A microscopic analysis of consistent word misperceptions

Tóth Attila Máté

Departamento de Filología Inglesa y Alemana y Traducción e Interpretación

Universidad del País Vasco (UPV/EHU)

Supervised by Prof. Martin Cooke and Dr. Mª Luisa García Lecumberri

2017

To my grandparents ...

Acknowledgements

First and foremost I would like to thank my supervisors, Prof. Martin Cooke and Dr. María Luisa García Lecumberri, as well as Dr. Jon Barker. Their help and guidance have been invaluable during this journey, and I am really grateful for having had the opportunity to work with them. They were truly an inspiration throughout, and I couldn't have managed without their knowledgable suggestions and support. I am also very grateful for having had the chance to be in such a friendly environment both in terms of the lab and Vitoria for the better part of my Ph.D. I have really enjoyed spending these years exploring the Basque Country and would like to thank Zsófi for accompanying me through most of this journey.

I would like to thank my colleagues in Laslab, who were very welcoming and with whom I've shared some great moments with. Thanks to Dr. Julian Villegas for putting me up when I arrived. Also thanks to Dr. Vincent Aubanel, Dr. Jian Gong, Dr. Yan Tang, Rubén Pérez Ramón and Ander Egurtzegi for the discussions, dinners and coffee. Thanks to Edurne for helping me deal with the Spanish bureaucracy beyond the call of duty.

I would like to thank everyone in the INSPIRE training network as well, who are to many too name. I have made some great friends, many professional acquaintances and attended a huge amount of workshops and conferences that were truly inspiring. I am really grateful for all the opportunities this program has afforded me.

Thanks to the guys back home: Ervin, Gábor, Feri and Satya.

Finally, I would also like to thank my whole family, especially my parents for their loving support during this entire process.

Abstract

Despite several decades of effort, our understanding of the way listeners process speech is still lacking, especially in the presence of real-world adversities. In addition to being a fundamental scientific problem, a better understanding of the mechanisms of human speech perception could benefit a variety of applications from hearing prosthetics to automatic recognition systems. One way to gain a better understanding of the human speech code is through the analysis of characteristic misperceptions. Past work has largely taken a macroscopic view by trying to model average intelligibility across a variety of adverse conditions. While certainly useful, these models operate on a level too coarse to provide insights into the underlying auditory processing mechanisms. The analysis of individual perception errors is likely to be much more informative in this regard. The main challenge in analysing misperceptions on an utterance-by-utterance basis whether collected in the lab or from everyday conversations is the inherent variability in perception, which often results in large differences in listener responses to the same physical stimuli. Some stimuli, however, elicit the same erroneous response from the majority of listeners. These consistent misperceptions are of great interest as they provide a solid basis for the analysis of target-confusion error patterns, as well as valuable diagnostic stimuli for the next generation of intelligibility models which aim to provide utterance-level predictions of listener responses. This dissertation presents the elicitation of a large-scale corpus of consistent misperceptions and its analysis from multiple perspectives. In Chapter 2 we present the corpus and its elicitation process. Our lab-based collection involved over 170 listeners screening over 300 000 speech tokens presented in 5 distinct

masker types chosen for their diversity. We also detail the adaptive token pruning and post processing steps used to maximise the yield of the collection, which resulted in over 3200 consistent misperceptions. In Chapter 3 we conduct a signal-independent analysis of the collected corpus. Phonetic transcriptions of the target word and majority confusion are aligned using a method sensitive to both stress and syllable structure. Using this method confusions can be analysed in terms of phone- and syllable-level factors, in addition to word-level characteristics. Through this analysis, we validate several trends reported in naturalistic misperception studies including lexical stress, phonetic similarity neighbourhood and word position, on a corpus with many fewer reliability issues compared to corpora derived from anecdotal collections. While naturalistic studies tend to explain confusion patterns in terms of the target word and the identity of its constituent phonemes, we show that the type of adversity (in our case masker type) has a great effect on the error patterns observed and that misperceptions are better understood in terms of the speech-masker interaction. In Chapter 4 we present a signal-dependent analysis of misperceptions using the glimpse decoding framework. Using as input the speech-noise mixture and an *a priori* segregation mask identifying coherent spectro-temporal regions, the glimpse decoder performs a joint search over the model and segregation space to return the most likely word-segregation pair. In this chapter, we present two distinct approaches to analysing misperceptions based on glimpse decoding. First, we evaluate the decoder's performance in explaining the misperceptions in our corpus and classify well-explained confusions based on the eliciting speech-masker interaction. We find that while a considerable subset of confusions can be traced to complex speech-noise interactions, many confusions arise due to acoustic similarity. An important finding of this analysis is that many confusions involve the misallocation of one or more speech fragments from the masker to the target or vice-versa. This motivates the second approach, where we use the decoder to force-align the speech fragments in the mixture to

the reported confusion. Through this analysis, we find that in speech-based maskers only a small percentage of cases can be attributed to energetic masking, while misperceptions often involve the misallocation of masker glimpses to the target. In chapter 5 we introduce signal modifications to the confusion-eliciting stimuli and reevaluate listeners percepts to determine the origin of the confusion. Modifications were selected to provide release either from energetic or from informational masking and involve SNR modification, signal resynthesis in target glimpses, and modification of target fundamental frequency. Based on the modification that was successful in revealing the target to the listeners we hypothesised the type of masking that caused the original confusion. The glimpse resynthesis condition proved to be the most successful modification in separating energetic and informational masking cases. Shifts in fundamental frequency had little effect on the elicited confusions. A relatively high release from masking was observed for the noise-based maskers in the resynthesis condition, suggesting that confusions in these conditions are not strictly due to simultaneous masking. Through the analysis of consistent misperceptions, this thesis attempts to extend the microscopic approach beyond modelling nonsense syllable confusions. We present an initial attempt at modelling word-level misperceptions from a glimpsing perspective. The understanding gained from constructing increasingly accurate end-to-end models of auditory perception will undoubtedly benefit speech and hearing related applications across the board.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Understanding speech perception

Communicating through speech is one of the most outstanding and unique feats of mankind. Yet the way listeners process speech, especially in less than ideal scenarios, remains an open scientific question. Our lack of understanding of speech perception is perhaps best illustrated by the fact that while we can construct systems that approach or even outperform listeners on other auditory tasks, for example, speaker identification [Hautamaki et al., 2010; Kahn et al., 2011], we have yet to propose a speech recognition system that achieves a performance suitable for practical applications, even in constrained speech tasks and especially in cases where the speech is degraded [Barker et al., 2015; Meyer et al., 2007]. While advances are made continuously — recent years have shown a jump in recognition rates with the introduction of Deep Neural Networks [Dahl et al., 2012; Hinton et al., 2012] — recognition rates for Automatic Speech Recognition (ASR) systems are still shy of human performance. One advantage of understanding how speech perception works in humans is that it can help bridge the human-machine gap in speech recognition and may result in successful commercial systems. A better understanding of human speech perception could also benefit other practical applications such as hearing prosthetics and speech transmission systems.

Commercial benefits notwithstanding, understanding human auditory perception is a valid scientific endeavour in its own right. There has been a consider-

able research focus in understanding speech perception starting from the 1950's. Early work focused on the phonetic level, trying to identify unique acoustic cues for each phonetic segment or determine the smallest perceptual unit. Liberman et al. [1952] found the same noise-burst can be perceived as /p/ or /k/ depending on the adjacent vowel, which contradicts the notion of a one-to-one correspondence between acoustic cues and perceived segments. The smallest perceptual unit has been the subject of an ongoing debate. Cooper [1974] found that it was impossible to separate the formant patterns of the vowel /i/ and the consonant /d/ in the CV /di/. When trying to divide the two phones, listeners either responded with /di/ or a non-speech sound suggesting that the syllable cannot be split into smaller segments perceptually. Savin and Bever [1970] report that listeners detect syllable targets faster than phoneme targets, and argue that the syllable is the primary unit of perception. In other experiments [Cutler et al., 1987; Mills, 1980], however, phonemes were recognised faster than syllables.

Later, the emphasis shifted from the perception of phones to that of entire words. One of the most consistent findings regarding word recognition is that listeners consider and select from multiple word hypotheses while receiving the acoustic input, rather then waiting until the end of the word to make a decision. Parallel activation of words has been shown through priming studies [Goldinger et al., 1989; Shillcock, 1990; Zwitserlood, 1989]. Word onsets consistent with multiple words or embedded words have been used as both intra-modal and cross-modal primes, which ease the recognition of semantically related words. Using eye-tracking [Allopenna et al., 1998; Tanenhaus et al., 1995], researchers have shown that subjects are slower at focusing on the target object when multiple objects are on display that partially match the acoustic input. Gating studies [Grosjean, 1980; Warren and Marslen-Wilson, 1987], in which increasing chunks of the target word are presented, have shown that listeners rule out competing words based on acoustic-phonetic detail before word offset. Several models of spoken word recognition — both theoretical and computational — have been proposed in order to provide a unified explanation of the behavioural findings [Luce and Pisoni, 1998; Marslen-Wilson and Welsh, 1978; McClelland and Elman, 1986; Norris, 1994]. There is agreement among these models in the activation and competition of word candidates. However, they differ in their input representation,

as well as the assumptions and implementation details behind activation and competition [Weber and Scharenborg, 2012].

Another line of research focuses on the physiology of hearing. Models of peripheral auditory processing have been developed to simulate the cochlear response to an incoming sound source or speech signal [Allen, 1985; Delgutte, 1986; Ghitza, 1992]. Less is known about the more central processing taking place in the auditory cortex. Studies on brain lesions have shown that damage to certain brain regions, such as Broca's and Wernicke's area, are associated with problems in speech production and perception respectively. More recently, emerging brain imaging techniques have allowed the examination of the brain during speech perception. Electroencephalography (EEG) studies have measured event-related potentials in response to violations of semantic [Kutas et al., 1980] and syntactic [Osterhout and Holcomb, 1992] expectations. Techniques such as positron-emission tomography (PET) and functional magnetic resonance imaging (fMRI) have also been used to study central auditory processing. These studies have established the role of the left inferior frontal lobe in semantic and syntactic processing, as well as the role of the right hemisphere in understanding context, figurative meaning and processing prosodic information [Bookheimer, 2002]. While increasingly detailed information is available on the areas responsible for certain functions, given the complexity and plasticity of the brain it is hard to draw conclusions on the exact processing mechanisms that occur in the activated brain regions.

A third approach aims to understand speech perception through its errors. A perception error is when a listener misperceives an utterance which was articulated clearly and correctly, possibly due to an internal or external disturbance. Errors in the perception process can be highly informative about the underlying mechanisms taking place. The first collection of misperceptions dates back to the end of the 19th century with the work of Meringer [Meringer, 1908; Meringer et al., 1895] based on a collection of 47 slips in German. He reports the robustness of the stressed vowel and finds that consonants are more error prone then vowels. Later, several such 'slip of the ear' studies [Bond, 1999b; Browman, 1980; Garnes and Bond, 1980; Labov, 2011; Tang and Nevins, 2012] were carried out, analysing speech misperception corpora compiled from anecdotal reports of 'slips' in everyday conversations. Most often these analyses involve the manual or automatic

alignment of the target and misperceived word and examination of segmental error patterns. Garnes and Bond [1980] published a collection of approximately 1000 misperceptions. They confirmed the robustness of the stressed vowel and showed the relevance of misperceptions to several other linguistic phenomena. From her collection of 200 misperceptions, Browman [1980] created CVC syllabic composites for stressed and unstressed syllables in monosyllabic and polysyllabic words. She found that error rates decreased from initial through medial to final syllables in polysyllabic words and onset through nucleus to coda position in syllables. She also reported lower error rates in the stressed syllable. By merging the above corpora to their own extensive collection [Tang, 2015; Tang and Nevins, 2012] created a large-scale corpus of naturalistic word misperceptions. They used an automatic alignment algorithm [Needleman and Wunsch, 1970] to align the segments of the target and confused word, which they modified slightly to align by syllables. They investigated factors starting at the phonetic level up to the word level, such as phonetic identity, stress, word position, adjacency, word frequency and neighbourhood density. Their results suggest that salient segments like the stressed syllable, vowels or consonants with high sonority are more likely to be correctly perceived.

The main argument in favour of the above studies is their ecological validity. However, several concerns have been raised about the reliability of the collection process. Cutler [1982] question the validity of speech error data in general, citing issues such as uncontrolled sampling, with only the most memorable slips reported, or potential confounds, such as mispronunciation or the listener recovering the meaning from context and not reporting the misperception.

In addition to collections of misperceptions 'in the wild', misperceptions have been collected in the laboratory. Different types of adverse conditions have been used to elicit misperceptions in the lab, including fast speech [Vitevich, 2002], faint speech [Cutler and Butterfield, 1992] as well as noise [Cooke, 2009; Tóth et al., 2015]. Laboratory collection allows for better control over potential confounds, such as the homogeneity and hearing status of the listener population or external disturbances such as noise and reverberation. It is also more reproducible as the misperception-inducing stimulus can be recorded. The work in this thesis falls under this third approach, focusing on analysing a large-scale collection of

noise-induced word misperceptions.

1.2 Speech perception in noise

With the increasingly widespread use of communication systems such as the telephone, the intelligibility and quality of speech over a transmission channel with noise and other possible distortions became an increasingly important research problem. While subjective listening tests provide the most direct and accurate way of quantifying intelligibility, these can be costly and time-consuming. Instead, Fletcher and Galt [1950] pioneered a new approach, creating the first method to predict speech intelligibility. Their work was later published by French and Steinberg [1947] and the calculations further refined by Kryter [1962] into the articulation index (AI). A standardised measure of intelligibility, the speech intelligibility index (SII) [ANSI, 1997] also evolved from this approach. AI based metrics were devised to handle additive noise. Later, the speech transmission index (STI) [Steeneken and Houtgast, 1980] was proposed to predict intelligibility for convolutional noise such as reverberation and other non-linear distortions by quantifying the reduction of speech modulations at each modulation frequency. The above intelligibility models make use of the long-term speech spectrum and consequently are not well suited for measuring intelligibility in time-varying maskers. To address this issue, several models have been proposed that calculate intelligibility in short temporal windows which are subsequently averaged [Rhebergen et al., 2005; Taal et al., 2010]. While objective intelligibility models match listener intelligibility for a variety of adverse conditions increasingly well, they provide no information on the type of confusions listeners make, or their cause.

More recently an alternative approach has been proposed which aims to model listener responses on a more fine-grained level. This approach, labelled 'microscopic' [Cooke, 2006], can be understood in several sense of the word [Jürgens, 2010]. First, instead of trying to provide a global intelligibility estimate, it aims to establish a mapping between the acoustic stimulus and the resulting percept for each *individual* utterance. Phatak and Allen [2007] and Phatak et al. [2008] studied listeners' confusions patterns of CVs in speech-shaped and white noise, aiming

to identify the auditory events corresponding to each phone. Li et al. [2010] used a three-dimensional deep search method to control the available speech cues in time, frequency and level by time truncation, high/low-pass filtering and noise masking to identify the relevant events for stop consonants. Varnet [2013] used the classification image technique to show that the second formant transition is key to be able to distinguish phones /b/ and /d/ in noise. Listeners were asked to identify one of the two signals, while random noise was added to each trial for a large number of trials. Relevant speech cues were identified by creating a correlation map between the noise field and the responses.

Another approach to 'microscopic' modelling is to create an end-to-end model of the auditory system, usually by connecting an auditory model front-end to a pattern recognition back-end, thus 'mimicking' peripheral and central auditory processing. This approach was first used by Ghitza [1994] as a means to evaluate the performance of auditory models. He compared the distribution of error patterns of the auditory model of interest connected to a Hidden Markov Model back-end to those of human subjects on a diagnostic rhyme test. Another example of this approach was proposed by Holube and Kollmeier [1996] who connected the auditory model proposed by Dau et al. [1996] to a dynamic time-warping recogniser, in order to predict recognition scores for normal hearing and hearing impaired listeners, also on a rhyme test. They found that their approach produced intelligibility results comparable to the AI or STI. However they did not analyse confusion patterns on a phone level. Jürgens and Brand [2009] extended the above approach by investigating different perceptual distance measures for the recogniser and evaluating model predictions on a phoneme level instead of overall intelligibility. Using a spectro-temporal excitation pattern as input, Cooke [2006] applied missing data speech recognition treating only time-frequency regions of high local SNR (i.e. glimpses) as speech evidence to identify consonants in a range of fluctuating maskers. The glimpsing model was found to be a good predictor of average intelligibility, but model predictions differed substantially from listeners' responses on a microscopic level.

Such functional models of the auditory system promise insights into human speech perception. However there are several issues involved. For example Zaar and Dau [2015] showed that listeners themselves are a considerable source of vari-

ability in their study of sources of variability in consonant perception. Focusing on consistent confusions — stimuli to which the majority of listeners respond with the same mistake — can reduce this variability. These cases can serve as valuable diagnostic stimuli for such 'microscopic' models as argued by Cooke [2009].

As shown above, most microscopic modelling approaches so far focus on phone level confusions. Also, when a stimulus elicits highly variable listener responses, it is unclear what the output of the model should be. To address these issues, in this work we aim to analyse a large-scale corpus of consistent word misperceptions. By focusing on misperceptions at the word as opposed to the phone level, we factor in some of the top-down processes that undoubtedly play a role in speech perception. In addition, by focusing the analysis on consistent confusions, listener variability is greatly reduced. Through both a signal dependent and independent analysis of confusions, this work aims to provide a first step towards a model that is capable of explaining listener misperceptions on an utterance-by-utterance basis.

Chapter 2

Elicitation of word misperceptions in noise

2.1 Introduction

[1] The development and testing of intelligibility models requires appropriate psychoacoustic data from listeners. Data used to test and evaluate macroscopic intelligibility models is often elicited using matrix-style sentences with a predictable syntactic structure presented in a closed-set task [Hagerman, 1982; Wagener et al., 2003]. Models are then fit based on listeners' average word recognition rates across the condition of interest.

Evaluating microscopic models of speech intelligibility is much less straightforward because of the variability present in individual listener responses. The distribution of responses to the same physical stimuli can show a significant spread across listeners. In addition, even the same listener can respond differently when presented with the exact same stimulus a second time [Zaar and Dau, 2015]. In the past, several methods have been introduced to capture response variability, including the confusion matrix [Miller and Nicely, 1955] or the confusion pattern [Allen, 2005; Phatak and Allen, 2007]. The confusion pattern — a tool representing the evolution of the response distribution for a single consonant across SNR values — was introduced by Allen [2005]. In theory, these methods could

[1] A version of this chapter has been published in JASA-EL.

be used to represent the variability stemming from across-listener differences as well [Zaar and Dau, 2015].

However, while these methods can work for nonsense syllable confusions in closed-set tasks, where the number of response alternatives is limited, it is unclear how these methods would generalise to higher-level linguistic units (e.g. words), where the number of response alternatives is much larger, especially in an open-set paradigm. One reason why microscopic investigations have been restricted to nonsense syllable confusions so far could be the lack of clear procedure through which microscopic models providing word-level predictions could be validated.

Consistent confusions — cases where a given stimulus elicits the same erroneous response from a large listener group — provide an attractive option for the development and testing of microscopic models. For these misperceptions, individual variability is minimised; thus they provide a clear target for microscopic models to aim for. In this chapter, we present our approach to collecting a large-scale dataset of robust, noise-induced word misperceptions in Spanish. In addition, we outline the measures we took to increase token finding efficiency and maximise the number of useful confusions.

In their seminal paper, Miller and Nicely [1955] noted that one of the reasons that individual phone confusions received little attention compared to average intelligibility prediction up to that point, could be the cost involved in collecting perceptual data suitable for such an investigation. Analysing perceptual confusions at the phone-level requires a considerable amount of data because the empirical probability of each possible phone confusion needs to be estimated. Further, as confusions of higher level speech units are investigated, the number of factors that could potentially impact the percept increases. Consequently, a large-scale confusions corpus is advantageous because it can provide statistical backing for the trends observed for a larger number of factors and interactions.

Considering the above, researchers have been continuously seeking more efficient ways of collection to increase the size of their datasets. Labov [2011] distributed pre-printed collection cards among linguistics faculty and sent frequent email reminders to jot down any misperceptions encountered during the day. Their collection process yielded around 870 misperceptions. Tang and Nevins [2012] assigned the collection of 5-10 confusions a week to linguistics students

enrolled in a course on speech misperceptions. Over the course of two years, their collection process resulted in 2857 misperceptions. To further increase the size of their corpus, Tang [2015] added available naturalistic collections to their own, standardising them into a single format.

A recent method — Web-based crowdsourcing — allows the scaling of the above approach even further. With the widespread use of the Internet, it has become possible to conduct perception experiments online, through the browser and headphones of the user, instead of in a formal lab setting. Such crowdsourced perception experiments have the potential to supply data from a large and varied sample of the population efficiently with a low financial investment. Several successful experiments conducted online in a variety of disciplines indicate the viability of this approach [Blin et al., 2008; Honing, 2006; Koekemoer et al., 2010].

Though certainly promising, there are also several drawbacks to the web-based crowdsourcing approach. First, the success of the collection depends on participants being aware of the experiment and interested enough to take part. Second, confounding factors such as the presentation quality and acoustic environment, as well as the language proficiency and hearing status of the listener are a lot more difficult to control compared to a laboratory setting

Cooke et al. [2011] evaluated the crowdsourcing approach by conducting a collection of consistent word misperceptions both online and in the lab. In an effort to separate listeners contributing low-quality data, online listeners underwent both an objective and a subjective assessment via test tokens and an online questionnaire. While Cooke et al. [2011] found that most consistent misperceptions were supplied by the formal group (129 exemplars), online listeners who met both subjective and objective criteria also contributed a significant amount of confusions (85 exemplars), demonstrating that crowdsourcing with the appropriate control measures is a viable method of collection. Interestingly, only about a quarter of confusions collected in the lab overlapped with the ones collected online, suggesting that some of the online confusions are due to adversities not present in a formal setting, such as a low-quality audio chain. Cooke et al. [2011] proposed that crowdsourcing could operate as a filter to preselect tokens which can later be retested under laboratory conditions. Considering that the formal

collection method yielded the highest number of consistent confusions and that data collected this way supports stronger scientific claims, we have opted to conduct the collection directly in the lab.

In this study, we aim to use a range of masker types with varying characteristics, as we suspect that maskers with different properties could elicit different types of misperceptions. In particular, we selected maskers that differ in their informational and modulation content. As our primary goal is generating many confusions and considering that confusions are most likely to arise from the phonological neighbourhood of the target word, we selected target words with a neighbourhood that is not the sparsest. Marian et al. [2012] have shown in their cross-linguistic study of phonological neighbourhood, that Spanish had the sparest neighbourhoods on average among the five languages studied, namely English, Dutch, French and German. In addition, the number of phonological neighbours diminished with word length. Thus we opted to use 1-3 syllable words as target utterances, to avoid words which — due to lack of competition — would potentially not generate any confusions. In the following, we present the details of our collection experiment.

2.2 Corpus elicitation

2.2.1 Speech material

Four talkers, two male and two female, were recorded reading a word list containing 3968 of the most frequent Spanish words of up to three syllables. Talkers were trained to avoid list intonation. Recordings took place in a sound-attenuated studio using an AKG 4500 microphone and RME Fireface 800 analogue-to-digital interface. The resulting recordings were manually segmented and downsampled to 16 kHz. Some 15 753 items remained after removal of 119 mispronounced or noise-contaminated items.

	Masker	Speech	Stationary	SNR range (dB)
SSN	Speech shaped noise	×	×	-7 to -4
BMN1	Speech modulated noise	×	✓	-13 to -7
BMN3	3-talker babble modulated noise	×	✓	-8 to -3
BAB4	4-talker natural babble	✓	✓	-3 to +1
BAB8	8-talker natural babble	✓	✓	-4 to +1

Table 2.1: *Maskers used in the experiment. The column headed 'Speech' indicates those maskers containing natural speech signals.*

2.2.2 Maskers

With the goal of promoting word misperceptions due to both energetic and informational masking, five maskers were used with varying modulation and information content (Table 2.1). One masker was stationary (SSN) while the rest differed in their depth of temporal modulation. In two cases (BAB4 and BAB8) maskers were composed of natural speech material, while for BMN1 and BMN3 the envelope of competing speech and 3-talker babble was used to modulate a speech-shaped noise carrier. Speech-plus-noise stimuli were presented to listeners at SNRs within the masker-specific ranges shown in the table. These values were chosen based on previous work [Cooke, 2009] and pilot tests as likely to elicit consistent misperceptions. All 5 maskers were generated using the speech material described in 2.2.1.

2.2.3 Participants

A total of 172 listeners provided responses to words in noise. Listeners were native monolingual Spanish or bilingual Spanish-Basque speakers studying at the University of the Basque Country in Vitoria, Spain (mean age 22 years, s.d. 4.8). Apart from three listeners from Spanish-speaking countries in South America, all participants were born in the Basque Country. Listeners gave written consent for anonymous use of their responses and were paid for their participation.

2.2.4 Adaptive stimulus pruning

Since consistent word confusions are quite rare, even in noise, the elicitation procedure employed adaptive token pruning techniques to decide which speech-in-noise

tokens were worth pursuing and to rapidly identify tokens which were unlikely to result in confusions and remove them from presentation. Corpus elicitation made use of a heuristics-based pruning technique designed to remove stimuli deemed unlikely to result in consistent misperceptions. The number of identical misperceptions for each stimulus was monitored online and an automated decision made following presentation as to whether to remove the stimulus from further consideration. Stimuli were pruned if any of the following conditions held:

- L1 listeners in a row identified the stimuli correctly in the first N presentations, or L2 listeners in a row identified the stimuli correctly after N presentations
- the responses of the first L3 listeners were all different
- the token had been presented Nmax times, at which point it was marked as exhausted

Tokens were discarded in the first two cases. Parameter values L1 = 2, L2 = 3, L3 = 4, N = 8, Nmax = 15 were chosen after pilot studies demonstrated their efficiency in maximising token turnover without discarding potentially-interesting items. For each pruned stimulus a replacement was generated online using the same SNR, masker type and speaker as the pruned stimulus. The replacement word and masker fragment were chosen at random. Note that while the pruning procedure might inadvertently remove a potential misperception some of the time, these losses are outweighed by the efficiency gains in discovering misperceptions. Indeed, subsequent analysis indicated that adaptive pruning enabled a near-tripling of the rate of discovery of consistent misperceptions.

2.2.5 Procedure

Over the course of two non-contiguous sessions lasting approximately one hour each, listeners identified blocks of 100 words in each of 20 conditions made up from all combinations of the four talkers and five maskers. Within each block, the target talker and masker type were held constant, and words were mixed with noise in a descending sequence of SNRs. For the first 5 stimuli, the SNR

decreased linearly from +30 dB to the upper SNR value shown in Table 2.1 to accustomise the listener to the target talker and masker type. For the remaining 95 stimuli the SNR was set randomly in the ranges corresponding to the masker type and presented in decreasing order of SNR, the goal being to explore a range of SNRs without large jumps between stimuli. As a consequence of pruning, the sequence of stimuli presented to each participant was assembled online and hence differed from listener to listener. The masker led and lagged the speech by 200 ms and 20 ms linear ramps were applied to the mixed token prior to presentation. Participants heard stimuli through Sennheiser HD 380 pro headphones at 75 ± 1.5 dB(A) while seated in a sound-attenuating booth. Listeners were instructed to type a single word in response to each stimulus although on a small proportion of occasions (1.1%) listeners typed more than one word. Listeners typed their responses into a textfield in a custom Java applet.

2.2.6 Postprocessing

Listeners' responses were subject to a number of post-processing steps designed to maximise the number of useful misperceptions. First, since on many occasions participants omitted stress marks or the diacritic in ñ, such words were identified and replaced whenever unambiguously possible (e.g., 'máximo' [maximum] for 'maximo', 'baño' [bath] for 'bano'). Second, words with orthographic errors (e.g., 'abestruz') but which resulted in a phoneme sequence identical to a unique existing word (e.g., 'avestruz' [ostrich]) were corrected. Finally, homophones (e.g., 'hola' [hello] and 'ola' [wave]) were replaced by the most frequent form. These steps were performed automatically using a combination of the GNU spell checker *aspell*, a rule-based Spanish semi-phonemic transcriber *HAPLO*, and the *CREA* Spanish word frequency list published by the Spanish Royal Language Academy [REAL, 2008]. Semi-phonemic (i.e., intermediate between broad and narrow) transcriptions were used since Spanish plosive realisations differ in a largely systematic manner according to the phonetic context. Further, words contrasting in the lateral versus central approximants ('ʎ', 'j') were treated as homophones since most Spanish speakers do not distinguish them.

To complement semi-phonemic transcriptions, syllable boundaries and stress

Table 2.2: *Counts of consistent misperceptions per test condition.*

Speaker	Gender	Masker					Totals	%
		SSN	BMN1	BMN3	BAB4	BAB8		
S1	M	176	222	223	128	119	868	27.00
S2	F	143	200	192	193	151	879	27.34
S3	F	122	171	159	137	64	653	20.31
S4	M	184	201	170	149	111	815	25.35
Totals		625	794	744	607	445	3215	100.00
%		19.44	24.70	23.14	18.88	13.84	100.00	

were obtained using *TIP* [Hernández-Figueroa et al., 2013], a Spanish word syllabification tool based on morphological analysis. Syllable boundaries are marked with a period, while '!' denotes the start of the stressed syllable. In addition, the phoneme alignment between target and misperception was computed using dynamic programming with a constraint that enforced alignment of consonants to consonants and vowels to vowels. Alignment costs for insertions and deletions were set to 7, while substitutions had a cost of 10.

2.3 Corpus description

Some 308 152 responses were collected during the elicitation process. In all, 53 039 individual speech-in-noise tokens were generated, of which 9288 survived pruning and were heard by at least 15 listeners. Of these, a minimal level of listener agreement of 6 listeners was applied in order to produce the final corpus. Some 3215 misperceptions meet this criterion and jointly make up the corpus of consistent misperceptions.

Table 2.2 summarises the number of misperceptions obtained in each test condition. All speaker/masker combinations contributed substantial numbers of misperceptions to the corpus, with somewhat more resulting from the two babble modulated noise conditions.

Figure 2.1 visualises counts of misperceptions as a function of phoneme alignment distance and consistency (expressed as the proportion of listeners reporting the same misperception), plotted separately for each of the five masker types. This plot suggests that while simple misalignments are frequent – 45% of misper-

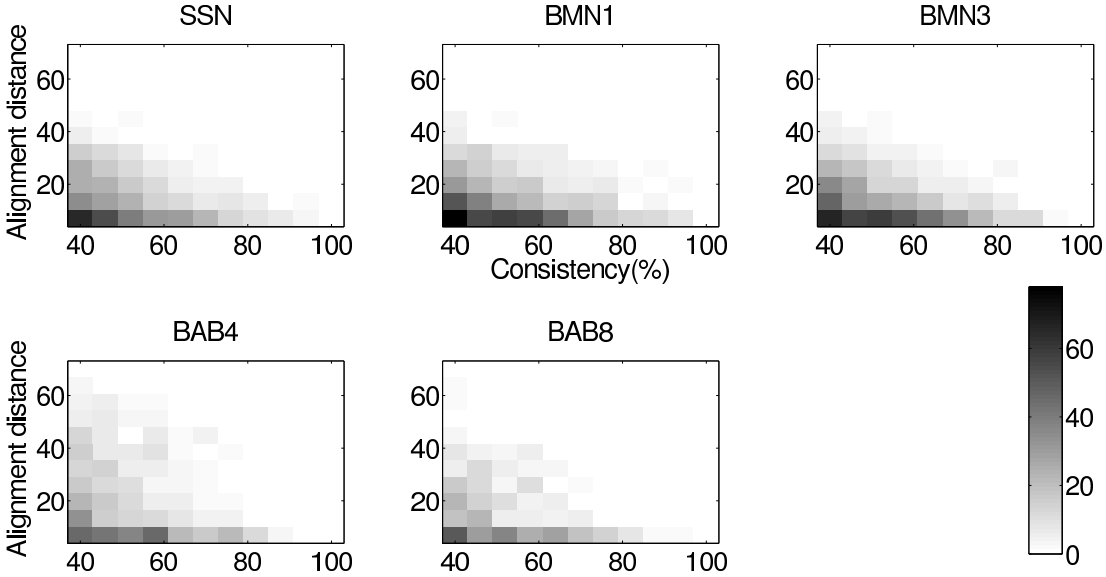


Figure 2.1: *Counts of misperceptions as a function of consistency and phoneme alignment distance for each masker.*

ceptions involve the insertion, deletion or substitution of a single phoneme – more complex confusions are also present. Highly-consistent complex misperceptions are less common. However, for the two natural babble maskers, and especially for the 4-talker case, such misperceptions exist, possibly due to the recruitment of phonetic material from speech-based maskers. We explore this issue further in subsequent chapters.

The corpus contains a substantial number of near-bimodal cases: on 189 occasions the number of listeners reporting the majority misperception differs from the second most frequent response (which might be the target) by two, 110 differ by one, while in 26 cases they are equally consistent.

2.4 Discussion

Studies investigating perception errors either rely on misperceptions collected from everyday conversations or confusions elicited in the lab. Both approaches, however, present certain disadvantages. Issues with anecdotal collections include uncontrolled sampling, the inability to record the acoustic environment for analysis and the many possible confounding factors associated with a naturalistic

setting. At the same time, experimental collections are often criticised for their lack of ecological validity. Consistent confusions present an interesting compromise between the two approaches. On the one hand, they can be elicited in a controlled laboratory environment and stimuli can be recorded for analysis and replication. On the other hand, the likelihood of listeners responding to a given stimulus with the same error in an open-set task is vanishingly small, suggesting that these misperceptions are characteristic of the speech perception process in general.

In this chapter, we have presented our approach to the collection of a corpus of consistent word misperceptions in Spanish. Through a listener cohort of considerable size, as well a set of heuristics aimed at maximising the confusion yield, a large-scale collection of consistent misperceptions was achieved. The methodology presented in this chapter has since been replicated in English [Marxer et al., 2016] and, on a smaller scale, in Dutch [Scharenborg et al., 2014].

The adaptive token pruning process led to a significant increase in token finding efficiency. While it is possible that the pruning process removed some tokens which would have otherwise lead to a consistent confusion, the overall gain in token finding efficiency justifies this approach. The parameters of the pruning were determined empirically in pilot studies. While other heuristics could also be considered, adding further criteria to the token pruning process would most likely lead to diminishing returns.

All talkers and maskers contributed a significant number of confusions. It has been shown that intelligibility amongst talkers varies considerably, even for healthy native speakers [Barker and Cooke, 2007]. This could explain why one of the female talkers contributed noticeably fewer confusions compared to the rest. While all maskers contributed a substantial amount of confusions, the SNR ranges conducive to generating misperceptions were quite different for each masker. In line with previous studies [Brungart, 2001; Festen and Plomp, 1990], we found that speech modulated noise was the least effective masker (requiring the lowest SNR range), while babble maskers were the most effective.

Understandably, many confusions are similar to the target, differing in only one or two phone edits. At the same time, Figure 2.1 illustrates that a considerable amount of complex cases are included in the corpus as well. These cases

are presumably the most interesting, since a high target-confusion edit distance could suggest that these misperceptions arise from a complex interplay between the masker and the target. In the next chapter, we start our analysis of the consistent confusions corpus from a signal-independent perspective. The corpus can be found at http://laslab.org/resources/confusions and is released under the Creative Commons CC BY licence.

Table 2.3: *Example corpus entry for the word "baño" [bath] in 4-talker babble noise, misperceived by 10 out of 15 listeners as "España" [Spain].*

Field	Description	Example
ID	integer used to identify the speech waveform corresponding to the entry	35877
Length	speech signal length in samples	8003
Masker	one of [SSN, BMN1, BMN3, BAB4, BAB8]	BAB4
Onset	starting location of the masker fragment within the masker waveform; along with the Length field this can be used to extract the masker waveform	562883
SNR	signal-to-noise ratio in dB	-0.545
Speaker	one of [s1, s2, s3, s4]	s2
Target	orthographic representation of target word	baño
Raw	raw responses prior to post-processing, one per listener	espana|baño|espana| espana|baño|bano| espana|españa ...
Responses	responses following post-processing, collected into groups; nonwords are identified with an asterisk; the first entry is the majority misperception	españa|baño| espainia*| baino*
N-Listeners	number of listeners who heard the token	15
Counts	for each processed response, in decreasing order	10 3 1 1
Confusion	most frequently-reported response	españa
Consistency	number of listeners reporting majority misperception	10
Target-X-Sampa	sequence of phonemes corresponding to the target in X-SAMPA notation with syllable boundaries and stress marked	! b a . J o
Target-IPA	sequence of phonemes corresponding to the target in IPA notation with syllable boundaries and stress marked	! b a . ɲ o
Target-frequency	normalised frequency (number of occurrences per 10^6 word-forms) of target word according to word-frequency list CREA [REAL, 2008]	44.64
Confusion-X-Sampa	as for *Target-X-Sampa*	e s ! p a . J a
Confusion-IPA	as for *Target-IPA*	e s ! p a . ɲ a
Confusion-frequency	as for *Target frequency*	525.66
Phoneme-distance	alignment distance computed using dynamic programming string alignment	34

ID	Target	Confusion	TargetIPA	ConfusionIPA	Masker	Cons. (%)
487	ladrones	ladrón	l a ! ð ɾ o . n e s	l a ! ð ɾ o n	BAB8	73
512	cobro	joven	! k o . β ɾ o	! x o . β e n	BAB8	47
538	último	últimos	! u l . t i . m o	! u l . t i . m o s	BAB8	53
543	doblar	leche	d o ! β l a ɾ	! l e . ʧ e	BAB4	60
556	estamos	cristal	e s ! t a . m o s	k ɾ i s ! t a l	BAB4	67
583	ésta	esto	! e s . t a	! e s . t o	BMN3	40
588	nombrar	sembrar	n o m ! b ɾ a ɾ	s e m ! b ɾ a ɾ	BAB8	40
589	escuche	escucha	e s ! k u . ʧ e	e s ! k u . ʧ a	BMN3	80
599	jurar	curar	x u ! ɾ a ɾ	k u ! ɾ a ɾ	BAB8	40
609	echo	mucho	! e . ʧ o	! m u . ʧ o	BMN3	40
619	gancho	ancho	! g a ɲ . ʧ o	! a ɲ . ʧ o	BMN3	47
627	caliente	calientes	k a ! l je n . t e	k a ! l je n . t e s	BAB4	53
629	casi	casa	! k a . s i	! k a . s a	BAB4	87
633	intenta	atenta	i n ! t e n . t a	a ! t e n . t a	BAB4	67
640	muerta	muerte	! m we ɾ . t a	! m we ɾ . t e	BAB8	47
645	valen	vale	! b a . l e n	! b a . l e	BAB4	60
652	vena	pena	! b e . n a	! p e . n a	BAB4	73
662	odian	odio	! o . ð ja n	! o . ð jo	BAB4	40
673	poco	coco	! p o . k o	! k o . k o	BAB4	47
704	caros	picaros	! k a . ɾ o s	p i ! k a . ɾ o s	BAB4	80
731	dirige	manzana	d i ! ɾ i . x e	m a n ! θ a . n a	BAB4	53
732	puntas	apuntas	! p u n . t a s	a ! p u n . t a s	BAB8	40
737	pocos	poco	! p o . k o s	! p o . k o	BMN3	47
759	suenan	suenas	! s we . n a n	! s we . n a s	BAB8	67
770	borde	bordes	! b o ɾ . ð e	! b o ɾ . ð e s	BAB4	47
779	mancha	escarcha	! m a ɲ . ʧ a	e s ! k a ɾ . ʧ a	BAB4	47
783	blusa	estaré	! b l u . s a	e s . t a ! ɾ e	BAB4	53
791	llegamos	digamos	ʝ e ! ɣ a . m o s	d i ! ɣ a . m o s	BAB4	40
800	sobrio	soja	! s o . β ɾ jo	! s o . x a	BAB4	53
807	iras	vidas	! i . ɾ a s	! b i . ð a s	BMN3	47
808	permiso	estaré	p e ɾ ! m i . s o	e s . t a ! ɾ e	BAB4	53
811	primos	chicos	! p ɾ i . m o s	! ʧ i . k o s	BAB4	67
814	puras	curas	! p u . ɾ a s	! k u . ɾ a s	BMN3	40
818	base	básico	! b a . s e	! b a . s i . k o	BAB4	80
847	muslo	musgo	! m u z . l o	! m u z . ɣ o	BMN3	60
875	vestido	vestir	b e s ! t i . ð o	b e s ! t i ɾ	BMN3	47
876	perdona	perdón	p e ɾ ! ð o . n a	p e ɾ ! ð o n	BMN3	73
881	decente	decir	d e ! θ e n . t e	d e ! θ i ɾ	BMN3	40
888	parezca	pared	p a ! ɾ e θ . k a	p a ! ɾ e ð	BMN3	80
890	huella	cuello	! we . ʝ a	! k we . ʝ o	BMN3	80
915	sangra	sangre	! s a ŋ . g ɾ a	! s a ŋ . g ɾ e	BMN3	73
916	cuido	ruido	! k wi . ð o	! r wi . ð o	BMN3	47
921	visto	vistos	! b i s . t o	! b i s . t o s	BAB8	40
929	rápida	rápidos	! r a . p i . ð a	! r a . p i . ð o s	BAB8	73
931	centros	centro	! θ e n . t ɾ o s	! θ e n . t ɾ o	BMN3	60
932	vieron	hierro	! b je . ɾ o n	! je . r o	BMN3	47

Table 2.4: *Examples of the Spanish confusions corpus. Columns correspond to confusion ID number, orthographic and phonetic transcriptions of the target and misperceived word, as well as the percentage of listeners who reported the majority confusion. Further examples are provided in Appendix A.*

Chapter 3

Signal-independent analysis of misperceptions

3.1 Introduction

The ultimate goal of microscopic intelligibility modelling is to predict listener responses to speech stimuli in adverse conditions at the utterance level. The exact way in which the adversity masks, distorts or otherwise interferes with the given speech signal will determine the acoustic evidence available to listeners and will shape their percept to a large extent. Thus, to allow for such fine-grained predictions, the input to microscopic models is often some sort of signal-level representation of the degraded stimulus. While sensory information is clearly central to the perception process, signal-independent factors have also been known to play a role [Benkí, 2002]. For example, lexical characteristics, such as word frequency, familiarity and phonological neighbourhood density have been shown to influence perception [Boothroyd and Nittrouer, 1988; Felty et al., 2013; Luce and Pisoni, 1998]. In addition, asymmetric confusions of consonants such as /v/-/ð/ [Miller and Nicely, 1955] and /θ/-/f/ [Miller and Nicely, 1955; Tang and Nevins, 2012] could indicate the existence of perceptual biases at the phone level.

