VOWEL DURATION AND CONSONANT VOICING: A PRODUCTION STUDY

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Humanities

2019

Stefano Coretta
School of Arts, Languages and Cultures

Contents

Abstract

This dissertation focusses on the so-called “voicing effect”, by which vowels tend to be shorter when followed by voiceless stops and to be longer when followed by voiced stops, as exemplified by the English word pair *bat* vs *bad*. While the presence of this effect is cross-linguistically widespread, less is known about the source(s) of this phenomenon and competing accounts have been proposed over the decades. In this work, I draw from acoustic and articulatory data of Italian, Polish, and English and offer an overarching account of which aspects of the production of voiceless vs voiced stops, and vowel/consonant sequences in general, contribute to the emergence of the voicing effect.

The results indicate that the voicing effect is the product of a mechanism of compensation between the duration of the vowel and that of the following stop closure. The acoustic temporal relations of consonants and vowels observed in disyllabic (CV́CV) words of Italian, Polish, and English suggest that the duration of the interval between the release of the two stops is not affected by the voicing of the second stop. The release-to-release interval has similar duration in words with a voiceless C2 and those with a voiced C2. Within this temporally stable interval, the timing of the closure onset (the VC boundary) determines the duration of both the vowel and the stop closure.

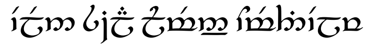
Ultrasound tongue imaging and electroglottographic data of Italian and Polish further show that the timing of the closure onset of voiced and voiced stops depends on articulatory factors related to the implementation of voicelessness and voicing. In particular, I argue that a delayed closure onset allows for enough tongue root advancement (known to facilitate voicing during closure) to be implemented during the production of the vowel in anticipation of the stop closure. Furthermore, glottal spreading typical of voiceless stops also can affect the timing of closure by anticipating the achievement of closure. These two factors, among other known factors, contribute to the observed pattern of short voiced closures and long preceding vowel duration, and, vice versa, long voiceless closures and short preceding vowel duration.

Declaration

I declare that no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgements



Elen síla lúmenn’ omentielvo.

A star shines on the hour of our meeting.

—J.R.R. Tolkien, The Fellowship of the Ring

I am for ever indebted to very many people, without whom I would not have seen the light at the end of this journey. A long journey that has witnessed bliss and sorrow, friends coming and parting, and lines of written words, over words, over words. But a joyful journey nonetheless, thanks to the cheer and comfort I was given by those who, in full or in good measure, walked next to me along such perilous path. Among these heroic figures, I wish to thank first and foremost my supervisors and mentors, Ricardo (Dr Bermúdez-Otero) and Patrycja (Dr Strycharczuk). With perpetual patience and kindness, they offered to me in countless occasions guidance and inspiration.

I also wish to thank my colleagues, as well as friends, from the Division and the Phonetics Lab, to which I owe so much in terms of time, thoughts, and laughter. Thanks to Simone, for the numerous chats about academia and life, for his moral support, and for the unicorn-shaped mug he had his coffee from, which gave to our time spent together a note of candour and light-heartedness. Thanks to Andrea (Dr Nini), who taught me about matters of study and of day-to-day living. Thanks must also go to Juliette, a companion of lunches and brunches, drinks and dinners, work sessions and pleasure trips, who brought constant light in cloudy Manchester and in my heart. Thanks to Jessica, for the lovely *chiacchierate* about all things and for her sparkling energy.

Thanks to the fellows I started this journey with: Stephen, to whom I am grateful

for his continuous help and for our creative exchanges; and Donald, to whom I am grateful for the cheerful wit and the appreciation for the good things in life. I have also received immense personal and academic support from George, whose exchanges of ideas were a continuous inspiration, and from Hannah, whose calmness and poise helped me cherish the good and the bad times. Last, but not least, a great thanks to all the other graduate students, some of whom were there before me, some of whom came later (Deepthi, Fernanda, Henri, Jane, Chris, Max, Lorenzo, Kai, Sarah, Mary, and many others whom I wished I had the chance to know better), and to all the other members of staff at LEL who contributed each with their own field to my broader understanding of linguistics. I am also indebted to Prof. Steven Lulich from Indiana University in Bloomington for hosting me in his Lab during October 2018 and letting me play with the 3D ultrasound machine (what a toy!), and to Prof. Małgorzata Ćavar for sharing her ideas on the tongue root. Thanks also to Sherman, Romario, Rebecca, Luma, and all the beautiful people in Bloomington that accompanied me for too short a time along the PhD path.

I would have not made it through without all of my friends in Manchester, especially (in order of appearance, because they are all dearest to me) David, Juan, Richard and Jaqui, Mark and Kat, and Crystal, who offered me unconditional friendship and help when I most needed them. Also thanks to Mariana, Maxime, Gül, Valentina, Ximena, Lucia, Edo, and to the friends I met in York but whom I found again in Manchester, thank you all for the great time.

Thanks to my teachers and friends from *Il Loto Blu*, whose love has sustained me for all these years. Thanks to Valentina A., with whom I share an irrepressible thrust to "acquire for ourselves a knowledge of higher worlds." Special thanks go to my friends in Verbania, who supported me and stood me (*supportato e sopportato*) since the dawn of time, and to Mattia, *sorella mia*, whose friendship, *in præsentia et in absentia*, has gracefully marked my University years from day one. Thanks to Sig.ra Racchelli, whose help and care provided me with the resources to survive difficult times. Thanks to my loving family on Earth and beyond, and a special thanks to my mother, whose continuous sacrifice encouraged me to keep on moving forward.

The tutelage and support of these and many more people came as those of *Kṛṣṇa*

upon *Arjuna* in the *Bhagavad-gitā*, during the battle at *Kurukṣetra*. Without their direction and kindness, I would have lost my way and strayed, never to find again the will to persist along the road. For all of you, a thousand thoughts, then a hundred, then another thousand, then a second hundred, then yet another thousand, then a hundred more.

असंशयं महाबाहो मनो दुर्निग्रहं चलम् ।
अभ्यासेन तु कौन्तेय वैराग्येण च गृह्यते ॥ ६-३५॥

asaṃśayaṃ mahābāho mano durnigrahaṃ calam |
abhyāsena tu kaunteya vairāgyeṇa ca gṛhyate || 6-35 ||

O mighty-armed one, it is undoubtedly very difficult to curb the flickering mind, but it can be controlled, O son of Kunti, by constant practice and by detachment. [BG 6.35]

Part I

Introduction

There is in all things a pattern that is part of our universe. It has symmetry, elegance, and grace—these qualities you find always in that the true artist captures. You can find it in the turning of the seasons, the way sand trails along a ridge, in the branch clusters of the *creosote* bush or the pattern of its leaves. We try to copy these patterns in our lives and in our society, seeking the rhythms, the dances, the forms that comfort. Yet, it is possible to see peril in the finding of ultimate perfection. It is clear that the ultimate pattern contains its own fixity. In such perfection, all things move towards death. —*from* The Collected Sayings of Muad'Dib *by the Princess Irulan*

— Frank Herbert, *Dune* (1965)

A careful analysis of the process of observation in atomic physics has shown that the subatomic particles have no meaning as isolated entities, but can only be understood as interconnections between the preparation of an experiment and the subsequent measurement.

—Fritjof Capra, *The Tao of Physics* (1975)

Chapter 1

The voicing effect and beyond

The sounds of a language form an incredibly complex system of relations and dependencies, both at a physical and a more abstract level. A topic that masterfully exemplifies the complexities of such a system and that has generated great interest over the decades is the somewhat elusive connection between vowel duration and consonant voicing. According to a robust cross-linguistic tendency, vowels are shorter when followed by a voiceless consonant and longer when the following consonant is voiced (Meyer 1903; Heffner 1937; House & Fairbanks 1953; Lisker 1957; Peterson & Lehiste 1960). This so-called "voicing effect" interacts with a variety of linguistic factors and scholars have sought its origins in properties of speech production, from aerodynamic mechanisms to gestural timing, and properties of speech perception (Belasco 1953; Zimmerman & Sapon 1958; Sharf 1962; Lindblom 1967; Halle et al. 1967; Javkin 1976; Kluender et al. 1988). While much progress has been made in understanding this link, after more than a century there is still disagreement as to what contributes to this phenomenon, as evidenced by the numerous accounts put forward.

Given the plurality of views concerning less understood aspects of the voicing effect, this thesis set out to investigate this phenomenon by employing a diverse set of techniques and sources of data. To keep this type of enquiry manageable, I decided to undertake this endeavour from a speech production outlook, an area which has fuelled a great part of the debate within the voicing effect literature. In particular, this thesis poses the question of what aspects of the articulation of vowel-consonant sequences can inform us about the influence of consonant voicing on the duration of vowels. In

answering this question, I collected data from a combination of acoustic, ultrasound tongue imaging, and electroglottographic techniques as part of two studies on Italian, Polish, and English. These languages make up an appropriate set in that they constitute a methodological window into the complex variation of the voicing effect as seen both across and within languages.

The dissertation is organised in three parts: an introduction (Part I), a collection of original manuscripts (Part II), and a conclusion (Part III).

The three chapters of Part I present ~~in turn~~ a review of the literature on the voicing effect and related issues (Chapter 1), a rationale for the current research including a discussion of the questions to be addressed (Chapter 2), and a description of the methodologies employed in the studies that make up this research (Chapter 3). The following sections introduce the phenomenon of the effect of consonant voicing on preceding vowel durations (the "voicing effect"). First, I will discuss how the voicing effect is cross-linguistically common (with the typological caveat that most investigated languages are from the Indo-European family), although alleged exceptions to its universality exist, both in terms of presence of the effect and of its magnitude (Section 1.1). This is followed by a discussion of the use of "voicing" as a comparative concept rather than as a phonetically-motivated descriptive category (Section 1.2), and by a presentation of other phonological and phonetic factors that are known to interact with the voicing effect, such as manner, prosody (Section 1.3), and processes of phonologisation (Section 1.4). The chapter proceeds with a critical review of the explanatory accounts proposed for the voicing effect, both from a production and a perception point of view (Section 1.5). ~~Finally,~~ the chapter concludes with a discussion of the effects of aspiration and ejection on vowel duration, and how these can shed light on the voicing effect (Section 1.6).

Chapter 2 provides a rationale for the current research. The research questions addressed in the dissertation are introduced and contextualised in relation to the topics touched upon in Chapter 1. A justification of the choice of data sources and languages used to answer the research questions is given here. This chapter also offers an overview of the phonologies of the chosen languages, namely Italian, Polish, and English. For each language, a brief description of the consonantal and vocalic phonemic systems is

given, with a focus on aspects of phonation contrasts, followed by a discussion of stress and rhythm. Section 2.3 offers a prospective overview of the main results, which are presented in full in Part II.

Chapter 3, the last of Part I, collates the methods employed in the studies that make up this research, namely an exploratory study of the voicing effect in Italian and Polish (Study I, Section 3.1) and a confirmatory study of the compensatory aspects of the voicing effect in English (Study II, Section 3.2). Note that each paper in Part II (Chapter 4 to Chapter 7) contains targeted methods sections that describe the subset of methods specific to the paper, so that a general overview of the methods is provided in Chapter 3. This chapter also introduces ultrasound tongue imaging and electroglottography, two articulatory techniques that allow us to learn in a non-invasive way about properties of tongue movement and vocal fold vibration. Finally, Section 3.3 discusses issues related to statistical methods, introduces principles and practices of Open Science as a remedy to some of these issues, and shows how Open Science has shaped the current research project.

Part II is a collection of original manuscripts in the form of standalone papers (Chapter 4 to Chapter 7), which report and discuss the conceptual and methodological contribution of the present work. The papers are connected in that they investigate related but self-contained aspects of the voicing effect. A "journal format" was chosen over a "book format" given the strong experimental and methodologically independent nature of the research behind each paper, and thanks to the fact that each can be read more or less independently of the others. Nonetheless, the four papers are laid out according to an order partly based on the chronological sequence of the research but also considering the logical ~~dependence~~ of the hypotheses ~~treated~~ in them. While the papers in Chapter 4 to Chapter 6 have a more conceptual focus, Chapter 7 centres around a novel methodological approach that enables a holistic analysis of vocal fold vibration data as obtained from electroglottography.

Chapter 4 (Paper I) describes an exploratory study of acoustic properties of the voicing effect in Italian and Polish disyllabic words, as investigated in Study I. Durational aspects of the voicing effect as evinced from acoustic data are surveyed in light of compensatory mechanisms between the duration of vowels and that of consonant closures.

This paper also provides a modern description of the voicing effect in Italian and Polish and discusses how the current results match or diverge from previous work.

The findings of Chapter 4 motivated a confirmatory study, Study II, which is described and discussed in Chapter 5 (Paper II). An articulatory account inferred from the acoustic data presented in Chapter 4 is proposed, which generates hypotheses regarding the durational behaviour of disyllabic vs monosyllabic words. These hypotheses, formulated in terms of acoustic durational patterns, are tested against acoustic data from English disyllabic and monosyllabic words. The paper also links differences in magnitude of the voicing effect in di- vs monosyllabic words to the interplay between the articulatory organisation of gestures and perceptual factors.

Two more papers present articulatory aspects of the voicing effect in Italian and Polish from Study I. This part of the study was carried out to explore voicing-driven differences in articulation during the production of vowel/consonant sequences that could favour the emergence of the voicing effect. Chapter 6 (Paper III) discusses ultrasound tongue imaging data and focusses on tongue root advancement, a mechanism known to facilitate voicing during closure. Both the static configuration of tongue root advancement at vowel onset and its dynamic development during the production of vowels followed by voiceless and voiced stops are discussed. Furthermore, the relation between the static and dynamic properties of tongue root advancement, vowel duration, and consonant voicing is studied. Chapter 7 (Paper IV) assesses a new technique for the dynamic analysis of electroglottographic data which combines established statistical methods. The application of this method is illustrated with an electroglottographic analysis of Italian and Polish, which investigates how vocal fold vibration during the production of vowels differs depending on the voicing status of the following consonant. Finally, the findings of this analysis are discussed in light of the voicing effect and how glottal spread, characteristic of voiceless consonants, might play a role in the emergence of the effect.

Part III summarises the results of this investigation by providing an overarching synthesis (Chapter 8) in response to the questions outlined in Chapter 2, and concludes with a discussion of limitations and future avenues of research (Chapter 9).

1.1 The voicing effect

Across a wide variety of languages, vowels tend to be shorter when followed by voiceless consonants, and longer when followed by voiced ones. This phenomenon has been called the "voicing effect" (Mitleb 1982) or "pre-fortis clipping" (Wells 1990). Among the earliest traceable mentions to this phenomenon there are Meyer (1903) for English (cited in Lindblom 1967), Meyer (1904) for German, Meyer & Gombocz (1909) for Hungarian, and Gregoire (1911) for French (all cited in Maddieson & Gandour 1976). After these, a great number of studies further confirmed the existence of the effect in these languages and reported it in an ever increasing list of others. Remarkably, no known language has been claimed to have the opposite effect, namely longer vowel durations before voiceless than before voiced consonants.[1]

English is the language that by far received the most attention in relation to the voicing effect (Heffner 1937; House & Fairbanks 1953; Lisker 1957; Zimmerman & Sapon 1958; Peterson & Lehiste 1960; House 1961; Sharf 1962, 1964; Lindblom 1967; Halle & Stevens 1967; Halle et al. 1967; Slis & Cohen 1969a,b; Chen 1970; Klatt 1973; Lisker 1974; Raphael 1975; Umeda 1975; Javkin 1976; Port & Dalby 1982; Mack 1982; Luce & Charles-Luce 1985; Summers 1987; Kluender et al. 1988; de Jong 1991; Laeufer 1992; Fowler 1992; de Jong 2004; Warren & Jacks 2005; Ko 2018; Glewwe 2018; Sanker 2019, among others). The presence of a voicing effect has been further corroborated in French by Belasco (1953), Chen (1970), and Laeufer (1992), in Hungarian by Sóskuthy (2013), and German (in the context of word-final voicing neutralisation, see Nicenboim et al. 2018 and references therein). Other known voicing-effect languages are Arabic (Hussein 1994, but cf. Mitleb 1982), Assamese and Bengali (Maddieson 1976), Dutch (Slis & Cohen 1969a), Georgian (Beguš 2017), Hindi (Maddieson & Gandour 1976; Ohala & Ohala 1992; Lampp & Reklis 2004; Durvasula & Luo 2012; Sanker 2018), Italian (Magno Caldognetto et al. 1979; Farnetani & Kori 1986; Esposito 2002), Icelandic (Einarsson 1927), Japanese (Port et al. 1987), Korean (Chen 1970), Lithuanian (Campos-Astorkiza 2007), Norwegian (Fintoft 1961), Swedish (Elert 1970), Spanish (Navarro Tomás 1916), Telugu (Sanker 2018), and Russian (Chen 1970).[2]

[1] This does not exclude that there might be or have been a language that shows this pattern.

[2] From a typological perspective, this sample is strongly biased towards the Indo-European language

While the voicing effect is cross-linguistically common, it is not universal, and some languages lack voicing-induced durational differences. Czech and Polish are generally reputed to be languages in which the duration of vowels does not significantly differ before voiceless and voiced stops. In fact, the results concerning the effect in these languages are mixed, and support can be found both for and against an effect of voicing on vowel duration. Keating (1984b) examined the duration of vowels in 3 Czech speakers. Vowels are 193.7 ms long when followed by /t/ and 204.2 ms when followed by /d/. This corresponds to a raw difference of 10.5 ms, which the author reports not to be significant ($t(30) = -0.37$, $p > 0.2$). Given the low number of speakers and the relative high standard error of the effect (about 28 ms, calculated as the mean difference over the *t*-value, Nicenboim et al. 2018), it is possible that the null result is due to low statistical power. Machač & Skarnitzl (2007) analyse 638 VCV sequences recorded from 53 speakers of Czech and found partial evidence for an effect of voicing in the language.

As for Polish, Slowiaczek & Dinnsen (1985) measures the duration of vowels in word-final syllables from 5 speakers, and vowels followed by an underlyingly voiced stop are 10–15 ms longer. Nowak (2006) investigates several properties of vowel duration in 4 speakers (from different parts of Poland), and finds that vowels followed by voiced stops are 4.5 ms longer (a significant difference). Malisz & Klessa (2008) analyse data from 40 speakers of "Standard Polish", and while they don't report estimates from the whole dataset, the means from 4 speakers suggest a difference in vowel duration before voiceless vs voiced stops of about 3.5 ms. On the other hand, an equal number of studies argue that voicing does not significantly affect vowel duration in Polish. Jassem & Richter (1989) do not replicate the results in Slowiaczek & Dinnsen (1985). Keating (1984b) reports a non-significant difference of 2 ms in the word pair /rata/ (167.4) and /rada/ (169.5 ms), based on data from 24 speakers living in Wrocław. Finally, Strycharczuk (2012) reports a non-significant effect in 6 Warsaw speakers in pre-sonorant word-final position. To summarise, the evidence concerning the presence or absence of the voicing effect in Czech and Polish is mixed and it is not possible to

family. Moreover, the five non-Indo-European languages in the list are all thriving and well-studied. This notwithstanding, the voicing effect is generally regarded as a very common and widespread phenomenon.

draw firm conclusions.

A second common stance about the voicing effect is that its magnitude differs across languages, and that the greatest effect is observed in English. The reported effect of voicing in word-final syllables in English varies between 35 and 150 ms (Heffner 1937; House & Fairbanks 1953; Zimmerman & Sapon 1958; Peterson & Lehiste 1960; Sharf 1962; Chen 1970; Klatt 1973; Mack 1982; Luce & Charles-Luce 1985; Laeufer 1992; Ko 2018). However, the effect is smaller in non-final syllables, with values between 18 and 35 ms (Sharf 1962; Klatt 1973; Davis & Summers 1989). Taking Italian for comparison, the mean difference in vowel duration before voiceless vs voiced stops in the first syllable of Italian disyllabic words is 22 in Farnetani & Kori (1986) and 24 ms in Esposito (2002). These values are within the range of the reported effect in English non-final syllables. It is thus possible that, once controlling for contextual factors, the apparent cross-linguistic differences in magnitude are, if not removed, at least reduced. A similar position is taken by Laeufer (1992), who directly compares French and English using carefully designed experimental materials. When the duration of a vowel is similar across languages, consonant voicing also has an effect which is comparable in degree.

1.2 Voicing as a physical property and as a linguistic category

The term “voicing”, as used in the literature on the voicing effect and related phenomena, can have substantially different referents. A first major distinction can be drawn between voicing as a physical property and voicing as a linguistic (abstract) category of lexical contrast. Within each of these two classes, further distinctions are possible. This section will review the physical and the linguistic sense in turn. After providing a physical definition of voicing as periodicity and vocal fold vibration, I will discuss some of the views on how voicing can be defined linguistically. Finally, I will introduce the notions of “descriptive category” and “comparative concept” (Haspelmath 2010) to show that the linguistic reading of voicing in this dissertation will be in the latter sense.

From a physical point of view, voicing can be defined either acoustically or physio-anatomically. Acoustically, voicing is the presence of periodicity in a speech signal. The physio-anatomical source of acoustic periodicity is the cyclic vibration of the vocal folds. A speech interval that is characterised by vocal fold vibration/periodicity is said to be voiced, while an interval that does not have vocal fold vibration is said to be voiceless. The terms “voiced” and “voiceless” in this sense refer to the physical properties of the speech interval.

The initiation of vocal fold vibration requires that the air pressure of the cavity below the vocal folds (broadly speaking, the lungs) is higher than that of the cavity above them (the oral tract). The positive trans-glottal air pressure differential is also necessary for the vibration to continue after it is initiated (van den Berg 1958; Rothenberg 1967). This property is formally known as the Aerodynamic Voicing Constraint (Ohala 2011). Vocal fold vibration in which no active articulatory adjustment is used to ensure the pressure differential is called *passive voicing*. The typical class of sounds characterised by passive voicing are sonorants (vowels, nasals, liquids). Similarly, the absence of voicing with no concurrent adjustments to prevent initiation and maintenance of vocal fold vibration is known as *passive devoicing*. Passive devoicing can be observed in the voiceless closure of stops, while a sign of passive voicing is the continuing presence of vocal fold vibration from the preceding voiced sound for some time into the closure (also know as voicing bleed, Davidson 2016).

Certain articulatory conditions, like a reduced oral tract volume or the full closure of stops consonants, hinder passive voicing by reducing the trans-glottal pressure differential. When the pressure differential is 0 (i.e. when pressure equalisation is reached), vocal fold vibration can no longer be maintained and it ceases. In such articulatory conditions, several articulatory adjustments can be implemented to counteract pressure equalisation. Active adjustments require muscular activity. Among the solutions to help sustaining vocal fold vibration there are (1) some that decrease supra-glottal pressure like tongue root advancement (Kent & Moll 1969; Perkell 1969; Westbury 1983), larynx lowering (Riordan 1980), opening of the velopharyngeal port (Yanagihara & Hyde 1966), or producing a retroflex occlusion (Sprouse et al. 2008), and (2) others that lower the pressure differential threshold required for fold vibration, like slackening of

the vocal folds (Halle & Stevens 1967) or producing a shorter stop closure (Lisker 1957). *Active voicing* is vocal fold vibration with articulatory adjustments that ensure continuing vibration.

Sounds that are intended not to have vocal fold vibration (i.e. are intended as voiceless) but that are characterised by favourable conditions for it tend to show passive voicing. In these cases, articulatory adjustments can be put in place to counteract the presence of passive voicing. For example, glottal abduction, larynx raising, vocal fold tensing and oral wall tensing can all prevent voicing by either decreasing the supraglottal volume or raising the pressure differential threshold. The resulting phenomenon is called *active devoicing* (Jansen 2004).

Rothenberg (1967) makes an important further distinction between purposive and non-purposive active articulatory adjustments. For example, tongue root advancement can be executed by muscular activity with the intent to maintaining voicing, in which case we would call it a *purposive* (active) gesture. If tongue root advancement is executed by muscular activity with an intent different from maintaining voicing (for example, aiding the creation of a tongue constriction movement), then this would be classified as a *non-purposive* (active) gesture in regards to voicing. While ascertaining whether an active gesture is purposive or non-purposive can be difficult, an active articulatory adjustment (executed by muscular activity) does not automatically imply the speaker's intention to achieve all of the benefits deriving from that adjustment. While discussing articulatory adjustments in this dissertation, a classification between passive, active, purposive and non-purposive adjustments will not be attempted, but the potential difference will be discussed when relevant. Finally, the terms movement, adjustment, and gesture will be used interchangeably throughout without any theoretical commitment to differences among these.

Turning now to the linguistic sense of voicing, different approaches have been proposed on how to classify phonological systems of phonation contrasts. This discussion will focus on systems that contrast *two* categories, and only on those approaches to voicing that are relevant to topics at matter. In particular, I will review three main ideas: the distinction between voicing and aspirating languages (Beckman et al. 2013), the tri-partite system of phonological categories (Keating 1984a), and the Articulatory

Phonology definition of voicing (Goldstein & Browman 1986). Secondly, I will show that a typological approach that distinguishes between language-specific and comparative entities spares us the need to find a definition of voicing that can simultaneously account for the different phonological systems. This approach forms the basis of the conceptual background of this dissertation.

The voicing-effect languages listed in Section 1.1 possess quite different phonation systems. A major distinction can be drawn between so-called "true voicing" languages and "aspirating" languages (Beckman et al. 2013). True voicing languages make use of the distinctive (privative) feature [voice], and segments specified with [voiced] are characterised by active voicing (vocal fold vibration). On the other hand, aspirating languages employ the feature [spread glottis]. Segments specified with [spread glottis] generally have long Voice Onset Time (VOT) values, while unspecified segments can show passive voicing (vocal fold vibration). Typical true-voicing languages are Italian, Spanish, and Russian, while Germanic languages like German, English, and Icelandic are aspirating languages. All of these languages are, even if at an allegedly different extent, voicing-effect languages. In true-voicing languages, vowels are longer when followed by [voice] segments, while in aspirating languages vowels are longer when followed by underspecified segments (segments without [spread glottis]).

Keating (1984a) is an attempt to define "voicing" in such a way that even systems that are very dissimilar at the physical level can be grouped together. This definition is restricted to languages that contrast only two categories of phonation. Keating (1984a) proposes that three levels of representation are necessary. One level is purely phonological, and abstracts away from real physical properties of the contrast. This phonological level of representation corresponds to the traditional [(±)voice] feature (either binary or privative). The second level of representation pertains to what Keating (1984a) calls "modified systematic phonetics." She proposes three phonetic categories based on VOT: {voiced}, {voiceless unaspirated}, and {voiceless aspirated}. The last level, "pseudo-physical", assigns a range of VOT values to such categories depending on the language and the phonological context. The [(±)voice] feature can then be interpreted as "more" or "less voiced", rather than as presence vs absence of vocal fold vibration.

Goldstein & Browman (1986), within their framework of Articulatory Phonology,

take a different stance and ascribe voicing to simply the presence or absence of a glottal opening-and-closing gesture. Voiceless segments are then characterised by the presence of such a gesture, while voiced segments by its absence. This definitions abstracts away from the presence/absence of vocal fold vibration, and allows us to group together for example an aspirating language like English and a true-voicing one like Italian.

