PERCEPTUAL CUE WEIGHTING OF INTERVOCALIC VELAR PLOSIVES IN ENGLISH

PERCEPTUAL CUE WEIGHTING OF INTERVOCALIC VELAR PLOSIVES IN ENGLISH

By OLIVIER KIBBINS MERCIER, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University
MASTER OF SCIENCE (2019)
Hamilton, Ontario (Cognitive Science of Language)

TITLE: PERCEPTUAL CUE WEIGHTING OF INTERVOCALIC VELAR PLOSIVES IN ENGLISH

AUTHOR: Olivier Kibbins Mercier, B.Sc. (University of Ottawa)

SUPERVISORS: Dr. Daniel Pape, Dr. Elisabet Service

NUMBER OF PAGE: x, 70

Lay Abstract

Speech signals contain information about the intended sound, which we use in order to maintain consistent speech perception despite variation among speakers. This thesis explores the perceptual weighting of the acoustic cues which distinguish between the stop consonants /k/ and /g/. English listeners were presented with biomechanical speech stimuli which varied by four acoustic cues (Voice Onset Time, Voicing Maintenance, Vowel Length, Consonant Length). The results reveal that Voicing Maintenance has a relatively higher influence than the other cues. However, the interaction between the cues is complex, and the sensitivity to different cues varies among individuals. These results suggest that rather than a single cue having influence on voicing perception, it is the complex interplay of all available cues.

Abstract

When producing speech, the acoustic signals contain several extractable parameters (cues) about the intended sound which are reliably mapped to achieve robust phoneme identification in speech perception, despite large acoustic variation in speech signals between and within individuals. This thesis presents findings on the perceptual cue weighting of intervocalic velar stop consonants /k/ and /g/ for English listeners, through use of biomechanical stimuli. Four acoustic cues were systematically manipulated to create the experimental stimuli: voicing maintenance throughout consonantal closure (VM), voice onset time (VOT), the duration of consonantal closure (CL) and the duration of the previous vowel segment (VL). This thesis presents the findings of three experiments. Experiment 1 investigates the cue weighting of three acoustic cues (VM, VL, CL) in the absence of a perceptual VOT cue. Experiment 2 defines a perceptually ambiguous VOT value [for the given biomechanical stimuli] and compares the influences of VOT and VM. Experiment 3 analyses how the perceptual cue weighting of the three acoustic cues (VM, VL, CL) is affected when a perceptually ambiguous VOT value is added to stimuli. The results from the three experiments reveal that VM has a significantly higher influence on voicing perception than the other cues, and only at low VM levels do other cues increase their influence. The cues interacted significantly with each other. The effect of VOT – known in literature as the main cue for consonantal voicing distinctions – was apparent at high VM values but significantly increased at low VM values. Experiment 2 gave additional evidence that individual listeners utilized very different strategies in determining voicing perception. Ultimately the results show that the perceptual cue weighting process is highly complex and cannot be attributed to only one or two perceptual cues (e.g. VOT).

Acknowledgements

I would like to thank my supervisor Daniel Pape for his helpful guidance throughout the project. I would also like to thank my thesis committee members Elisabet Service and Magda Stroinska for their valuable comments on my thesis drafts.

Table of Contents

List of Tables

List of Figures

List of all Abbreviations and symbols

VOT = voice onset time

VM = voicing maintenance throughout consonantal closure

CL = duration of the consonantal closure (consonant length)

VL = duration of the preceding vowel segment (vowel length)

Declaration of academic achievement

The conceptualization of this thesis was conducted by myself, in collaboration with Dr. Pape. I explored and analyzed the data under the supervision of Dr. Pape. The statistical analysis was completed by myself and Dr. Pape. Contributions to the each of the articles we intend to submit will be noted in the preface to those chapters. The present thesis was written by myself, with comments from Dr. Pape, Dr. Service, and Dr. Stroinska.

Chapter 1: Introduction

This thesis will explore the perceptual cue weighting of intervocalic velar stop consonants /k g/ in English listeners. Produced at the same place of articulation, these consonants vary only by the phonological dimension of voicing. This thesis explores 4 main acoustic cues which serve to differentiate the two sounds (/g k/) according to this dimension of phonological voicing: voice onset time (*henceforth VOT*), voicing maintenance throughout consonantal closure (*henceforth VM*), length of the preceding vowel (*henceforth VL*), and the length of the consonantal closure itself (*henceforth CL*). In English, VOT is traditionally considered to be the main cue for these voicing distinctions (Lisker & Abramson, 1964; Abramson & Whalen, 2017). The independent effects of the cues are well known, but their hierarchal interactions in English stop voicing are not. The effect of each cue, their interactions, and their hierarchy in consonant identification (perception) were analysed through the procedures of three main experiments and several pilot studies.

The conducted experiments use stimuli created and manipulated by biomechanical modelling (more below). The first experiment investigated cue weighting of three acoustic cues (VM, VL and CL) in the absence of a perceptual VOT cue. The second experiment aimed to define a perceptually ambiguous VOT value for the given biomechanical stimuli at which listeners change their perception from /k/ to /g/. The third experiment analysed how the cue weighting of three acoustic cues (VM, VL and CL) changes when ambiguous VOT values (derived from experiment 2) are presented within the stimuli compared to when they are absent (as they are in experiment 1). This chapter begins with a general overview of the theoretical domain, followed by a literature review of relevant studies, and concluding with the experimental design and hypotheses.

1.1. Background

A speech signal is a continuous stream of acoustic information which must be deciphered into discrete meaningful percepts. At the base level of human communication, this is considered to occur by placing sounds into their respective mental categories, known as phonemes. In principle, this task is accomplished by identifying concurrent acoustic signals which reveal recognizable information about the intended sound. However, natural speech tokens display a significant amount of variation, resulting from individual differences between subjects (differences in vocal tract size or muscle control), or from varying productions of the same sound in different contexts by the same subjects (Chodroff & Wilson, 2017; Clayards, 2017). Despite large acoustic variation in the speech signal for speakers, listeners still reliably maintain consistent phoneme identification under different speech conditions. The signals responsible for consistent perception of human speech sounds have been widely described across many languages and are known as perceptual cues. This thesis will explore the perceptual cues of significance for differentiating English stop consonants in their voicing dimension. It employs biomechanical speech modelling to analyse their effects, interactions, and hierarchical perceptual weighting under different speech conditions.

1.2. Synthetic Speech and Biomechanical Modelling

The invention of synthetic speech modelling has played a central role in identifying perceptual cues (i.e. Pape, Jesus, & Perrier, 2012). Several methods for creating experimental speech stimuli exist. The primary methods include Klatt articulatory synthesis, biomechanical modelling and manipulations of natural speech (Klatt, 1980; Pape, Jesus, & Birkholz, 2015). Early phonetic studies consisted mainly of systematically manipulating synthetic speech (using

Klatt synthesizers) to determine how this affected listener perception (i.e. Klatt & Klatt, 1990) on the one hand, and analysing acoustic spectra from natural speech productions on the other (i.e. Lisker & Abramson, 1964). These studies enumerated and described perceptual cues for all speech-sound classes. Biomechanical modelling allows for complex controlled manipulations while maintaining realistic articulatory transitions to account for coarticulation. All tongue movements, trajectories, and phoneme targets are designed to mimic natural speech targets (Pape, Jesus, & Perrier, 2012; Pape, Jesus, & Birkholz, 2015). By employing this new synthesis method, stimuli with precise manipulations for each predictor variable can be created, while ensuring control of desired or undesired variables and acoustic artifacts. This allows more precise control over more natural sounding stimuli in the research of perceptual cues.

1.3. Stop consonants & Voicing

Phonology is the study of the mental categorization and representation of recognizable sounds and serves to classify and describe the limited set of available phonemes in the world's languages (Handbook of the IPA, 1999). The International Phonetics Association classifies each consonant sound into a table with three dimensions, place of articulation (i.e. bilabial, alveolar, velar), manner of articulation (e.g., stops, fricatives, nasals), and voicing quality (voiced or voiceless). Phonetics is the study of the articulatory gestures used in natural speech production and the resulting acoustic output (Handbook of the IPA, 1999).

This thesis focuses on the cue-weighting for voicing perception in English stop consonant phonemes. There are six stop consonants in English: /p b t d k g/. They vary by two dimensions: place of articulation (bilabial, alveolar, velar) and voicing quality (voiced/voiceless). Stop consonants are characterized articulatorily by a complete closure in the vocal tract during

production. The articulatory closure produces a build-up of air posterior to the closure, and the consequent release usually creates a discernible 'burst' in the acoustic signal, although this burst signal is not always present. At the velar place of articulation (the focus of this study) the tongue body is raised to create a closure at the velum, and the two consonant sounds /k/ and /g/ differ only in their phonological voicing distinction (Handbook of the IPA, 1999). Phonological voicing specifically refers to a mental category which may or may not reflect true (phonetic) voicing (i.e. the actual vocal fold vibrations).

Frequently, natural speech productions do not match their phonological describing features. This is also the case in the voicing dimension (e.g., devoicing). For this reason, when discussing the voicing dimension it is important to distinguish between phonological and phonetic voicing. Phonetic voicing is the actual presence of glottal pulsing (vocal fold vibrations) during the consonantal production. Phonological voicing, on the other hand, refers to the phonological classification of the sounds, such as /k/ and /g/, as either voiced or voiceless, independent of the surface phonetic presence of that voicing pattern. This voicing terminology in phonology originates in languages where phonological voicing matches phonetic voicing, which is not always the case (Handbook of the IPA, 1999). In some languages (e.g., French & Italian) a phonologically voiced /g/ would typically be produced with continuous phonetic voicing throughout the consonantal closure, while a phonologically voiceless /k/ would typically be produced with partial or no glottal pulsing (Lisker & Abramson, 1964). In other languages (e.g., English, German and European Portuguese[1]; Lisker & Abramson, 1964; Pape & Jesus, 2014a), phonologically voiced stops tend to be phonetically devoiced. In these languages phonological

[1] Note that although European Portuguese is a Romance language in the same family as French and Italian, research (Pape & Jesus, 2014a) has shown that its voicing perception patterns align more closely with those of the Germanic languages (English and German), such that phonologically voiced stops tend to be phonetically devoiced, at least partially.

voicing serves to separate the mental representations of consonant pairs such as /k g/ and is not representative of actual phonetic voicing during production. In phonetic voicing, we therefore have an acoustic signal which causes differential perception depending on its presence or absence, and whose effect varies between languages. This is an example of a perceptual cue which serves to differentiate /k/ from /g/ along the voicing dimension.

1.4. Perceptual Cues in Stop Voicing

Every phoneme has its own defining set of perceptual cues. These are perceptually extractable acoustic signals which are responsible for the robust identification of natural sounds coming from different speaker sources or within-speaker variation. These acoustic cues provide information about the incoming sound and can work independently or in tandem to produce reliable and consistent responses. Perceptual cue-weighting refers to the complex and hierarchical manner in which acoustic cues interact to create a stable perceptual construct (Francis, Baldwin, & Nusbaum, 2000; Pape & Jesus, 2014a). No single cue is necessary for robust perception, since in the absence of one major cue, other cues are suitable/sufficient to guarantee a robust perceptual outcome. Cues which have no main effect on their own may influence other cues to increase or decrease their effect (Pape & Jesus, 2014a; Pape & Jesus, 2014b). Importantly, the perceptual weighting varies from language to language (Cho & Ladefoged, 1999; Pape & Jesus, 2014b), and even between individuals in the same language community (Clayards, 2017). Languages which share one or more phonemes may differ by which cues are defining for the phoneme, and the perceptual weight given to each cue (Oglesbee, 2008).

Differentiating intervocalic velar stop consonants by their dimension of phonological voicing invokes numerous acoustic cues. Production and perception studies employing natural and artificial speech conditions have uncovered such cues for phonemes across many of the world's languages.

Although voice onset time (*henceforth VOT)* is typically reported as the main perceptual cue (Lisker & Abramson, 1964; Abramson & Whalen, 2017), other cues also serve to distinguish voicing, either in combination with VOT or in its absence. These other cues include voicing maintenance throughout consonantal closure (*henceforth VM*), length of the preceding vowel (*henceforth VL*), the length of the consonantal closure itself (*henceforth CL*), aspiration, changes in fundamental frequency patterns, formant transitions, and acoustic burst features (Raphael, 1972; Francis, Baldwin, & Nusbaum, 2000; Li, Mennon, & Allen, 2010; Pape & Jesus, 2014a; Clayards, 2017).

1.4.1. Voicing Maintenance

Phonetically, voicing refers to the extent of vocal fold vibrations (glottal pulsing) throughout stop consonant production. Acoustically, the presence of phonetic voicing is represented by a characteristic acoustic energy during the stop closure, shown below in **Figure 1a** (top), and the absence of phonetic voicing is represented by an acoustic spectrum lacking energy at all frequency levels (i.e. silence), as shown in **Figure 1b** (middle)**.** The phonetic voicing dimension is not simply binary (absent or present) but can include any range of voicing maintenance (**Figure 1c,** bottom). When referring to values for VM, they will be presented as a percentage indicating the extent of phonetic voicing maintenance throughout the consonantal closure. Note that there may be confusion between the acoustic cue of voicing maintenance

(VM) and the perceptual voicing response (/g/ or /k/). All psychophysical figures involving participant responses will show the perceptual voicing response on the y-axis.

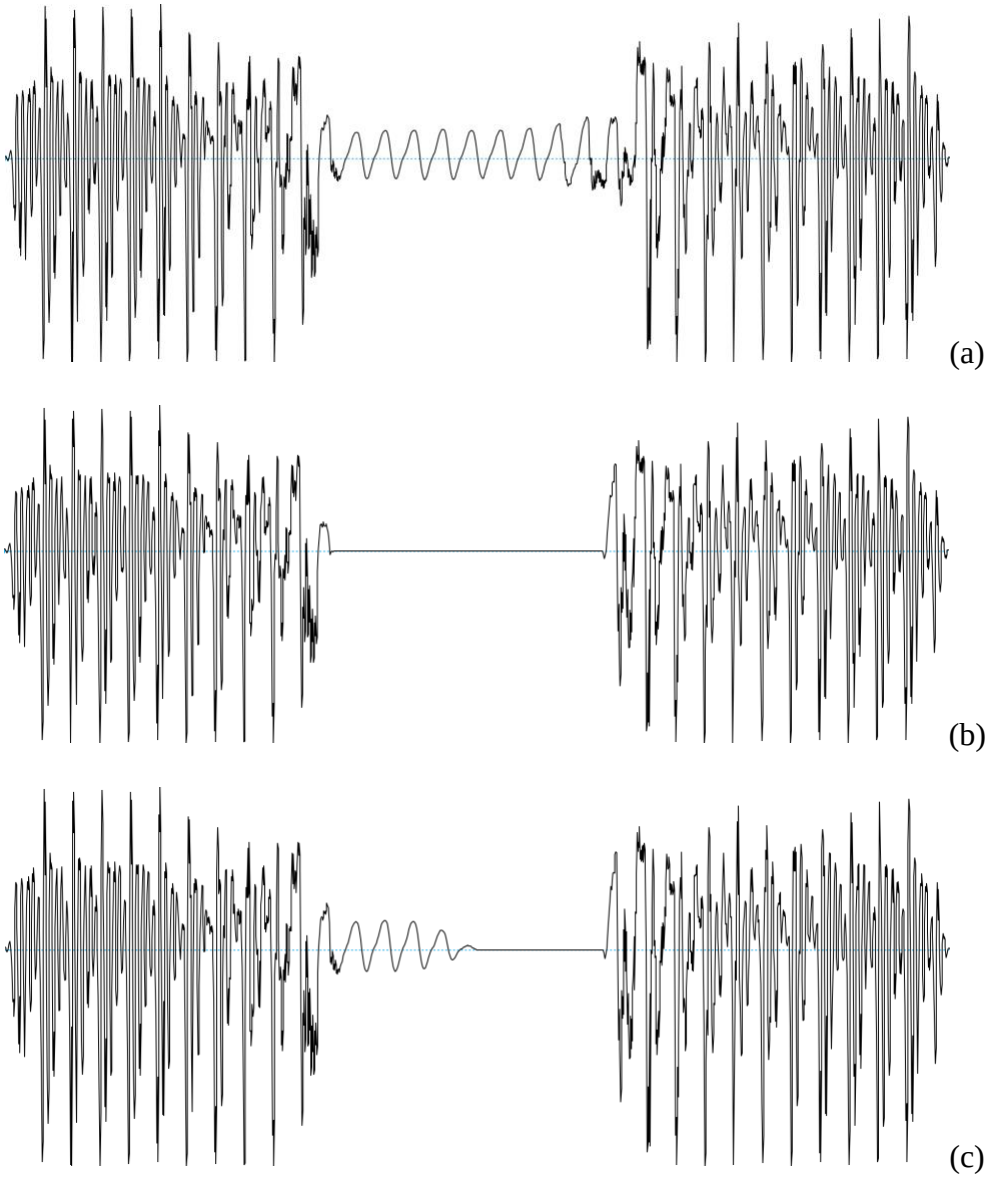


Figure 1: Illustration of the differences in voicing maintenance. Top shows presence of full voicing maintenance (100%), middle shows absence of voicing maintenance (0%), and bottom shows an intermediary value of 50%. Shown are the acoustic waveforms.

Production studies have shown that intervocalic phonologically voiced stops (/g/) can be produced with complete (100%) phonetic voicing throughout consonantal closure, with partial phonetic voicing, or with absent phonetic voicing (0%; Lisker & Abramson, 1964). The amount

of phonetic voicing required to produce a phonologically voiced perception varies by language and by individual (Oglesbee, 2008; Pape & Jesus, 2014a). For some languages (e.g. Spanish, Italian, French) a phonologically voiced /g/ will tend to be produced as phonetically voiced (high VM levels), and a phonologically voiceless /k/ will tend to be produced as phonetically voiceless (0% VM; Lisker & Abramson, 1964; Pape & Jesus, 2014b). Listeners of these languages likely rely perceptually on the presence or absence of this voicing maintenance to assess /k g/ sounds, thus assigning it high perceptual weight. In other languages (English, German, European Portuguese), the difference in voicing maintenance is not so straightforward between /k/ and /g/, and listeners must rely on other cues (Lisker & Abramson, 1964; Lisker, 1986; Pape & Jesus, 2014a). For these languages, the presence of phonetic voicing is typically an indication of a phonologically voiced phoneme, yet the absence of phonetic voicing does not necessarily signal a phonologically voiceless phoneme. Rather, phonologically voiced stops are produced with a range of VM values (0-100%). In other words, phonologically voiced /g/ can be produced with complete, partial, or absent phonetic voicing (devoicing), and a token with 0% VM could signal either /k/ or /g/. If a phonetically voiceless stop could perceptually indicate either /k/ (phonologically voiceless) or /g/ (phonologically voiced, but phonetically devoiced), then there must be other cues which serve to make the phonological distinction.