When facing uncertainty resulting from adverse conditions, listeners might rely on these signal-independent factors even more. Both Boothroyd and Nittrouer [1988] and later Benkí [2002] have shown that the recognition of mono-

syllabic words in noise is more accurate than that of nonsense syllables. Using a phonetically balanced set of words and nonsense syllables, they demonstrated that phones in nonsense syllables are perceived largely independently, while the perception of monosyllabic words benefits from a clear contextual advantage. In contrast, in their study of speech recognition under a processing load, Mattys et al. [2009] found that interfering noise with a substantial energetic masking component hinders access to lexical representations and caused listeners to fall back on acoustic cues for recognition.

The aim of this chapter was to investigate how noise-induced misperceptions are affected by factors independent from the particular speech-noise interaction. Using a custom procedure exploiting the robustness of the stressed vowel, we aligned the phonetic transcriptions of the target and misperceived word, which allowed us to investigate the effects of signal-independent factors on misperceptions across multiple levels of speech units. We also evaluated how the type of masker used for elicitation affected the resulting misperceptions, and contrasted our findings to the trends reported in previous naturalistic and experimental misperception studies.

The approach of the present chapter is comparable to previous studies analysing misperceptions collected in the wild [Bond, 1999b; Browman, 1980; Tang, 2015; Tang and Nevins, 2012]. When the collection of confusions is compiled from anecdotal reports, recordings of the utterance and its acoustic environment are often unavailable. This constrains studies investigating slips of the ear to a signal-independent analysis, which often involves contrasting the characteristics of the intended and perceived utterance and identifying the changes between them. In order to achieve this, typically, these studies align the phonetic transcriptions of the target and the misperceived word, either manually or via an algorithm. This step permits the analysis of perceptual error patterns below the word level.

Previous research has investigated misperceptions in the laboratory as well. Miller and Nicely [1955] were the first to advocate a more detailed examination of individual perception errors, as they argue that listener confusions are far from random and that the confusion patterns for each speech sound — which up to that point have been masked by the focus on global error rates — are highly informative about the underlying perception process. Their seminal work inspired

many further studies investigating consonant and vowel confusions [Dubno and Levitt, 1981; Gordon-Salant, 1986; Pickett, 1957]. While most of the experimental work has been done on nonsense syllables, word based misperception studies are also starting to emerge [Felty et al., 2013]. We start by giving a brief overview of the results of both naturalistic and experimental misperception studies organised according to the level of speech unit under investigation and proceed to detail our results following the same sequence.

3.2 Related work

3.2.1 Phone-level

The first studies on nonsense syllable perception were conducted by Fletcher and colleagues in Bell Labs almost a century ago [Fletcher and Galt, 1950]. With the rise of telephony, characterising the telephone's transmission channel both in terms of intelligibility and quality, became an important research problem. While the first studies measuring the intelligibility of speech transmitted over the telephone involved conversational speech, Fletcher soon realised that speech context decreases the efficiency of testing and increases the variability of phone errors, since listeners can take advantage of context to recover the target utterance. Instead, they chose to focus on how the perception of nonsense syllables and individual phones was affected by channel degradations. Through this work, Fletcher [1953] showed that the probability of accurately perceiving a nonsense syllable is roughly equal to the probability of correctly perceiving each individual phone in the syllable across speech levels.

Fletcher and colleges also investigated the spectral distribution of speech cues, by evaluating listeners' performance to nonsense syllables presented through a series of narrowband filters. Through introducing a non-linear transformation called the articulation index, they succeeded in making the contribution of each frequency band to the wideband articulation score additive.[1] An implicit as-

[1]Fletcher defined articulation as the probability of correctly transcribing phonemes and syllables which can be unmeaningful. The term intelligibility was used to describe listeners' performances on recognising words and sentences in adverse conditions.

sumption in this model is that the phonetic features present in each band are detected independently. The articulation index links the channel properties to the articulation score and has been shown to be accurate across a range of channel configurations.

While Fletcher already distinguished between consonant, vowel and syllable error rates, Miller and Nicely [1955] were the first to analyse error patterns in terms of consonant identity. They presented 16 English consonants preceding the vowel /a/ in high- and low-pass filtering conditions, as well as white noise masking. Consonant confusions were analysed along five articulatory feature dimensions, namely voicing, nasality, frication, duration and place of articulation. The amount of information transmitted by each articulatory feature was evaluated by calculating the mutual information between stimulus and response for each feature separately. They found that the selected articulatory features were perceived largely independently with voicing and nasality most accurately transmitted across all conditions, while place of articulation was most susceptible to errors, especially in the low-pass filtering and noise masking conditions.

Dubno and Levitt [1981], on the other hand, argued that features relevant for perception are acoustic rather than articulatory, as the former are grounded in physical properties of the speech signal. They proposed to explain consonant confusions in terms of the acoustic similarity between the presented and reported utterance. Similarity was evaluated along eleven acoustic parameters that have been shown to be relevant for perception in prior studies [Cooper et al., 1952; Delattre et al., 1955; Liberman et al., 1954]. The stimuli consisted of CV and VC nonsense syllables, constructed by combining consonants with vowels /a/,/i/ and /u/. Nonsense syllables were presented across five distinct speech levels, ranging from 20 to 52 dB SPL in both quiet and cafeteria noise, with a signal-to-noise ratio of 5 dB. While a few variables stood out as important, such as consonant duration, energy and consonant-to-noise ratio, no single set of acoustic parameters could be identified that accurately predicted confusion rates across all syllables and experimental conditions.

Along similar lines, Gordon-Salant [1986] also tried to determine whether perceptual confusions are better explained using articulatory or acoustic features without imposing any *a priori* framework. She formed CVs by pairing English

word-initial consonants with vowels /a/,/i/ and /u/ and presented them in twelve talker babble at 12, 6 and 0 dB SNR. Using multidimensional scaling, consonant confusion matrices were converted to perceptual distances in five dimensions. Though the articulatory feature best corresponding to each dimension could be identified, consonants did not form clusters around articulatory feature values in the perceptual domain. While acoustic parameters exhibited moderate to strong correlations with consonant coordinates in the perceptual space, the relative importance of acoustic cues varied greatly across noise and vowel context.

Later, in a series of studies, Allen and colleagues also aimed to determine the acoustic correlates of perceptual events that define consonants [Allen, 2005; Phatak and Allen, 2007; Phatak et al., 2008]. Phatak and Allen [2007] found that while vowels are uniformly masked by speech-shaped noise, consonants cluster into three sets with low, high and intermediate recognition scores. They found that high-scoring consonants — mainly consisting of fricatives and plosives — had considerable energy in the high-frequency regions and were not confused with consonants from the other two groups, while consonants in the low scoring group were often confused with intermediate scoring consonants but not vice-versa. They also found voicing confusions to be highly asymmetric, in favour of voiced consonants.

In an effort to replicate the original findings of Miller and Nicely [1955], Phatak et al. [2008] also analysed confusion patterns in a white noise masker. To determine whether the discrepancies in confusion patterns between their earlier study [Phatak and Allen, 2007] and Miller's original work stem from procedural differences or can be attributed to the different maskers used, Phatak et al. [2008] followed Miller's original experimental procedure as close as possible. Overall, Phatak et al. [2008] were successful in reproducing the findings of Miller and Nicely [1955], and concluded that the differences in confusion patterns compared to their earlier study [Phatak and Allen, 2007] can be attributed to the differences in the spectral distribution of energy between the two maskers. However, the asymmetric voicing confusions in favour of voiced fricatives, observed in both Phatak and Allen [2007] and Phatak et al. [2008] were not present in the study of Miller and Nicely [1955]. As similar results were reported in Grant and Walden [1996], this seems unlikely to be due to the stimuli or procedures used in the

studies by Phatak et al. [2008]. They argued that the reason for not observing such errors in the original study could be familiarity with the talker's voice, as the listeners in Miller and Nicely [1955] also served as the talkers.

One of the main innovations of the study by Miller and Nicely [1955] was approaching perceptual confusions from an information theoretical perspective, by evaluating the amount of information transmitted by each phonetic feature. Christiansen and Greenberg [2012] extended this approach by not only evaluating how consonant features are spectrally distributed, but also how featural cues are integrated across frequency. They presented 11 Danish consonants through narrowband slits, centred around 750 Hz, 1500 Hz and 3000 Hz, either individually or in combination. They measured the increase in information transmitted, as a function of the number of the slits used, as well as their combination. They found that voicing and manner cues are distributed redundantly across the spectrum, and the benefit of adding the third slit was only marginal. In contrast, place had largely independent cues in each slit. Thus, the recognition of place cues improved substantially with the addition of each frequency band.

As the basic unit of misperception in naturalistic studies is the word, less emphasis has been given to phone-level analyses of slips of the ear. Also, most naturalistic collections are too limited in size to provide enough statistical power to support such an analysis. That said, Bond [1999b] provided several observations at the phone level. For example, they reported that manner of articulation tends to be conserved between the intended and perceived consonant. In addition, they showed several examples of confusions involving the consonant /t/. However, these observations were not systematic and were not backed up by statistics. Nevertheless, several important trends started to emerge, later confirmed by larger datasets, such as the perceptual robustness of vowels and the stressed syllable.

Tang [2015] conducted a systematic investigation of confusions at the phone level, exploiting his large-scale collection of over 5000 misperceptions. Confirming results of prior studies [Bond, 1999b; Meringer, 1908], he found that segments with higher acoustic energy such as stressed syllables, vowels and voiced consonants, have lower error rates. Similar to Gordon-Salant [1986], Tang [2015] also evaluated how perceptual distances relate to phonetic distances. By convert-

ing phoneme confusion matrices to perceptual distances via Shepard's metric [Shepard, 1972] and applying multidimensional scaling and hierarchical clustering techniques, he compared clusters of speech sounds in phonetic and perceptual space. Tang [2015] found that while consonants form perceptual clusters around phonetic feature values such as frication, voicing, and nasality, there was only a modest correlation between perceptual and phonetic distances overall.

3.2.2 Syllable-level

While most early confusion studies employed a forced-choice task [Dubno and Levitt, 1981; Gordon-Salant, 1986; Miller and Nicely, 1955], later investigations switched to an open-set paradigm and used CVC syllables which allowed the examination of consonant position effects [Benkí, 2003; Woods et al., 2010]. Benkí [2003] presented a phonetically balanced set of CVCs to native listeners at four different signal-to-noise ratios. He obtained similar results as Miller and Nicely [1955] as well as Pickett [1957], in terms of the proportion of information transmitted, and confirmed that voicing and manner features were more accurately transmitted than place. Benkí [2003] also found that all three articulatory features were more accurately perceived in the onset position relative to coda, across all SNR conditions. Further, consonant deletions were more frequent in coda position, especially for sonorants and voiced stops. He concluded that consonants in the onset position had a perceptual advantage over coda.

Woods et al. [2010] also conducted a study of CVC confusions in speech-shaped noise. Reference level SNRs corresponding to roughly 65% recognition accuracy were determined for each consonant, in both onset and coda position prior to presentation. Psychometric functions were evaluated at reference level SNR, as well as 6 dB above and below reference, for each consonant and position. Reference level SNR values varied by more than 40 dB across consonant identity. A lower average reference SNR for onset position relative to coda indicated an initial consonant advantage, consistent with Benkí [2003]. In addition, Woods et al. [2010] reported that vowel identity significantly influenced consonant identification and that consonant confusions had a tendency to cluster around the same manner and voicing values with the exception of nasal-liquid and sibilant-affricate

clusters.

Browman [1980] examined how word and syllable position affected segmental error rates using her collection of 200 slips of the ear. In agreement with the experimental studies above, she found lower error rates in onset position with respect to coda in monosyllabic words. However, this trend is reversed in polysyllabic words where error rates progressively diminished from the initial to the final position, both within-syllable and within-word. In line with the findings of Garnes and Bond [1980], she found that stressed syllables were less error prone compared to unstressed ones. While the observations of Browman [1980] lacked statistical backing, Tang [2015] replicated the analysis on his extensive corpus and confirmed that the trends reported by Browman [1980] were significant.

3.2.3 Word-level

As most laboratory studies have tried to identify perceptually relevant speech cues by analysing nonsense syllable confusions, less emphasis has been placed on word-level misperceptions, where lexical factors also come into play. In order to quantify lexical advantage, Benkí [2003] compared the recognition of CVC words with CVC nonsense syllables. 120 stimuli of both words and nonsense syllables were presented to native listeners with normal hearing across four different SNRs. They found that the phonemes of nonsense syllables were perceived independently, while valid lexical items eased the recognition of the word-final consonant. This effect, however, was modulated by neighbourhood density, as listeners experienced a greater benefit when the neighbourhood of the stimulus word was sparse. Luce and Pisoni [1998] conducted a series of experiments, involving perceptual identification and lexical decision tasks to test the effects of word frequency and neighbourhood confusability on word recognition. They found that high-frequency words with sparse, low-frequency neighbourhoods supported better word identification. They reported that high density and frequency neighbourhoods also resulted in slower processing times for clean speech. Based on these findings they proposed the neighbourhood probability rule [1] and subse-

[1]The neighborhood probability rule is defined as $p(ID) = \frac{F(W)p(s)}{F(W)p(W)+\sum_{j=1}^{N} F(N_j)p(N_j)}$ where $p(ID)$ is the probability of identifying word W, $p(W)$ is the perceptual probability of W based

quently the neighbourhood activation model. Felty et al. [2013] presented a lab-based study of word misperceptions with target words sampled randomly from the Hoosier Mental Lexicon [Nusbaum et al., 1984], instead of constraining the stimuli to monosyllabic words. They compared targets and misperceptions across word length (measured in both the number of phones and syllables) and word frequency and evaluated the phonetic distance between targets and confusions. On average, they found misperceptions to be shorter and higher in frequency compared to target words. In addition, they reported that the word frequencies of targets and misperceptions are correlated, a finding also observed in naturalistic collections [Tang, 2015].

Conversely, as the basic units of slips of the ear are words, almost all naturalistic studies investigated misperceptions across lexical factors. Using a rather small (88 sample) subset of the corpus collected by Garnes and Bond [1980], Vitevich [2002] compared misperceptions and target words across several variables, including word length, familiarity, frequency, neighbourhood density and neighbourhood frequency. He found no significant differences between targets and misperceptions for any of the above variables. However, when comparing slip of the ear tokens (i.e. both target and misperceived words) to words randomly selected from the lexicon, he found that slip of the ear tokens had denser, higher frequency neighbourhoods. Bond [1999a] investigated the effects of morphology on misperceptions. She found that listeners perceived morphologically complex forms as simple, rather than the other way around. Further, they found that errors mostly involved inflectional instead of derivational affixes.

From the overview above, it is clear that significant factors influencing misperceptions can be identified across various levels of speech units. In the upcoming sections, we follow the same bottom-up progression for the analysis of our Spanish misperception corpus, starting with factors at the phone level. We start our analysis by detailing the alignment algorithm.

on the stimulus, $F(W)$ denotes the word frequency of W, $p(N_j)$ is the perceptual probability of the jth neighbour of word W and $F(N_j)$ denotes the word frequency of the jth neighbour.

3.3 Stress-based syllable and phoneme alignment

In order to be able to analyse confusion patterns below the word level, the phonetic transcriptions of the intended and misperceived utterance need to be aligned. Early naturalistic studies relied on manual alignment of target-confusion pairs. However, this approach lacks objectivity and is impractical for large-scale collections, which have become increasingly available. The alternative is to perform the alignment of segments automatically. A number of algorithms have been proposed in the past for the alignment of phone sequences. Covington [1996] used a depth-first search to align broad phonetic transcriptions of cognate pairs. Their algorithm is based on binary articulatory features and emphasises aligning segments of comparable syllabicity, distinguishing between three broad categories: consonants, vowels and glides. Somers [1999] proposed an algorithm similar to that of Covington [1996] to compare children's articulations to those of an adult model. However, he employed narrow instead of broad phonetic transcriptions and used the stressed vowel as an anchor point for the alignment. Kondrak [2003] used dynamic programming string alignment to align pairs of historical cognates based on multivalued articulatory features.

All three of the alignment algorithms above are feature-based. However, this approach can result in circularity, as noted by Tang and Nevins [2012]. If the algorithm imposes *a priori* constraints on the alignment, it can potentially bias the results. On the other hand, using a phonetically-blind procedure with no restrictions can lead to highly unlikely alignments. For example, blind alignment of the word 'intentar' to 'patentar' might map /i/ to /p/ and /n/ to /a/. Thus, when selecting the alignment algorithm, the trade-off between generating plausible alignments and introducing bias needs to be carefully considered.

Tang and Nevins [2012] used an algorithm by Needleman and Wunsch [1970] originally intended for DNA sequence alignment. While they placed no constraints on the alignment of consonants, each vowel was replaced by an anchor symbol to bias the algorithm to align by syllables and avoid mapping vowels to consonants.

We propose an alignment method that is sensitive to both stress and syllable structure. Alignment is achieved in two steps. First, the syllables of the tar-

get and misperceived word are aligned using lexical stress as an anchor point. Then, the phonetic[1] transcriptions of matched syllables are aligned according to syllable constituency, linking intended and perceived syllable segments in onset, nucleus and coda positions. Figure 3.1 and Figure 3.2 provide two examples of the alignment procedure. The first example corresponds to an onset and nucleus substitution in the stressed syllable, while the unstressed syllable remains unchanged. In the second, more complex example, only the stressed nucleus survives while the rest of the phones undergo a substitution and an entire syllable is inserted.

There are situations, however, where stress-based alignment is suboptimal, for example in cases where listeners report a salient word from a speech based masker in its entirety, and the target word has no impact whatsoever on the listener's percept. Cases where target and misperception share the same root but the stress in the misperception has shifted due to morphological variation also result in stress-based alignments that are not ideal. In order to make sure that our analysis is based on correct alignments we use the following heuristic:

1. if the stressed vowel is conserved between the target and the misperceived word we assume stress-based alignment is applicable. (As the stressed vowel is the most robust segment of the word, the odds of this occurring by chance are quite small).

2. cases where the stressed vowel is not conserved are screened manually and classified into:

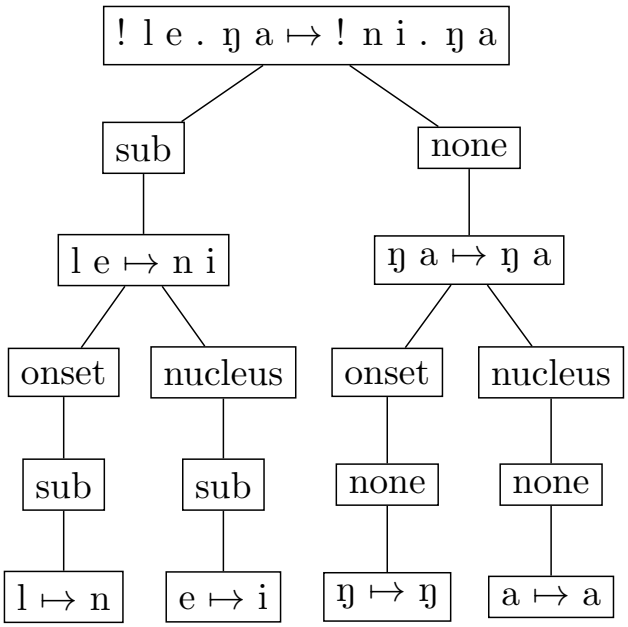
   (a) stressed-based alignment applicable even though stressed vowel undergoes a substitution. Figure 3.1 provides an example of such a case. Even though the vowel of the stressed syllable undergoes a substitution, stress based alignment is clearly applicable given that the second syllable remains unchanged.

   (b) stress-based alignment is not applicable. One example of such a case from the corpus would be the confusion 'muchacho'[boy]-'mucho'[many].

[1] In reality transcriptions are semi-phonemic. In the following we will use the term phonetic for simplicity.

Target			Confusion			
! l e . ɲa			! n i . ɲa			
onset	nucleus	coda	onset	nucleus	coda	stress
l	e	-	n	i	-	1
ɲ	a	-	ɲ	a	-	0

(a)

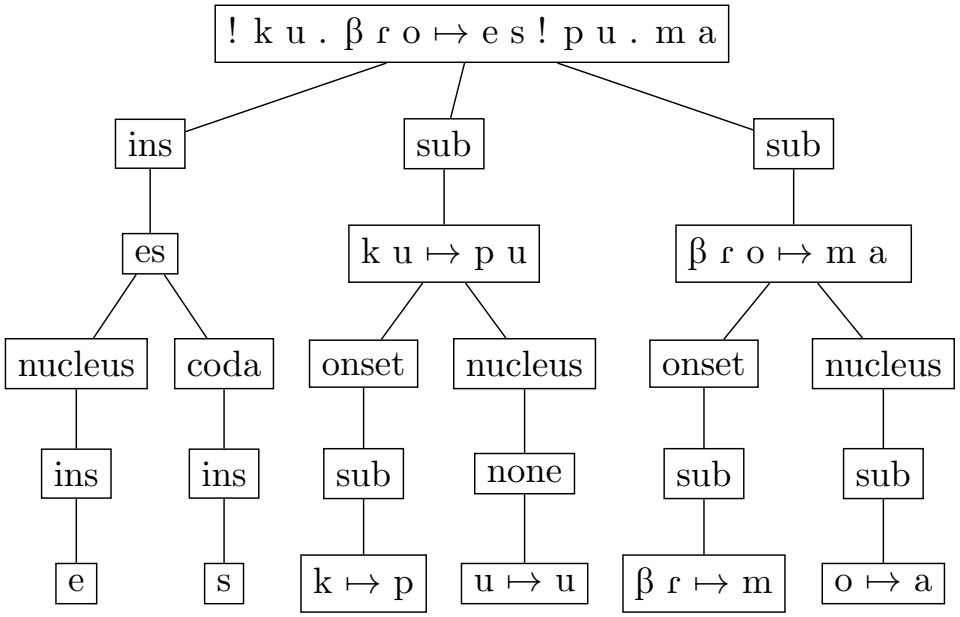


(b)

Figure 3.1: *Example of the stress-based alignment procedure for target misperception pairs 'leña'[firewood]–'niña'[girl]. Panel* **a** *and* **b** *provide the alignment in matrix and tree form respectively. Rows of the matrix corresponds to the aligned syllables while columns show syllable constituency for target word and misperception.*

Target			Confusion			
! k u . β ɾ o			e s ! p u . m a			
onset	nucleus	coda	onset	nucleus	coda	stress
-	-	-	-	e	s	0
k	u	-	p	u	-	1
β ɾ	o	-	m	a	-	0

(a)



(b)

Figure 3.2: *Example of the stress-based alignment procedure for a more complex case: 'cubro'[I cover]–'espuma'[foam]. Only the stressed vowel is conserved, while we observe a syllable insertion and the mapping of a consonant cluster to a single consonant.*

For this example the stress based alignment would align syllables [ʧ a] and [m u] which is incorrect: [m u ! ʧ a . ʧ o]↦ [! m u . ʧ o], where '.' denotes syllable boundaries and '!' marks the beginning of the stressed syllable.

1) and 2.a) are then combined and form the set of misperceptions of interest for this study.

3.4 Statistical methodology

The goal of the current analysis is to understand how different factors and their interactions affect the error patterns observed in misperceptions across multiple segmental levels. Since aligned segments can fall into one of four categories (i.e. substitution, deletion, insertion or unchanged) our outcome variable is of a categorical nature. While ANOVA has been used to judge the significance of categorical outcomes converted to proportions in the past, it has been shown that this can lead to spurious results [Jaeger, 2008]. More recently, the use of Generalised Linear Models with mixed effects was advocated in psycholinguistic research for analysing categorical data [Baayen, 2008]. As we assess the probability of each type of error separately, the outcome variable can be expressed as a linear combination of the predictors after applying the logit link ($logit(p) = ln(\frac{p}{1-p})$) function corresponding to binary outcomes. As the quantity $\frac{p}{1-p}$ is otherwise referred to as odds, each model coefficient β expresses the change in log-odds of the given type of error when the corresponding predictor is present. Thus positive coefficients are associated with an increase in log-odds and consequently likelihood, while negative coefficients correspond to a decrease in likelihood.

3.5 Outcome of stress-based alignment

Stress based alignment was applicable in 3082 cases corresponding to nearly 96% of the corpus. The 133 cases where the alignment was not applicable are listed in Appendix B. Misperceptions in the latter category are further classified based on the reason they defy stress based alignment. 72 cases correspond to overrides,

where a salient word was reported from the babble, while in 17 cases there is a shift in the stressed syllable due to morphology. In 44 cases the stress based alignment was suboptimal for another reason. The following analysis is focused on the 96% of the confusions in the corpus which lend themselves to stress based alignment.

Applying the proposed stress-based procedure to the 3082 target-misperception pairs resulted in 7215 syllabic and 17 602 phonetic aligned segments. One-to-many and many-to-one mappings of consonant clusters with no apparent consonant match can result in alignment ambiguity (see example of 'cubro'–'espuma' in Figure 4.1 panels **a** and **b**). As these cases constitute a tiny fraction of the total number of aligned phonetic segments (1.04%), they are omitted from the analysis. The comparison of aligned segments, both at the phone and syllable level, can produce one of four possible segmental outcomes. The target segment can remain unchanged (none), undergo a substitution (sub) or deletion (del), or a segment not originally present in the target can be inserted (ins). In the following analysis, we not only evaluate segmental error rates, defined as the proportion of errors over all occurrences $[(n_{sub} + n_{del} + n_{ins})/(n_{sub} + n_{del} + n_{ins} + n_{none})]$ across variables of interest, but also show the distribution of aligned segments in terms of the four possible outcomes above. The next section details the trends in perceptual error patterns at the phone level.

3.6 Effects of consonant identity

3.6.1 Results

Past studies [Cutler et al., 2004; Miller and Nicely, 1955; Phatak and Allen, 2007; Woods et al., 2010] investigating noise-induced misperceptions in nonsense syllables have demonstrated that consonants vary in terms of intelligibility and confusion patterns. This section analyses consonant errors patterns in the Spanish misperceptions corpus.

Figure 3.3 shows the distribution of consonant outcomes across the levels of articulatory features voicing, manner and place. Error rates showed significant differences across the levels of all three features. Regarding manner of

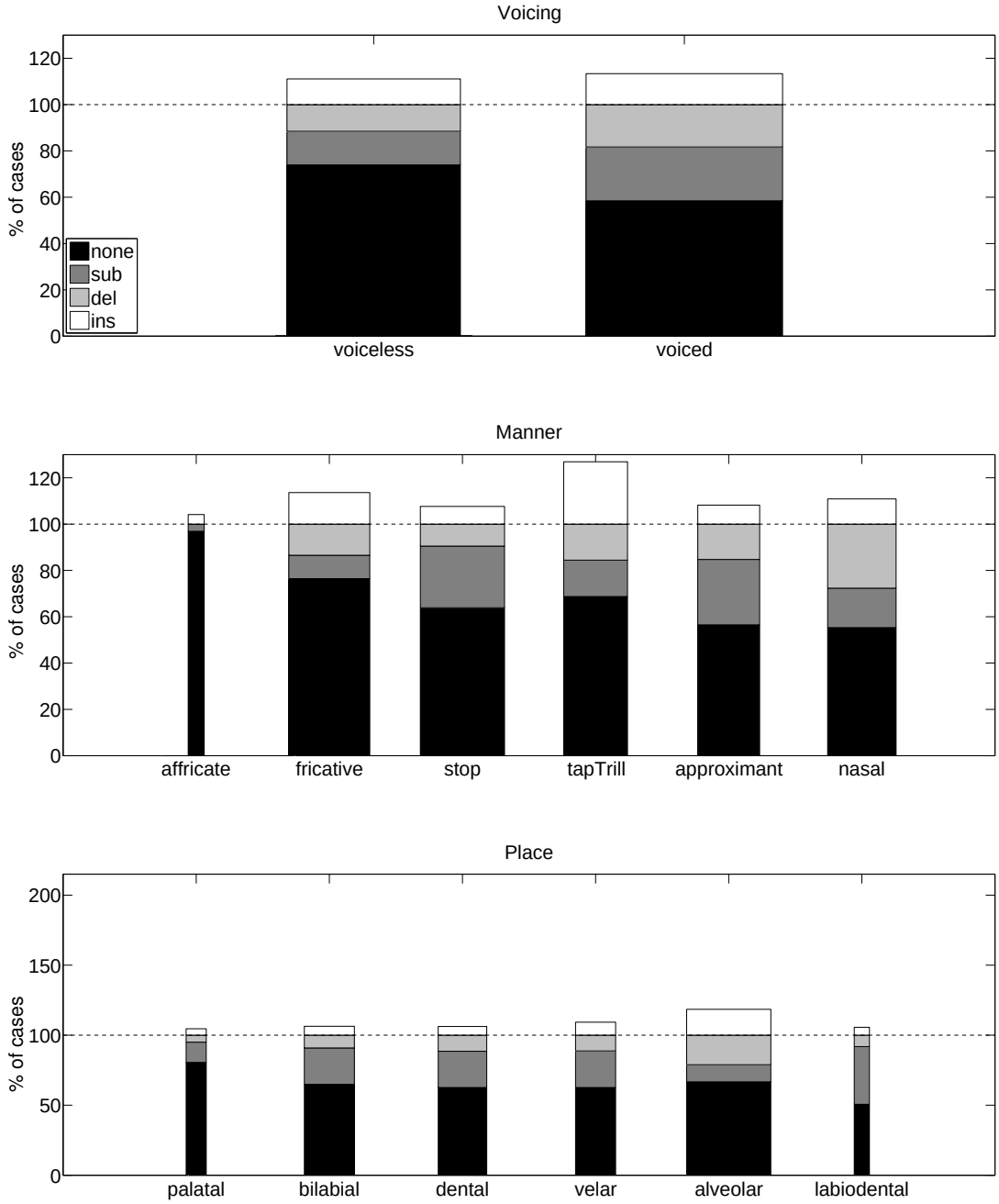


Figure 3.3: *Distribution of outcomes for each articulatory feature. Proportions are normalised on the number of times a given feature was present in the target word ($n_{none} + n_{sub} + n_{del}$). Bar widths are proportional to the square root of counts in each bin. For each feature, levels are ordered from most robust to most error prone from left to right.*

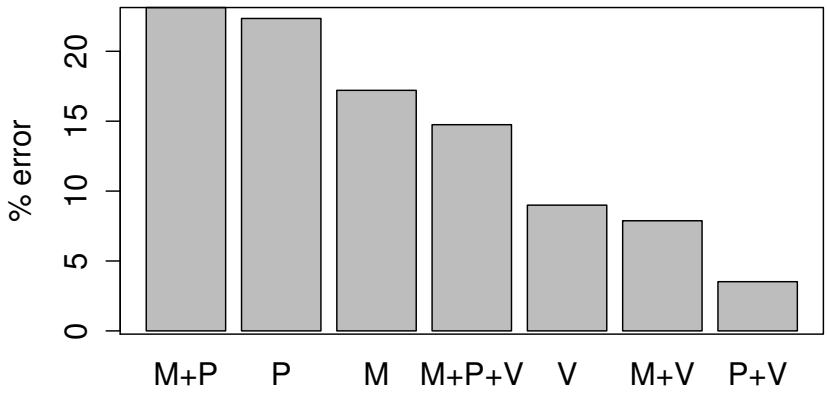


Figure 3.4: *Percentages of single and combined articulatory feature errors in consonant substitutions. M, P and V stand for single feature errors manner, place and voicing respectively, while errors with a '+' indicate combined feature errors.*

articulation, affricates had the smallest error rate (7%), followed by fricatives (33%), stops (41%), taps/trills (46%), approximants (48%) and nasals (50%). Differences in error rate were significant between affricates, fricatives, stops and taps/trills [$\chi^2_{min}(1) = 11.08, p < .001$], while differences between taps/trills, approximants and nasals were not significant [$p_{min} = 0.19$]. Regarding place of articulation, palatal consonants had the least amount of errors (23%), followed by bilabial (39%), dental (41%), velar (43%), alveolar (44%) and labiodental (52%) consonants. Bilabial, dental, velar and alveolar consonants produced a similar error rate [$p_{min} = 0.22$], while the differences in error rates were significant between palatal and bilabial [$\chi^2(1) = 27.38, p < .001$], as well as alveolar and labiodental [$\chi^2(1) = 4.33, p = 0.03$] consonants. We found voiced consonants to be significantly more error prone (48%) than voiceless ones (33%) [$\chi^2(1) = 236.09, p < .001$].

Figure 3.4 shows the types of articulatory feature errors listeners make in consonant substitutions. Combined feature error manner and place (P+M) (23%) followed by single feature errors of place (P) (22%) and manner (M) (17%) occurred most frequently. Consonant substitutions involving a voicing (V) error [M+P+V

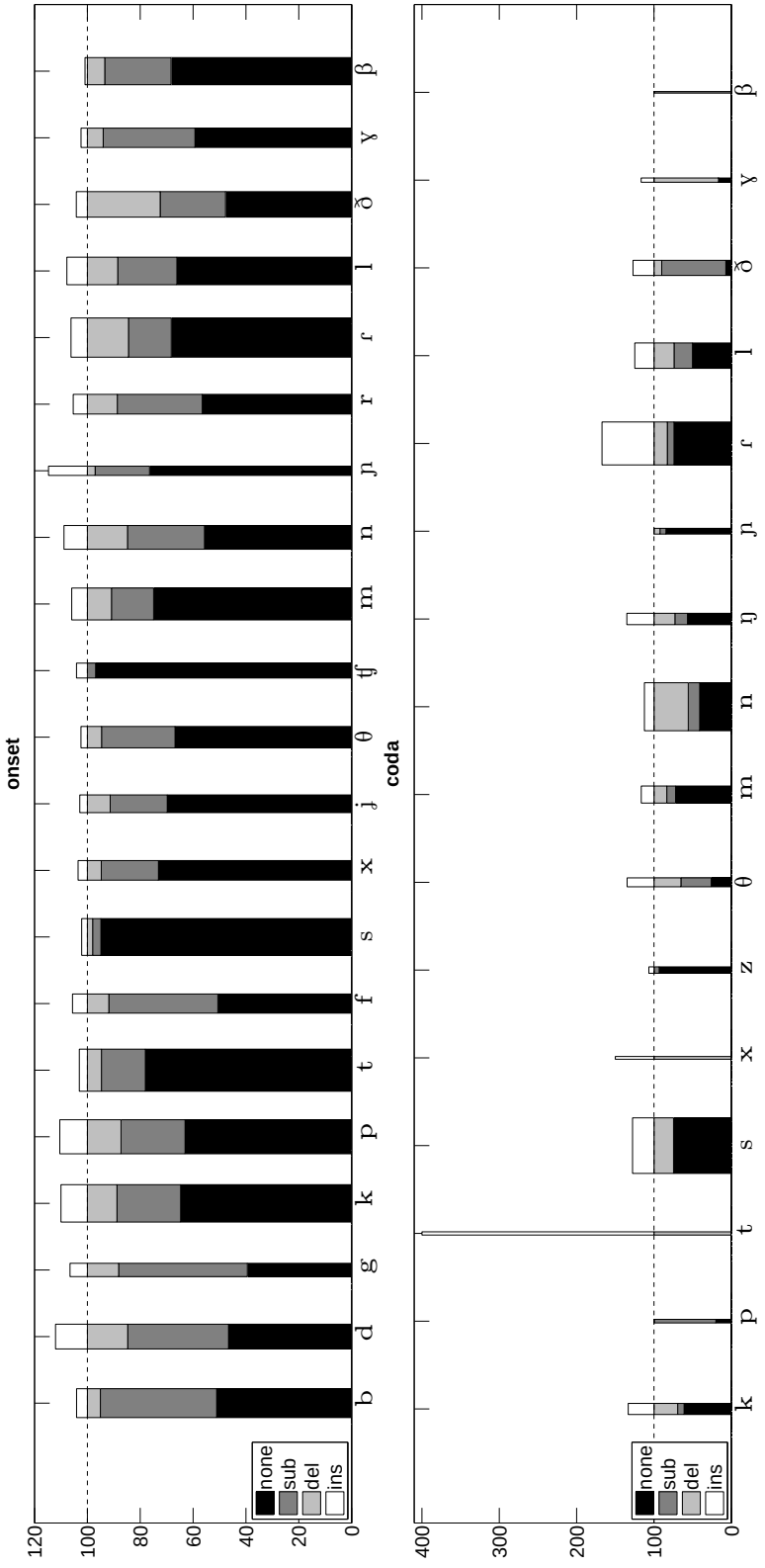


Figure 3.5: *Distribution of outcomes for consonants in onset and coda position respectively. Due to Spanish phonotactics fewer consonants are allowed in coda position compared to onset.*

(15%), V (9%), M+V (8%), P+V (4%)] were the least common. Significant differences were found between the error rates of single feature place and manner errors [$\chi^2(1) = 11.95, p < .001$], M+P+V and V categories [$\chi^2(1) = 24.96, p < .001$], as well as M+V and P+V [$\chi^2(1) = 29.82, p < .001$] while the differences between the other adjacent categories were not significant [$p_{min} = .23$].

Figure 3.5 shows the distribution of outcomes across consonant identity and syllable position. The most robust consonants include the voiceless palatal affricate /ʧ/ with (3%) error rate, followed by sibilant fricatives /z/ (7%) and /s/ (18%), the voiceless dental plosive /t/ (22%) and the voiced palatal nasal /ɲ/ (21%). On the other end of the spectrum, voiced plosives /b/ (49%), /d/ (53%) and /g/ (61%) were amongst the most error prone. Comparing consonants of the same identity in onset and coda position we find quite large differences, for example /s/ (5%-26%),/ð/ (52% -93%) or /ɲ/ with (24% -15%) which suggests an interaction between consonant identity and syllable position.