While the three approaches just reviewed try to posit categories that can both describe and categorise phonation systems, the typological approach adopted here keeps these two aims separate. Haspelmath (2010) introduces a helpful distinction between comparative concepts and descriptive categories. Following from what Haspelmath calls "categorial particularism", it is advocated that individual languages should be described in terms of language-specific categories. In this light, "voicing" in Italian is different from "voicing" in Spanish for the very fact that Italian and Spanish are two different linguistic systems. Typological comparison should not be based on (language-specific) "descriptive categories", but rather on "comparative concepts." Comparative concepts are created by the linguist who performs cross-linguistic analyses, and are not components of particular languages. I assume here that traditional phonological categories like "voicing", "vowel height", "place of articulation", can be thought of either as (language-particular) descriptive categories or comparative concepts, depending on the scientific enterprise.[3] In relation to the voicing effect, the use of "voicing" in this context as adopted in this work will be intended as a comparative concept and not as a descriptive category.

Note that the adoption of the distinction between descriptive categories and comparative concepts is a working assumption and a fully fledged argument in support of the distinction will not be pursued here. Rather, reference to comparative concepts here allows us to compare the voicing effect across languages where "voicing" behaves very differently, and allows us to make cross-linguistic generalisations that transcend language-specific descriptive categories. The holistic account expounded in Section 8.1 rests on this working assumption, and should be interpreted as applicable independently

[3]Haspelmath (2010) proposes to use capitalised names for descriptive categories (for example, "Italian Voicing"), but since this use is not common among phonologists/phoneticians, I will not adopt it here.

from language-specific voicing categories. I will not specify the sense of the term “voicing” when used (physical, descriptive, comparative), unless in cases where its interpretation is ambiguous.

1.3 The voicing effect and other phonological and phonetic factors

In Section 1.1, we saw that the voicing effect can differ depending on the language. In addition to language, this phenomenon is also modulated by other phonological and phonetic factors. For example, Umeda (1975) reports that the difference in vowel duration before voiceless vs voiced consonants is greater when the test word is pre-pausal. The effect of voicing also seems to be more robust in stressed than in unstressed vowels (Davis & Summers 1989). There is also indication that the effect is modulated by the position of the syllable in the word in English, so that word-final syllables show a greater effect than word-medial syllables (Sharf 1962; Klatt 1973; Davis & Summers 1989, although Abdelli-Beruh 2004 doesn’t find a significant difference across these contexts in French). Port (1981) further argues that the effect in word-initial stressed vowels is smaller along the hierarchy monosyllabic > disyllabic > trisyllabic words, which also reflects that of decreasing average vowel durations. Laeufer (1992) discusses the voicing effect as a function of vowel height, and shows that the effect is greater in low (intrinsically longer) vowels than in high (intrinsically shorter) vowels. Moreover, Sharf (1964) shows that the effect persists even in whispered (unvoiced) speech.

Manner of articulation of the consonant is a further relevant parameter. While most work seems to focus on stops, voicing of other types of consonants affects preceding vowel duration. For example, House & Fairbanks (1953) report that vowels are longer when followed by a voiced fricative than a voiceless one in English. They also argue that the durational difference is greater before fricatives than before stops. On average, vowels in House & Fairbanks (1953) are 84 ms longer when followed by a voiced stop (vs a voiceless stop) and 93 ms longer when followed by a voiced fricative (vs a voiceless fricative). Laeufer (1992) finds similar patterns in both English and French: vowels

followed by voiced fricatives are longer than when followed by voiceless fricatives (the average difference is 93 ms in English, 47 ms in French) and the effect of voicing with fricatives is greater than with stops (the average difference is 60 ms in English stops, 35 ms in French stops). Zimmerman & Sapon (1958) report vowel durations before voiceless and voiced stops and fricatives in English, and the difference is greater in the latter (95 vs 122 ms). On the other hand, Tanner et al. (2019) ~~rather~~ find the opposite effect in a survey of spontaneous speech from different varieties of English. Across English varieties, the effect of voicing is greater with stops than with fricatives by a mean factor of 1.3. To sum up, it is possible that the degree of the voicing effect is greater in fricatives than in stops, but it is difficult to make generalisation based on such small pool of studies.

The relation between the voicing effect in obstruents and durational effects of sonorant consonants further indicates mixed results. While only a few systematic investigations on the effect of sonorant voicing on vowel duration have been carried out, it was found that (1) nasals exercise an effect intermediate between that of voiceless and voiced stops but closer to that of the latter (House & Fairbanks 1953; Zimmerman & Sapon 1958); (2) nasals are preceded by vowels that are longer than those followed by voiced stops (Peterson & Lehiste 1960); or (3) the duration of vowels followed by nasals is indistinguishable from that of vowels followed by voiced stops (Lisker 1974). In House & Fairbanks (1953), vowels are on average 245 ms long when followed by a voiced stop, 232 ms long when followed by a nasal, and 161 ms when followed by a voiceless stop. Zimmerman & Sapon (1958) report English vowel durations of 218 ms before voiced stops and 200 ms before nasals, while vowels are 123 ms long when followed by voiceless stops. On the other hand, the duration of vowels in Peterson & Lehiste (1960) are 273 ms when followed by nasals, 265 ms when followed by voiced stops, and 171 ms when followed by voiceless stops. Lisker (1974) argues that the duration of vowels followed by voiced stops and nasals are virtually the same, but measurements are not provided. In sum, nasals seem to behave more like voiced stops than voiceless stops, but it is less clear whether vowels preceding them are longer or shorter than those followed by voiced stops.

1.4 On the phonologisation of the voicing effect

The voicing effect can take on a linguistic function resulting in the phonologisation of the durational differences, as argued for English (de Jong 1991, 2004; Solé et al. 2007; Sanker 2019). Some clarification is due here as to what is meant by phonologisation. The classical or structuralist definition of phonologisation states that this occurs when a contextual allophone becomes contrastive, or in other words it becomes a phoneme (Kiparsky 2015), generally after the disappearance or replacement of the conditioning context. Sanskrit velar palatalisation is a classical example of phonologisation (Hock 1991:149). At some point in the history of Sanskrit, the velar stops /k/ and /g/ where palatalised when followed by /i/ and /e/, creating an allophonic distinction between velars proper and palatal consonants of some sort. The subsequent change of /e/ to /a/ removed the context conditioning palatalisation, thus creating minimal pairs opposing /ka, ga/ and /tʃa, dʒa/. At this stage, the palatal allophones were phonologised. This conceptualisation of phonologisation amounts to saying that phonetic features that were previously computed procedurally (during phonological/phonetic derivation) from an underlying lexical representation are now instead already part of the lexical representation (which is, in structural terms, a string of phonemes).

Phonologisation assumes a different meaning within the framework of Lexical Phonology (Kiparsky 1988). Lexical Phonology argues that there exist two types of phonological processes: processes that apply at the lexical (stem and prosodic word) level, and processes that are post-lexical and apply across the board. According to the view of Lexical Phonology, a process is phonologised when it goes from being post-lexical to being lexical. To carry on with the Sanskrit example, phonologisation was initially post-lexical, in other words it was applied across the board during derivation after all lexical processes have been applied to the stem and word. During the course of sound change, the same process of velar palatalisation started being applied also at the lexical level (with the original copy of the process possibly still being applied post-lexically). Velar palatalisation has been phonologised, creating so called “quasi-phonemes” (categorical, distinctive units, not yet able to create lexical contrast, Janda 1999).

Kiparsky (2000) carries over the definition of phonologisation from Lexical Phonology onto Stratal Optimality Theory (Kiparsky 2000; Bermúdez-Otero 2017). Stratal OT assumes that the phonological module of grammar is stratified into three levels (called strata, or domains) as in Lexical Phonology: the stem, the word, and the phrasal level. OT constraints are independently ordered in each level, so that within each level different orders allow for different outputs to be selected. Stratal OT also stipulates that phonological constraints apply iteratively (cyclically) from the narrower domain, namely the stem, through the word domain, to the phrasal domain. Under cyclicity, the input of one domain is passed over to the next, and so on. For Kiparsky (2000), phonologisation occurs when the constraint ordering of the phrasal domain (the post-lexical level of Lexical Phonology) is carried over to the word and stem domains (the lexical level of Lexical Phonology).

An extension of Stratal OT, the life cycle of phonological processes (Bermúdez-Otero 2007, 2015), offers yet another definition of phonologisation and a more fine-grained terminological set. Bermúdez-Otero (2015) reserves the term "phonologisation" for when a physico-physiological (phonetic or mechanic) phenomenon comes under the control of the speaker/hearer and in fact becomes part of her grammar (more specifically, part of the phonetic module of the grammar). The process, once it has entered the grammar, can further its "ascent" through increasingly deeper grammatical modules. A (gradient) phonologised process is said to be "stabilised" (and thus categorical) once it is generated by a categorical phonological rule, which applies at the phrase level. At this stage, a stabilised process has entered the phonological module of the speaker/hearer. A stabilised process further undergoes "domain narrowing" when it starts being applied at the word level and then at the stem level. In the final step in the ascent of a sound pattern through the grammar, a phonological process comes under morphological and lexical control, until "it may die altogether, leaving behind no more than inert traces in underlying representations" (Bermúdez-Otero 2015:12).

A further definition of phonologisation stems from exemplar theories of speech perception and production (Johnson 1997; Pierrehumbert 2001; Sóskuthy et al. 2018; Ambridge 2018; Todd et al. 2019). A core tenet of these models is that speech tokens are stored in memory as so-called exemplars after having being experienced. Depend-

ing on the specifics of the particular model, exemplars are stored at varying degrees of granularity and richness of detail. Each exemplar consists of a (more or less) faithful representation of the actual token of experience that generated it, and it thus contains information from multiple levels and factors (phonetic, lexical, syntactic, sociolinguistic, contextual, and so on). Lexical and other linguistic units are represented as sets of exemplars, or exemplar clouds. The representational space of exemplar clouds is multi-dimensional and can be operationalised as a multivariate distribution. In modular approaches to grammar as briefly expounded above, sound alternations can be encoded (in terms of derivational rules and/or constraints) either at the phonological level or at the phonetic level of representation. On the other hand, as Sóskuthy (2013:183) illustrates, a consequence of the exemplar mode of representation is that all sound alternations are directly encoded by exemplars within the exemplar cloud, at one single level of representation. As soon as an exemplar with new phonetic characteristics is experienced and stored, the lexical representation of that lexical item already contains information on the sound alternation. In this sense, every type of variation is "phonologised" (represented) from the outset as soon as it is experienced by the speaker/hearer and stored in memory.

When the term phonologisation is employed in the phonetic literature of the voicing effect, it is generally not attributed to any specific phonological framework. This makes it less straightforward to interpret the term as the original author might have intended, but, as far as I can tell, most authors would interpret it at least as to mean that the effect is not just mechanical and/or low-level, but that it has assumed higher-level functions of some sort, whatever the specific function might be. Since the main focus of this work is on the source of the voicing effect rather than on what functions the voicing effect can assume in different languages, the topic of the phonologisation of the voicing effect will only briefly be touched upon in the rest of the dissertation. Note, however, that the account proposed in Chapter 8 is envisioned to be informed by some form of exemplar model of speech perception and production, where everything can be considered "phonologised" as soon as it is part of the lexical representation. In this sense, the effect is represented in the same way in all languages that have it, independent of its magnitude or function. A discussion of arguments for or against such

position are, however, beyond the scope of this dissertation.

Going back to the phonologisation (in the general sense) of the voicing effect in English, de Jong (2004) shows that the effect is greater in stressed syllables and under focus in English but not in Arabic (de Jong & Zawaydeh 2002), and argues that vowel duration is used contrastively as a cue to voicing in the former language. A further argument for the phonologisation of the durational difference in English is the stability of the effect across speaking tempos. Port & Dalby (1982) suggest that the ratio of the consonant and vowel durations is stable at faster and slower speaking rates, and that the CV ratio proves to be the primary acoustic correlate of voicing in word-final position. Luce & Charles-Luce (1985), however, claim that vowel duration is a more robust cue across tempos than the CV ratio and the duration of the stop closure. Finally, Ko (2018) compares CV ratio values in three speaking styles (normal, faster, and slower) and finds that the ratio changes as a function of speaking style and that the effect of style interacts with consonant voicing. In sum, there is contrasting evidence as to whether the relative magnitude of the effect is stable across speaking tempos or not, and as to whether this can be taken as evidence for or against the phonologisation of the effect in English.

The mechanisms behind the *emergence* of the voicing effect are in principle independent from those driving the subsequent phonologisation of the effect. In light of this, the next section reviews different proposals of what the source of the voicing effect might be, while leaving aside the further question of how the effect can be exploited phonologically once in place.

1.5 One phenomenon, many explanations

Over a century of research on the voicing effect has without doubt brought progress in our understanding of this complex phenomenon. While several proposals were put forward in the period between the 50s and the 70s, subsequent years focussed on testing or extending previous hypotheses and no final consensus has been reached. A broad distinction can be drawn between accounts that ascribe the voicing effect to articulatory or aerodynamic properties of speech production, and accounts that instead draw on biases of the perceptual system. No answer has been obtained as to which of the two

sides best accounts for all of the aspects of the voicing effect, and rather both views contribute in some respect to the overall picture. The following paragraphs review the most notable perception and production accounts, paving the way for a discussion of phonation effects related to that of voicing in the following section.

A perceptual-based explanation advocated by Javkin (1976) argues that the voicing effect emerges as a consequence of the difficulty in the perceptual identification of the vowel-consonant boundary in the context of voiced stops, and of the misinterpretation of voicing during closure. According to this account, speakers misperceive the periodic vibration of the vocal folds (voicing) during the closure of a voiced stop as being part of the preceding vowel. In the absence of contextual correction, this misperception can lead to the creation of a new production norm where the vowel is lengthened (Ohala 1989). Subsequent productions of vowels followed by voiced stops would thus be longer than vowels followed by voiceless stops. Although Javkin (1976) does not directly test the hypothesis that closure voicing is reinterpreted as being part of the preceding vowel, his study indicates that listeners perceive vowels to be longer when followed by voiced than when followed by voiceless stops, other things being equal. On the other hand, Sanker (2019) finds that vowels followed by voiced stops rather elicit fewer "long" responses, while more "long" responses are elicited in stimuli where the following consonant was spliced out. However, listeners were perceiving vowels with falling F0 to be longer than vowels with flat or raising F0, in partial accord with previous work (Lehiste 1976; Yu 2010; Cumming 2011).

To provide for a rationale of the language-specificity of the voicing effect, Kluender et al. (1988) propose that different languages can exploit the perceptual biases behind the effect at different degrees. As discussed in Section 1.4, the ratio between the duration of the closure and that of the vowel has been identified as one of the perceptual cues to voicing (Port & Dalby 1982; Lisker 1986). Listeners associate smaller values of the CV ratio to voiced stops, and, vice versa, greater values to voiceless stops. Kluender et al. (1988) argue that speakers can actively manipulate vowel durations to proportionally increase the difference in ratio between the two voicing categories, so that the ratio would be even smaller in the voiced context and even greater in the voiceless one. As a consequence, the perceptual distance between the voicing categories would be

enhanced, thus facilitating discrimination (Stevens & Keyser 1989; Kingston & Diehl 1994). According to this view, listeners' discrimination of vowel duration should show a contrast effect, by which longer closure durations elicit more "short vowel" responses and shorter closures more "long vowel" responses. However, Fowler (1992) shows that listeners judge vowels to be longer when the stop closure duration is increased, and that, similarly, stop closure is perceived to be longer when vowel duration is increased. These results indicate a mechanism of perceptual assimilation of the respective durations of vowels and stop closures and do not support a contrast effect.

While perceptual biases could be driving some aspects of the voicing effect and be responsible for its enhancement in some languages, production mechanisms are likely to provide the necessary variation that would be exploited by the perceptual system (Beguš 2017; Sanker 2019). Although individual production accounts differ in the details, two broad categories can be identified. Some accounts ascribe the source of the voicing effect to mechanisms of compensation within a certain property of speech (either duration or articulatory force), while others relate the emergence of the effect to timing aspects of articulatory gestures (either laryngeal or oral).

The *compensatory temporal adjustment* account (Lindblom 1967; Slis & Cohen 1969a,b; Lehiste 1970a,b) states that the relative durations of vowel and consonant in a VC sequence are correlated. A well-known fact about stop closure is that it is longer in voiceless stops and shorter in voiced stops (Lisker 1957; Summers 1987; Davis & Summers 1989; de Jong 1991). As a consequence, vowels are shorter when followed by the longer closure of voiceless stops, and they are longer when followed by the shorter closure of voiced stops. This compensatory pattern would be the consequence of keeping the duration of a particular speech interval fixed, while the duration of the closure changes depending on the voicing status of the stop. Proponents of this account have argued that compensation is implemented either at the level of the syllable (or of the VC sequence, Lindblom 1967; Farnetani & Kori 1986), or at the level of the word (Slis & Cohen 1969a,b; Lehiste 1970a,b). This formulation of the account, however, faces empirical and logical challenges. The duration of both the syllable and the word is affected by stop voicing (Chen 1970; Jacewicz et al. 2009), and it is not clear why compensation within a word should necessarily target the pre-consonantal vowel and

not other segments (these issues are discussed in more details in Chapter 4).

A second proposal attributes the voicing-driven duration differences of vowels to *articulatory energy expenditure*, rather than temporal aspects. Meyer (1903) and similarly Belasco (1953) propose that the articulatory force required to produce a syllable is constant, and thus it is distributed across segments according to their energy requirements. According to this hypothesis, voiceless stops are produced with more force than voiced stops, and hence some force is subtracted from the production of the preceding vowel to maintain the overall force constant. However, the concept of "articulatory force" lacks an empirically solid definition, and experimental results mentioned in Zimmerman & Sapon (1958) rather point to the absence of a relation between energy expenditure and vowel duration.

While the compensatory temporal adjustment and the energy expenditure accounts rely on compensatory mechanisms of duration or articulatory force, two other proposals concern aspects of gestural timing of the larynx and the consonant closing gesture. The *laryngeal adjustment* account (Halle & Stevens 1967; Halle et al. 1967; Chomsky & Halle 1968) is based on the idea that voicing during stop closure requires precise adjustments of the glottis in order to comply with aerodynamic constraints (Ohala 2011). Such an articulatory precision necessitates greater time to be implemented than the production of closure voicelessness. Because of these properties of laryngeal articulation, full closure can be achieved relatively faster in the context of voiceless stops (which require less precise control), while a delay of closure onset in voiced stops ensures enough time to produce the suitable glottal configuration. The preliminary electromyographic study of glottal muscular activity discussed in Chen (1970), however, does not suggest the presence of early laryngeal activity during the production of vowels followed by voiced stops compared to vowels followed by voiceless stops. Further articulatory evidence shows that there is rather a general increase of activity of certain laryngeal muscles (namely, the posterior crycoarytenoid and the crycothyroid) during the production of voiceless sounds (Hirose & Gay 1972; Kagaya & Hirose 1975; Hirose 1977; Löfqvist et al. 1989). No conclusive evidence can thus be adduced in support of the laryngeal adjustment hypothesis, although other laryngeal mechanisms cannot be ultimately excluded (Beguš 2017).

Another production account is based on the *rate of stop closure transition* (Öhman 1967b; Chen 1970). Voiceless stops are articulated with greater glottal opening relative to voiced stops (and vowels), so that a greater volume of air is admitted into the oral cavity. Öhman (1967b) argues that the production of the closure of voiceless stops would then require more muscular effort to counteract the increased intra-oral pressure generated by the greater airflow. As a consequence, the rate of the closing gesture of voiceless stops is higher than that of voiced stops. In other words, full closure will be achieved earlier relative to the onset of the closing gesture when the stop is voiceless than when it is voiced. Hence, vowels will be shorter when followed by voiceless stops than when followed by voiced stops. Chen (1970) observes that the difference in labial closure rate accounted for 20% of the difference in vowel duration. Subsequent work by Warren & Jacks (2005) further shows that the percentage of the difference which is accounted for rises to 80% when considering the movements of both the lips and the jaw.

In sum, four main production accounts (or variations thereof) can be found in the literature on the voicing effect. More specifically, two of these accounts posit a mechanism of compensation either between segmental durations (the compensatory temporal adjustment account) or articulatory force (the articulatory energy expenditure account), while two relate durational differences to the timing of laryngeal gestures (the laryngeal adjustment account) or oral gestures (the rate of stop closure transition account). In the following section I review the effects of two other phonation types (aspiration and ejection) on vowel duration, and how these shed light on the aforementioned accounts of the voicing effect.

1.6 Beyond voicing

Two phonation modes other than voicing are known to affect preceding vowel duration: aspiration and ejection. While this project focusses on the voicing effect, the closely related aspiration and ejection effects have consequences of theoretical importance. The results concerning the aspiration effect are mixed. Maddieson & Gandour (1976), Durvasula & Luo (2012), and Lampp & Reklis (2004) report longer vowels before aspirated

than before unaspirated stops in Hindi, and Maddieson (1976) finds a similar trend in Assamese, Bengali, and Marathi. Ohala & Ohala (1992), on the other hand, show that vowels have the same duration before unaspirated and aspirated stops in their sample of Hindi speakers. Sanker (2018) observes an effect of aspiration in Hindi long vowels but not in short vowels, while the effect is reversed in Telugu long vowels (vowels are shorter before aspirated than unaspirated stops), with no appreciable difference in short vowels. Note that these studies don't easily lend themselves to comparison, since the material and contexts used differ (for instance, vowel type, vowel phonological length, number of syllables, and context following the test word).

The trend of vowels being longer when followed by aspirated stops challenges some of the accounts presented in Section 1.5, as noted in Maddieson & Gandour (1976). The articulatory force expenditure hypothesis predicts vowels to be shorter before aspirated than before unaspirated stops since it is likely that aspirated stops require greater force than unaspirated ones. According to the laryngeal adjustment account, the duration of the vowels should not differ in voiceless unaspirated and aspirated stops since they both require glottal opening, rather than the precise adjustments characteristic of voiced stops. Since closure rate is determined by airflow, the rate of closure account expects vowels followed by aspirated stops to be shorter than or equal to vowels followed by unaspirated stops, since the former should be characterised by greater airflow and higher closure rates due to glottal spreading. While Maddieson & Gandour (1976) argue against a compensatory effect between vowel and consonant duration, the data in Durvasula & Luo (2012) are instead compatible with it (see Chapter 4). The results in Sanker (2018) on Hindi, but not Telugu, are also compatible with a compensatory mechanism. Closure duration is longer in unaspirated stops and shorter in aspirated stops in both languages, but the effect of aspiration on vowel duration has opposite directions depending on language, as discussed above.

An investigation of the effect of ejection in Georgian (Beguš 2017) shows that vowels are shortest when followed by aspirated stops, longer when followed by ejectives, and longest when followed by voiced stops. Georgian contrasts aspirated voiceless, ejective, and voiced unaspirated stops. The negative correlation between closure and vowel duration has greater magnitude in the context of voiced compared to that of as-

pirated and ejective stops. Moreover, the author shows that the closure effect on vowel duration coexists with a “Laryngeal Features” effect (both closure duration and phonation, when entered in a single regression model, lead to significant *p*-values). In other words, the variance in vowel duration is accounted for in part by the duration of the stop closure and in part by the voicing category of the post-vocalic stop. As discussed in Beguš (2017), these patterns are compatible with accounts of compensatory temporal adjustments, laryngeal adjustments, and rate of closure.

In conclusion, our partial understanding of the relation between vowel duration and consonant phonation is based on contrasting or complementary empirical evidence. This state of affairs can be taken as indication that, while most research (save a few recent exceptions) focussed on finding a unique and unified mechanism behind the voicing effect, we might rather seek multiple mechanisms that cooperate to produce the observed patterns. In light of this, this dissertation sets out to study the interrelations between different sources of evidence, and their interpretation. Chapter 2 presents a more detailed discussion of the rational behind the research of this dissertation, and a summary of the main results.

Chapter 2

Rationale ~~of~~ the current research

This research project has three broad aims. First, I sought to obtain acoustic and articulatory data using modern methods which can shed light on the production accounts of the voicing effect put forward in the past century. The second main objective was to enlighten the debate on reported cross-linguistic differences by conducting an analysis which encompasses three related but contrasting languages. The papers in Part II offer evidence in relation to these goals. The third aim was to carefully (re)consider previous and current results in light of the methodological debates related to the Open Science movement (Section 3.3). This three aims correspond to the following questions:

1. What is the diachronic articulatory source of the voicing effect, and what can synchronic acoustic and articulatory data tell us about the possible pathway to the emergence of the voicing effect?
2. How can the comparison of three contrasting languages enlighten the debate of the source of the voicing effect?
3. How can we effectively apply Open Science practices in phonetic research, and what level of confidence can we assign to the results?

This chapter is an overview of how these three questions have been addressed.

Another fundamental aspect of this research project is that it was developed in two stages: an exploratory (hypothesis-generating) stage, and a confirmatory (hypothesis-testing) stage (on the exploratory/confirmatory dichotomy, see Tukey 1980 and Section 3.3.3). These stages correspond to Study I and Study II respectively, an overview of

which is given in Section 3.1 and Section 3.2. The research questions at the exploratory stage (Study I) were formulated while being agnostic in regards to specific hypotheses. Rather, the literature reviewed in Chapter 1 formed the basis for a set of general questions about articulatory properties of VC sequences. These questions justified the experimental design of Study I (Section 3.1). More specific questions emerged while performing exploratory data analyses at this stage. New hypotheses were generated by the exploratory phase, which justified a confirmatory study (Study II, Section 3.2). Finally, questions pertaining to a comparison across languages and replication of previous results spanned across the two stages, and their discussion is brought up at different points across the dissertation.

Section 2.1 discusses in detail the research questions and a justification of methods. Section 2.2 is an overview of the chosen languages and of relevant aspects of their phonological systems. Finally, Section 2.3 is a preview of the results.

2.1 Research questions

The first research question concerns the source of the voicing effect, or, in other words, the diachronic pathway that led or can lead to the emergence of the voicing effect in any particular language. More specifically, the question asks how a language can develop the voicing effect and which speech aspects play a role in such development. The long-standing debate about the source of the voicing effect in light of the different proposals discussed in Section 1.5, whether articulatory or perceptual, is evidence for the difficulty of selecting a single property of speech that is behind the differential duration of vowels followed by voiceless vs voiced stops. Moreover, the existence of durational phenomena related to phonation types other than voicing, like aspiration and ejection (Section 1.6), call for an approach to the understanding of the voicing effect that is independent from voicing *per se*, while still limiting the investigation to the voicing contrast. Such an approach allows us to formulate an account that future research can generalise and apply to other durational phenomena (related to phonation or not). Furthermore, note that the focus of the current research is on how the voicing effect emerges in the first place, and not how individual languages exploit or not the effect to

enhance cues of phonological contrast.

A window into possible diachronic developments is offered by the investigation of cross-linguistic synchronic data, an approach taken here. This approach is justified by the idea that diachronic change draws upon synchronic variation and that synchronic variation is the outcome of diachronic change (Blevins 2004, 2006; Cristofaro 2012, 2014; Bermúdez-Otero 2015). The view of synchrony/diachrony entanglement enables the use of synchronic information to infer possible diachronic changes that might have led to the current synchronic state.

