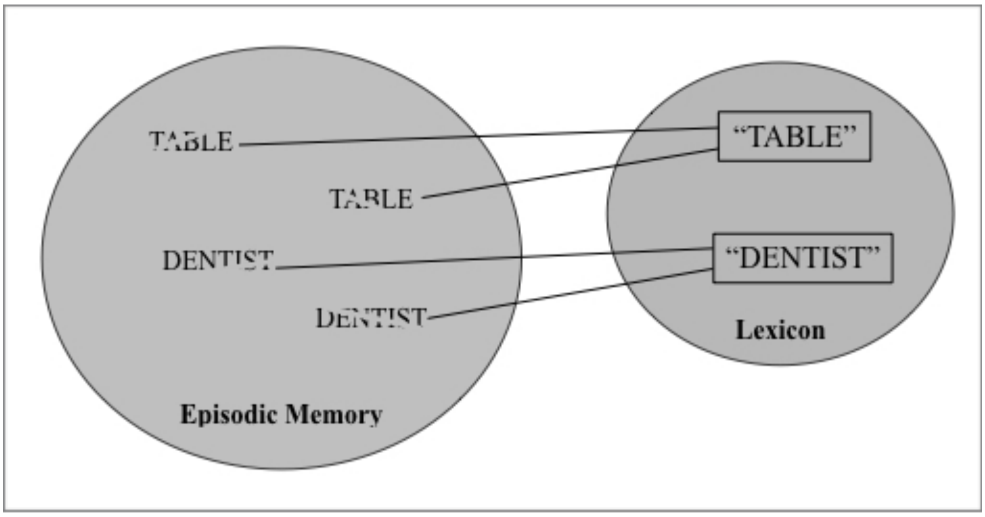

Figure 6.1B. A visualisation of the presence of only the degraded versions of words (glimpses) in the memory system. In this case, the memory system is more structured than in Fig.6.1A, including more specialised memory components and possible interactions between them.

Each illustration depicts a possible scenario in which the representations of the glimpses of words could be incorporated in memory. The first one (Figure 6.1A) does not impose any organization of the memory system into specialised components, but displays a relatively loose version of it, that can accommodate both the glimpses (episodic in nature) and the corresponding abstract representations.

In the second scenario (Figure 6.1B), the memory system is structured into an episodic component that contains the glimpses of the words and a lexical component that hosts the abstract lexical representations of the corresponding words. The arrows illustrate interactions between the two components. The issue of drawing boundaries between the auditory episodes and abstract lexical representations depends on the theoretical view endorsed regarding the organization of the memory system in general, and that of the lexicon, in particular. It is worth noting that even if a certain theoretical position is endorsed, a clear-cut definition of what constitutes purely lexical memory and purely episodic memory, as well as potential interactions between the two, has yet to emerge in the relevant literature. The theoretical impact of the effect will be discussed in the next section. The main purpose of the visual illustrations is to show how the memory system can accommodate the presence of degraded versions of spoken words without necessitating the inclusion of the masking sounds.

6.3.2. Theme 2: Mere co-occurrence is not always sufficient for a sound specificity effect - The case for integrality

Based on Chapter 4, this theme brings along another novel factor that seems to play a role in the emergence of a sound specificity effect, the *integrality* between words and sounds. Taken together, the findings of the two experiments make a compelling case for its role in the appearance of a sound specificity effect. Interestingly, the effect was manifested in both dependent variables, word recognition accuracy and response latency, and was comparable in size to the voice specificity effect Given that it is impossible to perceptually segregate the voice from the linguistic content in a spoken utterance, it is compelling to reason that implementing a degree of integration between words and sounds leads to a comparable specificity effect. The harder it becomes for the sounds to be segregated from the words, the easier it is to perceive the pair as a blended, integrated auditory item.

The integrality effect is reminiscent of the Gestalt *common fate* principle of grouping (e.g., Wertheimer, 1923; 1938). This relates to work by Bregman and colleagues, who adapted the principle to the auditory domain in order to provide a plausible account for how the auditory system analyses auditory scenes consisting of multiple elements, or “streams” of information (Bregman, 1990). What is particularly relevant here is the fact that the adaptation of the *common fate* principle concerns changes/manipulations in the sound over time, with the heuristics being: If different parts of the spectrum change in a correlated way, they are bound together into a common perceptual unit (Bregman, 1990).

In the case of integral words and sounds, the common fate heuristic is a domain-general principle that could easily explain the effect we observed. Co-occurring words and sounds constitute two different auditory "objects" that, in normal conditions, can be segregated with relative ease, as demonstrated by some of the results in **Chapter 3.** However, when modulated to undergo the same changes over time, apparently these two objects blend perceptually to form a relatively unified object, which in turn may promote a similarly unified encoding in memory.

It is worth noting that the “integrality effect" found in Experiment 4 supports the idea that sounds are encoded in memory. In this respect, it is consistent with the main claim made in Pufahl and Samuel (2014), and in contrast with the “glimpses hypothesis” supported by Experiment 3.

However, it is also an indication of the conditional nature of this encoding, given that the sound specificity effect disappeared once the integrality element was removed from the stimuli in Experiment 5.

To conclude this theme, I provide a set of visualisations illustrating several scenarios in which integral sounds could be incorporated into the memory system, as displayed in Figure 6.2A-D. The first scenario in Figure 6.2A illustrates the case where the integral sounds are encoded in episodic memory and interact with the respective abstract representation of the word in the lexicon.

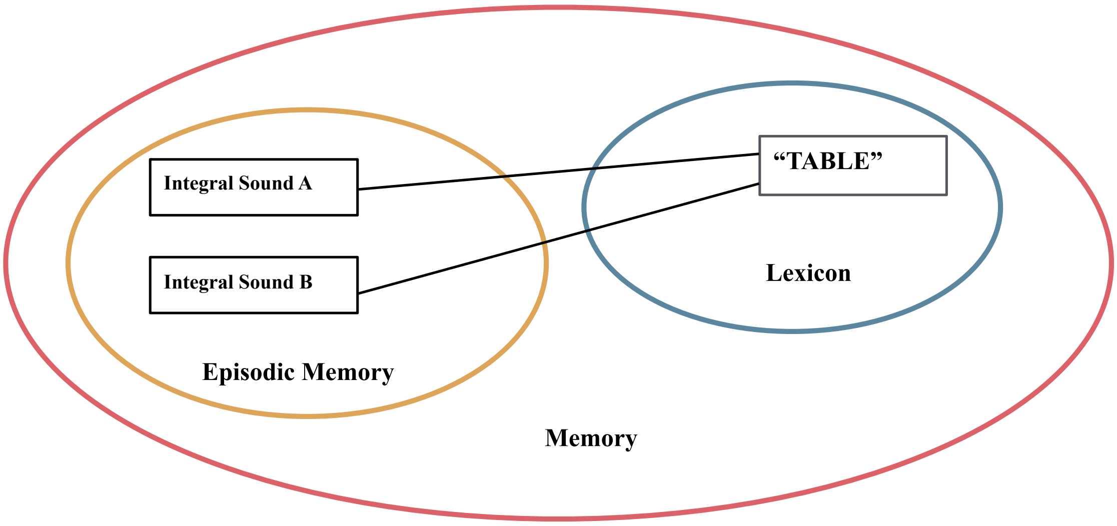

Figure 6.2A. A visualisation of the associations between the integral sounds in episodic memory and the abstract representation of a word in the lexicon. The auditory episode contains only the integral sounds, without instances of the word.

The scenario in Figure 6.2B depicts the case where integral sounds in episodic memory are directly associated with episodic instances of the word, which in turn are linked to the corresponding abstract representation in the lexicon. The lexicon in this case involves both abstract representations of the words and their episodic traces.

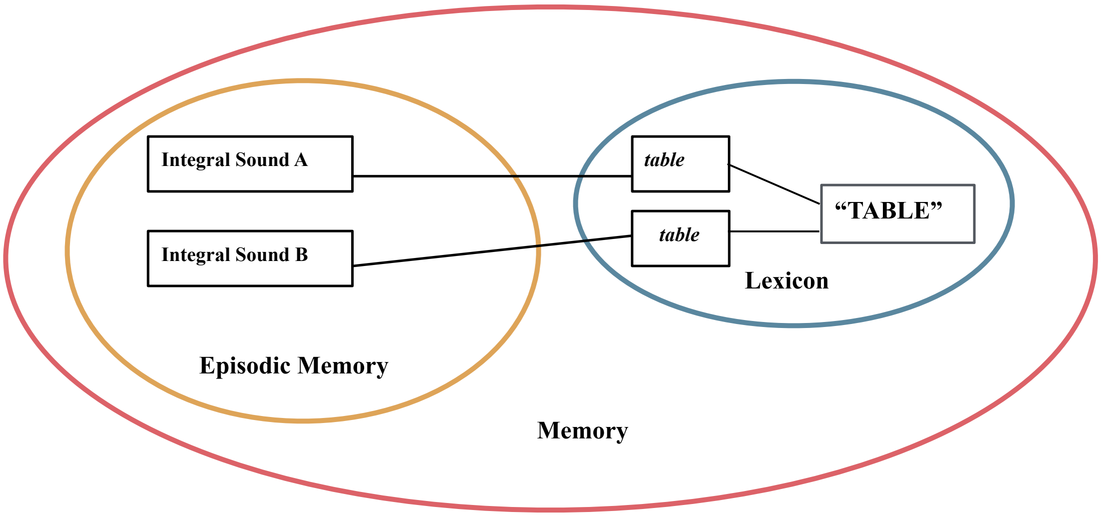

Figure 6.2B. A visualisation of the associations between the integral sounds in episodic memory and the abstract representation of a word in the lexicon. The auditory episode contains the episodic trace of integral sounds that interact first with non-lexical instances of the word, which in turn map to the corresponding abstract representation.

In Figure 6.2C, the auditory episode involves a non-lexical instance of the word paired with the integral sounds. The non-lexical nature of the word instance means that it is just an auditory trace that could be very short-lived, a detail that is depicted in light grey . The episodic instances of the word have associations with the corresponding abstract word representation in the lexicon. The last scenario in Figure 6.2D illustrates the case where the integral word-sound associations are included

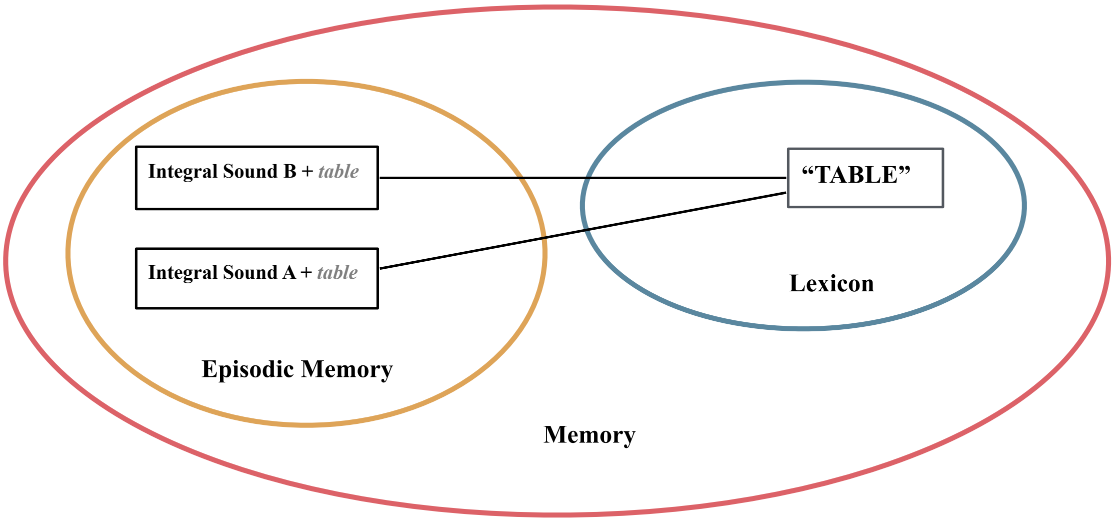

Figure 6.2C. A visualisation of the associations between the integral sounds in episodic memory and the abstract representation of a word in the lexicon. The auditory episode contains the episodic trace of integral sounds and non-lexical instances of the word. The latter represents only an auditory trace, that could be very short-lived.

in a so-called "episodic lexicon" within episodic memory, that may also contain other episodic instances of the word(s). The general episodic memory might involve non-integral versions of the sounds as well. The episodic instances of the word are connected to the respective abstract representation in the so-called "abstract lexicon".