The substitutions in Figure 3.5 for each consonant are further broken down using confusion matrices for both onset and coda position in Figure 3.6 and Figure 3.7 respectively. In onset position, consonants were substituted quite freely: on average 9.4 response alternatives scored over 2% and 11.4 over 1%. Nevertheless, several confusion clusters can be identified. Voiceless plosives /p/,/t/,/k/ are often substituted amongst each other. Plosives also show asymmetric voicing errors in the direction of voiced-voiceless. Asymmetric confusions can be observed for fricatives as well: /f/,/s/ and /x/ are most often confused with /k/ and /p/ and /θ/ with /t/. Spanish phonotactics allows for fewer consonants in coda position relative to onset. In coda, consonants are most often misperceived as /s/,/n/ or /ɾ/,which at least in part can be attributed to inflectional morphology as these phones serve as suffixes.

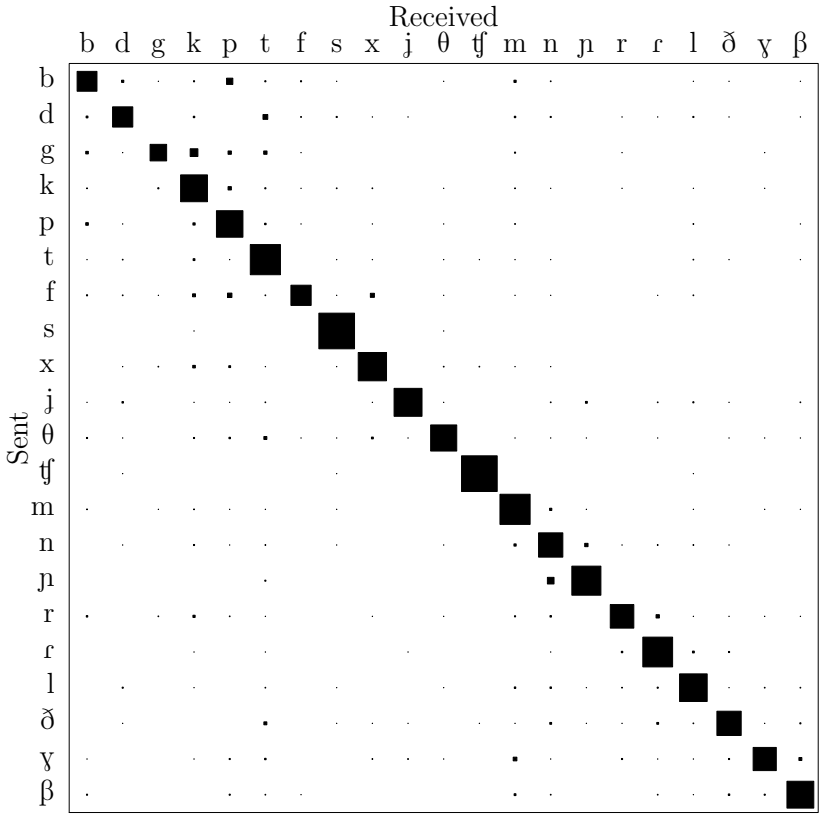


Figure 3.6: *Confusion matrix of phone substitutions in onset position. Here as in Figure 3.7, square area corresponds to proportion of cases as each row is normalised.*

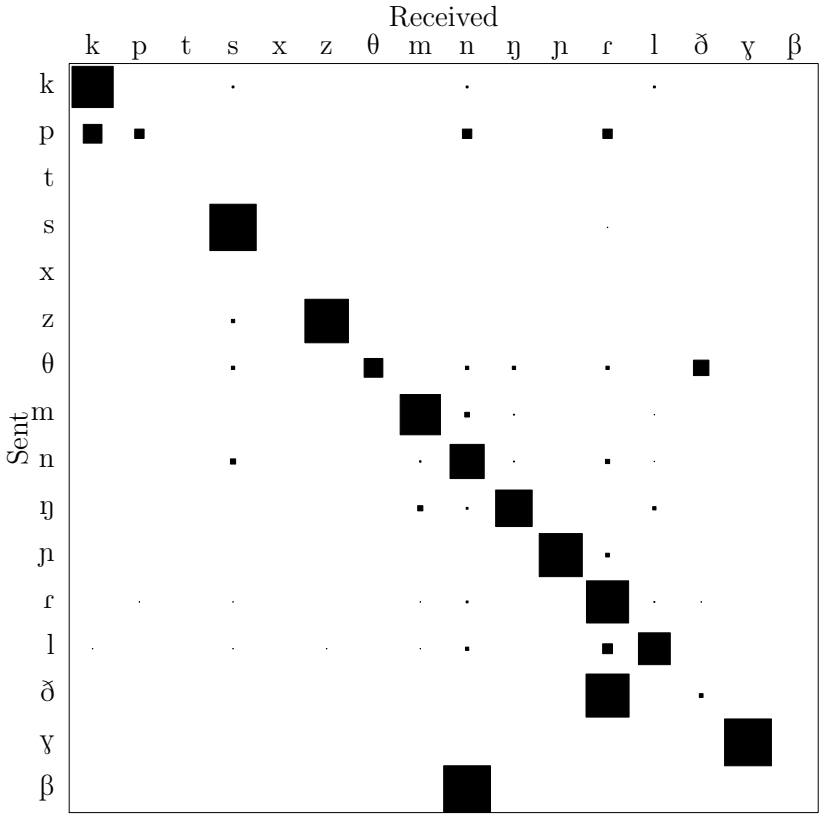


Figure 3.7: *Confusion matrix of phone substitutions in coda position.*

3.6.2 Interim discussion

Our findings at the phone level showed many similarities with confusion patterns reported in nonsense-syllable perception studies [Dubno and Levitt, 1981; Miller and Nicely, 1955; Phatak and Allen, 2007; Woods et al., 2010]. In their investigation of CVC confusions in speech-shaped noise, Woods et al. [2010] reported that in onset position, sibilants and affricates were most recognisable even in adverse SNRs, followed by liquids, plosives, nasals and non-sibilant fricatives. Phatak and Allen [2007] found that consonants could be categorised into high- /s,ʃ,z,ʒ,t/, low- /f,θ,v,ð,b,m/ and intermediate-scoring /n,p,g,k,d/ groups in their study of CV recognition in speech-shaped noise. In agreement with both these studies, our data showed that the voiceless palatal affricate /ʧ/, sibilant fricatives /s,z/, as well as the voiceless plosive /t/ had the lowest error rate. There was agreement on the other end of the spectrum as well. Non-sibilant fricatives /f/ and /θ/, the voiced plosive /b/ and approximant /ð/ displayed the highest error rates in our data. These consonants had above average baseline SNRs in Woods et al. [2010] and were also categorised as low-scoring in Phatak and Allen [2007]. An exception was /m/, which was classified as low-scoring for both the above studies, yet was among the more robust consonants (26% error rate) in our case. The above studies also reported that voiceless consonants were less error prone than voiced ones, a finding which was supported by our data as well.

Feature errors in consonant substitutions found in our corpus also matched those in prior experimental studies. In line with our findings, Woods et al. [2010] reported that single feature place errors were most frequent, followed by combined feature manner and place errors, single feature manner errors and finally errors involving voicing. While Dubno and Levitt [1981] presented voiced and unvoiced consonants separately, they found the same ordering of manner and place errors as in our corpus. Our findings in terms of consonant substitutions also showed agreements with the above studies. Woods et al. [2010] reported that intra-manner confusions are common for unvoiced plosives and nasals. Phatak and Allen [2007] reported that low-scoring consonants get confused with intermediate-scoring consonants but not vice-versa, which matched with some of the asymmetric confusions we observed, for voiceless plosives and non-sibilant fricatives /b/,/f/,/θ/.

As seen above, many segmental confusion patterns reported in experimental studies in response to nonsense syllable stimuli were similar to those stemming from word-level misperceptions. These results support the hypothesis that misperceptions are — at least to some extent — acoustically driven, as word level factors did not completely override the confusion patterns observed in nonsense syllable confusions.

At the same time, fewer similarities could be found with trends reported in naturalistic misperception studies at the phone level. For example, contrary to our findings, Tang [2015] reported that voiced consonants were less error prone than voiceless ones. We found differences in the error rates of manner categories as well. Tang [2015] reported that liquids and glides had the smallest error rate followed by nasals, fricatives and stops and finally affricates, with an error rate almost twice as high as the adjacent category. Tang [2015] argued that these findings could be explained in terms of sonority. As the validity of the concept of sonority has been called into question [Ohala and Kawasaki-Fukumori, 1997], Parker [2002] proposed a sonority scale grounded in acoustic correlates, most notably intensity. He proposed the following sonority hierarchy of consonants: Glide > Liquid > Nasal > Fricative > Affricate > Stop (where > indicates 'more sonorous than'). After discounting affricates as outliers, Tang [2015] found that the error rates of consonant manner groups roughly matched the sonority scale. Sonority could also account for the robustness of voiced consonants compared to voiceless ones. While higher acoustic energy resulting in more accurately perceived segments is a sensible explanation, our results, as well as the above-mentioned noise-induced nonsense-syllable studies did not support this claim. We will return to examine this discrepancy in Section 3.9 in more detail.

Despite the above differences, our results showed agreements with the findings of Tang [2015] in terms of place of articulation errors. Tang [2015] reported that error rates increased significantly from coronal through dorsal to labial segments. They argued that these findings were well explained by the featurally underspecified lexicon model proposed by Lahiri and Reetz [2002]. This model postulates that listeners handle the inherent variability (e.g. between dialects, speakers) of the speech signal by storing featurally underspecified representations in the mental lexicon. The phonological features extracted from the speech signal are

then mapped to the features stored in the lexicon with three possible outcomes: match, mismatch or no-mismatch. Lahiri and Reetz [2002] argued that labial and dorsal features are stored in the lexicon while coronals are underspecified, thus the probability of perceiving a labial or dorsal segment as coronal is higher than the other way around as it produces a no-mismatch. By collapsing the place of articulation into categories labial [bilabial, labiodental], coronal [dental, alveolar] and dorsal [palatal, velar], we saw that both dorsal and labial segments are perceived as coronal segments 36% of the time. However, coronals are perceived almost half as frequently as dorsal (19%) or labial (16%) segments. Thus, our data also seems to support the underspecification hypothesis.

3.7 Word-structure effects

3.7.1 Results

In this section, we investigated the effects of word-structure related and suprasegmental factors on misperceptions, focusing on segmental position and lexical stress. Syllable position within-word [initial, medial, final] and phone position within syllable (i.e. syllable constituency) [onset, nucleus, coda] were combined into a single factor with the following levels for simplicity: initial onset, medial onset, medial nucleus, medial coda and final coda. Figure 3.8 plots the outcomes across position for unstressed and stressed syllables in the upper and lower panel respectively. In order to determine the effect of position, stress and their interaction on the distribution of errors, we fitted three distinct logistic regression models, predicting the probability of observing each error type (i.e. substitutions, deletions and insertions) separately. The significance values of these two predictor variables and their interaction were assessed using likelihood ratio tests. We found that the best model included predictors Position and Stress as well as their interaction, for all three types of errors [$\chi^2_{min} = 369, p < .001$]. As the algorithm anchors the alignment on the stressed syllable, deletions and insertions are not possible in the stressed nucleus position (see lower panel of Figure 3.8).

Starting with the effects of stress, we find that insertions [$\beta = -1.08, p < .001$] and deletions [$\beta = -1.31, p < .001$] are less prevalent in stressed syllables com-

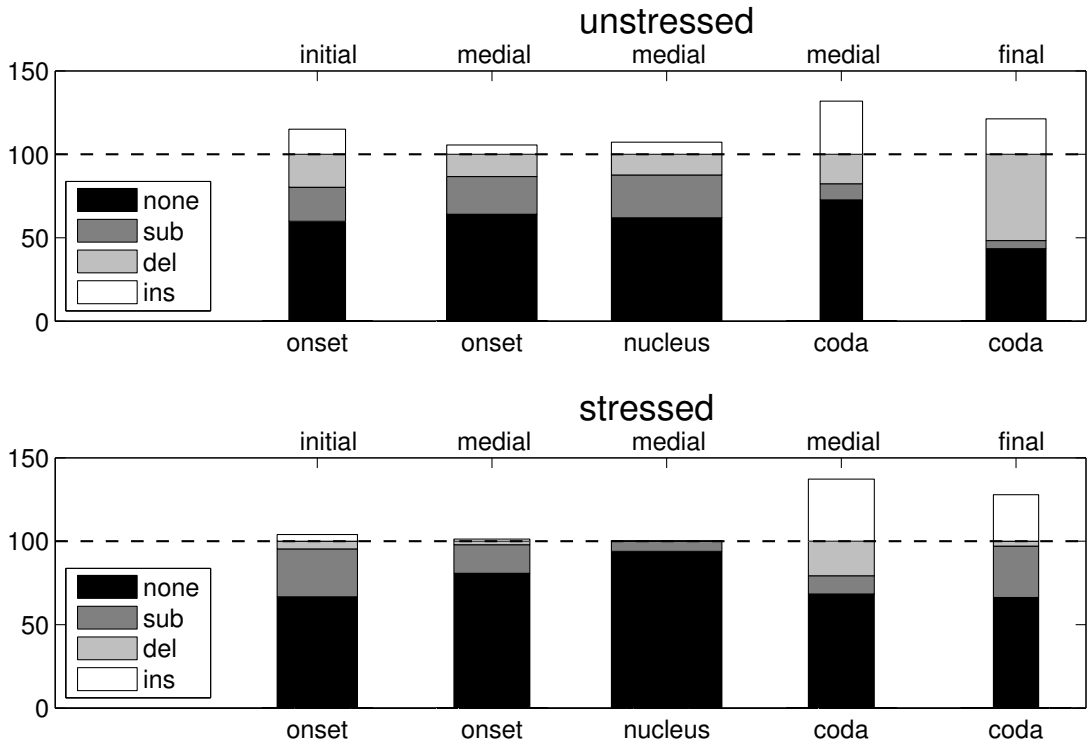


Figure 3.8: *Segmental outcomes across factors Position and Stress. The top and bottom panels show the distribution of outcomes in unstressed and stressed syllables respectively. Outcome proportions are normalised on the total number of sent phones i.e.* $n_{none} + n_{sub} + n_{del}$*. Bar widths are proportional to the square root of counts in each bin. The top and bottom labels across the x-axis corresponding to within word and within syllable position and jointly describe the levels of the factor position (e.g. initial onset)*

pared to unstressed ones. The effect of stress on substitutions is less straightforward and depends on word position.

Regarding word position, we find that deletions are most common in word-final coda [$\beta = 2.05, p < .001$], and are less pronounced in initial and medial positions. Insertions are also most frequent in coda position, both word-medial [$\beta = .92, p < .001$] and final [$\beta = .92, p < .001$]. Substitutions are most prevalent in nucleus [$\beta = .34, p < .001$] and onset position, with no significant difference between initial and medial onset [$\beta = .19, p = .06$].

Regarding the interaction between stress and word position, stress diminishes the proportion of deletions across all positions, except for word-medial coda [$\beta = 1.43, p < .001$]. The same is true for insertions except for both medial [$\beta = 1.50, p < .001$] and final [$\beta = .94, p < .001$] coda. Significantly fewer substitutions occur in the nucleus position for stressed syllables [$\beta = -2.16, p < .001$] compared to unstressed ones, while the proportion of substitutions increases in word-initial [$\beta = .48, p < .001$] and final position [$\beta = 1.53, p < .001$].

Our alignment algorithm also allows for treating syllables instead of phones as the primary segmental unit. This way we can investigate the distribution of outcomes on a syllable level. In this context, insertion and deletion refer to the insertion and deletion of an entire syllable, while substitution refers to any change occurring within the syllable. Figure 3.9 shows the distribution of syllable outcomes as a function of position relative to the stressed syllable. Due to the stress-based alignment, insertions and deletion are not possible in the stressed syllable position (see Figure 3.9 right panel). We found that pre-stressed position had a significantly smaller error rate (55%) compared to post-stressed position (79%) [$\chi^2(1) = 276.75, p < .001$].

3.7.2 Interim discussion

The robustness of the stressed syllable is one of the most prevalent findings of studies investigating slips of the ear [Browman, 1980; Garnes and Bond, 1980; Tang, 2015]. Pisoni [1981] argued that in order to understand speech, listeners take advantage of salient and reliable portions of the acoustic signal — such as the stressed syllable — which can activate other sources of knowledge (e.g.

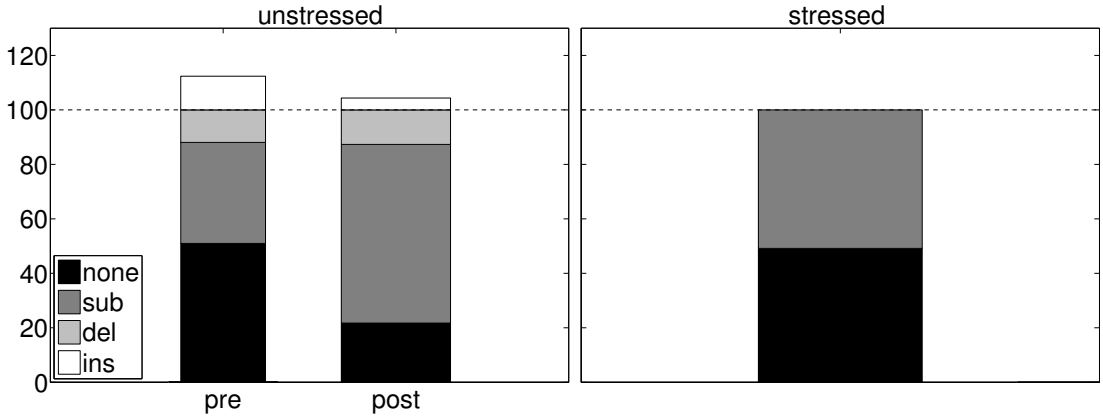


Figure 3.9: *Entire syllable changes across position relative to the stressed syllable. The distribution of outcomes in pre- and post stressed syllables are shown in the left panel, while the stressed syllable's outcome distribution is shown on the right. As the alignment is anchored on the stressed syllable, substitution is the only possible error type in the stressed position.*

lexical) that further aid the perception process. Our findings also supported the robustness of the stressed syllable. While the probability of errors was lower in the stressed syllable in general, we found that deletions are the least likely to occur. Tang and Nevins [2012] report a similar result in their analysis of phonetic adjacency: two adjacent vowels diminish the likelihood of a consonant deletion. These findings suggest that in salient acoustic environments such as adjacent vowels or the stressed syllable, listeners are unlikely to miss the presence of a phone entirely.

We observed high error rates in both word-initial and final positions. This is in contrast to previous naturalistic studies, such as Browman [1980] and Tang [2015] who reported lowest error rates for word-final syllables followed by word-medial and initial, and found that within-syllable, onset position was the most error-prone. Bond [1999b] also observed that word-initial consonant substitutions occurred twice as often as in any other position. One possibility is that this disparity is attributable to cross-linguistic differences in misperception corpora. Spanish is a highly inflected language, unlike English in which the above naturalistic corpora have been collected. This difference could be responsible for the higher error rates in word-medial and final positions for Spanish misperceptions. To investigate whether these differences were in fact due to a morphological effect,

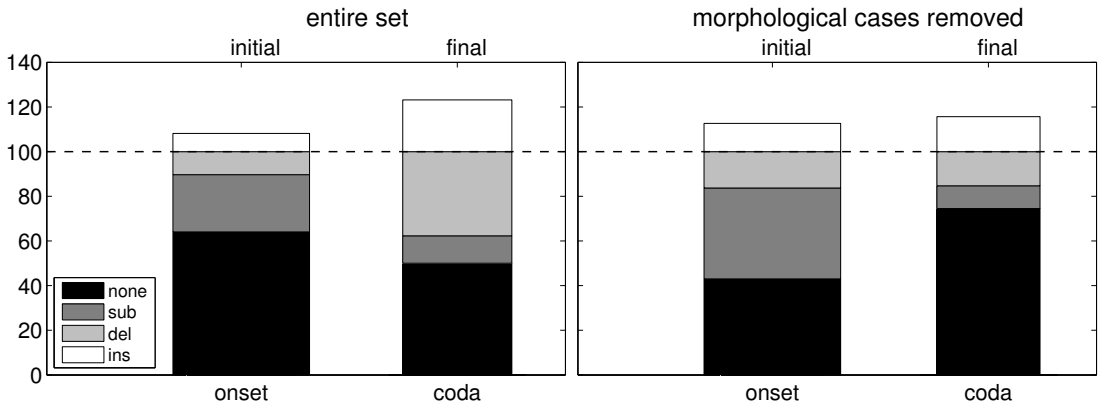


Figure 3.10: *The left panel shows the distribution of outcomes in word-initial onset and word-final coda for the entire set of confusions. Th right panel shows the distribution of outcomes in the same positions with the morphological cases filtered out.*

we have removed confusions where the target and the reported word are morphological variants. We flagged target-confusion pairs as morphological variants if they shared the same lexeme. Lexeme information was obtained from the ESPAL database [Duchon et al., 2013] and ambiguous cases were categorised manually.

Figure 3.10 plots the distribution of outcomes for initial onset and final coda position for the entire set of misperceptions (left panel), and with the morphological cases removed (right panel). Without the morphological cases, error rates showed a similar tendency to the findings of naturalistic studies reported above.

We observed a relatively high error rate in word-medial coda, especially in terms of insertions and deletions. This could be partly due to morphological variation, often resulting in changes in the suffix, e.g. "probado" [proven; /p ɾ o ! β a . ð o/] confused with "probar" [to test; /p ɾ o ! β a ɾ/]. This could also happen in conjunction with other changes to the word, for example "probado" [proven; /p ɾ o ! β a . ð o/] misperceived as "robar" [to steal; /r o ! β a ɾ/]. Another reason for high word-medial error rates could be the syllable boundary shift when a consonant is deleted in a cluster, as in the confusion between "cambia" [to change; /! k a m . b ja/] and "cama" [bed; /! k a . m a/].

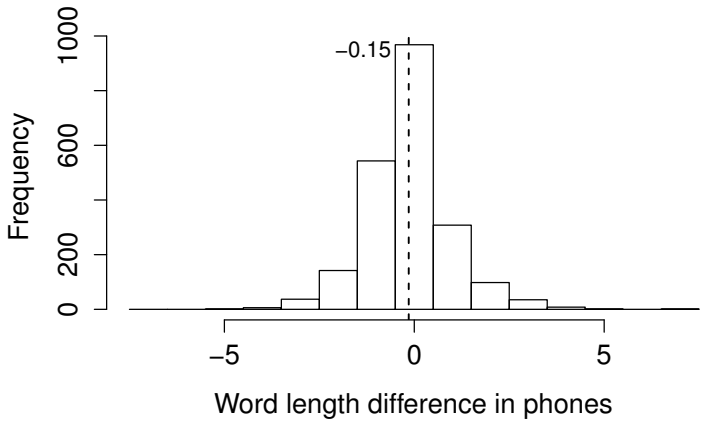


Figure 3.11: *Histogram of confusion-target phoneme length difference. The dashed vertical line corresponds to the sample mean.*

3.8 Word-level effects

3.8.1 Results

At the topmost level of our analysis, we investigated the effects of word-level factors on misperceptions. Starting with length, we found that misperceived words $[M = 5.11, SD = 1.28]$ were slightly shorter than targets $[M = 5.25, SD = 1.20]$, as evidenced by a paired sample t-test $[t(2147) = 6.00, p < .001]$. Figure 3.11 plots the histogram of word length difference measured as the number of phones between misperceptions and target words. The distribution is skewed to the right with a skewness of 0.34 and is significantly different from normal $[p < .001]$, shown by a D'Agostino-Pearson test.

Target-confusion pairs were also evaluated in terms of phonetic similarity. Similarity was measured using a simple distance metric — the Levenshtein distance — as in Felty et al. [2013], where each edit is equally penalised. Figure 3.12 shows the mean edit distance as a function of target word length with a solid line and the mean normalised edit distance with a dashed line. We found that the absolute number of errors per word increased with target word length, while the number of errors normalised on the length of the target word decreased.

As previous work has shown that high-frequency words exhibit many process-

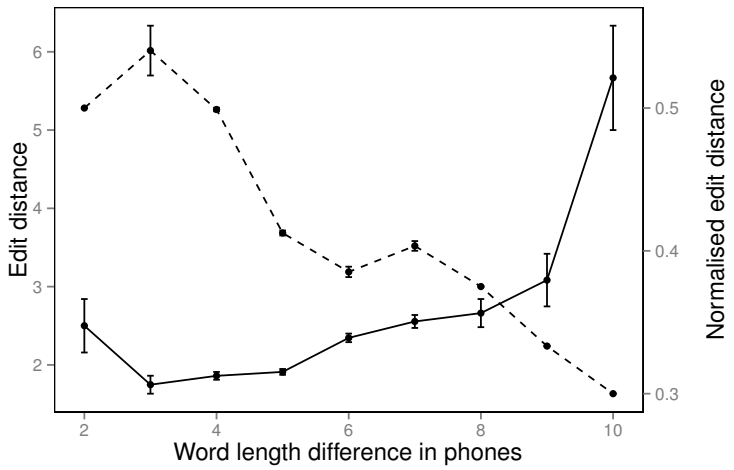


Figure 3.12: *Levenshtein distance (solid line) and normalised Levenshtein distance (dashed line) across target word length. Error bars correspond to* ± 1 *standard error.*

ing advantages [Scarborough et al., 1977; Stanners et al., 1975; Whaley, 1978], it is possible that when in doubt, listeners err on the side of high-frequency words. We found a significant difference in the lexical frequencies of target $[M = 1.25, SD = 0.63]$ and confused words $[M = 1.62, SD = 0.73]$ $[t(2139) = -18.83, p < .001]$. Figure 3.13 plots the histogram of word frequency difference between misperceptions and target words. The distribution is skewed to the left with a skewness of -0.14, and is also significantly different from normal $[p = .007]$. In addition, the frequencies[1] of intended and misperceived words show a weak correlation $[r = 0.19, p < .001]$.

The neighbourhood similarity structure of the target word could also affect confusion patterns. Phonological neighbourhood is often described by the number and frequency of the lexical items phonetically similar to the target word. Here, we used the common definition of phonological neighbourhood for simplicity, namely the set of words which differ from the target in a single phone edit. We found no significant difference between the neighbourhood frequency of intended $[M = 1.04, SD = 0.56]$ and misperceived $[M = 1.069, SD = 0.57]$ words $[t(2139) = -1.468, p = 0.14]$. However, we find that neighbourhood

[1]Word frequency was measured on a log-scale using $log_{10}(f + 1)$ where f is the raw metric expressed as number of occurrences of the given word per million.

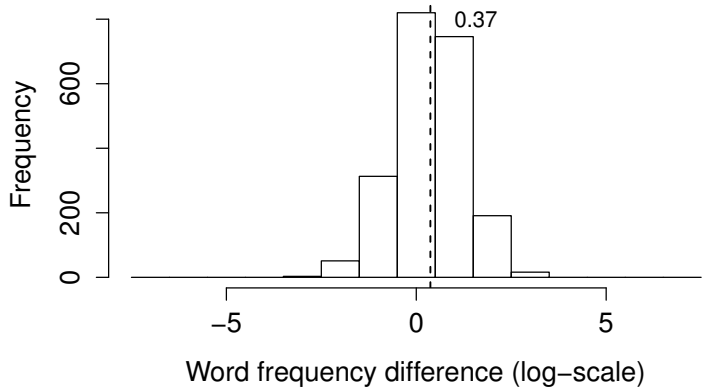


Figure 3.13: *Histogram of difference in word frequency between perceived and intended words. The dashed vertical line gives the sample mean.*

density is higher for confusions [$M = 21.02, SD = 16.67$] compared to targets [$M = 19.31, SD = 16.36$] [$t(2139) = -4.87, p < .001$].

Across the same variables, we also evaluated the differences between words from the list of targets in the elicitation experiment that contributed consistent confusions to those that did not. During the elicitation process, words were selected from the list randomly and mixed with the masker at the appropriate SNR to form the stimuli presented to listeners. Out of the 3968 words used in the original word set, 1856 words contributed at least one consistent misperception while 2112 words did not. We found no significant difference between the length of words resulting [$M = 5.52, SD = 1.33$] and not resulting [$M = 5.60, SD = 1.41$] in confusions [$t(3939.9) = -1.93, p = .053$]. Words not contributing any confusions had significantly higher frequencies [$M = 1.41, SD = 0.66$] compared to words that elicited at least one consistent confusion [$M = 1.30, SD = 0.63$] [$t(3929.8) = -5.25, p < .001$]. Words contributing confusions had both higher neighbourhood density [$M = 18.03, SD = 15.42$] and neighbourhood frequency [$M = 1.02, SD = 0.56$] compared to words contributing no confusions which had the following neighbourhood density [$M = 16.61, SD = 15.07$] and neighbourhood frequency values [$M = 0.96, SD = 0.57$]. The differences between the two group means were significant for both neighbourhood density [$t(3869.5) =$

$2.93, p < .001$] and neighbourhood frequency [$t(3916) = 3.58, p < .001$].

3.8.2 Interim discussion

As research on spoken word recognition — including misperception studies — has been mostly focused on monosyllable words [Benkí, 2003; Luce and Pisoni, 1998; Pollack et al., 1960], few analyses discussed the effects of word length. It is possible that longer words are subject to more lexical competition, as the number of potential embeddings increases with word length, especially in Spanish which has fewer phonemes (25) compared to English (44) and also longer average word length [Maddieson and Disner, 1984]. On the other hand, Pitt and Samuel [2006] argued that longer words will be more accurately perceived since they receive less competition from neighbours and accumulating phonological evidence supports stronger lexical activations. Pitt and Samuel [2006] used the Ganong [1980] paradigm to test this hypothesis.[1] In their experiment, listeners were asked to identify the final phone in words that ended in /s/ or /ʃ/. Stimuli were either monosyllable (e.g., miss, wish) or trisyllabic words (e.g., arthritis, abolish) with word-final consonants selected from an eight step /s/-/ʃ/ phonetic continuum. Pitt and Samuel [2006] reported that longer words generated a stronger lexical shift, further increased by an early uniqueness point. They concluded that length is an important variable in spoken word recognition which can influence lexical activation.

In line with the findings of Pitt and Samuel [2006], Felty et al. [2013] also reported that longer words were more accurately perceived in their study. Tang [2015] reported the same result. While we found that the mean word length was higher for words that did not contribute confusions, the result failed to reach significance. It is possible that failure to observe this effect was due to limiting the target words to maximum three syllables in the elicitation experiment.

Felty et al. [2013] also compared targets and misperceptions in terms of phone and syllable length. In more than 70% of the cases, the target and the misperceived word shared the same number of syllables, supporting the idea that listen-

[1] Listeners' tendency to interpret acoustically ambiguous phonemes consistent with the lexical context is referred to as the Ganong [1980] effect.

ers narrow down the initial list of word candidates using global information such as syllable structure, instead of a strict left-to-right decoding [Savin and Bever, 1970]. In addition, Felty et al. [2013] found that confusions were more likely to have fewer or the same number of phones as the target. In order to test whether shorter words were indeed stronger competitors with respect to longer words, they conducted a Monte Carlo simulation to verify whether this pattern arose due to chance. All of the 10 000 simulation runs produced a word length difference that was closer to zero compared to the observed difference, which allowed the authors to conclude that the length difference between the target and the misperceived word cannot be attributed to chance. Finally, they found that neither the number nor the proportion of phone errors was constant, as the number of errors increased with word length, while the proportion of errors normalised on length decreased.

As in Felty et al. [2013], we also found confusions to be shorter than targets and found similar results in terms of the number and proportion of errors with respect to word length. However, instead of attributing shorter confusions to listener bias, perhaps this trend is related to the confusion inducing adversity. Shorter confusions could result from the deletion of target material due to the energetic masking proprieties of noise. In Chapter 5 we return to this issue and show that irrespective of the type of masker used, confusions primarily due to energetic masking tend to be shorter than targets.

In the past, many misperception studies have investigated word frequency effects [Benkí, 2003; Felty et al., 2013; Luce and Pisoni, 1998; Pollack et al., 1959; Tang, 2015; Vitevich, 2002]. However, findings are not consistent, as certain studies report that misperceptions are higher in frequency compared to targets [Benkí, 2003; Felty et al., 2013; Luce and Pisoni, 1998], while others find no significant differences between the two groups.[Pollack et al., 1959; Tang, 2015; Vitevich, 2002]. For example, Tang [2015] observed that in their compilation of slip of the ear corpora, the direction of the frequency effect depends on the sub-corpora analysed, and found no significant differences overall. Further, they argued that misperceptions in Felty et al. [2013] appeared to be higher in frequency due to a confound with word length, as shorter words are also higher in frequency [Zipf, 1935]. Both Tang [2015] and Felty et al. [2013] agreed, however, that the frequen-

cies of the target and misperceived words were significantly correlated. Similar to Felty et al. [2013] we found that misperceptions had higher frequencies compared to targets and observed a significant correlation between target and misperceived words as well.

We found that words contributing consistent confusions had lower word frequency, higher neighbourhood density and higher neighbourhood frequency compared to those that did not. Thus, our findings matched the predictions of the neighbourhood probability rule proposed by Luce and Pisoni [1998], namely that low-frequency words with high neighbourhood density and frequency are more easily misperceived.

Luce and Pisoni [1998] argued that in addition to lexical frequency, the recognition of a given word will also be influenced by its phonological neighbourhood structure. Virtually all models of spoken word recognition agree that the incoming speech signal activates multiple lexical candidates which subsequently compete for recognition [Weber and Scharenborg, 2012]. These word candidates are likely to share a similar acoustic-phonetic representation, as they were activated by the same speech input. Thus, Luce and Pisoni [1998] hypothesised that words are organised according to acoustic-phonetic similarity in the mental lexicon and that the number and frequency of words similar to the target will affect its recognition. In order to investigate the effects of phonological neighbourhood, Luce and Pisoni [1998] conducted a series of perceptual identification and auditory lexical decision tasks. They found that high-frequency words are easier to recognise and resulted in faster reaction times in a sparse neighbourhood of low-frequency words than the other way around. While the original study [Luce and Pisoni, 1998] employed monosyllabic words, Vitevitch et al. [2008] found that similar effects hold for polysyllabic words.

3.9 Effects of Masker type

3.9.1 Results

One aim of the present analysis was to evaluate the effects of masker type on the misperceptions generated. Figure 3.14 shows the distribution of outcomes across

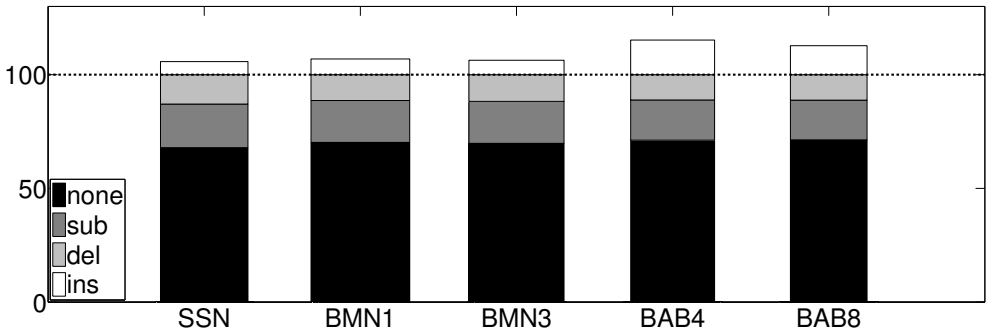


Figure 3.14: *Distribution of outcomes across masker type.*

the five maskers used. Discounting insertions, the outcomes did not show significant differences across masker type [$\chi^2(8) = 11.91, p = .15$]. When including insertions however, the difference becomes significant [$\chi^2(12) = 232.63, p < .001$], confirming the visual impression that speech-based maskers (BAB4 and BAB8) contribute a significantly larger proportion of insertions compared to speech-shaped noise and its amplitude modulated variants.

Figure 3.15 plots the phonetic similarity of target-misperception pairs across masker type. We found that misperceptions stemming from noise-based maskers were more similar to the target than those originating from speech-based ones, with similarity measured using the Levenshtein distance. Masker type had a significant effect on target-confusion edit distance, as shown by a one-way ANOVA [$F(4, 2143) = 18.76, p < .001$]. Post-hoc comparisons using the Tukey HSD test indicated that there was no significant difference in mean edit distance between the two speech-based maskers [$M_{BAB4} = 2.51, SD_{BAB4} = 1.52; M_{BAB8} = 2.33, SD_{BAB8} = 1.33$] and the three noise based maskers [$M_{SSN} = 2.02, SD_{SSN} = 1.01; M_{BMN1} = 1.94, SD_{BMN1} = 1.01; M_{BMN3} = 1.94, SD_{BMN3} = 1.04$]. However, there was a significant difference in mean edit distance for all pairwise comparisons between speech and noise based maskers.

We found that maskers affected the distribution of phone error rates as well. Figure 3.16 shows an association plot between consonant identity and masker type, split according to noise and speech based maskers. The association plot [Cohen, 1980; Friendly, 2000] is often used to better understand the relationships between two categorical variables, once the association has been established via

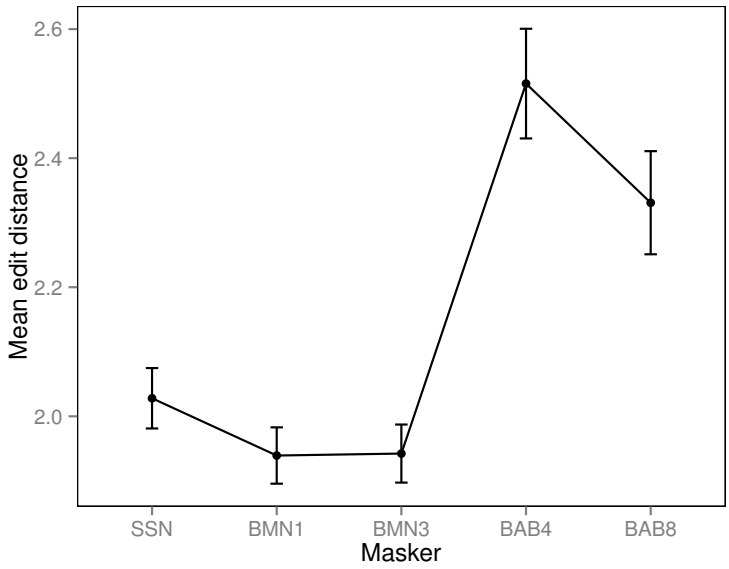


Figure 3.15: *Mean edit distance (Levenshtein) across masker type. Error bars correspond to* ± 1 *standard error.*

the χ^2 test. Association plots visualise the deviation from expected cell frequencies under independence. The height of each cell (signed) is proportional to the Pearson residual, and the width is proportional to the square root of the expected counts, resulting in the area of each cell being proportional to the raw residuals. Shading highlights the residuals that deviate most from independence with solid colour corresponding to residuals individually significant at approximately the 0.05 level. The association plot in Figure 3.16 shows that phones with high-frequency cues such as /ʧ/,/s/ and /t/ were less error prone in noise-based maskers than speech based ones. On the other hand liquids, approximants, as well as the nasal /n/ and plosives /p/ and /b/ are more error prone in noise based maskers.