In light of the complexity of the durational effects reviewed in Chapter 1, I further decided to limit the scope of the investigation to aspects of production, while keeping an open mind about perceptual factors, as discussed in Chapter 8. This choice was based in part on the relative paucity of recent articulatory data of the voicing effect in relation to, for example, acoustics and perception, and in part on the greater number of production accounts of the voicing effect relative to that of perception accounts. Furthermore, the production accounts reviewed in Section 1.5 deal either with oral (tongue) or laryngeal articulations. In order to identify potential properties of these two types of gestures it seemed a natural choice to use ultrasound tongue imaging and electroglottography in combination with acoustics as three sources of data. In particular, the research sought to obtain data on segment durations, timing of the consonantal gestures, and properties of vocal fold vibration, given the focus on these features in the literature reviewed in Chapter 1. Since the voicing effect (and related durational phenomena) has been prevalently if not exclusively defined and dealt with in terms of acoustic segmental durations, the same approach is used here, and acoustic durations will be at the core of the analyses presented in Part II. Since previous work on stop consonants has generated more coherent results than work on other manner of articulations, and since most hypotheses rest on aerodynamic properties of full stop closures, the focus of this dissertation will be limited to stop consonants.

A convenient way to investigate mechanic properties underlying the effect of voicing on vowel duration is to consider languages in which the effect has not been claimed to be phonologised (Section 1.4). Moreover, comparing two languages that differ in the presence or degree of vowel durational differences can uncover variation motivat-

ing cross-linguistic differences. Italian and Polish are two good candidates in that they satisfy both of these requirements. Moreover, their phonological systems allow for a somewhat direct comparison. For these reasons, an exploratory study of Italian and Polish (Study I) was carried out to examine the influence of voiceless and voiced stops on vowel duration. Section 3.1 contains a description of the methods employed in Study I, while Chapter 4, Chapter 6, and Chapter 7 report the results. Note that, with the terms "voiceless" and "voiced", I refer to the linguistic reading of "voicing", rather than to the physical implementation of such contrast, as detailed in Section 1.2. This approach is generally helpful in light of the distinction between aspirating vs true-voicing languages (Beckman et al. 2013), and in the case of English in particular (Docherty 1992), which is the subject of Study II. Furthermore, I focus here on voicing as a categorical lexical contrast, given this is the approach followed by most of the relevant literature. Future work is warranted to ascertain the role of a gradient/continuous operationalisation of voicing.

As a follow up of Study I, Study II set out to investigate in English the patterns observed in Study I in Italian and Polish. English was chosen as a further test language given the abundance of previous work dealing with different aspects of the English voicing effect. Moreover, virtually all the accounts reviewed in Section 1.5 were originally posited based on English data. A second reason behind this choice is that English allows us to look into differences between word-medial and word-final contexts.[1] This is warranted based on the reported difference in magnitude of the voicing effect in word-medial and word-final position, as mentioned in Section 1.3. An overview of the methods of Study II is given in Section 3.2, while Chapter 5 presents and discusses the study and its results.

The second question this dissertation set out to answer is concerned with a cross-linguistic comparison of the voicing effect. Building on the results discussed in the chapters of Part II, Section 8.1 offers a synthesis of the main topics touched upon in Part II. In turn, this forms the basis of the cross-linguistic comparison of Italian, Polish, and English in Section 8.2.

[1]Note that Polish would not be a good candidate because of word-final neutralisation of voicing (Gussmann 2007).

Lastly, the third objective is related to research practices and the Open Science movement. In light of the concepts and issues which will be reviewed in Section 3.3, the research described in this dissertation has been carried out according to principles of openness of data, transparency of analysis, and reproducibility and replicability of results. Section 3.3.4 in particular discusses how these principles were applied.

The following section gives a description of the main phonological features of Italian, Polish, and English, paving the way to the preview of the results in Section 2.3 and the discussion of the methods in Chapter 3.

2.2 Language sample

This section gives an overview of the phonological systems of Italian, Polish, and English, which will set the stage for the preview of the results in the following section and the discussion of these in the second part of the dissertation. Note that when referring to languages, the languoid model is implicitly assumed (Cysouw & Good 2013). A languoid is the pairing of a glossonym (a name that refers to a languoid or doculect) with a collection of doculects. In turn, a doculect is the pairing of a glossonym with a specific publication (in any form, for example a book with the grammatical description of the doculect, or an article focussing on a specific linguistic aspect). Languoids can be hierarchical, so that a languoid can be composed of other languoids, and so on. The doculects of this dissertation are referred to by the glossonyms *Italian*, *Polish*, and *Manchester English*. The Italian doculect is included in the languoid Italian [glottocode: ital1282], the Polish doculect in the languoid Polish [glottocode: poli1260], and the Manchester English doculect (*English* for short from now on) in Western Central English [glottocode: west2900].[2]

Vowel and consonant categories as used here should be interpreted as descriptive categories when language-specific phonemes are discussed, and as comparative concepts when cross-linguistic comparisons are carried out, as discussed for the category

[2]Languoid classification is controversial, as much as traditional language classification, so that classification decisions are taken here without fully committing to them. The classification adopted here does not directly bear on the research results. Future work is warranted for a more thorough classification.

Table 2.1: Italian consonant phonemes(adapted from Kramer 2009).

FL	labial	dental	alveolar	palatal	velar
stop	p, b	t, d	ts, dz	tʃ, dʒ	k, g
fricative	f, v		s, z	ʃ, (ʒ)	
nasal	m		n	ɲ	
lateral			l	ʎ	
rhotic			r		
approximant	w			j	

Table 2.2: Italian vocalic phonemes (adapted from Kramer 2009).

	front	central	back
high	i		u
mid-high	e		o
mid-low	ɛ		ɔ
low		a	

of voicing in Section 1.2. This approach follows from the view that phonemes make sense only within the linguistic system they are from (Trubetzkoy 1969; Haspelmath 2010). In this sense, they are descriptive categories. So the phoneme /a/ of Italian is different from the phoneme /a/ of Polish, even in the case they are phonetically similar, for the fact that they belong to two different linguistic systems. When effects like that of voicing are compared across languages, a category like /a/ is no longer to be intended as a descriptive category, but rather as a comparative concept.

The following sections introduce, for each language, the vowel and consonantal phonemic systems, with special attention to phonation contrasts in consonants, syllabic structure and stress patterns, and rhythmic class (Pike 1945).

2.2.1 Italian

Although the exact phonemic inventory of Italian is still debated, especially for consonants (Krämer 2009:44), a generally agreed upon phonemic set is given in Table 2.1 for consonants and Table 2.2 for vowels.

Italian contrasts consonants along five (phonological) places of articulation: labial (phonetically either bilabial or labiodental), dental, alveolar, palatal (palatal and post-alveolar), and velar. Stops (true stops and affricates) and fricatives contrast for voicing, although note that /z/ has limited functional load (Bertinetto & Loporcaro 2005) and /ʒ/ is relegated to loan words. The Italian voicing contrast is usually described in terms of an opposition between voiceless unaspirated consonants and fully voiced consonants (Vagges et al. 1978; Bortolini et al. 1995; Pape & Jesus 2014; Kirby 2016). Pape & Jesus (2014) shows that Italian speakers tend to perceive stops without a burst following the release as voiced consonants, independent of the duration of voicing during closure. While it is not clear which acoustic cue is employed by Italian speakers to discriminate voiceless and voiced consonants, Pape & Jesus (2014) find in their production study that Italian consistently articulate (velar) stops with full voicing during closure.

The vocalic system in Table 2.2 is found in stressed syllables, although the status of the mid-high and mid-low contrast is not straightforward (especially for the back vowels), and the mid vowels show a high degree of geographical and idiosyncratic variation (Renwick & Ladd 2016). In unstressed syllables, there is no contrast between mid-high and mid-low vowels, and these vowels are articulated as either mid-high or mid-low depending on the variety of Italian (Rogers 2004; Renwick & Ladd 2016). Although vowel duration is not contrastive (Rogers 2004; Krämer 2009; Renwick & Ladd 2016), vowels are longer when they appear in a stressed open syllable (/fa.to/ [faːto] ‘fate’) and shorter when the syllable is closed (/fat.to/ [fatto] ‘fact’).

Stress in Italian is contrastive (non-predicable), and main lexical stress is generally placed on one of the last three syllables (d’Imperio & Rosenthall 1999; Krämer 2009). The basic foot is a maximally bimoraic trochee (Krämer 2009). Italian is traditionally ascribed to the syllable-timed class of rhythmic typology (Pike 1945). However, properties of stress-timed languages (like vowel reduction) can also be observed in Italian, depending on the regional variety (White et al. 2009; Giordano & D’Anna 2010;

Table 2.3: Polish consonant phonemes (adapted from Jassem 2003).

	labial	dental	alveolar	alveopalatal	palatal	velar
plosive	p, b	t, d			c, ɟ	k, g
fricative	f, v	s, z	ʃ, ʒ	ɕ, ʑ		x
affricate		ts, dz	tʃ, dʒ	tɕ, dʑ		
nasal	m	n				ŋ
lateral		l				
rhotic			r			
approximant	w					

Table 2.4: Polish vocalic phonemes (adapted from Jassem 2003).

	front	central	back
high	i		u
mid-high		ɨ	
mid-low	ɛ ɛ̃		ɔ ɔ̃
low		a	

Pamies Bertrán 1999).

2.2.2 Polish

Polish consonants contrast six places of articulation (Jassem 2003): labial (bilabial and labiodental), dental, (post-)alveolar, alveopalatal, palatal, and velar. Similarly to Italian, Polish stops, fricatives, and affricates can either be voiceless or voiced. Keating (1984b) argues that the Polish voicing contrast is between fully voiced consonants and voiceless (short-lag VOT) consonants. Waniek-Klimczak (2011), on the other hand, suggest a possible change in progress by which the duration of VOT in Polish voiceless stops before stressed vowels is increasing. Moslin & Keating (1977) also suggest that the VOT values tend to be longer under certain prosodic conditions. In relation to this

Table 2.5: English consonant phonemes.

	labial	dental	alveolar	post-alveolar	palatal	velar	glottal
plosive	p, b		t, d			k, g	
fricative	f, v	θ, ð	s, z	ʃ, ʒ			h
affricate				tʃ, dʒ			
nasal	m		n			ŋ	
lateral			l				
rhotic				r			
approximant	w				j		

finding, Schwartz & Arndt (2018) report that the perception of the voicing contrast by Polish speakers is not hindered by the absence of pre-voicing in voiced stops. Finally, the voicing contrast is neutralised in absolute word-final position (Gussmann 2007), but it is maintained syllable-finally word-medially (Strycharczuk 2012). The Polish vocalic system is made of eight vowel phonemes, six oral and two nasalised: /i, ɛ, ɨ, a, ɔ, u/, /ɛ̃, ɔ̃/ (Jassem 2003; Gussmann 2007).

Polish lexical stress is fixed on the penultimate syllable, with exceptions having ante-penultimate stress being loan words (Gussmann 2007). The phonological nature of Polish lexical stress is still debated (see review in Łukaszewicz 2018). As for the class of rhythmic typology, Polish exhibits features from both stressed-timed and syllable-timed languages (Dauer 1987; Nespor 1990; Grabe & Low 2002; Arvaniti 2009).

2.2.3 English

In order to avoid influences of regional differences in English, especially in the vowel system, Study II (Section 3.2) was restricted to Manchester English.[3]

The consonant system of Manchester English minimally diverges from the general Southern British English system (Table 2.5), which is non-rhotic, with the notable exceptions of the so-called “T-glottaling” (realisation of /t/ in non-foot-initial position

[3]Due to the difficulty of recruiting speakers of Italian and Polish in Manchester and in the field in Italy, such approach was not possible for these languages.

as [ʔ]), “TH-fronting” (realisation of /θ, ð/ and [f, v]), “H-dropping”, and “velar nasal plus” (realisation of /ŋ/ as [ŋg/]) (Baranowski & Turton 2015; Baranowski et al. 2016; Bermúdez-Otero et al. 2016; Coretta & Canzi 2018; Bailey 2019a,b). The consonantal phonemes of Manchester English belong to one of seven places of articulation (labial, dental, alveolar, post-alveolar, palatal, velar, glottal) and seven manner of articulation (plosive, fricative, affricate, nasal, lateral, rhotic, approximant).

While voicing in Manchester English has not been systematically investigated, the literature on voicing in English in general is vast (for a review, see Davidson 2016). English obstruents (plosives, fricatives, affricates) contrast for what has been traditionally described as voicing, which is also reflected in the standard use of IPA voiceless and voiced symbols. However, the actual articulatory implementation of the contrast is constituted by a complex set of features and it is affected by other phonological factors, like syllabic structure and stress (Lisker 1986; Docherty 1992). Generally speaking, while the voicing contrast in word-medial position especially after stressed vowels is between a category with voicing during closure (voiced category) and one without it (voiceless category), in pre-stressed position and especially in word-initial position the contrast is between two voiceless categories that differ in voice onset time (short VOT vs long VOT, with no vibration of the vocal folds during closure in the former).

Another relevant dimension is the type of phonation used by speakers to encode the voicing contrast in English. For example, Gordeeva & Scobbie (2007, 2010, 2011) show that preaspiration, glottalisation, and ejection can be used by speakers as cues to the voicing contrast in fricatives and stops in Scottish English. Moreover, no evidence was found for a correlation between the type of phonation employed by each speaker and their general voice quality (Gordeeva & Scobbie 2011). The authors interpret this finding to mean that the speaker’s voice quality settings and the use of one phonation type over another are decoupled, and that preaspiration, glottalisation, and ejection play an important role in the speaker-specific phonologisation of the contrast. The sociolinguistic aspects of voicing investigated in these studies stress the multidimensional nature of the English voicing contrast.

Manchester English distinguishes short and long vowels (Table 2.6), which differ in duration and quality. The split between /ʊ/ and /ʌ/ (respectively FOOT and STRUT in

Table 2.6: Northern British English vowel monophthong phonemes (Orton 1962, Wells 1892).

	front	central	back
high	ɪ iː		ʊ uː
mid	ɛ	ə ɜː	ɔː
low	æ		ɒ ɑː

Well's lexical set, Wells 1982) present in many varieties of English is not in Manchester English (as in Northern English more generally), so that there is a single vowel category realised as [ʊ] (Baranowski & Turton 2015). Other features of the vocalic system in Manchester English are the fronting of /uː/, and the laxing of the *happy* vowel (the final vowel in words like *happy, city, duty*) to [ɛ] in word-final position.

Lexical stress is contrastive in English (Giegerich 1992). English is more or less uncontroversially regarded as a stress-timing language (Classe 1939; Pike 1945; Abercrombie 1967; Grabe & Low 2002).

2.3 Preview of results

This section presents an overview of the results derived from the investigation of acoustic durations and articulatory properties of vowel-consonant sequences of three related but contrasting languages (Italian, Polish, English) in Study I and II. The results suggest a composite production account of the voicing effect which synthesises previous independent and seemingly contrasting proposals. In particular, the proposed account revisits and combines elements from the compensatory temporal adjustment account, the laryngeal adjustment account, and rate of closure account (Section 1.5). The following paragraphs summarise the contribution of each original publication in Part II, while a full-fledged discussion of the holistic proposal will be given in Chapter 8.

Chapter 4 and Chapter 5 provide evidence for a revised compensatory adjustment

account of the voicing effect. Chapter 4 deals with Italian and Polish acoustic data, and it shows that vowel-consonant sequences are embedded within a speech interval that is temporally stable across voicing contexts. This paper discusses mechanisms of compensation between vowel and consonant closure duration within such interval. Chapter 5 extends these findings to English, by comparing durational properties of monosyllabic and disyllabic words. More specifically, I discuss how differences in the gestural organisation of mono- vs disyllabic words illuminates the debate on diachronic pathways and perceptual biases behind the voicing effect in these two phonological contexts. In Appendix B, I relate the current results with those from previous work, by means of a meta-analytical study of the English voicing effect.

Based on data from disyllabic words of Italian, Polish (Chapter 4), and English (Chapter 5), it is demonstrated that the duration of the speech interval between the releases of two stops flanking a stressed vowel is not affected by the voicing status of the post-vocalic consonant. By capitalising on known articulatory properties of vocalic and consonantal sequences (Öhman 1967b; Fowler 1983; O'Dell & Nieminen 2008; Saltzman et al. 2008), the temporal stability of the release-to-release interval is proposed to be a consequence of the isochrony of the vocalic gestures of the word and of the phasing of the consonantal gestures relative to vowels. While experimental testing of vowel-to-vowel isochrony and vowel-consonant phasing is warranted, Appendix D provides initial partial evidence. As a side effect of the release-to-release temporal stability, the timing of the VC boundary within such interval determines the respective durations of the vowel and the following consonant closure, the latter of which is known to be longer for voiceless than for voiced stops (Lisker 1957; Summers 1987; Davis & Summers 1989; de Jong 1991). As a consequence, shorter vowels are followed by the longer closures of voiceless stops, while longer vowels are followed by the shorter closure of voiced stops.

The results of English monosyllabic words, on the other hand, show that in this context the release-to-release interval is longer when the post-vocalic consonant is voiced (Chapter 5). The absence of release-to-release temporal stability in monosyllabic words is argued in Chapter 5 to be related to the absence of vowel-to-vowel isochrony, which in turn is a consequence of the lack of a second vowel functioning as a temporal anchor.

The respective durations of vowel and closure can thus be modified independently, fact that speakers can exploit to enhance the voicing contrast. Contrast enhancement can be obtained by manipulating the ratio between the duration of the vowel and that of the closure without the constraint of keeping the release-to-release duration stable, as in disyllabic words. The presence of the voicing effect in monosyllabic words is conjectured to have emerged as a consequence of mechanisms affecting the timing of the consonant closure onset, in accordance with the rate of closure and laryngeal adjustment accounts of the voicing effect. Chapter 6 and Chapter 7 offer insights about these accounts in relation to tongue root advancement and glottal spreading respectively.

In Chapter 6, the time of the boundary between a vowel and the following consonant (i.e. the stop closure onset) is shown to be modulated, among other known factors, by the position of the tongue root, as evidenced by tongue imaging data. In particular, I explore the link between vowel duration, closure duration and tongue root advancement, and discuss how the timing of consonant closure affects all three aspects. Tongue root advancement was observed during the closure of voiced stops in some but not all speakers of both Italian and Polish. Moreover, it was found that tongue root advancement is initiated during the production of the vowel preceding the target consonant and that the degree of advancement at stop closure onset is positively correlated with preceding vowel duration, such that longer vowels correspond to greater tongue root advancement. Together with the shorter duration of the closure of voiced stops, this pattern fits with the known role of tongue root advancement in the maintenance of voicing during stop closure (Kent & Moll 1969; Perkell 1969; Westbury 1983).

In Chapter 7, the analysis of vocal fold activity during the production of vowels shows that the latter portion of vowels followed by voiceless stops is produced with greater glottal spread in Italian than in Polish. This difference is taken as evidence for a language-specific implementation of the timing of glottal spreading. Increased glottal spread before voiceless stops is understood as the precursor of pre-aspiration, the presence of which has been reported in Italian (Ní Chasaide & Gobl 1993; Stevens & Hajek 2004a,b, 2010; Stevens 2010; Stevens & Reubold 2014). By combining previous work on pre-aspiration (Lisker 1974; Ní Chasaide 1985; Stevens et al. 2014), two alternative pathways of sound change development are proposed: either pre-aspiration is

enhanced by shortening the closure of the stop, or it is reduced or prevented altogether by producing an earlier stop closure. The latter solution would mask the acoustic effects of glottal spreading and result in a longer closure duration and shorter vowel duration, other things being equal.

Section 8.1 offers an answer to the first question set out in Section 2.1 by combining (1) a word-holistic articulatory account of gestural phasing, of which the release-to-release temporal stability is a consequence, and the modulating properties of (2) tongue root advancement and (3) glottal spreading on the timing of the vowel offset/closure onset in vowel-consonant sequences. It is proposed that these three interacting aspects play a role in driving the development of the voicing effect. As for the question of what cross-linguistic differences can be observed in relation to the voicing effect, the data from Italian, Polish, and English suggest that, when different phonological aspects are controlled for, the magnitude of the effect is similar across languages (Section 8.2).

The conceptual contribution of this dissertation, as summarised in the previous paragraphs, is accompanied by an advancement of methodologies in phonetic data analysis and research more generally. Chapter 7 and Appendix A introduce two methods for the analysis of electroglottographic data and tongue contours using generalised additive modelling. The application of generalised additive models on electroglottographic data allows us to obtain a dynamic and multidimensional view of vibratory properties of the vocal folds (Chapter 7). This constitutes an improvement from methods that reduce the multidimensionality of fold vibration to a single measure, like the closed quotient. Appendix A shows how generalised additive modelling can be used with tongue contour data in polar coordinates to control for a complex combination of effects. Modelling is exemplified by means of a comparison of tongue contours obtained from the time of maximum constriction of voiceless and voiced stops in Italian and Polish, which corroborates the between-speaker differences observed in Chapter 6.

This dissertation is also an example of how state-of-the-art research methods can be applied to linguistic research, as part of the third research aim outlined in Section 2.1. The methods adopted in this dissertation were influenced by the Open Science movement. All research materials (data, code, documentation) are made available on the Open Science Framework (Coretta 2020). In the interpretation of the results, more em-

phasis was given to the estimation of parameters in statistical models, and to the degree of uncertainty surrounding them. To facilitate this endeavour, Bayesian statistics was applied to address a subset of the research questions. Finally, custom research-management and analysis tools were developed in the form of R packages.

The following chapter includes an overview of the research methods of Study I (Section 3.1) and Study II (Section 3.2). Section 3.3 introduces the principles of Open Science and discusses how they shaped the current project.

Chapter 3

Methods

Each of the proposals regarding the origin of the voicing effect, reviewed in Chapter 1, stresses one particular aspect of the mechanisms that could lie behind this phenomenon. Whereas some of the hypotheses concern biases of the perceptual system, others depend on articulatory and aerodynamic properties of speech production. Crucially, all accounts have found only partial support in the literature. Over the years, evidence has accumulated for the articulatory accounts based on compensatory temporal adjustments, laryngeal adjustments, and rate of consonant closure. Given the complex nature that characterises the production accounts, this thesis sets out to investigate durational and dynamic aspects of the articulation of vowel-consonant sequences. The outcomes of a time-synchronised analysis of acoustic and articulatory data from Italian and Polish (Section 3.1) indicate that components of temporal compensation and gestural phasing are the likely source of the differences in vowel and closure durations. A follow-up acoustic study of English (Section 3.2) further corroborates these results, and offers new insights on the possible development of the voicing effect from the word-level structuring of vocalic and consonantal gestures.

An overview of these studies, with information on experimental materials, procedures, and data processing, is given in this chapter. The following sections constitute a synopsis of the methodologies presented in more specific details in the relevant papers. The data and code referred to here can be found in the dissertation repositories on the Open Science Framework and GitHub (see Section 3.3.4). Ethics clearance to undertake this work was obtained from the University Research Ethics Committee (UREC)

of the University of Manchester (REF 2016-009976).

3.1 Exploratory study of the voicing effect in Italian and Polish (Study I)

As discussed in Section 2.1, since the accounts this dissertation focusses on cover aspects of segmental duration, consonantal articulatory gestures, and voicing, data was obtained from three sources: (1) acoustics, (2) ultrasound tongue imaging, a non-invasive technique to image the tongue using ultrasonic equipment, and (3) electroglottography, an indirect and safe method to gather information on vocal fold activity. In the following sections, I offer an outline of the methodologies employed in Study I, while referring the reader to specific papers for a more in-depth description.

3.1.1 Participants

A total of 17 participants were recruited in Manchester (UK) and in Verbania (Italy). Eleven were native speakers of Italian (IT01-IT05, IT07, IT09, IT11-IT14), and six of Polish (PL02-PL07). Missing speaker codes refer to test participants or participants that produced unusable data because of individual anatomy or recording issues. Recordings were made in a sound-attenuated room at the Phonetics Laboratory of the University of Manchester, or in a quiet room at a field location in Verbania (IT03-IT07). Due to technical issues or poor signal quality, ultrasonic data of /u/ from two speakers (IT07, PL05) and electroglottographic data from two others (IT04, IT05) are missing. Participants' sociolinguistic data is given in Chapter 4. The participants were given an information sheet prior to the experiment and signed a consent form.

3.1.2 Ultrasound tongue imaging and electroglottography

2D Ultrasound tongue imaging (UTI) uses ultrasonography for charting the movements of the tongue into a two-dimensional image (Gick 2002; Stone 2005; Lulich et al. 2018). In medical sonography, ultrasonic waves (sound waves at high frequencies, ranging between 2 and 14 MHz) are emitted from piezoelectric components in a transducer. The

surface of the transducer is placed in contact with the subject's skin, and the waves irradiate from the transducer in a fan-like manner, travelling through the subject's soft tissue. When the surface of a material with different density is hit by the ultrasonic waves, some of the waves are partially reflected, and such 'echo' is registered by the probe. The information interpolated from these echoes can be plotted on a two-dimensional graph, where different material densities are represented by different shades (higher densities are brighter, while lower densities are darker). The graph, or ultrasound image, shows high density surfaces as very bright lines, surrounded by darker areas. By positioning the ultrasound probe in contact with the sub-mental triangle (the surface below the chin), sagittally oriented, it is possible to infer the cross-sectional profile of the tongue, which appears as a bright line in the resulting ultrasound image.

Electroglottography (Fabre 1957; Childers & Krishnamurthy 1985; Scherer & Titze 1987; Rothenberg & Mahshie 1988) is a technique that measures the size of contact between the vocal folds (the Vocal Folds Contact Area, VFCA). A high frequency low voltage electrical current is sent through two electrodes which are in contact with the surface of the neck, one on each side of the thyroid cartilage. The impedance of the current is directly correlated with VFCA, while its amplitude is inversely correlated with it (Titze 1990). Impedance increases with lower VFCA and decreases with higher VFCA. Conversely, amplitude decreases when the VFCA increases and it increases when the VFCA decreases. The EGG unit registers changes in impedance and converts it into amplitude values. The unit outputs a synchronised stereo recording which contains the EGG signal from the electrodes in one channel and the audio signal from the microphone in the other.

3.1.3 Equipment set-up

Figure 3.1 shows a schematics of the equipment set-up used in this study. The left part of the figure shows the ultrasonic components (Articulate Instruments Ltd™ 2011), while the EGG components (Glottal Enterprises) are shown on the right. Two separate Hewlett-Packard ProBook 6750b laptops with Microsoft Windows 7 were used for the acquisition of the UTI and EGG recordings. The main ultrasonic component was a TELEMED Echo Blaster 128 unit with a TELEMED C3.5/20/128Z-3 ultrasonic trans-

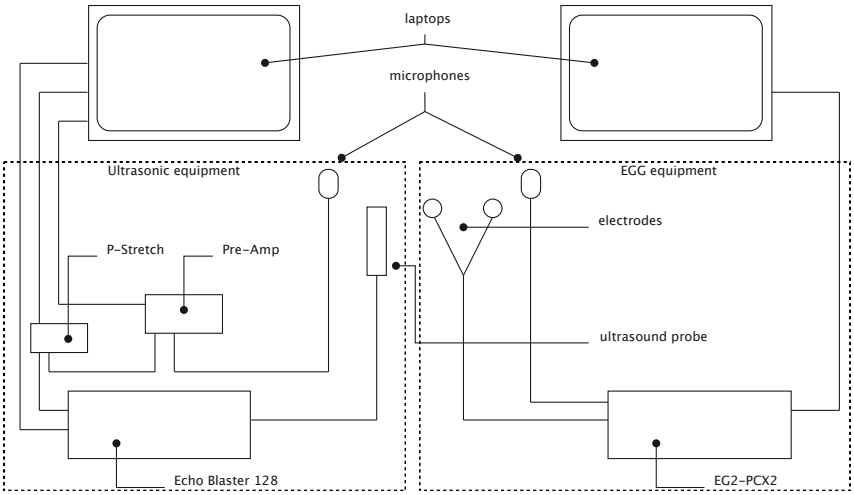


Figure 3.1: Schematics of the equipment set-up of Study I. Stabilisation headset not shown.

ducer (20mm radius, 2-4 MHz), plugged into one laptop via a USB cable. A P-Stretch unit (used for signal synchronisation) was connected to the ultrasound unit via a custom modification (Articulate Instruments Ltd™ 2011). The P-Stretch unit and a Movo LV4-O2 Lavalier microphone fed into a FocusRight Scarlett Solo pre-amplifier, which was plugged into the ultrasound laptop via USB. A second Movo microphone and the electrodes were connected to a Glottal Enterprises EG2-PCX2 unit, which was plugged into the second laptop (the audio signals from the UTI and the EGG units were used for synchronisation, see below).