1.4.2. Length of preceding vowel and consonantal closure

Cross-language production studies have shown that there are significant differences in stop consonant duration (CL) and duration of the preceding vowel (VL) between phonologically voiced and voiceless stops in all the above described languages (Raphael, 1972; Francis, Baldwin, & Nusbaum, 2000; Pape & Jesus, 2014). The VL and CL values introduce two

additional perceptual cues which can work either independently or in combination with the others to distinguish the /k/ and /g/ phonemes in English. These length effects are fairly robust; a prototypical /g/ is produced with relatively high VL and low CL, and a prototypical /k/ is produced with relatively low VL and high CL (Pape & Jesus, 2014a). In other words, increasing CL suggests a phonologically voiceless /k/, and increasing VL suggests a phonologically voiced /g/. This phenomenon is quasi-universal, in that many languages find this pattern but few others are exceptions (e.g. Russian).

The interactions of these length effects with other perceptual cues are less well known. A recent study (Pape & Jesus, 2014a) explored the perceptual cue weighting for VM, VL and CL in European Portuguese (results below in **Figure 2**) using biomechanical stimuli. The figure shows the probability of voicing perception (y-axis) by the listeners, with VM presented on the x-axis, and grouped by VL (across the 3 panels) and CL (within each panel). The research found that increasing VL values leads to a relative increase in /g/ responses (change in panels from left to right) and increasing CL values leads to a relative decrease in /g/ responses (change in lines within each panel). However, these cues don't act independently of other cues: the perceptual length effects are mostly occurring at lower values of VM (below 50%). Increasing VM leads to an increase in /g/ responses for all levels of VL and CL, but at different rates for each level of VM. The effects are not stable and linear, but rather are greatly affected by the values of other present cues. At low values of VM, the length effects of VL and CL are much greater than they are at higher values of VM. These findings indicate that VM has its strongest effect when values surpass a certain threshold, which appears to be located around 25% VM. Subsequent modelling found significant main effects for VM and VL, but interestingly not for CL. Though CL did not have a significant main effect in itself, there was an influence of CL on VL responses, which

suggests relative lengths of VL/CL serve to play a role in perception, instead of their absolute lengths (Tuller & Kelso, 1984; Pape & Jesus, 2014a). Additionally, the results demonstrate a perceptual bias towards /g/ responses for the given biomechanical stimuli. These effects and interactions reveal interesting information about the perceptual cue-weighting which occurs in European Portuguese. Though each cue has its own main effect, the effect changes based on the values of other present cues.

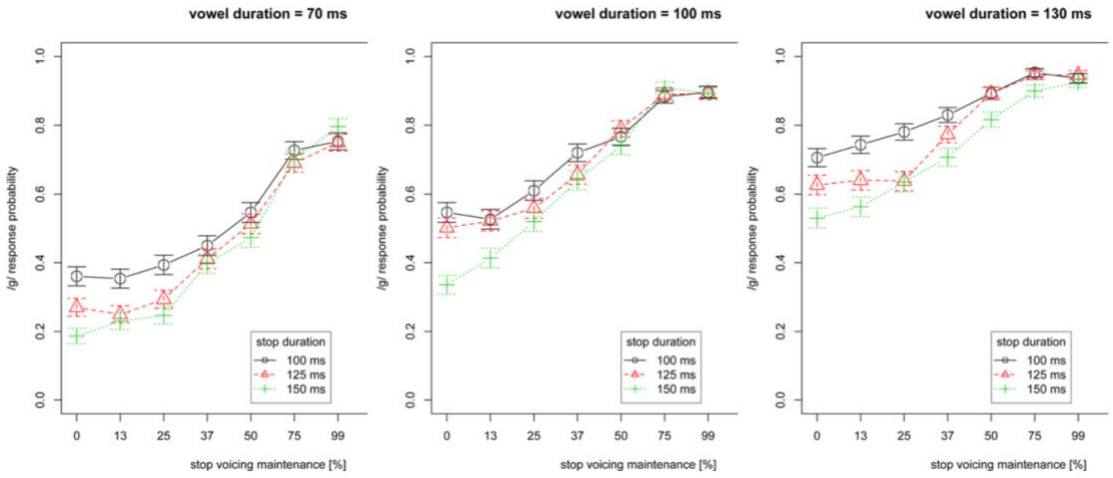


Figure 2: Experimental results from Pape & Jesus, 2014a. Each panel represents results for a particular VL value, while each line within the panels represent the results for a particular CL value. The x-axis represents the levels of the VM variable, and the y-axis shows the perceptual voicing response resulting from these variable levels.

1.4.3. Voice Onset Time

VOT is often cited as the main cue for phonological voicing distinctions in stop consonants, especially for languages such as English with devoicing in phonologically voiced phonemes (Lisker & Abramson, 1964; Abramson & Whalen, 2017). VOT was first defined by researchers at Haskins Laboratories (Lisker & Abramson, 1964; see also review in Abramson & Whalen, 2017) as the temporal delay between the release of articulatory closure (burst) and the onset of glottal pulsing (i.e. the voicing of the subsequent vowel). Generally, shorter VOT values are typical of /g/ productions, and perceptually indicate a higher probability of /g/ response. VOT

has been shown to be a universally reliable cue for phonological voicing distinctions in stop consonants across many of the world's languages with two- and three-way contrasts (explained below). Additionally, it can also distinguish between stops at different places of articulation (Lisker & Abramson, 1964; Nearey & Rochet, 1994), making it particularly useful as it categorizes consonants on two dimensions in one single value. Instead of requiring multiple information points to characterize stop consonants by place of articulation and by voicing, VOT offers a computationally simpler alternative with only one value.

When VOT is systematically varied between values for phonologically voiced and voiceless consonants, an interesting phenomenon is observed. The perceptual response to VOT increases does not show a linear relation with perception of voiced or voiceless stop consonants. Rather, there is an abrupt boundary were perception changes categorically from voiced to voiceless. On either side of the boundary perception remains relatively stable. In English, a wide range of low VOT values will signal perception of the voiced /g/, and a wide range of high VOT values will signal perception of the voiceless /k/. At an intermediary location between these two ranges of VOT values there will be a boundary where perception rapidly changes from one category to the other (steep slope). This phenomenon is known as categorical perception (Harnad, 2003), and a typical categorical perception curve is shown below in **Figure 3**.

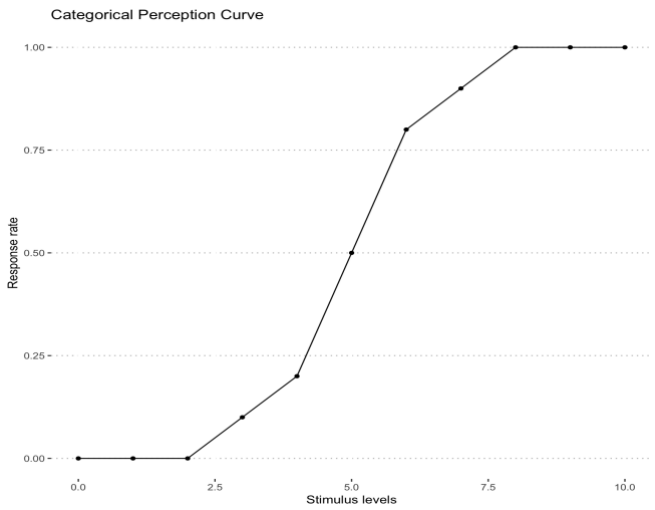


Figure 3: A typical/generic example of a categorical perception curve. Note that response values (y-axis) have sections which remain relatively stable (each section represents a perceptual category), and an intermediary boundary where response changes abruptly from one category to the next.

This phenomenon means that our perceptual system allows for a significant degree of acoustic variation and can place VOT productions with several different values into a single perceptual category with small perceptual differences inside categories. It also means that there is a small VOT value range surrounding the boundary where perceptual responses could fall into chance, due to the ambiguous nature of the signal. This is a general psychophysical response, specific to the nature of VOT.

Although VOT is a universally practical measure, the exact values which identify particular stop consonants vary among languages. The specific boundaries which distinguish /k/ and /g/ vary, and in fact fall into four possible ranges (Lisker & Abramson, 1964). The VOT categories include short-lag, long-lag, voicing-lead and voiced aspirated (the latter being relatively rare among the world's languages). Furthermore, languages differ by number of phonemic contrasts within the category. For instance, while English has a 2-way contrast between /k/ and /g/, some languages have 3-way contrasts (Korean; /g k kʰ/) or even 4-way contrasts (Hindi; /g gʰ k kʰ/). Two categories are most common in the world's languages,

languages seldom have 4-way contrasts due to the rarity of the voiced aspirated category (Abramson & Whalen, 2017). Interestingly, among 2-way contrast languages, rather than selecting categories which are most perceptually distinguishable (e.g. furthest values of VOT: voicing-lead and long-lag), they seem to select phonetically adjacent categories (either voicing-lead/short-lag, or short-lag/long-lag). The focus of this study is perception of English, therefore the summary of factors not relevant to English will be brief. The VOT categories which occur in English stop consonant productions are illustrated in **Figure 4**. The top panel shows voicing-lead, with continuous voicing maintenance throughout consonantal closure; the middle panel shows short-lag (VOT values of roughly 30ms or less), which has no voicing maintenance or aspiration; the bottom panel shows long-lag (VOT values of roughly 30ms or more), with no voicing maintenance but with aspiration.

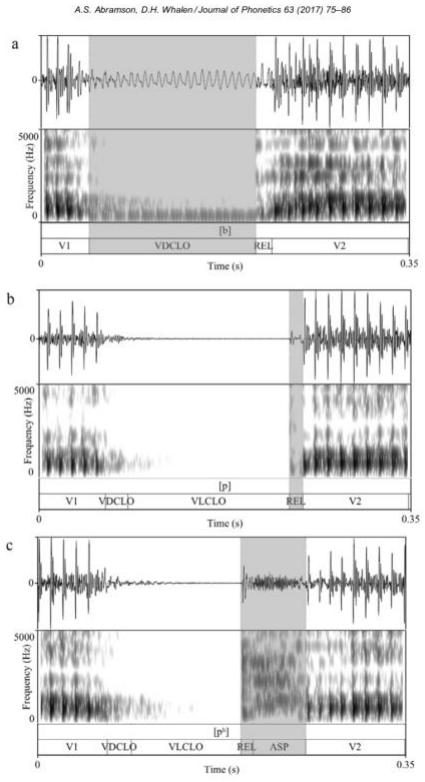


Figure 4: Figure from Abramson & Whalen, 2017. It displays the three categories of VOT present in English stop consonant productions: long-lag aspirated (bottom), short-lag unaspirated (middle), voicing-lead (top).

Though languages may have phonemes in common, they may differ with regards to which of these perceptual VOT categories serve to categorize each phoneme. For instance, in Spanish, /g/ VOT values fall into the voicing-lead category and values for /k/ into the short-lag category, while in English /g/ typically falls into the short-lag category (though it can be produced also with voicing-lead) and /k/ into the long-lag category (Lisker & Abramson, 1964). The perceptual boundaries between /k/ and /g/ differ between the languages, and a value considered ambiguous in one language may be prototypical of a sound in another. Each language has its own, different, perceptual boundary between /k/ and /g/.

1.5. Purpose

This thesis describes three experiments involving the cue-weighting of perceptual cues for voicing distinctions of intervocalic velar stops /k/ and /g/ in English. The ultimate goal is to identify and describe the effects, interactions and cue-weighting of the perceptual cues: VM, VL, CL, and VOT. The experiments use stimuli created by biomechanical modelling with cue values extracted from real speech production values.

The thesis focuses on the cues VOT, VM, VL and CL. Formant transition cues were excluded because they tend to differentiate stop consonant pairs by place of articulation while remaining the same across voicing for the same place of articulation (i.e. same for /g/ and /k/). The cue of fundamental frequency (*f0) at the onset of the following vowel* was excluded because it was not reproducible by biomechanical modelling, and its inclusion would have increased the stimulus complexity beyond its reasonable limit. The first experiment used stimuli without a burst typical of the stop consonants. As mentioned, stop consonants are typically produced with the burst signal, but they can also be produced with no discernable burst, and thus no reliable

signal for VOT. In these cases, it is useful to observe how the effects of the perceptual cue are shifted to accommodate the absence of VOT to continue to produce a robust perceptual experience. The goal was to examine how the perceptual cue of VM influenced voicing perception with VOT absent; CL and VL were added to see the interactions. These results were contrasted to those of EP listeners from Pape & Jesus, 2014a, who describe a similar identification task using the cues VM, VL and CL.

A second experiment determined the average perceptual ranges of VOT for /k/ and /g/ (for the given biomechanical stimuli) and examined where the perceptually ambiguous VOT values (between /k/ and /g/) are. Additionally, the experiment sought to find out how varying VOT values influence voicing perception. This was performed for stimuli with prototypical length cues (CL, VL & VM values representative of either /g/ or /k/) and stimuli with ambiguous length cues (intermediary CL, VL & VM values). This experiment explored not only the effect of VOT on voicing perception, but also how VOT interacts hierarchically with the other cues.

The third experiment incorporated the newly extracted ambiguous VOT value (from experiment 2) as an additional cue with the other manipulated cues VM, VL and CL. This stimulus was similar to that in experiment 1 but with an added burst signal (at ambiguous VOT values). This experiment examined the perceptual cue-weighting of the stimuli when a VOT value was present compared to experiment 1 results when VOT was absent.

The hypothesis was that VOT should have a strong effect on voicing perception when included, and its inclusion as an ambiguous value should cause the perceptual cue weighting to change from experiment 1 to experiment 3. Since phonological voicing distinctions are often made in the absence of appropriate phonetic voicing (VM) information, it was hypothesized that this variable would have the least effect for English listeners. The length effects of VL and CL

are well reported and should show a relatively strong effect. Specifically, we predicted VOT would have the most influence on perceptual voicing responses, followed by the effects of VL & CL, and finally VM would have the least effect.

Chapter 2: Experiment 1: Cue Weighting in the Absence of VOT

Experiment 1 investigated the cue weighting and hierarchy of three acoustic cues (VM, VL and CL) in the absence of a cue for VOT. The goal was to examine how the perceptual cue of VM influenced voicing perception with VOT absent; CL and VL were added to see the interactions.

2.1. Methods

2.1.1. Participants

Thirty-eight native English-speaking participants were recruited through the Linguistic Research Participation System (SONA) administrated by the Department of Linguistics and Languages at McMaster University. Speakers were selected as native English speakers if they learned English before the age of 5. All participants were undergraduate students (aged 17-24) who reported normal vision and hearing, and they received course credit for their participation.

2.1.2. Stimuli

The stimuli presented in all experimental conditions consisted of audio files generated by biomechanical modelling – an accurate way of modelling natural speech while controlling for various cues. Biomechanical modelling allows independent manipulations of vocal tract settings while maintaining realistic speech transitions. Following the steps outlined in previous work (Pape, Jesus, & Perrier, 2012; Pape & Jesus, 2014a), a physically realistic prototype stimulus /aCa/ where the central consonant (C) was a velar stop was created by biomechanical modelling and synthesized with a three mass vocal fold model. Manipulations varied the values for length of the central consonant (CL), the length of the preceding vowel (VL), and the amount of voicing throughout consonantal closure (VM). This ultimately produced stimuli of two speech sound

segments /aga/ and /aka/, which differ only by (1) their dimension of stop voicing, and (2) the length effect values (CL & VL) given to them. All other articulatory and acoustic parameters were exactly identical across all synthesized tokens.

The stimuli varied based on three acoustic cues: length of preceding vowel (VL; 70ms, 100ms, 130ms), length of consonantal closure (CL; 100ms, 125ms, 150ms), and the percent of voicing maintenance throughout the consonantal closure (VM; 0%, 25%, 50%, 75%, 100%). The levels chosen for each variable correspond to the extreme values (and their mean values) as reported in production studies for European Portuguese (Pape & Jesus, 2014a; Pape & Jesus, 2014b), and relevant literature for English perception (Raphael, 1972; Francis, Baldwin, & Nusbaum, 2000; Oglesbee, 2008). A pilot study confirmed that the extreme/outer values for each acoustic variable match the perceptual and production ranges for English speakers.

In total, there were 45 items (3x3x5). Of note is that they were all characterised by the absence of a typical burst during stop consonant production, thus removing the cue for VOT. Additional cues such as F0, and other speaker-specific factors were held constant across all stimuli to avoid the influence of these parameters of the voicing distinction, and to specifically control for the three variables of interest. The absence of natural imperfections and individual features causes the biomechanical stimuli to sound slightly unnatural. However, the level of precise control for each variable is only attainable with this type of advanced biomechanical modelling, as using natural speech stimuli would introduce many perceptual artifacts (i.e. confounding dimensions). **Figure 5** (from Pape & Jesus, 2014a) illustrates how the acoustic stimuli vary according to these parameters. Factor 1 refers to the changes in the duration of the consonantal closure (CL); factor 2 refers to the changes in the duration of the preceding vowel

(VL); factor 3 refers to manipulations made to the voicing maintenance throughout consonantal closure (VM).

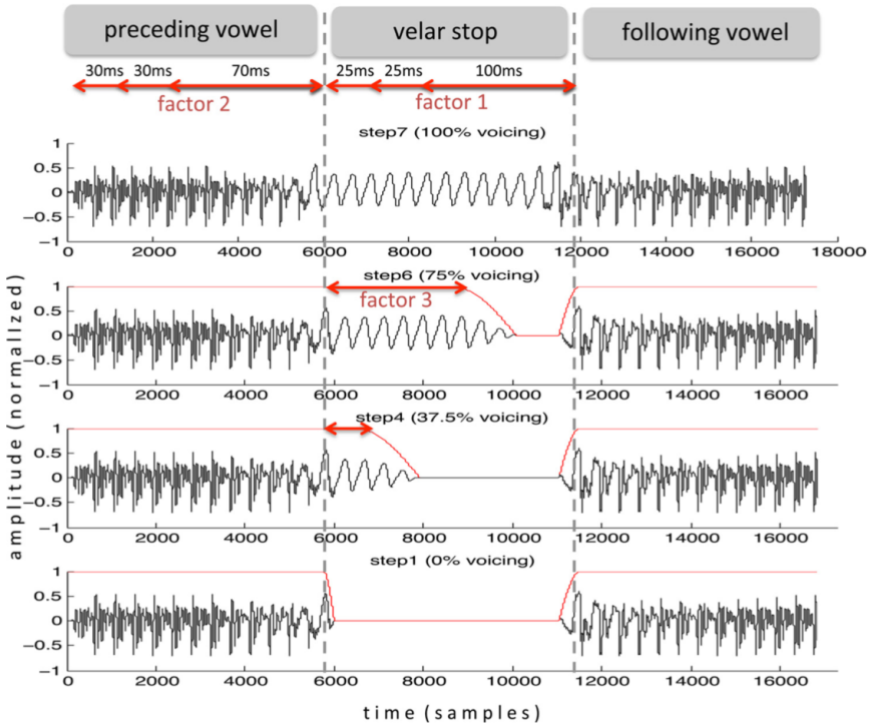


Figure 5: Figure from Pape & Jesus, 2014a. Manipulations made to control variable levels in stimuli. Factor 1 = VM; Factor 2 = VL; Factor 3 = CL.