3.9.2 Interim discussion

In the past, most studies investigating noise-induced misperceptions used a single type of masker. However, through comparisons across studies using different masker types, it has become increasingly apparent that the error distributions of

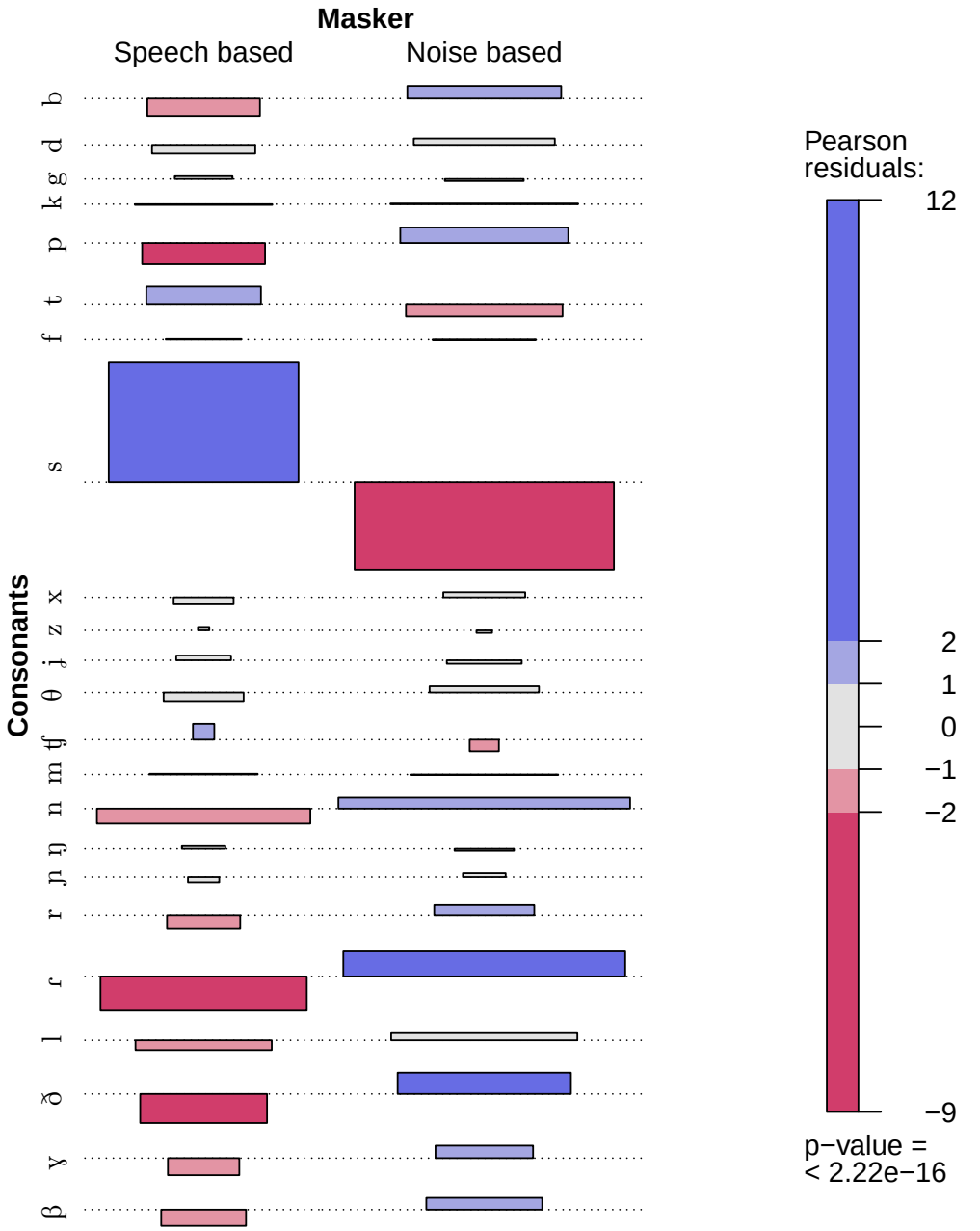


Figure 3.16: *Association plot for consonant error rate across phonetic identity and masker type grouped according to whether it was speech or noise based. Shading indicates cells which deviate most from independence. Cells with solid colour correspond to residuals individually significant at approximately the 0.05 significance level. The p-value corresponds to the significance value of the χ^2 test of independence between consonant identity and masker type.*

speech sounds are greatly affected by the type of masker used. Here we compared the effects of masker type on phone error rates within a single study. In particular, we found several key differences between misperceptions elicited by maskers constructed from speech and by maskers constructed from speech-shaped noise.

Insertions were more common in speech-based maskers compared to maskers constructed from speech-shaped noise. In addition, misperceptions stemming from speech-based maskers were more dissimilar to the target compared to the ones collected in noise-based maskers. One possible explanation for both of these observations is that in speech-based maskers, listeners can incorporate fragments from the background talkers into their resulting percept. In the next chapter, through a signal-dependent analysis of confusions, we will investigate the role misallocation plays in misperceptions in detail and demonstrate that listeners can indeed recruit background speech fragments when forming misperceptions.

As mentioned above, masker type affected the distribution of phone error rates, as the speech cues suffering the most degradation depended on the characteristics of the masker. For example, most of the energy in speech-shaped noise is concentrated between 500 Hz and 2 kHz. Consequently, fricatives occupying the high frequencies will stand out from the average masker level in this region. Phatak and Allen [2007] reported a boost of 10 dB in the SNR spectrum above 6 kHz in their CV misperception study using speech-shaped noise. Accordingly, they found that phones with high-frequency cues such as affricates, sibilant fricatives, and unvoiced plosives were most accurately perceived. A later study by Woods et al. [2010] investigating nonsense syllable confusions in speech-shaped noise also found similar results.

Conversely, in studies employing white noise, the perceptual advantage of consonants involving high-frequency cues was not observed. Miller and Nicely [1955] reported that features voicing and nasality were much less affected by white noise masking than frication or place of articulation. Similarly, Phatak et al. [2008] found that nasals constitute the lowest error consonant group in white noise, while sibilants and affricates were in the mid-to-high error group. When contrasting her findings to prior studies Gordon-Salant [1986] also noted that confusion patterns are dependent on the noise condition. Our results were in line with the findings above. As shown in Figure 3.16 maskers constructed

from speech-shaped noise were less effective at masking consonants with high-frequency cues /ʧ/,/t/ and /s/ compared to babble noise. At the same time, voiced consonants were more robust in speech based maskers.

These results highlight a key limitation of naturalistic studies. As in anecdotal collections, the error inducing adversity is not recorded, so the variability stemming from different types of adverse conditions is not apparent. The above findings suggest that the salience of speech sounds depends heavily on the characteristics of the adversity accompanying it.

3.10 General discussion

In this chapter, we conducted a signal-independent analysis of confusion patterns, investigating the effects of factors related to the target utterance which could potentially influence the listeners' percepts, across multiple levels of speech units. In addition, we examined the effects of the type of masker used for elicitation on the resulting misperceptions.

Among articulatory features voicing, manner and place, we found consonant confusions involving place errors to be most common, followed by errors involving manner and voicing. This trend has been reported consistently in both naturalistic [Bird, 1998; Garnes and Bond, 1980; Tang, 2015] and experimental [Benkí, 2003; Christiansen and Greenberg, 2012; Dubno and Levitt, 1981; Miller and Nicely, 1955; Woods et al., 2010] confusion studies. The fact that place confusions are so prevalent under a variety of different conditions suggests that place cues are inherently vulnerable. This could be explained by their lack of cross-spectral redundancy, in contrast with manner and voicing cues which have been shown to be more redundant and also more robust [Christiansen and Greenberg, 2012].

Overall, our findings suggested that consonants with significant high-frequency components were most accurately perceived. This contrasts with the results of Tang [2015], who found that consonant error rates can be explained in terms of sonority, with the most sonorous consonants being the least error prone. At the same time, our findings were in line with noise-induced misperception studies [Phatak and Allen, 2007; Woods et al., 2010], where the energy of the masker

used for elicitation had a similar spectral distribution. A closer examination of consonant error rates for speech- and speech-shaped noise-based maskers suggested that the high-frequency consonant advantage was only present for the noise-based maskers. This could also explain the above discrepancy of confusion patterns between naturalistic and experimental studies. It is not unreasonable to assume that misperceptions in slips of the ear corpora were collected in settings where the acoustic environment is more similar to the babble maskers used in this study (e.g. social gatherings [Cutler, 1982]) with respect to the noise-based ones. These findings highlight that in most cases the salience of a given phone is highly dependent on the adversities present, an aspect of perceptual confusions often overlooked by naturalistic studies.

In over 96% of the corpus, the vowel of the stressed syllable between the target and misperception was conserved. Overwhelming evidence supports the robustness of the stressed syllable and in particular, the stressed syllable nucleus [Browman, 1980; Garnes and Bond, 1980; Meringer et al., 1895; Tang, 2015]. As prior lab-based confusion studies either involved monosyllable words [Benkí, 2003; Pollack et al., 1959] or did not investigate confusion patterns below the word level [Felty et al., 2013], our study provides the first experimental support for these naturalistic findings.

Word-initial onset and word-final coda positions were the most error prone. These results seemingly contradict previous studies, which reported progressively decreasing error rates from onset to coda and word-initial to word-final position [Browman, 1980; Tang, 2015]. However, when removing confusions involving morphological variation, we obtained similar trends across word position as reported by the studies above. This suggests that the increased error rate in word-final position can be explained by morphological inflexions which occur frequently in the Spanish language. This hypothesis is further supported by the fact that a high proportion of deletions and insertions could be observed in coda position, potentially corresponding to a variation in the suffix.

We found that confusions were shorter, higher frequency words compared to target words, in accordance with Felty et al. [2013]. However, these two findings are potentially dependent, as Zipf [1935] has shown that shorter words are also higher in frequency. One possible explanation of these findings is that noise-

masking has a tendency to delete phonetic material from the target, which results in shorter confusions relative to targets overall. In chapter 5 we examine this argument in more detail. While misperception studies often compare the characteristics of the target and the confused word, the examination of the properties of words that contributed consistent confusions to words that did not, was also highly informative. In particular, we found that our results matched the predictions of the neighbourhood probability rule proposed by Luce and Pisoni [1998], as low-frequency words with dense, high-frequency neighbourhoods were more likely misperceived.

In light of the profusion of factors that govern speech perception, it is not surprising that most experimental investigations approached the problem by focusing on nonsense syllable confusions, thus excluding the effects of suprasegmental and lexical factors. While naturalistic studies usually investigate misperceptions at higher level speech units, these analyses are also often limited to examining a handful of factors. A few investigations [Tang, 2015; Tang and Nevins, 2012] including our current study, have begun to explore the effects of factors across multiple levels of speech units in a single collection. However, even these large-scale studies are unable to encompass all the possible factors and interactions that are known to affect speech perception. Consequently, the questions of how the factors at various levels interact to affect the percept of the listener and what relative importance can be associated to each level, remain largely unaddressed. An added difficulty is that listeners are unlikely to use fixed perceptual strategies, and possibly vary their approach depending on the adverse condition [Mattys et al., 2009]. Further work is needed to explore the interactions that exist among factors at varying levels of speech units and the way in which listeners integrate acoustic cues with lexical knowledge to recognise speech in adverse conditions.

In this chapter, we have explored how signal-independent factors affect misperceptions. However, when trying to understand how perceptual errors arise, signal-dependent factors also need to be taken into account. The analysis of error patterns across masker type in this chapter already provided an indication that confusion patters are greatly affected by the type of adversity giving rise to them. In the next chapter, we attempt to explain misperceptions from a signal-dependent perspective.

Chapter 4

Signal-dependent analysis of misperceptions

4.1 Introduction

While there has been a longstanding interest in devising models which provide predictions of average intelligibility for a variety of adverse conditions [Christiansen et al., 2010; Taal et al., 2010], recently a more in-depth investigation of perceptual errors has received increasing attention [Cooke, 2006; Holube and Kollmeier, 1996; Jürgens and Brand, 2009; Li et al., 2010; Zaar and Dau, 2015]. However, perhaps due to its novelty, this microscopic approach has been understood in multiple different ways. Some studies [Jürgens and Brand, 2009; Zaar and Dau, 2015] define the approach as investigating error rates and confusion patterns of elementary speech units such as phones instead of quantifying intelligibility on the sentence and word level. Others placed the emphasis on analysing each stimulus-response pair *individually*, by establishing a mapping between the listeners' reported percepts and the spectro-temporal characteristics of the eliciting waveform [Li et al., 2010; Phatak and Allen, 2007] . Finally, end-to-end models of speech perception, providing utterance level predictions of listener responses given the input mixture, have also been referred to as microscopic [Cooke, 2006; Holube and Kollmeier, 1996; Jürgens and Brand, 2009].

The analysis in the previous chapter can be considered microscopic according

to the first definition. We have shown how segmental, suprasegmental, lexical and morphological factors affected confusion patterns across multiple levels of speech units. On its own, however, such a signal-independent analysis is insufficient to adequately explain the cause of each individual misperception. Listeners' percepts will, in large part, be determined by the spectro-temporal details of the stimulus waveform, and differences — even across different exemplars of the same utterance and noise type — can result in significant perceptual variability.

In a recent study, Zaar and Dau [2015] aimed to quantify the relative importance of sources of variability in consonant perception. They presented 15 Danish consonants paired with the vowel /i/ spoken by two native talkers to eight normal hearing listeners. Stimuli were presented across six SNR conditions ranging from -15 dB to 12 dB, as well as a quiet condition. Pronunciation variation across and within talkers, acoustic differences between time-shifted exemplars of the noise-masker and listener-related differences were all considered as potential sources of perceptual variability. Conditions were compared using the perceptual distance measure proposed by Scheidiger and Allen [2013], where each set of responses was coded as a vector with a dimensionality equal to the number of response alternatives (i.e. the number of consonants used in the experiment). The perceptual distance between conditions was then quantified using the normalised angular distance between two vectors. They found that the largest amount of variability could be attributed to across- followed by within-talker articulatory differences. Surprisingly, different time-shifted exemplars of the same stationary masker also resulted in significant perceptual variability. While within-listener differences proved to be the smallest source of variability, across listener differences were also large, similar in magnitude to talker-related variability.

In our current study, we control for listener-related variability by constraining the analysis to misperceptions consistently reported across a large listener group. However, the variability stemming from waveform-level characteristics of the speech-noise interaction remains. Consequently, this chapter presents a signal-dependent investigation of misperceptions, in line with the two later definitions of the microscopic approach. We show that the glimpse decoder [Barker et al., 2005] — a human-inspired noise robust speech recognition framework — can serve both as a tool to determine the time-frequency regions in the mix-

ture listeners treated as speech evidence given their reported percept, as well as the potential basis for a microscopic model of speech perception. First, we treat the decoder as a microscopic model and evaluate the percentage of listener confusions well-explained by the decoder in an open-set paradigm. As a byproduct of the recognition process, the decoder returns the set of glimpses — salient spectro-temporal regions corresponding to a single source — that best support the confusion. By evaluating the origin of each glimpse, we can determine the amount of masker material incorporated into each misperception, and classify well-explained confusions based on the type of masker interference involved. Using an unmodified speech recogniser, we also evaluate the number of confusions due to acoustic similarity. Then, we invert the process by conditioning the decoding on the misperceived word, in order to determine the set of glimpses that best explain each confusion through forced alignment. By determining the spectro-temporal regions that best support the listener's percept, we investigate the role misallocation of speech fragments plays in generating misperceptions in babble maskers.

In sum, this chapter aims to conduct a signal-dependent analysis, in order to explain misperceptions based on the type of masker interference giving rise to them. We start by giving a brief outline of the different ways unwanted sources can impede the correct identification of an utterance.

Extraneous sources can interfere with the perception of the target message at multiple stages of auditory processing, starting at the auditory periphery. Upon reaching the listener's ear, signals stemming from various sources in the acoustic scene combine additively. Instead of the entire mixture, however, the listener is more often interested in inferring the properties of a single constituent source. This problem is fundamentally ill-posed, as many source configurations can give rise to the same mixture signal. Weaker target cues occupying the same spectro-temporal region as a more energetic masker component can become undetectable. This is exacerbated by the fact that the ear applies pseudo-logarithmic compression, which can make the contribution of the weaker signal to the resulting nerve excitation even more insignificant. As a result, masked target components are often effectively missing, and listeners are forced to reconstruct the target message based on partial evidence. This phenomenon is known as energetic masking

[Pollack, 1975].

However, in several masking conditions, the intelligibility loss exceeds that which would be expected from overlapping excitation patterns alone. Almost a century ago, Wegel and Lane [1924] had already made the distinction between masking due to signals exciting the same region in the basilar membrane and masking resulting from 'conflicting sensations in the brain'. Later, in their study of spondee identification in multiple masking conditions, Carhart et al. [1969] also noted that listeners experienced excess masking when speech signals were included in the masker complex. They argued that this added masking effect resulted from listeners facing the additional task of correctly allocating signal components before they can recognise the target message. This phenomenon, known as 'informational masking' [Pollack, 1975], is often used to label the compound effect of all processes that result in intelligibility loss beyond energetic masking. Informational masking has been linked to several processes beyond the auditory periphery such as perceptual grouping [Bregman, 1990], source segregation [Brungart and Simpson, 2002; Brungart et al., 2006], auditory selective attention [Cherry, 1953] and cognitive load [Mattys and Wiget, 2011]. (However, since none of the above processes involve overlapping patterns of excitation, Tanner [1958] noted that the term *masking* can be somewhat misleading in this case.)

In order to understand their relative contributions to the overall masking effect and the underlying processes involved, researchers have tried to isolate energetic and informational masking through multiple experimental paradigms [Bronkhorst and Plomp, 1988; Festen and Plomp, 1990; Hirsh, 1950; Rhebergen et al., 2005; Versfeld and Dreschler, 2002] (for a brief overview see Chapter 5). Through this work, misallocation of signal components has emerged as an integral constituent of informational masking for both speech and non-speech stimuli. As such, when applying modifications that perceptually segregate the target from the background sources, the amount of informational masking can be greatly reduced. For example, Neff [1995] has shown that introducing manipulations that promote perceptual segregation of the target results in better identification performance in a tone detection context. Four types of manipulations were selected based on low-level auditory grouping cues identified by Bregman [1990]. Two manipulations

used target tones with shorter durations — 10 and 100 ms targets in a 200 ms masker tone. The third manipulation involved changes in signal quality, by employing a narrow band noise target instead of a tone. The fourth condition separated the target and masker signal dichotically. These modifications resulted in a release from masking of up to 25 dB. However, Neff [1995] also found that as the number of masking components increased, the unmasking resulting from the modifications diminished, since the increasing number of tones in the masker complex shifted the type of masking from mainly informational to energetic, where perceptual segregation provides less benefit.

Perceptual segregation reduces the amount of informational masking in a speech context as well. For example, Festen and Plomp [1990] observed that competing speech from a talker of the opposing gender results in less masking relative to the same-sex condition. Brungart [2001] also studied the factors that influenced the listener's percept when listening to speech in the presence of a single competing talker. They found that in a single competing talker scenario, the effect of masker interference can largely be attributable to energetic masking. Similar to Festen and Plomp [1990] they reported that the voice characteristics of the competing talker had a significant effect, with less masking produced when the competing speaker was of opposing gender compared to a competing talker of the same-sex. Contrary to masking by speech-shaped or speech-modulated noise, the recognition performance in the presence of a single competing talker did not decrease monotonically with decreasing SNR. In fact, recognition plateaued between 0 and 9 dB. In some cases, even an increase in performance can be observed with lower SNR, because level differences help listeners segregate one voice from the other, even if the target voice is presented at the lower speech level.

In a later study, Brungart et al. [2006] has also shown that in a competing talker scenario, the intelligibility loss is primarily attributable to informational masking, more specifically to the inability to correctly segregate target information from the competing speech fragments. By resynthesising the mixture exclusively in regions dominated by the target, listeners achieved near perfect intelligibility compared to only around 50% 20% and 10% recognition scores in the single, two and four competing talker conditions respectively.

To illustrate how energetic masking and misallocation of speech fragments

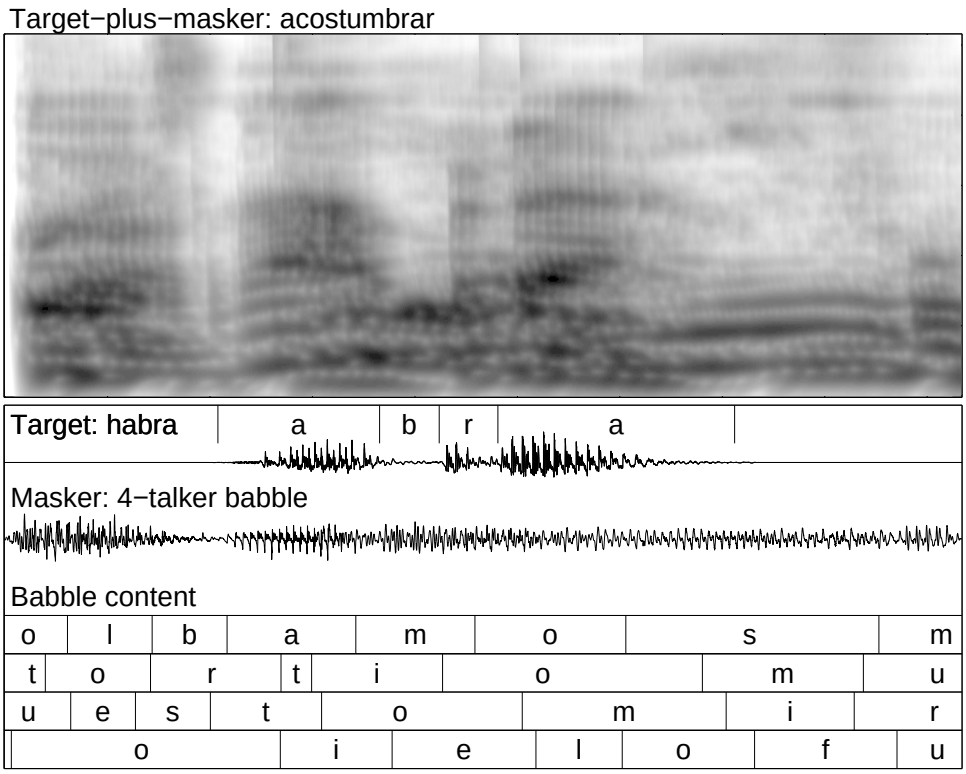


Figure 4.1: *An example robust misperception. Upper: Auditory spectrogram of a speech-in-babble mixture (see 4.1 for details). Lower: target and masker waveforms with phonemic content of target and each individual talker in the babble masker.*

can lead to misperceptions and intelligibility loss, consider the following example from our corpus in Figure 4.1. The target word "habrá" [there will be; /abra/] is reported as "acostumbrar" [to get used to; /akostumbrar/] by 9 listeners when presented in 4-talker babble at a signal-to-noise ratio (SNR) of −0.8 dB. Phoneme transcriptions for the target word and the four individual babble tracks are also shown. While the sequence /bra/ is shared by the target and confusion, it is evident that additional processes are needed to explain the misperception. First, there is some evidence for the incorporation of babble segments corresponding to /o/, /st/, /m/, and later /r/ in temporal locations which are consistent with their inserted positions in the perceived word /akostumbrar/. In these cases, it is possible that some part of the segment is energetically-dominant in the mixture during the relevant intervals. Second, elements of the initial /a/ of the target

word may have been sufficiently masked to render its identity uncertain. Finally, for the initial segments /ak/, while there is no equivalent segment in the babble, segments with vowel and voiceless plosive characteristics occur at the right place. Here, listeners may be using lexical information to hypothesise "acostumbrar" in the absence of a lexical candidate congruent with the acoustic evidence. Thus, it is plausible that the confusion arises through an interplay of energetic masking, the incorporation (i.e. misallocation) of phonetic detail from the masker, and the failure to include certain details from the target itself.

Despite these challenges, in everyday life listeners can comprehend speech and hold conversations across a variety of multi-source environments, many of which would stump state-of-the-art recognition systems. How are listeners able to achieve this feat? Given the complexity of auditory perception, most psycholinguistic research has focused on uncovering the mechanisms involved in understanding a single isolated talker. However, listeners are undoubtedly equipped with a set of additional perceptual strategies allowing them to maintain intelligibility despite the adversities present.

It has been suggested that listeners take advantage of glimpses — salient spectro-temporal regions stemming from a single source — in order to sustain comprehension in adverse conditions [Buss et al., 2004; Cooke, 2006; Miller and Licklider, 1950]. Listeners' ability to piece together salient target regions into a coherent percept has been evaluated for temporal, spectral and spectro-temporal glimpses. By multiplying speech with a square wave with a given frequency and duty cycle, multiple studies have measured listener performance when presented with glimpses of the entire undistorted speech spectrum. In a recent study, in addition to the effects of the interruption parameters, Wang and Humes [2010] also analysed whether linguistic factors affect glimpse integration. Target words were divided into lexically easy and lexically hard based on their word frequency and neighbourhood similarity structure following Luce and Pisoni [1998]. In addition, effects of talker gender and presentation level were examined. Wang and Humes [2010] reported that the proportion of speech glimpsing, as well as lexical difficulty, had the largest impact on recognition. They also found a significant interaction with lexically easy words requiring less acoustic information to be recognised relative to hard words. Glimpse integration has been investigated

across frequency as well. Warren et al. [1995] measured listeners performance in response to bandpass filtered sentences using narrowband slits. Using 1/3 octave-bands and 1/20 octave bands with centre frequencies ranging from 370 to 6000 Hz in two consecutive experiments, Warren et al. [1995] found that intelligibility remains surprisingly high, with up to 95% and 77% respectively for the band centred around 1500 Hz. Finally, glimpse integration has also been analysed for spectro-temporally complex distributions, which are closer to the ones occurring as a result of natural fluctuations in target and masker energy. Howard-Jones and Rosen [1993] filtered pink noise into either 2,4,8 or 16 bands and amplitude-modulated neighbouring bands synchronously or with a 180° phase shift, resulting in coherent glimpses of the target spectrum or a glimpse distribution resembling a checkerboard with varying square size. Speech reception thresholds were improved in the synchronous condition by 23 dB compared to the unmodulated case. However, unmasking in the asynchronous condition was only evident for the case of two bands. At the same time, Buss et al. [2004], using amplitude modulated speech in a steady state masker, found that listeners are able to integrate asynchronous glimpses even when filtered into 16 narrow bands. A recent study by Ozmeral et al. [2012] argued that the lack of benefit provided by asynchronous narrowband glimpses is not perceptually-driven, but due to the peripheral spread of masking occurring for narrowband filters. By eliminating this effect via a dichotic presentation of even and odd frequency bands, they found significantly greater masking release compared to the diotic condition. However, performance still declined as the number of bands increased, suggesting that listeners inability to take advantage of narrowband asynchronous glimpses cannot be attributed to masking spread alone. Other studies have also supplied evidence in favour of the glimpsing hypothesis. The study by Brungart et al. [2006] mentioned above, involving resynthesis of speech exclusively in regions dominated by the target has demonstrated that enough information exists in target glimpses to support the correct identification of the target. Finally, glimpse proportion — the area of spectro-temporal plane glimpsing with respect to the entire mixture — has been successfully used as a predictor of speech intelligibility for both stationary and fluctuating maskers [Tang et al., 2016].

Two key properties of speech that allow recognition based on target glimpses

are sparseness and redundancy. Speech remains intelligible when reduced to three frequency modulated sinusoids following the formant frequencies (i.e. sine wave speech [Bailey et al., 1977; Remez et al., 1981]) or four amplitude modulated frequency channels excited by noise (i.e. cochlear implant speech [Shannon et al., 1995]). In addition, speech is a sparse signal highly modulated in both time and frequency. High-energy regions such as vowel harmonics or high-frequency bursts alternate with regions of lower energy or complete silence, such as the pause before the onset of a plosive. Sparsity can provide listeners with glimpsing opportunities even when the global SNR is quite adverse.

The glimpsing model of speech perception in noise provides an explanation to how listeners can maintain intelligibility when facing partial speech evidence. However, in the presence of interfering speech-like sources, listeners need to correctly segregate target components prior to recognition. In order to accomplish this, Bregman [1990] argues that listeners rely on both source and schema driven processes. It has been suggested that listeners perform bottom-up grouping of coherent spectro-temporal regions exploiting cues such as co-modulation across frequency bands, common onset, location and fundamental frequency for voiced regions. At the same time, there is evidence of listeners relying on top-down processes as well [Remez et al., 1981; Scheffers, 1983].

Over the last few decades, ASR research has put considerable effort into recognising speech in non-ideal conditions. For automatic recognition systems, non-stationary maskers such as competing speech have proved to be the most challenging scenarios. While some of the proposed methods are incompatible with human auditory processing, as they rely on multiple microphones [Hori et al., 2017; Sullivan and Stern, 1993] or use non-auditory features [Ali et al., 2014], other approaches — in light of listeners' high performance — seek inspiration from human speech perception. One such approach, the glimpse decoder, is based on a glimpsing model of speech perception and — similar to listeners — integrates both source and model driven processes. Introduced by Barker et al. [2005], the glimpse decoder is a modified statistical speech recognition framework originally intended for decoding speech in the presence of non-stationary maskers. For our purposes, the advantage of such a listener-inspired approach is that it can also serve as a computational model of speech perception. In this chapter we will

use glimpse decoding as our microscopic modelling approach, both in terms of providing predictions to individual listener responses, as well as determining the spectro-temporal regions listeners most likely treated as target evidence. In the following section, we start by giving a brief outline of glimpse decoding theory.

4.2 Glimpse decoding

4.2.1 Theory

Glimpse decoding is based on the missing data speech recognition framework introduced by Cooke et al. [1994, 2001]. Instead of trying to extract the target signal from the mixture prior to recognition, the missing data approach bases recognition on glimpses of the target, where the signal is largely uncorrupted by the masker and speech separation is superfluous. Non-glimpsing regions in the mixture are either treated as missing or used as an upper bound for speech energy. The missing data approach requires:

1. identification of time-frequency regions in the mixture dominated by target energy

2. modification of the recognition algorithm to handle missing data

The conventional speech recognition problem can be formulated by finding the most likely word sequence W given the series of observation vectors $\mathbf{X}$ corresponding to the clean speech source.

$$\widehat{W} = \underset{W}{\operatorname{argmax}}\, P(W|\mathbf{X}) \tag{4.1}$$

When the observations are corrupted by noise, the problem becomes:

$$\widehat{W} = \underset{W}{\operatorname{argmax}}\, P(W|\mathbf{Y}) \tag{4.2}$$

where $\mathbf{Y}$ represents the series of observation vectors corresponding to the noise mixture. Each observation vector y can be partitioned into reliable components dominated by target energy y_r and unreliable components dominated by the

masker y_u. When used as a robust speech recognition technique, reliable and unreliable regions in the mixture need to be estimated, for example by using primitive auditory grouping cues [Bregman, 1990]. Here, as we use glimpse decoding as a microscopic modelling approach, we can assume *a priori* knowledge of each source in the mixture, and target glimpses can be precisely determined. Once each input feature vector is separated into reliable and unreliable components, the recognition algorithm needs to be modified to handle missing data.

In a traditional HMM-based recogniser, each speech unit (e.g. word, triphone, etc ...) is modelled by an HMM with a given number of states. Each state represents the learned feature distribution of a particular segment of the modelled speech unit. When these feature distributions are approximated using multivariate Gaussian mixtures, the output probability of feature vector y stemming from state C can be computed in the following way:

$$f(y|C) = \sum_{k=1}^{M} P(k|C) f(y|k, C) \tag{4.3}$$

where $P(k|C)$ denotes the coefficients of the Gaussian mixtures. In a noisy mixture signal Y, where unreliable feature values y_u are dominated by masker energy, we would like to base the decision solely on the marginal distribution of reliable components $f(y_r|C)$. This can be achieved by integrating over the unreliable components:

$$f(y_r|C) = \int f(y_u, y_r|C) dy_u \tag{4.4}$$

Exploiting the independence of mixture components this can be written as:

$$f(y_r|C) = \sum_{k=1}^{M} P(k|C) f(y_r|k, C) \int f(y_u|k, C) dy_u \tag{4.5}$$

Thus, recognition can be based entirely on reliable target information by integrating over unreliable feature values [Ahmad and Tresp, 1993]. In certain cases, such as band limited speech, it is appropriate to treat unreliable regions as entirely missing, since they carry no information about the target signal. In the case of noise masking however, the feature values of masker dominated regions can serve

as an upper bound for the underlying target values. Missing data approaches exploit the fact that the value of target feature is between zero and the observed value and the integral can be evaluated using the multivariate error function:

$$\int f(y_u|k,C)dy_u = \frac{1}{2}\left[erf\left(\frac{x_{high,u}-\mu_{u,k}}{\sqrt{2}\sigma_{u,k}}\right) - erf\left(\frac{x_{low,u}-\mu_{u,k}}{\sqrt{2}\sigma_{u,k}}\right)\right] \tag{4.6}$$

where $x_{high,u}$ and $x_{low,u}$ correspond to the upper and lower bound of speech energy in the unreliable regions. Missing data recognition serves as the basis of glimpse decoding. However, instead of partitioning the mixture into reliable and unreliable components prior to recognition, glimpse decoding extends the search into the segregation space, seeking the most likely model-segregation pair simultaneously. To illustrate the connection between missing data recognition and glimpse decoding, consider the spectro-temporal representation of a mixture signal consisting of T frames of F frequency components. Each time-frequency "pixel" of the mixture can either stem from the target or one of the masking sources or both in some cases. In theory, we could simply find the most likely model-segregation pair by applying the missing data approach to the $2^{T\times F}$ possible segregation hypotheses. Of course, this brute-force approach is impractical, considering that the recognition of a single hypothesis is already computationally expensive. Instead, in the following section, we show the series of steps Barker et al. [2005] took to make the computation feasible.

The glimpse decoding problem can be formulated as a simultaneous search over the model and the segregation space for the most likely utterance-segregation hypothesis given the noisy mixture:

$$\widehat{W},\widehat{S} = \underset{W,S}{\operatorname{argmax}}\, P(W,S|\mathbf{Y}) \tag{4.7}$$

As noted above, the search over all possible segregations of a given input mixture is computationally-complex. However, many of these segregations can be ruled out as improbable, since the majority neighbouring spectro-temporal pixels are likely to stem from the same source. By partitioning the time-frequency plane into coherent regions (or glimpses) corresponding to each source in the mix, the size of the segregation space can be drastically reduced. Introducing the set of a

priori glimpses G, equation 4.7 becomes

$$\widehat{W}, \widehat{S} = \underset{W,S\in\mathcal{P}(G)}{\operatorname{argmax}} P(W, S|Y, G) \tag{4.8}$$

where S is understood as the set of partitions – i.e., members of the powerset of glimpses, $\mathcal{P}(G)$ – of the observations into those belonging to the target speech and those belonging to the masker. Note that while the spectro-temporal plane is partitioned *a priori*, the segregation corresponding to the target word is determined through the decoding process (contrary to missing data recognition where target glimpses need to be identified *a priori*). In practice, this can be achieved by taking advantage of the primitive grouping processes introduced by Bregman [1990] such as harmonicity, common onset and offset or location. Again for our modelling approach, we have *a priori* knowledge of the each mixture component so the set of glimpses corresponding to each source can be determined exactly.

Figure 4.2 illustrates how once the *a priori* glimpses have been determined glimpse decoding can be thought of as applying missing data recognition to each possible segregation hypothesis. While the introduction of the *a priori* glimpse set dramatically reduces the size of the search space, the number of possible segregations is still exponential in the number of glimpses. As the number of glimpses is expected to increase linearly with the length of the utterance, the associated complexity can still become prohibitive.

When using a hidden Hidden-Markov model implementation, Barker et al. [2005] proposed a method exploiting the Markovian assumption to share much of the computation, instead of decoding each segregation hypothesis independently. To illustrate this method, consider two segregation hypotheses that are identical until time T, where they differ in whether glimpse N is allocated to the target or one of the masker sources. As the decoding proceeds from left to right, the computation can be shared up until time T, at which point the computation can be split into two branches, allocating glimpse N to either the target or one of the masker sources. When the end of the glimpse is reached, the likelihood score corresponding to the two segregation hypotheses can be compared. Since under this implementation we consider that the Markovian assumption holds, namely that future states only depend on the current state, a decision can be made on

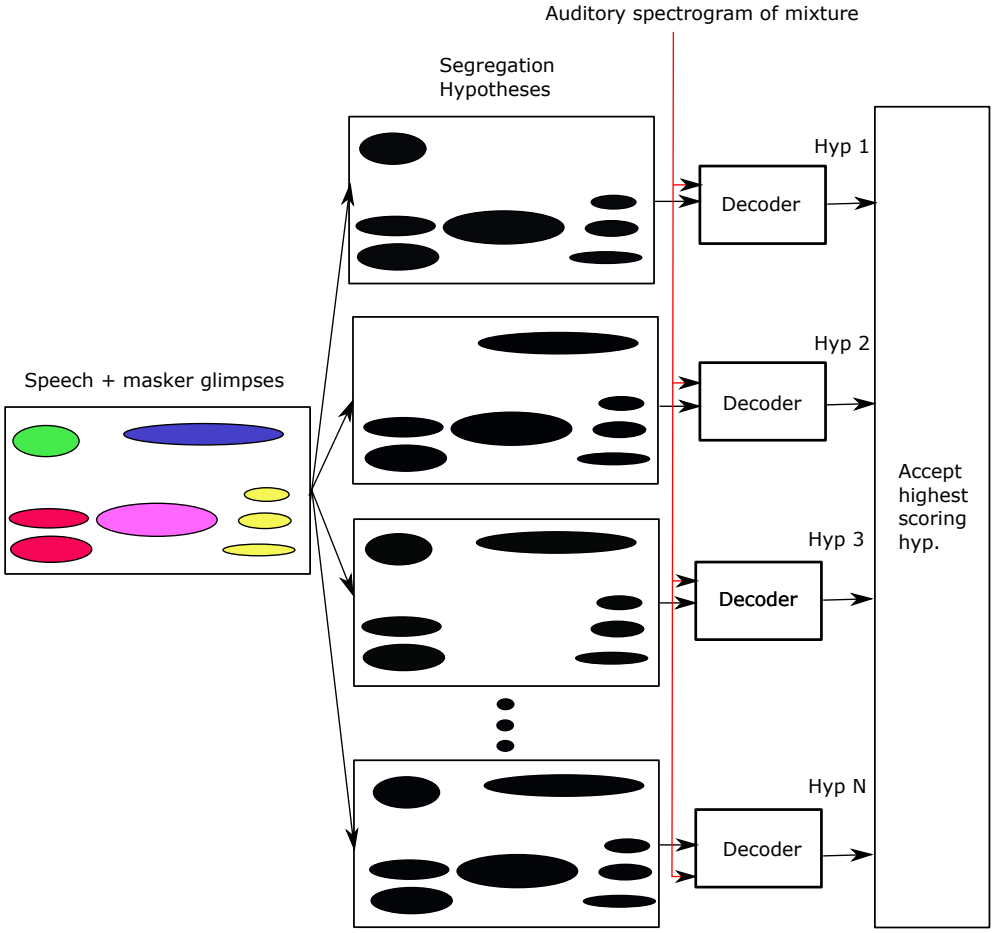


Figure 4.2: *Illustrating the connection between missing data recognition and glimpse decoding. Glimpse decoding is equivalent to applying missing data recognition to each possible segregation hypothesis. Figure reproduced from Barker et al. [2005]*

the origin of glimpse N and the lower scoring branch can be eliminated. With this method, at any given moment the number of concurrent computations will be exponential in the number of active glimpses at time T. It is important to note that while this last step reduces computation, the results are equivalent to the optimal approach of considering each hypothesis independently. While the segregation model was introduced to make the computation feasible, it is of great interest in our current approach as it highlights the spectro-temporal regions in the mixture supporting each word-hypothesis. In the next section, we supply the implementation details of the glimpse decoder.

4.2.2 Implementation

Figure 4.3 outlines the glimpse decoding process. In stage I, we compute an auditory representation of the input mixture. During stage II, prior knowledge of the target speech and masker components is used to generate the entire glimpse set. In stage III, the decoder performs a joint search over the model and segregation space to return the most likely word-segregation pair. These three stages are detailed in the section below.