The subject wore a metallic headset produced by Articulate Instruments Ltd™ (2008) (Figure 3.2), which stabilises the position of the ultrasound probe (allowing free head movement), and the Velcro strap with the EGG electrodes around their neck. The electrodes were located on each side of the thyroid cartilage, at the level of the glottis. The microphones were clipped to the headset on either side, at identical heights. Before the reading task, the participant's occlusal plane was obtained using a bite plate (Scobbie et al. 2011). This procedure allows data to be rotated along the occlusal plane and provides us with a reference plane.

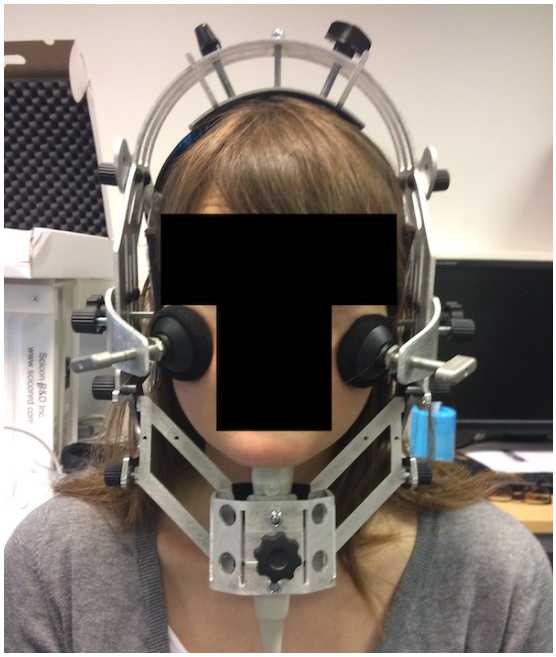


Figure 3.2: Ultrasonic probe stabilisation headset.

3.1.4 Procedure and data processing

The participants read sentence stimuli containing test words presented on the screen via the software Articulate Assistant Advanced™ (AAA v2.17.2, Articulate Instruments Ltd™ 2011). The test words were $C_1V_1C_2V_2$ words, where C_1 = /p/, V_1 = /a, o, u/, C_2 = /t, d, k, g/, and V_2 = V_1. Note that in both languages C2 is the onset of the second syllable. The choice of the segmental make-up of the test words was constrained by the use of ultrasound tongue imaging. The words were embedded in the frame sentence *Dico X lentamente* 'I say X slowly', and *Mówię X teraz* 'Say X now' for Italian and Polish respectively. Chapter 4 provides more details on the rationale behind the material design.

The UTI+audio and EGG+audio signals were acquired and recorded by means of AAA and Praat (Boersma & Weenink 2018) respectively. Since the signals from the ultrasonic machine and the electroglottograph are recorded simultaneously but separately, data from both sources were synchronised after acquisition. Synchronisation was achieved using the cross-correlation of the audio signals obtained from the separate sources (Grimaldi et al. 2008). The interval between the start of the cross-correlated signal and the time of the signal maximum amplitude is equal to the lag between

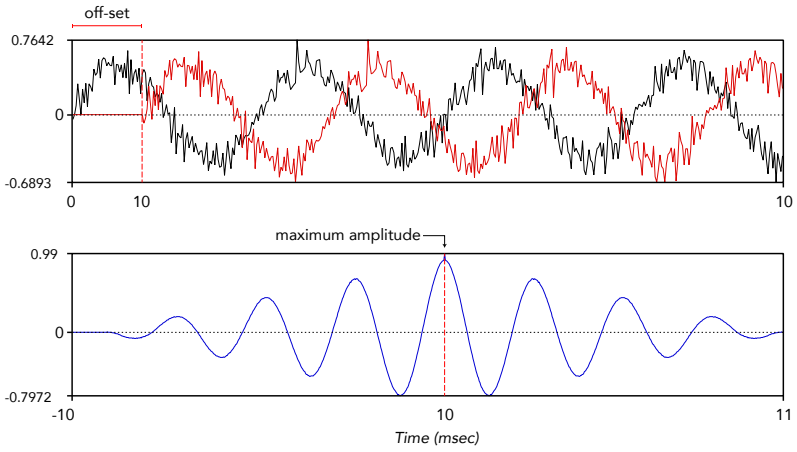


Figure 3.3: Synchronisation of two sounds by cross-correlation. The top panel shows the waveforms of two identical sounds, one of which has been offset by 10 ms. The bottom panel shows the cross-correlation of the two sounds in the top panel. The offset of the sounds corresponds to the time of maximum amplitude in the cross-correlated sound.

the two original sounds (Figure 3.3). Synchronisation of the original sound files was achieved by trimming the beginning of the longer sound by the lag obtained from cross-correlation. A Praat script was written to automate the synchronisation process (`sync-egg.praat` in Coretta 2018a).

A time-aligned transcript of the recordings was obtained with a force-alignment procedure using the SPeech Phonetisation Alignment and Syllabification (SPPAS) software (Bigi 2015). SPPAS is a language-agnostic system which comes with pre-packaged models for a variety of languages, among which Italian, Polish, and English. Since the audio recorded with the EGG system was nosier than that recorded with the UTI system, the latter was used in all subsequent acoustic-based analyses. The output of the force-alignment is a Praat TextGrid with time-aligned interval tiers containing the annotations of intonational phrase units (utterances), words, and phones. The automatic annotation was then manually checked by the author and corrected if necessary. The placement of segment boundaries followed the suggestions in Machač & Skarnitzl (2009). See Chapter 4 for details.

Detection of the release of C1 and C2 was accomplished through the algorithmic

procedure described in Ananthapadmanabha et al. (2014). I have written a custom implementation of the procedure in Praat for this study (`release-detection-c1.praat` and `burst-detection.praat` in Coretta 2018a). The output of the automatic detection was manually checked and corrected. The times of the following landmarks were extracted via a custom Praat script (`get-duration.praat` in Coretta 2018a): sentence onset and offset, target word onset and offset, C1 release, V1 onset, V1 offset/C2 closure onset, C2 release, V2 onset.

UTI data processing was performed in AAA. Spline curves were fitted to the tongue surface images using a built-in automatic batch procedure, within a search area defined by the interval between the onset of the CV sequence preceding the target word and the offset of the that following it (*Di[co X le]natamente*, *Mów[ię X te]raz*). The search area was created via Praat scripting (with `search-area.praat` in Coretta 2018a) and imported in AAA for the batch procedure. The automatic fitting procedure was monitored by the author and manual correction of the fitted splines was applied if necessary. Splines were fitted to the tongue contours at the original frame rate, which varied between 43 and 68 frames per second depending on the participant. The ranges of other UTI settings were: 88-114 scan lines, 980–988 pixels per scan line, field of view 71–93°, pixel offset 109–263, scan depth 75–180 mm. After fitting, the splines were linearly interpolated from the original frame rate to a sampling rate of 100 kHz (this is a default feature in AAA).

Subsequent data processing followed the method described in Strycharczuk & Scobbie (2015) (see also Chapter 6 and Appendix A) and it was based on the upsampled (100 kHz) spline data. Tongue displacement was obtained with a built-in procedure by tracking the time-varying displacement of the interpolated tongue splines along fan-lines from a fan-like coordinate system (Scobbie et al. 2011). Tongue displacement was measured for (1) the tongue tip, (2) the tongue dorsum, and (3) the tongue root. These were broadly defined as follows: (1) tongue tip as the region of the tongue that produces the closure of coronal stops, (2) tongue dorsum as the region that produces the closure of velar stops, (3) tongue root as the region between the hyoid bone shadow and the tongue dorsum region. Within each tongue region, the fan-line with the highest standard deviation was chosen as the vector for calculating tongue displacement (these fan-lines

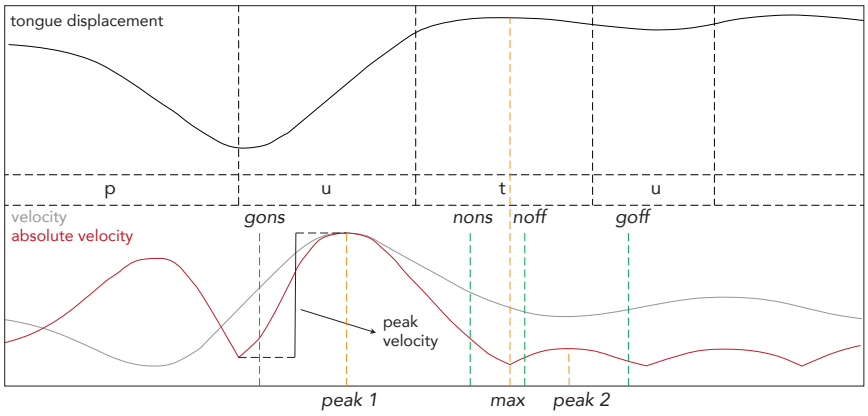


Figure 3.4: Gestural landmarks of tongue movements in the word *putu*. The top panel shows tongue tip displacement, while the bottom panel shows tangential and absolute velocity.

were manually chosen for each speaker individually). A Savitzky–Golay smoothing filter (second-order, frame length 75 ms) was applied to raw tongue displacement along the chosen fan-line to generate smoothed displacement values. Tangential velocity was calculated from the smoothed displacement signal of the tongue tip and tongue dorsum with a Savitzky–Golay filter (second-order, frame length 75 ms), as implemented in AAA.

The absolute values of the tangential velocity were used for the identification of gestural landmarks using a built-in algorithm. The times of the following gestural landmarks were obtained for the tongue tip and the tongue dorsum (Figure 3.4 exemplifies the tongue tip case): (a) maximum tongue displacement (MAX), (b) peak velocity before MAX (PEAK_1), (c) peak velocity after MAX (PEAK_2), (d) gesture onset (GONS), corresponding to the time when absolute velocity of tongue displacement reaches 20% of the peak absolute velocity before PEAK_1, (e) gesture nucleus (plateau) onset (NONS), when velocity is at 20% of the peak velocity between PEAK_1 and MAX, (f) nucleus offset (NOFF), when velocity is at 20% of the peak velocity between MAX and PEAK_2, and (g) gesture offset (GOFF), when velocity is at 20% of peak velocity after PEAK_2 (Kroos et al. 1997; Gafos et al. 2010). The time resolution of the gestural landmarks is 100 kHz (the sampling rate of the upsampled spline data).

The Cartesian coordinates of the fitted splines, together with tongue displacement,

tangential velocity, and absolute velocity of the tongue root, dorsum and tip, were obtained at the times corresponding to the following gestural landmarks: C2 closure onset, tongue maximum displacement during C2 closure (MAX), peak absolute velocity before and after tongue maximum displacement (PEAK_1 and PEAK_2), tongue closing gesture onset (GONS), closing gesture nucleus onset (NONS) and offset (NOFF). Tongue displacement, and the tangential and absolute velocities of the tongue root, dorsum, and tip were also extracted as time-series along the entire duration of V1. The coordinates of the splines were converted from Cartesian to polar for statistical analysis (see Appendix A). As discussed in Section 2.1, Study I set out to gather data on acoustic segmental durations, the timing of the consonantal gestures, and properties of vocal fold vibration. For this reason, tongue movement velocity was not analysed as part of the current investigation, and future work on this aspect is warranted.

Finally, wavegram data was extracted from the EGG data, following the method proposed by Herbst et al. (2010). In brief, a wavegram is a spectrogram-like representation of the EGG signal, where individual glottal cycles are sequentially placed on the *x*-axis, the normalised time of the samples taken from within each glottal cycle is the *y*-axis, and the third dimension, represented by colours of different shading, is the normalised amplitude of the EGG signal. Wavegram data was obtained from the entire duration of V1. See Chapter 7 for the detailed procedure.

3.2 Compensatory aspects of the effect of voicing on vowel duration in English (Study II)

The results from Study I indicated that the temporal distance between two consecutive stop releases in disyllabic words of Italian and Polish is not affected by the voicing of the second stop (Chapter 4). Within this temporally stable interval, differences in the timing of oral closure determines the respective durations of the vowel and the consonant closure. Thus, a second study was carried out to assess whether the durational pattern found in Study I would generalise to English, and to investigate differences between disyllabic and monosyllabic words as predicted by a word-holistic account of gestural

phasing (Chapter 5). It was expected that, while the duration of the release-to-release interval would be insensitive to the voicing status of the second consonant, the interval would be longer in monosyllabic words with a voiced consonant. Chapter 5 discusses the empirical and theoretical foundation of the research hypotheses and methods of Study II more in detail. A brief synthesis of the methodologies of the study is reported here.

Fifteen university students were recorded in a sound attenuated room in the Phonetics Laboratory of the University of Manchester while reading sentence stimuli containing test words, presented on a computer screen with PsychoPy (Peirce 2009). The participants were native speakers of British English, aged between 20 and 29, born and raised within Greater Manchester.[1] The test words were built according to the following structure: $C_1\acute{V}_1C_2$(VC), where C_1 = /t/, V_1 = /iː, ɜː, ɑː/, C_2 = /p, b, k, g/, and (VC) = /əs/. The issue of the syllabification of word-medial consonants in English is far from settled and still a controversial topic in English phonology. All possible syllabification options continue to be defended (Bermúdez-Otero 2013): onset-maximal syllabification, rhyme-maximal syllabification, and ambisyllabicity. Bermúdez-Otero (2013) offers initial phonological evidence for preferring an onset-maximal approach. This view is adopted as a working assumption in this work, and C2 in CVCV words are taken to be onsets. Future articulatory work is needed to shed light on the syllabification issue. The target words were embedded within five different frame sentences, so that each speaker would read a total of 120 stimuli (24 words × 5 frames). The audio recordings were force-aligned and the times of acoustic landmarks were extracted according to the same procedure as in Study I (with `make-textgrid.praat` and `get-measurements.praat` in Coretta 2019a).

3.3 Open Science

Open Science is a movement that stresses the importance of a more honest and transparent scientific attitude by promoting a series of research principles and by warning

[1] No sociolinguistic information was collected for this study, given that sociolinguistic considerations were not part of the study aims and given the sensitivity of the data.

from common, although not necessarily intentional, questionable practices and misconceptions. The term Open Science as a whole refers to the fundamental concepts of "openness, transparency, rigour, reproducibility, replicability, and accumulation of knowledge" (Crüwell et al. 2018:3). The goodness of the latter depends in great part on the reproducibility and replicability of the studies that contribute to knowledge accumulation. While the terms "reproducibility" and "replicability" are generally used interchangeably, they refer to two different ideas. A study is *replicable* when researchers can independently run the study on new subjects/data and obtain the same results (in brief, same analysis, different data/researchers). The *reproducibility* of a study is, instead, related to the ability of independent researchers to run the original analysis on the original data and obtain the same results as those presented by the original authors, pending enough information on the analysis procedures is given (in brief, same analysis, same data).

A sense of need for Open Science, now increasingly spreading to different disciplines and enterprises, arose primarily from the ongoing so-called "replication crisis" (Pashler & Wagenmakers 2012; Schooler 2014), which has attracted the most attention within the circles of medical and psychological sciences. Recent attempts to replicate results from high-impact studies in psychology have demonstrated an alarmingly high rate of failure to replicate. For example, in a replication attempt of 100 psychology studies, only 39% of the original results were rated by annotators as successfully replicated (Open Science Collaboration 2015). Failure to replicate previous results have been claimed to be a consequence of low statistical power (Button et al. 2013), and of so-called questionable research and measurement practices (Simmons et al. 2011; Morin 2015; Flake & Fried 2019). The following sections discuss these problems in turn.

3.3.1 "With great power comes great replicability"

One of the issues that can affect statistical analysis is related to errors in rejecting the null hypothesis.[2] A researcher could falsely reject the null hypothesis when in fact is

[2]The quote in the title is from a 2016 twitter status by Nathan C. Hall (`https://twitter.com/prof_nch/status/790744443313852417?s=20`).

correct (Type I errors, an effect is found when there is none), or they could falsely fail to reject the null hypothesis when in fact it should have been (Type II errors, an effect is not found when there is one). Type I and Type II errors do occur and cannot be totally prevented. Rather, the aim is to keep their rate of occurrence as low as possible. The generally accepted rates of Type I and Type II errors are 0.05 and 0.2 respectively (usually referred to as the α and β levels). This means that, in a series of imaginary multiple replications of a study, 5% of the times the null hypothesis will be falsely rejected, and 20% of the times will falsely be not rejected. A concept closely related to Type II errors is statistical power, which is the probability of correctly rejecting the null hypothesis when it is false (calculated as $B = 1 - \beta$). In other words, power is the probability of detecting an effect equal or greater than a specified effect size. Given the standard $\beta = 0.2$, an accepted (minimum) power threshold is 80% (which means that an effect equal or greater than a chosen size will be detected 80% of the time).

Two other types of statistical errors are the Type S (sign) and Type M (magnitude) errors (Gelman & Tuerlinckx 2000; Gelman & Carlin 2014). Type S errors refer to the probability of the estimated effect having the wrong sign (for example, finding a positive effect when in reality the effect is negative), while Type M errors correspond to the exaggeration ratio (the ratio between the estimated and the real effect). When the statistical power of a study is low (below 50%), Gelman & Carlin (2014) show that the exaggeration ratio (Type M error) is particularly high (from 2.5 up to 10 times the true effect size). Type S errors (wrong sign) are more common at lower power levels (below 10%), although these can easily arise due to small sample sizes and high variance.

Several researchers have shown that the average statistical power of studies in different disciplines is very low (35% or below) and that the last 50 years did not witness an improvement. Bakker et al. (2012) show that the median statistical power in psychology is 35%, while Button et al. (2013) reports a median of 21% obtained from 48 neuroscience meta-analyses. In Dumas-Mallet et al. (2017), half of the surveyed biomedical studies (N = 660) have power below 20%, while the median ranges between 9% and 30% depending on the subfield. Rossi (1990) and Marszalek et al. (2011) show that from the 70s to date there hasn't been an increase in power and sample sizes. Tressoldi & Giofré (2015) also find that only 2.9% of 853 studies in psychology report a prospec-

tive power analysis for sample size determination, i.e. the estimation of the smallest sample size necessary to obtain a certain power level before the experiment is run. In sum, low statistical power (well below the recommended 80% threshold) seems to be the norm.

3.3.2 The dark side of research

Questionable research and measurement practices are practices that negatively affect the scientific enterprise, but that are employed (most of the time unintentionally) by a surprisingly high number of researchers (John et al. 2012). Silberzahn et al. (2018) asked 29 teams (61 analysts) to answer the same research question given the same data set, and showed that data analysis can be highly subjective. A total of 21 unique combinations of predictors were used across the 29 teams, leading to diverging results (20 teams obtained a significant result, while 9 did not). At various stages of the study timeline, a researcher can exploit the so-called "researcher's degrees of freedom" to obtain a significant result (Simmons et al. 2011). The researcher's degrees of freedom create a "garden of forking paths" (Gelman & Loken 2013), that the researcher can explore until the results are satisfactory (i.e., they lead to high-impact or expected findings).

P-hacking is a general term that refers to the process of choosing and reporting those analyses that change a non-significant *p*-value to a significant one (Simmons et al. 2011; Wagenmakers 2007; Motulsky 2014). *P*-hacking can be achieved by several means, for example by trying different dependent variables, including and/or excluding predictors, selective inclusion/exclusion of subjects and observations, or sequential testing (collecting data until the results are significant). Another common practice is to back-engineer a hypothesis after obtaining unexpected results, also known as Hypothesising After the Results are Known (HARKing, Kerr 1998). Lieber (2009) warns against "double dipping", or the use of the same data to generate a hypothesis and test it. Morin (2015) and Flake & Fried (2019) more specifically discuss questionable practices related to how research variables are measured and operationalised. The literature reviewed in Flake & Fried (2019) suggests that a very high percentage of published papers contains measures that are created on the fly but lack any reference to reliability tests. Researchers have also been found to manipulate validated scales to obtain desired results.

Cognitive biases and statistical misconceptions can also have a negative impact on research conduct. Wagenmakers et al. (2012) discuss the effects of congnitive biases like the confirmation bias (the tendency to look for facts and interpretations that confirm one's prior conviction, Nickerson 1998) and the hindsight bias (the tendency to find an event less surprising after it has occurred, Roese & Vohs 2012). Greenland (2017) defines further common distortions pertaining to methodological approaches, like statistical reification (interpreting statistical results as reflections of an actual physical reality). Finally, Wagenmakers (2007) and Motulsky (2014) examine mistaken beliefs about the meaning of *p*-values and statistical significance (like interpreting *p*-values as an index to statistical evidence or the idea that *p*-values inform us about the likelihood of the null-hypothesis given the data).

A bias in the observed effects can also arise at the stage of publication. A publication bias has been observed in that significant and novel results are generally favoured over null results or replications (Easterbrook et al. 1991; Ioannidis 2005; Song et al. 2010; Kicinski 2013; Nissen et al. 2016). Rosenthal (1979) called the bias against publishing null results the "file drawer" problem. Studies that don't lead to a significant result are stored in a metaphorical file drawer and forgotten. This practice not only can bias meta-analytical effect sizes, but also allows for waste of resources when studies with undisclosed null results are repeatedly performed. The questionable research and measurement practices described above, together with publication bias, conspire to unduly increase confidence in our research outcomes. A final exculpatory note is due, though, in that these practices are not necessarily intentional or fraudulent, and in some cases lie within a "grey area" of accepted standard procedures.

3.3.3 Where we stand and where we are heading

Given the similarities in methods between the psychological sciences and phonetics/phonology, it is reasonable to assume that the situation does not fare better in the latter. As mentioned above, sample size, coupled with the effects of increased variance due to between-subject designs, can have a big impact on statistical power. Kirby & Sonderegger (2018) suggest that the number of participants in phonetic studies is generally low, and that, even with nominally high-powered sample sizes, estimation

of small effect sizes is subject to the power-related issues discussed above (especially Type S/M errors). Nicenboim et al. (2018) further show how low statistical power has adverse effects on the investigation of phonetic phenomena characterised by small effect sizes, like incomplete neutralisation. Winter (2015) further argues that the common practice of using few items (e.g. word types) and a high number of repetitions increases statistical certainty of the estimates of idiosyncratic differences between items rather than those of the sought effects. Roettger (2019) discusses how the inherently multidimensional nature of speech favours exploration of the researcher's degrees of freedom, by allowing the researcher to navigate through a variety of choices of phonetic correlates and their operationalisation.

In a review of 113 studies of acoustic correlates of word stress in a variety of languages, published between 1955 and 2017, Roettger & Gordon (2017) show that the majority of studies include 1 to 10 speakers (mode = 1), 1 to 40 lexical items, and 1 to 6 repetitions. A follow-up analysis conducted on the same data indicates that the median number of participants per study is 5 (see Appendix E). A few recent studies (2010 onwards) constitute a clear exception by having more than 30 participants. However, no apparent trend of increasing average number of speakers can be observed and the situation has been fairly stable over the years. Finally, the language endangerment status has a small but negligible negative effect on participants' number in vulnerable and definitely endangered languages, but not so much in severely and critically endangered ones. It is reasonable to assume that, based on this cursory analysis, sample size in phonetic studies is generally very low, independent from publication year and endangerment status.

As a partial remedy to the issues discussed so far, researchers have proposed two solutions: pre-registrations and Registered Reports. Pre-registration of a study consists in the researchers' commitment to an experimental and analytical protocol before collecting and seeing the data (Wagenmakers et al. 2012; van 't Veer & Giner-Sorolla 2016). Pre-registering a study establishes a clear separation between confirmatory (hypothesis-testing) analyses and exploratory (hypothesis-generating) research. While both types of research are essential to scientific progress (Tukey 1980), presenting exploratory analyses as confirmatory is detrimental to it. Pre-registrations ensure researchers comply

to such demarcation, while leaving space to generate new hypotheses via exploratory research. A more recent initiative proposes Registered Reports as a publication format that can counteract questionable research practices and the exploitation of the researcher's degrees of freedom (Chambers et al. 2015). At the time of writing, no journal specialised in phonetics/phonology offers this article format, although it is currently under implementation at the Journal of the Association for Laboratory Phonology and a few other journals focussed on other linguistic fields.[3]

Another incentive to developing a transparent research attitude comes from aspects of reproducibility. As discussed above, a research analysis is reproducible when different researchers obtain the same results as in the published study by running the same analysis on the same data. Ensuring full reproducibility also means ensuring computational reproducibility, or in other words enabling researchers to perform the original analysis in an identical computational environment (Schwab et al. 2000; Fomel & Claerbout 2009). Peng (2009) mentions exposed cases of fraudulent data manipulation and unintentional analysis errors that call for policies of reproducibility to ensure accountability of published results. Our field is not immune from these issues (see for example the "Yokuts vowels" case, Weigel 2002, 2005; Blevins 2004), and the idea of reproducibility is not new to linguistics in general (Bird & Simons 2003; Thieberger 2004; Maxwell & Amith 2005; Maxwell 2013; Cysouw 2015; Gawne et al. 2017) nor to phonetics/phonology specifically (Abari 2012; Roettger 2019).

The objective of making research accountable can be achieved by publicly sharing data (subject to ethical restrictions), analysis code, and detailed information on the software that produced the results (Sandve et al. 2013). Sharing data is also fundamental for the accumulation of knowledge, for example in the context of meta-analytical studies. Several services are now available which offer free online data storage and versioning, like the Open Science Framework, GitHub, and DataHub. Extensive documentation of code takes on an important role, and the paradigm of literate programming offers a practical solution (Knuth 1984). Within the literate programming framework, code and documentation coexist within a single source file, and code snippets are in-

[3]See the spreadsheet at this link for a curated list: `https://docs.google.com/spreadsheets/d/17dLaqKXcjyWk1thG8y5C3_fHXXNEqQMcGWDY62BOc0Q/edit?usp=sharing`.

terweaved with their documentation. Reproducible reporting further implements this concept (Peng 2015) by automating the generation and inclusion of summary tables, statistics, and figures in a paper using statistical software like R (R Core Team 2019). In a reproducible report, data and results are computationally linked via the statistical software, and changes in data or analyses are reflected in changes in the results appearing in the text. This workflow reduces chances of reporting errors and facilitates validation of the data analyses by other researchers.

3.3.4 Putting this into practice

The research project behind this dissertation has been carried out with principles of Open Science in mind. The reader might notice, however, a certain arc of development in putting these concepts into practice, since my understanding of Open Science practices evolved during the realisation of the project. This last section of the Introduction discusses how the principles and methods expounded in the above sections were applied in this project.