2.1.3. Procedure

The experiment was designed to determine the significant effects of the examined acoustic cues and their interactions on the perception of stop voicing, all in the absence of the main cue of VOT, classically considered to be dominant. All combinations of the manipulated variables were tested against each other.

Participants were placed in a soundproof room located in the Phonetics Lab in the ARiEAL research centre at McMaster University. The software interface used to present the experiment was Alvin 2 (Hillenbrand, Gayvert, & Clark, 2015) on a laptop computer with all

unrelated processes disabled. Stimuli were presented through a pair of Sennheiser HD 598 headphones with linear frequency response, connected to a Focusrite Scarlett 2i2 audio interface. Instructions were given verbally with a printed guideline. Participants were informed that their task was to listen to the synthetic /aka/ and /aga/ stimuli and indicate which they perceived: i.e. to perform a binary forced-choice identification task based on phonological consonant voicing. Responses were made by clicking on the corresponding button in the experimental interface, shown in **Figure 6**. There was no time limit to the trials, but participants were instructed to answer as quickly and accurately as possible. After each response was made, the next trial began after a 1500ms interval. There were 15 repetitions per stimulus, for a total of 675 trials. Trial repetition was not possible. The experiment lasted on average 30 minutes, including a short practice session of 10 trials.

Figure 6: Screenshot of the experimental interface. Same for all experimental conditions.

2.2. Results

Of the 38 participants, 16 were eliminated for less than adequate response patterns (examples illustrated in **Figure 7**). These included responses not showing substantial deviation from chance (i.e. approaching 50% random response pattern; left panel), response patterns that selected only one option for every single response (right panel), or participants who did not complete the procedure properly or to completion.

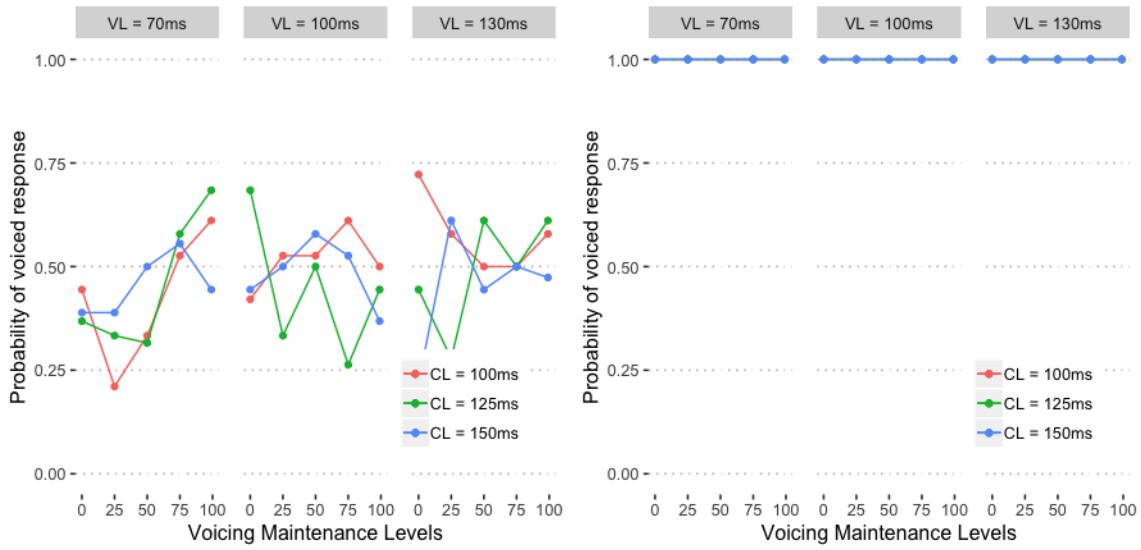


Figure 7: Examples of excluded participant responses. Left panel shows 50% (chance) response rate; right panel shows 100% response rate.

In total, thus data from 42% of the initial participants was excluded. A higher than average rate of random responses was likely due to the difficult nature of the experiment, because the discernible burst – and therefore main VOT cue – was absent from the signal. This number of excluded participants is higher also when contrasting with a similar experimental procedure involving European Portuguese listeners (Pape & Jesus, 2014a). In their experiment they reported 7 excluded participants from a total of 38 (18.4%). The relatively higher number for English listeners could perhaps be due to their increased tendency to favor the perceptual cue

of VOT and their lower sensitivity to voicing (VM) and duration cues (CL and VL), thus making the task extremely difficult for these participants.

Trials with reaction time values greater than 2.5 standard deviations from each participant mean were excluded from the analysis. The data for the remaining useable trials is illustrated in **Figure 8**, contrasting effects across all acoustic dimensions in a 3x3 matrix. The figure shows the probability of voicing perception by the listeners (y-axis), with VM presented on the x-axis, and grouped by VL (across panels) and CL (within panels). From this figure several trends can be observed. The most striking observation is the clear perceptual preference for /g/ responses, with the probability of a /g/ response never going below 25% for any condition. Increasing VM has a clear effect of increasing the probability of a /g/ response, for all levels of VL and CL. Interestingly, this increase is not linear, but rises more steeply for values of VM between 25-50% and seems to taper off (i.e. ceiling-effect) around 75% VM. An increase in VL (panels from left to right) causes a slight increase in /g/ responses, while an increase in CL (lines within each panel) seems to have little or no effect.

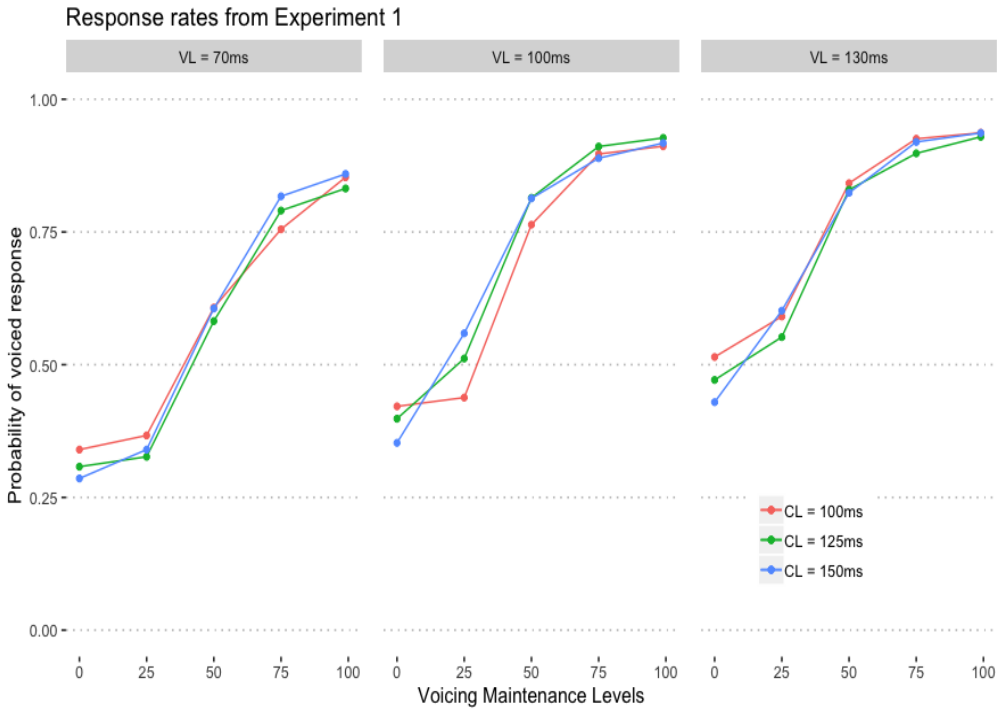


Figure 8: 3x3 matrix displaying response patterns grouped by variable levels. The x-axis for every panel represents the changes in voicing maintenance (VM), and the y-axis represents the perceptual response rate (% /g/ response). Panels from left to right show changes in VL, and lines within each panel show changes in CL.

Figure 9 (the curve for CL = 125ms from the middle panel above, VL = 100ms) shows the effect of VM when the length variables are held at ambiguous values (i.e. the mid points for all duration values of the stimuli; CL = 125ms and VL = 100ms). Notably, the probability of /g/ response approaches chance (50%) when VM = 25%. From these results we extrapolate that an ambiguous VOT value would be best determined when the other acoustic cues of vowel length, consonant length and voicing maintenance have values of VL = 100ms, CL = 125ms, and VM = 25% (approximately). This will be of relevance in the next chapter.

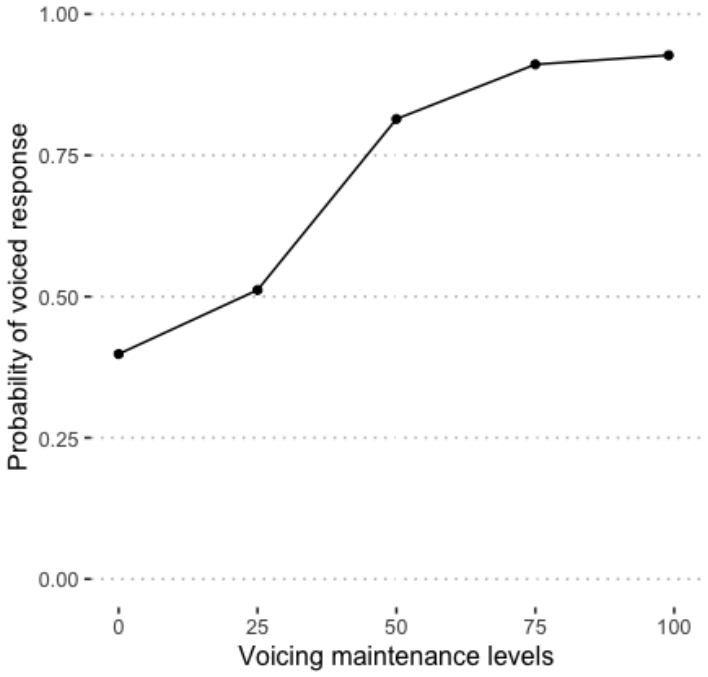


Figure 9: Effect of VM when length duration held at ambiguous values (CL = 125ms, VL = 100ms). Note that 50% response rate (chance) is reached when VM = 25% (approximately).

2.2.2. Model (Main Effects, Interactions)

The analysis for the significance of participant response patterns towards voicing perception was performed by generalized linear mixed-effects modelling, function glmer (Bates, Maechler, Bolker, & Walker, 2015) in the R environment (R Core Team, 2018). A binary logit model was used to best account for the nature of the dependent variable (a binary choice between /g/ and /k/) (Baayen, 2008). The model was worked down step-wise by complexity until the most complex model with lowest AIC (Akaike's Information Criterion) was obtained. AIC is a measure of the model's fit to the data, with lower values indicating better fit. As per Burnham, Anderson, & Huyvaert, 2011, this is the best model to report when desiring to weight the importance of each variable to the model. Rather than favouring more complex models as many

approaches do, this approach examines the model fit while simultaneously penalizing model complexity.

We tested the statistical validity of the above observations with the /g/ or /k/ binomial response as the dependent variable. Fixed effects of VL (3 levels: 70, 100 and 130ms), CL (3 levels: 100, 125 and 150ms), and VM (5 levels: 0, 25, 50, 75 and 100%) were tested with a significance threshold of $p < 0.05$. *Subject* was selected as a random factor and was included with random intercepts and random slopes for VM and VL, as these led to the most complex model with the lowest AIC. An increase or decrease in the complexity of this model led to an increase in AIC. A likelihood ratio test was performed comparing the full model to a null model, composed of only the random factors and not including the three predictor variables of interest. The full model was significantly different from the null model ($chi_2 = 63.32$, $p < 0.001$), indicating that the main effects and interactions of the predictor variables were significant. In the full model, the factors VM ($z = 11.991$, $p < 0.001$) and VL ($z = 4.994$, $p < 0.001$) had significant main effects, but the effect of CL ($z = 0.559$, $p = 0.575964$) failed to reach significance. The main effects are illustrated in **Figure 10**.

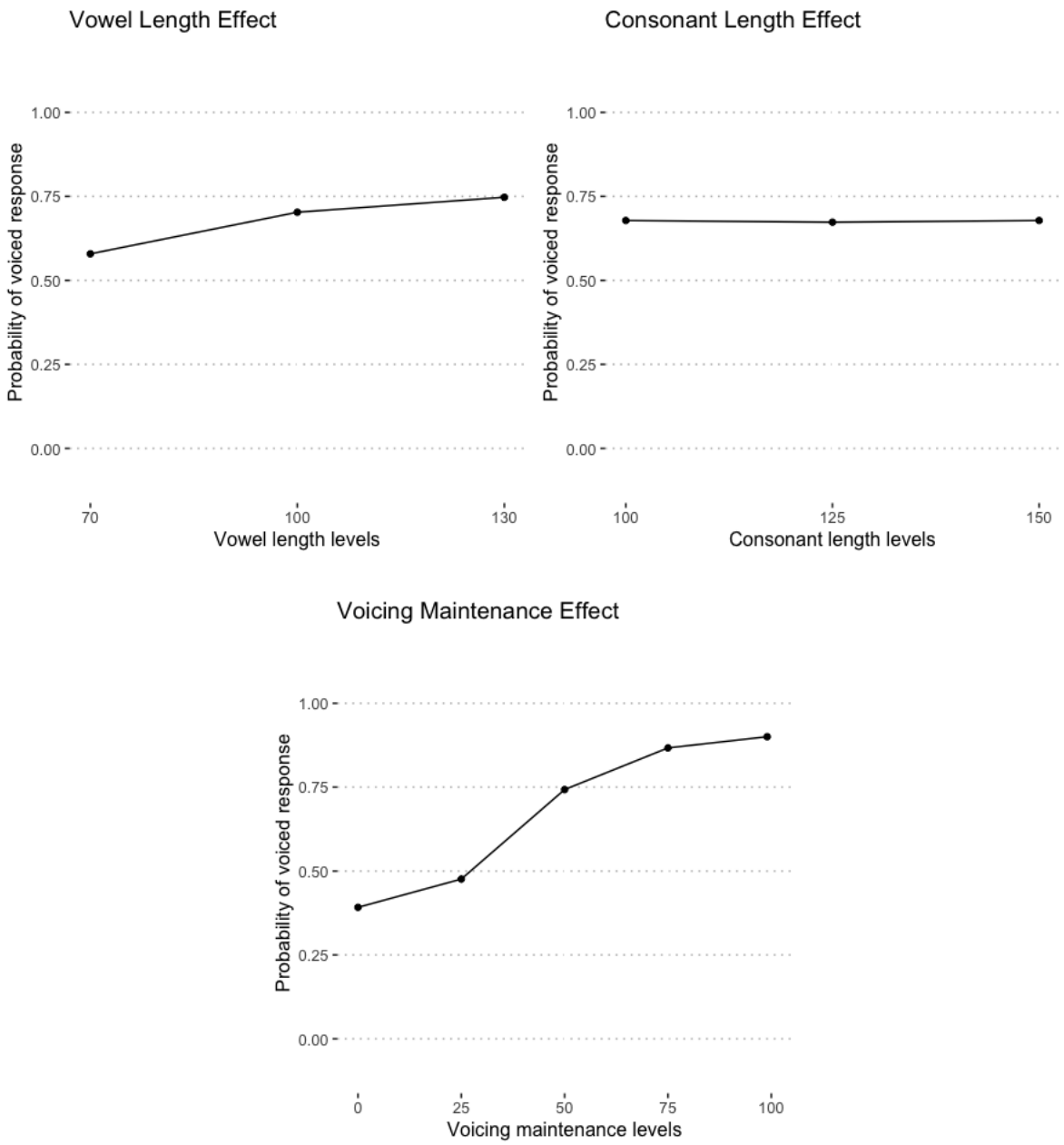


Figure 10: Main effects for the cues of VL (top left), CL (top right), and VM (bottom). The x-axis represents the change in the respective variable, and the y-axis shows the subsequent change in response rate.

The only significant interactions were VL:VM ($z = 3.365$, $p < 0.001$) and CL:VM ($z = 2.331$, $p < 0.05$), illustrated in **Figure 11**. For the VL:VM interaction, it appears that VL has a greater effect at VM levels between 25-50% (indicated by a steeper slope for these VM values). For the CL:VM interaction, CL appears to have a very minimal effect (horizontal/zero slope),

except at VM levels of 0% where there is a decrease /g/ response and of 25% where there is a slight increase. These results indicate that VM has the relatively stronger perceptual effect, and that CL and VL mainly take over at low levels of VM (i.e. below 50%). Significance and z-scores for all main effects and interactions are listed in **Table 1.** Note that while CL has no statistically significant main effect in itself, its influence is clear when considering its interaction with VM.