4.2.2.1 Stage I: Input representation

The target speech and masker waveform are scaled to the presentation-level SNR to obtain the mixture eliciting the consistent confusion. The mixture is then fed into an auditory model which outputs a spectro-temporal representation of the auditory nerve excitation generated by the stimulus. This 'auditory ratemap' representation is computed by passing the mixture signal through a bank of 39 gammatone filters with centre frequencies between 50–8000 Hz equally-spaced on an ERB-rate scale. The instantaneous Hilbert envelope is extracted at the output of each filter, which is subsequently temporally-smoothed, log-compressed and downsampled at 100 Hz. This auditory spectrogram serves as one of the inputs to the glimpse decoder (corresponding to the noisy observation $\mathbf{Y}$).

4.2.2.2 Stage II: *A priori* glimpse generation

A member of the set of *a priori* glimpses G is defined as a connected spectro-temporal region (with 8-connectivity) originating from a single source, with a positive local SNR throughout the region. For the noise-based maskers (SSN, BMN1, BMN3) the mixture y is a sum of two sources, the target and the noise-masker. For speech- based maskers the mixture y is the sum of N sources:

$$y = \sum_{i=1}^{N} s_i \tag{4.9}$$

where s_i is the waveform corresponding to talker i and N is the total number of talkers in mixture y, including the target and each babble component (i.e.,

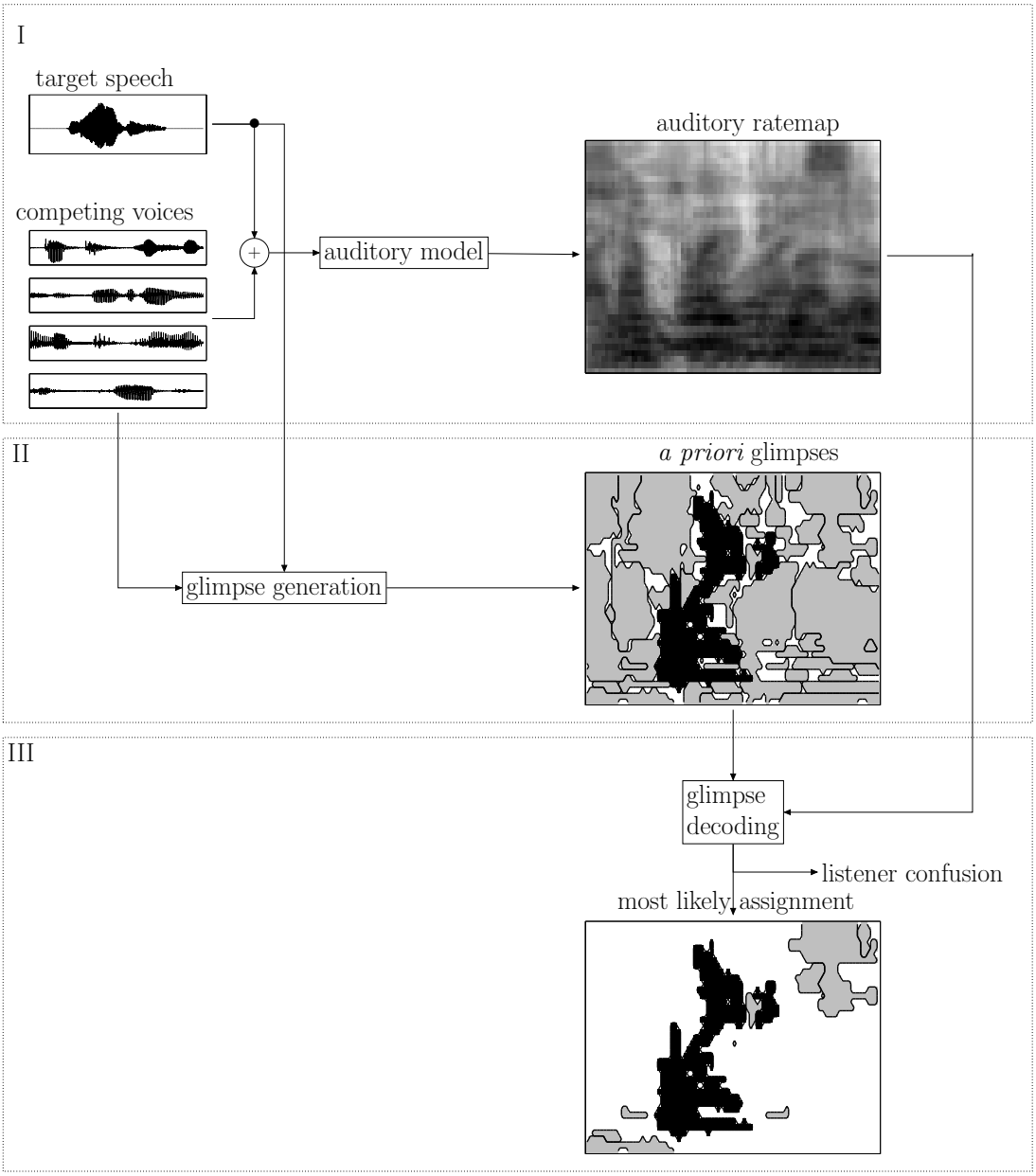


Figure 4.3: *Overview of the glimpse decoding process. I: computation of auditory ratemap; II: generation of a priori glimpses; III: joint search of the model and segregation space for the most likely hypothesis. The glimpses shown in black come from the target (presented) word while those in grey come from the background babble.*

$N = 1 + N_{babble}$ where N_{babble} is either 4 or 8 in the current study). For a given source j in the mixture, we compute separate ratemaps for s_j and the sum of the remaining $N - 1$ signals. The ratemaps are then compared to identify the spectro-temporal regions where the jth source is dominant, i.e.

$$g_{RM}(s_j) > g_{RM}\Big(\sum_{i=1, i\neq j}^{N} s_i\Big) \tag{4.10}$$

where the function g_{RM} maps a time-domain signal to the auditory representation defined in Section 4.2.2.1, and where the comparison is done for each time-frequency 'pixel'. A glimpse of a given source is then defined as a connected spectro-temporal region satisfying the inequality above. This process is repeated for each of the sources in y. The set of fragments obtained for each source are combined to form G — the set of fragments input to the decoder.

$$G = G_T \cup G_M \tag{4.11}$$

where

$$G_M = \bigcup_{i=1}^{N_{babble}} G_{M_i} \tag{4.12}$$

G_T denotes glimpses originating in the target source and G_{M_i} those stemming from the ith babble component. Tiny glimpses which are unlikely to be used by listeners are eliminated from G (in the current study an area threshold of 6 time-frequency pixels is used). An example is shown in the second panel of Figure 4.3. Black glimpses correspond to the target word while those in grey come from one of the background sources. Regions in white correspond to spectro-temporal locations where none of the sources are dominant.

4.2.2.3 Stage III: Glimpse decoding

With the auditory ratemap representation and the predetermined set of glimpses as input, the decoder outputs the most likely word-model and its corresponding segregation. The decoder's acoustic models are speaker-independent 3-state triphone models trained on over 12 000 instances of Spanish word utterances using the same speech material as in the corpus collection described in Section 2.

10-component Gaussian mixtures, with model- and state-level tying, are used to represent the feature distribution of each state.

4.3 Automatic confusion categorisation

4.3.1 Category membership criteria

When decoder predictions match listener responses reasonably well, we can determine the type of masker interference eliciting the confusion. By examining whether each glimpse in the returned segregation originated in the target or masker source, we can determine the amount of masker involvement in each confusion. This way, misperceptions can be placed on a continuum based on the amount of masker information incorporated into the percept. On one end of the spectrum, it is possible that confusions are best explained using target glimpses alone. In these cases, the amount of information contained in the target glimpses was insufficient to support the correct identification of the utterance. We refer to these cases as *reinterpretations* since the listener is forced to generate a new word hypothesis by reinterpreting the partial target evidence available. While maskers such as SSN, BMN1 and BMN3 might elicit this type of confusion, it is also possible that competing speech glimpses can be successfully excluded from the material used to determine the confusion, due to insufficient similarity with the target glimpses in properties such as F0 or formant continuity. On the other end of the spectrum, it is possible that listeners based their response on glimpses stemming from the masker entirely. One can envisage confusions where an acoustically salient word stemming from one of the babble components 'hijacked' the listener's attention and was reported in its entirety instead of the target. These are referred to as *overrides*, and require the masker to contain speech material. Intermediate cases, where the confusion makes use of glimpses of both target and masker are referred to as *blends*. Here, low-level auditory grouping processes probably failed, resulting in incorrect allocation of target and masker glimpses. These confusions are of great interest as the grouping of target and masker material into a single coherent precept probably requires special circumstances. In sum, the amount of masker information incorporated into the percept forms the

basis of the categorisation scheme.

When the confusion is highly dissimilar to the target, we can assume that the masker interference played a key role in forming the misperception. However, a significant portion of the corpus consists of misperceptions where the target and confused word are highly similar, differing in one or two phone edits. In these cases, confusions are most likely explained by the acoustic similarity between the two words, with perhaps some uncertainty contributed by the masker. Finally, some confusions might be attributed to the signal-independent factors reviewed in the previous chapter or other factors we have not considered here. These confusions are labelled as *unexplained*.

We introduce the following criteria for confusion categorisation. If the confused word is ranked in the top three candidates by the baseline recogniser in response to the noise mixture, confusions are classified as due to acoustic similarity. If confusions are ranked within the top 20 candidates by the glimpse decoder and halve their rank compared to baseline, they are classified as well-explained by the decoder, as they likely originate in the speech-noise interaction. This latter criterion is added to classify confusions as well-explained only if they are significantly better explained by decoder instead of acoustic similarity. Well-explained confusions are then further categorised based on the ontology defined above. Confusions are classified as reinterpretations if more than 90 % of the material originates from the target, overrides if it is less than 10 % and otherwise as blends.

4.3.2 Results

1344 confusions ($\approx 42\,\%$) from our corpus can be explained based on the above ontology using glimpse decoding or acoustic similarity, while 1903 cases remain *unexplained* (this figure includes 683 out-of-vocabulary items). Table 4.1 provides a breakdown by masker type. Clearly, the number of confusions in each category depends on the eliciting masker type. Speech-based maskers are more conducive to generating confusions with masker involvement, which is understandable since in these cases speech information can be recruited from the masker. Overrides occur exclusively in BAB4. Further, the frequency of reinterpretations for the

	Reinterpret.	Override	Blend	Acoustic sim.	Unexplained
SSN	19 (6.2)	0	2 (2.1)	211 (22.9)	400 (21.0)
BMN1	134 (43.4)	0	14 (14.9)	232 (25.1)	414 (21.8)
BMN3	69 (22.3)	0	9 (9.6)	268 (29.1)	409 (21.5)
BAB4	42 (13.5)	19	37 (39.4)	116 (12.6)	403 (21.2)
BAB8	45 (14.5)	0	32 (34.0)	95 (10.3)	277 (14.6)
sum	309	19	94	922	1903

Table 4.1: *Counts and percentages of masker types inducing each confusion category.*

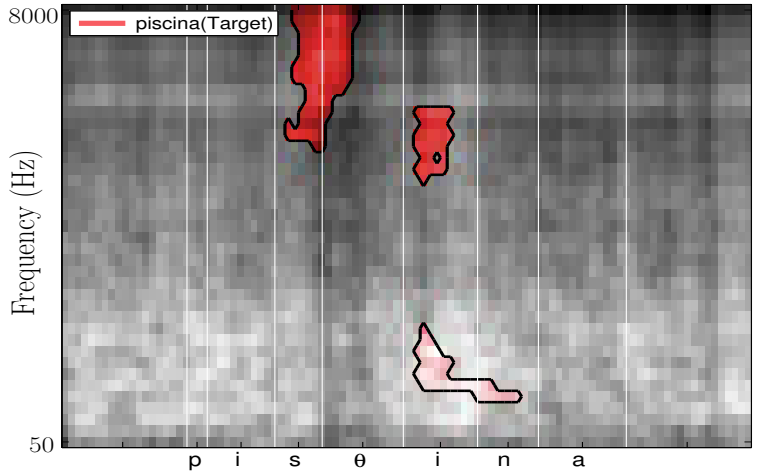
noise-based maskers appears to vary with the depth of amplitude modulation present in the masker (i.e., most for BMN1, least for SSN). The unexplained cases are equally-distributed across masker types apart from a slight reduction for BAB8.

Figures 4.4-4.6 provide an example for each type of well-explained confusion from our ontology. Figure 4.4 shows a reinterpretation. The target word 'piscina' ['swimming pool'], when mixed with BMN3 at −4.11 dB, is reported as 'distinto' ['distinct'] by seven of 15 listeners (other responses include 'destino', 'estino', 'distintos' and 'instinto'). Figures 4.5 illustrates a blend. The target is 'muda' ['mute'] mixed with BAB4 at −2.81 dB, for which 11 listeners reported 'muchas' ['lots']. Figure 4.6 shows an override, where the target 'vuestra' ['yours'] mixed with BAB4 at −2.45 dB is reported as 'manzana' ['apple'] by 9 listeners. For all three examples, the first row shows the phonetic transcriptions of the target and confusion. The second row displays the mixture in the auditory representation used. The x-axis shows the temporal location of phones constituting the target word, determined using the baseline recogniser by force-aligning the HMM state sequence to the target word in clean. Glimpses originating from different background talkers are shown with distinct colours; target glimpses are marked in red. The word utterances corresponding to each talker in the mix are also shown. The glimpses that the decoder selects as part of the best segregation hypothesis for listeners reported word are shown with a thick black border. The glimpses included in the most likely segregation clearly illustrate the masker involvement for the blend and override example.

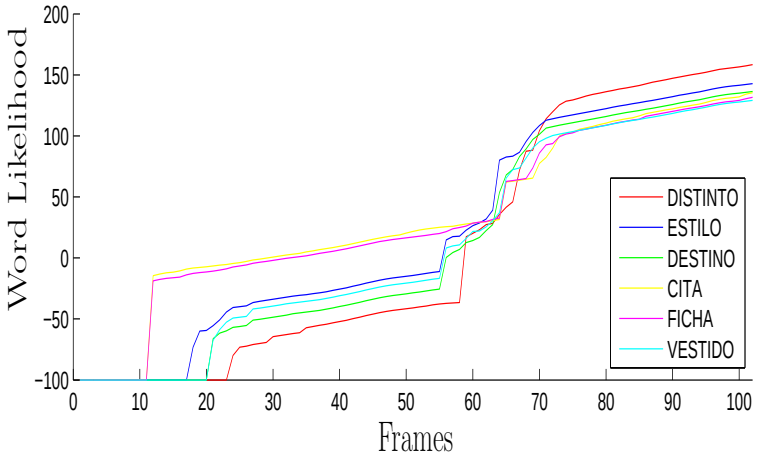
The third row shows the likelihood evolution of the top 6 word candidates

/pisθina/ ↦ /distinto/

Cochleogram, glimpse mask and most likely segregation



Word activations

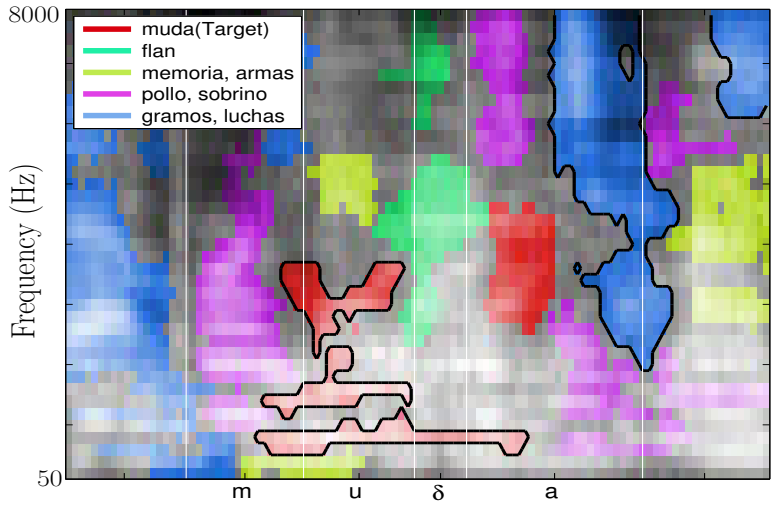


Consistency	Masker	SNR	Other responses
7	BMN3	-4.1 dB	destino(2), piscina, instinto, festín, estinto, distintos, cristina

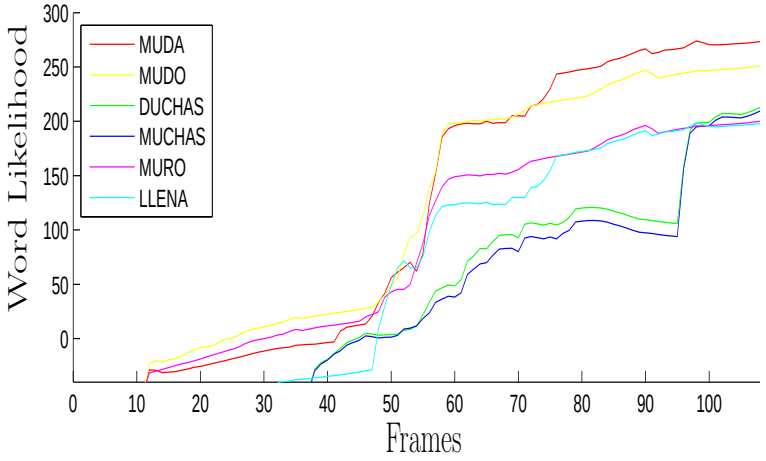
Figure 4.4: *Reinterpretation example. The second row shows the auditory representation used, as well as target and masker glimpses. Log energy values are coded using the lightness dimension, glimpses from different words are distinguished by hue and the segregation hypothesis corresponding to the confusion is shown with a solid border. Vertical lines indicate phone boundaries. Likelihood scores for the top 6 candidates are shown in each 10 ms time frame. The bottom row details the consistency of the majority confusion, masker type, SNR and other responses.*

/muðə/ ↦ /mutʃas/

Cochleogram, glimpse mask and most likely segregation



Word activations

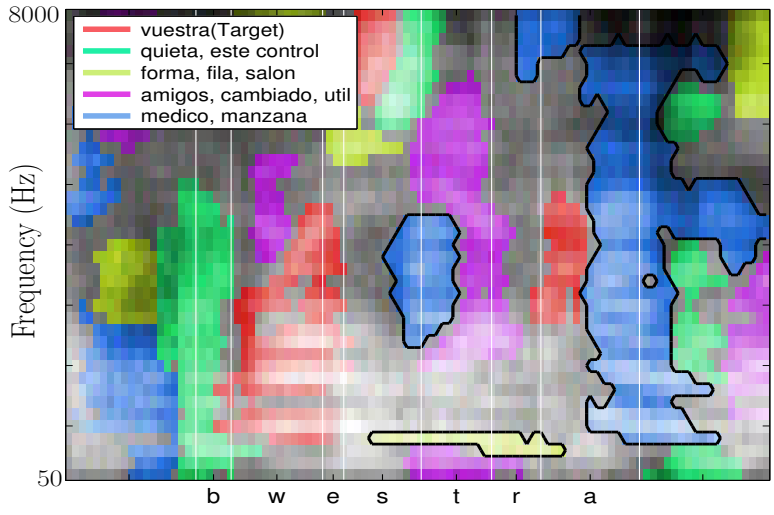


Consistency	Masker	SNR	Other responses
11	BAB4	-2.8 dB	vueltas, mula, cuchara, ampolla

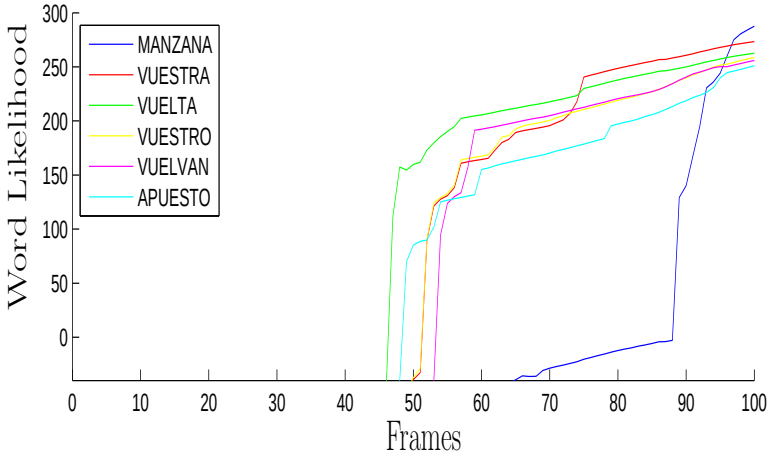
Figure 4.5: *Blend example. Details as for Figure 4.4.*

/bwestra/ ↦ /manθana/

Cochleogram, glimpse mask and most likely segregation



Word activations



Consistency	Masker	SNR	Other responses
9	BAB4	-2.5 dB	puesta(2), manzanas, jueces, cuesta, buena

Figure 4.6: *Override example. Details as for Figure 4.4.*

selected by the decoder. For the reinterpretation example, the decoder correctly predicts listeners' confusion, as it is ranked number one after decoding. Note that the target glimpses correspond to the speech fragment /stɪ/ as shown on the x-axis. This syllable is present in all listener responses. Word activations for the top 6 candidates show a boost in likelihood for those 4 cases containing /stɪ/ at around frames 55-57. For the blend case, listeners appear to recruit glimpses from both 'muda' and the syllable /tʃas/ from 'luchas' ['struggles'] in the babble (shown in blue) to form the confusion. Note that in this case the original target ('muda') is highest ranked after glimpse decoding, while the percept reported by listeners 'muchas' is ranked 4th. For the override example, activation of 'manzana' shows a steep increase in likelihood near the end of the mixture since this word occurs in the latter part of the masker. Here too, the decoder prediction matches listener responses.

Figure 4.7 shows all the confusions in the corpus at a glance as ranked by the baseline recogniser. The x-axis shows the rank of the *confused word* when the target word is presented to the recogniser in the clean condition. This can be seen as a measure of acoustic similarity between the target and the confusion. Thus, if in response to the target word in clean, the confusion is ranked highly, it is likely that the two words are acoustically similar. The fact that the density of points increases towards the left of the figure shows that the likelihood of a confusion increases with acoustic similarity. The y-axis shows the rank of the confused word determined by the baseline recogniser in response to the mixture. Coloured markers show confusions well-explained by the decoder, with red marking reinterpretations, orange marking blends and blue marking overrides. The shape of the marker indicates the masker type. Acoustic similarity cases are shown in black dots, while unexplained cases are shown as grey. Well-explained confusions – which by definition must rank in the top 20 and show a halving of rank – have a wide range of ranks in noise prior to the application of glimpse decoding. Indeed, the mean rank of the reported confusion in noise is 1767, demonstrating that any attempt to explain confusions without somehow separating target and masker components is unlikely to be successful in most cases.

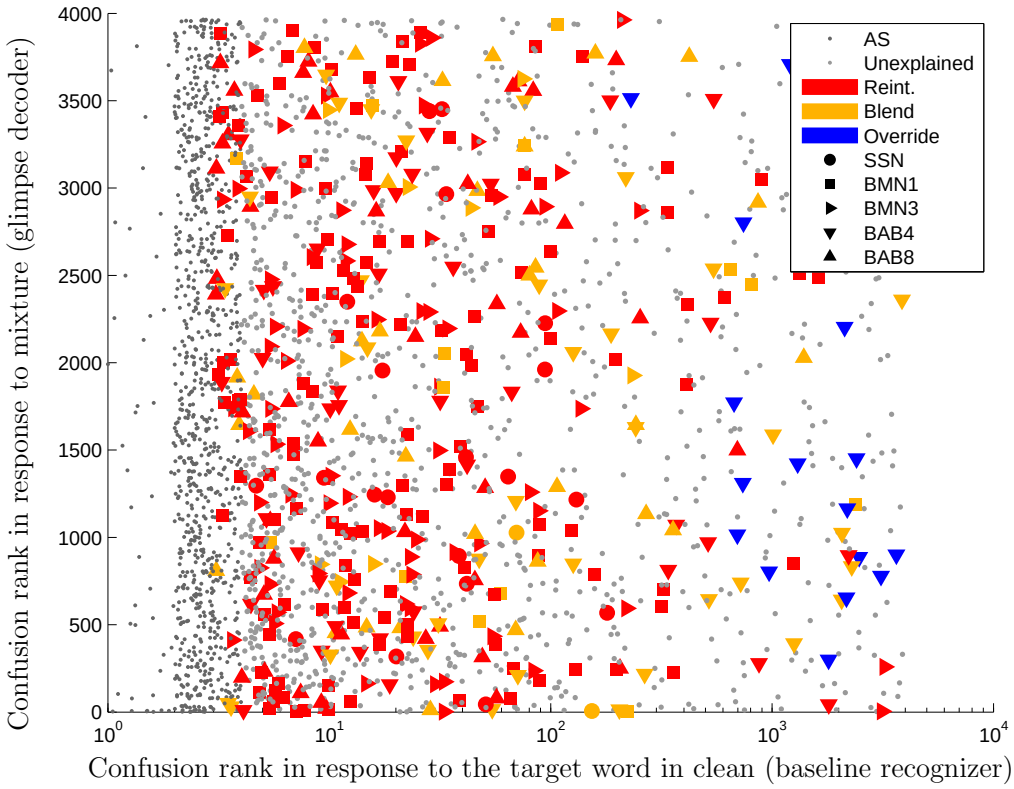


Figure 4.7: *Confusions plotted according to their rank in quiet and noise. Confusions well-explained by the decoder are shown with coloured markers; confusions corresponding to acoustic similarity (AS) are marked with black dots; unexplained cases are shown in grey dots; out of vocabulary cases are omitted. The masker type inducing the confusion is denoted by marker shape. The x-axis is logarithmic to improve visual separation. To avoid clutter, the masker type in which the confusion occurred is depicted only for reinterpretations, blends and overrides. A small jitter has been added in both dimensions to reduce overplotting.*

4.3.3 Interim discussion

A significant fraction of consistent listener confusions could be accounted for by our proposed classification scheme. Nevertheless, many misperceptions remain unexplained, suggesting that our confusion ontology needs to be expanded to take into account other possible causes. One possible direction of future work could be to integrate the signal-independent factors identified in the previous chapter into the model.

While all the maskers in our analysis gave rise to confusions, their quantity and category show some dependency on masker type. Two determining factors of the masker in this aspect seem to be temporal modulation, allowing for glimpsing opportunities, and the presence of speech information in the masker from which listeners can recruit. Unsurprisingly, blends and overrides appear almost exclusively in babble maskers. Overrides, in particular, require low-order babble and are universally caused by the BAB4 masker. Their rarity results from a design decision to avoid using even lower-order babble (e.g., 1- or 2-talker), as from a speech perception perspective these cases are the least interesting. When constructed from many voices (e.g. BAB8) salient words did not emerge, most likely due to the smaller, more fragmented masker glimpses. In studying the way speech and noise interact at a fine-grained level, blends and reinterpretations are of more interest than overrides, and the corpus contains sufficient numbers of both types to support further development of end-to-end models of speech perception. Understandably the majority of reinterpretations were elicited from noise-based maskers, with the modulation depth of the masker conducive to generating reinterpretations. Blends, on the other hand, require maskers with informational content, and both speech-based maskers (BAB4 and BAB8) seem to contribute an approximately equal amount.

Several studies investigating speech perception in multi-talker scenarios have postulated that listeners erroneously reporting masker words is indicative of informational masking, while randomly distributed errors are associated with primarily energetic masking conditions [Brungart, 2001; Brungart et al., 2006; Kidd et al., 2016]. While such errors are present in our analysis as well (i.e. overrides), we have shown that the recruitment of material from the masker is not limited

to lexical items, as listeners can recruit phonemic and sub-phonemic cues as well, and blend them with target glimpses to form a consistent percept. From a speech perception perspective, confusions stemming from misallocation of low-level signal components are the most interesting. While it is possible that severe energetic masking results in randomly distributed error patterns, we have shown that — at least when the stimulus supports a consistent percept — the error responses are not random and tend to build on the phonetic fragments of the target which are glimpsed. In the next section, we analyse how misallocations can lead to misperceptions in speech-based maskers in more detail.

4.4 Quantifying misallocation

4.4.1 Babble subset

This section presents an in-depth analysis of the role misallocations plays in generating misperceptions. We restrict the analysis to speech-based maskers where these types of confusions are most likely to arise, with a special focus on the BAB4 masker.

Masker	N	mean SNR (dB)	SNR range (dB)
4-talker babble (BAB4)	610	-0.66	-3 to +1
8-talker babble (BAB8)	447	-0.51	-4 to +1

Table 4.2: *Details of babble maskers.*

4.4.2 Selecting $\widehat{S}$ through forced alignment

In the previous section, we have shown how the glimpse decoder performs a joint search over the model and segregation space to find the most likely model-segregation pair given the set of *a priori* glimpses and the noisy input mixture.

$$\widehat{W}, \widehat{S} = \underset{W, S \in \mathcal{P}(G)}{\operatorname{argmax}} P(W, S | \mathbf{Y}, G) \tag{4.13}$$

Instead of a joint search, in this section we will condition the decoding of the misperceived word, restricting the search to the segregation space. In other words,

we use the glimpse decoder to find the set of glimpses that best support the percept reported by listeners through forced-alignment. In a conventional HMM recogniser, forced-alignment consists of finding the most likely state sequence Q given the utterance W and the sequence of acoustic feature vectors $\mathbf{X}$.

$$\widehat{Q} = \underset{Q}{\operatorname{argmax}}\, P(Q|W, \mathbf{X}) \tag{4.14}$$

In our case we look for the most likely HMM state sequence $\widehat{Q}$ and segregation given W, G and Y:

$$\widehat{S}, \widehat{Q} = \underset{S,Q}{\operatorname{argmax}}\, P(Q, S|W, \mathbf{Y}, G) \tag{4.15}$$

Thus, in addition to the noise-mixture $\mathbf{Y}$ and glimpse set G, the majority confusion W will also serve as input to the forced alignment process. Figure 4.8 demonstrates how the decoding process is modified to obtain the segregation hypothesis $\widehat{S}$ best supporting each confusion.

4.4.3 Target and masker proportion

Since – from stage II – we know the origin of each glimpse in G (target or masker), we also know the origin of each glimpse in the best segregation $\widehat{S}$. Thus, we can quantify the amount of misallocation leading to each misperception. In order to do so, we introduce the metrics TP and MP, which quantify the proportion of target and masker glimpses incorporated into each misperception. The Target Proportion (TP) is defined as:

$$TP = f_A(\widehat{S} \cap G_T)/f_A(G_T) \tag{4.16}$$

where the function f_A computes the total number of spectro-temporal pixels in the ratemap representation for a given set of glimpses. TP denotes the total area of target glimpses included in the most likely segregation hypothesis $\widehat{S}$ divided by the total area of available target glimpses in the input set G. The Masker

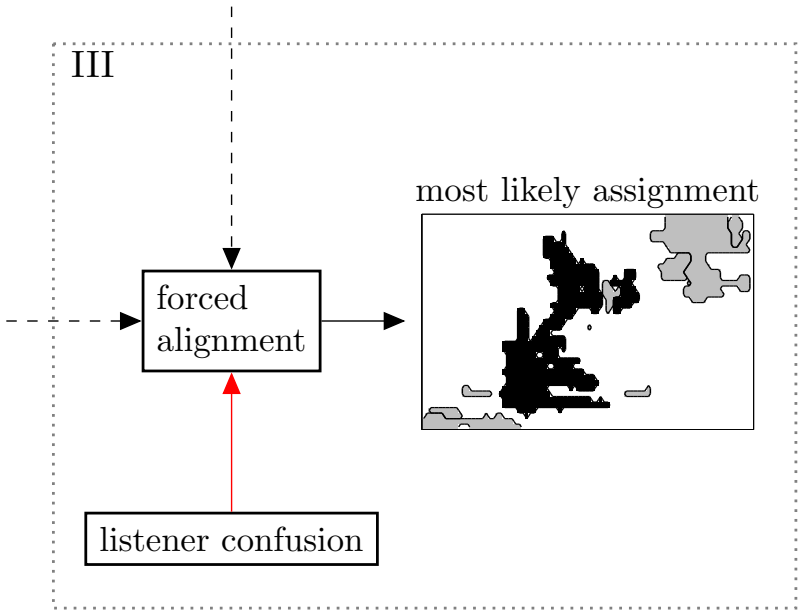


Figure 4.8: *Identification of time-frequency glimpses contributing to listeners' misperceptions. Stages I and II are identical to Figure 4.3 and are omitted. Stage III: forced alignment of glimpses given the listener confusion. The glimpses shown in black come from the target (presented) word while those in grey come from the background babble.*

Proportion (MP) is defined similarly:

$$MP = f_A(\widehat{S} \cap G_M)/f_A(G_M) \tag{4.17}$$

With the above metrics, we can quantify the different types of allocation errors listeners make. A misallocation can occur in one of two ways: either a masker glimpse is considered part of the speech hypothesis (Type I error) or vice-versa, an available target glimpse is excluded from the speech evidence (Type II). MP quantifies the amount of Type I allocation errors while $1 - TP$ quantifies Type II errors.

4.4.4 Results

The left panel of Figure 4.9 compares the two maskers in terms of the proportion of the spectro-temporal area covered by target glimpses. BAB4 produces a little less

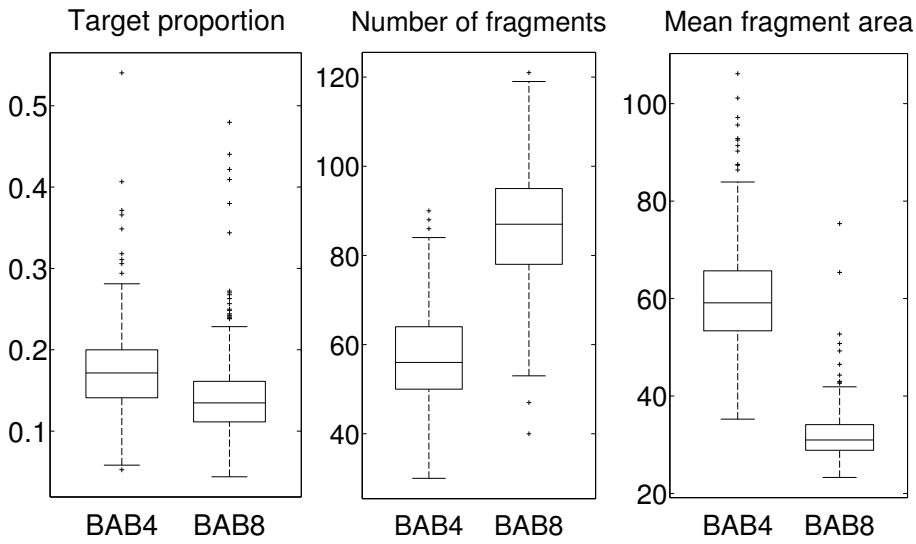


Figure 4.9: *Boxplots showing glimpse properties for speech-based maskers. Left: Proportion of the target escaping masking; middle: total (target+masker) glimpse counts; right: mean glimpse spectro-temporal area.*

energetic masking than BAB8 [$t(936.02) = 9.79$, $p < .001$] in spite of its slightly more adverse mean SNR (Table 4.2). As shown in the central panel of Figure 4.9, the number of glimpses in G (i.e., from both target and masker combined) is substantially higher for BAB8 [$t(838.83) = -38.64$, $p < .001$]. On the other hand the mean glimpse area (right column) is higher for BAB4 [$t(942.77) = 60.37$, $p < .001$]. Thus, we can characterise BAB4 as producing a smaller number of larger glimpses, and vice versa for BAB8. This difference seems likely to affect the types of allocation errors listeners make.

Figure 4.10 shows masker and target proportions, MP and TP, for each confusion, along with marginal densities for both maskers. To ease analysis, the scatterplot has been partitioned into regions corresponding to the different misallocation error types introduced above. Table 4.3 defines these unequal-sized quadrants and details the counts and proportions of confusions that fall into each. More specifically

Q1 covers those cases where very few Type I or II misallocation errors occur: over 95% of available target glimpses are used, and fewer than 5% of masker glimpses are deemed by the decoder to contribute to the confusion.

Q2 represents confusions which have little masker involvement but some loss of available target glimpses (i.e., predominantly Type II errors). Nearly all confusions of this type stem from the BAB4 masker.

Q3 is the region where most of the available target glimpses are used but are accompanied by varying amounts of masker glimpses (i.e., predominantly Type I errors). Here, both masker types elicit similar numbers of misperceptions.

Q4 contains the bulk of confusions and represents a combination of Type I and II errors, with some loss of information from the target and inclusion of masker glimpses.

	TP	**MP**	**BAB4**	**BAB8**	**Type**
Q1	$> .95$	$< .05$	42 (7%)	4 (0.9%)	None
Q2	$\leq .95$	$< .05$	81 (13%)	3 (0.7%)	Type II
Q3	$> .95$	$\geq .05$	116 (19%)	100 (22%)	Type I
Q4	$\leq .95$	$\geq .05$	371 (61%)	340 (76%)	Type I & II

Table 4.3: *Counts and proportions of confusions by error type.*

The distribution of well-explained confusions in Figure 4.7 suggests a correlation between the phonetic distance between target and confusion and masker involvement. To verify this, correlations were computed between the masker proportion MP and phonetic alignment distance. Alignment distance is computed through dynamic programming-based string alignment between the phonetic transcriptions of the target and confused word using penalties of (7,7,10) for insertions, deletions and substitutions respectively. These penalties were selected so that a substitution has a lower penalty than a deletion plus an insertion. A significant positive correlation was found for BAB4 [$r(608) = .39$, $p < .001$] but not for BAB8 [$p = .19$] between phonetic alignment distance and masker involvement. It is clear from Table 4.3 that BAB4 is the most diverse, in terms of the types of allocation errors involved. In the remainder of this chapter, we focus entirely on misperceptions due to BAB4.