Study I was not pre-registered, since the absence of a specific hypothesis did not allow for the formulation of a corresponding analysis. Sample size was determined on the basis of practical considerations concerning the time required for processing ultrasound tongue imaging data. Several different parameters and models have been explored in search of relevant patterns, while only a few of these have been reported in the papers of this study. The models of Study I were conducted within a Neyman-Pearson (frequentist) framework of statistical inference (Perezgonzalez 2015), while a Bayes factor analysis (Kass & Raftery 1995) was performed in those cases that demanded it (Chapter 4). To compensate for the asymmetry between the number of models run and those reported, less attention has been given to *p*-values, while greater focus was placed on effect sizes and their hypothesis-generating value. The design and analyses of Study II ~~have been~~ pre-registered on the Open Science Framework (`https://osf.io/2m39u/`). Sample size determination followed the method of the Region Of Practical Equivalence (Vasishth et al. 2018b). Data from Study II was subject to a full Bayesian analysis, and greater emphasis was given on the posterior probability distributions of the effect estimates rather than on their means (Chapter 5).

Data, code, and information about software can be found in the research compendium of this project on the Open Science Framework (Coretta 2020, `https://osf.io/w92me/`). This OSF repository acts as a hub for the research compendia of the project and of the individual papers. The data has been packaged and documented using R (Marwick et al. 2017). The data packages (coretta2018itapol and coretta2019eng) are available on GitHub and links to them are given in the OSF repository (see the Data components in the repository).[4] Following the recommendations in Berez-Kroeker et al. (2018), each data component is given a standard bibliographical citation (Coretta 2018a, 2019a).

Data processing using Praat scripting followed the principles of literate programming, i.e. code and documentation coexist in a single source file. The source files were written in literate markdown, a flavour of markdown that allows extracting code from markdown text to build the scripts from the source file. The scripts were extracted from the literate source file using the Literate Markdown Tangle software[5] and were run with a custom R package, speakr (Coretta 2019d). The documentation of the scripts was generated with Pandoc[6] and a custom Praat syntax highlighting extension (available in speakr). All analyses and derived figures in this dissertation can be reproduced in R (R Core Team 2019) with the scripts found in the papers' respective compendia.

The following software and packages were used: tidyverse (Wickham 2017), lme4 (Bates et al. 2015), lmerTest (Kuznetsova et al. 2017), effects (Fox & Weisberg 2019), broom.mixed (Bolker & Robinson 2019), mgcv (Wood 2017), itsadug (van Rij et al. 2017), Stan (Stan Development Team 2017), brms (Bürkner 2017), tidybayes (Kay 2019). I developed two R packages for the visualisation of generalised additive models using tidyverse software (tidymv, Coretta 2019e) and the analysis of ultrasound tongue imaging data with mgcv (rticulate, Coretta 2018b). This dissertation was written in R Markdown, and typeset with knitr and bookdown (Xie 2014, 2016; Xie et al. 2018; Xie 2019).

[4] The raw ultrasound tongue imaging AAA files were not uploaded given their size (almost 40 GB) and can be obtained from the author.

[5] `https://github.com/driusan/lmt`

[6] `https://pandoc.org`

Part II

Original publications

Chapter 4

An exploratory study of voicing-related differences in vowel duration as compensatory temporal adjustment in Italian and Polish [Paper I]

This paper has been published in Glossa as:
Coretta, Stefano. 2019. An exploratory study of voicing-related differences in vowel duration as compensatory temporal adjustment in Italian and Polish. *Glossa*(4)1. 1–25.
DOI: `https://doi.org/10.5334/gjgl.869`.
When citing, please refer to the published version.

Abstract

Over a century of phonetic research has established the cross-linguistic existence of the so called 'voicing effect', by which vowels tend to be shorter when followed by voiceless stops and longer when the following stop is voiced. However, no agreement is found among scholars regarding the source of this effect, and several causal accounts have been advanced. A notable one is the compensatory temporal adjustment account, according to which the duration of the vowel is inversely correlated with the stop clo-

sure duration (voiceless stops having longer closure durations than voiced stops). The compensatory account has been criticised due to lack of empirical support and its vagueness regarding the temporal interval within which compensation is implemented. The results from an exploratory study of Italian and Polish suggest that the duration of the interval between two consecutive stop releases in CVCV words in these languages is not affected by the voicing of the second stop. The durational difference of the first vowel and the stop closure would then follow from differences in timing of the VC boundary within this interval. While other aspects, like production mechanisms related to laryngeal features effects and perceptual biases cannot be ruled out, the data discussed here are compatible with a production account based on compensatory mechanisms.

4.1 Introduction

Almost a hundred years of research have consistently shown that consonantal voicing has an effect on preceding vowel duration: vowels followed by voiced obstruents are longer than when followed by voiceless ones (Meyer 1904; Heffner 1937; House & Fairbanks 1953; Belasco 1953; Peterson & Lehiste 1960; Halle & Stevens 1967; Chen 1970; Klatt 1973; Lisker 1974; Laeufer 1992; Fowler 1992; Hussein 1994; Lampp & Reklis 2004; Warren & Jacks 2005; Durvasula & Luo 2012). This so called "voicing effect" has been found in a considerable variety of languages.[1] These include (but are not limited to) English, German, French, Spanish, Hindi, Russian, Italian, Arabic, and Korean (see Maddieson & Gandour 1976 for a more comprehensive, but still not exhaustive list).[2] Despite of the plethora of evidence in support of the *existence* of the voicing effect, agreement hasn't been reached regarding its *source*.

Several proposals have been put forward in relation to the possible source of the voicing effect (see Sóskuthy 2013 and Beguš 2017 for an overview). Some of the proposed mechanisms for the emergence of the voicing effect refer to properties of speech

[1] One of the first attestations of the term 'voicing effect' can be attributed to Mitleb (1982). Another term used to refer to the same phenomenon is 'pre-fortis clipping', probably introduced by Wells (1990).

[2] A typological note. Most languages reported having a voicing effect come from the Indo-European family. Others are from a pool of widely studied languages. It is thus of vital importance that future studies look at other language families and underdocumented/underdescribed languages.

production. A notable production account, which will be the focus of this study, is based on compensatory temporal adjustments (Lindblom 1967; Slis & Cohen 1969a,b; Lehiste 1970a,b). According to this account, the voicing effect follows from the reorganisation of gestures within a unit of speech the duration of which is not affected by stop voicing. The duration of such a unit is held constant across voicing contexts, while the duration of voiceless and voiced obstruents differs. The closure of voiceless stops is longer than that of voiced stops (Lisker 1957; Summers 1987; Davis & Summers 1989; de Jong 1991). As a consequence, vowels followed by voiceless stops (which have a long closure) are shorter than vowels followed by voiced stops (which have a short closure). Advocates of a compensatory mechanism propose two prosodic units as the scope of the temporal adjustment: the syllable (and, equivalently, the VC sequence or vowel-to-vowel interval, Lindblom 1967; Farnetani & Kori 1986), and the word (Slis & Cohen 1969a,b; Lehiste 1970a,b). However, the compensatory temporal adjustment account has been criticised in subsequent work.

Empirical evidence and logic challenge the proposal that the syllable or the word have a constant duration and hence drive compensation. First, Lindblom's 1967 argument that the duration of the syllable is constant is not supported by the findings in Chen (1970) and Jacewicz et al. (2009). Chen (1970) rejects a syllable-based compensatory mechanism in the light of the fact that the duration of the syllable is affected by consonant voicing. Jacewicz et al. (2009) further show that the duration of monosyllabic words in American English changes depending on the voicing of the coda consonant. Second, although the results in Slis & Cohen (1969a) suggest that the duration of disyllabic words in Dutch is constant whether the second stop is voiceless or voiced, it does not follow from this fact that compensation should necessarily target the vowel preceding the stop. Indeed, it is logically possible that the following unstressed vowel could be the target of the compensation, therefore differences in preceding vowel duration still call for an explanation.

The compensatory temporal adjustment account has been further challenged on the basis of the so-called "aspiration effect" (Maddieson & Gandour 1976), by which vowels are longer when followed by aspirated stops than when followed by unaspirated stops. In Hindi, vowels before voiceless unaspirated stops are short, vowels followed

by voiced aspirated stops are long, and vowels followed by voiced unaspirated and voiceless aspirated stops are in between and have similar durations. Maddieson & Gandour (1976) find no compensatory pattern between vowel and consonant duration. The consonant /t/, which has the shortest duration, is preceded by the shortest vowel, and vowels before /d/ and /tʰ/ have the same duration although the durations of the two consonants are different. Maddieson & Gandour (1976) argue that a compensatory explanation for differences in vowel duration cannot be maintained.

However, a re-evaluation of the way consonant duration is measured in Maddieson & Gandour (1976) might actually turn their findings in favour of a compensatory account. Due to difficulties in detecting the release of the consonant of interest, consonant duration in Maddieson & Gandour (1976) is measured from the closure of the relevant consonant to the release of the following, (e.g., in *ab sāth kaho*, the duration of /tʰ/ in *sāth* is calculated as the interval between the closure of /tʰ/ and the release of /k/). This measure includes the burst and aspiration (if present) of the consonant following the target vowel. Slis & Cohen (1969a), however, state that the inverse relation between vowel duration and the following consonant applies to *closure* duration, and not to the entire *consonant* duration.[3] If an inverse relation exists between vowel and closure duration, the inclusion of burst and/or aspiration clearly alters this relationship.

Indeed, the study on Hindi voicing and aspiration effects conducted by Durvasula & Luo (2012) indicates that closure duration, measured from closure onset to closure offset, decreases according to the hierarchy voiceless unaspirated > voiced unaspirated > voiceless aspirated > voiced aspirated, which closely resembles the order of increasing vowel duration in Maddieson & Gandour (1976). Nonetheless, Durvasula & Luo (2012) do not find a negative correlation between vowel duration and consonant closure duration, but rather a (small) *positive effect*. Vowel duration increases with closure duration when voicing and aspiration are taken into account. However, as noted in Beguš (2017), it is likely that this result is a consequence of not controlling for speech rate. A small negative effect of closure duration can turn positive if the effect of speech rate

[3]In this paper, I use the term *relation* to mean a categorical pattern of entailment (like in 'a long vowel entails a short closure'), while the term *correlation* is reserved to a statistical correlation of two continuous variables.

(which is positive) is greater, given the cumulative nature of these effects.

de Jong (1991) finds partial support for a compensatory mechanism between vowel and closure duration in an electro-magnetic-articulometric study of two American English speakers. The duration of vowels in nuclear accented, pre-, and post-nuclear accented position is weakly negatively correlated with closure duration (the slope coefficients range between -0.12 and -0.35, meaning that the amount of durational compensation is between 10% and 35%). While the magnitude of the correlation is too weak to univocally support compensation, the direction of the correlation is correct (i.e. a negative correlation).

Further evidence for a compensatory account and a negative correlation between vowel and closure duration comes from the effect of a third type of consonants, namely ejectives. Beguš (2017) finds that in Georgian (which contrasts aspirated, voiced, and ejective consonants) vowels are short when followed by voiceless aspirated stops, longer before ejective stops, and longest when followed by voiced stops. Crucially, stop closure duration follows the reversed pattern: closure is short in voiced stops, longer in ejectives, and longest in voiceless aspirated stops. Moreover, vowel duration is inversely correlated with closure across the three phonation types. Beguš (2017) mentions the possibility that the negative correlation is an artefact of the vowel and closure intervals sharing a boundary. This annotation bias could generate negative correlations (by which the vowel would shorten and the closure would lengthen by the same amount when, for example, the boundary is placed to the left of the "actual" boundary). However, Beguš shows with a cross-annotator analysis that this was not the case. Beguš (2017) argues that these findings support temporal compensation (although not univocally, see Beguš 2017:Section V, and Section 4.4.2 of this paper).

To summarise, a mechanism of compensatory temporal adjustment has been proposed as the pathway to the emergence of the voicing effect. According to such an account, the difference in vowel duration before consonants varying in voicing (and possibly other phonation types) is the outcome of a compensation between vowel and closure duration. After reviewing the critiques advanced by Chen (1970) and Maddieson & Gandour (1976), and in face of the results in Slis & Cohen (1969a), de Jong (1991) and Beguš (2017), a temporal compensation mechanism gains credibility. However,

issues about the actual implementation of the compensation mechanism still remain. While compensatory temporal adjustments are plausible in light of the reviewed literature, we are still left with the necessity of identifying a speech interval the duration of which is not affected by the voicing of the post-vocalic consonant, and within which compensation can be logically implemented.

4.1.1 The present study

This paper reports on selected results from a broader exploratory study that investigates the relationship between vowel duration and consonant voicing from both an acoustic and articulatory perspective. Synchronised recordings of audio, ultrasound tongue imaging, and electroglottography were carried out to enable a data-driven approach to the analysis of features related to the voicing effect in the context of disyllabic (CV́CV) words in Italian and Polish.[4] This study was not designed to test the compensatory account, but rather to collect synchronised articulatory and acoustic data on the voicing effect. Moreover, the design of the study has been constrained by the use of ultrasound articulatory techniques (see Section 4.2). Since the tongue imaging and electroglottographic data don't bear on the main argument put forward here, only the results from acoustics will be discussed.

Italian and Polish reportedly differ in the magnitude (or presence) of the effect of stop voicing on vowel duration. On the other hand, the typical realisation of phonological voiced stops in these languages are similar (but see Huszthy 2016 and Schwartz & Arndt 2018 for a phonological and phonetic discussion on laryngeal aspects of Italian and Polish respectively).[5] Cyran (2011) argues for a distinction between voicing and aspirating varieties of Polish, based on phonological arguments. Waniek-Klimczak (2011), on the other hand, cautiously argues that a possible change in progress in Polish is affecting the VOT values of voiceless stops in pre-stressed position.

The non-clear status of Polish laryngeal phonology/phonetics could be seen as a

[4] As per Cysouw & Good (2013), the glossonyms *Italian* and *Polish* as used here to refer, respectively, to the languoids Italian [Glottocode: `ital1282`] and Polish [Glottocode: `poli1260`].

[5] Polish neutralises the voicing contrast word-finally, although the contrast is maintained word-medially (Gussmann 2007).

hindrance affecting the comparison with Italian. However, based on data from Italian, Kirby & Ladd propose that the distinction between voicing and aspirating languages itself (Beckman et al. 2013) cannot be straightforwardly mapped onto phonetics, and they remind us that "the production of laryngeal contrasts of all kinds are considerably more complex" than generally described in the phonological literature (Kirby & Ladd 2016:2409). Since this study focusses on the effect of post-stressed stops on preceding vowel durations, we believe that the comparison between Italian and Polish is still feasible, even in the case Polish voiceless pre-stressed stops are articulated with longer VOT values. Given that Italian and Polish share some features of the segmental and prosodic make-up of their phonological systems, the design of the experimental material and comparison of the results were facilitated. For these reasons, these languages offer an opportunity to investigate differences that could reveal mechanisms underlying the voicing effect, at least on a general level.

Italian has been unanimously reported as a voicing-effect language (Magno Caldognetto et al. 1979; Farnetani & Kori 1986; Esposito 2002). The mean difference in vowel duration when followed by voiceless vs voiced consonants ranges between 22 and 24 ms in these studies, with longer vowels followed by voiced consonants. The mean differences are based on 3 speakers in Farnetani & Kori 1986 and 7 speakers in Esposito 2002. Magno Caldognetto et al. (1979) don't report estimates of vowel duration, just the direction of the effect, but the study is based on 10 speakers.

The results regarding the presence and magnitude of the effect in Polish are instead mixed. Slowiaczek & Dinnsen (1985) find that vowels followed by word-final underlyingly voiced stops are 10–15 ms longer in 5 Polish speakers, although Jassem & Richter (1989) did not replicate their results. Similarly, Keating (1984b) reports a difference of 2 ms in the duration of stressed vowels in disyllabic words from 24 speakers, which the author argues to be non-significant. On the other hand, Nowak (2006) finds that vowels followed by voiced stops are 4.5 ms longer in the 4 speakers recorded. Malisz & Klessa (2008) argue based on data from 40 speakers that the magnitude of the voicing effect in Polish is highly idiosyncratic, and claim that their results are inconclusive on this matter. While they do not report estimates from the 40 speakers, a table with mean vowel durations from 4 suggests a mean difference before voiceless vs voiced stops of

3.5 ms. Finally, Strycharczuk (2012) reports a non-significant effect in 6 speakers in pre-sonorant word-final position.

The variety of results concerning the voicing effect in Polish could be related to differences in methodology. However, no clear pattern between studies which find a voicing effect and those which don't can be identified. For example, the studies reviewed here looked at either word-final or word-medial stops, controlled or read speech, speakers with a low or advanced proficiency in English. However, in all the individual cases both a positive and a negative result are reported depending on the study. What might be more relevant, though, is that the estimates of the difference in vowel duration are generally very low, between 3.5 and 15 ms. Given the small magnitude of the difference, it is likely that the failure to obtain significant *p*-values in some studies are due to low statistical power, rather than because of absence of the effect (as also hinted in Beguš 2017, see arguments in Roettger 2019 and Nicenboim et al. 2018).

The acoustic data from the study discussed here suggests that (1) a voicing effect can be detected both in Italian and Polish, and that (2) the duration of the interval between two consecutive stop releases (the release-to-release interval) is not affected by the voicing of the second consonant in both languages. This finding is compatible with a compensatory temporal adjustment account by which the timing of the closure onset of the stop following the vowel within said interval determines the respective durations of vowel and closure.

4.2 Method

4.2.1 Participants

Participants were sought in Manchester (UK), and in Verbania (Italy). Seventeen subjects in total participated in this study. Eleven subjects are native speakers of Italian (5 female, 6 male), while six are native speakers of Polish (3 female, 3 male). The Italian speakers are from the North and Centre of Italy (8 speakers from Northern Italy, 3 from Central Italy). The Polish group has 2 speakers from Western Poland, 3 speakers from Central Poland, and 1 speaker from Eastern Poland. For more information on the

sociolinguistic details of the speakers, see Section 4.6. Ethical clearance for this study was obtained from the University of Manchester (REF 2016-0099-76). The participants signed a written consent and received a monetary compensation of £10.

4.2.2 Equipment

The acquisition of the audio signal was achieved with the software Articulate Assistant Advanced™ (AAA, v2.17.2, Articulate Instruments Ltd™ 2011) running on a Hewlett-Packard ProBook 6750b laptop with Microsoft Windows 7. Audio recordings were sampled at 22050 Hz (16-bit) and saved in a proprietary format (`.aa0`). A FocusRight Scarlett Solo pre-amplifier and a Movo LV4-O2 Lavalier microphone were used for audio recording. The microphone was placed at the level of the participant's mouth on one side, at a distance of about 10 cm. The microphone was clipped onto a metal headset worn by the participant, which was part of the ultrasonic equipment.

4.2.3 Materials

The target stimuli were disyllabic words with $C_1V_1C_2V_2$ structure, where C_1 = /p/, V_1 = /a, o, u/, C_2 = /t, d, k, g/, and $V_2 = V_1$ (e.g. /pata/, /pada/, /poto/, etc.).[6] Most are nonce words, although inevitably some combinations produce real words both in Italian (4 words) and Polish (2 words, see Table 4.1). The lexical stress of the target words was placed by speakers of both Italian and Polish on V_1, as intended.

The make-up of the target words was constrained by the design of the experiment, which included ultrasound tongue imaging (UTI). Front vowels are difficult to be imaged with UTI, since their articulation involves tongue surface positions which are particularly far from the ultrasonic probe, hence reducing the visibility of the tongue

[6]Italian has both a mid-low [ɔ] and a mid-high [o] back vowel in its vowel inventory. These vowels are traditionally described as two distinct phonemes (Krämer 2009), although both their phonemic status and their phonetic substance are subject to a high degree of geographical and idiosyncratic variability (Renwick & Ladd 2016). As a rule of thumb, stressed open syllables in Italian (like the ones used in this study) have [ɔː] (vowels in penultimate stressed open syllables are long) rather than [oː] (Renwick & Ladd 2016). On the other hand, Polish has only a mid-low back vowel phoneme /ɔ/ (Gussmann 2007). For the sake of typographical simplicity, the symbol /o/ will be used here for both languages.

Table 4.1: Target words. Asterisks indicate real words.

Italian			Polish		
pata	poto*	putu	pata	poto	putu
pada	podo	pudu	pada*	podo	pudu
paca*	poco*	pucu	paka*	poko	puku
paga*	pogo	pugu	paga	pogo	pugu

contour. For this reason, only central and back vowels were included. Since one of the variables of interest in the study was the closing gesture of C_2, only lingual consonants were used. A labial stop was chosen as the first consonant to reduce possible coarticulation with the following vowel (although see Vazquez-Alvarez & Hewlett 2007). The number of target words was kept low to reduce the time required for completing the task, since the ultrasonic equipment can get very uncomfortable for the speaker when worn for more than 15/20 minutes.

The target words were embedded in a frame sentence. Controlling for meaning, segmental and prosodic make-up between languages proved to be difficult. The frames are *Dico X lentamente* 'I say X slowly' in Italian (following Hajek & Stevens 2008), and *Mówię X teraz* 'I say X now' in Polish. These sentences were chosen in order to maintain a similar intonation contour across languages.

4.2.4 Procedure

The participant was asked to read the sentences with the target words which were presented on the computer screen. The order of the sentences was randomised for each participant. Participants read the list of randomised sentence stimuli 6 times. Due to software constraints, the order of the list was kept the same across the six repetitions within each participant. The reading task lasted between 15 and 20 minutes, with optional short breaks between one repetition and the other. The total session time was around 45 minutes. Before the start of the experiment, the participants were spoken to in their mother tongue to try and reduce exposure to English prior to being recorded. Instructions were also given in their respective mother tongues. Each speaker read a

Table 4.2: Criteria for the identification of acoustics landmarks.

landmark		**criteria**
vowel onset	(V1 onset)	Appearance of higher formants in the spectrogram following the release of /p/ (C1)
vowel offset	(V1 offset)	Disappearance of the higher formants in the spectrogram preceding the target consonant (C2)
consonant onset	(C2 onset)	Corresponds to V1 offset
closure onset	(C2 closure onset)	Corresponds to V1 offset
consonant offset	(C2 offset)	Appearance of higher formants of the vowel following C2 (V2); corresponds to V2 onset
consonant release	(C1/C2 release)	Automatic detection + manual correction (Ananthapadmanabha et al. 2014)

total of 12 sentences for 6 times (with the exceptions of IT02, who repeated the 12 sentences 5 times), which yields a grand total of 1212 tokens (792 from Italian, 420 from Polish).

The experiment was carried out in two locations: in the sound attenuated booth of the Phonetics Laboratory at the University of Manchester, and in a quiet room in a field location in Italy (Verbania, Northern Italy). In both locations the equipment and procedures were the same. Data collection started in December 2016 and ended in March 2018.

4.2.5 Data processing and measurements

The audio recordings were exported from AAA in the `.wav` format for further processing. The sample and bit rate were kept as upon recording (22050 Hz, 16-bit). A forced aligned transcription was accomplished through the SPeech Phonetisation Alignment

and Syllabification software (SPPAS, Bigi 2015). The outcome of the automatic annotation was manually corrected for the relevant boundaries, according to the criteria in Table 4.2 based on Machač & Skarnitzl (2009). Segmentation boundaries not used in the analyses have not been checked to speed up processing. The releases of C1 and C2 were detected automatically by means of a Praat scripting implementation of the algorithm described in Ananthapadmanabha et al. (2014), and subsequently corrected if necessary. The identification of the stop release was not possible in 99 tokens (8%) of C1 and 265 tokens (22%) of C2 out of 1212. This was due either to the absence of a clear burst in the waveform and spectrogram, or the realisation of voiced stops as voiced fricatives. Most of the fricativised tokens come from three speakers of Central Italian, IT12, IT13, and IT14, a variety of Italian known to show processes of lenition (Hualde & Nadeu 2011).

Moreover IT12 and IT14 produced several tokens of voiceless stops with voicing during closure (in some cases the closure was completely voiced). These tokens have been used in the analyses (as voiceless tokens), because (1) the actual presence or absence of voicing during closure does not bear on the compensatory account discussed here (which concerns supraglottal gestures) and laryngeal gestures can be implemented almost entirely independently from oral gestures, and (2) the voicing effect has been shown to exist even in whispered speech, where vocal fold vibration is entirely absent (Sharf 1964).[7]

The durations in milliseconds of the following intervals were extracted with a series of custom Praat scripts from the annotated acoustic landmarks: word duration, vowel duration (V1 onset to V1 offset), consonant closure duration (V1 offset to C2 release), and release-to-release duration (C1 release to C2 release). Sentence duration was measured in seconds. Figure 4.1 shows an example of the segmentation of /pata/ (a) and /pada/ (b) from an Italian speaker. Syllable rate (syllables per second) was used as a proxy to speech rate (Plug & Smith 2018), and was calculated as the number of syl-

[7] A reviewer makes interesting phonological remarks. The presence of lenition and voicing of voiceless stops in some varieties of Italian and its absence in Polish could be related to differences in laryngeal phonology and prosodic structure between these languages, namely the absence of a feature [voice] in Italian and the absence of true trochees in Polish. This hypothesis is compatible with work by Schwartz & Arndt (2018) and Schwartz (2016), to which the reader is referred.

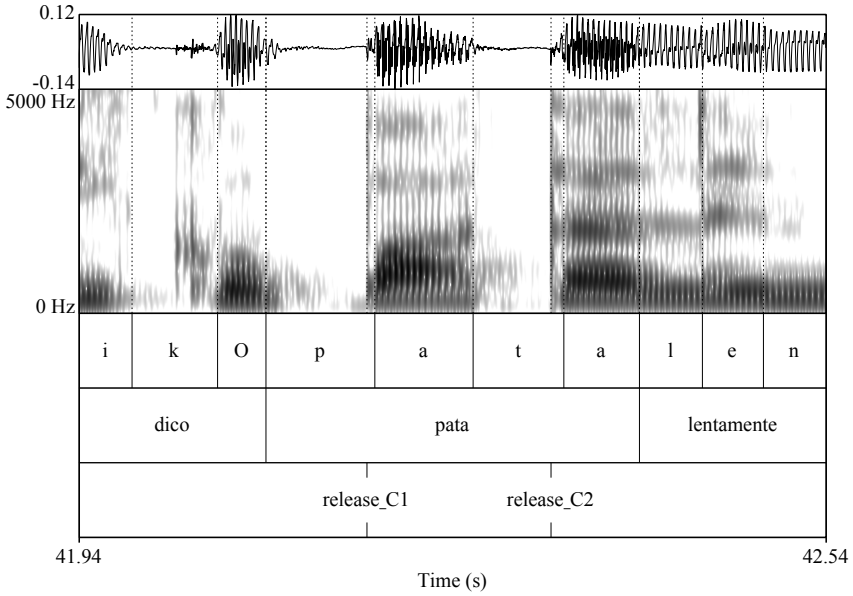
lables divided by the duration of the sentence in seconds (8 syllables in Italian, 6 in Polish). All further data processing and visualisation was done in R v3.5.2 (R Core Team 2018; Wickham 2017).