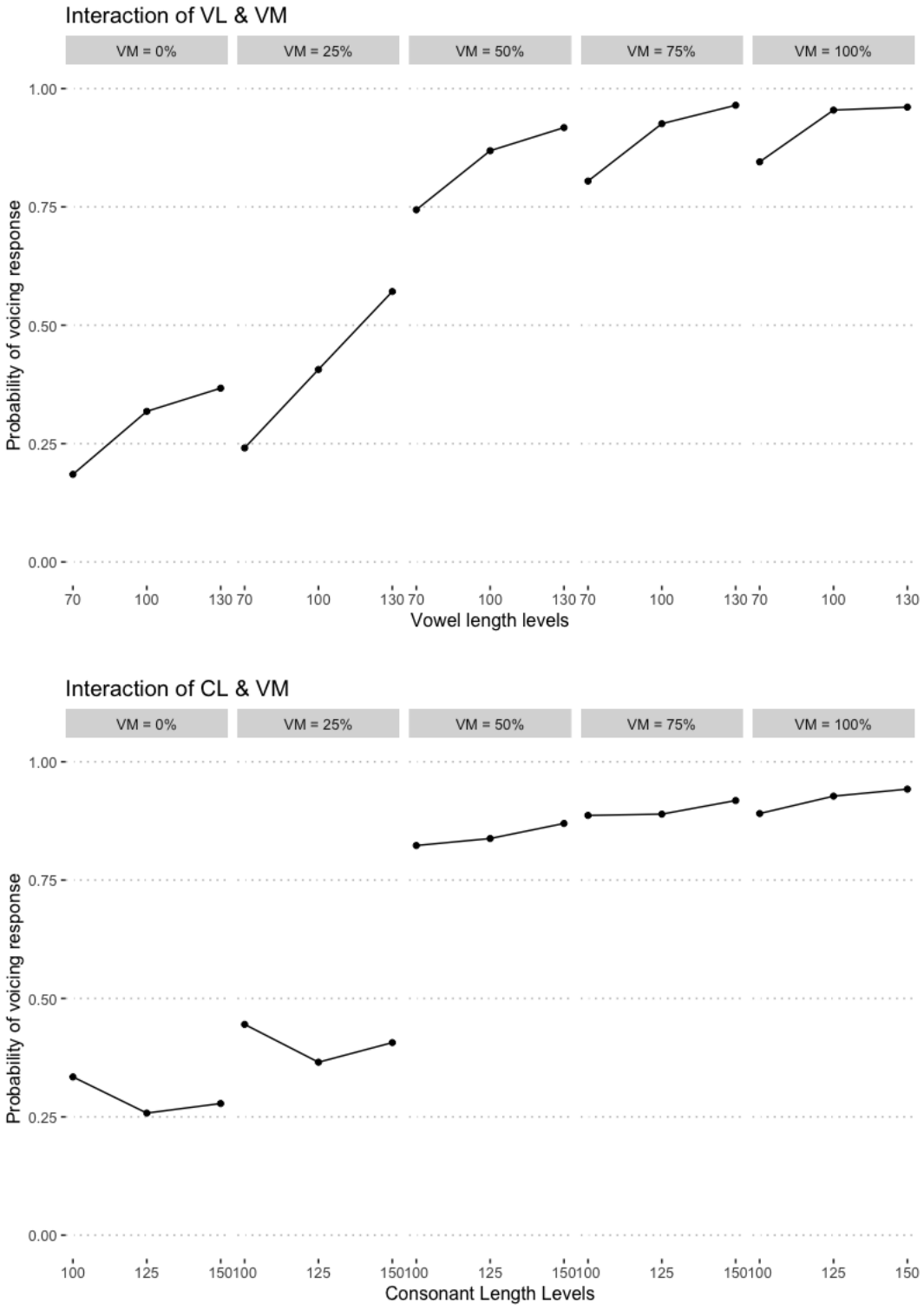


Figure 11: Significant interactions of VM:VL (top) and VM:CL (bottom). Note that although CL has no significant main effect for the whole dataset, it still has an influence when considering its interaction with VM.

Effect/Interaction	Estimate	Std. Error	z value	Pr(>|z|)
VL	0.57859	0.11585	4.994	5.90e-07 ***
CL	0.01235	0.02209	0.559	0.575964
VM	1.34878	0.11248	11.991	< 2e-16 ***
VL:CL	-0.01546	0.02207	-0.701	0.483551
VL:VM	0.08015	0.02382	3.365	0.000765***
CL:VM	0.05360	0.02299	2.331	0.019736 *
VL:CL:VM	-0.01228	0.02288	-0.537	0.591454

Table 1: All main effects and interactions from the statistical modelling of experiment 1 results.

2.2.3. Cue weighting

Weighting of predictor variable importance to the model was performed following Burnham and colleagues (Burnham, Anderson, & Huyvaert, 2011). The process involves scaling down the model by each desired interaction and main effect and comparing each resulting AIC to that of the original full model (AIC = 14928.53 for the full model). The difference in AIC computes what is referred to as ΔAIC and is a measure of the importance for each predictor variable (the relevant scales are presented in the paper). Removal of VM caused a ΔAIC of 11.66, removal of VL led to ΔAIC of 6.04, and removal of CL led to ΔAIC of 1.06. These results are summarized in **Table 2.** This analysis method confirms what can be seen visually in the data, that VM has the strongest effect, followed by VL, and finally CL with the lowest effect.

Cue	**AIC**	**ΔAIC**	**Cue Ranking**
VM	14940.2	11.66432	1
CL	14929.59	1.058651	3

VL	14934.58	6.042195	2

Table 2: Cue weighting results from experiment 1.

2.3. Discussion

Many of the results of this experiment are congruent with those of Pape and Jesus (2014a) and will be discussed together. In both experiments the response patterns indicate a perceptual bias for /g/. This is likely a symptom of the biomechanical modelling and the specific manipulations made to the stimuli. The absence of a discernable burst from the signal, and thus the VOT cue, as well as aspiration and *f0*, are likely to have influenced voicing perceptions towards /g/. However, the difference in native languages by the participants has some perceptual significance, in that European Portuguese listeners showed little difficulty in identification (18.4% showed inadequate response patterns) while English listeners had considerably more trouble (42%).

The relatively higher number of removed English listeners is likely due to their reliance of *f0*, which was absent. Additionally, English listeners have an increased tendency to favor the perceptual cue of VOT and be less sensitive to voicing (VM) and duration cues (CL and VL), which creates a difficult task often leading to random response rates. English is one of the aforementioned languages where phonetic voicing during the production of stop consonants tends not to match with phonological voicing (i.e. devoicing of /g/). Phonologically voiced stops can be produced with a range of VM values, including 0%. For this reason, it is assumed that VM would have significantly less of an effect on voicing perception than the other cues of VOT, VL and CL. The number of English listeners who are insensitive to the manipulated cues is an indication of the strong influence of VOT (and the consequence of its removal).

The effects of VM are similar in both the results from this experiment and those of Pape and Jesus (2014a), serving to increase probability of voicing perception as it increases in level. The VM effect is interesting because it appears to show its own form of categorical perception. Low VM values (below 25%) indicate lower probability of voiced responses, while high VM values (above 75%) indicate higher probability of voiced responses. There is a point around 25-50% where perception changes categorically from /k/ to /g/. This response pattern for VM is expected (higher VM levels indicating voiced responses) for some languages (as in European Portuguese) but the strength of its effect is surprising, since English listeners tend to focus on other cues, such as VOT, in making their perceptual voicing responses. Additionally, it was previously discussed that low VM values could be interpreted ambiguously between /k/ and /g/ by English listeners.

The robust effect of VL is clear when looking across the panels in **Figure 8.** The shape of the plot remains very similar, but perception is shifted up towards the voiced response for each level of VL. The figure indicates the lack of influence of CL (which failed to reach statistical significance), in that all plots within a panel (where only differences between data points are by levels of CL) are essentially identical. This is quite surprising, since the effect of CL is well reported in English (Raphael, 1972; Francis, Baldwin, & Nusbaum, 2000; Pape & Jesus, 2014). These main effects are illustrated in **Figure 10** and correspond to the European Portuguese responses reported by Pape and Jesus (2014a). Their model also failed to reach significance for the CL cue. The interactions (as shown in **Figure 11**) reveal interesting information about the effect of VM. At low levels of VM (below 50%) the effects of VL and CL appear to have much more influence than at high levels of VM. The effect of VL is much more prominent at low levels of VM (below 50%) and tapers off when VM reaches 75%. More importantly, CL also

appears to have some effect at low VM values (0%). Recall that increasing CL values are reported to favor voiceless (/k/) perception in the literature.

The cue-weighting analysis indicated VM to have the largest influence, followed by VL and CL. Indeed, it appears that while VM was the main influence on voicing perception, the length cues VL and CL noticeably increase their effects when voicing maintenance values are low. These results are in line with relevant literature regarding Canadian English perception. Phonologically voiced /g/ can be produced with a range of VM values, but the voiceless /k/ is rarely produced with any voicing maintenance. For this reason, high values of VM (above 50%) would indicate /g/ perceptually, but the lower values of VM are more ambiguous. Combined with these low VM values the remaining cues (in this case VL and CL) take over to provide a more accurate perceptual response.

Chapter 3: Experiments 2a, 2b and 2c: Ambiguous VOT Value and the Perceptual VOT Boundary

This chapter will review the results from three between-subjects groups that were studied (Experiments 2a, 2b and 2c). Experiment 2a was designed to determine the perceptual boundary of velar stops for English speakers in the context of ambiguous length cues (VL and CL) and varying voicing maintenance. Experiments 2b and 2c were designed to determine if these perceptual differences were primarily attributable to length effects or voicing maintenance differences. For this reason, stimuli with prototypical values for both /g/ and /k/ productions were used with varying voicing maintenance levels.

3.1. Methods

3.1.1. Participants

Thirty-two native English speaking participants were recruited through the Linguistic Research Participation System administrated by the Department of Linguistics and Languages at McMaster University. Speakers were considered as native English speakers if they had learned English before the age of 5. All participants were undergraduate students (aged 17-24), reported normal vision and hearing, and received course credit for their participation. All 32 participated in Experiment 2a, 7 also participated in Experiment 2b and 15 in Experiment 2c (to be explained in the next section).

3.1.2. Stimuli

In this experiment, voice onset time (VOT) was introduced as a systematically changing variable for the purpose of determining a perceptual boundary. VOT values of 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110 and 120ms were included (the higher values were included when

permitted within the length of the consonantal closure, e.g. CL of 100ms cannot accommodate a VOT of 120ms). These values fall within the production and perceptual ranges for English as confirmed by literature (Lisker & Abramson, 1964; Abramson & Whalen, 2017) and pilot studies. Three new sets of stimuli were created from the original set of stimuli from Experiment 1, one for each of the new Experiments 2a, 2b and 2c (explained further below).

The VOT signal added to the original stimuli was selected from a pool of recorded tokens produced by native Canadian-English speakers in a soundproof room. The best tokens were selected based on signal quality and VOT values. The recording was made on a notebook computer using a Shure SM-27 microphone connected to a Tascam US-122MKII audio interface with a sampling frequency of 44,100 Hz and 16 bits. A sample of real English words (e.g., "escort") designed to elicit observable bursts & VOT measures and to produce un-aspirated /k/ and /g/ prototypes were used. When a voiceless stop consonant (/k/) is preceded by a fricative (/s/), it is produced with little or no aspiration. The unaspirated prototypes were selected because the presence of aspiration is a perceptual cue in itself. In English /k/ can be heavily aspirated in certain phonotactic positions, while /g/ is produced with little or no aspiration.

The recorded stimuli were processed and manipulated in the phonetic software Praat (Boersma & Weenink, 2019) and in the audio processing software Audacity (Audacity Team, 2019). The VOT measure from the selected originally recorded sound was 32ms, similar to values reported in literature for English velar stops (i.e. Lisker & Abramson, 1964; Nakai & Scobbie, 2016; Chodroff & Wilson, 2017). The original recording used for the VOT manipulations is illustrated in **Figure 12** (spoken word is "escort")**.**

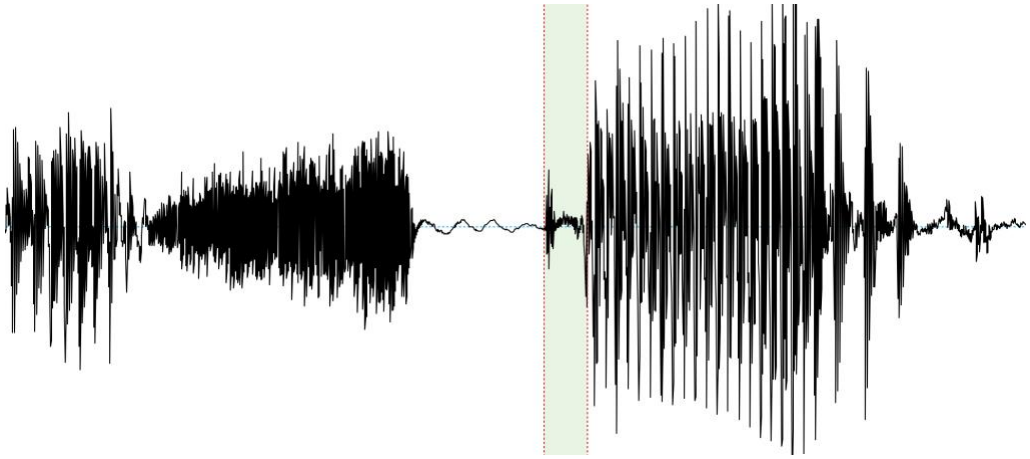


Figure 12: Illustration of original VOT recording. Spoken word is "escort". VOT segment from the consonantal burst to the onset of vowel vibration is highlighted.

To produce the desired range of VOT values, the selected original recording was temporally manipulated[2] in Praat. Rather than selecting only the VOT section for manipulation, the entire word recording was used in order to preserve the acoustic context of the VOT segment. In the resulting VOT audio files, the peak amplitudes of each of the bursts were normalized to half the amplitude of the vowel signal to match typical production values. The addition of VOT is illustrated in **Figure 13** by comparing the original stimuli from experiment 1 (without VOT) to stimuli from experiment 2 with manipulated VOT values. Stimuli with fixed values for the other cues are shown to emphasize the change in VOT (VM = 0%, CL = 125ms, VL = 100ms).

[2] Temporal manipulation in Praat by PSOLA algorithm to increase or decrease the duration of each sound file by a calculated ratio in order for its VOT segment to match the desired VOT duration outcomes (i.e. 5-120ms). The calculation for the manipulated change in duration of the entire sound file was made by dividing the desired VOT value by 32 (the original VOT duration). This provided a ratio between the original file duration and the duration required for each desired VOT value.

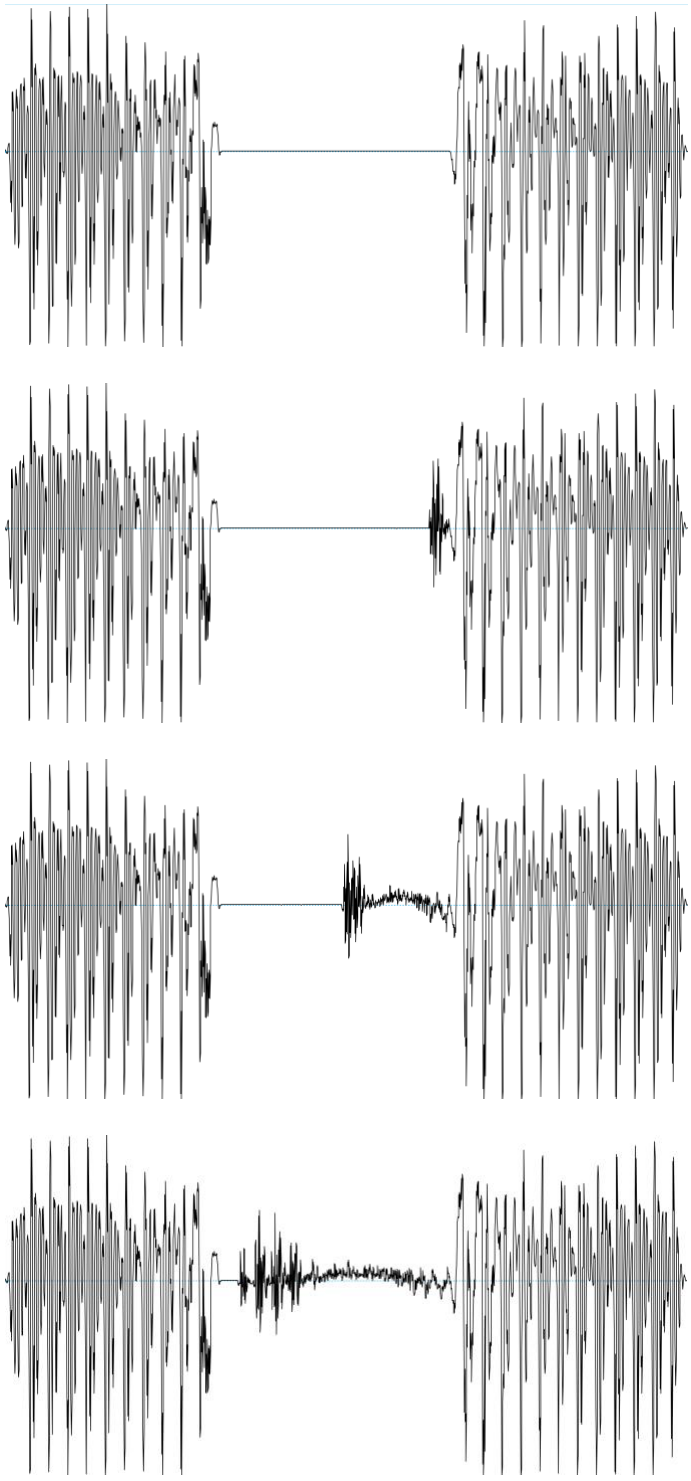
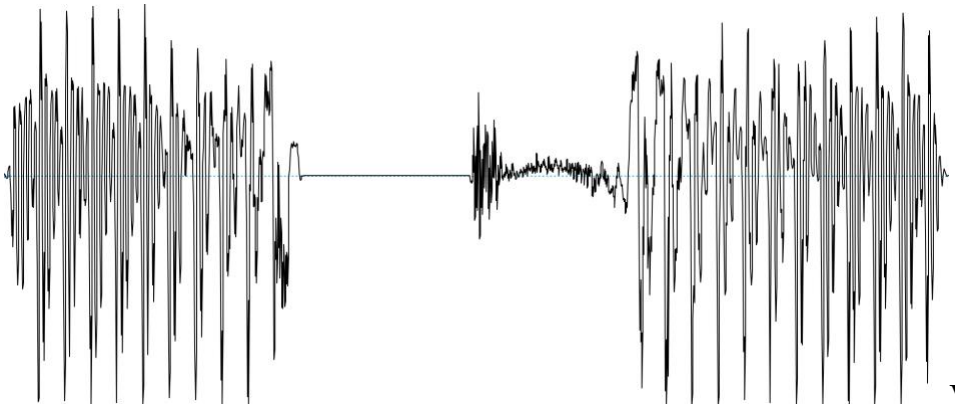


Figure 13: Illustration of the VOT manipulations in the stimuli. Top panel shows original stimulus without VOT, second panel shows 10ms VOT, third panel shows 50ms VOT, and final panel shows 100ms VOT.

In the *Ambiguous* sub-experiment (2a), stimuli with ambiguous VL and CL cues (middle values, e.g. VL = 100 and CL = 125) were selected at VM levels of 00, 50 and 100%. A burst with varying VOT measures from 05ms to 100ms was mixed into the signal. These stimulus values were chosen to investigate how varying VOT influences voicing perception when the length variables are held constant at ambiguous values (CL = 125ms, VL = 100ms) and are consequently neutralized and made uninformative as perceptual cues. The three levels of voicing maintenance (VM = 0, 50, 100%) were chosen to determine whether the effect of VOT or the change in VM was more important in determining the perceptual voicing response. These results would provide an insight into generating perceptually ambiguous VOT values (between /k/ and /g/) and determining the average perceptual ranges of VOT for /k/ and /g/ for the given biomechanical stimuli. The total number of different stimuli was 33 (1 x 1 x 3 x 11 = 33). **Figure 14** illustrates the VM changes (CL and VL are held constant at their middle values) while keeping VOT at 50ms. Note that the VOT in the experimental stimuli ranges from 5-120ms.