Figure 4.11 shows the distribution of target and masker proportions, MP and TP, for the 610 confusions stemming from BAB4. In around 150 cases, (a

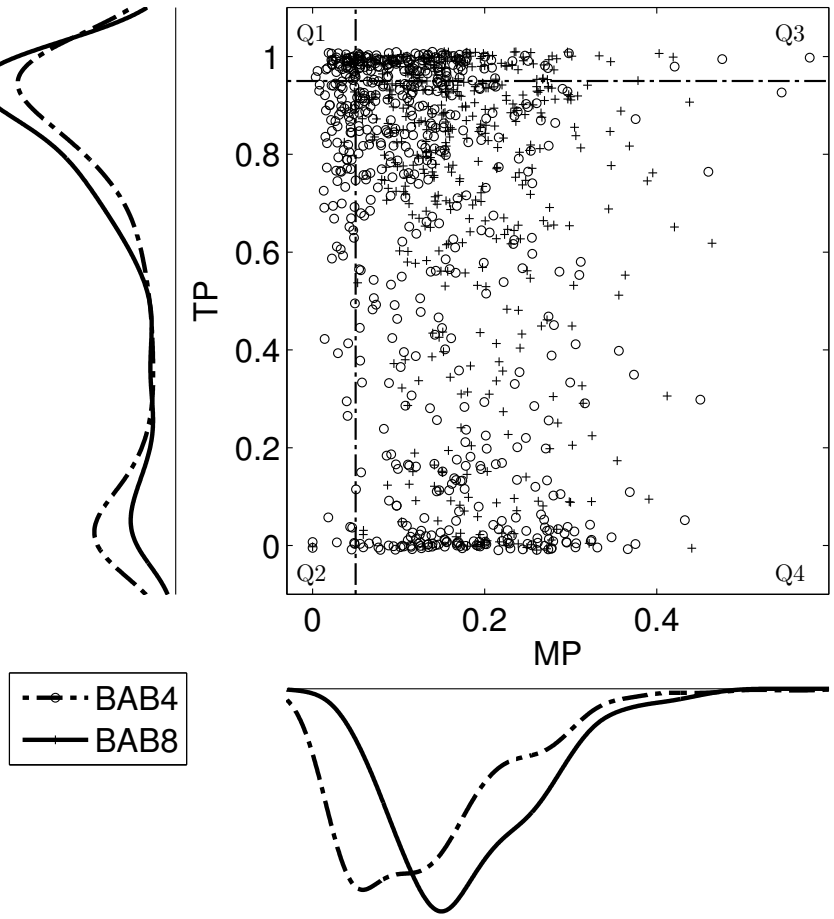


Figure 4.10: *Masker (MP) and Target Proportions (TP) for each individual confusion. The scatterplot is partitioned according to types of allocation error into four quadrants whose boundaries are marked with dotted lines. Marginal densities are also shown. A slight jitter has been added for confusions with* $TP \sim 0$ *and* $TP \sim 1$.

quarter of the total) nearly all of the available target glimpses are used in the misperception, according to the decoder, while half of the confusions make use of at least 80% of the target glimpses. However, about 100 misperceptions use no material from the target at all, corresponding to the override cases shown in the previous section. On average, misperceptions make use of 13% of masker glimpses, and in only 2% of cases is more than a third of masker material used. This is not surprising since the masker consists of 4 talkers in parallel, any one of which could in principle contribute sufficient phonetic information to create a

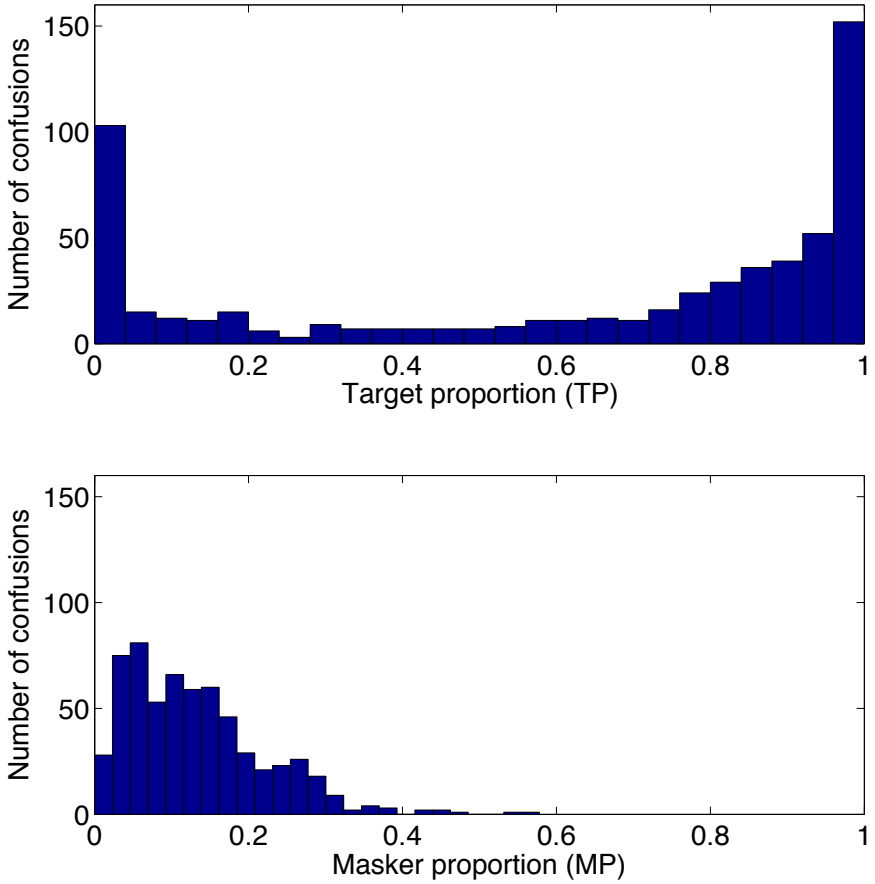


Figure 4.11: *Distribution of target and masker proportions across confusions for the BAB4 masker*

confusion.

The proportion of time-frequency pixels taken up by the glimpses in the most likely segregation $\widehat{S}$ relative to the area of the target glimpses is shown in Fig. 4.12. On average the best hypothesis incorporates 23% more of the spectro-temporal plane than occupied by target glimpses, suggesting that the decoder frequently makes use of a substantial amount of information from the background babble.

Figure 4.13 shows a similar scatter plot as Figure 4.10 for the BAB4 subset, except in addition to target and masker proportion, we also encode three classes of phoneme distance, using a classification scheme similar to that of earlier slips of the ear studies [Bond, 1999b; Garnes and Bond, 1980]. *Single* cases are those involving the deletion, insertion, or substitution of a single phoneme segment (e.g.,

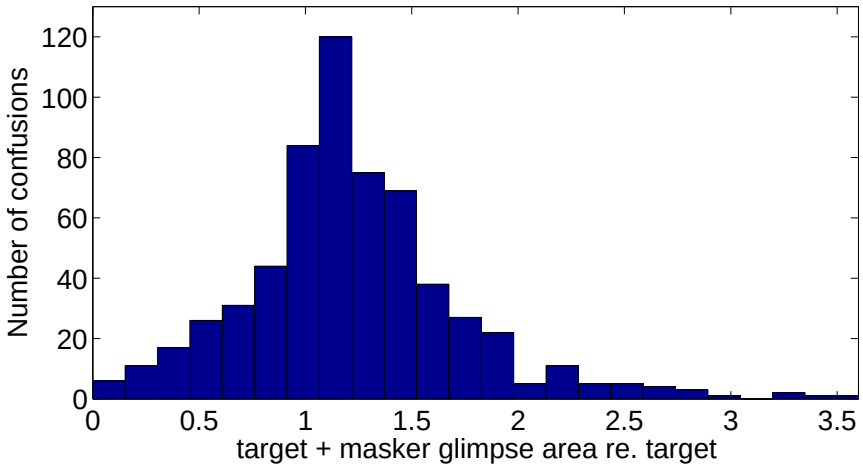


Figure 4.12: *Distribution of the area of glimpses in $\widehat{S}$ relative to the area of glimpses in G_T for BAB4 confusions*

socios ↦ sucios); *dual* cases correspond to changes involving a pair of segments (e.g., sección ↦ disección); all others are denoted *complex* (e.g., antes ↦ alcohol). While *single* cases tend to involve high values of target proportion, there remains a substantial number of cases where the target proportion is reduced. Conversely, *complex* cases typically correspond to low values of TP, but again there is a significant spread. In many such cases, the misperception appears to be due to the masker material overriding the target signal entirely ($TP = 0$). Similarly, the amount of masker involved for all three classes is highly-variable across tokens. These findings suggest that while phoneme distance is correlated with target and masker proportion across the corpus, this kind of segmental metric alone is a poor predictor of the involvement of target and masker glimpses for any given misperceived token.

4.4.5 Interim discussion

Misallocation of signal components can play a key role in the intelligibility loss resulting from listening to speech in the presence of other talkers. The aim of the above analysis was to determine the extent to which misallocations played a role in generating misperceptions in babble noise, by finding the set of glimpses that listeners most likely treated as evidence for their reported percept. In their

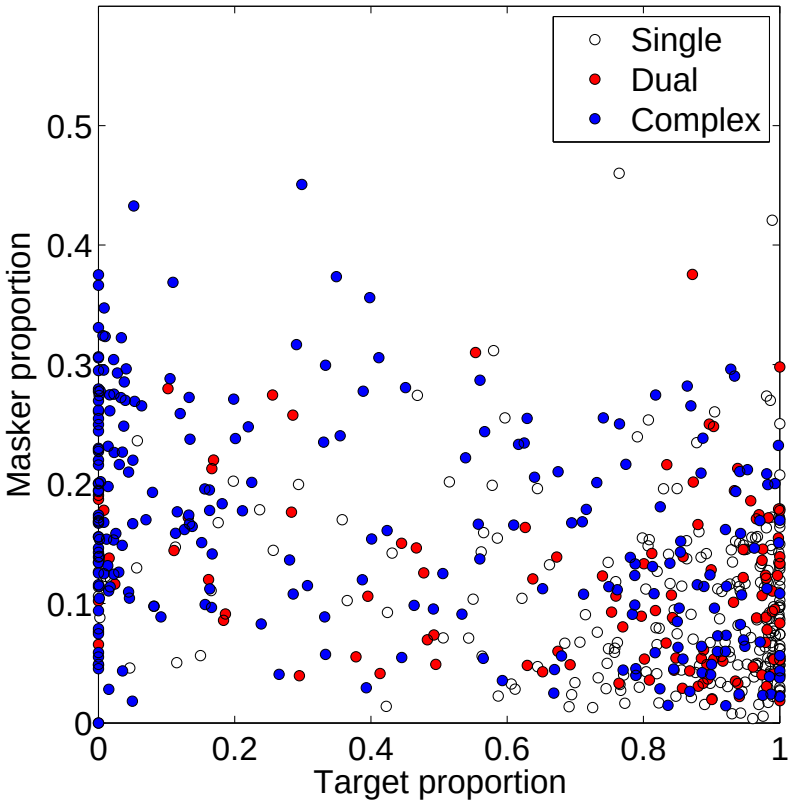


Figure 4.13: *Joint distribution of target and masker proportion for BAB4 confusions, along with target-confusion phoneme distance class.*

analysis aimed at isolating the energetic masking effects of competing background talkers, Brungart et al. [2006] argued that segregation errors can involve incorporating material from the masker into their percept or ignoring relevant information from the target. In the current Chapter, we have introduced metrics to quantify the degree to which these two types of segregation errors are involved in the formation of each misperception. The absence of points around $MP \approx 0$ in Figure 4.10 indicates that most misperceptions stemming from the two babble conditions incorporated material from the masker. In fact, 88% of the confusions involved a Type I misallocation with 20% being exclusively Type I. As virtually all confusions stemming from BAB8 involved a Type I misallocation, the abundance of small masker fragments contributed by eight talkers might be

easier to incorporate, especially in an inflected language such as Spanish. At the extreme of Type I misallocation, a cluster of values near $TP = 0$ are best explained by entirely using masker glimpses. These confusions probably resulted from the listener reporting a word from one of the background talkers and correspond to the 'override' category defined in the previous section. These cases also seem to stem almost exclusively from BAB4. The larger average glimpse size in this masker probably allowed for entire salient words to emerge from the background. Fewer confusions involved strictly type II misallocations, and these cases also originated predominantly from BAB4. Misperceptions involving neither type I nor type II errors may have resulted from energetic masking, perhaps in combination with phonetic similarity. A key finding of the current analysis was the remarkable paucity of this type of misperception which involves no masker glimpses. Again, almost all such cases arose from the BAB4 masker, perhaps since its larger glimpses were more effective at masking phonetic information in the target.

The amount of information borrowed from the masker does not need to be large. We found that on average less than one fifth of the babble masker was incorporated into the misperceptions in the four-talker case, which is reasonable considering that the material contributed by either talker could form the basis of an entire utterance. Nevertheless, misallocating one or more masker glimpses to the interpreted percept, particularly when these glimpses are relatively large as in the BAB4 case, is likely to affect the final outcome for listeners, with a larger phonetic distance from the target word resulting from the inclusion of larger quantities of BAB4 material.

Phoneme distance between the target and confused word explains some proportion of the misallocation effect, but the spread of individual cases is too wide for a segmental metric such as this to be a robust predictor. This is likely to be caused by the strictly temporal nature of the segmental metric. In contrast, the glimpse decoder took into account the spectro-temporal decomposition of the signal. It is possible that more sophisticated forms of alignment, which take into account the segmental constituents of the babble itself (as shown in Fig. 4.1) may lead to better predictions.

4.5 General discussion

In this chapter, we conducted a microscopic investigation of speech misperceptions, by coupling an auditory representation to a multi-source decoder based on the glimpsing model of speech perception. Prior studies implementing a similar approach [Cooke, 2006; Holube and Kollmeier, 1996; Jürgens and Brand, 2009; Zaar and Dau, 2017] have focused exclusively on the prediction of nonsense syllable confusions. Here, we attempted to explain word-level misperceptions from a glimpsing perspective.

In the first part of the chapter, we used the glimpse decoder in addition to an unmodified speech recogniser to sort misperceptions based on whether they were caused by acoustic similarity or a more complex interplay between the target and masker signal. The latter cases, which were well-explained by the decoder, were placed on a continuum based on the amount of masker material incorporated into the percept and classified into reinterpretations, blends and overrides. In the second part of the chapter, we used the glimpse decoder to force align the glimpses in the mixture to the percept reported by listeners in order to quantify the amount of allocation errors involved in each misperception.

In the literature, informational masking is often defined as the masker component contributing to intelligibility loss beyond energetic masking. While this definition is often useful, it conceals the many underlying processes which jointly result in the informational masking effect. Previous work suggested that informational masking can be separated into two major components. On the one hand, it seems that informational masking is closely linked to top-down auditory attention. It has been shown that familiarity with the voice of the target talker [Brungart, 2001; Freyman et al., 2004], as well as knowledge about when [Varghese et al., 2012] and where [Kidd et al., 2005] to listen, can provide a substantial release from informational masking. In addition, the analyses of listener errors in maskers with an informational masking component often showed that many mistakes correspond to listeners reporting words from the masker rather than the target [Brungart et al., 2006; Kidd et al., 2016]. This suggests that in these cases, listeners were either unable to correctly select or to sustain attention on the appropriate speech stream. On the other hand, studies involving non-native

and time-reversed speech [Rhebergen et al., 2005] have shown that the the masker does not necessarily need to be composed of intelligible speech to produce an informational masking effect. Bottom-up auditory grouping cues such as common onset, amplitude modulation and harmonicity have been shown to impact the perception of speech sounds [Darwin, 1981, 1984]. These grouping cues are likely to affect how different signals in the mixture are segregated prior to the formation of auditory streams and objects which the listener can selectively attend to. Errors in this segregation process can also lead to intelligibility loss and probably constitute a crucial component of the informational masking effect.

We have shown how misallocation of low-level speech fragments can generate errors, either by incorporating speech fragments from the background voices into the percept or distracting from vital target glimpses. At the same time, we also observed several misperceptions where a salient word in the background was reported in its entirety. Thus, the findings in our above analysis support the notion that errors in speech segregation and selective auditory attention both contribute to the informational masking effect.

Even though the glimpse decoder proved to be an invaluable tool in explaining the cause of misperceptions in our corpus — in large part through providing the spectro-temporal segregation best-explaining listeners' percepts — it could be argued that glimpse decoding has been less successful from a microscopic modelling standpoint. Only 13% percent of the confusions were well-explained by the decoder according to our criteria, which could question the validity of glimpse decoding as a microscopic modelling approach. However, despite this apparent lack of performance, we argue that glimpse decoding is, in fact, a promising microscopic modelling approach. Prior models of microscopic speech perception have been largely evaluated using nonsense-syllable stimuli in a closed-set paradigm. The approach presented here is a first attempt to the author's knowledge to model listener misperceptions at a word level. This is a significantly more challenging task. First, the number of response alternatives is immensely larger for words compared to nonsense syllables, especially in an open-set task. Second, many of the factors influencing speech perception, including the ones presented in the above chapter such as word position, stress, lexical frequency, as well as factors related to phonological neighbourhood apply only to words. In order to provide

accurate predictions of listeners' percepts, these factors would need to be incorporated into the model. In addition, many other well-known auditory mechanisms, such as forward masking or fragment grouping based on harmonicity and other bottom-up cues were also absent from the model. In light of the above, it is perhaps surprising that the decoder was able to explain so many confusions based on the acoustics alone. Future work could extend the model by incorporating these factors. For example, the trends uncovered in the signal-independent analysis could be added to the model as prior probabilities. The emergence of consistent confusion corpora in multiple languages [Marxer et al., 2016; Scharenborg et al., 2014] is already available and can provide ample diagnostic material, supporting the further testing and development of microscopic models.

Consistent confusions require specific configurations of the speech and masker signal to arise. Changing the SNR or some other aspect of the stimulus will potentially alter the listener's original percept. In the next chapter, we explore the effects signal modifications have on a set of consistent confusions, with the aim of separating confusions caused by energetic and informational masking.

Chapter 5

Determining the origin of confusions through signal modifications

5.1 Introduction

In previous chapters we have shown that consistent confusions can most often be explained by the underlying speech-masker interaction. The interference from the masker — whether by obscuring vital target cues through energetic masking, inducing segregation errors or both — can shift listeners' percepts away from the target to another word hypothesis which they deem more plausible. This raises the question: what aspect of the speech-masker interaction caused the confusion and how does the confusion eliciting stimulus need to be modified to allow listeners to correctly identify the intended utterance.

The goal of the present chapter was to determine whether confusions originated from energetic or informational masking using a follow-up perceptual experiment. Starting from the original confusion-inducing stimuli, we introduced signal modifications selected for their diverse masking release properties and subsequently re-evaluated listeners' percepts. By determining which modifications were successful in allowing listeners to correctly identify the target word across different masking conditions, we could hypothesise the type of masking that caused

the misperceptions in the first place.

Several studies in the literature followed a similar methodology of re-evaluating listener responses after applying signal modifications to the original stimuli. Li et al. [2010] applied signal modifications in 3 dimensions: time truncation, spectral filtering and noise masking, in an effort to triangulate the position of stop consonant cues. They showed that stops are characterised by a short burst, followed by the second formant transition, though the latter is not necessary for the perception of /ta/ and /ka/. Similar to the current study, Cooke [2009] confirmed the feasibility of a large-scale collection of consistent confusions by measuring the rate at which such confusions occur and also applied signal modifications to uncover their cause. Cooke [2009] found relatively few confusions to result from energetic masking, suggesting that the majority of confusions arose from more complex speech-masker interactions. Varnet [2013] aimed to uncover the spectro-temporal location of perceptual cues relevant in distinguishing between syllables /aba/ and /ada/ using a slightly different approach. Instead of introducing systematic signal modifications, they applied the classification image technique, which involves establishing a correlation map between each individual noise field and the corresponding listener response. By presenting a large number of stimuli (over 5000 noise exemplars for each target syllable), they identified the spectro-temporal regions and masker levels which resulted in a perceptual difference. Using this technique, they confirmed the result of Liberman et al. [1954], namely that the second formant transition is key in distinguishing between consonants /b/ and /d/.

When trying to understand how the presence of a masker signal can alter a listener's percept, one of the main questions is whether the confusion was caused by energetic or informational masking. In the past, several experimental paradigms have been introduced in an effort to isolate the masking contribution of the energetic and the informational components of the interfering signal. Speech modulated noise has been used instead of speech-shaped noise to obtain a better approximation of the energetic masking effects of competing speech, by matching not only the long-term spectrum but also the broadband intensity fluctuations of the speech envelope [Festen and Plomp, 1990; Versfeld and Dreschler, 2002]. For example, Festen and Plomp [1990] compared speech modulated noise to steady

state noise and competing speech, in order to determine the release from masking due to masker fluctuations for listeners with normal hearing and sensorineural hearing loss. They showed that listeners with hearing loss do not benefit from the same unmasking caused by masker fluctuations as normal hearing listeners do.

While speech-modulated noise does contain temporal modulations, it has many fewer modulations in frequency compared to real speech, which can provide listeners with additional glimpsing opportunities. Time-reversed speech, which has a spectro-temporal profile similar to normal speech, has also been used as a masker and is expected to produce similar energetic masking effects while remaining unintelligible. However, there are also several drawbacks to this approach. First, time reversal introduces important changes to the speech envelope. For example, plosives — which are characterised by a sudden onset and a gradual decay — when reversed result in abrupt offsets, which can cause a large amount of forward masking. Second, not all informational masking effects can be attributed solely to the intelligibility of the masking signal, as time-reversed speech can also have a significant informational masking component. Rhebergen et al. [2005] used normal and time-reversed speech as a distractor in both a foreign language and listeners' native language. As foreign speech is unintelligible when presented in both a forward and backwards direction, the authors argued that the differences in masking could be attributed to the forward masking effect. They found that the non-native speech masker when time-reversed resulted in speech reception thresholds (SRTs) 2 dB higher compared to when presented in the forward direction. For native speech, they found that the difference in speech reception thresholds (SRT) between time-reversed and normal presentation was around 4.4 dB, the combined result of a release from informational masking due to intelligibility and an increase in forward masking. In addition, they found that SRTs were 3.6 dB higher for native versus non-native time-reversed speech. The fact that both native and foreign speech is unintelligible when time reversed suggests that the masker does not necessarily have to be intelligible to produce informational masking.

Other studies have exploited binaural effects in order to try to isolate the energetic and informational components of the masker. Spatial separation has

been known to result in masking release independent of the type of masker used [Bronkhorst and Plomp, 1988; Hirsh, 1950; Peissig, 1997]. However, Freyman et al. [1999] found that this unmasking is greatly reduced in a reverberant condition for a steady-state masker. By adding a single simulated reflection, the masking release of a spatial separation of 60° was reduced from 8 dB to 1 dB or less, compared to the anechoic condition. While they observed a reduction of unmasking of similar magnitude from 14 dB to 9 dB between the two conditions when the masker signal was a female competing talker, the benefit of spatial separation was present in both the anechoic and the reverberant condition. Freyman et al. [1999] concluded that when the masker has an informational component, spatial separation, whether actual or perceived, can provide additional segregation cues which listeners can take advantage of.

Brungart and Simpson [2002] used a hybrid monaural-dichotic paradigm to separate informational and energetic masking effects. By presenting the target in a single ear and maskers in both ears, they investigated within-ear and across-ear speech segregation. They found that listeners had little difficulty in segregating the target when the masker was absent or was steady-state noise in the unattended ear. However, when speech and speech-like signals — such as time-reversed speech — were presented across ear, listeners' ability to segregate the target in the attended ear was seriously degraded, suggesting that within-ear and across-ear segregation is difficult to perform simultaneously. This highlights that correct segregation of the target talker from similar masking signals relies on limited attentional resources.

Informational and energetic masking effects have also been separated using more elaborate signal processing techniques. Arbogast et al. [2002] isolated these two components of the masker using cochlear implant simulation. They reduced spectral overlap between target and masker signals by allocating different frequency bands for each, thus minimising the energetic masking component. They confirmed that spatial separation provides the biggest benefit for informational masking release. A 90° separation produced an 18 dB advantage over the non-separated condition for a different-band speech masker expected to produce only informational masking, compared to a 7 dB gain for the same-band noise masker expected to mainly result in energetic masking. Brungart et al. [2006] approached

the problem from the opposite direction, by eliminating the informational masking component. By resynthesising the mixture signal only in the spectro-temporal regions that exceeded a local SNR threshold, only target information surviving energetic masking remained. They found that eliminating the regions dominated by the masker resulted in intelligibility improvements ranging from 50% to 90% compared to the unsegregated condition depending on the number of talkers. The resynthesis condition resulted in almost 100% intelligibility, suggesting that for the conditions tested — namely 1,2 and 3 competing talkers — loss of intelligibility is entirely attributable to informational masking.

All of the techniques mentioned above were designed to provide release from either informational or energetic masking effects. Kidd et al. [2016] proposed to apply several of the above techniques including same- and different-sex talkers, spatial separation and time-reversal in conjunction with glimpse resynthesis to better understand the how these modifications affect the energetic and informational component of the masker.

The resynthesis condition served as a control for energetic masking. Based on listeners' performances in the resynthesis condition, Kidd et al. [2016] concluded that the amount of energetic masking produced by each modification, including the unmodified condition was about the same. Consequently, they confirmed that all modifications i.e. different sex talkers, spatial separation and time-reversal produced a substantial release from informational masking. They also found that time-reversal resulted in the smallest release from energetic masking. Finally, large individual differences were observed in listeners' ability to take advantage of the cues provided by the masking release conditions.

Brungart et al. [2013] studied listeners' performance in a variety of listening tasks with differing levels of complexity with a single competing talker or a continuous noise masker present. In line with previous findings, they reported that listeners perform better in simple tasks in the competing talker condition containing unrelated speech, than in a continuous noise condition when the amount of energetic masking produced by the two conditions is similar. However, Brungart et al. [2013] argued that this comes with the cost of allocating additional cognitive resources in order to be able to segregate the target from the speech based masker, which is more difficult compared to when the masker is noise-based. In a series of

listening tasks of increasing complexity, Brungart et al. [2013] showed that with increasing task complexity, the performance of listeners showed a sharper decline for the competing talker relative to the continuous masker condition. Their results seem to confirm the hypothesis that understanding the target utterance in a speech-based masker requires more cognitive resources.

Our goal in the present Chapter is to determine the extent to which energetic and informational masking effects are responsible for generating misperceptions across the five masker types used in our corpus. Inspired by some of the studies above, we used signal modifications that specifically target release from either the energetic or the informational component of the masker, and evaluate the contribution of each component based on listeners responses. Modifications involved a simple increase SNR, glimpse resynthesis and shifting of the target F0. We present the details of the modifications in the following section.

5.2 Modifying speech-in-noise confusions

5.2.1 Control condition

Our first experimental condition involved presenting the confusion inducing stimuli as in the original elicitation experiment, with no modifications. This condition allowed us to select the subset of tokens for analysis which successfully reproduced the original confusion with the same consistency as in the original collection. This allowed us to ensure that the changes in the percepts reported by listeners were indeed caused by the introduced modifications. At the same time, this condition provided a means to measure the rate at which consistent confusions can be reproduced for a different listener cohort.

5.2.2 SNR increase

The first modification employed was a 3 dB increase in SNR. For a single competing talker where the masking effect is expected to be primarily informational, Brungart [2001] has shown that listener performance is relatively constant in a range from -12 to 0 dB, while performance for speech-shaped noise increases

manipulation	condition(s)
none	original (control)
SNR increase	SNR increased by 3 dB
F0 shift	-1, 1, 2, 3 semitones
Glimpse resynthesis	target glimpses alone
	target glimpses+low-level noise

Table 5.1: *Experimental conditions*

monotonically in the same range. As our misperceptions were largely collected in a similar SNR range, we expected a 3 dB increase to primarily result in a release from energetic masking. Thus, we hypothesised that if listeners respond with fewer instances of the prior confusion and more of the correct target word following this manipulation, the original confusion was probably caused by energetic masking.

5.2.3 Resynthesis from glimpses

One way to assess the extent to which listeners are utilising information from the masker in reporting a confused percept is to resynthesise just those parts of the target signal that are deemed to survive energetic masking. In this way, no parts of the masker are presented to the listener. We hypothesised that if listeners continue to report the original confusion following resynthesis, the misperception is probably caused by energetic masking. Listeners reporting the correct target word instead implies that sufficient information exists in the target glimpses for correct identification. We interpret this as a consequence of removing the informational masking effect of those parts of the stimulus not belonging to the target, a form of release from informational masking. A third possible outcome is that listeners report something other than the original confusion or the target word.

Resynthesis from glimpses is performed by first determining target glimpses — spectro-temporal regions in an auditory representation where the target word is more energetic than the masker [Cooke, 2006] — then passing the speech-plus-masker signal through a zero-phase gammatone filterbank, selectively gating glimpsed regions in each frequency channel, and summing across channels. The zero-phase filterbank ensures that the resynthesised signal possesses the same

phase structure as the original signal, and is implemented following Weintraub [1985], by filtering the signal, time-reversing the output, filtering the signal for a second time, and time-reversing the output again. Two experimental conditions were tested, one in which glimpses alone are presented (3rd panel, Figure 5.1), the other where a speech-shaped noise is added at 12 dB SNR to the un-glimpsed spectro-temporal regions (4th panel, Figure 5.1). This latter condition was used to mask possible artefacts caused by the discontinuities that appear when resynthesising speech from the target glimpses.

5.2.4 F0 shift

In the previous chapter, we have shown that confusions might result from allocating parts of the masker to the speech hypothesis. One way in which this is thought to be catalysed in listeners is via similarity in F0 between target and masker. For instance, it is more difficult to identify simultaneously-presented vowels if they have the same F0 [Bird and Darwin, 1998; Scheffers, 1983]. Another study by Brungart [2001] showed that speech masked by the same talker is the least intelligible, followed by same sex and opposing-sex talkers. By modifying the F0 difference between the target and masker, we hypothesised that any confusions that reverted to the correct target word are dominated by informational masking. Four conditions were tested, corresponding to shifting the F0 of any voiced regions of the target word by -1, $+1$, $+2$ and $+3$ semitones. Larger shifts were avoided, as they tended to change the perceived gender of the male target talkers. We chose to manipulate the F0 of the target word since unlike the masker, it has at most a single F0. STRAIGHT [Kawahara et al., 1999] was used to achieve F0 shifts. Figure 5.2 depicts the F0 shift cases.

5.3 Perception experiment

5.3.1 Stimuli

A subset of 800 tokens was selected from the Spanish Confusions corpus presented in Chapter 2. As the goal of this chapter is to introduce signal modifications

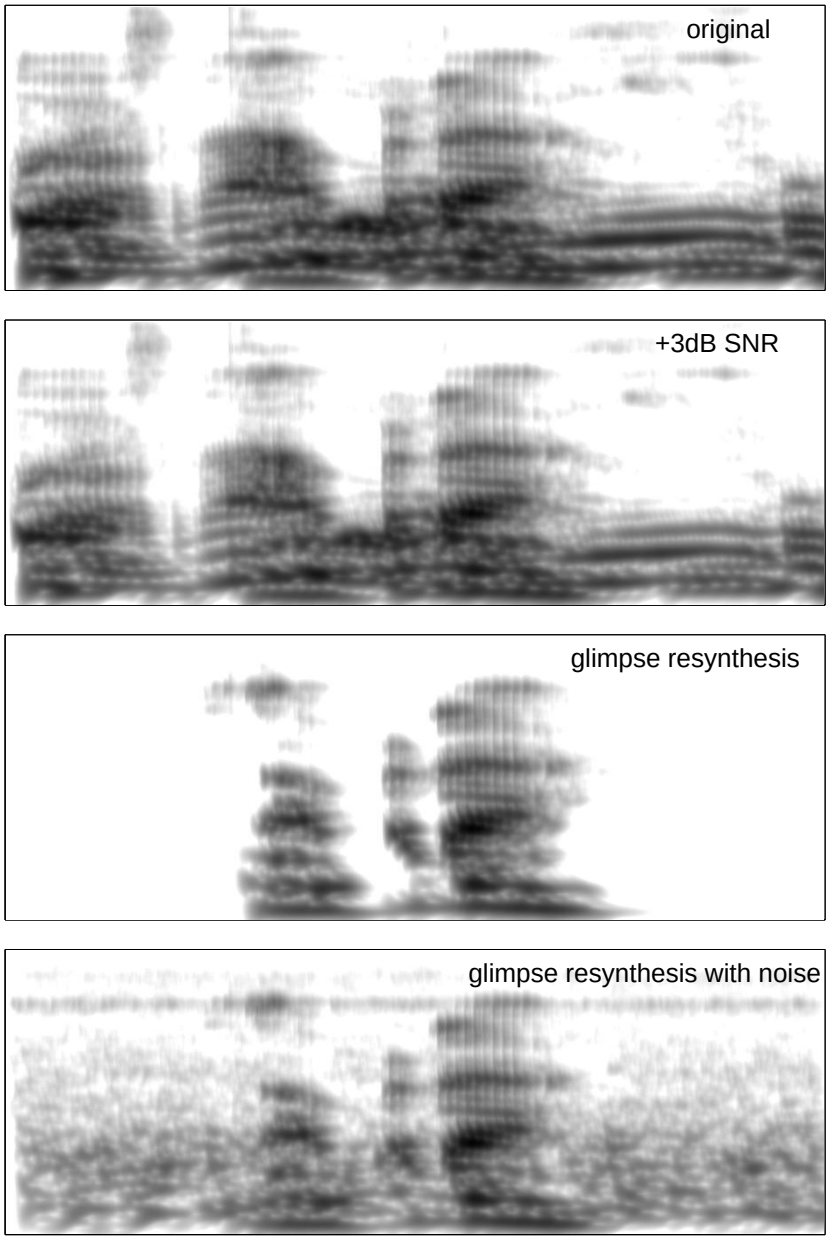


Figure 5.1: *Auditory spectrograms showing the original speech-in-noise token (target word "habrá", majority confusion "acostumbrar") and some of the experimental manipulations described in the text.*

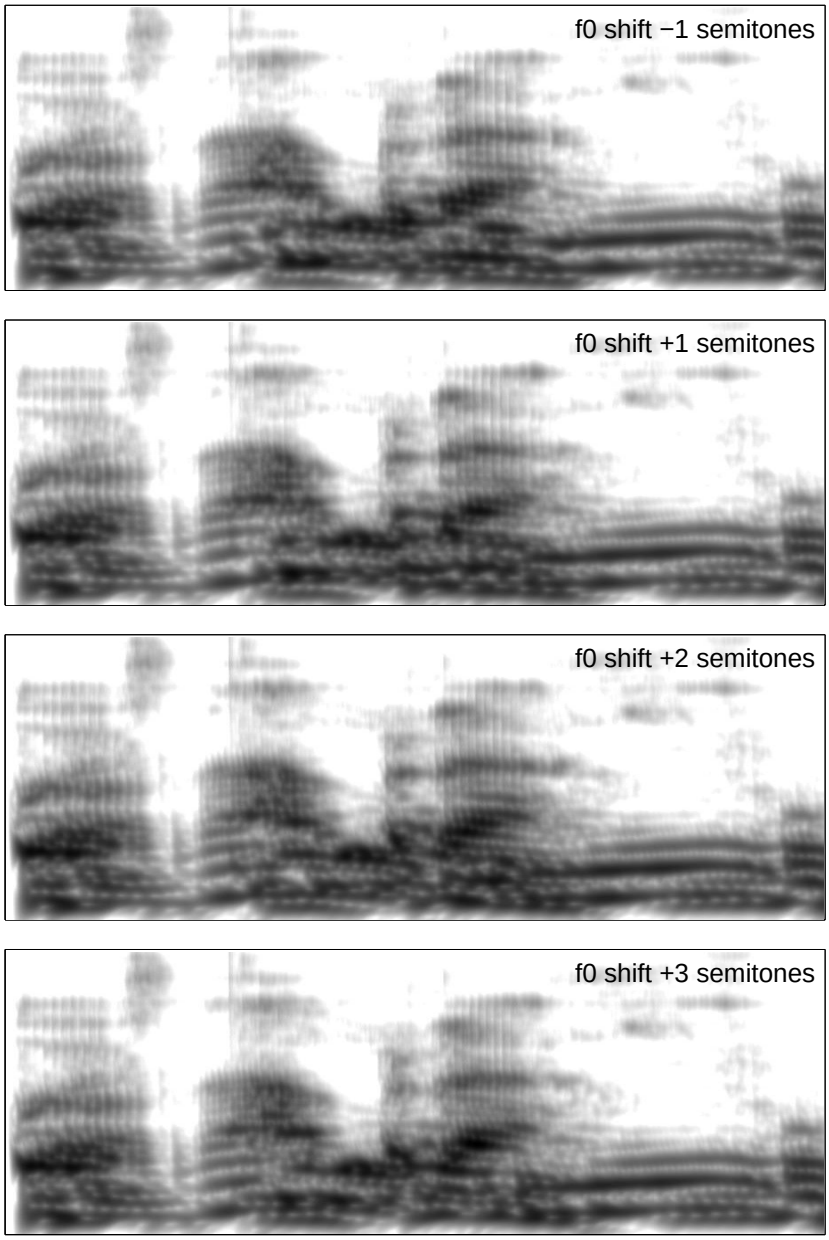


Figure 5.2: *Auditory spectrograms showing the F0 manipulations (target word "habrá", majority confusion "acostumbrar")*

which undo the effects of the masker interference, we tried to avoid selecting a subset where the majority of confusions originated in acoustic similarity. In order to achieve this, confusions were selected randomly after excluding those cases which differed from the target in the insertion, deletion or substitution of a single phoneme, as such cases had a higher probability of being caused by acoustic similarity, especially in an inflected language like Spanish (e.g., gender: "guapa/guapo"; number: "casa/casas"; person/tense: "veré/verá"). Tokens selected for the current experiment were balanced across the four talkers and five masker types. Tokens were presented in the 8 conditions listed in Table 5.1 based on the manipulations described in Section 5.2.

5.3.2 Listeners

72 monolingual Spanish or bilingual Spanish-Basque adults (age: $\mu = 26$ $\sigma = 4.6$) took part in the experiment after screening for hearing loss at 20 dB HL. Participants gave written consent and were paid for their participation.

5.3.3 Procedure

Of the 6400 unique stimuli (800 tokens x 8 manipulations), each listener screened 1600 stimuli in total in two 1 hr sessions, separated by a break of at least an hour. The 3 dB increase, control and two resynthesis conditions were screened in the first session, and those involving F0 shifts in the second session. The experiment was conducted using custom MATLAB software in a sound-attenuated studio booth over Sennheiser HD 380 Pro headphones. Listeners were instructed to identify a single word after hearing each stimulus exactly once and to type in their first impression. Stimuli were blocked by target talker and masker type, resulting in 20 blocks of 40 stimuli in each session. Prior to each block listeners heard four practice stimuli at a high SNR to familiarise themselves with the voice of the target talker and masker type (for the conditions where the masker was present). Block order was randomised first on speaker followed by masker, so that blocks of the same target speaker were presented successively, in order to minimise the switching between target talkers. The order of stimulus presentation in each block was randomised. Each individual stimulus (i.e. token-condition combination) was

heard by at least 15 listeners. Except for the SNR increase condition, presentation level SNRs were maintained for the remaining modifications. For the F0 shift conditions, the presentation level SNR was set after the F0 modification was applied to the target voice. In the glimpse resynthesis condition, the glimpse threshold was set to 0 dB. Consequently, regions with a positive local SNR were considered glimpses.

5.4 Modification conditions

5.4.1 Results

5.4.1.1 Test-retest rate

The majority confusion in the unmodified condition matched the majority response in the original experiment in 636 cases (79.5%) of the sub-corpus used in this study. To ensure that the subsequent analyses are based on highly-robust confusions, we additionally insisted upon a minimum listener agreement of 40% as in Chapter 2, which reduced the number of tokens to 505 (63%). The remaining analyses are based on this subset. We employed the following terminology to describe the relationship between the original confusion and the majority response elicited by the modified stimulus: listeners either MAINTAIN the original confusion, REVERT to the correct target word, or produce OTHER responses.

5.4.1.2 SNR increase

Following a 3 dB increase in the target relative to the masker, listeners MAINTAIN the original confusion in 339 cases (67.1%), REVERT to the correct target in 127 (25.2%) cases, and produce OTHER responses to the remaining 39 (7.7%) tokens. Figure 5.3 shows the breakdown of these responses across masker type. It is evident that the largest proportion of reversions to the correct target word occurred for the SSN (36%) and BMN3 (33%) maskers. The response categories differed significantly across masker type for the SNR increase condition [$\chi^2(8, N = 505) = 28.98, p < .001$]. In a series of follow-up χ^2 tests, we found that after applying Bonferroni adjustment to the significance criteria, differences in response

categories were significant between BMN1 and SSN [$\chi^2(8, N = 219) = 20.30, p < .001$], as well as BMN1 and BMN3 [$\chi^2(8, N = 217) = 16.43, p < .001$], while all other contrasts failed to reach significance at the .005 level.