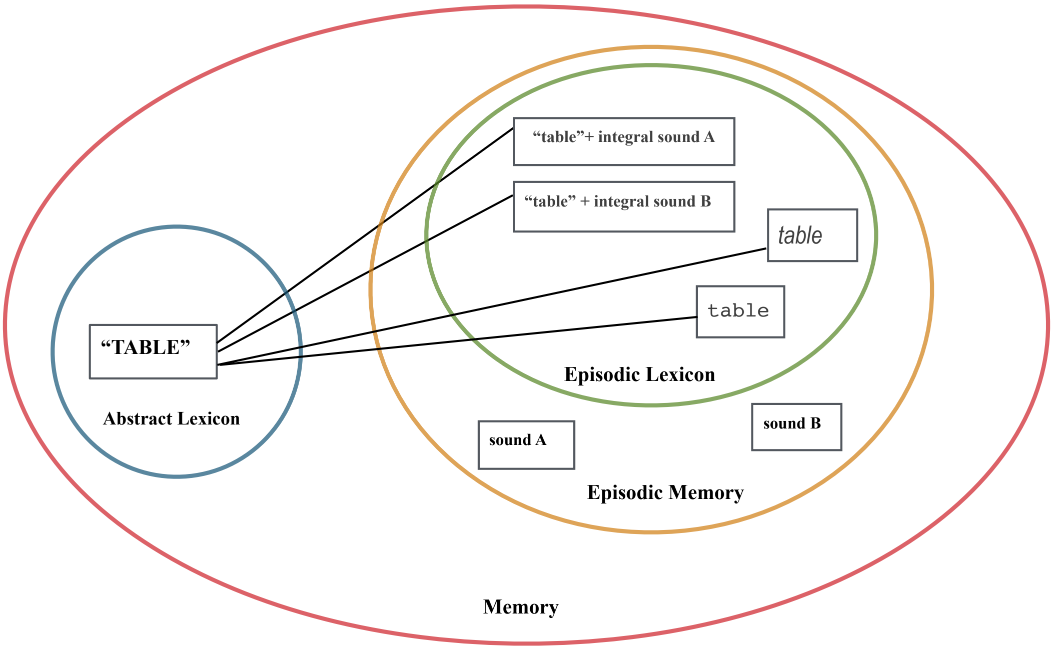

Figure 6.2D. A visualisation of the interactions between the auditory episodes consisting of integral word-sound pairs in a lexicon that is episodic in nature, and the abstract representation of the corresponding word in the abstract lexicon. The auditory episodes may be considered as containing the episodic trace of integral sounds and non-lexical instances of the word.

6.3.3. Theme 3: No role of the unique pairing between words and sounds in the appearance of a sound specificity effect

This theme emerged from the results in Chapter 5. The absence of a sound specificity effect when there is a unique pairing between words and sounds is puzzling, especially considering the fact that another study reported an effect in a similar context (Pufahl and Samuel, 2014). Several potential reasons for the discrepancy were discussed in Chapter 5, hence I will not repeat them here. It is worth stating that it still seems interesting to examine the role of increasing the number of paired sounds in the emergence of a sound specificity effect. It could be particularly informative if future studies determined a threshold for the number of sounds needed to elicit a relatively robust effect. Ideally, it would be more convenient to use highly familiar environmental sounds, that are not too masking and too alerting, as to distract the perceptual pairing with the word during encoding. There are several studies that have examined indexical effects using more than two speakers (e.g., Nygaard et al., 1994;). Therefore analogous studies with multiple sounds could also be informative, without necessarily involving a unique, one-to-one pairing between the words and sounds.

6.3.4. Theme 4: Sound specificity effects: General contextual effects?

Sound specificity effects arise as a result of changes in the auditory context of the word. A relevant question to consider is whether these effects are simply an instance general contextual effects or another type of indexical effects. There is evidence that the change in the physical context in which words are first encountered impairs later recall performance. One classical example is the study by Godden and Baddeley (1975). In this experiment, the participants were trained divers. They listened to a list of words either on land, or 20 feet under water. Every participant was tested in each of four different combinations of exposure/test: 1) land/land; 2) land/water; 3) water/water; 4) water/land Divers recalled significantly fewer words when the context of test was different from that of exposure, compared to when the context was the same in both phases. This finding was interpreted as supporting a context-dependent memory model, in which long-term memory was sensitive to changes in the environmental context in which words were first encountered, for their subsequent recall.

An explanation of this effect can be found in the “retrieval failure theory”. There is evidence that information is more likely to be retrieved from long-term memory if appropriate retrieval cues are present (e.g., Tulving & Watkins, 1974). Tulving (1974) argued that information can be retrieved more easily if the cues present when the information was encoded are also present when it is retrieved. Tulving suggested that information about the physical surroundings (external context) and about the individual’s psychological or physical state (internal context) is stored at the same time as a targeted information is learned or encoded. During recall, if the original state/context in which the encoded information was encountered is reinstated via appropriate cues, recall is facilitated. The absence of relevant cues on the other hand leads to a retrieval failure, which in turn manifests itself in the observed recall memory impairment.

It is worth mentioning that the memory task used in the present thesis is a recognition task, and it has been argued that it is relatively insensitive to retrieval issues. Namely, early theories of recognition memory viewed the process of recognition failure for a previously encountered item as reflecting the lack of appropriate stored information in the memory system. Hence, it is not an issue of retrieval in a recognition task, but only one of making a decision after evaluating stored information in memory (for a review, see McCormack, 1972). However, Tulving (1974) argued that cue-dependent retrieval issues apply to recognition tasks as well, citing evidence in favour of this hypothesis (e.g., Tulving & Thompson, 1973). The purpose of this brief overview is to draw attention to potential parallels between the findings from the Godden and Baddeley (1975) and the main results obtained in this thesis. More specifically, the evidence from their study casts some doubts on the claim that the appearance of sound specificity effects requires the involvement of sound details in lexical representations. Apparently, long-term memory can be sensitive to a wider range of contextual changes, not necessarily pertaining to the auditory domain, such as the physical environment in which spoken words are encountered (land or water). Does this mean that these broader scale details would also need to be incorporated alongside the words in the lexicon? The present evidence can only suggest the impact of contextual information in the word encoding and retention in memory at some level. However, whether this level is the lexical one seems highly arguable.

So, to tentatively answer the question posed at the beginning of this theme:

- Yes, sound specificity effects could be considered as a type of general contextual effects, rather than another type of indexical effects, unless compelling evidence that helps to clearly dissociate between the two alternatives arises.

The fact that they appear as a result of experimental manipulations and tasks typically used to measure indexical effects is not enough to readily put them in the same category as the voice effects. The main results of this thesis, supported by the general literature on indexical effects, highlight a major discrepancy between the two types of effects: context-sensitivity. Sound effects seem more susceptible to the experimental context in which they are probed than do indexical effects. In addition, there is other evidence in the literature that suggests differences in the processing of voice information and that of other non-vocal sounds. This last observation brings us to the next theme, in which I discuss in a comparative fashion the status of voices and sounds in the auditory and memory systems, as well as their impact in the processing and representation of spoken words in memory.

6.3.5. Theme 5: Voices and sounds in the auditory and memory systems

This theme has been present throughout the studies in this thesis. The first question to arise is:

- Are sounds like voices for the auditory and memory systems?

Based on the evidence so far, a tentative answer is “No”. Perhaps, the most obvious difference between sounds and voices is that voices are intrinsic, whereas sounds are extrinsic to the speech signal. The voice is the sole, unique carrier of speech, while co-occurring sounds are just additional, external auditory elements that happen to occasionally co-occur with speech. Second, as demonstrated by the main findings of this thesis, the fragility of the sound specificity effects and their context-dependent nature indicate a discrepancy between them and the voice effects. It is only when the integrality of between words and voices is simulated that a relatively robust and comparable sound specificity effect appears. On the other hand, as evident in the literature and also shown in the first study of this thesis, voice effects are typically more robust and less dependent on the context in which they are probed.

Another argument can be made on the basis of evidence from neuroimaging research that suggests a special status for voice processing in the brain. Namely, studies have consistently reported particular brain regions located along the superior temporal sulcus/gyrus (STS/STG), that selectively respond to human voices (e.g., Belin et al., 2000; Belin et al., 2002; Belin et al., 2004 ; Fecteau et al., 2004; Meyer et al., 2005; Stevens, 2004; von Kriegstein et al., 2003; von Kriegstein & Giraud, 2004; von Kriegstein et al., 2005). Additionally, it has also been shown that that these regions respond better to speaking voices than to non-speech human vocalisation (like laughter and cries) and other natural sounds (Belin et al., 2000). Although the increased sensitivity to verbal stimuli indicates that there is no strict functional selectivity for indexical-only properties of speech, selective activations along the STS in response to non-verbal human vocalisation compared to acoustically matched non-vocal sounds has also been observed (Belin et al., 2002). Furthermore, there is evidence suggesting that this region is species-specific, such that it responds selectively to human non-verbal vocalisation compared to animal vocalisation and other non-vocal sounds (Fecteau et al., 2004). In summary, this evidence supports the idea that human voices are selectively processed by specialised "voice" areas in the brain.

Other evidence supporting the idea that voices and sounds are processed and stored in different memory systems comes from several experiments in Pufahl and Samuel (2014). While their first experiment indicated that a change in a co-occurring sound impaired word identification performance, this effect was not symmetrical with respect to the sound identification performance. In a second experiment, using the same stimuli as in the first, they examined whether a change in a co-occurring spoken word would affect sound source identification to a similar extent that a change in a co-occurring background sound affected spoken word identification. The only difference from the first experiment was that participants had to focus on the sounds instead of the words when performing the experimental tasks in exposure and test phases. There were four different conditions with respect to the change from exposure to test: 1) no change (i.e., the same sound exemplar and the talker voice); 2) sound exemplar change; 3) voice change; and 4) both sound and voice exemplar change. The crucial finding was the absence of an effect on the overall sound source identification as a result of the change in voice. The absence of a symmetrical episodic effect persisted in two more experiments: 1) one that investigated the effect of various degrees of repetitions during exposure on the magnitude of the targeted specificity effect (the fourth experiment in the series), and 2) another one that examined the same question , but with a blind population, that presumably relies more on identifying environmental sounds (the fifth experiment in the series). It was only in the last experiment, in which they introduced a more extreme change in the auditory episode of the sound to increase the acoustic variability between the two instances heard at exposure and test, that an effect on sound identification emerged. Specifically, this enhanced episodic variability involved the change in both the voice and content of the co-occurring word. The authors interpreted this result as supporting the hypothesis that indexical and sound specificity effects result from a general mechanism that applies to all auditory inputs. This conclusion might be premature, however, since the effect emerged in only one of the experimental contexts. Rather, taken together, these results from Pufahl and Samuel (2014) are in line with what I have been arguing so far regarding the vulnerability and context-sensitivity of sound specificity effects.

The next critical question I would like to address is:

- Is there compelling evidence that the sounds are retained in the lexicon?

This question assumes the existence of a memory structure that hosts lexical representations of spoken words, for which there is considerable converging evidence across fields (see Gow, 2012 for a review). Based on the results of this thesis and the literature, the evidence that sounds are retained in lexicon is not particularly strong. Perhaps a more constructive way of addressing this issue is to try to answer the question: What would it take for the sounds to be encoded in the lexicon? First, one would have to show that sound specificity effects are widely replicable, which considering how recent these effects are and their context-sensitive nature, does not seem likely. Second, it would need to be shown that such effects persist in time and are consolidated in long-term memory the same way that new words, and to some extent voices, are. There has not been any reported empirical work so far that indicates this. Third, the processing of sounds would need to show sensitivity to the same factors that affect spoken word processing, such as frequency of occurrence, neighbourhood density, lexical competition. Evidence suggesting that any of these factors plays a role in the processing of sounds co-occurring with words has yet to start emerging.

A related question is:

- Is there compelling evidence showing that the voices are indeed retained in the lexicon? Like in the case of sounds, this question assumes the existence of a lexicon.