4.2.6 Statistical analysis

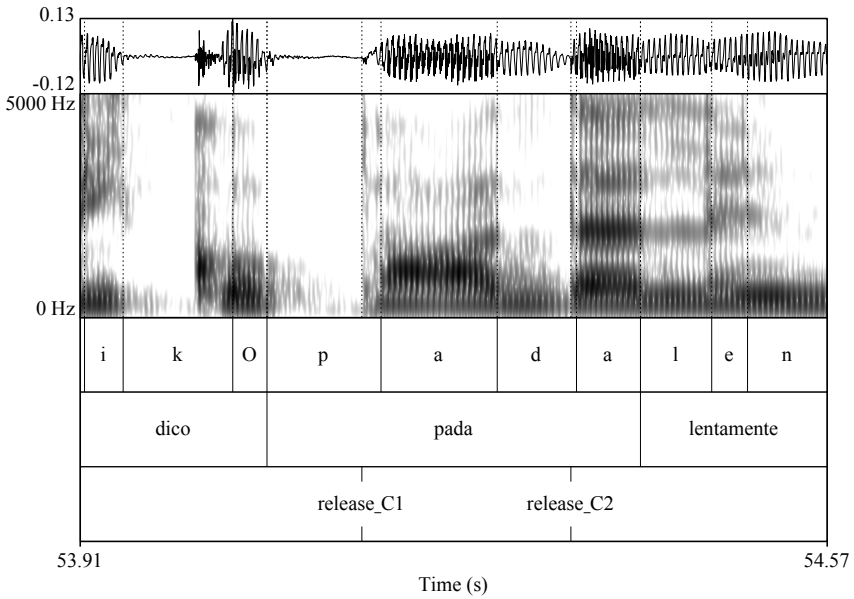
Given the data-driven nature of the study, all statistical analyses reported here are to be considered exploratory (hypothesis-generating) rather than confirmatory (hypothesis-driven, Kerr 1998; Gelman & Loken 2013; Roettger 2019). The durational measurements were analysed with linear mixed-effects models using lme4 v1.1-19 in R (Bates et al. 2015), and model estimates were extracted with the effects package v4.1-0 (Fox 2003). All factors were coded with treatment contrasts and the following reference levels: voiceless (vs voiced), /a/ (vs /o/, /u/), coronal (vs velar), Italian (vs Polish). Speech rate has been centred when included in the models to make the intercept estimates more interpretable. The models were fitted by Restricted Maximum Likelihood estimation (REML). The estimates in the results section refer to these reference levels unless interactions are discussed. *P*-values for the individual terms were obtained with lmerTest v3.0-1, which uses the Satterthwaite's approximation to degrees of freedom (Kuznetsova et al. 2017; Luke 2017). A result is considered significant if the *p*-value is below the alpha level ($\alpha = 0.05$). The choice of not using likelihood ratio tests for statistical inference is based on Luke (2017) who argues that this approach can lead to inflated Type I error rates. In any case, Luke (2017:1501) also warns that 'results [from mixed-effects models] should be interpreted with caution, regardless of the method adopted for obtaining *p*-values'. Inspection of residual plots and QQ plots of the models described below indicated absence of patterns in the residuals.

Bayes factors were used to test whether word and release-to-release duration are not affected by C2 voicing (i.e., the effect of C2 voicing on duration is 0).[8] For each set of

[8]The choice of Bayes factors over other information criteria, like AIC, is a practical one. First, Bayes factors can be used to identify the relative strength of the evidence for each hypothesis. The higher the Bayes factor of H_{01}, the stronger the evidence for H_0 according to the data. Second, a Bayes factor near 1 indicates that the data is compatible with both hypotheses (even when AIC indicates a preference of one over the other), in which case it is not possible to chose among them. Note that the AICs of the



(a) /pata/



(b) /pada/

Figure 4.1: Segmentation example of the words *pata* and *pada* uttered by the Italian speaker IT09 (the times on the *x*-axis refer to the times in the concatenated audio file).

null/alternative hypotheses, a full model (with the predictor of interest) and a null model (excluding it) were fitted separately using the Maximum Likelihood estimation (ML, Bates et al. 2015:34). The Bayes Information Criterion (BIC) approximation was then used to obtain Bayes factors (Raftery 1995, 1999; Wagenmakers 2007; Jarosz & Wiley 2014). The approximation is calculated according to the equation in 4.1 (Wagenmakers 2007:796).

$$BF_{01} \approx exp(\Delta BIC_{10}/2) \tag{4.1}$$

where $\Delta BIC_{10} = BIC_1 - BIC_0$, BIC_1 is the BIC of the full model, and BIC_0 is the BIC of the null model. Values of $BF_{01} > 1$ indicate a preference of H_0 over H_1. The interpretation of the Bayes factors follows the recommendations in Raftery (1995:p. 139): 1–3 = weak evidence, 3–20 = positive evidence, 20–150 = strong evidence, > 150 = very strong evidence.

The extracted measurements were filtered before statistical analysis. Measures of vowel duration, closure duration, word duration, and release-to-release duration that are 3 standard deviations lower or higher than the respective means were excluded from the final dataset (this procedure generally corresponds to a loss of around 2.5% of the data). One sentence (sentence 48 of IT07, *Dico pada lentamente*) included a speech error and has been excluded. After excluding missing measurements, these operations yield a total of 920 tokens of vowel and closure durations, 1176 tokens of word duration, and 848 tokens of release-to-release duration.

4.2.7 Open Science statement

Following recommendations for Open Science in Crüwell et al. (2018) and Berez-Kroeker et al. (2018), the data and code used to produce the analyses discussed in this paper are available on the Open Science Framework at `https://osf.io/quy7k/`.

word duration and release-to-release duration models reported below are lower when C2 voicing is not included as a predictor than when it is included, although the difference in AIC between the null and full models is very small (below 2).

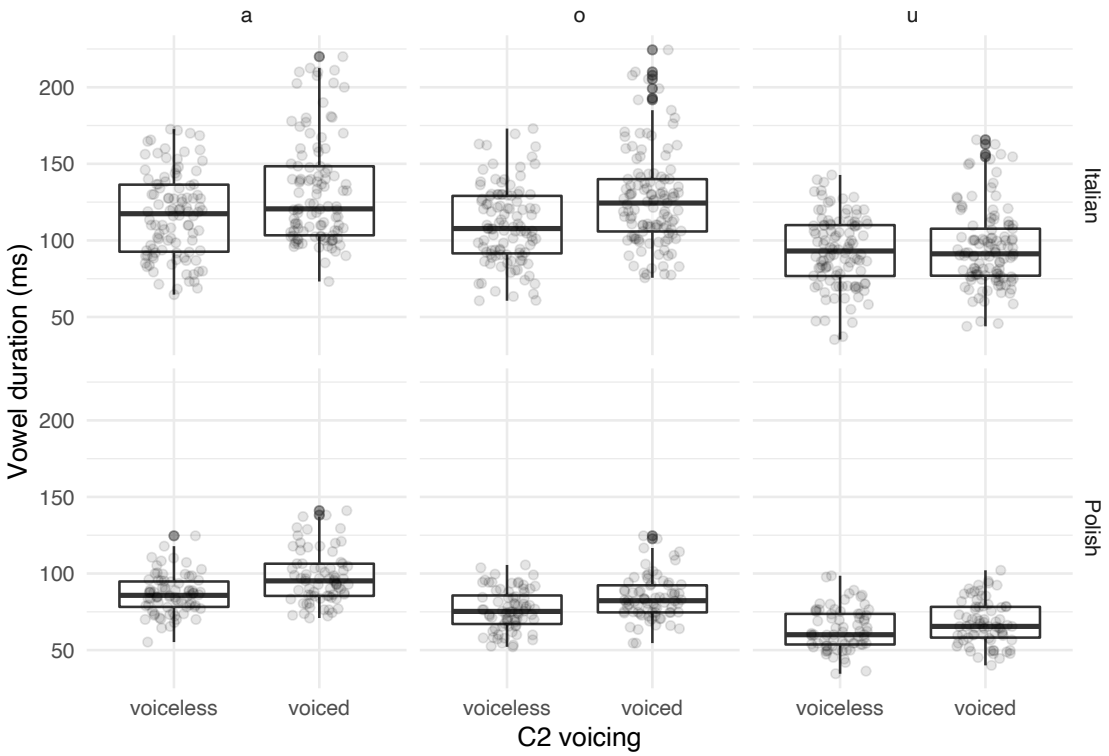


Figure 4.2: Raw data and boxplots of the duration in milliseconds of vowels in Italian (top row) and Polish (bottom row), for the vowels /a, o, u/ when followed by a voiceless or voiced stop.

4.3 Results

The following sections report the results of the study in relation to the durations of vowels, consonant closure, word, and the release-to-release interval. When discussing the output of statistical modelling, only the relevant predictors and interactions will be presented.

4.3.1 Vowel duration

Figure 4.2 shows boxplots and raw data of vowel duration for the three vowels /a, o, u/ when followed by voiceless or voiced stops in Italian and Polish. Vowels tend to be longer when followed by a voiced stop in both languages. The effect appears to be greater in Italian than in Polish, especially for the vowels /a/ and /o/. There is no evident effect of C2 voicing in /u/ in Italian, but the effect is discernible in Polish /u/. In Italian,

vowels have a mean duration of 106.16 ms (SD = 27.08) before voiceless stops, and a mean duration of 117.66 ms (SD = 34.63) before voiced stops. Polish vowels are on average 75.57 ms long (SD = 16.16) when followed by a voiceless stop, and 83.11 ms long (SD = 19.37) if a voiced stop follows. The difference in vowel duration based on the raw means is 11.5 ms in Italian and 7.54 ms in Polish.

A linear mixed-effects model with vowel duration as the outcome variable was fitted with the following predictors: fixed effects for C2 voicing (voiceless, voiced), C2 place of articulation (coronal, velar), vowel (a, o, u), language (Italian, Polish), and speech rate (as syllables per second, centred); by-speaker and by-word random intercepts with by-speaker random slopes for C2 voicing. All possible interactions between C2 voicing, vowel, and language were included. The following terms are significant according to *t*-tests with Satterthwaite's approximation to degrees of freedom (Table 4.3): C2 voicing, C2 place, vowel, language, and speech rate. Only the interaction between C2 voicing and vowel is significant. Vowels are 16.28 ms longer (SE = 4.42) when followed by a voiced stop (C2 voicing), and 8 ms shorter (SE = 1.63) when followed by a velar stop. The effect of C2 voicing is smaller with /u/ (around 3 ms, $\hat{\beta}$ = -13.1 ms, SE = 5.56). Polish has on average shorter vowels than Italian ($\hat{\beta}$ = -24.05 ms, SE = 7.83), and the effect of voicing is estimated to be about 10.55 ms, although note that the interaction between language and C2 voicing is not significant. Speech rate has a negative effect on vowel duration, such that faster rates correlate with shorter vowel durations ($\hat{\beta}$ = -16.23 ms, SE = 1.26).

4.3.2 Consonant closure duration

Figure 4.3 illustrates stop closure durations with boxplots and individual raw data points. A pattern opposite to that with vowel duration can be noticed: closure duration is shorter for voiced than for voiceless stops. The closure of voiced stops in Italian is 106.16 ms long (SD = 27.08), while the voiceless stops have a mean closure duration of 117.66 ms (SD = 34.63). In Polish, the closure duration is 75.57 ms (SD = 16.16) in voiced stops and 83.11 ms (SD = 19.37) in voiceless stops. The difference in closure duration based on the raw means is 13.33 ms in Italian and 10.87 ms in Polish. The same model specification as with vowel duration has been fitted with consonant

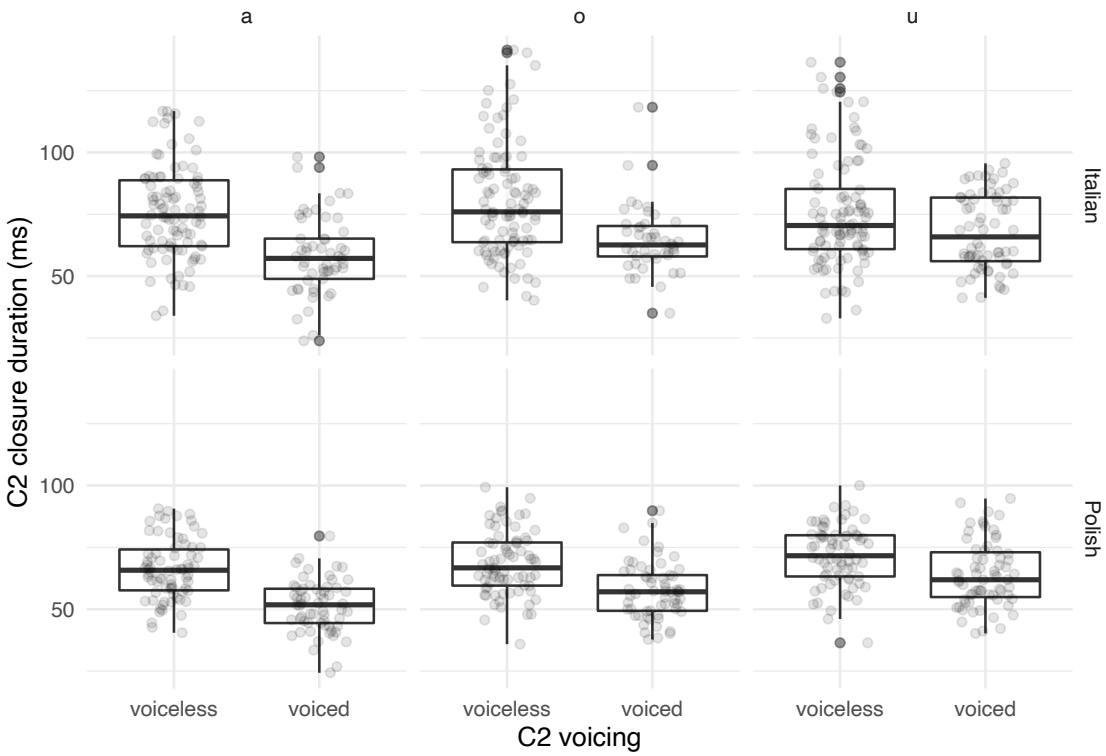


Figure 4.3: Raw data and boxplots of closure duration in milliseconds of voiceless and voiced stops in Italian (top row) and Polish (bottom row) when preceded by the vowels /a, o, u/.

Table 4.3: Summary of the linear mixed-effects model fitted to vowel duration (see Section 4.3.1).

Predictor	Estimate	SE	CI low	CI up	df	t-value	p-value	< α
Intercept	118.06	4.94	108.38	127.74	23.89	23.91	0.00	*
Voicing = voiced	16.28	4.42	7.62	24.95	15.39	3.68	0.00	*
Vowel = /o/	-7.50	3.93	-15.21	0.21	10.31	-1.91	0.08	
Vowel = /u/	-25.71	3.94	-33.44	-17.98	10.43	-6.52	0.00	*
Lang = Polish	-24.05	7.83	-39.40	-8.70	22.38	-3.07	0.01	*
Place = velar	-7.95	1.63	-11.15	-4.75	10.99	-4.87	0.00	*
Speech rate	-16.23	1.26	-18.70	-13.77	854.64	-12.92	0.00	*
Voiced × /o/	2.09	5.54	-8.77	12.96	10.18	0.38	0.71	
Voiced × /u/	-13.09	5.56	-23.99	-2.20	10.30	-2.36	0.04	*
Voiced × Polish	-5.73	6.61	-18.69	7.23	18.00	-0.87	0.40	
/o/ × Polish	-2.50	5.66	-13.60	8.60	11.09	-0.44	0.67	
/u/ × Polish	1.12	5.68	-10.01	12.26	11.23	0.20	0.85	
Voiced × /o/ × Polish	-6.16	8.00	-21.85	9.53	11.06	-0.77	0.46	
Voiced × /u/ × Polish	6.40	8.03	-9.34	22.13	11.19	0.80	0.44	

closure duration as the outcome variable. C2 voicing, C2 place, and speech rate are significant (Table 4.4). Stop closure is 17.5 ms shorter (SE = 4) if the stop is voiced and 3.5 ms longer (SE = 1.5) if velar. Finally, faster speech rates correlate with shorter closure durations ($\hat{\beta}$ = -8.5 ms, SE = 1 ms).

4.3.3 Vowel and closure duration

A model addressing the relationship between vowel and stop closure duration was fitted with the following terms and interactions: vowel duration as the outcome variable; as fixed effects, closure duration, vowel, speech rate (centred); all logical interactions between closure duration, vowel, and speech rate; by-speaker and by-word random intercepts (Table 4.5). Closure duration has a significant effect on vowel duration ($\hat{\beta}$ = -0.19 ms, SE = 0.06 ms). The effect with /u/ is greater than with /a/ and /o/ ($\hat{\beta}$ = -0.23 ms, SE = 0.08 ms). In general, closure duration is inversely proportional to vowel duration. However, such a correlation is quite weak, as shown by the small estimates. A 1 ms increase in closure duration corresponds to a 0.2–0.45 ms decrease in vowel duration. These estimates can be interpreted in terms of percentages of compensation, which

Table 4.4: Summary of the linear mixed-effects model fitted to closure duration (see Section 4.3.2).

Predictor	Estimate	SE	CI low	CI up	df	t-value	p-value	< α
Intercept	73.25	4.28	64.86	81.63	22.38	17.11	0.00	*
Voicing = voiced	-17.70	4.06	-25.66	-9.74	18.63	-4.36	0.00	*
Vowel = /o/	3.75	3.26	-2.64	10.13	9.43	1.15	0.28	
Vowel = /u/	-1.91	3.27	-8.32	4.49	9.56	-0.58	0.57	
Lang = Polish	-7.03	6.82	-20.40	6.34	20.82	-1.03	0.31	
Place = velar	3.80	1.38	1.09	6.51	10.94	2.74	0.02	*
Speech rate	-7.86	1.13	-10.08	-5.64	488.51	-6.94	0.00	*
Voiced × /o/	1.91	4.88	-7.65	11.47	11.80	0.39	0.70	
Voiced × /u/	10.88	4.79	1.50	20.27	10.97	2.27	0.04	*
Voiced × Polish	2.30	6.07	-9.59	14.19	19.83	0.38	0.71	
/o/ × Polish	-1.04	4.67	-10.19	8.10	9.94	-0.22	0.83	
/u/ × Polish	6.94	4.68	-2.24	16.12	10.09	1.48	0.17	
Voiced × /o/ × Polish	1.36	6.84	-12.04	14.77	11.44	0.20	0.85	
Voiced × /u/ × Polish	-3.08	6.77	-16.35	10.20	11.01	-0.45	0.66	

range between 20 and 45%. Note, moreover, that the negative correlation found here could be a consequence of annotation bias, since the vowel and closure share a boundary. Faster speech rates elicit a bigger effect than slower speech rates, as indicated by the significant interaction between closure duration and speech rate ($\hat{\beta}$ = -0.2 ms, SE = 0.06 ms). The effect of the interaction is reduced when the vowel is /u/ ($\hat{\beta}$ = 0.17 ms, SE = 0.08 ms). Figure 4.4 shows for each vowel /a, o, u/ the individual data points and the regression lines with 95% confidence intervals extracted from the mixed-effects model.

4.3.4 Word duration

Words with a voiceless C2 are on average 393.72 ms long (SD = 79.05) in Italian and 387.72 ms long (SD = 73.45) in Polish. Words with a voiced stop have a mean duration of 357.07 ms (SD = 39.14) in Italian and 361.87 ms (SD = 38.51) in Polish. The following full and null models were fitted to test the effect of C2 voicing on word duration. The full model is made up of the following fixed effects: C2 voicing, C2 place, vowel, language, and speech rate. The model also includes by-speaker and by-word random

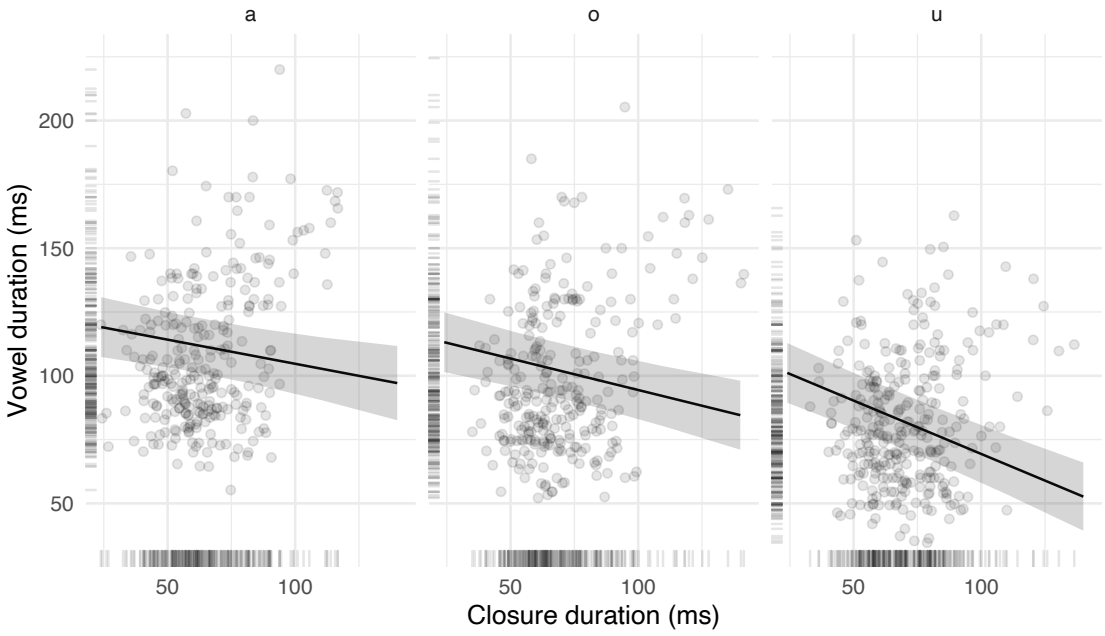


Figure 4.4: Raw data, estimated regression lines, and 95 per cent confidence intervals of the effect of closure duration on vowel duration for the vowels /a, o, u/ (from a mixed-effects model fitted to data pooled from Italian and Polish, see text for details).

Table 4.5: Summary of the linear mixed-effects model for testing the correlation between vowel and closure duration (see Section 4.3.3).

Predictor	Estimate	SE	CI low	CI up	df	t-value	p-value	< α
Intercept	123.62	6.76	110.36	136.87	56.24	18.27	0.00	*
Closure dur.	-0.19	0.06	-0.32	-0.06	816.53	-2.93	0.00	*
Vowel = /o/	-4.54	6.31	-16.90	7.82	127.47	-0.72	0.47	
Vowel = /u/	-12.47	6.40	-25.00	0.07	134.65	-1.95	0.05	
Speech rate	-5.16	4.28	-13.55	3.23	827.04	-1.21	0.23	
Closure × /o/	-0.06	0.08	-0.22	0.10	829.38	-0.71	0.48	
Closure × /u/	-0.23	0.08	-0.39	-0.07	831.49	-2.82	0.00	*
C2 closure × sp. rate	-0.20	0.06	-0.32	-0.08	826.97	-3.18	0.00	*
/o/ × sp. rate	-3.75	5.19	-13.92	6.42	819.79	-0.72	0.47	
/u/ × sp. rate	-10.13	5.50	-20.91	0.64	822.55	-1.84	0.07	
Closure × /o/ × sp. rate	0.09	0.07	-0.06	0.23	820.74	1.17	0.24	
Closure × /u/ × sp. rate	0.17	0.08	0.01	0.32	823.88	2.14	0.03	*

intercepts, and a by-speaker random slope for C2 voicing. The null model is the same as the full model with the exclusion of the fixed effect of C2 voicing. The Bayes factor

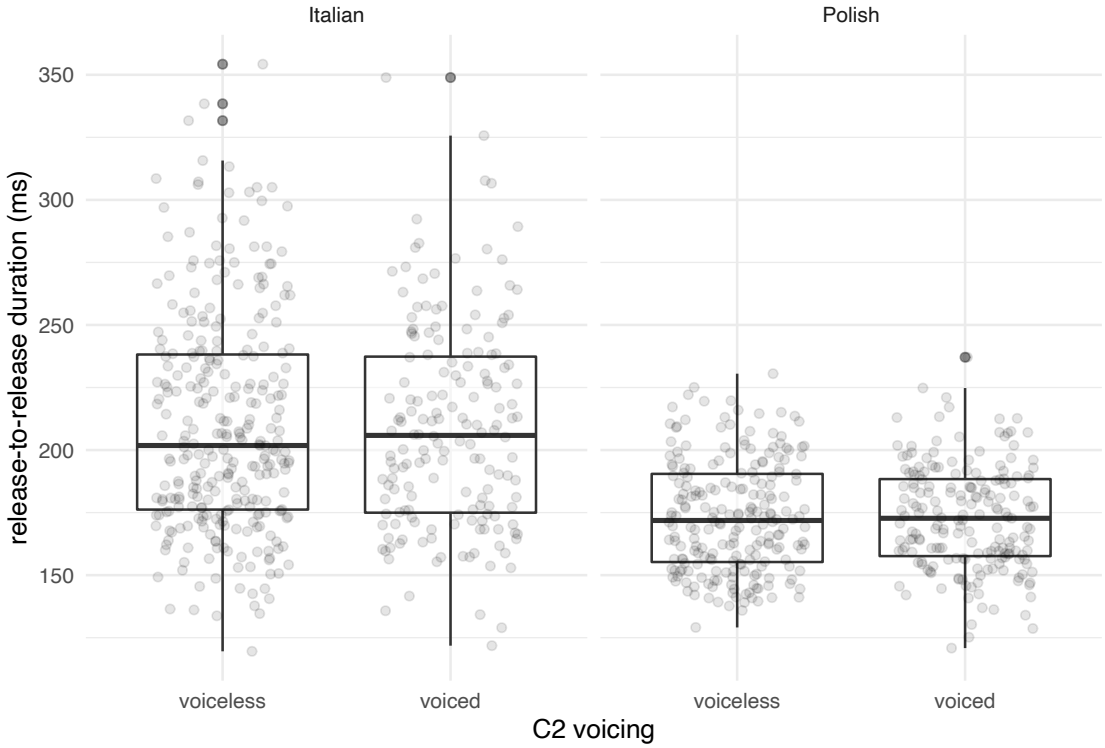


Figure 4.5: Raw data and boxplots of the duration in milliseconds of the release-to-release interval in Italian (left) and Polish (right) when C2 is voiceless or voiced.

of the null against the full model is 19. Thus, the null model (in which there is no effect of C2 voicing, $\beta = 0$) is 19 times more likely under the observed data than the full model. This indicates that there is positive evidence for a null effect of C2 voicing on word duration.

4.3.5 Release-to-release interval duration

In Figure 4.5, boxplots and raw data points show the duration of the release-to-release interval in words with a voiceless vs a voiced C2 stop, in Italian and Polish. It can be seen that the distributions, medians, and quartiles of the durations in the voiceless and voiced condition do not differ much in either language. In Italian, the mean duration of the release-to-release interval is 209.88 ms (SD = 43.84) if C2 is voiceless, and 208.6 ms (SD = 41.34) if voiced. In Polish, the mean durations are respectively 173.13 (SD = 22.44) and 172.67 (SD = 20.47) ms. The specifications of the null and full models for

the release-to-release duration are the same as for word duration. The Bayes factor of the null model against the full model is 21, which means that the null model (without C2 voicing) is 21 times more likely than the model with C2 voicing as a predictor. The Bayes factor suggests there is strong evidence that duration of the release-to-release interval is not affected by C2 voicing.