VM = 0%

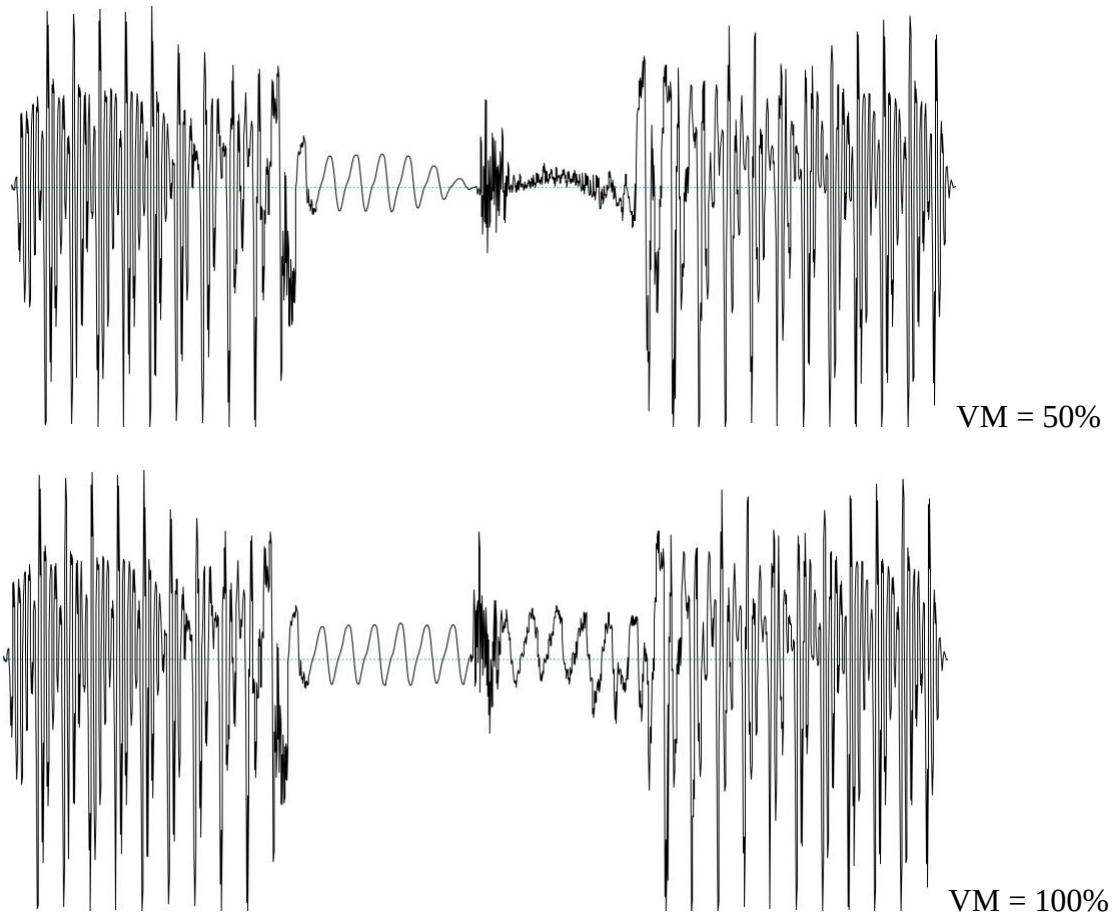


Figure 14: Stimuli from Experiment 2a. Visualizes the variation in VM levels (0%, 50%, 100%) while maintaining the other cues constant.

The Experiment 2b investigated responses to stimuli with prototypical /k/ (VL = 70, CL = 150, VM = 0) and /g/ (VL = 130, CL = 100, VM = 100) values for the selected cues. These are essentially biomechanical replicas of /k/ and /g/ with matched respective acoustic cue values for VM, VL and CL. To study the varying effect of VOT on these prototypes, a range of VOT values from 5ms to 120ms (only up to 90ms VOT for /g/ prototype due to short consonant length) were added into the stimuli. There was a total of 22 different stimuli (10 + 12). An illustration of the stimuli is provided in **Figure 15**. Note that VOT values (held constant at 50ms in the images) also vary from 5-120ms.

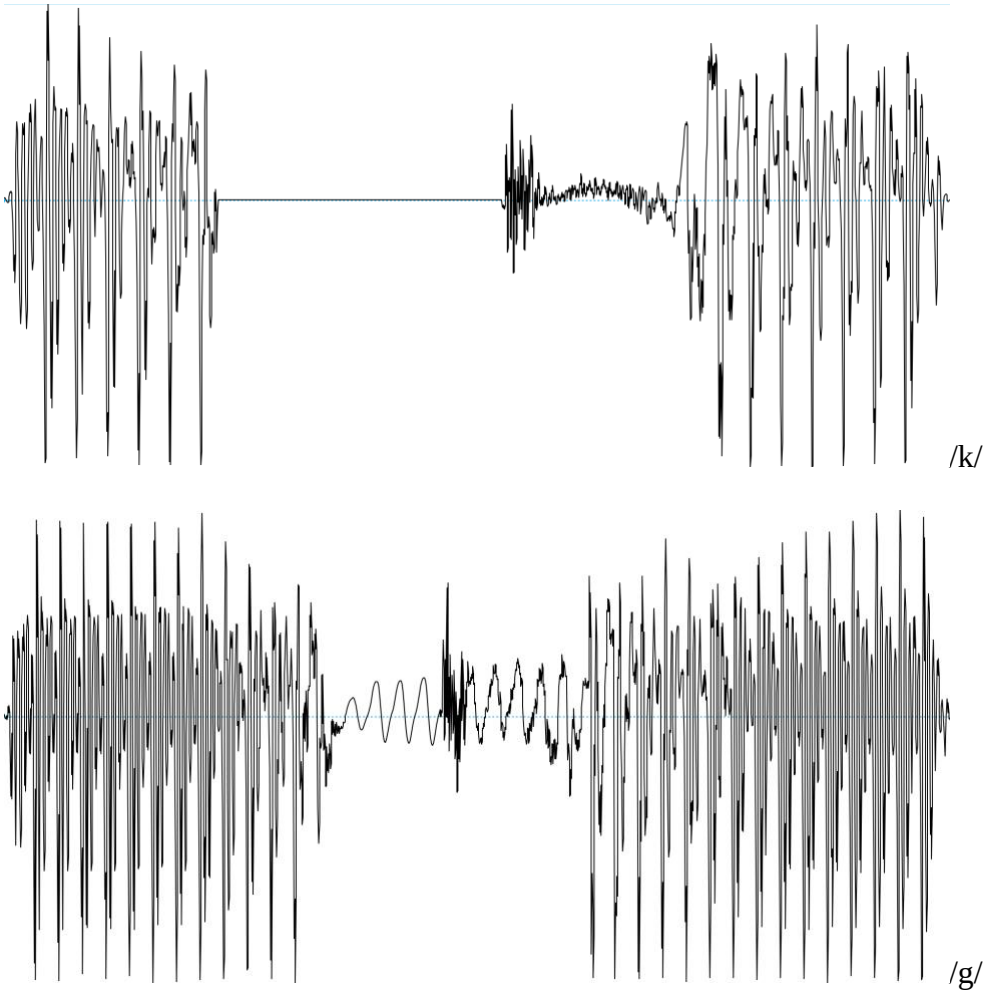


Figure 15: Stimuli in Experiment 2b. *Prototype of /k/ is shown on the top and /g/ on the bottom.*

In Experiment 2c the same prototypical /k/ and /g/ stimuli as used in Experiment 2b were used but with the added dimension of varying VM (three levels of VM = 0, 50, 100) for each prototype. Again, VOT measures were added from 05ms to 120ms (only up to 90ms VOT for /g/ prototype due to shorter consonant length). Since prototypical /k/ and /g/ length values for VL and CL were tested together with prototypical /k/ and /g/ values of VM, the resulting perceptual response towards the prototypes cannot be attributed to one cue over the other. The addition of VM levels was to disentangle the length effects (VL and CL) from voicing maintenance (VM). A

total of 66 different stimuli (12*3 + 10*3) were used. **Figure 16** provides an illustration of the added dimension of VM to the /k/ prototypical stimuli (the same process was applied to the /g/ prototype). Note again that VOT is constant at 50ms in the images but varies from 5-120ms in the stimulus set.

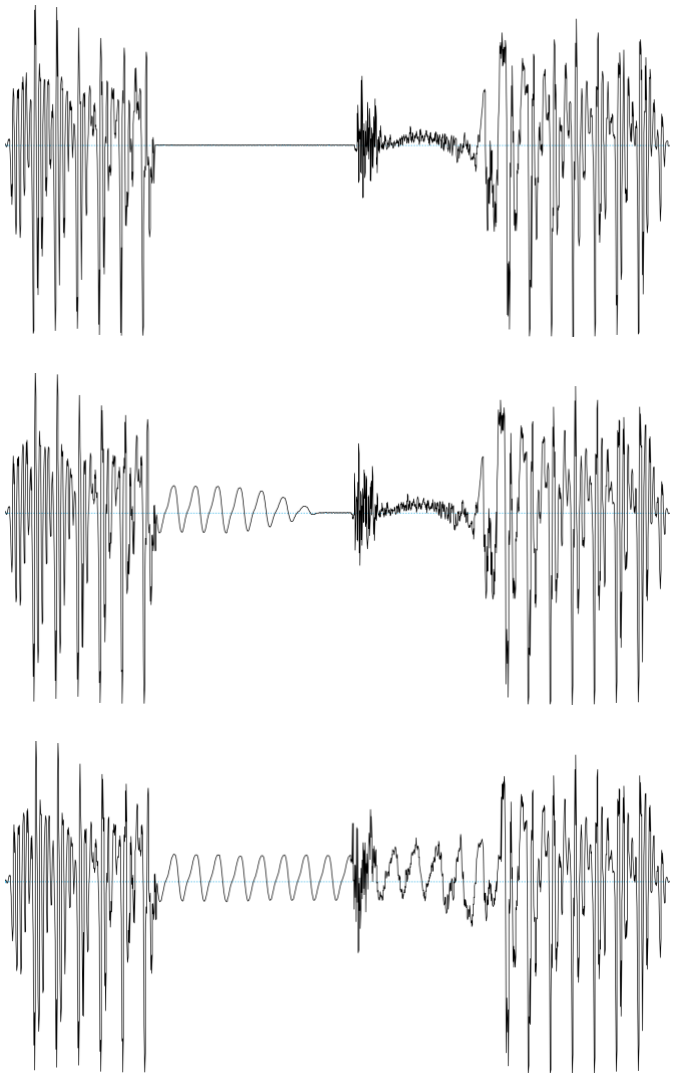


Figure 16: VM manipulations in Experiment 2c (VOT held at 50ms). Visualization of the varying levels of VM for prototype /k/.

3.1.3. Procedure

The location and equipment used for this experimental task were almost entirely identical to that of Experiment 1 (Chapter 2). Only the differences in procedure will be noted here. In

Experiment 2a, the 33 stimuli were presented 15 times, a total of 495 trials lasting 20 minutes on average. In Experiment 2b, the 22 stimuli were presented 15 times, a total of 330 trials lasting 12 minutes on average. In Experiment 2c, the 66 stimuli were presented 10 times, for a total of 660 trials lasting 25 minutes on average.

3.2. Results

3.2.1. Experiment 2a: Ambiguous length cues

This experimental condition allowed visualization of the VOT effect when CL and VL cues were held at ambiguous values (i.e. ambiguous vowel and consonant durations between /g/ and /k/ productions), and how this effect varied over three levels of VM (0%, 50%, 100%). Two participants (out of 32 total) were removed from the analysis for the same criteria as outlined in Experiment 1 (Chapter 2). Note the greatly reduced number of ineligible participants following the inclusion of VOT (and consequently a burst signal) to the stimuli. The data for all remaining participants together in this condition is illustrated below in **Figure 17.** The figure shows the probability of voicing perception by the listeners (y-axis), with VOT presented on the x-axis, and grouped by VM. To avoid confusion, note that the x-axis has changed from VM in Experiment 1 results (Chapter 2) to VOT in the current chapter. Similar to the results from the previous experiment, the effect of VM had a strong influence on /g/ response probability. Across all participants, the increase of VOT appears to cause the probability of /g/ response to decrease. This effect is apparent for all levels of VM, but at 0% VM the impact of VOT seems to reach a ceiling around 40ms, with the probability of /g/ response never going below 20% (another sign of the clear perceptual bias for /g/, even when VOT is included). The strength of the effect of VOT seems to be greatest at low VM values, and it decreases as VM approaches 100%.

Figure 17: Data for Experiment 2a. The x-axis refers to levels of VOT, and the y-axis refers to the /g/ response rate. Panels from left to right reflect the change in VM levels.

Analysis of the individual response patterns of the participants revealed an interesting result; the participants seemed to be split into three qualitatively different groups based on their response patterns. *Group A (n = 9)* seemed to have a relatively much higher sensitivity to VOT variation across all levels of VM, while also having relatively low sensitivity to the voicing cue. *Group B* (n = 14) showed far less sensitivity to VOT and their responses appear mostly influenced by changes in VM. *Group C* (n = 7) responses were influenced primarily by VM, but at ambiguous VM levels (VM = 50%) they responded relatively strongly to VOT. The response patterns for these three groups are shown in **Figure 18.** The figure shows the probability of voicing perception by the listeners grouped by VM (y-axis), with VOT presented on the x-axis.

Group A responses are shown in red, group B responses are shown in green, and group C responses are shown in blue.

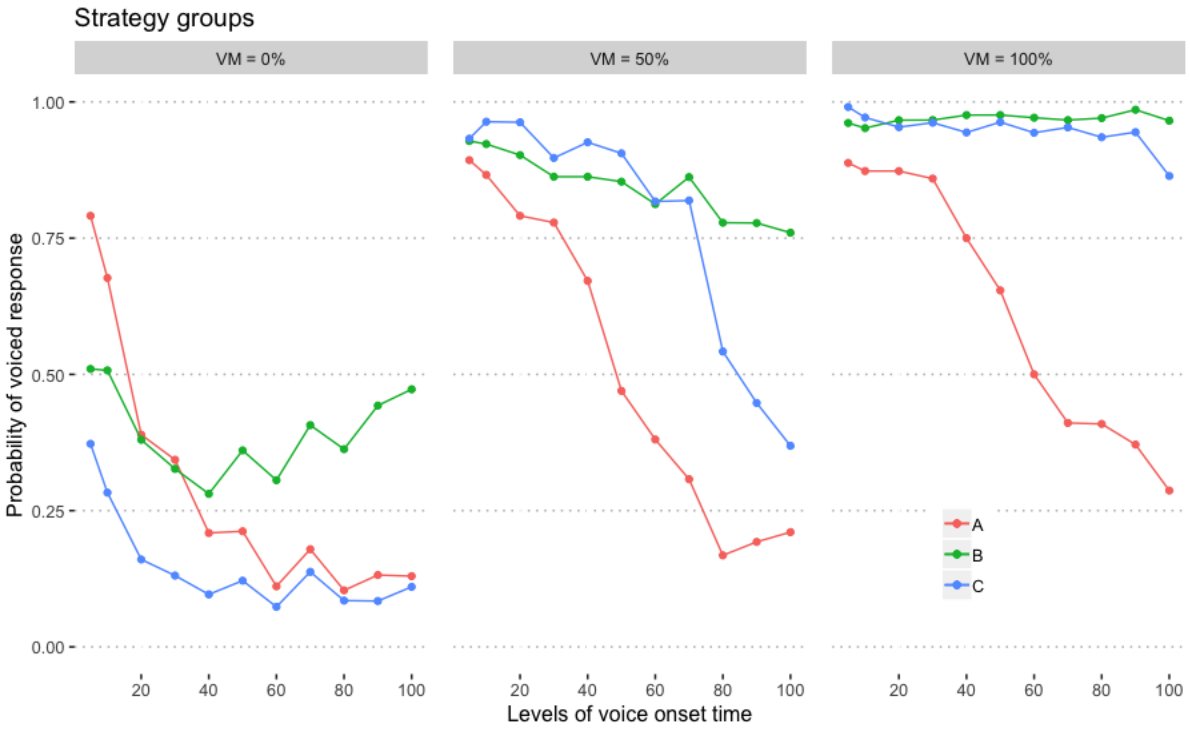


Figure 18: Results from Experiment 2a split into 3 groups of responders by response pattern. Panels from left to right indicate changing levels of VM.

In the determination of the perceptual VOT boundary, we chose to analyse responses from Group A, who exhibited a response strategy almost entirely based on VOT. The reasoning for this choice is as follows. In determining a perceptual VOT boundary, it makes more sense to use values from individuals who specifically attend to variation in VOT. The perceptual results from other groups are more highly influenced by VM (and less so by VOT), and thus make those groups less suitable for an analysis of VOT perception. As described in the results from Experiment 1 (Chapter 2), the ideal extraction of an ambiguous VOT value would occur when VL = 100ms, CL = 125ms, and VM = 25%. Since the extracted VOT value is meant to be ambiguous between voicing perception for /g/ and /k/, the cue values found to be ambiguous

were selected. The reasoning for the ambiguous value of VM being 25% and not its midpoint of 50% is explained in Chapter 2. As this portion of the experiment was previously only intended as a pilot study, only VM levels of 0, 50 and 100 were tested. To account for this lack of data at the 25% voicing maintenance point, ambiguous VOT duration values were extracted from VM = 0% and VM = 50% subsets and averaged to interpolate to model the VM = 25% data points. Although the change between 0-50% should not be assumed to be linear, we believe that this is the best averaging method available given the status of our limited data set.

To determine the ambiguous VOT boundary in this experiment the following procedure was applied: for each VM level subset in Group A, the VOT value when /g/ probability was at 50% was determined. This procedure is illustrated in **Figure 19**. The 50% /g/ response crossover at VM = 0% was 17ms, and at VM = 50% was 51ms. The final ambiguous value was determined to be the arithmetic mean, thus 34ms. This value, determined based on our biomechanical VCV stimuli, is in accordance with many perception and production studies[3] (Lisker & Abramson, 1964; Nakai & Scobbie, 2016; Chodroff & Wilson, 2017).

[3] Nakai & Scobbie, 2016, found a perceptual VOT boundary of 27ms between /g/ and /k/. Chodroff & Wilson, 2017, found a mean VOT of 56 for /k/, and a mean VOT of 17 for /g/ in isolated speech. The arithmetic mean of these values (the presumptive VOT boundary) is 36.5ms.