Typically, indexical effects have been interpreted as evidence that voices are included in the long-term lexical memory However, there is evidence , mainly from neuroimaging studies, that paints a more complex picture Such studies have repeatedly suggested that the processing of linguistic and voice information is served by at least partially dissociable neural substrates. For instance, a number of studies using working memory tasks, have indicated that words and voices are processed in parallel during an early, pre-attentive stage, but that there is a dissociation between the two at the memory encoding stage (Stevens, 2004; von Kriegstein et al., 2003, 2005; von Kriegstein & Giraud, 2004). Importantly, this dissociation is especially robust in the STS/STG "voice" area. For example, using a two-back task, Stevens (2004) observed that memory for voices relative to words activates different areas of STG: right superior and middle frontal gyri, posterior cingulate and right angular gyrus, than the memory for words relative to voices: left inferior frontal gyrus and bilateral supramarginal gyri. Other studies have observed that in 'voice compared with word' recognition tasks, dorsolateral, orbital and preorbital frontal regions, parietal regions and the cerebellum were activated in addition to the STG areas (von Kriegstein et al., 2003, 2005; von Kriegstein & Giraud, 2004). On the other hand, the 'word compared with voice' tasks, elicited activation in the left middle temporal and lingual cortices (von Kriegstein et al., 2003), hence indicating the existence of separate neural substrates for the processing of linguistic and indexical information in recognition task.

I would like to conclude this section by noting that what I have discussed so far concerns mainly the "when" question concerning the appearance of sound specificity effects. Namely, I have argued for plausible contexts that reveal such effects, and conditions that constrain when speech and sound stay together. This issue is partly independent from where the auditory episodes of spoken words are stored. Two other questions of interest are: *where* in memory these effects reside, and *how* they contribute to spoken word recognition.

6.4. Implications for accounts of spoken word recognition and the lexicon

Main models of spoken word recognition were reviewed in **Chapter 1.** In this section, I will contrast the main findings from this thesis with each account and argue that the hybrid view seems to be the one that best accommodates the data.

6.4.1. The Abstract Account

A view of spoken word recognition that includes only abstract representations of spoken words cannot accommodate the findings of this thesis. The robust voice specificity effect observed in the first study of the thesis, joins the extensive array of voice specificity effects in the literature in demonstrating that the indexical information of the speech signal is not discarded, but makes it into the processing stages and subsequently affects spoken word recognition. The sound specificity effects observed in two other studies take this claim a step further and suggest that spoken word recognition seems to be sensitive not only to speech-intrinsic changes in the surface properties of the signal, but in some contexts, to external changes in irrelevant, co-existing sounds as well.

6.4.2. The Episodic Account

In principle, the episodic view of speech recognition can accommodate the findings of this thesis. For example, in one of the classical exemplar-based episodic models, Goldinger's (1998) "Echo" model, the mapping of a word to the lexicon is conceptualised as a vector in a multi-dimensional space. These vectors could in principle be extended to include more dimensions beyond the ones that refer to linguistic and indexical information, to accommodate acoustic variability, such as that elicited by background sounds. However, the context-sensitive nature of sound specificity effects may pose a challenge to such an implementation. These exemplar models need vast memory resources (for a critical review, see Goldinger, 2007). Accommodating the additional episodic variability manifested by the sound specificity effects, in a context-constrained fashion, would potentially require even more resources. Such a requirement may in turn lead to memory storage and resource-sharing issues in the network.

Other plausible episodic models that can accommodate the present findings are connectionist models that rely on a distributed view of the mental lexicon, wherein co-activation is extended to the entire co-occurring variation available in the auditory stream (Elman, 2004;2009; Gaskell & Marslen-Wilson, 1997). For example, Elman's (2004, 2009) simple recurrent network (SRN) eliminates the need for having a lexicon altogether, and assumes a distributed representation of word knowledge in which categories emerge over time based on the distributional properties of the input that the system receives. According to this perspective, words serve as cues or pointers to the co-occurring information with which they have appeared, and can activate this information based on the frequency of their co-occurrence.

Finally, even if able to accommodate the present effects, such models would still have to face the criticisms that surround them in the literature. Namely, it has been frequently argued that episodic-only models cannot account for the ample amount of evidence that supports the abstract representations (Cutler, 2008, Goldinger, 2007; Pisoni & Levi, 2007). Therefore, the recent trend in the literature favours the need for designing hybrid models of spoken word recognition that can accommodate the co-existence of abstract lexical representations and episodic information.

6.4.3. The Hybrid Account

The results observed in this thesis cannot directly dissociate between the episodic and hybrid views, since both views accept the encoding of episodic information and its impact on spoken word recognition. However, as mentioned above, there is a growing consensus in the literature favouring hybrid views of spoken word processing, wherein the episodic and abstract information can co-exist (Cutler, 2008; Goldinger, 2007; Pisoni & Levi, 2007). A hybrid view of spoken word recognition could accommodate the findings of this thesis, without disregarding important evidence in the literature supporting the existence of abstract representations. As Cutler (2008) notes, the functional significance of abstract representations is undeniable, since abstraction and generalisation are crucial factors in the efficiency of cognitive processing in general and speech processing in particular. However, as also noted in **Chapter1**, integrating abstract representations and stored episodic information in a single model is a challenge, and despite a few attempts, there is as yet little directly relevant evidence. I will briefly mention the three recent relevant attempts (a more detailed review was provided in **Chapter 1**) and discuss whether and how they can accommodate the sound specificity effects found in this thesis.

6.4.3.1. The complementary systems approach

Originally proposed by McClelland et al. (1995) as a *complementary learning systems* (CLS) model of memory, this approach was adapted for spoken word perception by Goldinger (2007). It is based on a *complementary-systems* perspective, wherein reciprocal computational neural networks represent hippocampal and cortical memory systems. The hybrid memory system posited by the model eliminates the abstract-episodic opposition. Namely, detailed episodic traces (hippocampal system) and holographic, abstract traces (cortical system) combine to simulate behavior in real time, thus allowing perceptual or memory data to appear relatively “episodic”. Importantly, the two memory systems are inter-connected, such that the traces in each result from the complementary interactivity between the systems. Goldinger (2007) successfully simulated the model with a voice-sensitive priming task on bisyllabic words and observed that same-voice trials led to fastest settling times (a measure of the network’s performance), whereas larger voice changes induced a steady decline in performance. Therefore, the simulation of voice-sensitive priming demonstrated that activity in the hippocampal network can simulate changes in the indexical properties of words.

In principle, Goldinger (2007)’s proposal could be extended to accommodate the sound specificity effects observed here. The episodic traces in the hippocampal system could be modified to include more episodic detail corresponding to the sounds, as illustrated in some of the earlier visualisations In almost all of them, except for Figure 6.1A, episodic information of the auditory episode of a word resides in an episodic memory structure and is connected/interacts with more abstract lexical information that resides in a lexical memory structure.

6.4.3.2. The adaptive resonance theory approach

The second relevant attempt at a hybrid model was put forward by McLennan et al. (2003) and was inspired by their findings from a series of repetition priming experiments that examined whether surface representations of spoken words are mapped onto underlying, abstract representations. More specifically, they tested the hypothesis that flaps (neutralised allophones of intervocalic /t/s and /d/s) are mapped onto their underlying phonemic counterparts. The overall results supported the co-existence of both surface and underlying form-based representations, which motivated them to propose an explanatory account that was adapted from Grossberg's ARTPHONE neural model (Grossberg et al., 1997). In such a model, the acoustic–phonetic input consisting of relatively rich and specific surface representations, resonates with 'chunks' belonging to more abstract phonological representations, as well as 'chunks' corresponding to less abstract, allophonic representations. These resonances in turn serve as the basis for long-term repetition priming (the task predominantly used in McLennan et al.. According to McLennan et al., perception may be better conceived of as a resonance between the learned expectation and the sensory input, such that the percept may not necessarily exist in either the sensory data or the long-term representation, but instead, in some mixture of the two. While possibly a plausible approach for integrating abstract and episodic information, it has not yet been implemented to predict indexical effects at a lexical level. In this respect, the CLS approach adapted by Goldinger (2007) provides more direct evidence towards building plausible hybrid models. At present, it is unclear how such an approach could be extended to explain the sound specificity effect.

6.4.3.3. The socially-weighted dual-route approach

The third and also most recent attempt is by Sumner et al. (2014). This is a dual-route approach of speech perception that advocates the integration of linguistic and talker-related information from a socio-linguistic perspective. They argue that the perception of spoken words is socially weighted and propose a dual-route approach to speech perception in which listeners map acoustic patterns in speech to linguistic and social representations simultaneously. Accordingly, socially salient tokens are encoded with greater strength (by increased attention to the stimulus) than both typical and atypical non-salient tokens. In this view, a representation derived from one instance of a strongly encoded socially salient token may be as robust as one derived from a large number of less salient, default tokens. An interesting aspect of this approach is, that contrary to typical views that try to explain the many-to-one mapping of variable signals to a single linguistic representation, it endorses a one-to-many perspective, in which a single speech string is mapped to multiple linguistic and social representations simultaneously. A visual illustration of the approach taken from Sumner et al., (2014), is displayed in Figure 6.3.

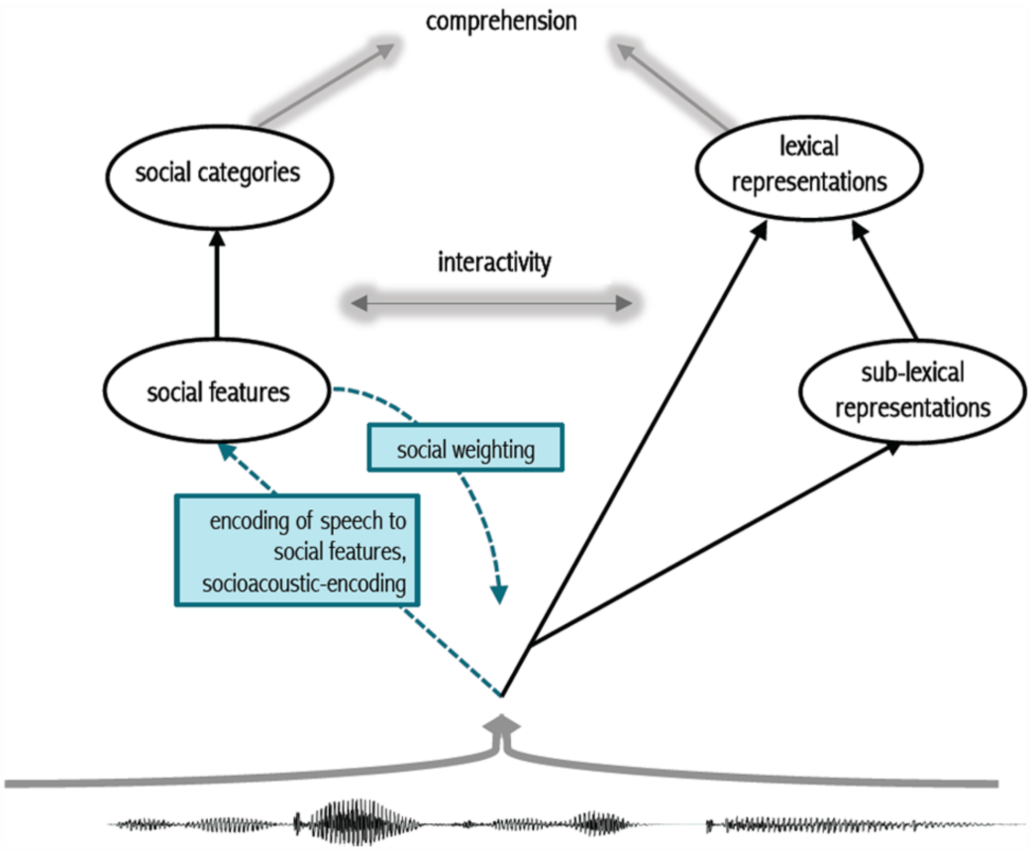

Figure 6.3. (Image and explanation taken from Sumner et al., 2014). In tandem with the encoding of speech to sounds and words (right), acoustic patterns in speech are encoded to social representations (left). Socially weighted encoding results from the heightened activation of social representations that modulates attention to the speech signal. This in turn results in the deep encoding of socially salient acoustic patterns along with linguistic representations, but also independent of them.