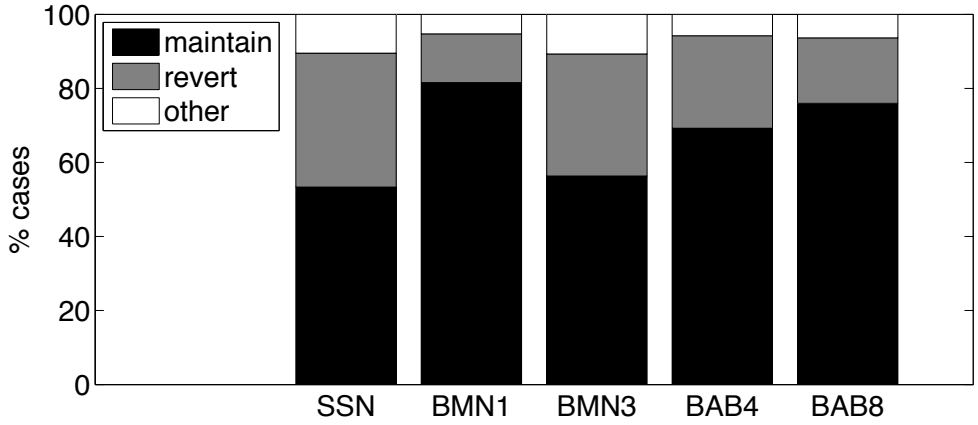


Figure 5.3: *Percentages of* MAINTAIN, REVERT *and* OTHER *responses per masker type for the SNR increase condition.*

5.4.1.3 Glimpse resynthesis

Figure 5.4 shows the distribution of MAINTAIN, REVERT, and OTHER responses following glimpse resynthesis with and without low-level noise, as a function of masker type. Here we see a striking difference between pure energetic maskers (SSN, BMN1, BMN3) and those which contain speech, and hence also have an informational masking component (BAB4, BAB8). The former group had a larger proportion of cases where the original confusion is maintained, while babble-based maskers lead to many REVERT cases. In order to determine the significant associations between resynthesis condition, masker and response type, a hierarchical log-linear analysis [Agresti, 2006] was conducted. A backward elimination procedure was used to select the best model. The model was assessed with the likelihood ratio chi-square test, which tests the difference between the observed counts and those predicted by the model. Thus, non-significant p values are associated with good models. The best model [$G^2(12) = 4.27, p = .98$] included significant interactions between masker and response type [partial $\chi^2(8) = 212.00, p < .001$], as well as response type and resynthesis condition [partial

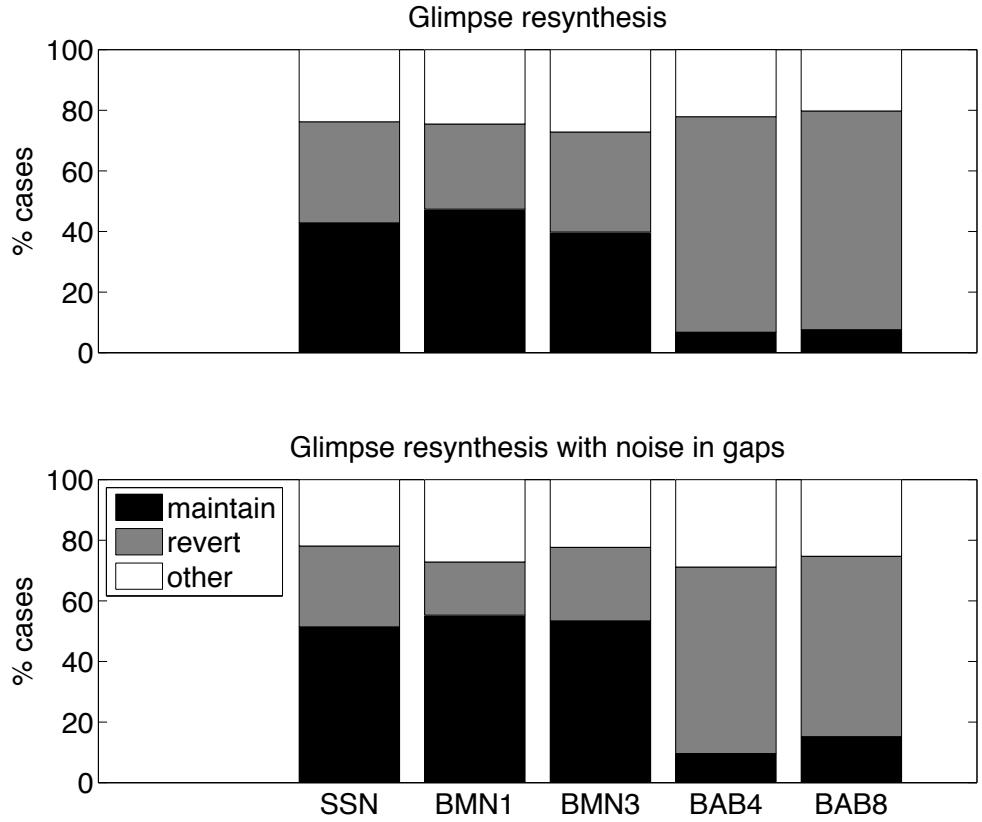


Figure 5.4: *Percentages of* MAINTAIN, REVERT *and* OTHER *responses per masker type for the glimpse resynthesis conditions.*

$\chi^2(2) = 10.61, p < .01$] and the corresponding main effects, out of which masker [partial $\chi^2(4) = 14.15, p < .01$] and response type [partial $\chi^2(2) = 44.03, p < .001$] were significant while resynthesis condition was not [partial $\chi^2(1) = 0, p = 1$]. The former significant interaction supported the differences of response categories across masker type mentioned above. The latter indicated that the distribution of responses are significantly different for the two resynthesis conditions, possibly since the conditions with noise in the gaps seemed to contribute slightly more MAINTAIN responses. However, a Cramer's V value of 0.1 confirms the visual impression, that the effect is indeed very small.

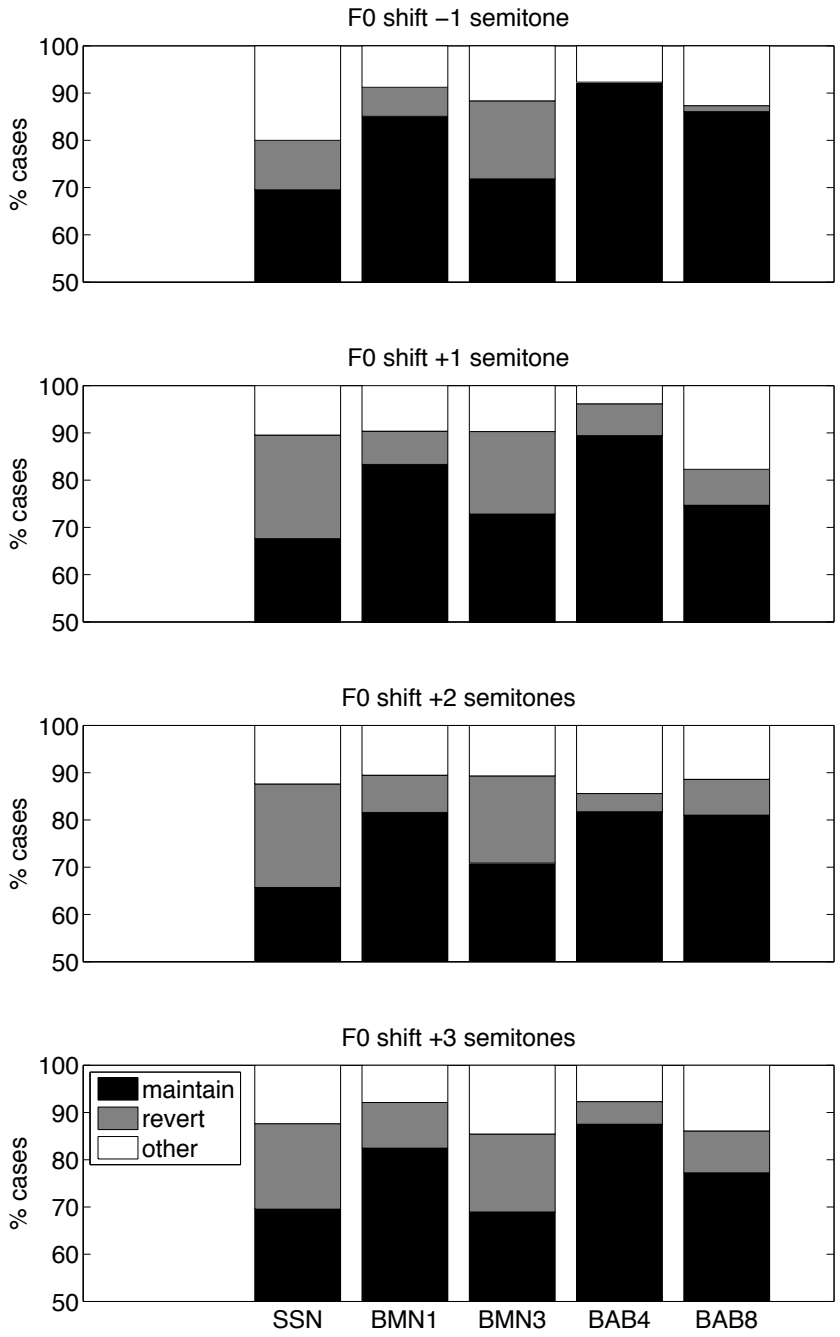


Figure 5.5: *Distribution of responses as a function of F0 shift. Note the change in axis range.*

5.4.1.4 F0 shifts

Figure 5.5 shows the responses for the F0 shift cases. Overall, just over half of the 505 robust tokens (274; 54%) were unaffected by shifts in F0. In the remaining 231 cases at least one of the shifts had an effect. We used a log-linear analysis as for the resynthesis conditions to determine associations of masker, response type and the amount of F0 shift. The best model [$G^2(45) = 38.11, p = .76$] included a significant interaction between masker and response type [partial $\chi^2(8) = 90.65, p < .001$] and the corresponding main effects for masker [partial $\chi^2(4) = 28.29, p < .001$] and response type [partial $\chi^2(2) = 1691.45, p < .001$]. The factor F0 shift was not included in the best fitting model, indicating no significant main or interactive effects. These results show that the number of MAINTAIN, REVERT and OTHER responses did not differ significantly as a function of F0 shift. However, the differences in responses across masker type as shown in Figure 5.5 were found significant with the largest numbers of REVERT responses seen for the SSN and BMN3 maskers. Intriguingly, this is the same pattern as observed in the SNR increase case.

5.4.2 Interim discussion

In Section 5.2 we hypothesised that the condition involving a +3 dB increase in SNR would primarily result in a release from energetic masking. This hypothesis is supported by the fact that the largest release from masking in this condition was observed for the two of the three noise-based maskers (SSN and BMN3) with the least amount of modulation. This increase in SNR was probably sufficient to uncover target glimpses critical to the comprehension of the utterance for the cases where listeners were able to recover the target. On the other hand, BMN1 showed the smallest proportion of REVERT cases in this condition, possibly since a 3 dB increase in SNR is not enough to bridge the gap between noise and speech energy in the masked regions caused by the large temporal modulations of speech modulated noise. This is in accordance with Eisenberg et al. [1995], who reported a steeper increase in keyword recognition performance in sentences with increasing SNR for the steady state masker compared to amplitude modulated noise in a ±2 dB range around the SRT value. Coincidentally, in their study these ranges

corresponded to -8 to -4 dB for the steady state masker and -12 to -8 dB for the amplitude modulated masker, which roughly match the SNR ranges used in our original experiment for these two maskers. At the same time, the SNR increase condition also resulted in a considerable amount of REVERT cases for speech-based maskers as well. It is unclear whether these cases can be attributed to a release from energetic masking, as the level differences between the target speech and background talkers could provide listeners with additional segregation cues. Brungart [2001] found that level differences between target and masker voices can support identification, even when the target is presented at a lower speech level.

The glimpse resynthesis condition provided a clearer separation between informational and energetic masking effects. We hypothesised that for the resynthesis conditions, REVERT and MAINTAIN responses correspond to confusions caused by informational and energetic masking respectively. This hypothesis was supported by the distribution of response types across maskers, as shown in Section 5.4.1.3. The release from masking provided by this modification was significantly larger for babble maskers relative to maskers constructed from speech-based noise. Consistent with the literature [Brungart, 2001; Brungart et al., 2006; Freyman et al., 2004; Roman et al., 2003], our results in the glimpse reynthesis condition suggested that the dominant form of masking is informational in a multi-talker scenario. Specifically, these findings are in line with Brungart et al. [2006], who found near perfect recognition scores for speech masked with up to 4 competing talkers, after applying a similar resynthesis with a 0 dB glimpse threshold. While our results indicate that a small proportion of confusions in babble were caused by energetic masking, Brungart et al. [2006] noted that due to the restricted response alternatives in their study the effects of energetic masking might have been somewhat underestimated.

The glimpse resynthesis condition also resulted in an appreciable proportion of REVERT cases in noise-based maskers. Since in the resynthesis conditions listeners had access to the exact same target glimpses as in the control condition — the only difference being the presence or the absence of the masker in the unglimpsed regions — these confusions cannot be explained by simultaneous energetic masking. One possibility is that these cases were originally caused by forward masking [Oxenham, 2001; Plack and Oxenham, 1998] — the saturation of inner hair cells

following the onset of nerve firing, causing the inner hair cells to be unable to respond to stimuli for a short interval after masker offset — which is absent in the resynthesis condition. Another explanation is that with the noise removed, listeners experience less cognitive load resulting in better identification. A similar result was reported in Brungart et al. [2006], who found that, though substantially less beneficial compared to the multi-talker condition, glimpse resynthesis still provided a masking release of 2-5 dB in a steady-state and modulated noise masker condition.

One of the resynthesis conditions involved adding a low-level noise in the spectro-temporal regions where the target was not glimpsing, in order to mask the artefacts caused by the discontinuities resulting from resynthesising speech in disjoint target glimpses. By eliminating these artefacts, we expected this condition to result in more REVERT cases. Instead, even the low-level masking noise (12 dB SNR) added in the time-frequency gaps resulted in a slight increase in MAINTAIN responses for all maskers. As this masker could not have impacted the target glimpses directly, again, the slight increase in MAINTAIN responses could be explained by forward masking or increased cognitive load due to the presence of low-level noise. Another explanation is that the presence of even a low-level noise is required to perceive the original confusions because at least in part, it was influenced by the phonemic restoration effect [Warren et al., 1970].

Among the modifications considered, F0 manipulations had the smallest effect on the confusions, in agreement with Cooke [2009] who also found that F0 changes had little impact on the original percept. Contrary to expectations, we did not find evidence that F0 manipulations provide more release from informational masking compared to energetic masking. Instead, F0 modifications resulted in the most REVERT cases for maskers SSN and BMN3. As forward masking exhibits a sharper tuning curve than simultaneous masking [Moore and Glasberg, 1986], it is possible that shifts in F0 resulted in a release from forward masking for some of these cases.

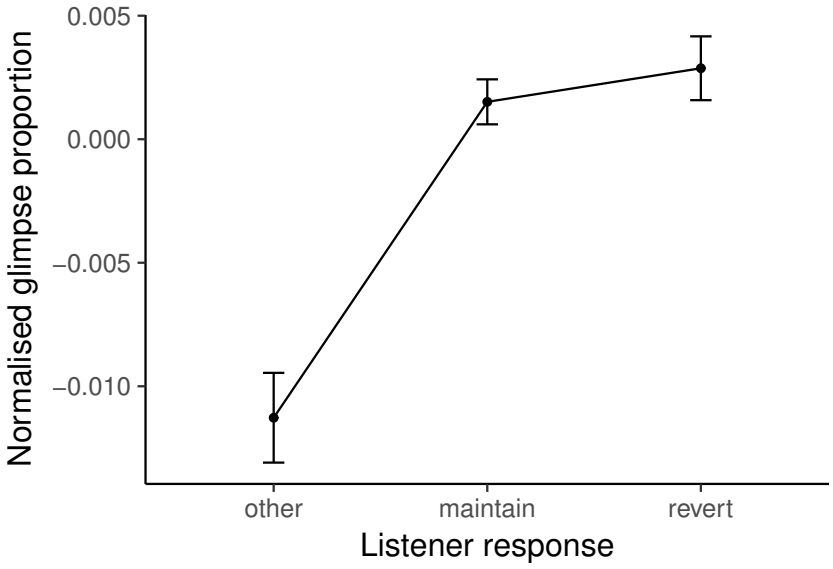


Figure 5.6: *Normalised glimpse proportion for response types pooled across conditions and maskers. Error bars indicate standard error of the mean.*

5.5 Glimpse proportion and word length differences

5.5.1 Results

5.5.1.1 Glimpse proportion

We speculated that the amount of spectro-temporal plane glimpsing or the target glimpse proportion (GP), is a good predictor of the response categories observed. Since average GP is highly variable across modifications (e.g. SNR increase results in a higher GP compared to the other conditions) and masker type, we subtract the mean GP calculated for each masker and condition from the GP of each corresponding stimulus. This modified GP-metric differed significantly between the three response categories [$F(2, 3532) = 22.06, p < .001$], a post hoc Tukey test revealing it to be significantly lower [$p < .001$] for OTHER responses then for REVERT and MAINTAIN cases, which did not differ significantly.

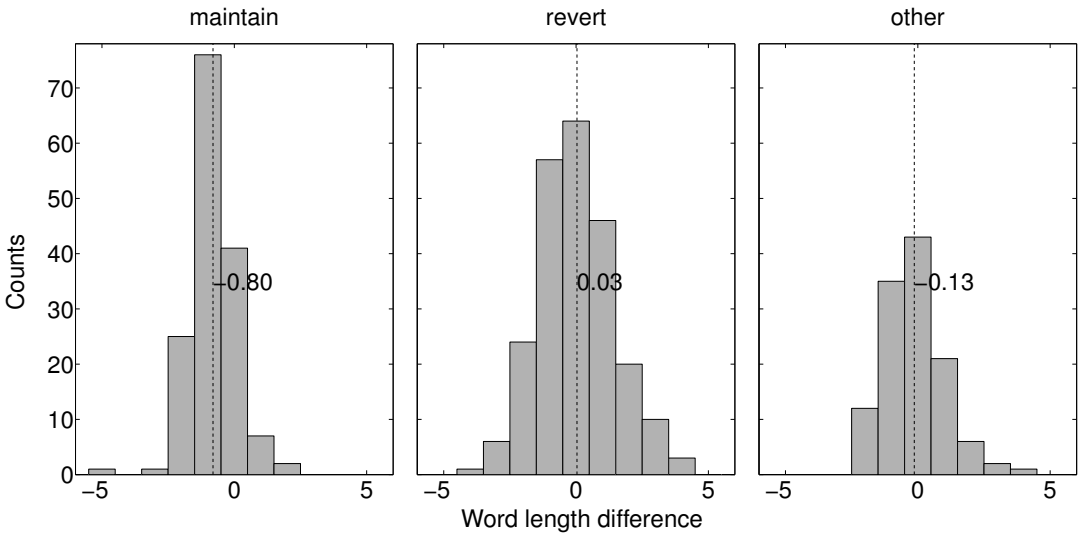


Figure 5.7: *Distribution of word length difference measured in number of phones between confusion and target for* MAINTAIN, REVERT *and* OTHER *cases in the glimpse resynthesis condition. The vertical dotted line shows the sample mean.*

5.5.1.2 Target-confusion length difference

While we expected energetic masking to primarily result in the deletion of target speech fragments, as shown in the previous chapter, maskers with informational content can both delete and contribute phonetic material to the listener's percept. It is possible that this asymmetry appears in the resulting misperceptions. Thus, we hypothesised that confusions stemming from energetic masking would, on average, be shorter then their respective targets, while confusions stemming from informational masking are expected to have a more centred distribution of word length difference. Using the glimpse resynthesis condition, we separated confusions caused by energetic (MAINTAIN) and informational (REVERT) masking based on listeners' responses and evaluated the word length difference in phones between the transcriptions of the target word and the misperception.

Figure 5.7 shows that the distribution of confusion-target word length difference for MAINTAIN and REVERT cases largely matches our predictions: misperceptions caused by energetic masking tend to be shorter than the corresponding target word $[t(152) = -10.91, p < .001]$ while for misperceptions caused by informational masking we can make no such claim $[t(231) = 0.344, p = .70]$.

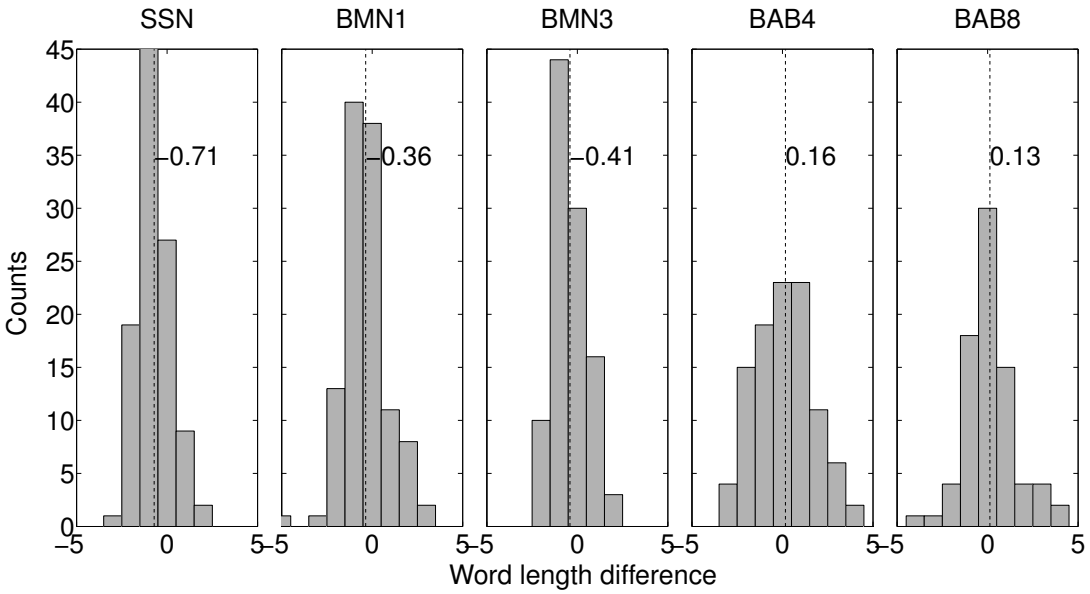


Figure 5.8: *Distribution of word length difference between confusion and target across masker type. The vertical dotted line indicates the sample mean.*

Figure 5.8 plots the distributions of confusion-target word length difference across masker type. For noise-based maskers, which are expected to result in mostly energetic masking, confusions are significantly shorter than targets as shown by a one sample t-test. $[t_{SSN}(104) = -7.64, p < .001; t_{BMN1}(113) = -3.11, p < .001; t_{BMN3}(102) = -4.29, p < .001]$. For speech-based maskers, which primarily result in informational masking, distributions are again centred around zero $[t_{BAB4}(104) = 0.95, p = .34; t_{BAB8}(113) = 0.79, p = .42]$.

Finally Figure 5.9 shows the distribution of phoneme length difference across the responses elicited in the glimpse resynthesis condition and masker type.

5.5.2 Interim discussion

We hypothesised that the proportion of spectro-temporal plane glimpsing normalised on the given condition and masker type would be a good predictor of listeners' responses. Interestingly, we found that the normalised glimpse proportion of REVERT and MAINTAIN cases did not differ significantly, and had a higher than average normalised GP value. On the other hand, cases in the OTHER

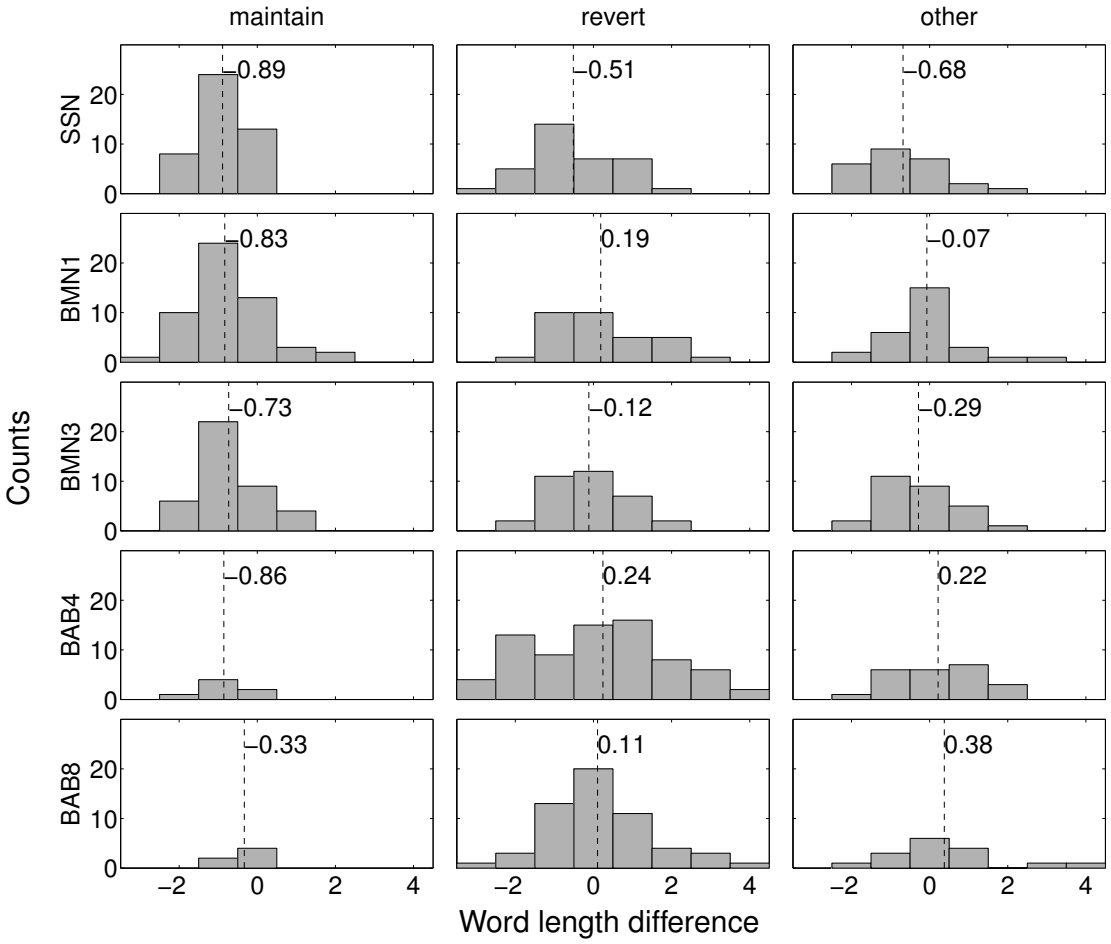


Figure 5.9: *Distribution of word length difference measured in number of phones between confusion and target across* MAINTAIN *,* REVERT *and* OTHER *cases in the glimpse resynthesis condition and masker type. The vertical dotted line indicates the sample mean.*

category had a GP value significantly lower than average.

It is clear why REVERT cases, where listeners succeed in recovering the target message corresponded to high normalised GP values. The fact that the normalised GP values of MAINTAIN responses were similarly high is more intriguing. One possibility is that consistent confusions also require a certain amount of salient target fragments to support a common percept across listeners, even though the identified target fragments might be reinterpreted as another word or

blended with masker material. The lower GP could result in higher uncertainty and consequently more guessing on the part of the listeners, resulting in OTHER responses, which in most cases were not consistent. This claim is supported by the fact that the three response types also differed significantly in consistency $[F(2, 3532) = 217.26, p < .001]$, with OTHER responses having significantly lower consistency $[p < .001]$ then REVERT and MAINTAIN responses, which did not have any significant differences between them.

While energetic masking resulted in the deletion of target speech, misallocating fragments from background talkers can add phonetic material to the percept. Consequently, we hypothesised that confusions stemming from energetic maskers would be shorter then targets, while informational maskers can both contribute or eliminate material, so we expected distributions of word-length difference to be more centred. Our hypothesis was confirmed both regarding the distribution of the word length difference for the response alternatives in the glimpse resynthesis condition, as well as across maskers. Shorter confusions relative to targets were observed for the MAINTAIN responses in the glimpse resynthesis condition corresponding to the energetic masking cases.

In the glimpse resynthesis condition, MAINTAIN responses corresponding to the energetic masking cases resulted in shorter confusions compared to targets. While speech-based maskers tend to have a zero centred word-length difference distribution, for the few MAINTAIN cases they do contribute the confusions are shorter then targets $[t(12) = -3.41, p = .005]$, providing further evidence that these cases stem from energetic masking. Likewise, REVERT cases stemming from noise based maskers — with the exception of SSN — result in a centred distribution.

5.6 General discussion

Multiple studies in the past have tried to identify the perceptual cues relevant for the identification of a particular sound, by introducing systematic modifications to the stimuli and subsequently reevaluating listeners' percepts [Cooke, 2009; Li et al., 2010; Varnet, 2013]. However, most of these studies are limited in scope to nonsense syllable perception. Following the work of Cooke [2009], in

this study, we aimed to identify the role energetic and informational masking played in generating confusions using this approach. Stimuli were presented in four types of conditions: a control condition, as well as conditions involving an increase in SNR, glimpse resynthesis and modification of target fundamental frequency. Modifications were selected to provide release either from energetic or informational masking, in order to separate confusions based on the type of masker interference causing them.

We found that for almost 80% of tokens the majority listener response matched the consistent confusion reported in the original experiment and 60% of the cases had at least the same threshold for consistency (40%). The fact that the majority of consistent confusions could be replicated across different listener groups provides further evidence that these cases are characteristic of general speech processing mechanisms and the given listener population.

The glimpse resynthesis condition caused most confusions to revert in speech-based maskers indicating that the vast majority of these misperceptions are caused by informational masking. These findings are in line with the results reported by Brungart et al. [2006], who found the effects of energetic masking to be minimal in speech-on-speech masking using a similar glimpse resynthesis technique. At the same time, the resynthesis condition also generated a substantial proportion of reverts in the noise masking conditions, suggesting that these confusions result from non-simultaneous masking. One possible explanation is that these confusions originate from forward masking. The increase in cognitive load associated with the presence of noise could also contribute to these confusions.

Surprisingly, MAINTAIN and REVERT responses across maskers and conditions had similar normalised glimpse proportions, as we would have expected REVERT cases to score highest on this metric. However, OTHER cases scored lower compared to the two groups above. It seems possible that OTHER cases correspond to a higher degree of listener uncertainty, potentially resulting from the lower proportion of available target glimpses. This hypothesis is further supported by the fact that in general, OTHER responses also showed lower consistency compared to MAINTAIN and REVERT cases.

Finally, the processes through which energetic and informational masking interfere with the target utterance are reflected in confusion word-forms. Energetic

masking cases primarily resulted in the deletion of phonetic material in the target, as evidenced by the word length difference between target and confusion. At the same time, informational masking can also contribute material which resulted in a more centred distribution for the same metric.

One unexpected finding of the above analysis was the relatively small impact F0 modifications had on the resulting misperceptions. This is in contrast with several previous studies, which have found that F0 differences between target and background voices improved intelligibility of the target, both regarding the recognition of isolated vowels [Assmann and Summerfield, 1990; Culling and Darwin, 1993; Scheffers, 1983] and in terms of sentences [Brokx and Nooteboom, 1982; Darwin and Hukin, 2000]. A possible explanation for this discrepancy is that F0 differences help listeners in tracking the source of interest over time, as in sentence recognition context, making it easier for them to sustain their attention on the target stream. On the other hand, our results suggests that differences in F0 do not interfere with grouping smaller speech fragments from the masker to the target. In line with our findings, in his original study of English misperceptions, Cooke [2009] also found little effect of F0 changes overall.

Apart from the F0 modification condition — which had the smallest impact on confusions — for the remaining conditions, signal modifications were not introduced incrementally. Future work in this direction could employ gradual shifts, which would allow us to witness the evolution of the percept on a fine-grained level. While this approach has been applied to nonsense syllable confusions [Li et al., 2010], identifying the boundaries of perceptual categories for word-level misperceptions could further our understanding of the mechanisms through which misperceptions are formed.

Chapter 6

Conclusions

In this thesis, we presented the elicitation of a corpus of consistent word misperceptions and its analysis from a microscopic perspective. Consistent confusions are an interesting speech perception phenomenon which are characteristic misperceptions of a listener population, and can serve as valuable diagnostic stimuli for microscopic intelligibility models. Most existing research on microscopic speech intelligibility modelling has focused exclusively on nonsense syllable confusions. Together with the study by Cooke [2009], this work represents one of the first attempts to explain word-level misperceptions using a microscopic modelling approach.

The elicitation of a large-scale corpus of consistent word misperceptions was presented in Chapter 2. 172 Spanish listeners participated in the elicitation experiment, responding to over 50 000 unique speech-in-noise tokens presented in an open-set task. Stimuli were constructed using four talkers and five distinct masker types mixed at SNR ranges favourable for confusion elicitation. Overall, 300 000 responses were collected. Online token pruning heuristics and a series of post-processing steps were applied to maximise the number of useful confusions. After applying an across-listener consistency criterion of 40%, the final corpus consisted of over 3200 robust word misperceptions. This experiment has demonstrated the feasibility of collecting a large-scale corpus of robust word misperceptions in a lab setting. The adaptive token pruning process was a key component of the collection process, allowing for a roughly three-fold increase in consistent confusion finding efficiency.

In Chapter 3 we examined the confusions collected in Chapter 2 from a signal-independent perspective. In this analysis, we explored how the characteristics of the target word and its constituent phones affected the confusion patterns observed. We also evaluated the overall effect of masker type on the resulting misperceptions. Past studies have proposed various methods to align the phonetic transcriptions of the target and misperceived word in order to identify sub-lexical confusion patterns. However, there are certain drawbacks associated with most of these approaches, including the potential for circularity or suboptimal alignments. To address these limitations, we proposed a novel alignment method based on syllable constituency and lexical stress, which takes advantage of the robustness of the stressed syllable nucleus, which has been well documented in the literature.

We investigated the effects of factors on perceptual confusion patterns across multiple levels of speech units, starting from articulatory features, through sub-lexical factors including word-position and stress to lexical characteristics. In line with prior experimental and naturalistic studies, we found that place cues were most vulnerable, followed by articulatory features manner and voicing. We found consonants with high-frequency energy to be especially robust, consistent with multiple laboratory collections, but contrary to naturalistic reports. Further examination suggested that the high-frequency advantage was mostly present for maskers constructed from speech-shaped noise, where these consonants stand out from the average masker level. In babble noise, high-frequency consonants were less accurately perceived, possibly since babble noise can contribute similar high-frequency consonant fragments which could result in a potential confusion. We confirmed the results of prior naturalistic studies regarding the robustness of the stressed syllables and the progressively decreasing segment error rates from word-initial to word-final position — after accounting for morphological effects. The agreement of these results across languages could indicate that these trends are not language specific, but are characteristic of common word recognition mechanisms. We found that words not contributing any confusions had higher word frequency, as well as a lower neighbourhood density and frequency, compared to words that did, providing support for the neighbourhood probability rule proposed by Luce and Pisoni [1998].

Finally, by investigating the effects of the masker type used for elicitation, we

found that confusions stemming from babble maskers had a higher proportion of insertion errors and a higher edit distance from the target compared to confusions stemming from noise-based maskers. These results provided a first indication that in maskers constructed from speech, listeners could be forming misperceptions by borrowing phonetic material from the background talkers.

In Chapter 4 we conducted a signal-dependent analysis of misperceptions from a microscopic modelling perspective. Using the glimpse decoder introduced by Barker et al. [2005], we conducted an automatic classification of confusions based on their cause. Confusions were classified as either stemming from acoustic similarity or originating from the speech-noise interaction. These latter cases were further classified into reinterpretations, blends and overrides, based on the amount of masker material incorporated into the confusion. The involvement of the masker could be measured as the proportion of spectro-temporal 'area' corresponding to the masker glimpses in the segregation hypothesis that best supported the misperceived word. Using this classification procedure we were able to explain 40% of the misperceptions in our corpus. In addition, we illustrated how misperceptions could be formed through the misallocation of low-level speech fragments.

In the second part of the chapter, we investigated the role misallocation plays in the generation of misperceptions in more detail. By supplying the decoder with the majority confusion reported by listeners, we could constrain the search to the segregation space, effectively force-aligning the glimpses to the precept reported by listeners. In order to quantify the amount of masker involvement for each confusion, we defined metrics target and masker proportion, corresponding to the area of target glimpses included in the most likely segregation over the total area of target fragments (likewise for masker proportion). Using these metrics, we quantified the amount of allocation errors involved in each misperception in speech-based maskers. Our findings suggested that most confusions in babble involved misallocations to some degree. Our study confirmed the notion that the misallocation of low-level speech fragments is a major component of informational masking.

In order to better understand which aspect of the target-masker interaction resulted in the misperception, in Chapter 5, we asked the question, how does the

confusion eliciting stimulus need to be modified for listeners to correctly identify the target? We introduced a series of signal modifications to a subset of the confusion eliciting tokens and presented them to listeners in a follow-up perceptual experiment. The selected modifications aimed to provide a release from either energetic or informational masking and involved a 3 dB increase in SNR, speech resynthesis in the target glimpses and shifting the F0 of the target word. The stimuli used in the original experiment were also presented to listeners as a control condition. We found that the majority confusion in the follow-up experiment matched the original confusion roughly 80% of the time. When applying the same consistency criterion (40%) as in the original experiment, we obtained a retest rate of 63%. These results suggested that consistent confusions can be replicated across different samples of the listener population. Listeners' responses to the modified stimuli were classified into three categories corresponding to maintaining the confusion in the original experiment, reverting to the target word or forming a new percept. As expected, the SNR increase condition resulted in a release primarily from energetic masking, with most revert cases occurring for maskers SSN and BMN3. BMN1 had the smallest proportion of reverts, possibly because of a 3 dB increase is insufficient to uncover masked regions due to the large temporal modulations of this masker. A substantial amount of revert cases were also observed for babble maskers in this condition, suggesting that an increase in SNR might also help the correct segregation of the target word. The glimpse resynthesis condition provided a clear separation of confusions caused by energetic and informational masking. This modification caused the majority of misperceptions stemming from speech based maskers to revert to the target, suggesting that these cases originated from informational masking. At the same time, this glimpse resynthesis also caused a substantial proportion of reverts for the noise-based maskers. This suggested that not all confusions stemming from noise-based maskers are caused by simultaneous masking. In line with the findings of Cooke [2009], the four conditions involving F0 modifications affected the smallest proportion of confusions. Shifting the fundamental frequency of the target did not cause a large proportion of reverts for the speech-based masker, suggesting that a mismatch in F0 is not prohibitive for listeners when grouping speech fragments to form a percept.

In this thesis, we presented a signal-dependent and signal-independent analysis of misperceptions as separate approaches. Further investigations could merge the two approaches, possibly by incorporating the signal-independent factors identified in Chapter 3 into a microscopic speech perception model.

In sum, speech misperceptions have the potential to help us understand the processes involved in human speech perception. Consistent misperceptions are especially helpful in this regard, eliminating the variability stemming from individual differences, which in turn, makes it easier to analyse confusion patterns at higher levels of speech units such as the word. Striving to create end-to-end models of auditory perception not only promises better understanding of the underlying processes, but also has the potential to revolutionise technologies related to speech and hearing. As long as listeners outperform machines in speech recognition, insights into human speech processing have the capacity to improve recognition rates. Further, a detailed understanding of the errors listeners make across adverse conditions will undoubtedly benefit hearing prosthetics. Although still in its infancy, microscopic modelling may well help us gain an increasingly accurate understanding of human speech perception in the future.