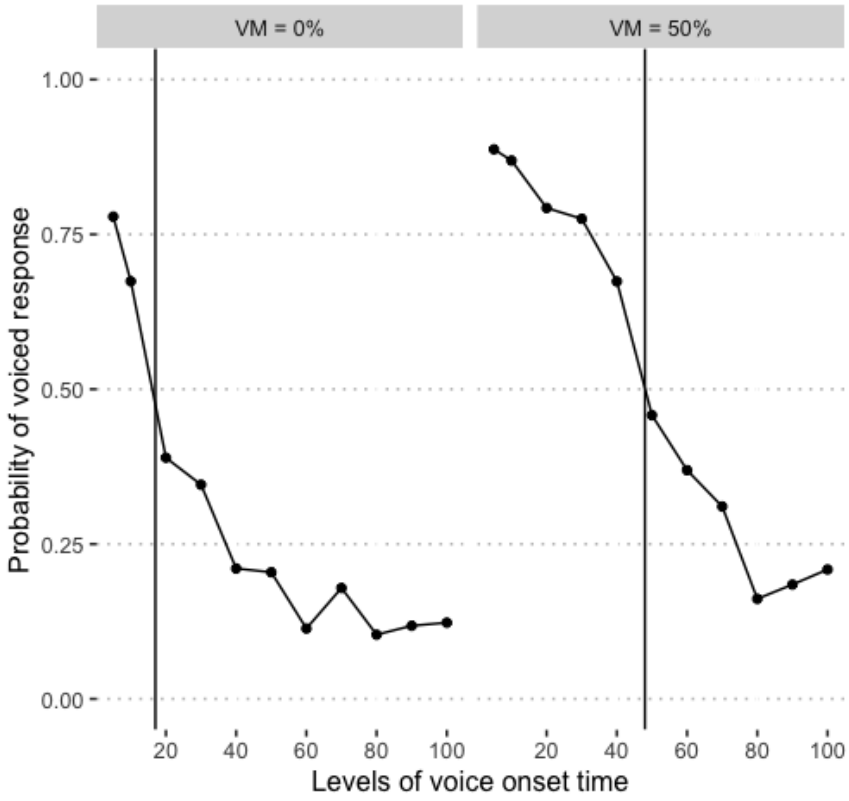


Figure 19: Illustration of the process for extracting the ambiguous VOT value. Vertical lines represent which VOT values result in the 50% response rate for /g/.

3.2.2. Experiment 2b: Prototypical values for VM, VL and CL

Experiment 2b investigated the effect of systematically varying VOT on the probability of /g/ response when the stimulus was either presented with prototypical /k/ durations and voicing (VL = 70, CL = 150, VM = 0) or prototypical /g/ durations and voicing (VL = 130, CL = 100, VM = 100). Though only a pilot study with 7 participants, the results in figure 20 offer clear evidence that the stimuli were perceived as their intended prototypes based on their duration and voicing values only (cues of VM, VL and CL). Thus an effect of VOT was nearly absent. Please note that this experiment did not contrast the length cues (VL and CL) with the voicing maintenance cue (VM), as our experimental design only tested a specific combination of the cue

factors (i.e. the prototypical values for /k/ and /g/). In other words, the difference in response patterns between each prototype could have been due to a difference in CL length (CL = 150ms for /k/ versus CL = 100ms for /g/) VL length (VL = 70ms for /k/ versus VL = 130ms for /g/), or due to the difference in voicing maintenance (VM = 0% for /k/ versus VM = 100% for /g/). This issue will be addressed with Experiment 2c, which aims to disentangle these effects.

Experiment 2b results are illustrated in **Figure 20**. The figure shows the probability of voicing perception by the listeners (y-axis), with VOT presented on the x-axis. The results for the /k/ prototype are in blue, and the results for the /g/ prototype are in red. Note that the /g/ prototype, with CL = 100, could only accommodate VOT values up to 90ms, which accounts for the varying lengths of data in the figure. The results indicate that, for the present biomechanical stimuli, the variables manipulated to create the prototype stimuli (VL, CL and VM) appear to have greater weight than VOT for the perception of voicing. These results are interesting, as the literature has suggested VOT to be the main cue in voicing perception (Lisker & Abramson, 1964; Abramson & Whalen, 2017).

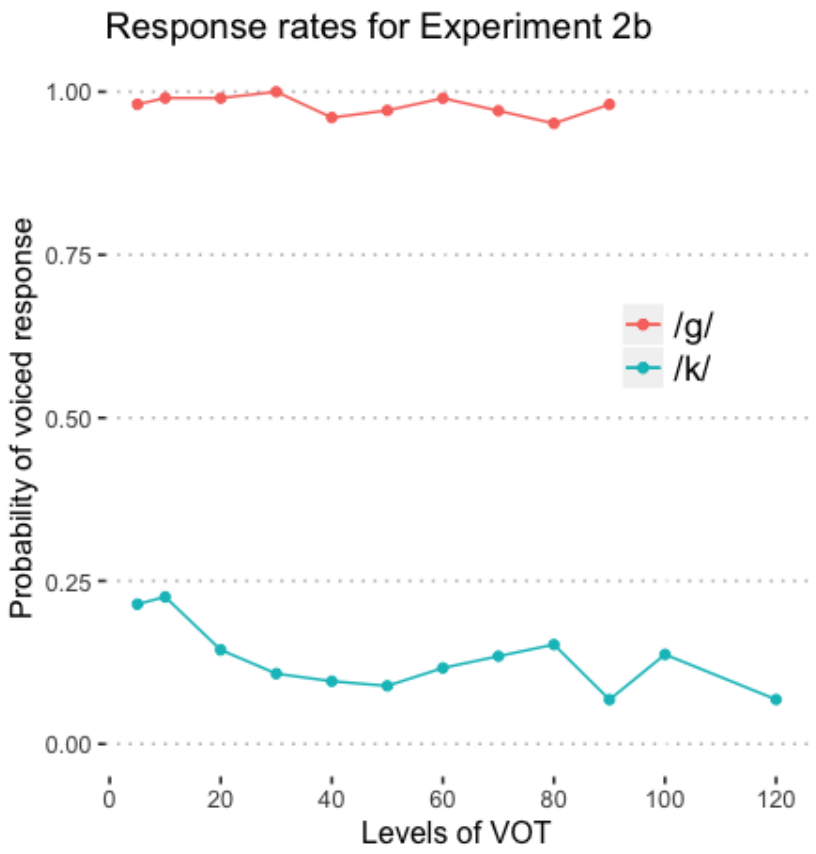


Figure 20: Results from Experiment 2b. Response differences between /k/ and /g/ prototypes.

3.2.3. Experiment 2c: Prototypical values for VL and CL with varying levels of VM

The results from Experiment 2b revealed that the prototypes were easily distinguished from one another, and that the manipulated variables VM, VL and CL together appear to have higher weight than VOT for /k/ vs. /g/ distinctions. Again, the lack of influence from VOT is surprising due to it being reported as a central perceptual cue to voicing distinctions in English (Lisker & Abramson, 1964; Abramson & Whalen, 2017). Experiment 2c was designed to address whether the perceptual differences between the prototypes are due to voicing maintenance or to the VL and CL length effects. This experiment used stimuli with the prototypical length cue values for /k/ (VL = 70ms, CL = 150ms) and for /g/ (VL = 130ms, CL = 100ms), but introduced

varying levels of voicing maintenance (0, 50 and 100%). The effect of voicing maintenance was confounded with those of vowel and consonant length in Experiment 2b. In Experiment 2c, the prototypes for /k/ and /g/ across the levels of VM were only distinguished by the values of their length cues: VL (VL = 70ms for /k/, VL = 130ms for /g/) and CL (CL = 150ms for /k/, and CL = 100ms for /g/) while VM was varied. Experiment 2c was also designed as a pilot study. Results were gathered from 15 participants.

The response patterns for the two length prototypes (as illustrated in **Figure 21)** were very similar across all VM conditions. The figure shows the probability of voicing perception by the listeners (y-axis), with VOT presented on the x-axis, and responses grouped by the three VM conditions. The response patterns for both prototypes were influenced towards /g/ perception mainly by VM increases, indicating that perception relies more heavily on VM than the length effects of VL and CL. The formerly reported differences between prototypes in **Figure 20** (Experiment 2b results) were essentially the same as the differences in response patterns between VM values of 0% and 100% in the Experiment 2c results (**Figure 21**), indicating the length cues on their own had little effect when compared to VM.

These results are in line with the previous findings from Experiment 1 (Chapter 2). Note that there was no experimental setting which explored one of the length cues (CL or VL) while keeping the other constant, so these effects cannot be disentangled from each other. The results of Experiment 1 (Chapter 2) showed that VL had a statistically significant effect, while CL did not. Based on this, we assume that the length effects (when looked at in tandem) were mostly an effect of VL, but since it was not explicitly explored, we can only surmise. It is worthwhile to note that although CL did not have a statistically significant effect, its influence on perception is still apparent when looking at the interactions with VM (see **Figure 11** in Chapter 2).

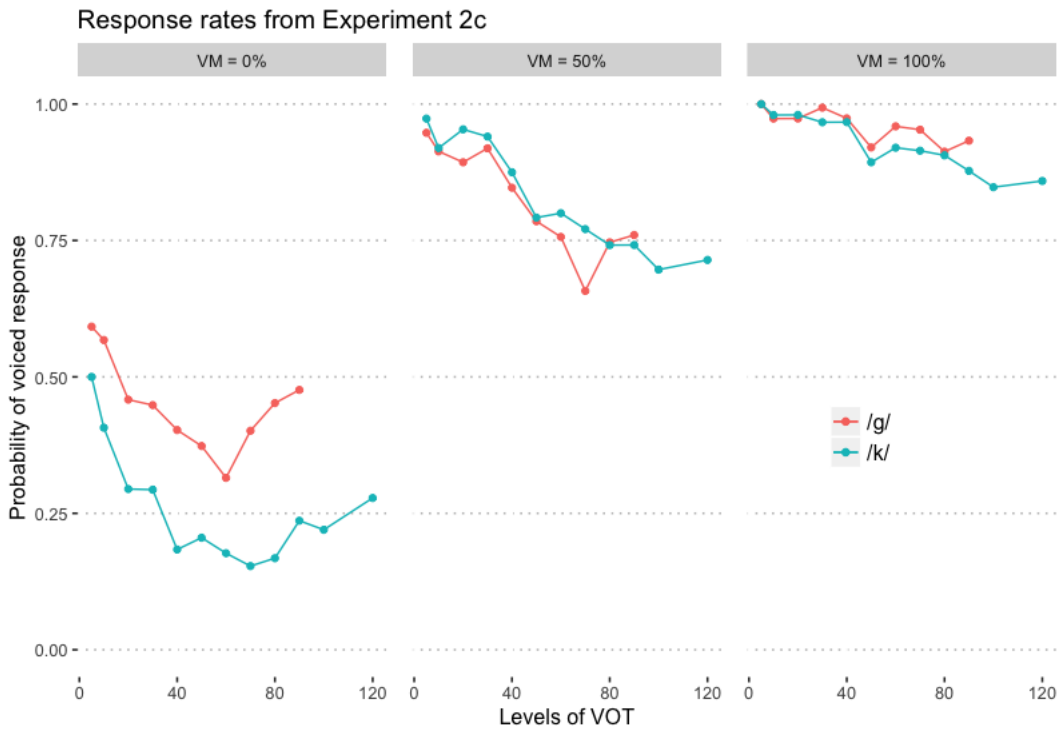


Figure 21: Results for Experiment 2c *comparing /k/ and /g/ responses across three levels of VM. Panels from left to right show results for each level of VM (0, 50 and 100%).*

3.3. Discussion of Experiments 2a, 2b and 2c

Experiment 2a. The results from sub-experiment 2a provided further detail about the effect of VM as reported in Experiment 1. Recall from the discussion in Chapter 2 that while VM was found to have the most influence on the probability of voiced perception, at low levels of VM the other cues of VL and CL took over to increase their effect. Experiment 2a controlled for the effects of VL and CL, since they were held constant at their ambiguous middle values. This allowed the direct comparison of the effects of VM and VOT. As VOT is widely reported as the most influential cue in stop consonant voicing perception for English, it was expected to have a greater relative effect. These expectations were partly met. The effect of VOT was clearly apparent when looking at **Figure 17**. For all levels of VM, increasing VOT values decreased the

probability of a voiced response. However, VM still maintained an overwhelming influence on the probability of voiced responses. Congruent with the results from Experiment 1, at low levels of VM, other cues became essential. The effect of VOT was strongest when voicing maintenance was totally absent (VM = 0%) and weakened as VM increased.

The expected responses should have shown the greatest influence by the VOT cue, with small increases in voicing response as VM levels increased. This would appear as similar to results in **Figure 8** (where VM had the greatest influence, and the increases in VL levels caused probability of voiced response to increase while the same probability curve shape was maintained), but with VOT replacing the role of VM, and VM replacing the role of VL. Note that in this case, however, the categorical perception curve would be flipped, since a longer VOT is a cue for /k/. Thus, increasing VOT should serve to *decrease* voicing perception.

Splitting participant responses by their perceptual patterns *post hoc* revealed an interesting phenomenon: individual participants appeared to exhibit different perceptual strategies for voicing identification (**Figure 18**). As described above, three response patterns were found: group A (n = 9) was influenced almost solely by VOT; group B (n = 14) was influenced almost solely by VM; and group C (n = 7) was influenced primarily by VM but shifted to VOT at ambiguous (50%) VM values. These results are interesting when considering the previous results. Both Experiment 1 and the Experiment 2a indicated that while VM showed the highest influence at high levels, the other cues increased their effect at low VM levels. Group A demonstrates the response patterns that were initially expected to occur (where VOT has the highest influence as a perceptual cue). More specifically, this group showed a high influence of VOT values, as would be expected for English participants based on the literature. Group B demonstrated very little influence of VOT, and responses were mainly a function of the VM

levels. These strategy differences may be due to individual differences in perceptual weighting within a language community, or to interference from perceptual cues relevant to known languages other than English. For instance, a bilingual French speaker could be more influenced by VM due to its prominence in their other known language. The ultimately extracted value of 34ms for an ambiguous VOT is interesting because it matches that of reported literature (Lisker & Abramson, 1964; Nakai & Scobbie, 2016; Chodroff & Wilson, 2017).

Experiment 2b. Experiment 2b investigated how VOT manipulations affected response patterns for each prototypical stimulus. The respective extreme values for three prototypical cues (VM, VL, and CL) were used to create /g/ and /k/. The results (**Figure 20**) revealed that these prototypical cue differences had a much greater effect than VOT in differentiating the prototypes perceptually. The perceptual difference caused by these cues influenced voicing perception much more than VOT did. Note that because voicing maintenance (VM) and the length cues (VL and CL) were tested together, further experimentation was required to disentangle their individual effects and determine which one was most influential.

Experiment 2c. The issue of differences between voicing and length cues was addressed with Experiment 2c, which varied VM when comparing /k/ and /g/ stimuli determined by fixed prototypical CL and VL settings. These results (**Figure 21**) make it clear that the perceptual difference between prototypes in Experiment 2b was almost entirely due to differences in VM. Only at the lowest value of VM (0%) did the length cues cause a noticeable difference in perception. The effect of VOT was also more prominent when VM levels were lower. This again is consistent with the previous results, which indicated that VM has the strongest influence at high levels, whereas at low VM levels, the other cues jump in to increase their effect.

Note that the response patterns at 0% VM start to deviate from their expected course around 60ms. This rise is unintuitive considering the literature regarding VOT perception, which suggests that increases in VOT would decrease voicing perceptions. An explanation for this is likely to lie in the manipulated VOT stimuli. Manipulating the original VOT recording length of 34ms to higher lengths (up to 120ms) causes the sound to artificially 'stretch'. This stretch affects the sound of the burst segment in the VOT so that it sounds less like the characteristic 'click' of natural productions and more like a metallic spring. We assume this artefact from the stimulus manipulations to have caused the unexpected rise.

The trend in the results so far is that VM has a greater effect than VOT, VL, and CL for the perception of Canadian English intervocalic stops. Since VOT is typically reported as the primary cue in stop consonant voicing perception in English, this finding is surprising. It is only at low values of VM (which could indicate both /g/ or /k/) that the other cues come in and increase their effect to reduce the ambiguity. Note that there was no experimental condition where the effect of VOT was explicitly tested against the effects of VL and CL. This means that in the grand scheme of perceptual cue weighting, VM is placed at the top, and we know that VL has a greater effect than CL (recall statistical significance and effects from Experiment 1), but we cannot know where to place VOT along this scale. VOT likely has a greater effect than that of VL and CL, as is suggested by the literature, but this was not unambiguously measured here.

Chapter 4: Experiment 3: Cue Weighting with Ambiguous VOT

Experiment 3 incorporated the newly extracted ambiguous VOT value from Experiment 2 as an additional cue with the other manipulated cues VM, VL and CL. A similar stimulus set to that in Experiment 1 was created but with an added burst signal to indicate the ambiguous VOT value. Experiment 3 examined the perceptual cue-weighting of the stimuli when a VOT value was present, which was compared to when VOT was absent (as in Experiment 1, Chapter 2).

4.1. Methods

4.1.1. Participants

Nineteen native English speaking participants were recruited through the Linguistic Research Participation System administrated by the Department of Linguistics and Languages at McMaster University. English speakers were selected if they learned the language before the age of 5. Again, participants were undergraduate students (aged 17-24) who reported normal hearing and vision, and they received course credit for their participation.

4.1.2. Stimuli

Experiment 3 employed a similar cue-weighting process to that of Experiment 1 but additionally introduced an ambiguous VOT value to all stimuli. The stimulus set from Experiment 1 (see Chapter 2) was modified to include an unaspirated VOT of 34ms as described in the previous section (Chapter 3, under results for Experiment 2a). The original VOT recording used to create stimuli in Experiment 2 was 32ms. A 2ms period of silence was added at the end of the VOT duration (preceding the onset of vowel voicing) to create an ambiguous VOT sample of 34ms, as described in the previous experiment.

This resulted in a new stimulus set which included the same manipulated cues of VL (70, 100 and 130ms), CL (100, 125 and 150ms), VM (0, 25, 50, 75 and 100%) as in Experiment 1, but with the added ambiguous VOT of 34ms, identical for all stimuli. The total number of different stimuli was 45. The reasoning behind the addition of the ambiguous VOT to all stimuli was as follows. Experiment 1 (Chapter 2) explored the cue-weighting of the cues VM, VL and CL in the absence of VOT. Experiments 2a, 2b and 2c (Chapter 3) discovered that varying VOT had a relatively small influence on voicing perception when compared to VM, VL and CL. Furthermore, these experiments found that within those latter cues, VM was the main influencer. In the third experiment (current chapter), the goal was to revisit the cue-weighting of VM, VL and CL, but in the presence of an ambiguous as opposed to missing VOT signal. Although VOT was found to have a relatively low influence on voicing perception compared to VM in Experiments 2b and 2c, its effect is clearly visible in Experiment 2a (see **Figures 17 and 18**). The aim was to compare the results to those of Experiment 1, in order to show how the effects and interactions of the cues change in the presence of VOT. Additionally, to compare how the inclusion of a VOT cue influenced experimental difficulty, as evidenced by the number of removed participants.