Although awaiting direct empirical validation, this approach offers an interesting perspective on how the abstract/episodic debate could be approached. Accordingly, it also presents a plausible opportunity to accommodate speech-extrinsic, contextual information that is not discarded, but makes it into the processing stages of spoken words. Consider the following observations arising from the evidence up to date on speech-extrinsic effects. Namely, the sound/noise details that co-occur with spoken words seem to display the following properties:

- Can be perceptually integrated with spoken words (Cooper et al., 2015)
- Can affect spoken word identification (Pufahl & Samuel, 2014)
- Can be integrated with newly learned words in memory (Creel at al., 2012)
- Can affect spoken word recognition in some contexts, but not in others (this thesis)
- Can behave similarly to, but not the same as indexical information (this thesis, Pufahl & Samuel, 2014)
- There is no compelling evidence for their presence in the lexicon (no such evidence for the presence of the indexical information, either)

Taken together, these observations lead to wondering whether perhaps it could be more informative to shift the focus from *where* to *how* sound-word integration occurs. Specifically, thinking in terms of how the sound information could be processed alongside the linguistic and indexical information to impact speech comprehension might be more constructive than trying to argue *where* in memory it resides The approach proposed by Sumner et al. (2014) seems particularly relevant here. This account offers the flexibility of adding a third simultaneous route to the model in a way that does not seem to dispute the above considerations.

In Figure 6.4, I have sketched an extended version of the illustration in Sumner at et al. (2014) displayed in Figure 6.3. The parts belonging to the original figure are highlighted in grey, and the additional components in blue. The labels of the original illustration are also slightly modified, such that I have included the term "indexical" alongside "social", to allow for the processing of indexical information in a broad sense, not confined to only social features and categories. I have also added a "recognition" term next to "comprehension" to accommodate the recognition process. Note that in this case, the assumption is that recognising the word entails understanding/comprehending it, but the model could be modified to comprise a distinction, if necessary. As illustrated, this triple-route approach can accommodate speech-extrinsic specificity detail, in those cases when it affects spoken word recognition/comprehension. Analogous to the "social weighting" posited by Sumner et al. (2014), a so-called "contextual" weighing could take place when sounds/noise co-occur with speech. Accordingly, contextually weighted encoding that modulates attention to the speech signal, may result from the enhanced activation of contextual representations (e.g., from salient sound categories). The entire process may then lead to the deep encoding of socially and contextually salient acoustic patterns along with linguistic representations in a simultaneous fashion.

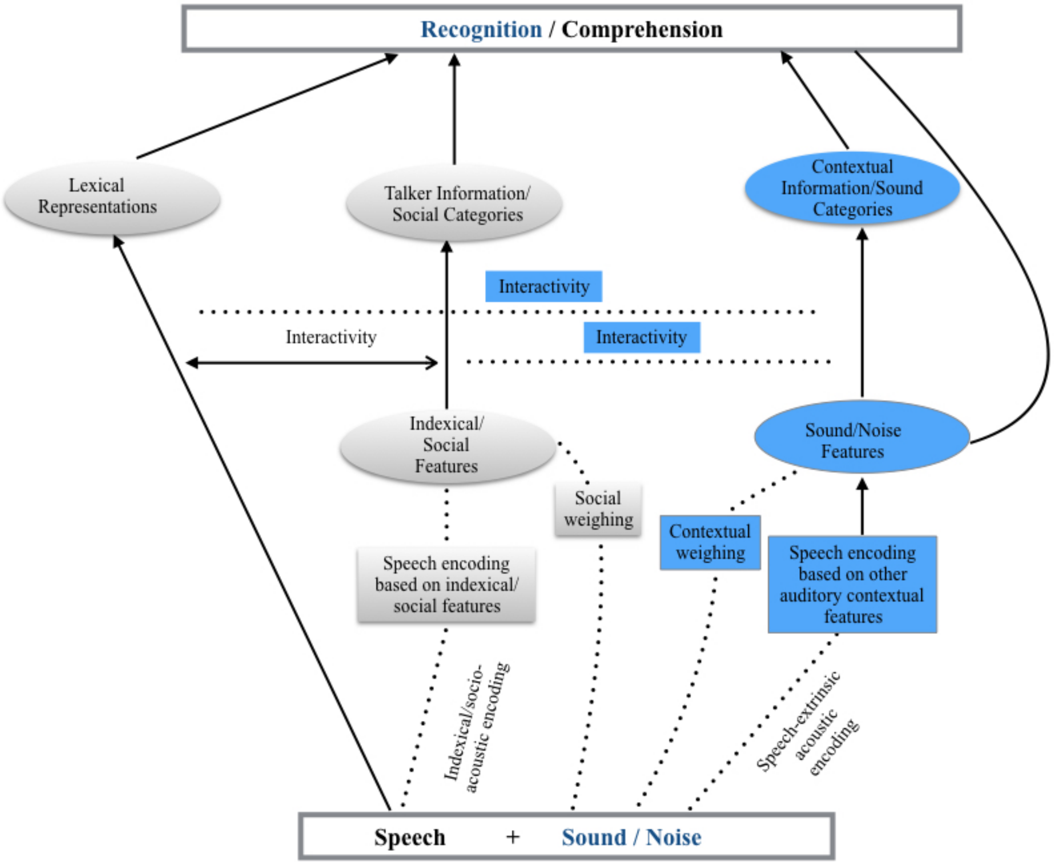
Recognition / Comprehension
Lexical Representations
Talker Information/ Social Categories
Contextual Information/Sound Categories
Interactivity
Interactivity
Interactivity
Indexical/ Social Features
Sound/Noise Features
Social weighing
Contextual weighing
Speech encoding based on indexical/ social features
Speech encoding based on other auditory contextual features
Indexical/socio-acoustic encoding
Speech-extrinsic acoustic encoding
Speech + Sound / Noise

Figure 6.4 (continued from the previous page). A visual illustration of how Sumner (2014)'s Dual-Route Approach can be adapted to accommodate speech-extrinsic auditory information present in its auditory context. This adaptation leads to a triple-route approach, in which linguistic, indexical and speech-extrinsic information can be processed simultaneously, affecting recognition/comprehension of the spoken word. I have not dissociated between recognition and comprehension in this instance, since for the sake of simplicity, I am assuming that recognising the word entails comprehending it. However, it should be possible to modify that level, if required. The parts highlighted in blue represents the third route that I posit adding to Sumner et al. (2014)' approach (the components in grey), in order to accommodate the present sound specificity effects. The dotted lines in the case of the third route represent the evidence that speech-extrinsic sound/noise information may not always make it to affect processing, given the seemingly transient and context-selective nature of the respective specificity effects.

6.5. Potential limitations

In this section, I will address three potential limitations of the research presented in this thesis. These include: the suggestion that explicit memory tasks may not be the most efficient and appropriate for revealing specificity effects; the relatively constrained implications arising from the glimpse account; the failure to replicate the original sound specificity effect reported by Pufahl and Samuel (2014); and finally, the limitations in explaining what I refer to as the "when, how, and where" conundrum, that deals with the main questions revolving around sound specificity effects.

6.5.1. The memory task

When it comes to the methodology used in measuring indexical and sound specificity effects, the memory test has been a matter of controversy. As discussed in **Chapter 2**, the debate has focused on two main issues: 1) the reliability of the explicit vs. implicit tasks: implicit tasks have been found more sensitive; and 2) the suitability (validity) of these tasks to measure specificity effects that have implications for lexical representations in memory: explicit tasks are considered to tap into episodic memory, rather than the lexicon (Goh, 2005; Pufahl & Samuel, 2014).

With respect to the first matter, the results in this thesis indicates the opposite pattern. Namely, the recognition memory task (explicit) was successful in revealing the voice specificity effect (Experiment 1: **Chapter 2**), as well as sound specificity effects in two cases (**Chapter** 3: Experiment 3, and **Chapter 4**: Experiment 4). On the other hand, the identification memory task used in one study failed to reveal the anticipated sound specificity effect (**Chapter 5**: Experiment 7).

The issue of suitability is more complicated. As I argued in **Chapter 2**, the validity of the dependent measure depends on theoretical assumptions about the structural organization of episodic and lexical memory. Judging from the ongoing debate in the literature, it is still unclear what constitutes strictly lexical vs episodic memory, how these two types of memory interact with each other, and whether there is a need for drawing boundaries at all. Perhaps the best arguments to be made in favour of using a recognition task are: 1) ultimately, the task is about the word per se, and recognising the word entails lexically accessing it; and 2) both explicit and implicit tasks have been used extensively to measure specificity effects and inform models of spoken word recognition.

6.5.2. The fragility of the glimpse account

As noted earlier in the discussion, the case of highly contrasted glimpses provides evidence in support of the possibility that the degraded versions of the words are retained in memory, without the actual sounds being encoded. In doing so, it offers an alternative explanation to the claims that mere co-occurrence is sufficient for the emergence of a sound specificity effect. However, this account is fragile in two respects:

1) The calculated glimpse proportions from the two sounds were quantitatively different in both the experiment that revealed an effect (**Chapter 3**: Experiment 3) and the one that did not (**Chapter 3**: Experiment 2 (A and B)). I argued that the crucial factor in explaining the appearance of an effect was the *quality* of glimpses, rather than their quantitative measure. Specifically, in the case of an effect, the glimpses of the same word resulting from the masking of the two sounds were both quantitatively and qualitatively different, due to the masking contrast created as a result of a joint change in both the sound pitch and its temporal overlap with the word.

2) The absence of a study where only the degraded versions (glimpses) of the words are played as stimuli, instead of the word-sound pairs, also contributes to the limitation of the account. Similarly, another study in which the same sound would have led to different glimpses of the same word (i.e., the same car horn sound, in different temporal overlaps with the word) could have strengthened the case for the encoding of the glimpses in memory.

Therefore, despite making a plausible case for the role of energetic masking in sound specificity effects, this account needs more elaboration and additional empirical work.

6.5.3. The context uniqueness puzzle

The failure to replicate the original sound specificity effect found by Pufahl and Samuel (2014) despite using the same environmental sounds, the same encoding and memory tasks, the same filtering technique, and a similar (though, not the same) experimental design, is a puzzling result and could be considered as a potential weakness of the present work (**Chapter 5**: Experiment 7). However, it is important to consider the following two observations that arise from both these studies. First, the lack of an effect does not constitute a problem only for the present work, but also for the Pufahl and Samuel's original study. Given the novelty of the effect, any failure to replicate it weakens the strength of claims made with regard to its implications. Second, taken together, the pattern of results in these studies suggests that these types of effects are small and fragile in general.

6.5.4. The "when, how, and where" conundrum

This issue is complex and almost impossible to address within a single study. As already mentioned, the results in this thesis provide more evidence for the *when* question regarding the emergence of sound specificity effects. Specifically, these results identified conditions that constrain *when* the speech and sound stay together.

As to *where* in memory these effects may reside, this work provides little direct evidence. Although the implications of specificity effects in general have been frequently interpreted in terms of lexical representations, it is far from clear whether the sounds, or even voices, co-exist with the words in long-term lexical memory I suggested several ways in which these effects could be accommodated in the memory system by providing visualisations, but these are speculative at this stage.

With respect to *how* speech-extrinsic auditory information is processed alongside linguistic and indexical information, again, the present work provides little-to-no direct evidence. Perhaps the only insight is the integrality effect, as reminiscent of the *common fate* principle of integrating different sources of information that follow the same pattern of change over time.

I addressed this question indirectly in section 6.4.3.3, where I discussed how the sound specificity effect could be accommodated in a recently proposed dual-route theoretical approach of speech perception (Sumner et al., 2014). This account attempts to explain how different sources of information in the context of spoken words, namely linguistic and indexical, can be integrated in parallel during processing.

As it currently stands in the literature, a better understanding and interpretation of sound specificity effects in the face of this multi-faceted conundrum will need to be supported by considerably more evidence in future studies.

6.6. Future directions

Speech-extrinsic specificity effects are a very recent trend in the literature of specificity effects, albeit an exciting one, that opens the path for interesting questions regarding spoken word recognition and representation in memory. In this final section, I outline four future directions for the field.

6.6.1. The time-course and consolidation of sound specificity effects in memory

As noted earlier, it is difficult to make plausible claims regarding the status of sound specificity effects in memory without evidence for their time-course and consolidation patterns. At the moment, such evidence is lacking. There is only some evidence from a study using the Garner speeded classification paradigm (Garner, 1974; Exp.1), suggesting that speech-extrinsic information (noise) may be processed alongside indexical information relatively early, at a perceptual stage (Cooper et al., 2015).

It would be interesting to examine whether a *time*-course *hypothesis* like the one proposed for indexical effects applies to sound specificity effects as well (Luce et al., 2003; McLennan & Luce, 2005). Future studies in the spirit of McLennan & Luce's (2005), which examined whether processing time mediates the emergence of sound specificity effects, could provide insights on whether these effects appear early or late in processing.