4.4 Discussion

A study of articulatory and acoustic aspects of the effect of consonant voicing on vowel duration in Italian and Polish has been carried out to look for a possible source of such an effect in speech production. Only the results from the acoustic part of the study bear on the main argument of this paper. The following sections discuss, in turn, the results regarding the effect of voicing on vowel duration in Italian and Polish and how the finding that the duration of the interval between the two consecutive consonant releases in CV́CV words is compatible with a compensatory temporal adjustment account of the voicing effect. The section concludes by discussing the limitations and open issues of this study.

4.4.1 Voicing effect in Italian and Polish

The results of vowel duration and C2 voicing indicate that vowels are longer when followed by voiced than when followed by voiceless stops both in Italian and Polish. The estimated effect is around 16 ms when C2 is voiced for Italian. This value is not too far from the estimates of previous works on this language (Magno Caldognetto et al. 1979; Farnetani & Kori 1986; Esposito 2002), the range of which is between 22 and 24 ms. The higher estimates of these studies compared to the one here could be related to differences in experimental design, or Type M (magnitude) errors due to low statistical power (see Kirby & Sonderegger 2018). The estimate of the effect of voicing on C2 closure duration is around -18 ms. Crucially, the effect of voicing on vowel and closure duration have very similar magnitudes and opposite signs. These results suggest a compensatory mechanism between vowel and closure duration.

Furthermore, the effect of voicing on the duration of Italian /u/ is smaller than with

/a/ and /o/ (about 3 vs 16 ms respectively), a fact already observed by Ferrero et al. 1978. While it is not clear why the duration of this particular vowel should not be affected by C2 voicing, the data reported here indicate that the magnitude of the difference in closure duration when the preceding vowel is /u/ is smaller than with /a/ and /o/ (about 7 vs 17 ms respectively). If vowel duration compensates for closure duration, then a smaller difference in closure duration should correspond to a small difference in vowel duration, as the estimates seem to suggest.

The interpretation of the Polish results is less straightforward. Previous studies found either no voicing effect or a small effect in Polish (3.5–4.5 ms). In particular, Malisz & Klessa (2008) say that the effect seems to be very idiosyncratic in the 40 speakers of their analysis. The estimated effect found in the 6 Polish speakers of the present study is about 10.5 ms, and the difference based on the means of the raw vowel durations is 7.5 ms. Recall, however, that the interaction between language and C2 voicing (which gives the estimate of 10.54) is not significant (see the full model summary in Table 4.3). It is likely, though, that the non-significance might be related to low power. Indeed, the raw mean difference of 7.5 ms in Polish—although still higher than what found in previous studies—might be more informative.

More specifically, when one compares the raw mean duration differences of vowels with the raw mean duration differences of consonant closures, a pattern can be seen. The mean differences of Italian vowels and closures (11.5 and 13.33, respectively) are bigger than those of Polish (7.54 and 10.87), even if by just a small amount. It is plausible that the smaller effect of C2 voicing on preceding vowel duration in Polish is related to the smaller effect on closure duration, if we assume a temporal mechanism of compensation between the closure and the vowel. These patterns will need to be confirmed with a more balanced sample of Italian and Polish speakers.

On the other hand, while the estimated differences in vowel durations can be interpreted in reference to Italian and Polish as two independent linguistic objects, the patterns observed in the individual speakers does not indicate a systematic relation between magnitude of the effect and language. Figure 4.6 shows the random coefficients of the effect of C2 voicing on vowel duration for the individual speakers, extracted from the mixed-effects model presented in Section 4.3.1. Black indicates Italian speak-

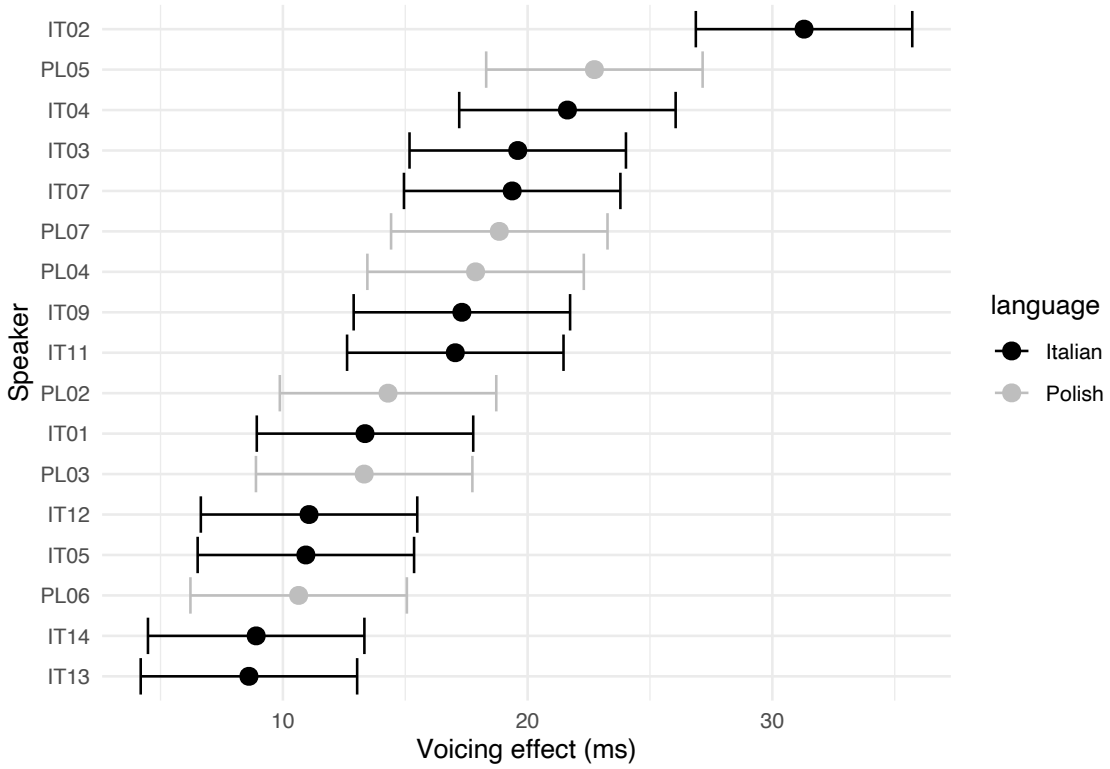


Figure 4.6: By-speaker random coefficients and error bars for the effect of C2 voicing on vowel duration, extracted from a mixed-effect model (Section 3.1).

ers, while grey is for Polish speakers. As can be seen, speakers of both languages are scattered along the values of the voicing effect. These results are in agreement with the idiosyncrasy of the voicing effect of Polish found in Malisz & Klessa (2008). While large-scale studies could reveal clear language-level patterns, the data discussed here point to a scenario in which the speaker's individual behaviour is substantial. Future studies could thus look into the respective role of individual-level and community-level factors and how these contribute to the magnitude of the durational differences across speakers and languages.

4.4.2 Compensatory temporal adjustment

Vowels followed by voiced stops are long, while vowels followed by voiceless stops are short. The closure duration of voiced stops is short compared to that of voiceless stops. There seems to be an inverse relation between vowel duration and closure duration, by which a long vowel entails a short closure (and vice versa), and a short vowel entails a long closure (and vice versa).

The data and statistical analyses of this study suggest that the duration of the interval between the releases of two consecutive consonants in CV́CV words (the release-to-release interval) is not affected by the phonological voicing of the second consonant (C2) in Italian and Polish. In accordance with a compensatory temporal adjustment account (Slis & Cohen 1969a; Lehiste 1970a), the difference in vowel duration and closure durations before voiceless vs voiced stops can be seen as the outcome of differences in timing of the vowel offset/closure onset (more neutrally, the VC boundary). In other words, the timing of the VC boundary within the temporally stable release-to-release interval determines the duration of both the vowel and the stop closure. An earlier VC boundary relative to the onset of the preceding vowel results in a shorter vowel and a longer stop closure. On the other hand, a later VC boundary produces a longer vowel and a shorter closure. Figure 4.7 illustrates this compensatory mechanism. Note that the term "temporal stability" (and "temporally stable") as used here means that the underlying statistical distribution of the interval duration is stable *across contexts of C2 voicing*. No specific statement is implied about the variance of the duration around the mean, across or within phonological contexts.

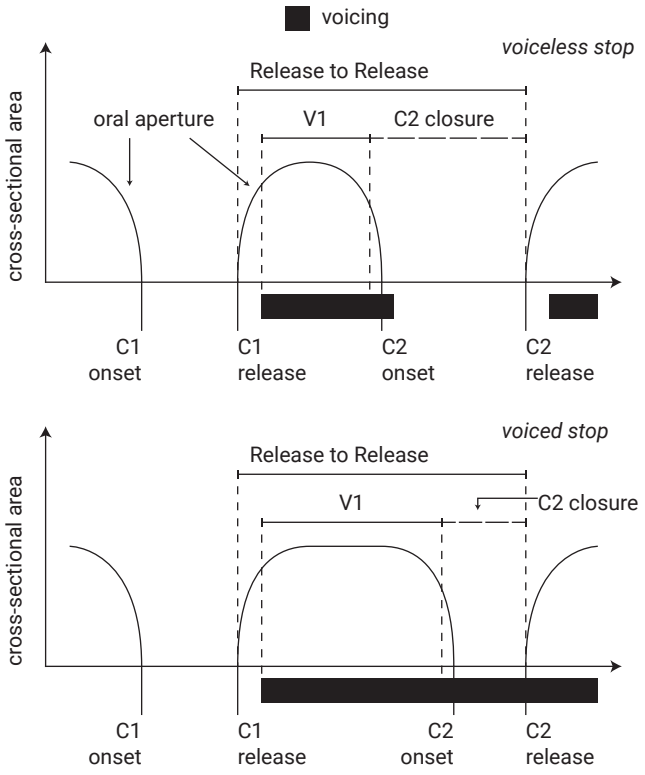


Figure 4.7: A schematic representation of the oral cavity cross-sectional area, as inferred from acoustics. Design based on Esposito (2002). The top panel shows a CV́C sequence with a voiceless C2, the bottom panel with a voiced C2. Oral cavity aperture (on the *y*-axis, as the inverse of oral constriction) through time (on the *x*-axis) is represented by the black line. Lower values represent a more constricted oral tract (a contoid configuration), while higher values indicate a more open oral tract (a vocoid configuration). The black bars below the time axis represent voicing (vocal fold vibration). Various landmarks and intervals are indicated in the schematic.

The invariance of the release-to-release interval allows us to refine the logistics of the compensatory account by narrowing the scope of the temporal adjustment action. A limitation of this account, as proposed by Slis & Cohen (1969a) and Lehiste (1970a), is the lack of a precise identification of the word-internal mechanics of compensation. As already discussed in Section 4.1, it is not clear why the adjustment should target the preceding stressed vowel, rather than the following unstressed vowel or any other segment in the word. Since the release-to-release interval includes just the vocoid gesture between the release of C1 and the VC boundary, and the consonant closure, it follows that differences in the timing of the VC boundary must be reflected in differences in both vowel and closure durations.

Under an account of temporal compensation, the voicing effect can be interpreted as a by-product of gestural phasing and mechanisms operating on the timing of the VC boundary. The temporal stability of the release-to-release interval across voicing contexts allows us to refine the compensatory mechanism by providing a temporal anchor. On the other hand, it is important to note that the release-to-release interval should not necessarily have a special status in such a compensatory account, but rather can be used as a proxy to the understanding of a full gestural mechanism of compensation. Indeed, the temporal stability of this interval should be derivable from a theory of gestural phasing, rather than one that simply states that the interval is stable across voicing contexts.

The non-exclusivity of the release-to-release interval is also shown by the fact that excluding the VOT from it still indicates that C2 voicing is not affecting the interval duration. The duration of the vowel onset to release interval (the release-to-release minus VOT) is stable across voicing contexts (Bayes factor = 9). However, the duration of release-to-release interval has relatively more cohesion than that of the vowel onset to release interval, as indicated by two measures of relative dispersion (the coefficient of variation CV and the coefficient of quartile variation CQV, see Bonett 2006).[9] On the other hand, the duration of the interval between the vowel onset of V1 to the vowel

[9]The CV of the release-to-release duration is 0.203, while that of the vowel onset to release duration is 0.232. The CQV is 0.127 for the release-to-release and 0.136 for the vowel onset to release. Lower values mean less dispersion/more cohesion.

onset of V2 does change depending on C2 voicing (the interval is around 20 ms longer if C2 is voiceless). This fact is simply a consequence of including the VOT of C2 in the measure. Voiceless stops have longer VOT values, which increases the duration of the interval. The difficulty in identifying a clear-cut time point corresponding to vowel onset could explain the relative higher dispersion of the vowel onset to release interval duration. For these reasons, the release-to-release interval is probably a better measure of temporal stability than the vowel onset to release, given its inherent higher cohesion.

It is possible that the temporal stability of the release-to-release interval is not an antecedent, but rather a consequence of manipulating vowel and closure durations. If this were the case, the differential duration of vowels and closures would not be the result of a compensatory mechanism. The present data cannot disambiguate between these two scenarios, and future studies should look into investigating independent reasons for the release-to-release interval stability across voicing contexts. The account of gestural phasing proposed by Tilsen (2013, 2016) is promising, in that the temporal stability of the release-to-release interval would directly follow from the relative phasing of the vowels in CVCV words (see Figure 6 in Tilsen 2013). Articulatory work on the gestural coordination of sequences besides the traditional syllable might reveal a principled organisation that results in the temporal patterns observed in this study and in other durational phenomena.

However, even if independent reasons for the interval stability can be identified, other mechanisms, unrelated to compensatory effects, would still be required to explain the differential timing of the VC boundary within that interval. Accounts compatible with other aspects of production and perception would not be ultimately ruled out, as thoroughly discussed in Beguš (2017). For example, the laryngeal adjustment hypothesis (Halle et al. 1967) states that adjustments of the glottis for obstruent voicing require more time to be implemented, so that stop closure onset (VC boundary) for a voiced stops will be achieved later than that of a voiceless stop, relative to the onset of the preceding vowel. Tongue root advancement (Rothenberg 1967; Westbury 1983; Ohala 2011) could also play a role in modulating the time required before closure can be implemented. Another account (Chen 1970) makes direct reference to velocity of the closing gesture, which is faster in voiceless than in voiced stops (Summers 1987;

de Jong 1991), so that the VC boundary within the release-to-release would be timed earlier in the former than in the latter case. Moreover, perceptual explanations of the voicing effect have been proposed in Javkin (1976) and Kluender et al. (1988), and these perceptual factors might play a role in the enhancement of the effect (see Kingston & Diehl 1994, Port & Dalby 1982, and Luce & Charles-Luce 1985, see Fowler 1992 for a critique to Kluender et al. 1988). Finally, whether the timing of the VC boundary depends on modulations of the vocalic or consonantal gesture, or both, is another aspect that should be investigated further (see de Jong 1991 for an example).

A comment is also due in relation to possible coexisting effects on vowel duration. Beguš (2017) finds that, even when C2 closure duration is controlled for, C2 phonation (ejective, voiceless, voiced) in Georgian is still a significant predictor. The author argues for a separate laryngeal features effect, which operates in addition to a closure duration effect. In the present study, C2 voicing (voiceless, voiced) and its interactions are not significant when included in the model discussed in section Section 4.3.3, which has vowel duration as outcome and C2 closure duration as one of the predictors.[10] However, even when multicollinearity between predictors is minimal, presence or lack of statistical significance of multiple terms cannot unequivocally inform us on the actual contribution of those terms, since it is possible that unknown relations between terms mask underlying mechanisms (for a discussion see McElreath 2015). The diachronic development of context-driven statistical sub-distributions can override the original causal link (Sóskuthy 2013). Under this scenario, it is not possible to discern which of the competing predictors is diachronically responsible for the relation, and either or both the compensatory mechanism and the laryngeal features could have had a role in generating the synchronic patterns (this kind of reasoning is compatible for example with exemplar theories of speech perception and production, see among others Johnson 1997; Sóskuthy et al. 2018; Ambridge 2018; Todd et al. 2019).

Since diverging results have been obtained in relation to the significance of C2 phonation in addition to C2 closure durations, these aspects need to be further investigated in future studies, although to ascertain whether they are artefacts of statistical procedures or if they reflect an underlying state of affairs might still prove difficult. To

[10]Multicollinearity is not an issue here, since the VIFs are all below 3 (Zuur et al. 2010).

conclude, lack of significance of a separate laryngeal features effect in this study cannot be taken as evidence for its absence in the present data, and a compensatory mechanism could coexist with mechanisms directly related to laryngeal features, which would in turn explain the differential timing of the VC boundary.

4.4.3 Limitations and future work

The generalisations put forward in this paper strictly apply to disyllabic words with a stressed vowel in the first syllable, flanked by single stops. First, it is possible that the pattern found in this context does not occur in sequences including an unstressed vowel. For example, it is known that the difference in closure duration between voiceless and voiced stops is not stable when the stops precede a stressed vowel, although vowels preceding pre-stress stops have slightly different durations (Davis & Summers 1989). According to the mechanism proposed here, the absence of differences in closure duration should correspond to the absence of differences in vowel duration. Second, it is known that the magnitude of the effect of voicing is modulated by other prosodic characteristics, like the number of syllables in the word, presence/absence of focus, and position within the sentence (Sharf 1962; Klatt 1973; Laeufer 1992; de Jong 2004). Third, the constraints on experimental material enforced by the use of ultrasound tongue imaging have been previously mentioned in Section 4.2.3. Given these constraints, temporal information from other vowels (like front vowels), places and manners of articulation is a desideratum. Data from different contexts and different languages is thus needed to assess the generality of the claims put forward in this paper.

Another issue is the interaction of the temporal compensation and speech rate. The magnitude of compensation between vowel and closure duration found in de Jong (1991) and here is somewhat small (between 12% and 40%). Ideally, given the temporal stability of the release-to-release interval relative to C2 voicing, the compensation rates should approximate 100%. However, it is possible that the correlation between vowel and closure duration is modulated in complex ways by the individual effects of speech rate on the vowel and the closure. For example, Ko (2018) finds that the vowel/closure ratio differs depending on speaking rate and that there is an interaction between the voicing of the consonant and speaking rate. When the consonant is voice-

less, the vowel/closure ratio is smaller when speaking rate is slow, while slow speaking rate induces larger vowel/closure values when the consonant is voiced. Experimental work is required which addresses the differential effect of speaking rate on vowel and consonant closures, and how these interact with a possible compensatory mechanism.

Some concern could be raised in relation to possible influences of English on the native productions of participants recorded in the English-speaking context of the University of Manchester Laboratory. However, as reported in Section 4.2.4, conversations during the session prior to the experiment and instructions were in the participant's native language. Antoniou et al. (2010) show that, in a situational language context study of Greek-English bilinguals, being exposed to the native language during the experiment elicited Greek native-like phonetic values even when the dominant language at the time of recording was English (the bilingual speakers acquired English as a second language, being Greek their first). A small effect of L2 could persist in proficient L2 speakers, as found by Schwartz et al. (2015). The five Polish speakers with a highly proficient level of English investigated in that study showed a 10 ms increase in VOT values compared to the quasi-monolingual base level. While previous studies focussed on VOT, future work should directly test the influence of English on the magnitude of the voicing effect of one's native language.

The compensatory temporal adjustment account presented here extends to other durational effects discussed in the literature. In particular, the account bears predictions on the direction of the durational difference led by phonation types different from voicing, like aspiration and ejection. For example, the mix of results with regard to the effect of aspiration (Durvasula & Luo 2012) suggests that the conditions for a temporal adjustment might differ across the contexts and languages studied. In light of the results in Beguš (2017), future studies will also have to investigate the durational invariance of speech intervals in relation to a variety of phonation contrasts.

4.5 Conclusions

The results of this exploratory study of the effect of voicing on vowel duration are congruent with a compensatory temporal adjustment account of such effect. Acoustic

data from seventeen speakers of Italian and Polish show that the temporal distance between two consecutive stop releases is not affected by the voicing of the second stop in CV́CV words. The temporal invariance of the release-to-release interval, together with a difference in timing of the VC boundary, can cause vowels to be shorter when followed by voiceless stops (which have a long closure) and longer when followed by voiced stops (the closure of which is short).

As discussed in Section 4.4.2, the temporal patterns reported here do not univocally exclude other possible sources for the duration differential. Multiple mechanisms (both articulatory and perceptual) could conspire together to produce the observed patterns. Such a pluralist view has already been proposed for the voicing effect (for example, Beguš 2017 and Sanker 2018), and for other related phenomena, like vowel duration in incomplete neutralisation (Winter & Roettger 2011). For a review of explanatory pluralism in the cognitive sciences, see Dale et al. 2009 and references therein. Indeed, a hybrid account, which takes into consideration and synthesises aspects of multiple proposed accounts, is probably warranted, given the diversity of compatible results obtained so far. Future work will need to investigate further aspects of the patterns found in this study, with a particular focus on the effects of different segmental and prosodic structures and different laryngeal contrasts on the release-to-release interval, and in relation to other attributes of consonant effects on vowel duration.

4.6 Socio-linguistic information of participants

See Table 4.6.

Table 4.6: Participants' sociolinguistic information. The column 'Spent most time in' gives the city in which the participant spent most of their life. The last column ('> 6 mo') indicates whether the participant has spent more than 6 months abroad.

ID	Age	Sex	Native L	Other Ls	City of birth	Spent most time in	> 6 mo
IT01	29	Male	Italian	English, Spanish	Verbania	Verbania	Yes
IT02	26	Male	Italian	Friulian, English, Ladin-Venetan	Udine	Tricesimo	Yes
IT03	28	Female	Italian	English, German	Verbania	Verbania	No
IT04	54	Female	Italian	Calabrese	Verbania	Verbania	No
IT05	28	Female	Italian	English	Verbania	Verbania	No
IT09	35	Female	Italian	English	Vignola	Vignola	Yes
IT11	24	Male	Italian	English	Monza	Monza	Yes
IT12	26	Male	Italian	English	Rome	Rome	Yes
IT13	20	Female	Italian	English, French, Arabic, Farsi	Ancona	Chiaravalle	Yes
IT14	32	Male	Italian	English, Spanish	Frosinone	Frosinone	Yes
PL02	32	Female	Polish	English, Norwegian, French, German, Dutch	Koło	Poznań	Yes
PL03	26	Male	Polish	Russian, English, French, German	Nowa Sol	Poznań	Yes
PL04	34	Female	Polish	Spanish, English, French	Warsaw	Warsaw	No
PL05	42	Male	Polish	English, French	Przasnysz	Warsaw	No
PL06	33	Male	Polish	English	Zgierz	Zgierz	Yes
PL07	32	Female	Polish	English, Russian	Bielsk Podlaski	Bielsk Podlaski	Yes

Chapter 5

Temporal (in)stability in English monosyllabic and disyllabic words: Insights on the effect of voicing on vowel duration [Paper II]

Coretta, Stefano. 2019. Temporal (in)stability in English monosyllabic and disyllabic words: Insights on the effect of voicing on vowel duration. Manuscript. DOI: https://doi.org/10.31219/osf.io/jvwa8.

Abstract

English is one in the wide range of languages in which the duration of vowels is modulated by the voicing of the following consonant: Vowels are shorter when followed by voiceless stops, and longer when followed by voiced stops. The so-called voicing effect has been attributed to a variety of mechanisms. Temporal compensation between the duration of the vowel and the following stop closure is one of these mechanisms. Based on acoustic data from Italian and Polish disyllabic words, the compensatory mechanism has been proposed to be a consequence of the temporal stability of the interval between the consonant releases flanking the vowel. The timing of the VC boundary within this interval determines the respective durations of the vowel and the stop closure. In this paper, it is shown that the duration of the release-to-release interval is not affected by

the voicing of the second consonant in English disyllabic words, but that it is in monosyllabic words. It is argued that the stability of the interval can be derived from the isochronous phasing of the vocalic gestures in the VCV sequence of disyllabic words. The absence of the temporal anchor of a second vowel in monosyllabic words, on the other hand, allows the vocalic and the consonant gesture durations to be modified independently. Other aspects of production and perception behind the voicing effect can coexist with a temporal compensation mechanism and cannot be excluded.

5.1 Introduction

A well-known cross-linguistic tendency is that vowels have shorter durations when followed by voiceless stops and longer durations when followed by voiced stops. This so-called "voicing effect" has been long documented in a wide range of languages across different linguistic families (Maddieson & Gandour 1976; Beguš 2017). Several hypotheses have been proposed as to the origin of this phenomenon, from articulatory mechanisms to perceptual biases; however, no one particular account has gained universal support.

One such hypothesis, the compensatory temporal adjustment account, states that the voicing effect involves a compensatory mechanism between vowel and consonant closure duration. Vowels are shorter when followed by voiceless stops because the latter have longer closure durations, and, vice versa, vowels are longer before voiced stops because the latter have shorter closure durations. However, the compensatory account fails to clearly identify a speech interval within which compensation is implemented. Both the syllable (Lindblom 1967; Farnetani & Kori 1986) and the word (Slis & Cohen 1969a,b; Lehiste 1970a,b) have been proposed as such intervals, but these have been subsequently criticised on empirical and logical grounds (Chen 1970; Jacewicz et al. 2009; Maddieson & Gandour 1976; Coretta 2019b).

In an exploratory study of acoustic durations in Italian and Polish trochaic CV́CV words, Coretta (2019b) finds that the duration of the interval between the two consonant releases is not affected by the voicing status of the second consonant. The duration of the release-to-release interval in words where the second consonant is voiceless (like

/pata/) is not significantly different from that in words where the second consonant is voiced (for example, /pada/). The temporal stability of the release-to-release interval is compatible with a compensatory temporal adjustment account of the voicing effect (Lindblom 1967; Slis & Cohen 1969a,b; Lehiste 1970a,b), and it offers a resolution to the drawbacks of previous versions of the account.

Given the temporal stability of the release-to-release interval, the timing of the vowel/consonant (VC) boundary (corresponding to the vowel offset and the consonant closure onset) within that interval will determine the respective durations of vowel and consonant closure. Since the VC boundary in voiceless stops is timed earlier within the release-to-release relative to voiced stops, the vowel is shorter and closures is longer when the post-vocalic stop is voiceless than when it is voiced. This interpretation agrees with known differences of closure durations in voiceless vs voiced stops (Lisker 1957; Summers 1987; Davis & Summers 1989; de Jong 1991), namely that voiceless closures are longer than voiced ones. Thus, a possible diachronic pathway to the voicing effect in disyllabic words is one in which vowel and closure duration differences emerge from changes in the timing of the VC boundary within the release-to-release interval which affect the voiceless and voiced contexts differently.

Note that the release-to-release interval in itself does not have a special status. The proposed account of compensatory temporal adjustment can be understood in relation to the acoustic duration of vowels, hence the scope of compensation can (but need not) be defined in terms of acoustic intervals. The interval found to be temporally stable across voicing contexts in disyllabic words is the release-to-release interval. However, it is desirable to derive the isochrony of this acoustic interval from properties of articulatory coordination. A tentative account of the underlying gestural coordination from which the release-to-release isochrony could be derived is offered here.