4.1.3. Procedure

All elements of the procedure were identical to that of Experiment 1.

4.2. Results

Of the 19 participants, 3 were removed for responses not showing substantial deviation from chance or failing to complete the procedure properly. Note the rate of random responses is

much lower than in the first condition, indicating the relative ease of the task when VOT is included as a perceptual cue[4]. **Figure 22** shows the data for the remaining 16 participants grouped by VL and CL responses. The figure shows the probability of voicing perception by the listeners (y-axis), with VM presented on the x-axis, and grouped by VL (across panel) and CL (within panels). Note that the x-axis has returned to VM (as in Chapter 2) from VOT in the previous chapter. This figure indicates several effects and interactions in the data. As in Experiment 1 (Chapter 2), there is a bias for more /g/ responses compared to /k/ responses. Increasing VM has a clear effect of increasing /g/ responses in all conditions, although not in a linear manner. The increase is most dramatic between low VM levels (particularly between 25-50%), similar to results from Experiment 1. The increase of VL causes a slight increase in /g/ responses, while changes in CL have little effect.

[4] In Experiment 1 (Chapter 2), there were 42% of participants removed due to random response patterns, and the difficulty in the task was attributed to the lack of VOT in the stimulus. In Experiment 3, the number of discarded participants was reduced to 16%. The only difference between these experimental conditions was the presence or absence of a VOT signal (ambiguous VOT of 34ms). This discrepancy between the numbers of random response patterns from listeners shows the importance of VOT for English listeners, since its addition caused a significant decrease in the rate of random responses.

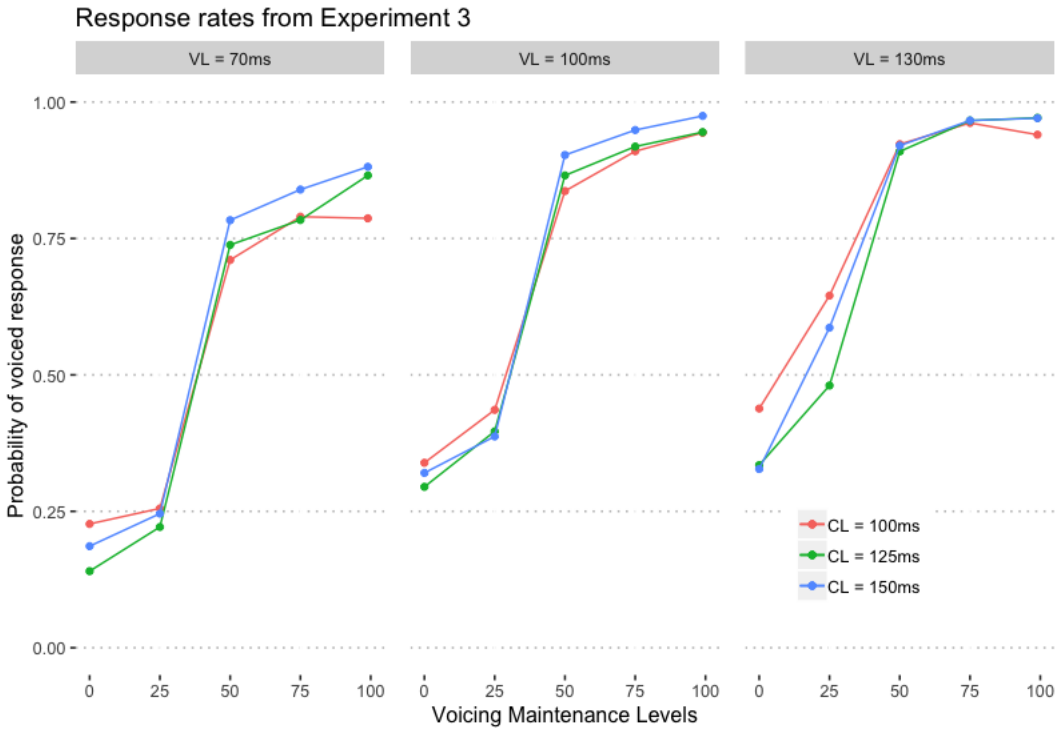


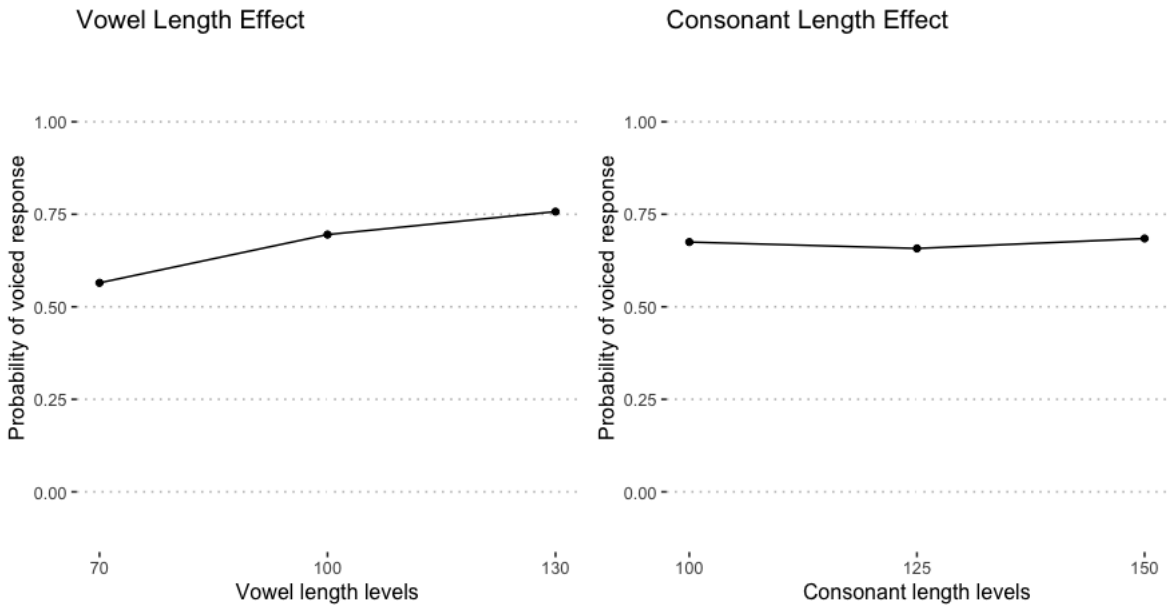
Figure 22: 3x3 matrix of the results from Experiment 3.

4.2.2. Statistical analysis

The significance of the main effects and interactions was determined by generalized linear mixed-effects modelling, using function glmer (Bates, Maechler, Bolker, & Walker, 2015) in the R environment (R Core Team, 2018). A binary logit model was used to best account for the binary dependent variable (/g/ or /k/ response). The model was again worked down in complexity until the most complex model with lowest AIC was reached (Burnham, Anderson, & Huyvaert, 2011). This model included all main effects and interactions for variables VL (70, 100 and 130ms), CL (100, 125 and 150ms), and VM (0, 25, 50, 75 and 100%), which were tested with a significance threshold of $p < 0.05$. *Subject* was selected as a random factor and was included with random intercepts and random slopes for VM, and VL, as these led to the most

complex model with the lowest AIC. An increase or decrease in the complexity of this model led to an increase in AIC.

A likelihood ratio test was performed comparing the full model to a null model including only the random effects. The effects and interactions of the full model were statistically significant. In the full model, all variables VM ($z = 8.751$, $p < 0.001$), VL ($z = 5.175$, $p < 0.001$) and CL ($z = 2.845$, $p = 0.004442$) had significant main effects. The main effects are illustrated in **Figure 23** (note that although the CL effect appears minimal in the figure, its effect reached statistical significance). The only significant interactions were VL:VM ($z = 3.365$, $p < 0.001$) and CL:VM ($z = 2.331$, $p < 0.05$), illustrated in **Figure 24**. For the VL:VM interaction, it appears that VL has a greater effect at a positive but low VM ratio, when VM = 25% (indicated by the steeper slope). For the CL:VM interaction, at high values of VM (over 50%) the effect of CL is quite minimal, but at low VM values (0-25%) its effect is more prominent, serving to decrease /g/ response probability as CL increases. Significance and z-scores for all main effects and interactions are outlined in **Table 3.**



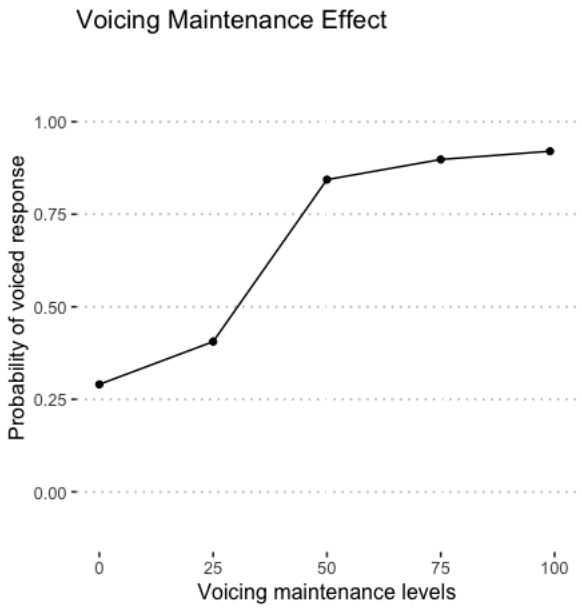


Figure 23: Main effects of the vowel length, consonant length and voicing maintenance factors on voiced responses in Experiment 3.

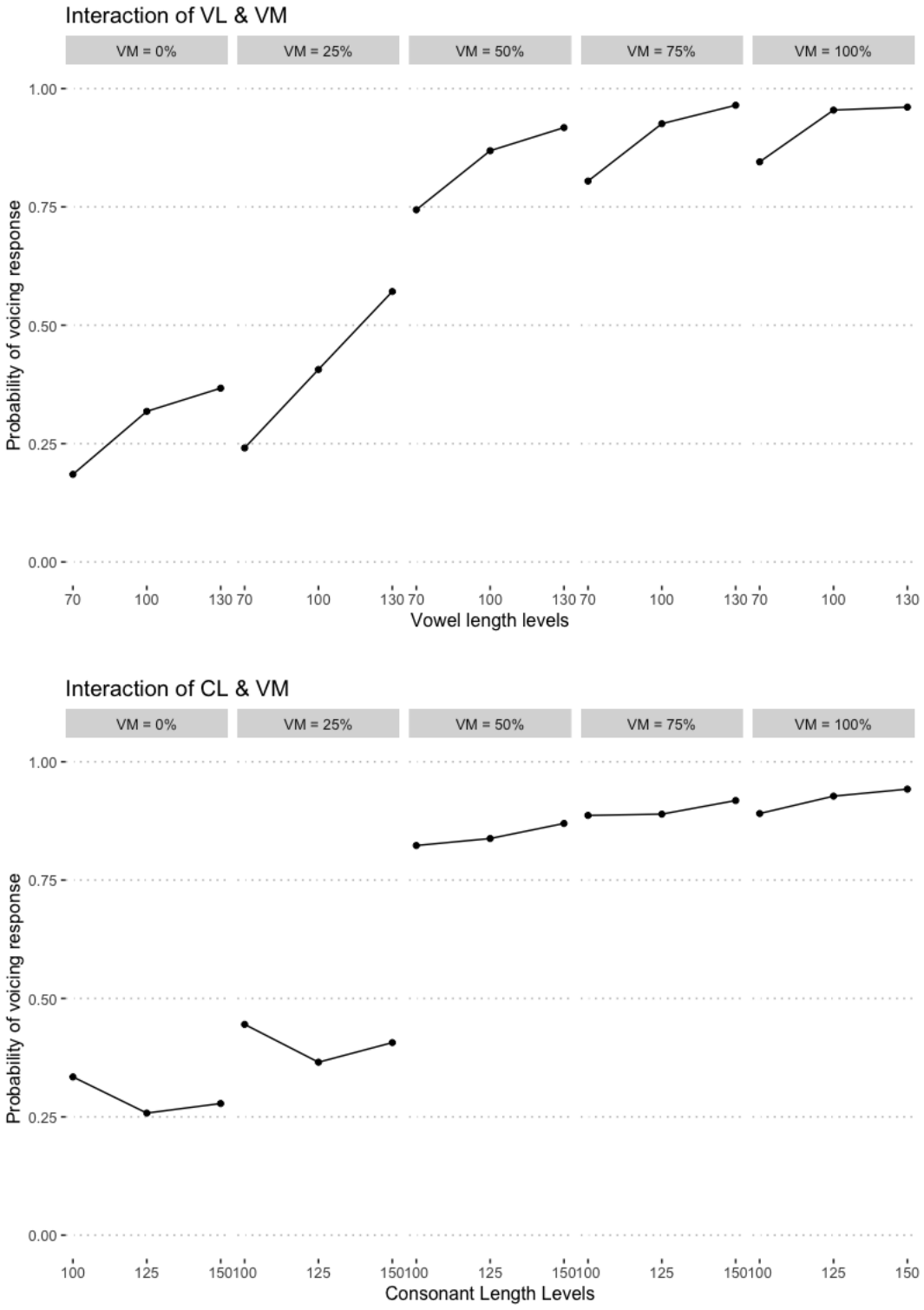


Figure 24 : Interactions in Experiment 3 voiced responses. VL:VM on the top and CL:VM on the bottom.

Effect/Interaction	Estimate	Std. Error	z value	Pr(>|z|)
VL	0.68592	0.13254	5.175	2.28e-07 ***
CL	0.09299	0.03268	2.845	0.004442 **
VM	1.92981	0.22052	8.751	< 2e-16 ***
VL:CL	-0.03519	0.03249	-1.083	0.278785
VL:VM	0.14486	0.03742	3.871	0.000108***
CL:VM	0.20121	0.03587	5.610	2.03e-08 ***
VL:CL:VM	0.02009	0.03559	0.565	0.572385

Table 3: Main effects and interactions from experiment 3.

4.2.3. Cue Weighting

Following Burnham et al. (2012), variable importance to model fit was calculated by change in AIC resulting from the removal of that variable in the model. The change in AIC is referred to as ΔAIC. The removal of VM caused a ΔAIC of 41, the removal of CL caused a ΔAIC of 11, and the removal of VL caused a ΔAIC of 34. The results, summarized in **Table 4**, confirm the results that VM had the strongest influence, followed by VL, then finally CL with the lowest effect. It is important to note that since there are significant interactions between these variables, the interpretation of this ranking is not so straightforward.

Cue	**AIC**	**ΔAIC**	**Cue Ranking**
VM	8373	41	1
CL	8343	11	3

VL	8366	34	2

Table 4: Cue weighting results from Experiment 3.

4.2.4. Full Analysis

In this 'full analysis', the data from Experiment 1 (Chapter 2) are compared to those from Experiment 3 (Chapter 4) to determine the perceptual changes caused by the inclusion of an ambiguous VOT value (Chapter 4) versus its absence (Chapter 2) in the stimuli. Participant responses from both data sets are compared in **Figure 25**. The 3x3 matrix illustrates the effect of voicing maintenance (x-axis) on participant response rates (y-axis), grouped by CL (columns) and VL (rows).

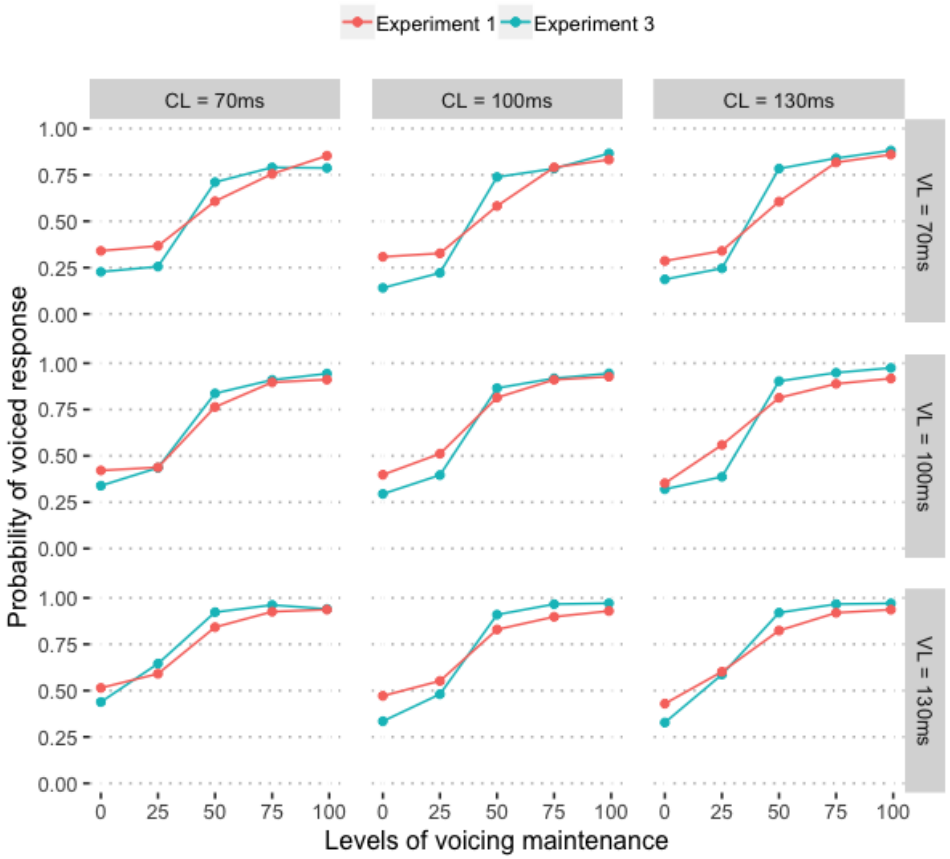


Figure 25 : Comparison of results from Experiment 1 and Experiment 3. The 3x3 matrix illustrates the effect of voicing maintenance (x-axis) on participant response rates (y-axis), grouped by CL (columns) and VL (rows).

The results and implications of the significant main effects and interactions from the main cues of VM, VL and CL have so far been presented separately for the experiments. This section focuses on the comparison between the results of Experiment 1 and Experiment 3 as a consequence of the addition of VOT. The differences between the main effects are shown in **Figure 26,** and the comparison of the significant interactions is shown in **Figure 27.** While VL and CL seem to have nearly identical effects in both sets of results, the change in VM is more apparent. The inclusion of an ambiguous VOT (difference from Experiment 1 to Experiment 3) appears to strengthen the influence of VM. This is shown by lower probability of voiced response at low values of VM (0-25%) and a higher probability of voiced response at high values of VM (50-100%, although the effect tapers off at higher VM values) in Experiment 3 compared to Experiment 1. This perceptual response change is characteristic of the influence of VM on voicing perceptions.