It will also be important to see whether these effects persist in long-term memory. Like in the case of indexical studies, the retention intervals used in the few existing sound specificity studies are relatively short (usually less than one hour). Even in the case of indexical effects, only a few studies have directly addressed this issue. For example, Goldinger (1996) found voice specificity effects in a word identification identification-in-noise task one week post-study. Recently, Brown and Gaskell (2014) showed that voice-specificity effects on recognition of novel words emerged immediately after study and remained generally stable over the course of a week. Hence, future studies of speech-extrinsic specificity effects will need to address this issue.

Finally, it would be interesting to see whether speech-extrinsic auditory information is consolidated in memory. For indexical effects, Brown and Gaskell (2014) showed that the encoding of newly learned words seemed to initially retain detailed episodic information such as talker identity. These representations could be maintained for at least a week after the words were learned, but did not show any consolidation advantage when tested at later time intervals. Regarding speech-extrinsic details, Creel et al. (2012) also showed that the encoding of newly learned word appeared to retain speech-extrinsic contextual detail (background noise) immediately after study. However, this study did not investigate the retention time interval of these specificity effects in memory and their consolidation pattern. Future studies that follow up on this issue would be informative.

6.6.2. Identifying the locus of sound specificity effects in memory

Identifying *where* in memory sound specificity effects reside will be an important, albeit challenging task. Similar to the case of indexical effects, the debate will ultimately concentrate on dissociating whether they belong exclusively in an episodic memory system/subsystem, or in one that allows some sort of interaction, or complementary co-existence with the long-term lexical system. Judging from a recent trend in designing models of spoken word recognition/learning that rely on the *complementary systems* approach (Goldinger, 2007; Davis & Gaskell, 2009)*,* the complementary learning systems (CLS) models of memory (McClelland et al., 1995) may provide useful insights into such a debate and aid towards understanding the locus of sound specificity effects in memory.

6.6.3. Testing different populations

Testing different populations could also prove informative with respect to how speech-extrinsic auditory variability impacts spoken word processing. Non-native (L2) listeners constitute an interesting population that has already been investigated with respect to talker-related indexical variability. For example, Bradlow and Pisoni (1999) showed that the ability to benefit from surface phonetic information, such as a consistent talker across items, is a skill that is present in both first and second language perception. However, in that study, non-native listeners were affected more by the lexical nature of the stimuli, such that they had particular difficulty with lexically hard words even when familiarity with the items was controlled. This last finding suggests that non-native word recognition may be compromised when the task requires fine phonetic discrimination at the segmental level. Another study by Bent et al. (2010) explored how across-talker differences influence non-native vowel perception in American English (native) and Korean listeners (non-native). Results demonstrated that Korean listeners' error patterns for four vowels were strongly influenced by variability in vowel production that was within the normal range for the American English talkers. These results suggest that non-native listeners are strongly influenced by cross-talker variability perhaps because of the difficulty they have forming native-like vowel categories. Using a speeded classification paradigm (Garner, 1974), Vaughn and Brouwer (2012) tested English monolinguals and Mandarin-English bilinguals to examine how different types of indexical information, talker information and the language being spoken, are perceptually integrated in bilingual speech. Variability in characteristics of the talker (gender and talker identity) and in the language being spoken (Mandarin vs. English) was manipulated. Listeners from both groups classified short, meaningful sentences obtained from different Mandarin-English bilingual talkers on these indexical dimensions. Results showed that gender information and language were processed in an integral manner for both groups, but only bilinguals demonstrated symmetrical interference for talker identity-language classification.

When it comes to speech-extrinsic acoustic variability, there are several aspects that make non-native listeners an interesting population in which to examine arising effects. For example, it is well-known that non-native speech processing in noise is more difficult than the native counterpart (e.g., Garcia Lecumberri et al., 2010). Further, native and and non-native listeners seem to use different types of information to interpret speech in noise: native listeners normally use higher-level lexical information more than non-native listeners, and this reliance is further increased in the presence of noise (Mattys et al., 2010). In general, lexical information is less available to non-native listeners, such as vocabulary size, relative lexical frequencies of occurrence, transitional probabilities, and contextual plausibility (Mattys et al., 2010). Therefore, L2 listeners may reveal a relatively different pattern of sound specificity effects compared to their native counterparts. Namely, they may not segregate the sounds from speech as easily or efficiently as native listeners and may thus encode them pair-wise in memory more readily. Similarly, non-native listeners may use the paired sound as a source of contextual cue to facilitate word retrieval in word identification/identification tasks. Finally, the proficiency level of the second language, as well as the degree of similarity between the first and second languages could be factors in the emergence and strength of a sound specificity effect.

Another population of interest to investigate the role of speech-extrinsic auditory variability is speech processing is children. Unlike adults, children are still acquiring language, and hence their lexical representations may be less stable and less robust than those of adults. Therefore, they may be affected by the variability in talker-related surface properties of speech to a relatively different extent compared to adults. Form existing evidence in the literature, we know that children make use of talker-related indexical information during speech comprehension and learning of new words, although in a relatively constrained fashion. For example, using eye-tracking methodology, Creel and colleagues have conducted several studies investigating how children integrate linguistic and indexical information during on-line speech processing. Results have indicated that similar to adults, children, store real-world knowledge of the role activated by a talker's voice (e.g., male talker identifying his role as a pirate and female talker identifying her role as a princess) and actively use this information during speech comprehension (Borovsky & Creel, 2014). In another study, Creel (2012) also showed that children can encode information about talkers while simultaneously learning new words, suggesting that their language input may be conditioned on talker context quite early in language learning. Another recent study examined the familiar talker advantage in school-age children (Levi, 2015). Children were first familiarized with the voices of three different talkers and were subsequently tested on the speech produced by six talkers, only three of whom were familiar. Results revealed that children displayed the familiar talker advantage in their speech processing, such that their performance was higher in the case of familiar talkers. However, this benefit was limited to highly familiar lexical items, which could be attributed to differences in the representation of highly familiar and less familiar lexical items.

Similar research questions to those concerning children's performance in the context of speech-intrinsic, talker-related variability, can also be examined in the context of speech-extrinsic auditory variability. Children may show a relatively different or similar pattern in their speech processing performance to that of adults with respect to the latter variability. Perhaps, they make different use of auditory contextual cues (co-occurring sounds/noises) during speech comprehension and learning of new words. These cues may have a facilitatory, or an inhibitory effect on their speech performance. For instance, children may integrate co-occurring sound information into their memory representations more readily than adults. Alternatively, they may segregate such information to a better extent than adults. It would be interesting and informative to see future studies that address these issues.

6.7. General conclusions

In summary, this thesis investigated the co-existence of spoken words and environmental sounds in memory, from the perspective of how variability in the co-occurring sound affects the recognition of the word. By manipulating the context of speech-sound co-existence, it identified two major constraints. The first one regards degradation of the same word by different masking sounds, which when contrasted enough, can lead to a sound specificity effect. This effect indicates that perhaps it may be the unique degraded versions of the words that are retained in memory, rather than the co-occurring sounds per se. The second constraint arises from the integrality element in the speech-sound co-occurrence, that was shown to play an intrinsic role in the emergence of a sound specificity effect. Namely, when the sound is rendered integral to the word, such that segregation of the two is difficult, a relatively robust effect of the sound variability on word recognition memory emerges. This effect is reminiscent of a perceived functional/causal link between word and sound, similar to the case of word and voice.

Overall, the present work shows that like in the case of indexical variability, spoken word recognition is also sensitive to variation arising from external auditory sources. However, unlike indexical effects, sound specificity effects are fragile and conditional. Listeners seem to be able to encode details of sounds co-occurring with speech in their memory representations, but only in certain occasions. Importantly, mere co-occurrence is not always sufficient in eliciting an effect of the sound variability on word recognition performance. For these reasons, the present results restrain the scope of any claims concerning further expansion of the mental lexicon. As it currently stands, we may have to wait for more compelling evidence to emerge in the literature, before seriously challenging this memory structure with the inclusion of speech-extrinsic auditory information.

Appendix A

List of all the word stimuli and their respective average frequency values

Word	Animacy	CELEX MLN	CELEX Log
dolphin	Animate	0.48	3
eagle	Animate	0.95	9
squirrel	Animate	0.78	6
rabbit	Animate	1.28	19
baby	Animate	2.41	258
doctor	Animate	2.26	184
teacher	Animate	2.21	162
student	Animate	2.48	304
actor	Animate	1.92	84
singer	Animate	1.08	12
tiger	Animate	1.08	12
monkey	Animate	1.26	18
writer	Animate	1.82	66
donkey	Animate	1.15	14
zebra	Animate	0.3	2
hamster	Animate	0.6	4
panther	Animate	0.9	8
parrot	Animate	0.6	4
penguin	Animate	0.7	5
pigeon	Animate	1.04	11
scorpion	Animate	0.3	2
spider	Animate	0.85	7
turtle	Animate	0.6	4
lizard	Animate	0.6	4
dentist	Animate	0.95	9
waiter	Animate	1.34	22
dancer	Animate	1.15	14
artist	Animate	1.87	74
painter	Animate	1.48	30
plumber	Animate	0.6	4

lawyer	Animate	1.71	51
driver	Animate	1.75	56
worker	Animate	2.31	204
banker	Animate	1.15	14
sculptor	Animate	0.7	5
soldier	Animate	1.92	83
athlete	Animate	1.23	17
chemist	Animate	0.85	7
scholar	Animate	1.26	18
leopard	Animate	0.9	8
	Average Frequency	1.22	45.45
	Standard Deviation	0.60	73.70
basket	Inanimate	1.38	24
biscuit	Inanimate	1.18	15
sofa	Inanimate	1.34	22
table	Inanimate	2.37	235
bottle	Inanimate	2.06	116
apple	Inanimate	1.48	30
orange	Inanimate	1.3	20
olive	Inanimate	1.11	13
lemon	Inanimate	1.18	15
chapel	Inanimate	1.34	22
cabin	Inanimate	1.48	30
oven	Inanimate	1.3	20
pencil	Inanimate	1.28	19
pillow	Inanimate	1.28	19
candle	Inanimate	1.2	16
onion	Inanimate	1.2	16
taxi	Inanimate	1.53	34
coffee	Inanimate	1.96	92
window	Inanimate	2.3	200
jacket	Inanimate	1.62	42
bucket	Inanimate	1.3	20

sugar	Inanimate	1.76	57
berry	Inanimate	1	10
paper	Inanimate	2.35	225
mirror	Inanimate	1.69	49
butter	Inanimate	1.43	27
carriage	Inanimate	1.2	16
peanut	Inanimate	0.7	5
panel	Inanimate	1.4	25
pepper	Inanimate	0.95	9
sausage	Inanimate	1.08	12
ribbon	Inanimate	1.04	11
building	Inanimate	2.25	177
bracelet	Inanimate	0.78	6
necklace	Inanimate	0.6	4
collar	Inanimate	1.34	22
blanket	Inanimate	1.46	29
freezer	Inanimate	0.6	4
heater	Inanimate	0.7	5
carpet	Inanimate	1.48	30
	Average Frequency	1.38	43.58
	Standard Deviation	0.45	60.45

Appendix B

Environmental Sounds Used in the Experiments 6 and 7

Table 5.11. The list of the environmental sounds used in the experimental trials, organised by their animacy

Animate Sounds - 40 categories - 2 exemplars in each, A and B	Inanimate Sounds - 40 categories - 2 exemplars in each, A and B
bear	accordian
bee	alarm clock
canary	bell
cat	bike bell
chick	boiling water
chicken	camera
chimp	can open
cicada	car horn
cow	cash register
coyote	chainsaw
cricket	chimes
crow	coins
cuckoo	cow bell
dog	cymbal
dolphin	doorbell
donkey	drumroll
dove	flute
duck	glass breaking
eagle	harmonica
elephant	harp
fly	helicopter
frog	jackhammer
goat	musicbox
goose	page turn
gorilla	partyfavor
horse	phone
lamb	piano
lion	pingpong
loon	saw

(**Table 5.11** continued from above)

Animate Sounds - 40 categories - 2 exemplars in each, A and B	Inanimate Sounds - 40 categories - 2 exemplars in each, A and B
mosquito	ship
mouse	shufflecards
owl	siren
parrot	steeldrum
pig	tambourine
rat	train
rattlesnake	trumpet
seagull	tuba
seal	typewriter
turkey	violin
woodpecker	zipper

Appendix C

Filter-wise average individual word intelligibilities and selected filters in Exp. 6-7

Table 5.12. Mean word identification accuracy values (% correct) for each individual word and each filter-bank condition in Phase 1 of Experiment 6.