Appendix A

Examples from the confusions corpus

ID	Target	Confusion	TargetIPA	ConfusionIPA	Masker	Cons.(%)
1181	cruza	bruja	! k ɾ u . θ a	! b ɾ u . x a	BAB4	40
1189	suelta	suelo	! s we l . t a	! s we . l o	BAB4	40
1214	mamá	escama	m a ! m a	e s ! k a . m a	BAB4	53
1269	pájaros	paja	! p a . x a . ɾ o s	! p a . x a	BAB4	53
1275	cómodo	como	! k o . m o . ð o	! k o . m o	SSN	47
1277	escuchen	escucha	e s ! k u . ʧ e n	e s ! k u . ʧ a	SSN	87
1288	harás	gas	a ! ɾ a s	! g a s	SSN	40
1291	visita	caros	b i ! s i . t a	! k a . ɾ o s	BAB4	60
1305	brilla	envidia	! b ɾ i . ʝ a	e m ! b i . ð ja	BAB4	47
1309	muelle	traje	! m we . ʝ e	! t ɾ a . x e	BAB4	40
1319	vino	chicos	! b i . n o	! ʧ i . k o s	BAB4	60
1331	pesos	besos	! p e . s o s	! b e . s o s	SSN	93
1340	hacemos	acentos	a ! θ e . m o s	a ! θ e n . t o s	SSN	47
1344	vaya	baño	! b a . ʝ a	! b a . ɲ o	SSN	47
1351	juré	puré	x u ! ɾ e	p u ! ɾ e	SSN	60
1360	corto	corta	! k o ɾ . t o	! k o ɾ . t a	SSN	47
1382	plumas	lunes	! p l u . m a s	! l u . n e s	SSN	47
1401	verán	verano	b e ! ɾ a n	b e ! ɾ a . n o	SSN	60
1419	sabían	sabía	s a ! β i . a n	s a ! β i . a	SSN	87
1439	debajo	llevar	d e ! β a . x o	ʝ e ! β a ɾ	SSN	40
1447	doblar	burlar	d o ! β l a ɾ	b u ɾ ! l a ɾ	SSN	67
1448	querían	querías	k e ! ɾ i . a n	k e ! ɾ i . a s	BAB8	47
1452	paré	pared	p a ! ɾ e	p a ! ɾ e ð	SSN	47
1460	raro	rara	! r a . ɾ o	! r a . ɾ a	BAB8	40
1477	estudio	estudios	e s ! t u . ð jo	e s ! t u . ð jo s	BAB8	60
1496	falla	fallo	! f a . ʝ a	! f a . ʝ o	SSN	47
1519	cuántas	cuatro	! k wa n . t a s	! k wa . t ɾ o	SSN	53
1526	primer	primero	p ɾ i ! m e ɾ	p ɾ i ! m e . ɾ o	SSN	60
1531	drogas	flores	! d ɾ o . ɣ a s	! f l o . ɾ e s	SSN	47
1539	viví	vivir	b i ! β i	b i ! β i ɾ	BMN1	47
1542	sabría	sabía	s a ! β ɾ i . a	s a ! β i . a	SSN	47
1564	nacido	nativo	n a ! θ i . ð o	n a ! t i . β o	BMN1	40
1576	ganado	ganar	g a ! n a . ð o	g a ! n a ɾ	BMN1	47
1583	obtener	contener	o β . t e ! n e ɾ	k o n . t e ! n e ɾ	BMN3	60
1589	buques	bunque	! b u . k e s	! b u ŋ . k e	BMN1	73
1595	viví	vivir	b i ! β i	b i ! β i ɾ	BMN3	67
1596	contento	intento	k o n ! t e n . t o	i n ! t e n . t o	BMN1	40
1597	privado	primero	p ɾ i ! β a . ð o	p ɾ i ! m e . ɾ o	BMN1	47
1604	pienses	piensas	! p je n . s e s	! p je n . s a s	BMN1	80
1617	entera	entero	e n ! t e . ɾ a	e n ! t e . ɾ o	BMN1	67
1623	mueven	mueve	! m we . β e n	! m we . β e	BMN1	40

Table 1: *Examples of the Spanish confusions corpus. Columns correspond to confusion ID number, orthographic and phonetic transcriptions of the target and misperceived word, as well as the percentage of listeners who reported the majority confusion.*

ID	Target	Confusion	TargetIPA	ConfusionIPA	Masker	Cons.(%)
14588	si	sigo	! s i	! s i . ɣ o	BAB4	53
14589	casi	casa	! k a . s i	! k a . s a	BAB4	73
14606	hermanos	hermano	e ɾ ! m a . n o s	e ɾ ! m a . n o	BMN3	73
14613	casada	casarse	k a ! s a . ð a	k a ! s a ɾ . s e	BAB4	40
14619	salid	salir	s a ! l i ð	s a ! l i ɾ	BMN1	67
14623	honrar	comprar	o n ! r a ɾ	k o m ! p ɾ a ɾ	BMN1	47
14626	hacías	hacía	a ! θ i . a s	a ! θ i . a	BAB4	60
14630	besa	pesa	! b e . s a	! p e . s a	BAB4	53
14644	jura	cura	! x u . ɾ a	! k u . ɾ a	BAB4	60
14654	vista	pista	! b i s . t a	! p i s . t a	BAB4	73
14695	honrar	comprar	o n ! r a ɾ	k o m ! p ɾ a ɾ	SSN	53
14696	ida	vida	! i . ð a	! b i . ð a	SSN	47
14710	preciso	piso	p ɾ e ! θ i . s o	! p i . s o	SSN	53
14715	gordos	cortos	! g o ɾ . ð o s	! k o ɾ . t o s	SSN	53
14724	local	tocar	l o ! k a l	t o ! k a ɾ	BMN3	67
14758	juntas	puntas	! x u n . t a s	! p u n . t a s	SSN	53
14766	dado	cantar	! d a . ð o	k a n ! t a ɾ	BMN1	47
14771	vulgar	buscar	b u l ! ɣ a ɾ	b u s ! k a ɾ	BAB8	53
14789	bajas	bajos	! b a . x a s	! b a . x o s	BAB8	47
14839	jura	cura	! x u . ɾ a	! k u . ɾ a	SSN	47
14848	humana	humano	u ! m a . n a	u ! m a . n o	SSN	60
14871	desnudo	desnuda	d e z ! n u . ð o	d e z ! n u . ð a	BAB8	40
14889	gatos	datos	! g a . t o s	! d a . t o s	BAB4	60
14913	loca	locas	! l o . k a	! l o . k a s	BAB4	47
14917	tía	magia	! t i . a	! m a . x ja	BAB4	40
14924	viva	manzana	! b i . β a	m a n ! θ a . n a	BAB4	60
14926	disparen	dispare	d i s ! p a . ɾ e n	d i s ! p a . ɾ e	BMN1	93
14932	cortan	corta	! k o ɾ . t a n	! k o ɾ . t a	BMN1	60
14940	fumo	humo	! f u . m o	! u . m o	BAB4	80
14957	marina	gallina	m a ! ɾ i . n a	g a ! ʝ i . n a	BMN3	47
14959	enseñar	siempre	e n . s e ! ɲ a ɾ	! s je m . p ɾ e	BMN3	40
14983	sordo	solo	! s o ɾ . ð o	! s o . l o	BMN1	60
15011	tráfico	trágico	! t ɾ a . f i . k o	! t ɾ a . x i . k o	BMN3	60
15016	fieles	quieres	! f je . l e s	! k je . ɾ e s	BMN3	47
15020	buen	bueno	! b we n	! b we . n o	BMN3	80
15025	honor	color	o ! n o ɾ	k o ! l o ɾ	BMN3	40
15029	usado	usar	u ! s a . ð o	u ! s a ɾ	BMN3	87
15098	danza	lanza	! d a n . θ a	! l a n . θ a	BMN1	40
15117	rancho	gancho	! r a ɲ . ʧ o	! g a ɲ . ʧ o	SSN	40
15131	trabaja	trabajo	t ɾ a ! β a . x a	t ɾ a ! β a . x o	SSN	60
15148	leyes	celos	! l e . ʝ e s	! θ e . l o s	BAB4	67

Table 2: *Examples of the Spanish confusions corpus. Columns correspond to confusion ID number, orthographic and phonetic transcriptions of the target and misperceived word, as well as the percentage of listeners who reported the majority confusion.*

ID	Target	Confusion	TargetIPA	ConfusionIPA	Masker	Cons.(%)
45071	huelen	vuelve	! we . l e n	! b we l . β e	BMN1	67
45110	clara	claro	! k l a . ɾ a	! k l a . ɾ o	BAB8	47
45117	bajan	baja	! b a . x a n	! b a . x a	BMN3	47
45133	coger	mujer	k o ! x e ɾ	m u ! x e ɾ	BMN3	53
45135	escuche	escucha	e s ! k u . ʧ e	e s ! k u . ʧ a	BMN3	60
45138	nuclear	leal	n u . k l e ! a ɾ	l e ! a l	BMN3	47
45225	peón	león	p e ! o n	l e ! o n	BMN1	73
45226	tiendas	piernas	! t je n . d a s	! p je ɾ . n a s	BMN1	40
45231	cuestión	gestión	k we s ! t jo n	x e s ! t jo n	BMN3	53
45250	dejado	dejar	d e ! x a . ð o	d e ! x a ɾ	BMN3	47
45256	robo	ropa	! r o . β o	! r o . p a	BMN3	47
45260	parecen	pared	p a ! ɾ e . θ e n	p a ! ɾ e ð	BMN3	47
45263	vosotros	nosotros	b o ! s o . t ɾ o s	n o ! s o . t ɾ o s	BMN3	100
45266	talla	calla	! t a . ʝ a	! k a . ʝ a	BAB4	60
45274	dejad	dejar	d e ! x a ð	d e ! x a ɾ	BAB4	53
45320	murieron	morir	m u ! ɾ je . ɾ o n	m o ! ɾ i ɾ	BMN3	40
45334	preciosa	precioso	p ɾ e ! θ jo . s a	p ɾ e ! θ jo . s o	BAB8	40
45347	sede	ser	! s e . ð e	! s e ɾ	SSN	73
45356	importan	importa	i m ! p o ɾ . t a n	i m ! p o ɾ . t a	SSN	47
45362	nuestra	nuestro	! n we s . t ɾ a	! n we s . t ɾ o	SSN	73
45400	sueña	sueños	! s we . ɲ a	! s we . ɲ o s	BAB4	80
45428	hablaba	hablar	a ! β l a . β a	a ! β l a ɾ	BMN3	40
45444	ir	pis	! i ɾ	! p i s	BMN1	40
45449	taxis	ducharse	! t a . k s i s	d u ! ʧ a ɾ . s e	BAB4	73
45451	cuerdas	puerta	! k we ɾ . ð a s	! p we ɾ . t a	BAB4	73
45456	boca	ropa	! b o . k a	! r o . p a	BAB4	40
45467	falda	falta	! f a l . d a	! f a l . t a	BAB4	67
45479	justos	músculos	! x u s . t o s	! m u s . k u . l o s	SSN	40
45509	chico	chica	! ʧ i . k o	! ʧ i . k a	BAB8	47
45522	salsa	salsas	! s a l . s a	! s a l . s a s	BAB8	53
45524	bueno	control	! b we . n o	k o n ! t ɾ o l	BAB8	40
45560	pasen	pase	! p a . s e n	! p a . s e	BMN3	40
45571	dejé	tejer	d e ! x e	t e ! x e ɾ	BMN3	53
45634	frasco	casco	! f ɾ a s . k o	! k a s . k o	BAB4	53
45639	valla	bailas	! b a . ʝ a	! b aI . l a s	BAB4	53
45640	sentar	cantar	s e n ! t a ɾ	k a n ! t a ɾ	BAB4	40
45662	quiere	quiero	! k je . ɾ e	! k je . ɾ o	BAB4	60
45663	sucio	formas	! s u . θ jo	! f o ɾ . m a s	BAB4	47
45690	formal	fumar	f o ɾ ! m a l	f u ! m a ɾ	SSN	60
45697	menos	niños	! m e . n o s	! n i . ɲ o s	SSN	60
45707	probado	robado	p ɾ o ! β a . ð o	r o ! β a . ð o	BAB4	53

Table 3: *Examples of the Spanish confusions corpus. Columns correspond to confusion ID number, orthographic and phonetic transcriptions of the target and misperceived word, as well as the percentage of listeners who reported the majority confusion.*

Appendix B

Confusions defying stress-based alignment

ID	Target	Confusion	ID	Target	Confusion
38012	antes	alcohol	23171	cuento	mensaje
11124	fila	entramos	23148	salimos	mensaje
40089	puntos	manzana	22479	horrible	leche
30603	siglo	alcohol	18825	valió	ropas
1924	locura	leche	18631	sentir	cachas
14520	cerdo	entramos	17551	fuente	chicos
47593	crió	alcohol	45663	sucio	formas
24464	guardan	pozo	42186	ola	chico
8981	reír	años	35049	medias	traje
22144	guardias	olor	21654	basta	partir
543	doblar	leche	9836	ángel	adjetivo
39398	sabrá	choca	8468	sentó	pared
27816	grito	alcohol	7788	novia	vacas
18626	creó	duchas	7647	súper	entramos
16971	vuestra	manzana	46100	tenían	acción
13331	nombres	leche	46068	aunque	cama
1291	visita	caros	45524	bueno	control
808	permiso	estaré	43983	rompí	formas
783	blusa	estaré	4227	interior	gustas
731	dirige	manzana	4186	pulmón	baños
34701	pluma	cartera	38007	rubio	formas
28930	progreso	chicos	376	encuentra	rubias
21668	para	leche	3742	son	solemos
20906	sacan	alcohol	37364	imán	muchas
1878	sabía	entramos	34699	noticia	formas
16207	entienden	acción	34692	células	atención
15996	mitad	leche	31291	preguntas	comer
8966	golfo	autobús	30547	deseas	pozo
48579	tribunal	leche	25131	juré	formas
48196	comienzo	fresca	24131	remo	duchas
46949	fotos	acción	17277	pedí	aplasta
46539	dime	manzana	16197	murió	celos
41679	ocurre	pensión	15150	éste	ducha
3669	multa	tensión	10857	litros	comen
32353	gruñón	manzana	22158	nunca	jabón
2972	llegando	fumas	16605	urgencias	ropa

Table 4: *Cases where stress based alignment is not applicable due to listeners reporting a salient word from the background. Confusions can be located in the online corpus resource using the ID number.*

ID	Target	Confusion
28492	viaja	viajar
44259	corta	cortar
43306	robo	robar
36699	molesto	molestar
30173	busco	buscaron
21099	personal	persona
11640	conducta	conductor
20328	idea	ideal
43987	sabes	saber
7484	espacial	espacio
3664	ésta	estar
33607	filtro	filtrar
26051	debéis	debes
14142	robo	robar
13244	asusta	asustar
11575	vengar	venga
10089	acerca	acercar

Table 5: *Cases where stress based alignment is not applicable due to shift in stress caused by morphological variation.*

ID	Target	Confusion	ID	Target	Confusion
41520	sector	insecto	17955	llegan	tierra
290	muchacho	mucho	7096	fe	feliz
27529	circo	ciudad	45890	logramos	lobos
19735	río	frio	42480	verá	primavera
8465	nueve	duchas	42460	números	niños
35096	seres	seis	37217	robado	ropa
26870	peces	pegamos	35682	desgracia	queso
38178	médicos	niños	33363	cambien	camping
33605	sopa	soplar	31654	pegó	auto
30590	privada	triste	30414	rotas	rodillas
43045	precio	empezar	28670	muñeca	muy
30731	oír	huir	25499	acepta	acertar
25861	vago	vacío	25242	yegua	idioma
17692	forman	hoy	23420	burla	pueblo
26165	cobarde	escoba	20090	sincero	enfadado
40569	hacía	cama	16078	policía	infierno
38987	auto	audición	14959	enseñar	siempre
38738	belleza	invierno	14917	tía	magia
287	cañón	rayo	13810	volví	serie
25680	niña	escritor	12918	tierno	viaje
2528	venda	joven	10651	cubro	jabón
24598	mago	no	14371	sincero	cama

Table 6: *Cases where stress based alignment is not applicable due to other reasons.*

References

A. Agresti. Loglinear Models for Contingency Tables. In *An Introduction to Categorical Data Analysis*, pages 204–243. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006. 117

S. Ahmad and V. Tresp. Some solutions to the missing feature problem in vision. In *NIPS*, pages 393–393. Morgan Kaufmann Publishers, 1993. 75

H. Ali, N. Ahmad, X. Zhou, K. Iqbal, and S. M. Ali. DWT features performance analysis for automatic speech recognition of Urdu. *SpringerPlus*, 3:204, 2014. 73

J. B. Allen. Cochlear modeling. *IEEE ASSP Magazine*, 2(1):3–29, 1985. 3

J. B. Allen. Consonant recognition and the articulation index. *J. Acoust. Soc. Am.*, 117(4):2212–2223, 2005. 9, 27

P. D. Allopenna, J. S. Magnuson, and M. K. Tanenhaus. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4):419–439, 1998. 2

ANSI. S3. 5-1997, Methods for the calculation of the speech intelligibility index, 1997. 5

T. L. Arbogast, C. R. Mason, and G. Kidd. The effect of spatial separation on informational and energetic masking of speech. *J. Acoust. Soc. Am.*, 112(5): 2086–2098, 2002. 108

P. F. Assmann and Q. Summerfield. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, 88(2): 680–697, 1990. 129

R. H. Baayen. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press, Cambridge, 2008. 36

P. J. Bailey, Q. Summerfield, M. Dorman, et al. On the identification of sine-wave analogues of certain speech sounds. Technical report, Report no: SR-51/52, Haskins Labs, 1977. 73

J. Barker and M. Cooke. Modelling speaker intelligibility in noise. *Speech Comm.*, 49(5):402–417, 2007. 18

J. Barker, M. P. Cooke, and D. P. Ellis. Decoding speech in the presence of other sources. *Speech Comm.*, 45(1):5–25, 2005. xv, 66, 73, 76, 77, 78, 133

J. Barker, R. Marxer, E. Vincent, and S. Watanabe. The third CHiME speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511. IEEE, 2015. 1

J. R. Benkí. Effects of signal-independent factors in speech perception. In *Proceedings of the 28th annual meeting of the Berkeley Linguistics Society*, pages 63–74, 2002. 23

J. R. Benkí. Analysis of English nonsense syllable recognition in noise. *Phonetica*, 60(2):129–157, 2003. 29, 61

J. R. Benkí. Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition. *J. Acoust. Soc. Am.*, 113(3):1689–1705, 2003. 30, 54, 55, 62

H. Bird. Slips of the ear as evidence for the postperceptual priority of grammaticality. *Linguistics*, 36(3):469–516, 1998. 61

J. Bird and C. Darwin. Effects of a difference in fundamental frequency in separating two sentences. *Psychophysical and Physiological Advances in Hearing*, pages 263–269, 1998. 112

L. Blin, O. Boeffard, and V. Barreaud. Web-based listening test system for speech synthesis and speech conversion evaluation. In *LREC*, pages 2270–2274, 2008. 11

Z. S. Bond. Morphological Errors in Casual Conversation. *Brain and Language*, 68(12):144 – 150, 1999a. 31

Z. S. Bond. *Slips of the ear: Errors in the perception of casual conversation.* Academic Press, London, 1999b. 3, 24, 28, 49, 97

S. Bookheimer. Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience*, 25(1):151–188, 2002. 3

A. Boothroyd and S. Nittrouer. Mathematical treatment of context effects in phoneme and word recognition. *J. Acoust. Soc. Am.*, 84(1):101–114, 1988. 23

A. Bregman. *Auditory Scene Analysis: The perceptual organization of sound. 1990.* MIT Press, Cambridge, MA, 1990. 68, 73, 75, 77

J. Brokx and S. Nooteboom. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10(1):23–36, 1982. 129

A. Bronkhorst and R. Plomp. The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J. Acoust. Soc. Am.*, 83(4): 1508–1516, 1988. 68, 108

C. Browman. Perceptual processing: Evidence from slips of the ear. In Victoria A.Fromkin, editor, *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand.* Academic Press, New York, 1980. 3, 4, 24, 30, 48, 49, 62

D. Brungart, N. Iyer, E. R. Thompson, B. D. Simpson, S. Gordon-Salant, J. Schurman, C. Vogel, and K. Grant. Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli.

In *Proceedings of Meetings on Acoustics ICA2013*, volume 19. ASA, 2013. 109, 110

D. S. Brungart. Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.*, 109(3):1101–1109, 2001. 18, 69, 90, 101, 110, 112, 121

D. S. Brungart and B. D. Simpson. Within-ear and across-ear interference in a cocktail-party listening task. *J. Acoust. Soc. Am.*, 112(6):2985–2995, 2002. 68, 108

D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J. Acoust. Soc. Am.*, 120(6):4007–4018, 2006. 68, 69, 72, 90, 99, 101, 108, 121, 122, 128

E. Buss, J. W. Hall III, and J. H. Grose. Spectral integration of synchronous and asynchronous cues to consonant identification. *J. Acoust. Soc. Am.*, 115 (5):2278–2285, 2004. 71, 72

R. Carhart, T. W. Tillman, and E. S. Greetis. Perceptual masking in multiple sound backgrounds. *J. Acoust. Soc. Am.*, 45(3):694–703, 1969. 68

E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, 25(5):975–979, 1953. 68

C. Christiansen, M. S. Pedersen, and T. Dau. Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Comm.*, 52(7-8):678–692, 2010. 65

T. U. Christiansen and S. Greenberg. Perceptual Confusions Among Consonants, Revisited Cross-Spectral Integration of Phonetic-Feature Information and Consonant Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):147–161, 2012. 28, 61

A. Cohen. On the graphical display of the significant components in two-way contingency tables. *Communications in Statistics-Theory and Methods*, 9(10): 1025–1041, 1980. 57

M. Cooke. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 119(3):1562–1573, 2006. 5, 6, 65, 71, 101, 111

M. Cooke. Discovering consistent word confusions in noise. In *Proc. Interspeech*, pages 1887–1890. ISCA, 2009. 4, 7, 13, 106, 122, 127, 129, 131, 134

M. Cooke, P. D. Green, and M. Crawford. Handling missing data in speech recognition. In *ICSLP*, 1994. 74

M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Comm.*, 34(3): 267–285, 2001. 74

M. Cooke, J. Barker, M. L. G. Lecumberri, and K. Wasilewski. Crowdsourcing for Word Recognition in Noise. In *Proc. Interspeech*, pages 3049–3052, 2011. 11

F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.*, 24(6):597–606, 1952. 26

R. M. Cooper. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing . *Cognitive Psychology* , 6(1):84 – 107, 1974. . 2

M. A. Covington. An Algorithm to Align Words for Historical Comparison. *Comput. Linguist.*, 22(4):481–496, dec 1996. 32

J. F. Culling and C. Darwin. Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *J. Acoust. Soc. Am.*, 93(6):3454–3467, 1993. 129

A. Cutler. The reliability of speech error data. In A. Cutler, editor, *Slips of the Tongue and Language Production*. Walter de Gruyter/Mouton, Amsterdam, 1982. 4, 62

A. Cutler and S. Butterfield. Rhythmic Cues to Speech Segmentation: Evidence from Juncture Misperception. *Journal of Memory and Language*, 31:218–236, 1992. 4

A. Cutler, D. Norris, and J. N. Williams. A note on the role of phonological expectations in speech segmentation. *Journal of Memory and Language*, 26(4): 480–487, 1987. 2

A. Cutler, A. Weber, R. Smits, and N. Cooper. Patterns of English phoneme confusions by native and non-native listeners. *J. Acoust. Soc. Am.*, 116(6): 3668–3678, 2004. 37

G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, 2012. 1

C. Darwin and R. Hukin. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Am.*, 107(2):970–977, 2000. 129

C. J. Darwin. Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology*, 33(2):185–207, 1981. 102

C. J. Darwin. Auditory processing and speech-perception. *Attention and Performance*, 10:197–209, 1984. 102

T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the effectivesignal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.*, 99(6):3615–3622, 1996. 6

P. C. Delattre, A. M. Liberman, and F. S. Cooper. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.*, 27(4):769–773, 1955. 26

B. Delgutte. Analysis of French stop consonants using a model of the peripheral auditory system. *Invariance and Variability in Speech Processes*, pages 163–177, 1986. 3

J. R. Dubno and H. Levitt. Predicting consonant confusions from acoustic analysis. *J. Acoust. Soc. Am.*, 69(1):249–261, 1981. 25, 26, 29, 44, 61

A. Duchon, M. Perea, A. Sebastiá n-Gallés, Nuria and Martí, and M. Carreiras. EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, 45(4):1246–1258, 2013. 50

L. S. Eisenberg, D. D. Dirks, and T. S. Bell. Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing. *Journal of Speech, Language, and Hearing Research*, 38(1):222–233, 1995. 120

R. Felty, A. Buchwald, T. M. Gruenenfelder, and D. B. Pisoni. Misperceptions of spoken words: Data from a random sample of American English words. *J. Acoust. Soc. Am.*, 134(1):572–585, 2013. 23, 25, 31, 51, 54, 55, 56, 62

J. M. Festen and R. Plomp. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.*, 88(4):1725–1736, 1990. 18, 68, 69, 106

H. Fletcher. *Speech and hearing in communication*. D. van Nostrand, New York-London, 1953. 25

H. Fletcher and R. H. Galt. The perception of speech and its relation to telephony. *J. Acoust. Soc. Am.*, 22(2):89–151, 1950. 5, 25

N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19(1):90–119, 1947. 5

R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton. The role of perceived spatial separation in the unmasking of speech. *J. Acoust. Soc. Am.*, 106(6):3578–3588, 1999. 108

R. L. Freyman, U. Balakrishnan, and K. S. Helfer. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *J. Acoust. Soc. Am.*, 115(5):2246–2256, 2004. 101, 121

M. Friendly. *Visualizing categorical data*. Sas Institute, Cary, NC, 2000. 57

W. F. Ganong. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1):110, 1980. 54

S. Garnes and Z. S. Bond. A Slip of the Ear? A Snip of the Ear? A Slip of the Year? . In Victoria A.Fromkin, editor, *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*. Academic Press, New York, 1980. 3, 4, 30, 31, 48, 61, 62, 97

O. Ghitza. Auditory nerve representation as a basis for speech processing. *Advances in Speech Signal Processing*, pages 453–485, 1992. 3

O. Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):115–132, 1994. 6

S. D. Goldinger, P. A. Luce, and D. B. Pisoni. Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28(5):501–518, 1989. 2

S. Gordon-Salant. Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing. *J. Acoust. Soc. Am.*, 80(6): 1599–1607, 1986. 25, 26, 28, 29, 60

K. W. Grant and B. E. Walden. Evaluating the articulation index for auditory–visual consonant recognition. *J. Acoust. Soc. Am.*, 100(4):2415–2424, 1996. 27

F. Grosjean. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28(4):267–283, 1980. 2

B. Hagerman. Sentences for testing speech intelligibility in noise. *Scandinavian Audiology*, 11(2):79–87, 1982. 9

V. Hautamaki, T. Kinnunen, M. Nosratighods, K.-A. Lee, B. Ma, and H. Li. Approaching human listener accuracy with modern speaker verification. In *Proc. Interspeech*, pages 1473–1476, 2010. 1

Z. Hernández-Figueroa, F. J. Carreras-Riudavets, and G. Rodríguez-Rodríguez. Automatic syllabification for Spanish using lemmatization and derivation to solve the prefixs prominence issue. *Expert Systems with Applications*, 40(17): 7122–7131, 2013. 16

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. 1

I. J. Hirsh. The relation between localization and intelligibility. *J. Acoust. Soc. Am.*, 22(2):196–200, 1950. 68, 108

I. Holube and B. Kollmeier. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *J. Acoust. Soc. Am.*, 100(3):1703–1716, 1996. 6, 65, 101

H. Honing. Evidence for tempo-specific timing in music using a web-based experimental setup. *Journal of Experimental Psychology: Human Perception and Performance*, 32(3):780, 2006. 11

T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe. Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend. *Computer Speech & Language*, 2017. 73

P. A. Howard-Jones and S. Rosen. Uncomodulated glimpsing in checkerboard-noise. *J. Acoust. Soc. Am.*, 93(5):2915–2922, 1993. 72

T. F. Jaeger. Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, 59(4):434–446, 2008. 36

T. Jürgens. *A microscopic model of speech recognition for listeners with normal and impaired hearing*. PhD thesis, Oldenburg, Univ., Diss., 2010, 2010. 5

T. Jürgens and T. Brand. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *J. Acoust. Soc. Am.*, 126(5):2635–2648, 2009. 6, 65, 101

J. Kahn, N. Audibert, S. Rossato, and J. F. Bonastre. Speaker verification by inexperienced and experienced listeners vs. speaker verification system. In *Proc. ICASSP*, pages 5912–5915. IEEE, 2011. 1

H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné . Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Comm.* , 27(34):187 – 207, 1999. 112

G. Kidd, T. L. Arbogast, C. R. Mason, and F. J. Gallun. The advantage of knowing where to listen a. *J. Acoust. Soc. Am.*, 118(6):3804–3815, 2005. 101

G. Kidd, C. R. Mason, J. Swaminathan, E. Roverud, K. K. Clayton, and V. Best. Determining the energetic and informational components of speech-on-speech masking. *J. Acoust. Soc. Am.*, 140(1):132–144, 2016. 90, 101, 109

D. Koekemoer, J. Clark, et al. Intercontinental hearing assessment–a study in tele-audiology. *Journal of Telemedicine and Telecare*, 16(5):248–252, 2010. 11

G. Kondrak. Phonetic Alignment and Similarity. *Computers and the Humanities*, 37(3):273–291, 2003. 32

K. D. Kryter. Methods for the calculation and use of the articulation index. *J. Acoust. Soc. Am.*, 34(11):1689–1697, 1962. 5

M. Kutas, S. A. Hillyard, et al. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980. 3

W. Labov. *Principles of Linguistic Change, Cognitive and Cultural Factors*. Principles of Linguistic Change. Wiley, Hoboken, New Jersey, 2011. 3, 10

A. Lahiri and H. Reetz. Underspecified recognition. *Laboratory Phonology*, 7: 637–675, 2002. 45, 46

F. Li, A. Menon, and J. B. Allen. A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *J. Acoust. Soc. Am.*, 127(4):2599–2610, 2010. 6, 65, 106, 127, 129

A. M. Liberman, P. Delattre, and F. S. Cooper. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, 65(4):497–516, 1952. 2

A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8):1, 1954. 26, 106

P. A. Luce and D. B. Pisoni. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1):1, 1998. 2, 23, 30, 54, 55, 56, 63, 71, 132

I. Maddieson and S. F. Disner. *Patterns of sounds*. Cambridge university press, Cambridge, 1984. 54

V. Marian, J. Bartolotti, S. Chabal, and A. Shook. CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one*, 7(8):e43230, 2012. 12

W. D. Marslen-Wilson and A. Welsh. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1): 29–63, 1978. 2

R. Marxer, J. Barker, M. Cooke, and M. L. Garcia Lecumberri. A corpus of noise-induced word misperceptions for English. *J. Acoust. Soc. Am.*, 140(5): EL458–EL463, 2016. 18, 103

S. L. Mattys and L. Wiget. Effects of cognitive load on speech recognition. *Journal of Memory and Language*, 65(2):145–160, 2011. 68

S. L. Mattys, J. Brooks, and M. Cooke. Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, 59(3):203–243, 2009. 24, 63

J. L. McClelland and J. L. Elman. The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1–86, 1986. 2

R. Meringer. *Aus dem leben der sprache*. B. Behr, 1908. 3, 28

R. Meringer, C. Mayer, A. Cutler, and D. Fay. *Versprechen und Verlesen. Eine psychologisch-linguistische Studie.([With the assistance of] Carl Mayer.)*, volume 2. John Benjamins Publishing, Amsterdam, 1895. 3, 62

B. T. Meyer, M. Wächter, T. Brand, and B. Kollmeier. Phoneme confusions in human and automatic speech recognition. In *Proc. Interspeech*, pages 1485–1488, 2007. 1

G. A. Miller and J. C. Licklider. The intelligibility of interrupted speech. *J. Acoust. Soc. Am.*, 22(2):167–173, 1950. 71

G. A. Miller and P. E. Nicely. An Analysis of Perceptual Confusions Among Some English Consonants. *J. Acoust. Soc. Am.*, 27(2):338–352, 1955. 9, 10, 23, 24, 26, 27, 28, 29, 37, 44, 60, 61

C. B. Mills. Effects of context on reaction time to phonemes. *Journal of Verbal Learning and Verbal Behavior*, 19(1):75–83, 1980. 2

B. C. J. Moore and B. R. Glasberg. Comparisons of frequency selectivity in simultaneous and forward masking for subjects with unilateral cochlear impairments. *J. Acoust. Soc. Am.*, 80(1):93–107, 1986. 122

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins . *Journal of Molecular Biology* , 48(3):443 – 453, 1970. . 4, 32

D. L. Neff. Signal properties that reduce masking by simultaneous, random-frequency maskers. *J. Acoust. Soc. Am.*, 98(4):1909–1920, 1995. 68, 69

D. Norris. Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234, 1994. 2

H. C. Nusbaum, D. B. Pisoni, and C. K. Davis. Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report*, 10(10):357–376, 1984. 31

J. Ohala and H. Kawasaki-Fukumori. Alternatives to the sonority hierarchy for explaining segmental sequential constraints. *Language and its ecology: Essays in memory of Einar Haugen*, 100:343, 1997. 45

L. Osterhout and P. J. Holcomb. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6):785–806, 1992. 3

A. J. Oxenham. Forward masking: Adaptation or integration? *J. Acoust. Soc. Am.*, 109(2):732–741, 2001. 121

E. J. Ozmeral, E. Buss, and J. W. Hall III. Asynchronous glimpsing of speech: Spread of masking and task set-size. *J. Acoust. Soc. Am.*, 132(2):1152–1164, 2012. 72

S. G. Parker. *Quantifying the sonority hierarchy*. PhD thesis, University of Massachusetts Amherst, 2002. 45

J. Peissig. Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *J. Acoust. Soc. Am.*, 101(3): 1660–1670, 1997. 108

S. A. Phatak and J. B. Allen. Consonant and vowel confusions in speech-weighted noise. *J. Acoust. Soc. Am.*, 121(4):2312–2326, 2007. 5, 9, 27, 37, 44, 60, 61, 65

S. A. Phatak, A. Lovitt, and J. B. Allen. Consonant confusions in white noise. *J. Acoust. Soc. Am.*, 124(2):1220–1233, 2008. 5, 27, 28, 60

J. Pickett. Perception of vowels heard in noises of various spectra. *J. Acoust. Soc. Am.*, 29(5):613–620, 1957. 25, 29

D. B. Pisoni. Some current theoretical issues in speech perception. *Cognition*, 10 (1-3):249, 1981. 48

M. A. Pitt and A. G. Samuel. Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception and Performance*, 32 (5):1120, 2006. 54

C. J. Plack and A. J. Oxenham. Basilar-membrane nonlinearity and the growth of forward masking. *J. Acoust. Soc. Am.*, 103(3):1598–1608, 1998. 121

I. Pollack. Auditory informational masking. *J. Acoust. Soc. Am.*, 57(S1):S5–S5, 1975. 68

I. Pollack, H. Rubenstein, and L. Decker. Intelligibility of Known and Unknown Message Sets. *J. Acoust. Soc. Am.*, 31(3):273–279, 1959. 55, 62

I. Pollack, H. Rubenstein, and L. Decker. Analysis of Incorrect Responses to an Unknown Message Set. *J. Acoust. Soc. Am.*, 32(4):454–457, 1960. 54

REAL. Corpus diacrónico del español. http://www.rae.es, 2008. 15, 20

R. E. Remez, P. E. Rubin, D. B. Pisoni, T. D. Carrell, et al. Speech perception without traditional speech cues. *Science*, 212(4497):947–949, 1981. 73

K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler. Release from informational masking by time reversal of native and non-native interfering speech. *J. Acoust. Soc. Am.*, 118(3):1274–1277, 2005. 5, 68, 102, 107

N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *J. Acoust. Soc. Am.*, 114(4):2236–2252, 2003. 121

H. B. Savin and T. G. Bever. The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, 9(3):295–302, 1970. 2, 55

D. L. Scarborough, C. Cortese, and H. S. Scarborough. Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human perception and performance*, 3(1):1, 1977. 52

O. Scharenborg, E. Sanders, and B. Cranen. Collecting a corpus of Dutch noise-inducedslips of the ear'. In *Proc. Interspeech*, pages 2600–2604, 2014. 18, 103

M. T. Scheffers. Simulation of auditory analysis of pitch: An elaboration on the DWS pitch meter. *J. Acoust. Soc. Am.*, 74(6):1716–1725, 1983. 73, 112, 129

C. Scheidiger and J. B. Allen. Effects of NALR on consonant-vowel perception. In *the 4th International Symposium on Auditory and Audiological Research (ISAAR-2013), Nyborg, Denmark*, 2013. 66

R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303, 1995. 73

R. N. Shepard. Psychological representation of speech sounds. *Human communication: A unified view*, pages 67–113, 1972. 29

R. Shillcock. Lexical Hypotheses in Continuous Speech. In Altmann, Gerry T. M., editor, *Cognitive Models of Speech Processing*, pages 24–49. MIT Press, Cambridge, MA, USA, 1990. 2

H. L. Somers. Aligning Phonetic Segments for Children's Articulation Assessment. *Comput. Linguist.*, 25(2):267–275, jun 1999. 32

R. F. Stanners, J. E. Jastrzembski, and A. Westbrook. Frequency and visual quality in a word-nonword classification task. *Journal of Verbal Learning and Verbal Behavior*, 14(3):259–264, 1975. 52

H. J. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.*, 67(1):318–326, 1980. 5

T. M. Sullivan and R. M. Stern. Multi-microphone correlation-based processing for robust speech recognition. In *Proc. ICASSP*, volume 2, pages 91–94. IEEE, 1993. 73

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short time objective intelligibility measure for time-frequency weighted noisy speech. In *Proc. ICASSP*, pages 4214–4217, 2010. 5, 65

M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632, 1995. 2

K. Tang. *Naturalistic Speech Misperception*. PhD thesis, University College London, 2015. 4, 11, 24, 28, 29, 30, 31, 45, 48, 49, 54, 55, 61, 62, 63

K. Tang and A. Nevins. Naturalistic Speech Misperception - a Computational Corpus-based Study. In *Proceedings of the 43rd Meeting of the North East Linguistic Society*, 2012. 3, 4, 10, 23, 24, 32, 49, 63

Y. Tang, M. Cooke, et al. Glimpse-based metrics for predicting speech intelligibility in additive noise conditions. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 2488–2492. International Speech and Communication Association, 2016. 72

W. Tanner. What is masking? *J. Acoust. Soc. Am.*, 30(10):919–921, 1958. 68

M. A. Tóth, M. L. García Lecumberri, Y. Tang, and M. Cooke. A corpus of noise-induced word misperceptions for Spanish. *J. Acoust. Soc. Am.*, 137(2): EL184–EL189, 2015. 4

L. A. Varghese, E. J. Ozmeral, V. Best, and B. G. Shinn-Cunningham. How visual cues for when to listen aid selective auditory attention. *Journal of the Association for Research in Otolaryngology*, 13(3):359–368, 2012. 101

L. Varnet. Using auditory classification images for the identification of fine acoustic cues used in speech perception. *Frontiers in Human Neuroscience*, 7:865, 2013. 6, 106, 127

N. J. Versfeld and W. A. Dreschler. The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *J. Acoust. Soc. Am.*, 111(1):401–408, 2002. 68, 106

M. S. Vitevich. Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear. *Language and Speech*, 45: 407–434, 2002. 4, 31, 55

M. S. Vitevitch, M. K. Stamer, and J. A. Sereno. Word length and lexical competition: Longer is the same as shorter. *Language and Speech*, 51(4):361–383, 2008. 56

K. Wagener, J. L. Josvassen, and Ardenkjær, Regitze. Design, optimization and evaluation of a Danish sentence test in noise: Diseño, optimización y evaluación de la prueba Danesa de frases en ruido. *International Journal of Audiology*, 42 (1):10–17, 2003. 9

X. Wang and L. E. Humes. Factors influencing recognition of interrupted speech. *J. Acoust. Soc. Am.*, 128(4):2100–2111, 2010. 71

P. Warren and W. Marslen-Wilson. Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics*, 41(3):262–275, 1987. 2

R. M. Warren et al. Perceptual restoration of missing speech sounds. *Science*, 167(3917):392–393, 1970. 122

R. M. Warren, K. R. Riener, J. A. Bashford, and B. S. Brubaker. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Attention, Perception, & Psychophysics*, 57(2):175–182, 1995. 72

A. Weber and O. Scharenborg. Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3):387–401, 2012. 3, 56

R. Wegel and C. Lane. The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Physical Review*, 23(2): 266, 1924. 68

M. Weintraub. *A Theory And Computational Model Of Auditory Monaural Sound Separation*. PhD thesis, Stanford University, 1985. 112

C. P. Whaley. Wordnonword classification time. *Journal of Verbal Learning and Verbal Behavior*, 17(2):143–154, 1978. 52

D. L. Woods, E. W. Yund, T. J. Herron, and M. A. I. U. Cruadhlaoich. Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise. *J. Acoust. Soc. Am.*, 127(3):1609–1623, 2010. 29, 37, 44, 60, 61

J. Zaar and T. Dau. Sources of variability in consonant perception of normal-hearing listeners. *J. Acoust. Soc. Am.*, 138(3):1253–1267, 2015. 6, 9, 10, 65, 66

J. Zaar and T. Dau. Predicting consonant recognition and confusions in normal-hearing listeners. *J. Acoust. Soc. Am.*, 141(2):1051–1064, 2017. 101

G. K. Zipf. *The psycho-biology of language.* Houghton, Mifflin, Boston, 1935. 55, 62

P. Zwitserlood. The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, 32(1):25–64, 1989. 2