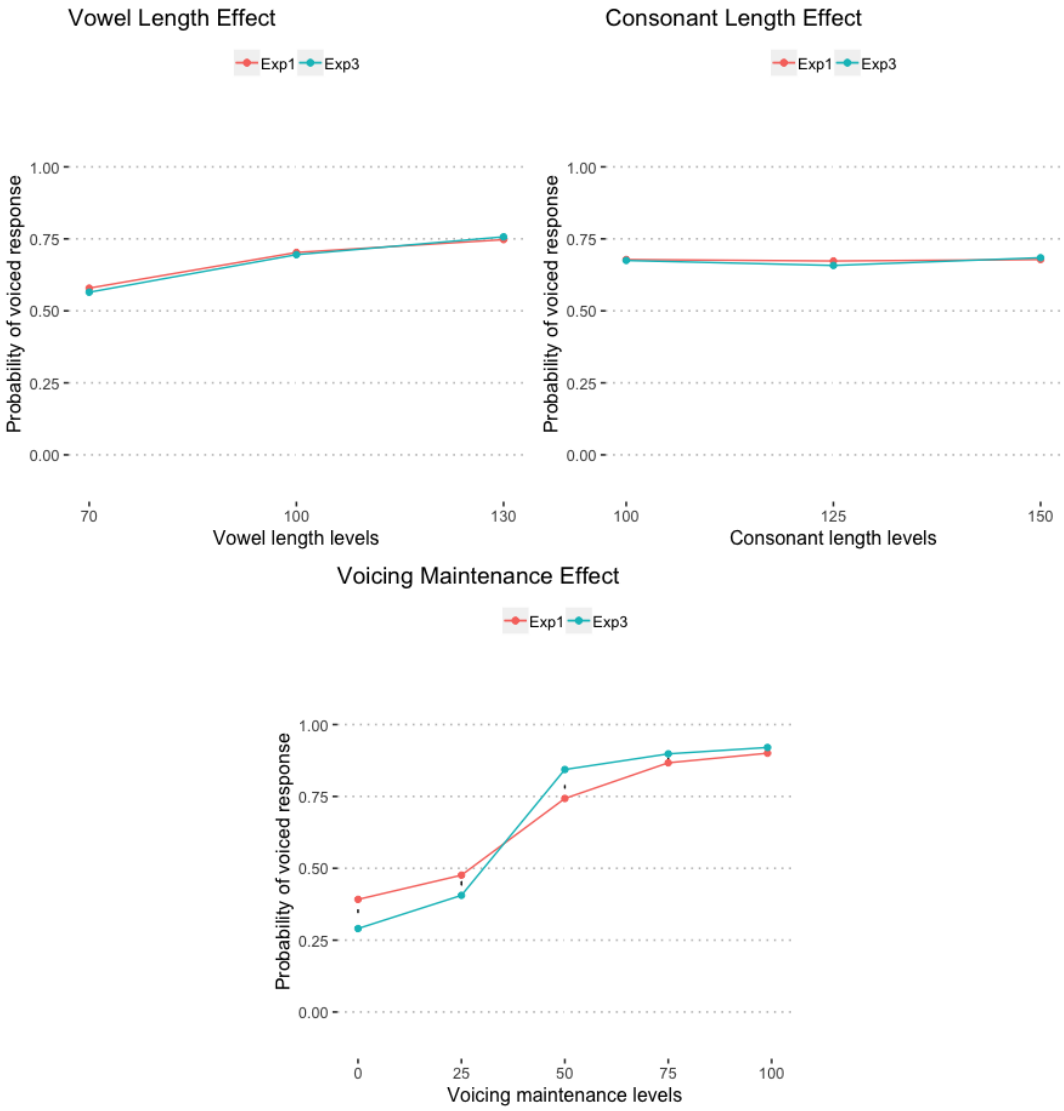


Figure 26: Comparison of main effects in Experiment 1 and Experiment 3. Top left shows VL, top right shows CL, and bottom shows VM. Experiment 1 is in red and Experiment 2 in blue.

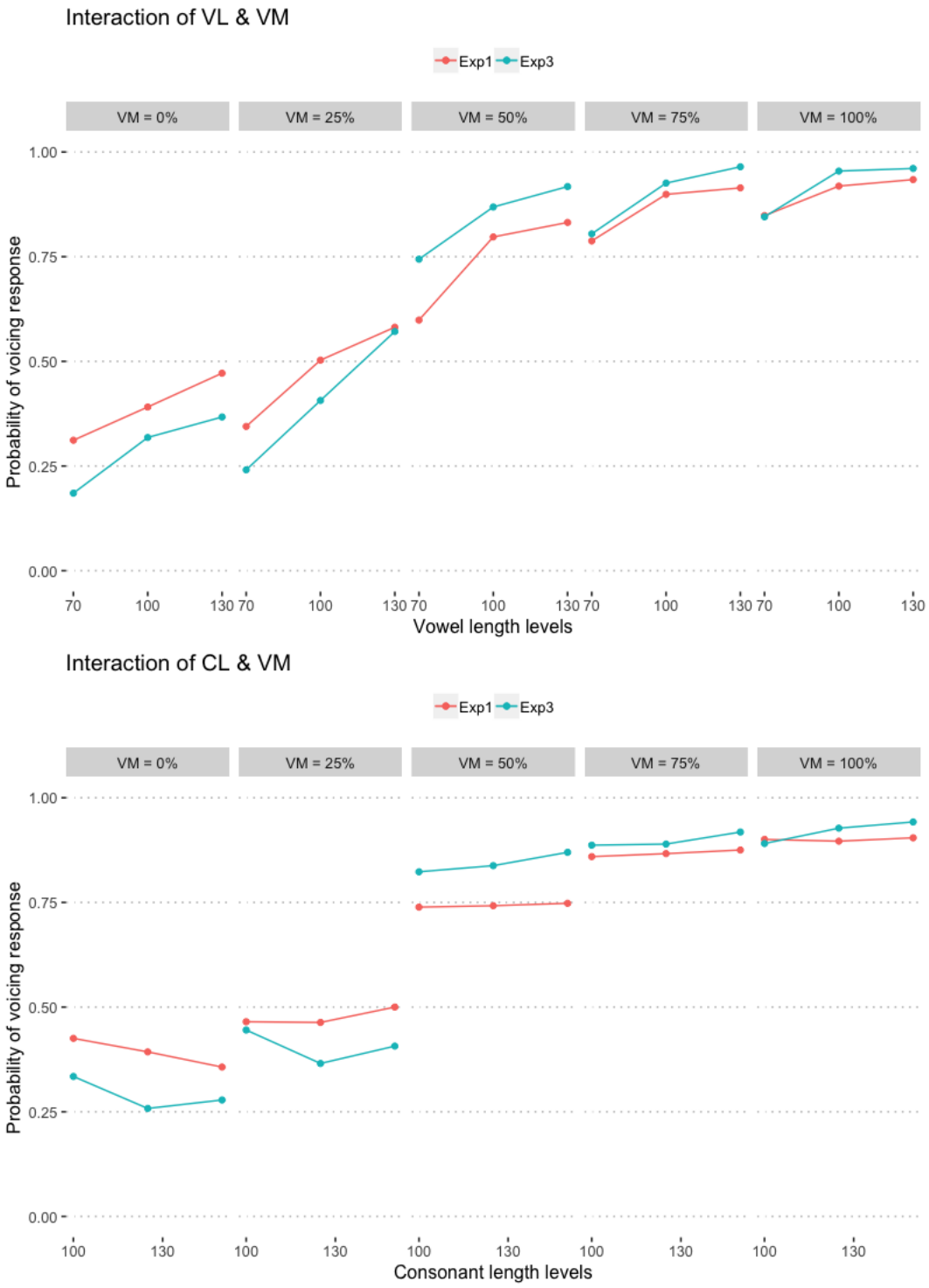


Figure 27: Comparison of significant interactions from Experiment 1 and Experiment 3. VL:VM on the top and CL:VM on the bottom. Experiment 1 in red and Experiment 2 in blue.

4.3. Discussion

We have seen in the previous experiments that the effect of VM is overwhelming compared to the effects of VOT, VL and CL. It is only at low values of VM (which could perceptually signal either a rather "voiced" /k/ or a strongly devoiced /g/) that the effect of the other cues becomes more prominent. The current experiment introduced a perceptually ambiguous VOT value to all stimuli from Experiment 1. This allows insight into how the perceptual cue weighting of the cues VM, VL and CL changes in the presence of an ambiguous VOT cue, in other words, when a listener cannot rely on the VOT cue alone (because it is ambiguous) and would presumably turn to use other perceptual cues to achieve a robust phoneme identification. The patterns in the main effects and interactions are very similar to those of Experiment 1, with the major exception that the effect of CL reached statistical significance in this experiment. The results continue to suggest that VM has the strongest influence, and that only when VM values are low do the other cues take over to guarantee a robust voicing perception.

Further interesting findings come from the comparison of Experiment 1 and Experiment 3 results (**Figures 26** & **27**). Although the patterns of the main effects and interactions are very similar across both experiments, the addition of VOT caused a shift in the influence of VM on response patterns. This is illustrated most prominently in the interactions (**Figure 27**). At low VM values (0-25%), the perceptual responses from Experiment 3 were shifted towards a lower probability of voiced response, and at high VM values (50-100%), Experiment 3 responses were shifted to a higher probability of voiced judgments. In other words, the response rates in Experiment 3 showed an increased influence of VM, with high values suggesting /g/ perceptually, and low values signalling /k/. This resulting change in VM influence was more

prominent for the low values than the high values. It is likely that the added VOT created more ambiguity in the signal, which led to VM taking over more as a cue to determine voicing perception. Only at low VM values do the other cues increase in influence. In other words, when the VOT cue is ambiguous and unreliable, and cannot be used alone, the other perceptual cues function to achieve a robust phoneme identification.

Chapter 5: General Discussion and Conclusions

The described experimental results provide an interesting picture of the weighting of the selected cues for perception of voicing in inter-vocalic velar consonants in English. We had originally hypothesized that, for the Canadian English variety of the English language, the cue of VOT would have the highest perceptual weight, followed by the length effects of VL & CL, and finally with VM having the smallest effect of the reported cues. However, it appears as though in our data VM has the highest perceptual weight, and only at low VM values do the other cues noticeably show their effect. In other words, strong cue-weighting between the available cues takes place in the examined Canadian English variety. This phenomenon is surprising considering the widely reported strength of VOT in English stop voicing perception.

It is important to note that these experimental results provide the perceptual weighting of the acoustic cues by English speakers *for our given biomechanical stimuli*. The stimuli themselves were created and manipulated to have ecological validity compared to natural English production and perception. However, due to the synthetic nature of the stimuli, the results must be interpreted in context. This means that the perceptually ambiguous VOT value of 34ms is specific to the given stimuli and may differ in value for natural productions and perceptions. Additionally, the perceptual cue-weighting is also specific to the stimuli. The relative lack of effect from VOT and the length cues of VL & CL when compared to that of VM can be a factor of the manipulated stimuli, rather than signalling the lack of these cues' effect for English listeners. However, it is highly unlikely that English listeners do not respond perceptually to VOT, VL and CL, even at high values of VM. It is important to note that the methods used for experimental stimulus manipulation are the best currently available for the desired stimulus outputs.

As was presented in Experiment 2a, three distinct response patterns were found within the pool of participant responses (**Figure 18**). These included participants who relied almost solely on VOT (Group A, n = 9), participants who relied almost solely on VM (Group B, n = 14) and those who relied on VM at high levels and relied on VOT at low VM levels (Group C, n = 7). The results indicate that VOT has a great effect (most influential) at least for some participants, although overall VM seems to have the highest effect. The difference in perceptual cue weighting between groups is evidence of individual differences in the strategies used to identify the presented stimuli. This finding of differing strategies is significant because typical studies generalize over all participants, rather than looking for varying response patterns within the participants (i.e. cluster analysis). One of the main findings of this experiment is that different listeners rely on different combinations of cues for a robust voicing decision. Rather than one universally "main" cue (as VOT is typically reported to be), the results indicate a complicated interplay of all acoustically available cues, and each class of listeners use this set of cues differently.

The results lead to the conclusion that for *our given biomechanical stimuli*, the VM cue has the strongest perceptual weight. It is only at low VM values, when the signal becomes more ambiguous, that the other cues' influence takes over. The cue-weighting for our given biomechanical stimuli is as follows. In general, the manipulations of VM cause the greatest shift in perceptual response. High levels of VM (75%-100%) lead to listener response rates at almost 100% probability of a voiced response. When VM values are lowered to 50% and below, the effect of the other cues increases. Additionally, the change in VM values from 25-50% causes a rapid spike in perceptual change, reminiscent of the categorical perception curves expected from VOT responses.

VOT retains its effect at high VM levels, but its effect is significantly greater as VM levels are lowered. VL acts similarly, in that its effect is still present at high values of VM, but its effect is greatly increased at low VM values. Note that there were no experimental conditions which compared the effect of VOT to the effect of VL and CL while keeping VM constant. For this reason, in the perceptual cue-weighting hierarchy, VOT cannot be specifically placed ahead or behind of these length cues. Another experiment is required to definitively state whether VOT would be right after VM in influence, or fall after VL/CL.

This study attempts to fill the gaps in knowledge about the interplay and cue-weighting of VM, VL, CL and VOT in Canadian English velar stop consonant voicing perception. The results demonstrate that, for Canadian English listeners and our stimuli dataset, VM has the strongest effect, followed by VL, and then CL. Where VOT places exactly along this spectrum is yet unclear, although it would be expected to fall right after VM. The strength of VM is surprising, since its presence or absence is not always congruent with phonological voicing in English. However, the effect of VM is reduced at low values (which are perceptually ambiguous between /g/ and /k/) and other cues increase their influence. These findings are important because they suggest that VOT is not, as classically reported, the major perceptual cue in English, at least when VM is dominantly present as a cue[5]. In the absence of VM cues it is likely that VOT would emerge as the dominant cue. Again, it is important to note that because they are significant interactions between the variables, the interpretation of their ranking is not so straightforward. Due to the complex manipulations to experimental the stimulus, the results should be understood to reflect phoneme perception in relative isolation, as opposed to a natural speech context. In

[5] Acoustic energy of the voicing maintenance was created at 25% of the vowel segment intensity. This value matches those reported in production studies for English. However, real-life values can occur below 25%, rendering the VM cue dominant in the stimuli.

continuous speech, the interplay of cues is likely much more complicated, with many more cues interacting than were included in this study.

Ultimately, this study contributes new evidence towards the perceptual cue-weighting in English stop consonant voicing and demonstrates that in the absence of a major perceptual cue (such as a facilitating burst and consequently an extractable VOT cue), multiple acoustic cues are used in combination with varying cue-weighting to provide consistent and stable stop voicing perception.

Bibliography

Abramson, A., & Whalen, D. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*(63), 75-86.

Audacity Team (2019) Audacity®. Version 2.3.2. Audio editor and recorder. Available from: http://audacityteam.org/ (Accessed 2018).

Baayen, H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. *Cambridge University Press, Cambridge*.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*(67), 1-48.

Boersma, Paul & Weenink, David (2019). Praat: doing phonetics by computer [Computer program]. Version 6.0.56, retrieved 20 June 2019 from http://www.praat.org/

Burnham, K., Anderson, D., & Huyvaert, K. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav Ecol Sociobiol*(65), 23-35.

Cho, T., & Ladefoged, P. (1999). Variation and Universals in VOT: evidence from 18 languages. *Journal of Phonetics*, 207-229.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Jounral of Phonetics*(61), 30-47.

Clayards, M. (2017). IndividualTalker andToken Covariation in the Production of Multiple Cues to Stop Voicing. *Phonetica, 75*, 1-23.

Francis, A., Baldwin, K., & Nusbaum, C. (2000). Effects of training on attention to acoustic cues. *Atten., Percept., Psychophys., 62*, 1668-1680.

Handbook of the International Phonetic Association: a guide to the use of the international phonetic alphabet. (1999). Cambridge: Cambridge University Press.

Harnad, S. (2003). Categorical Perception. In *Encyclopedia of Cognitive Science.* Nature Publishing Group: Macmillan.

Klatt, D. (1980). Software for a Cascade/Parallel Formant Synthesizer. *The Journal of the Acoustical Society of America, 67*(3), 971-995.

Klatt, D., & Klatt, L. (1990). Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers. *The Journal of the Acoustical Society of America, 87*(2), 820-857.

Li, F., Mennon, A., & Allen, J. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America, 127*(4), 2599-2610.

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang Speech, 29*(1), 3-11.

Lisker, L., & Abramson, A. (1964). A Cross-Language Study of VOicing in Initial Stops: Acoustical Measurements. *WORD, 3*(20), 384-422.

Nakai, S., & Scobbie, J. (2016). The VOT Category Boundary in Word-Initial Stops: Counter-Evidence Against Rate Normalization in English Spontaneous Speech. *Jounral of the Association for Laboratory Phonology, 1*(7), 1-31.

Nearey, T., & Rochet, B. (1994). Effects of Place of Articulation and Vowel Context on VOT Production and Perception for French and English Stops. *Journal of the INternational Phonetic Association, 24*(1), 1-18.

Oglesbee, E. (2008). Multidimensional stop categorization in English, Spanish, Korean, Japanese, and Canadian French. *Ph.D. thesis, Indiana University*.

Pape, D., & Jesus, L. (2014). Cue-weighting in the perception of intervocalic stop voicing in European Portuguese. *The Journal of the Acoustical Society of America*(136), 1334.

Pape, D., & Jesus, L. (2014). Production and Perception of velar stop (de)voicing in European Portuguese and Italian. *EURASIP Journal on Audio, Speech, and Music Processing, 6*.

Pape, D., Jesus, L., & Birkholz, P. (2015). Intervocalic fricative perception in European Portuguese: An articulatory synthesis study. *Speech Communication, 74*, 93-103.

Pape, D., Jesus, L., & Perrier, P. (2012). Constructing Physically Realistic VCV Stimuli for the Perception of Stop Voicing in European Portuguese. *Computational Processing of the Portuguese Language, edited by H. Caseli, A. Teixeira, and A. Villavicencio, Lecture Notes in Computer Sceince, vol 7243.*

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raphael, L. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America, 51*(4:2), 1276-1303.

Tuller, B., & Kelso, J. (1984). The timing of articulatory gestures. Evidence for relational invariants. *Journal of the Acoustical Society of America, 76*, 1030-1036.