Word	File Name	Filter 1	Filter 1.25	Filter 1.5	Filter 1.75	Filter 2	Filter 3
TIGER	FA01	100.00	100.00	100.00	87.50	87.50	80
DONKEY	FA02	100.00	100.00	100.00	62.50	62.50	20
PENGUEN	FA03	100.00	88.89	66.67	62.50	87.50	0
BANKER	FA04	0.00	0.00	0.00	0	0	20
STUDENT	FA05	100.00	100.00	77.78	87.50	75.00	40
DOLPHIN	FA06	66.67	66.67	100.00	50	25.00	20
EAGLE	FA07	100.00	100.00	100.00	87.50	87.50	60
PLUMBER	FA08	33.33	22.22	11.11	0	0	0
DENTIST	FA09	100.00	100.00	100.00	100	87.50	40
SQUIRREL	FA10	100.00	88.89	77.78	75.00	50.00	40
ZEBRA	FA11	100.00	88.89	55.56	25.00	12.50	0
ARTIST	FA12	100.00	88.89	66.67	62.50	87.50	60
SCHOLAR	FA13	50.00	55.56	33.33	50	25.00	0
CHEMIST	FA14	0.00	11.11	0.00	0	12.50	0
DRIVER	FA15	50.00	55.56	55.56	25.00	12.50	0
ATHLETE	FA16	83.33	44.44	33.33	12.50	12.50	0
LAWYER	FA17	83.33	55.56	44.44	62.50	62.50	20
PANTHER	FA18	100.00	66.67	0.00	37.50	12.50	0
SINGER	FA19	33.33	22.22	0.00	0	0	20
LEOPARD	FA20	83.33	66.67	66.67	62.50	50.00	20
WORKER	FA21	100.00	88.89	66.67	50.00	37.50	0
DANCER	FA22	100.00	88.89	77.78	75.00	75.00	40
RABBIT	FA23	16.67	77.78	22.22	25.00	0	0
WRITER	FA24	100.00	66.67	77.78	75.00	50.00	20
HAMSTER	FA25	83.33	88.89	44.44	25.00	37.50	0
SOLDIER	FA26	100.00	55.56	77.78	75.00	62.50	20
MONKEY	FA27	100.00	100.00	100.00	87.50	75.00	0
PARROT	FA28	100.00	100.00	77.78	100.00	25.00	0
TURTLE	FA29	100.00	88.89	88.89	50	25.00	0

(**Table 5.12** continued from above)

SCULPTOR	FA30	100.00	66.67	55.56	50.00	75.00	0
BABY	FA31	100.00	66.67	88.89	62.5	62.50	0
TEACHER	FA32	100.00	100.00	100.00	100	100	60
PIGEON	FA33	100.00	100.00	88.89	87.50	100	40
LIZZARD	FA34	100.00	100.00	100.00	75.00	87.50	80
DOCTOR	FA35	100.00	66.67	55.56	25.00	75.00	20
SPIDER	FA36	100.00	100.00	88.89	75.00	100	80
PAINTER	FA37	100.00	100.00	100.00	100	87.50	80
WAITER	FA38	83.33	100.00	100.00	87.5	87.50	40
ACTOR	FA39	100.00	100.00	100.00	75.00	75.00	20
SCORPION	FA40	100.00	100.00	100.00	100	87.50	80
PEANUT	FI01	100.00	55.56	33.33	37.50	37.50	0
MIRROR	FI02	100.00	55.56	0.00	37.50	37.50	0
SAUSAGE	FI03	100.00	100.00	88.89	87.50	62.50	40
LEMON	FI04	33.33	22.22	11.11	0	0	20
TABLE	FI05	100.00	88.89	100.00	87.50	87.50	80
APPLE	FI06	83.33	100.00	88.89	50	12.50	20
CHAPEL	FI07	50.00	77.78	22.22	12.50	12.50	20
ORANGE	FI08	100.00	77.78	55.56	50.00	25.00	20
PAPER	FI09	100.00	100.00	77.78	12.50	37.50	0
BLANKET	FI10	100.00	88.89	66.67	50.00	50.00	0
CARRIAGE	FI11	83.33	44.44	33.33	37.50	0	0
HEATER	FI12	100.00	100.00	100.00	75.00	62.50	40
TAXI	FI13	100.00	88.89	100.00	100	100	100
SUGAR	FI14	100.00	88.89	55.56	50.00	25.00	0
OVEN	FI15	83.33	66.67	88.89	37.50	50.00	20
JACKET	FI16	100.00	88.89	88.89	100	62.50	20
BISCUIT	FI17	100.00	88.89	77.78	62.50	75.00	40
CUPBOARD	FI18	66.67	22.22	44.44	25.00	25.00	0
BERRY	FI19	83.33	66.67	55.56	12.50	25.00	0
RIBBON	FI20	100.00	77.78	66.67	62.50	37.50	40
NECKLACE	FI21	100.00	100.00	88.89	37.50	62.50	0
CABIN	FI22	100.00	88.89	88.89	100	75.00	0

(**Table 5.12** continued from above)

BUCKET	FI23	83.33	88.89	55.56	50.00	50.00	0
CARPET	FI24	33.33	55.56	66.67	37.50	37.50	0
OLIVE	FI25	66.67	55.56	44.44	37.50	0	0
COLLAR	FI26	33.33	11.11	22.22	0	0	0
PANEL	FI27	16.67	0.00	11.11	50.00	12.50	0
BRACELET	FI28	100.00	100.00	88.89	50.00	62.50	0
WINDOW	FI29	100.00	100.00	88.89	50.00	62.50	20
FREEZER	FI30	83.33	88.89	88.89	50.00	37.50	0
ONION	FI31	100.00	77.78	77.78	87.50	50.00	0
PENCIL	FI32	83.33	100.00	100.00	87.5	75.00	20
SOFA	FI33	100.00	77.78	33.33	25.00	25.00	20
BOTTLE	FI34	100.00	100.00	44.44	37.5	12.50	0
PEPPER	FI35	100.00	55.56	44.44	25.00	12.50	0
CANDLE	FI36	0.00	11.11	11.11	0	0	0
BUILDING	FI37	100.00	88.89	100.00	100	37.50	0
COFFEE	FI38	66.67	22.22	0.00	0	12.50	0
PILLOW	FI39	100.00	77.78	44.44	12.50	12.5	0
BASKET	FI40	100.00	88.89	66.67	25.00	37.50	0
Mean Accuracy		**84.17**	**74.58**	**64.03**	**52.5**	**45.94**	**19.25**

Table 5.13. The selected mean word identification accuracy values (% correct) for each individual word and the corresponding filters at the end of Phase 1 in Experiment 6.

Word_Test	File Name	Accuracy Chosen	Filter Chosen
TIGER	FA01	80	3
DONKEY	FA02	62.5	1.75
PENGUEN	FA03	62.5	1.75
BANKER	FA04	0	1
STUDENT	FA05	75	2
DOLPHIN	FA06	100	1.5
EAGLE	FA07	87.5	2
PLUMBER	FA08	33.33	1
DENTIST	FA09	87.5	2
SQUIRREL	FA10	75	1.75
ZEBRA	FA11	88.89	1.25
ARTIST	FA12	66.67	1.5
SCHOLAR	FA13	50	1
CHEMIST	FA14	0	1
DRIVER	FA15	50	1
ATHLETE	FA16	83.33	1
LAWYER	FA17	83.33	1
PANTHER	FA18	66.67	1.25
SINGER	FA19	33.33	1
LEOPARD	FA20	66.67	1.5
WORKER	FA21	66.67	1.5
DANCER	FA22	75	1.75
RABBIT	FA23	77.78	1.25
WRITER	FA24	77.78	1.5
HAMSTER	FA25	88.89	1.25
SOLDIER	FA26	77.78	1.5
MONKEY	FA27	75	2
PARROT	FA28	100	1.75
TURTLE	FA29	88.89	1.5
SCULPTOR	FA30	66.67	1.25
BABY	FA31	62.5	1.75
TEACHER	FA32	60	3

(**Table 5.13** continued from above)

PIGEON	FA33	87.5	1.75
LIZZARD	FA34	80	3
DOCTOR	FA35	66.67	1.25
SPIDER	FA36	80	3
PAINTER	FA37	80	3
WAITER	FA38	87.5	2
ACTOR	FA39	75	1.75
SCORPION	FA40	80	3
PEANUT	FI01	55.56	1.25
MIRROR	FI02	55.56	1.25
SAUSAGE	FI03	87.5	1.75
LEMON	FI04	33.33	1
TABLE	FI05	80	3
APPLE	FI06	88.89	1.5
CHAPEL	FI07	50	1
ORANGE	FI08	77.78	1.25
PAPER	FI09	77.78	1.5
BLANKET	FI10	66.67	1.5
CARRIAGE	FI11	83.33	1
HEATER	FI12	75	1.75
TAXI	FI13	100	3
SUGAR	FI14	55.56	1.5
OVEN	FI15	66.67	1.25
JACKET	FI16	62.5	2
BISCUIT	FI17	62.5	1.75
CUPBOARD	FI18	66.67	1
BERRY	FI19	66.67	1.25
RIBBON	FI20	66.67	1.5
NECKLACE	FI21	88.89	1.5
CABIN	FI22	75	2
BUCKET	FI23	88.89	1.25
CARPET	FI24	66.67	1.5
OLIVE	FI25	66.67	1
COLLAR	FI26	33.33	1

(**Table 5.13** continued from above)

PANEL	FI27	16.67	1
BRACELET	FI28	50	1.75
WINDOW	FI29	50	1.75
FREEZER	FI30	88.89	1.5
ONION	FI31	77.78	1.5
PENCIL	FI32	75	2
SOFA	FI33	77.78	1.25
BOTTLE	FI34	44.44	1.5
PEPPER	FI35	55.56	1.25
CANDLE	FI36	0	1
BUILDING	FI37	100	1.75
COFFEE	FI38	66.67	1
PILLOW	FI39	77.78	1.25
BASKET	FI40	66.67	1.5
Mean Accuracy		**68.14 % correct**	

Table 5.14. The average word identification accuracy after the first piloting in Phase 2 of Experiment 6, the updated filters at each intermediate piloting step, and the mean accuracy after the last filter update. The version of the updated filters are noted by "v1, v2, v3", and the number in brackets indicate the number of participants that completed the respective piloting phase.

Word	File Name	Phase 1_Filter Chosen	Phase 2_Accuracy (5)	Phase 2_Updated Filters_v1	Updated Filters_v2	Updated Filters_v3	Accuracy_v3(7)
TIGER	FA01	3	100.00		3.5	4	85.71
DONKEY	FA02	1.75	100.00				100.00
PENGUEN	FA03	1.75	80.00				100.00
BANKER	FA04	1	20.00	0.25	0.25	0.25	28.57
STUDENT	FA05	2	60.00				85.71
DOLPHIN	FA06	1.5	80.00				57.14
EAGLE	FA07	2	80.00				85.71
PLUMBER	FA08	1	0.00	0.75	0.75	0.75	42.86
DENTIST	FA09	2	80.00				100.00
SQUIRREL	FA10	1.75	80.00				71.43
ZEBRA	FA11	1.25	60.00				14.29
ARTIST	FA12	1.5	100.00	1.75	1.75	1.75	85.71
SCHOLAR	FA13	1	40.00				42.86
CHEMIST	FA14	1	20.00	0.25	0.25	0.5	85.71
DRIVER	FA15	1	40.00				57.14
ATHLETE	FA16	1	60.00				71.43
LAWYER	FA17	1	100.00	1.25	1.25	1.25	42.86
PANTHER	FA18	1.25	40.00				71.43
SINGER	FA19	1	40.00	0.75	0.75	0.5	71.43
LEOPARD	FA20	1.5	80.00				42.86
WORKER	FA21	1.5	60.00				57.14
DANCER	FA22	1.75	80.00				85.71
RABBIT	FA23	1.25	80.00				14.29

(**Table 5.14** continued from above)

WRITER	FA24	1.5	60.00				85.71
HAMSTER	FA25	1.25	80.00				42.86
SOLDIER	FA26	1.5	100.00	1.75	1.75	1.75	57.14
MONKEY	FA27	2	80.00				85.71
PARROT	FA28	1.75	40.00	1.5	1.5	1.5	57.14
TURTLE	FA29	1.5	40.00	1.25	1.25	1.25	71.43
SCULPTOR	FA30	1.25	60.00				57.14
BABY	FA31	1.75	40.00	1.5	1.5	1.5	100.00
TEACHER	FA32	3	80.00				100.00
PIGEON	FA33	1.75	100.00	2	2	2	85.71
LIZZARD	FA34	3	80.00				85.71
DOCTOR	FA35	1.25	80.00				57.14
SPIDER	FA36	3	60.00				42.86
PAINTER	FA37	3	60.00				85.71
WAITER	FA38	2	100.00				71.43
ACTOR	FA39	1.75	80.00				85.71
SCORPION	FA40	3	80.00				85.71
PEANUT	FI01	1.25	80.00				57.14
MIRROR	FI02	1.25	80.00				42.86
SAUSAGE	FI03	1.75	100.00				57.14
LEMON	FI04	1	60.00			0.75	42.86
TABLE	FI05	3	100.00		3.5	4	57.14
APPLE	FI06	1.5	80.00				71.43
CHAPEL	FI07	1	0.00				42.86

(**Table 5.14** continued from above)

ORANGE	FI08	1.25	80.00				85.71
PAPER	FI09	1.5	80.00				85.71
BLANKET	FI10	1.5	80.00				57.14
CARRIAGE	FI11	1	100.00				85.71
HEATER	FI12	1.75	80.00				100.00
TAXI	FI13	3	100.00		3.5	4	71.43
SUGAR	FI14	1.5	40.00	1.25	1.25	1.25	57.14
OVEN	FI15	1.25	80.00				57.14
JACKET	FI16	2	80.00				85.71
BISCUIT	FI17	1.75	80.00				71.43
CUPBOARD	FI18	1	20.00				57.14
BERRY	FI19	1.25	40.00	1	1	1	85.71
RIBBON	FI20	1.5	100.00				85.71
NECKLACE	FI21	1.5	60.00				28.57
CABIN	FI22	2	60.00				71.43
BUCKET	FI23	1.25	100.00	1.5	1.5	1.5	42.86
CARPET	FI24	1.5	60.00				42.86
OLIVE	FI25	1	40.00				57.14
COLLAR	FI26	1	0.00	0.75	0.75	0.75	57.14
PANEL	FI27	1	20.00	0.5	0.5	0.5	100.00
BRACELET	FI28	1.75	60.00				71.43
WINDOW	FI29	1.75	80.00				42.86
FREEZER	FI30	1.5	100.00				85.71
ONION	FI31	1.5	80.00				100.00
PENCIL	FI32	2	80.00				71.43

(**Table 5.14** continued from above)

SOFA	FI33	1.25	100.00				57.14
BOTTLE	FI34	1.5	20.00	1.25	1.25	1.25	42.86
PEPPER	FI35	1.25	80.00				71.43
CANDLE	FI36	1	0.00	0.25	0.25	0.5	85.71
BUILDING	FI37	1.75	60.00				57.14
COFFEE	FI38	1	20.00				28.57
PILLOW	FI39	1.25	60.00				14.29
BASKET	FI40	1.5	40.00	1.25	1.25	1.25	28.57
Mean Accuracy			**66.00**				**65.89**

Table 5.15. The final selected filters for each word individually, decided after Phase 2 of Experiment 6 was completed and a few more filter changes were made

Word_Test	File Name	Final Filters_Phase 2
TIGER	FA01	4
DONKEY	FA02	2
PENGUEN	FA03	1.75
BANKER	FA04	0.25
STUDENT	FA05	2
DOLPHIN	FA06	1.5
EAGLE	FA07	2
PLUMBER	FA08	0.5
DENTIST	FA09	2
SQUIRREL	FA10	1.75
ZEBRA	FA11	1.25
ARTIST	FA12	1.75
SCHOLAR	FA13	0.75
CHEMIST	FA14	0.5
DRIVER	FA15	0.75
ATHLETE	FA16	1
LAWYER	FA17	1
PANTHER	FA18	1.25
SINGER	FA19	0.5
LEOPARD	FA20	1.25
WORKER	FA21	1.5
DANCER	FA22	2
RABBIT	FA23	1.25
WRITER	FA24	1.5
HAMSTER	FA25	1.25
SOLDIER	FA26	1.75
MONKEY	FA27	2
PARROT	FA28	1.5
TURTLE	FA29	1.25
SCULPTOR	FA30	1.25
BABY	FA31	1.5
TEACHER	FA32	3
PIGEON	FA33	2

(**Table 5.15** continued from above)

LIZZARD	FA34	3
DOCTOR	FA35	1.25
SPIDER	FA36	2
PAINTER	FA37	3
WAITER	FA38	2
ACTOR	FA39	1.75
SCORPION	FA40	3
PEANUT	FI01	1.25
MIRROR	FI02	1
SAUSAGE	FI03	1.75
LEMON	FI04	0.5
TABLE	FI05	3.5
APPLE	FI06	1.5
CHAPEL	FI07	0.5
ORANGE	FI08	1.25
PAPER	FI09	1.5
BLANKET	FI10	1.25
CARRIAGE	FI11	1
HEATER	FI12	1.75
TAXI	FI13	4
SUGAR	FI14	1.25
OVEN	FI15	1
JACKET	FI16	2
BISCUIT	FI17	1.75
CUPBOARD	FI18	0.75
BERRY	FI19	1
RIBBON	FI20	1.5
NECKLACE	FI21	1.5
CABIN	FI22	2
BUCKET	FI23	1.25
CARPET	FI24	1.25
OLIVE	FI25	0.75
COLLAR	FI26	0.5
PANEL	FI27	0.75

(**Table 5.15** continued from above)

BRACELET	FI28	1.75
WINDOW	FI29	1.5
FREEZER	FI30	1.5
ONION	FI31	1.5
PENCIL	FI32	2
SOFA	FI33	1.25
BOTTLE	FI34	1.25
PEPPER	FI35	1.25
CANDLE	FI36	0.5
BUILDING	FI37	1.75
COFFEE	FI38	0.5
PILLOW	FI39	1
BASKET	FI40	1.25

Appendix D

Questionnaire about the environmental sounds used in Experiment 7

Name:

Date:

Participant Number:

Questionnaire

Please take some time to reflect on the environmental sounds you heard throughout the experiment and answer the following questions as accurately as you can.

1. How do you think the environmental sounds were distributed throughout the experiment, with respect to their source? Choose one of the following options:

   A. There were more sounds from animate than inanimate sources

   B. There were more sounds from inanimate than animate sources

   C. There were equal (or roughly so) numbers from both sources

   D. Don’t know / Didn’t notice

2. i) Did you hear more than one exemplar of a certain sound between the two phases of the experiment (e.g., a dog barking sound in the first part and another dog barking in the second part)? Yes/No

ii) If yes, how often did this happen? Choose one of the following options

A. Always

B. Very Often

C. Often

D. A couple of times

E. Never

3. Please list as many sound names as you can remember

4. Did you hear the sounds produced from the following sources?

1) Bagpipe: Yes / No / Don’t know

2) Bee: Yes / No / Don’t know

3) Bouncing ball: Yes / No / Don’t know

4) Canary: Yes / No / Don’t know

5) Chimpanzee: Yes / No / Don’t know

6) Cicada: Yes / No / Don’t know

7) Coyote: Yes / No / Don’t know

8) Cricket: Yes / No / Don’t know

9) Crow: Yes / No / Don’t know

10) Cuckoo: Yes / No / Don’t know

11) Dog: Yes / No / Don’t know

12) Dolphin: Yes / No / Don’t know

13) Donkey: Yes / No / Don’t know

14) Dove: Yes / No / Don’t know

15) Eagle: Yes / No / Don’t know

16) Elephant: Yes / No / Don’t know

17) Fly: Yes / No / Don’t know

18) Frog: Yes / No / Don’t know

19) Footsteps: Yes / No / Don’t know

20) Goose: Yes / No / Don’t know

21) Gorilla: Yes / No / Don’t know

22) Horse: Yes / No / Don’t know

23) Lamb: Yes / No / Don’t know

24) Laughter: Yes / No / Don’t know

25) Lion: Yes / No / Don’t know

26) Loon: Yes / No / Don’t know

27) Mosquito: Yes / No / Don’t know

28) Mouse: Yes / No / Don’t know

29) Nightingale: Yes / No / Don’t know

30) Owl: Yes / No / Don’t know

31) Parrot: Yes / No / Don’t know

32) Pig: Yes /No / Don’t know

33) Raccoon: Yes / No / Don’t know

34) Rain: Yes / No / Don’t know

35) Rattlesnake: Yes / No / Don’t know

36) Seagull: Yes / No / Don’t know

37) Seal: Yes / No / Don’t know

38) Turkey: Yes / No / Don’t know

39) Woodpecker: Yes / No / Don’t know

40) Accordion: Yes/ No / Don’t know

41) Alarm clock: Yes / No / Don’t know

42) Bike bell: Yes / No / Don’t know

43) Boiling water: Yes / No / Don’t know

44) Camera: Yes / No / Don’t know

45) Opening can: Yes / No / Don’t know

46) Car horn: Yes / No / Don’t know

47) Cash register: Yes / No / Don’t know

48) Chainsaw: Yes / No / Don’t know

49) Chimes: Yes / No / Don’t know

50) Clarinet: Yes / No / Don’t know

51) Coins: Yes / No / Don’t know

52) Cymbal: Yes / No / Don’t know

53) Flute: Yes / No / Don’t know

54) Glass breaking: Yes / No / Don’t know

55) Guitar: Yes / No / Don’t know

56) Door bell: Yes / No / Don’t know

57) Drum roll: Yes / No / Don’t know

58) Harmonica: Yes / No / Don’t know

59) Harp: Yes / No / Don’t know

60) Helicopter: Yes / No / Don’t know

61) Jackhammer: Yes / No / Don’t know

62) Music box: Yes / No / Don’t know

63) Page-turn: Yes / No / Don’t know

64) Phone: Yes / No / Don’t know

65) Piano: Yes / No / Don’t know

66) Ping pong: Yes / No / Don’t know

67) Saw: Yes / No / Don’t know

68) Saxophone: Yes / No / Don’t know

69) Scream: Yes / No / Don’t know

70) Ship: Yes / No / Don’t know

71) Sneeze: Yes / No / Don’t know

72) Snort: Yes / No / Don’t know

73) Shuffling cards: Yes / No / Don’t know

74) Siren: Yes / No / Don’t know

75) Tambourine: Yes / No / Don’t know

76) Train: Yes / No / Don’t know

77) Trumpet: Yes / No / Don’t know

78) Typewriter: Yes / No / Don’t know

79) Wind: Yes / No / Don’t know

80) Zipper: Yes / No / Don’t know

References

Abercrombie, David (1967). Elements of General Phonetics. Edinburgh: Edinburgh University.

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

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset- time. *The Journal of the Acoustical Society of America, 115(6)*, 3171-3183.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52,* 163–187.

Assmann, P. F. , & Summerfield, Q. (2004). "The perception of speech under adverse acoustic conditions," in *Speech Processing in the Auditory System*, Springer Handbook of Auditory Research, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer-Verlag, Berlin), Vol. 18.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Barr, D.J., Levy, R., Scheepers, C., & Tilly, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. *Nature, 403,* 309–312.

Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal-lobe response to vocal sounds. *Cognitive Brain Research 13*, 17–26.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Science, 8,* 129–135.

Bent, T., Kewley-Port, D., and Ferguson, S. H. (2010). Across-talker effects on non-native listeners' vowel perception in noise. *Journal of the Acoustical Society of America, 128*(5)*,* 3142-3151.

Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America, 67,* 648–662.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer [Computer program]. Retrieved from www.praat.org.

Borovsky, A., & Creel, S.C. (2014). Children and adults integrate talker and verb information in online processing. *Developmental Psychology, 50*(5),1600-1613.

Bregman, A. S. (1990). *Auditory Scene Analysis.* Cambridge, MA: MIT Press.

Bradlow, A.R., Torretta, G.M., & Pisoni, D.B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20*, 255 - 273.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics,* 61(2)*,* 206-219.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America, 106*(4), 2074-2085.

Bricker, P.D., & Pruzansky, S. (1976). Speaker recognition. In: Lass, NJ., editor. Contemporary issues in experimental phonetics. New York: Academic Press, 295-326.

Brouwer S., Van Engen K. J., Calandruccio L., Bradlow A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content. *Journal of the Acoustical Society of America, 131,* 1449-1459.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America, 109,* 1101-1109.

Campeanu S., Craik F.I. & Alain C. (2013). Voice congruency facilitates word recognition. *PLoS ONE* .

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language, 32,* 193-210.

Cooke, M. P. (2003). Glimpsing speech. *Journal of Phonetics, 31,* 579-584.

Cooke M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America, 119,* 1562-1573.

Cooke, M., García Lecumberri, M.L.,Barker, J. (2008). The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. *Journal of the Acoustical Society of America, 123,* 414-427.

Cooper, A., Brouwer, S., & Bradlow, A. R. (2015). Interdependent processing and encoding of speech and concurrent background noise. *Attention, Perception & Psychophysics, 77(4),* 1342-1357.

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology, 26,* 274–284.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heading the voice of experience: The role of talker variation in lexical access. *Cognition, 108*, 633–664.

Creel, S. C., & Tumlin, M.A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language, 65*, 264–285.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2012). Word learning under adverse listening conditions: Context-specific recognition. *Language and Cognitive Processes, 27,* 1021-1038.

Creel, S. C. (2012). Preschoolers' use of talker information in on-line comprehension. *Child Development*, *83*(6), 2042–56.

Creelman, C. D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America, 29,* 655.

Cutler, A. (2008). The abstract representations in speech processing. *The Quarterly Journal of Experimental Psychology, 61*(11), 1601-1619.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 496–509.

Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society of London B*, *363,* 1011-1021.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioral evidence. *Philosophical Transactions of the Royal Society B, 364*(1536), 3773–3800.

Durlach, N.I., Mason, C.R., Kidd, G. Jr., Arbogast, T.L., Colburn, H.S., & Shinn-Cunningham, B.G. (2003). Note on informational masking. *Journal of the Acoustical Society of America, 113,* 2984-2987.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*, 224–238.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences, 8*(7), 301–306.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33*(4), 547–582.

Fecteau, S., Armony, J.L., Joanette, Y., & Belin, P. (2004). Is voice processing species-specific in human auditory cortex? *NeuroImage, 23,* 840–848.

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics, 59*, 839–849.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*, 116-124.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3–28.

Fowler, C. A., & Rosenblum, L. D. (1991). Perception of the phonetic gesture. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory* (pp. 33–59). Hillsdale, NJ: Erlbaum.

Fowler, C. A., & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 123–136). Hillsdale, NJ: Erlbaum.

Gaskell, M. G.,& Marslen-Wilson,W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613–656.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science, 23*(4), 439-462.

Gaskell, G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology, 45*, 220–566.

Garcia Lecumberri, M.L., Cooke, M., Cutler, A. (2010). Non-native speech perception in adverse conditions: a review. *Speech Communication, 52*, 864-886.

Garner, W (1974). The processing of information and structure. Potomac, MD: Erlbaum.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325-331.

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory,* *and Cognition, 31*(1), 40-53.

Goldinger, S. D. (1992). Words and voices:Implicit and explicit memory for spoken words (Research on speech perception Tech. Rep. No. 7).Bloomington, IN: Indiana UniversityPress.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on serial recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 152-162.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics, 31*, 305-320.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes: An episodic theory of lexical access. *Psychological Review,105,* 251–279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In J. Trouvain & W. J. Barry (Eds.), Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007) (pp. 49 – 54). Dudweiler, Germany: Pirrot.

Gow, D.W. (2012). The cortical organization of lexical knowledge: A dual lexicon model of spoken language processing. *Brain & Language, 121*, 273–288.

Grossberg,S., Boardman, I. and Cohen, M. (1997). Neural dynamics of variable-rate speech categoriza- tion. *Journal ofExperimental Psychology: Human Perception and Performance,* 23, 481-503.

Hawkins, S., & Smith, R. (2001). Polysp: A polysystemic, phonetically rich approach to speech understanding. *Italian Journal of Linguistics—Rivista di Linguistica 13,* 99-188.

Hecker, M. (1971). Speaker recognition: An interpretive survey of the literature: ASHA Monographs, No. 16.Washington, DC: American Speech and Hearing Association.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Science, 4*(1), 131–138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(12), 67–99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 77-109). Cambridge, MA: MIT Press, Bradford Books.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,*411–428.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*, 19–24.

Howard-Jones, P. A., & Rosen, S. (1993a). The perception of speech in fluctuating noise. *Acustica 78,* 258-272.

Jackson, A., & Morton, J. (1984). Facilitation of auditory word recognition. *Memory* & *Cognition, 12,* 568-574.

Jacoby, L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning* & *Verbal Behavior, 22,* 485-508.

Jacoby, L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory, (Vol 18, pp. 1-47). New York: Academic Press.

Jerger S, Martin R, Pearson D.A, & Dinh T. (1995). Childhood hearing impairment: Auditory and linguistic interactions during multi-dimensional speech processing. *Journal of Speech and Hearing Research, 38,* 930–948.

Jerger S, Pirozzolo F, Jerger J, Elizondo R, Desai S, Wright E, & Reynosa R. (1993). Developmental trends in the interaction between auditory and linguistic processing. *Perception & Psychophysics, 54,* 310–320.

Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *Journal of the Acoustical Society of America,130(3),* 1475-1487.

Jørgensen, S., Ewert, S., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *Journal of the Acoustical Society of America, 134*(1)*,* 436-446.

Ju M., & Luce, P.A. (2006). Representational specificity of within-category phonetic variation in the long-term mental lexicon. *Journal of Experimental Psychology: Human Perception and Performance, 32(1),* 120-138.

Jusczyk, P.W. and Luce, P.A. (2002). Speech perception. In S. Yantis, & H.E. Pashler (Eds.) *Stevens' handbook of experimental psychology*, (pp. 493-536). 3rd edn., Vol. 1. New York: John Wiley and Sons.

Kidd, G. Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (2007). Informational masking. In W. A. Yost, A. N. Popper, and R. R. Fay (Eds.), *Springer Handbook of Auditory*

Research, Vol. 29: Auditory Perception of Sound Sources, (pp. 143-189). New York, NY: Springer.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustment to multiple speakers.*Journal of Memory and Language,56,* 1-15.

Kraljic, T. & Samuel, A.G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13,* 262–268.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology,51,* 141-178.

Kreiman, J. (1997). Listening to voices: Theory and practice in voice perception research. In K. Johnson, & J.W. Mullennix (Eds.), *Talker variability in speech processing,* (pp. 85–108). San Diego, CA: Academic Press.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29,* 98–104.

Levi, S. V. (2015). Talker familiarity and spoken word recognition in school-age children. *Journal of Child Language*, *42*(4), 843–872.

Lidestam, B., Holgersson, J., & Moradi, S. (2014). Comparison of informational vs. energetic masking effects on speechreading performance. *Frontiers in psychology, 5,* 639.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62,* 615–625.

Luce, P. A., & Lyons, E. (1998). Specificity of memory representation for spoken words. *Memory & Cognition, 26,* 708–715.

Luce, P. A., McLennan, C. T., & Charles-Luce, J. (2003). Abstractness and specificity in spoken word recognition: Indexical and allophonic variability in long-term repetition priming. In J. Bowers & C. Marsolek (Eds.), *Rethinking implicit memory* (pp. 197-214). New York: Oxford University Press.

Luce, P. A., & McLennan, C. T. (2005). Spoken word recognition: The challenge of variation. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception,* (pp. 591-609). Malden, MA: Blackwell Publishing Ltd.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10,* 29-63.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25,* 71-102.

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: On the importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15,* 576-585.

MATLAB and Statistics Toolbox (Release 2014b). Computer program. The MathWorks, Inc., Natick, Massachusetts, United States.

Mattys, S.L. & Liss, J.M. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as from artificial normality. *Perception & Psychophysics*, *70*, 1235-1242.

Mattys, S., Brooks, J., and Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology, 59,* 203-243.

Mattys S. L., Carroll L. M., Li C. K. W, & Chan S. L. Y. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech Communication, 52*, 887–899.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101-B111.

McClelland, J. L., & Elman, J. L. (1986). Interactive processes in speech recognition: The TRACE model. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition,* (pp. 58-121). Cambridge, MA: MIT Press.

McClelland, J., McNaughton, B., O'Reilly, R. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419-437.

McLennan, C. T., & González, J. (2012). Examining talker effects in the perception of native- and foreign-accented speech. *Attention, Perception, & Psychophysics, 74,* 824-830.

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31,* 306-321.

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 539-553.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* B32-B42.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30,* 1113–1126.

McQueen, J. M., & Cutler, A. (2010). Cognitive processes in speech perception. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), The handbook of phonetic sciences (2nd ed., pp. 489-520). Oxford: Blackwell.

Mehler, J. (1981). The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society, Series B, 295,* 333-352.

Meyer, M., Zysset, S., von Cramon, D.Y., Alter, K. (2005). Distinct fMRI responses to laughter, speech, and sounds along the human peri-sylvian cortex. *Cognitive Brain Research, 24,* 291-306.

Milberg, W., Blumstein, S., & Dworetzky, B. (1988). Phonological factors in lexical access: Evidence from an auditory lexical decision task. *Bulletin of the Psychonomic Society, 26,* 305-308.

Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America, 22,* 167-173.

Moore, B. C. J. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics 31,* 563–574.

Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Development, 72,* 834-843.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47,* 379390.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365-378.

Naveh-Benjamin M. & Craik F.I. 1995. Memory for context and its use in item memory: Comparisons of younger and older persons.. *Psychology and Aging*, *10* (2), 284-293.

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

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189-234.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23,* 299-325.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,*204-238.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115(2),* 357-95.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60(3),* 355-376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5,* 42-46.

O'Keefe, D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures, 1*, 291-299.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309-328.

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

Pisoni D. B., & Levi S. V. (2007). Some observations on representations and representational specificity in speech perception and spoken word recognition. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 3-18). Oxford: Oxford University Press.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9-32). San Diego, CA: Academic Press.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication, 13,* 109-125.

Pollack, I. (1975). Auditory informational masking. *Journal of the Acoustical Society of America, 57,* S5.

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology, 70,* 1-30.

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.

Roediger, III, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45(9),* 1043-1056.

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *Journal of the Acoustical Society of America, 133,* 2431-2443.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110,* 474-494.

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(5),* 915-930.

Scott, S. K. (2005). Auditory processing – speech, space and auditory objects. *Current Opinion in Neurobiology, 15*(2), 197–201.

Scott, S. K., & Wise, R. J. S. (2004). The functional neuroanatomy of prelexical processing n speech perception. *Cognition, 92*(1–2), 13–45.

Scott, S. K., Rosen, S., Beaman, C. P., Davis, J. P., & Wise, R. J. S. (2009). The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. *Journal of the Acoustical Society of America, 125,* 1737-1743.

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

Sheffert, S. M. (1998a). Contributions of surface and conceptual information to recognition memory. *Perception & Psychophysics, 60,* 1141– 1152.

Sheffert, S. M. (1998b). Format-specificity effects on auditory word priming. *Memory & Cognition, 26,* 591–598.

Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language, 34,* 665–685.

Sommers, M. S. (1999). Perceptual specificity and implicit memory priming in older and younger adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1236–1255.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111,* 1872–1891.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64,* 1358–1368.

Stevens, A.A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research, 18,* 162–171.

Sumner, M., Kim, S. K., King, E., and McGowan, K. (2014). The socially-weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4,* 1 – 13.

Plaut, D., & McClelland, J. L. (2010). Locating object knowledge in the brain: Comment on Bower's (2009) attempt to review the grandmother cell hypothesis. *Psychological Review, 117*(1), 284–288.

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception & Psychophysics, 77*(5), 1674-1684.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), Organization of Memory (pp. 382-402). New York, NY: Academic Press, Inc.

Tulving, E. (1974). Cue-dependent forgetting. *American Scientist*, *62*, 74-82.

Tulving, E., & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memorv. *Psychological Review, 80*, 352-373.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I. *Recognition of backward voices. Journal of Phonetics, 13,* 19–38.

Van Lancker, D., Kreiman, J., & Wickens, T. (1985). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics, 13,* 39 52.

Vaughn, C., & Brouwer, S. (2013). Perceptual integration of indexical information in bilingual speech. *Proceedings of the Acoustical Society of America*. Montreal, Canada.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 333–342.

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.

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.

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.

Watson, C. S. (2005). Some comments on informational masking. *Acta Acoustica, 91,* 502- 512.

Watson, D., Tanenhaus, M., & Gunlogson, C. (2008). Interpreting pitch accents in online comprehension: H/ vs. L + H/. *Cognitive Science, 32,* 1232–1244.

Zue, VW., Schwartz, RM. (1980). Acoustic processing and phonetic analysis. In: Lea, WA., editor. Trends in speech recognition. Englewood Cliffs, NJ: Prentice-Hall, 101-124.