5.1.1 A gestural account of the voicing effect

The task-dynamic model (Saltzman et al. 2008) of Articulatory Phonology (Browman & Goldstein 1986, 1988, 1992) states that any two gestures can be implemented according to two modes. Either they are initiated in synchrony or they are implemented sequentially. These modes of gestural phasing (in-phase and anti-phase) can account

for a variety of patterns of articulatory timing. Relevant to our discussion is that onset consonants are generally produced in-phase with the following vowel, meaning that the vocalic and consonantal gestures are initiated together. This mechanism gives rise to the so-called C-centre effects observed with onsets, by which the acoustic duration of a vowel depends on the number of onset consonants (Browman & Goldstein 1988; Marin & Pouplier 2010; Hermes et al. 2013; Marin & Pouplier 2014).

According to Öhman (1966, 1967a), the speech stream is composed by a series of continuous vocalic gestures interrupted by gestures of oral constriction (consonants). Fowler (1983) further proposes that the vocalic gestures of a VCV sequence are characterised by a cyclic pattern of production, so that the temporal distance between the two vowels is constant, independent of the nature of the intervening consonant. However, the temporal distance of the V-to-V interval is modulated by the number of intervening consonants (Zmarich et al. 2011; Zeroual et al. 2015). Figure 5.1 (a) illustrates this point.

The schematics at the top of Figure 5.1 (a) shows an abstract representation (based on Marin & Pouplier 2010) of a word-like series of consonants and vowels, *pata* (nonce words are used as examples to enable the creation of a full set of minimal pairs, as needed). The *x*-axis represents time, while the *y*-axis can be interpreted as oral aperture for vowels and oral constriction for consonants. As per C-centre effects, the onset of the consonants /p/ and /t/ are aligned with the onset of the respective following vowels (the consonants are produced in-phase with the vowels). The bottom scheme of Figure 5.1 (a) represents the word *patta*, syllabified as /pat.ta/. As a representational device, the geminate stop is given as two separate consecutive gestures (the actual details of gestural implementation depend on one's chosen gestural account and are not directly relevant to the present discussion). In *patta*, the onset of /p/ and of the second /t/ are, as in *pata*, aligned with the onset of the respective following vowels. However, the first /t/ is instead produced anti-phase with the preceding vowel. As it can be seen by comparing the top and bottom scheme of Figure 5.1 (a), the temporal distance between the two vowels differs in *pata* and *patta*, as per the results in Zmarich et al. (2011) and Zeroual et al. (2015).

On the other hand, the distance of the vowels can still be expected to be stable

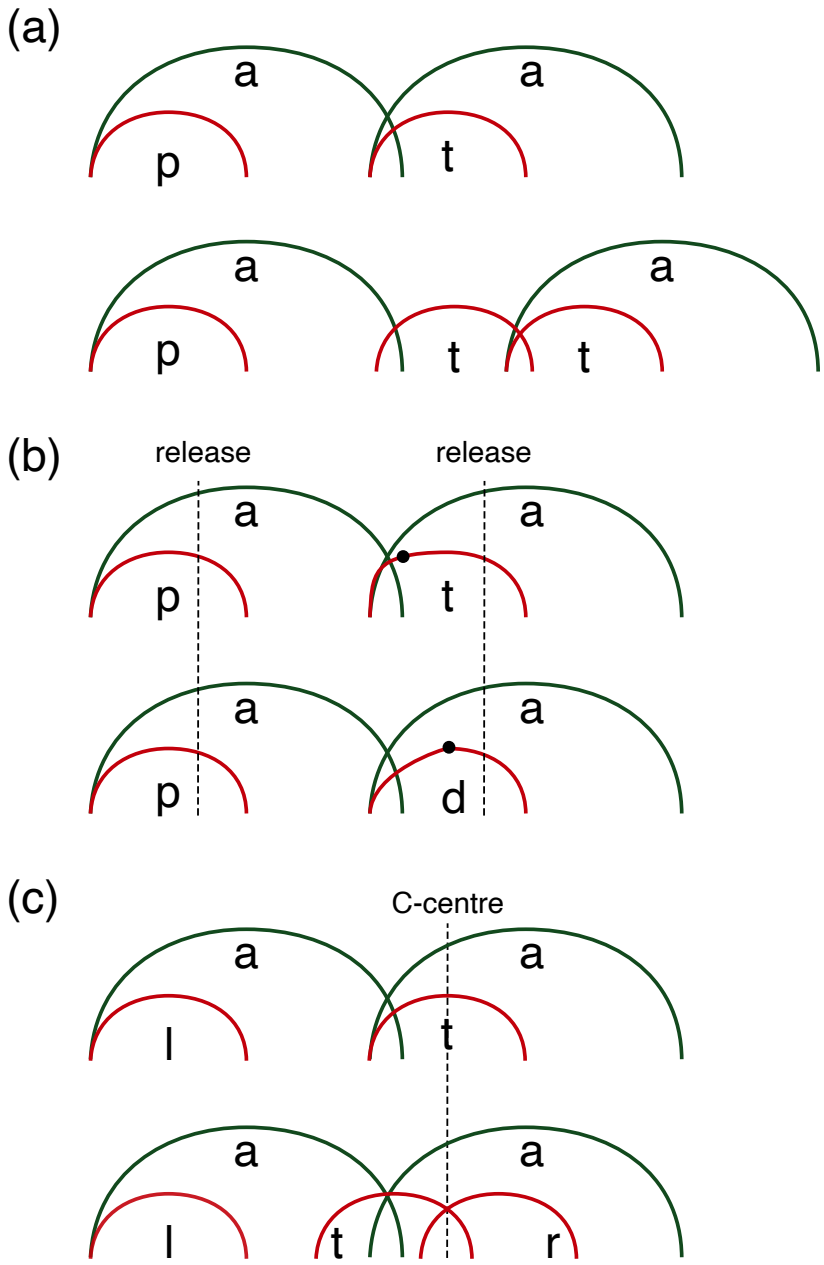


Figure 5.1: Schematics of the gestural phasing of vocalic and consonantal gestures in different contexts. The *x**-axis is time, while the *y*-axis can be interpreted as oral aperture for vowels and oral constriction for consonants. (a) shows singleton vs geminate stops, (b) voiceless and voiced stops, and (c) singleton vs tautosyllabic cluster. Note that in (a) the distance between the vowels increases in the geminate context, while it is stable in (b) and (c). The dots in (b) on the consonant gesture lines indicate the time of closure onset.

when there is a single intervocalic consonant that alternates in voicing. This is shown in Figure 5.1 (b). In *pata* and *pada*, the consonants are in-phase with the respective following vowels, as we have seen in Figure 5.1 (a). Summers (1987) and de Jong (1991) show that the closing gesture of voiceless stops has greater velocity than that of voiced stops. Assuming that the closing gesture of both voiceless and voiced stops is initiated in synchrony with that of the following vowel as per the in-phase alignment, full oral closure will be achieved earlier in voiceless than in voiced stops relative to the beginning of the preceding vocalic gesture. Under this scenario, the temporal distance of the vowels does not differ in *pata* vs *pada*. The result is, everything else being equal, a shorter vowel and a longer (full) closure in voiceless stops, and a longer vowel and a shorter closure in voiced stops. Warren & Hay (2006) offer evidence from lip data, where the jaw and lip closing gesture duration accounts for about 80% of the vowel duration difference.

Moreover, the combined action of the isochrony of the vowel-to-vowel interval and the in-phase alignment of the onset consonant would also be responsible for the isochrony of the release-to-release interval in CVCV words. The first consonant and vowel are produced in-phase with each other, and these are sequentially followed by the second consonant and vowel, again produced in-phase with each other. Then, the differential duration found in the voicing effect would be a consequence of the different velocity of the closing gesture in voiceless vs voiced consonants. Assuming that the closing gesture of both voiceless and voiced stops is initiated in synchrony with that of the following vowel (as per the in-phase alignment), full oral closure is achieved earlier in voiceless stops relative to its timing in voiced stops.

Finally, further evidence for a vowel-based rhythmic gestural implementation of speech comes from work by Farnetani & Kori (1986) and Celata & Mairano (2014). These studies investigate the relation between vowel duration and syllable structure in Italian. In the first study, it was found that vowels followed by a singleton stop (for example in /la.da/) are longer than vowels followed by a cluster belonging to the following syllable (/la.dra/). This pattern can easily be derived from a scenario in which the distance between the vowels is the same in the two contexts (/la.da/ and /la.dra/), and the onset consonants follow a C-centre alignment. This is represented in Figure 5.1

(c). Celata & Mairano (2014) also show that the duration of the consonant/consonant cluster is negatively correlated with the duration of the preceding vowel (although the magnitude of the correlation is low to moderate).

5.1.2 The voicing effect in English

English is one of the most investigated languages in relation to the voicing effect (Meyer 1904; Heffner 1937; House & Fairbanks 1953; Belasco 1953; Peterson & Lehiste 1960; Halle & Stevens 1967; Chen 1970; Klatt 1973; Lisker 1974; Laeufer 1992; Fowler 1992; Hussein 1994; Lampp & Reklis 2004; Warren & Jacks 2005; Durvasula & Luo 2012; Ko 2018). English is also the language in which the voicing effect has the greatest magnitude relative to that of other languages. This special status of English is traditionally attributed to the phonologisation of the voicing effect in this language (Sharf 1964; de Jong 2004). Vowel duration and the vowel-to-consonant duration ratio are considered to be among the most stable cues to consonantal voicing (Peterson & Lehiste 1960; Raphael 1972; Port & Dalby 1982). Kluender et al. (1988) proposed that the difference in vowel duration before voiceless vs voiced stops could have been enhanced and exploited to cue the voicing contrast. This could explain the greater effect of English compared for example to the effect in Italian, in which voicing is most robustly cued by vocal fold vibration during closure (Pape & Jesus 2014).

Indeed, previous studies on English report a difference in vowel duration before voiceless vs voiced stops which ranges between 20 and 150 ms, while the values for the effect in Italian are lower, between 15 and 25 ms (Magno Caldognetto et al. 1979; Farnetani & Kori 1986; Esposito 2002; Coretta 2019b). A Bayesian meta-analysis of the voicing effect (see Appendix B) returned a 95% credible interval for the effect of voicing in English monosyllabic words between 55 and 95 ms, with a meta-analytical mean of 75 ms. In other words, we can be 95% confident that the effect is between 55-95 ms. On the other hand, the meta-analytical estimate of the voicing effect for disyllabic words is lower, at about 25 ms (around 50 ms less than in monosyllabic words). This estimate is closer to the effect sizes reported for Italian. Note also that the Italian values refer to the effect as observed in disyllabic words.

However, it is possible that the alleged differences in magnitude between English

and other languages are a product of the different contexts under examination (Laeufer 1992). Ko (2018), in a more recent investigation of the voicing effect in English monosyllabic words, finds a substantially lower difference in vowel duration (35 ms). The Bayesian meta-analysis (see Appendix B) further suggests a potential for publication bias, which means that the meta-analytical estimate (75 ms) could be an overestimation. Finally, the surveyed studies have a very low number of participants (mean = 3.4, SD = 2.5), which can lead to so-called Type M errors (estimate magnitude errors) and overestimation of the effect (Kirby & Sonderegger 2018; Roettger 2019). In sum, it is generally assumed that the voicing-driven differences in vowel duration are greater in English than in other languages, although the empirical foundation of this conception is not entirely straightforward. Although not the focus of this study, arguments based on differences in effect size will become relevant when discussing the results.

5.1.3 Research hypotheses

One of the aims of this study is to test whether the same temporal stability observed for the release-to-release interval in Italian and Polish disyllabic words can also be observed in English. While the temporal stability of the release-to-release interval is expected in English disyllabic words, monosyllabic words are predicted not to show such stability, for the following reasons.

As discussed above, an essential component of the release-to-release temporal stability in disyllabic words is the presence of a direct relation between the two vowels in these words. Since monosyllabic words don't have a second vowel, there is no direct vowel-to-vowel relation to derive the release-to-release stability from. If coda consonants are produced anti-phase with the preceding vowel and the closing gesture of the consonant starts at a specific time after the production of the vowel independent of consonant voicing, then the greater velocity of the closing gesture in voiceless stops would result in a shorter vowel and a longer closure relative to voiced stops. However, the articulatory data in de Jong (1991) suggests that the timing of the onset of the stop closing gesture differs depending on voicing in English monosyllabic words.

Furthermore, Jacewicz et al. (2009) report that, in American English, monosyllabic words are longer when the second consonant is voiced. Based on this finding, it is ex-

pected that the release-to-release duration should be longer when C2 is voiced. Jacewicz et al. (2009) attribute the difference in monosyllabic word duration to the difference in vowel duration before voiceless vs voiced stops. Thus, we can expect the magnitude of the difference in release-to-release duration in monosyllabic words to be close to the difference in vowel duration. This hypothesis also fits with the reported greater effect of voicing on vowel duration in monosyllabic than disyllabic words. Section 5.4.3 will discuss a possible solution based on perceptual mechanisms that can reconcile the expected absence of release-to-release temporal stability in English monosyllabic words with the expected presence of the voicing effect.

The data in Coretta (2019b) suggests that the intrinsic duration of vowels and consonants can contribute to the duration of the release-to-release interval. In particular, release-to-release intervals containing a high vowel have shorter durations than those with a low vowel. This is not surprising, given the well-known tendency of high vowels to be shorter than low vowels (Hertrich & Ackermann 1997; Esposito 2002; Mortensen & Tøndering 2013; Toivonen et al. 2015; Kawahara et al. 2017). As for the consonantal place of articulation, the release-to-release is shorter in Italian and Polish when the second consonant is velar compared to when it is coronal. This could be a consequence of the fact that the closure of velar stops is shorter than that of other stops. For example, Sharf's (1962) data on closure duration in English suggests that the closure of labial stops (60-90 ms) is about 10 ms longer than that of velar stops (55-75 ms). It can be expected that release-to-release intervals with a velar stop in English will be about 10 ms shorter than intervals with a labial stop.

Another set of objectives concerns the effect of voicing on vowel durations. A conceptual replication of previous studies' effect sizes is sought, with special attention to differences between monosyllabic and disyllabic words. Only a few studies directly compare the effect in different syllabic positions (for example, Sharf 1962 and Klatt 1973). The reported effects are in the range of 50-55 ms in word-final (closed-syllable) position and 20-25 in word-medial (open-syllable) position. The Bayesian meta-analysis of the voicing effect indicates a mean difference of 50 ms (75 ms in word-final position vs 25 ms word-medially).

The data in Sharf (1962) and Luce & Charles-Luce (1985) indicate that voiced stop

closures in English are 10-13 ms shorter than voiceless stop closures. Other surveyed studies did not report effect sizes for stop closure durations (Port & Rotunno 1979; Summers 1987; de Jong 1991). For this study, no specific hypothesis was set in relation to stop closure durations. Closure duration was nonetheless investigated as a conceptual replication of previous work. Moreover, the estimated posterior distributions can be used to set up priors in the statistical analyses of future studies, in accordance to the Bayesian principle of knowledge update via accumulation of evidence (Etz et al. 2018).

To summarise, the following research questions and respective hypotheses can be formulated:

1. Is the duration of the interval between two consecutive stop releases (the release-to-release interval) in monosyllabic and disyllabic words affected by the voicing of C2 in English?

- H1a: The duration of the release-to-release interval is not affected by C2 voicing in disyllabic words.
- H1b: The release-to-release interval is longer in monosyllabic words with a voiced C2 than in monosyllabic words with a voiceless C2.

2. Is the duration of the release-to-release interval affected by (a) the number of syllables of the word, (b) the quality of V1, and (c) the place of C2?

- H2a: The release-to-release interval is longer in monosyllabic than in disyllabic words.
- H2b: The duration of the release-to-release interval decreases according to the hierarchy /ɑ:/, /ɜ:/, /i:/.
- H2c: The release-to-release interval is longer when C2 is labial.

3. What is the estimated difference in the effect of voicing on vowel and stop closure duration between monosyllabic and disyllabic words?

- H3: The effect of voicing on vowel duration is greater in monosyllabic than in disyllabic words (no specific hypothesis in relation to closure duration).

5.2 Methods

The following subsections describe the experimental and statistical methods of this study. The research design and data analyses were pre-registered on the Open Science Framework prior to data collection (`https://osf.io/hwr94/`). The research compendium of this paper with data (Coretta 2019a) and analysis scripts (`https://osf.io/32fst/`) is also available on the Open Science Framework. Choices on experimental design and analysis were made within the Bayesian framework of statistical inference (see Section 5.2.1 and Section 5.2.7 for details).

5.2.1 Sample size and stopping rule

Sample size and a stopping rule were decided prior to data collection with a Bayesian method of sample determination based on the Region Of Practical Equivalence (ROPE, Kruschke 2015; Vasishth et al. 2018b). A "no-effect" region of values around 0 is first identified. This null region (the ROPE) can be thought of as a Bayesian 95% credible interval of a distribution, the values within which can be interpreted as a negligible or null effect. For this study, a ROPE between −10 and +10 ms has been chosen. The width of 20 ms is based on the estimates of the just noticeable difference in Huggins (1972) and Nooteboom & Doodeman (1980). Differences in release-to-release durations below 10 ms (either positive or negative) will be interpreted as compatible with a null effect.

Once a ROPE width is set, the goal is to collect data during sequential testing until the width of the 95% credible interval (CI) of the tested effect is equal to or less than the ROPE width (in this study, 20 ms). In other words, the objective is to reach estimate precision, rather than significance (as in frequentist null hypothesis testing). Inference can then be made based on the credible interval of the sought effect. When the precision goal is reached (the CI width is equal or lower than the ROPE width), three possible scenarios can arise: (1) the CI of the effect completely overlaps with the ROPE around 0, in which case the data supports a practically equivalent null effect; (2) the CI of the effect completely lies outside the ROPE, which indicates that the data support the effect to be within that CI; (3) the CI partially overlaps with the ROPE, in which case no decision can be made on whether the data support one hypothesis over the other,

although it still possible to infer the sign of the effect (if the CI partially overlaps with the right side of the ROPE without including 0, there is evidence for a positive effect, while if the CI overlaps with the left side of the ROPE without including 0, there is evidence for a negative effect).

An initial minimum of 20 participants was chosen for sequential testing. Due to resource and time constraints specific to this particular study, a second condition had to be included in the stopping rule such that data collection would be have to stop on 5 April 2019, independent of the ROPE condition.

5.2.2 Participants

The participants of this study were 15 native speakers of British English, who were born and raised in the Greater Manchester area. The speakers were all undergraduate students at the University of Manchester with no reported hearing or speaking disorders, and with normal or corrected to normal vision. The participants signed a written consent form and received £5 for participation.

5.2.3 Equipment

Audio recordings were obtained in a sound-attenuated room in the Phonetics Laboratory of the University of Manchester, with a Zoom H4n Pro recorder and a RØDE Lavalier microphone, at a sample rate of 44100 Hz (16-bit, downsampled to 22050 Hz for analysis). The Lavalier microphone was clipped on the participants clothes, about 20 cm from the mouth, displaced a few centimetres to one side.

5.2.4 Materials

The test words were $C_1\acute{V}_1C_2$(VC) words, where C_1 = /t/, V_1 = /iː, ɜː, ɑː/, C_2 = /p, b, k, g/, and (VC) = /əs/. /əs/ was chosen for its lower parsability as a native suffix, in order to prevent morphological complexity in disyllabic words. This structure specification generates 24 test words, shown in Table 5.1. All of these are nonce words, with the exception of *turk* and *tarp*, and of *teek* via the homophone *teak*. Building stimuli from

Table 5.1: Test $C_1\acute{V}_1C_2$(VC) words.

teep	teepus	teek	teekus
teeb	teebus	teeg	teegus
terp	terpus	terk	terkus
terb	terbus	terg	tergus
tarp	tarpus	tark	tarkus
tarb	tarbus	targ	targus

a structure template rather than from the lexicon ensures greater experimental and statistical control. Moreover, the use of nonce words removes or reduces confounds from some usage variables, like for example lexical frequency.[1] Each word was embedded in the following frame sentences: *I'll say X this Thursday*, *You'll say X this Monday*, *She'll say X this Sunday*, *We'll say X this Friday*, *They'll say X this Tuesday*. Each word + frame combination was included once in the stimuli list, so that each speaker read a total of 120 sentence stimuli (24 words × 5 frames). A total of 1800 observations were recorded (120 stimuli × 15 speakers).

5.2.5 Procedure

The experimental procedure was first explained to the participants prior to recording. The participants also familiarised themselves with the materials by reading them aloud. They were instructed not to insert pauses anywhere within the sentence stimuli and to keep a similar intonation contour for the total duration of the experiment. They were also given the chance to take any number of breaks at any point during recording. Misreadings or speech errors were corrected by asking the participant to repeat the stimulus. The reading task took around 6 to 10 minutes, while the total experiment session lasted about 25 minutes. Data collection started on 19 February 2019 and ended on 5 April 2019.

[1]The three real words in the materials have low lexical frequency (Zipf 1-7 log-frequency: *tarp* 2.23, *teak* 2.76, and *turk* 2.91) according to the SUBTLEX-UK corpus (Van Heuven et al. 2014).

5.2.6 Data processing and measurements

A forced-aligned transcription was obtained with the SPeech Phonetisation Alignment and Syllabification software (SPPAS, Bigi 2015). The automatic annotation was corrected by the author according to the principles of phonetic segmentation detailed in Machač & Skarnitzl (2009). A custom Praat script was written to automatically detect the burst onset of the consonants in the test words, using the algorithm in Ananthapadmanabha et al. (2014). The output was checked and manually corrected by the author when necessary.

The following measures were obtained via a custom Praat script:

- Duration of the release-to-release interval: from the release of C1 to the release of C2.
- V1 duration: from appearance to disappearance of higher formant structure in the spectrogram in correspondence of V1 (Machač & Skarnitzl 2009).
- C2 closure duration: from disappearance of higher formant structure in the V1C2 sequence to the release of C2 (Machač & Skarnitzl 2009).
- Speech rate: calculated as the number of syllables per second (number of syllables in the sentence divided by the sentence duration in seconds, Plug & Smith 2018).

5.2.7 Statistical analysis

The choice of Bayesian over frequentist statistics stems from a recent discussion of the problems associated with the reliance of *p*-values in statistical inference (Wagenmakers 2007; Munafò et al. 2017; Kirby & Sonderegger 2018; Roettger 2019). Bayesian statistics also offers a straightforward framework for investigating the absence of differences across conditions (a "null effect") based on the ROPE (Section 5.2.1), as it is in part the case in this study. Another favourable aspect of Bayesian methods is that more focus is given to the distributions of the enquired effects, rather than on point estimates (which are less informative when matters of statistical power are taken into consideration, see a discussion of Type S-M errors in Kirby & Sonderegger 2018) and an arbitrary significance cut-off point. Furthermore, Bayesian inference is centred around an incremental

procedure of reallocation of credibility between natural states and on evidence based on observed data (Kruschke 2015), rather than on a series of hypothetical experimental replications (Wagenmakers 2007).[2] For an introduction to Bayesian statistics in phonetics, see Vasishth et al. (2018a), and Nicenboim et al. (2018), while for a general introduction see Etz et al. (2018), McElreath (2015), Kruschke (2015), and references therein. While a thorough discussion of Bayesian methods would be beyond the scope of this paper, it is relevant to provide the less familiar reader with the basic tools for interpreting analyses and results.

Particular weight will be given to the estimated distributions of the sought effects in presenting the results of this study. The estimated distribution of an effect (or parameter) is the posterior distribution of that effect (or parameter). The posterior distribution is an approximation of the parameter distribution, and it takes into account the specified prior for that parameter, i.e. the theoretical probability of the parameter as known or derived by the researcher. The inclusion of priors in the analysis is at the heart of Bayesian modelling, which relies on prior knowledge for the estimation of parameter values. For each relevant term in the models, the 95% credible intervals (CI) should be taken as a summary of the posterior distribution, and inference should be based on the posterior rather than on the point estimate (the posterior mean, represented here with $\bar{\theta}$). A 95% CI can be interpreted as the 95% probability that a parameter lies within that interval range. For example, if the 95% CI is between 10 and 30 ms, there is a 95% probability that the true parameter value is between 10 and 30 ms, with extreme values being less likely than values in the centre of the interval.

In each model, priors are specified for each of the parameters to be estimated. The priors are in the form of particular distributions, like the Gaussian (normal) or the Cauchy distribution. A prior defines the prior knowledge of where the parameter might lie within a range of values. For example, a prior as a normal distribution with mean 200 ms and standard deviation 50 indicates the researcher's belief that the parameter lies between 100 and 300 ms with 95% probability (i.e., the mean minus twice the standard deviation, and the mean plus twice the standard deviation).

[2] I am not advocating here against *p*-values in absolute terms. On the contrary, *p*-values are still useful in that they provide us with a practical solution in situations that involve, for example, decision-making.

Statistical analysis was performed in R v3.5.3 (R Core Team 2019). Bayesian regression models were fit with brms (Bürkner 2017, 2018). Each model was run with four MCMC chains and 2000 iterations per chain, of which 1000 for warm-up. A Gaussian (normal) distribution was used in all the models as the response distribution. All factors were coded using treatment contrasts (the first level in this list was set as the reference level): number of syllables (disyllabic, monosyllabic), vowel (/ɑ:/, /ɜ:/, /i:/), C2 voicing (voiceless, voiced), C2 place of articulation (velar, labial). Speech rate was centred when included in the models so that the intercept could be interpreted as the intercept at mean speech rate. A seed (1234) was set in all models to ensure reproducibility of the output. The priors used in the models reported here will be discussed along with the results in the following sections.

A concern could be raised that the priors might have greater influence on the posterior distributions than the observed data. A sensitivity analysis based on posterior *z*-scores and shrinkage (Betancourt 2018) indicates that the models discussed in this study are highly informed by the observed data and don't heavily rely on prior specifications.

5.3 Results

This section reports the results of the Bayesian models, grouped by outcome variable (release-to-release, vowel duration, closure duration). A description of the model structure and priors is given for each model, followed by the presentation of the posterior distributions of the relevant terms. Each model is assigned a number (1 to 5), and the text refers to these.

Model convergence was reached in all the reported models ($\hat{R}$ = 1) and no major divergences in the MCMC chains were observed. The posterior predictive check plots indicate that the observed distributions are slightly positively skewed so that a log-normal distribution would have been more appropriate. Previous work has shown that speech-units duration does follow, as a general trend, a log-normal distribution (Rosen 2005; Ratnikova 2017), and the practice of transforming duration data to the logarithmic scale is not uncommon (Gahl & Baayen 2019). However, the deviations

from a Gaussian distribution here were minimal, and an informal comparison of one of the models fitted with a log-normal distribution led to virtually identical results.

5.3.1 Release-to-release duration

A Bayesian regression was fit to model the duration of the release-to-release interval (model 1). The following terms were included as fixed effects: C2 voicing (voiceless, voiced), number of syllables (disyllabic, monosyllabic), centred speech rate, an interaction between C2 voicing and number of syllables. A by-speaker and by-word random intercept, and a by-speaker random coefficient for C2 voicing were entered as random effects. The following priors were used. Two weakly informative priors based on the results from Coretta (2019b) were chosen for the intercept and the effect of C2 voicing. The former prior is a normal distribution with mean 200 ms and SD = 50, while the latter a normal distribution with mean 0 ms and SD = 25. A weakly informative prior as a normal distribution with mean 50 ms and SD = 25 was specified for the effect of number of syllables. The prior is based on differences in vowel duration between mono- vs disyllabic words, which range between 30 and 100 ms (Sharf 1962; Klatt 1973). The same prior was used for the interaction between C2 voicing and number of syllables, based on the reported differences in voicing effect in mono- vs disyllabic words (Sharf 1962; Klatt 1973). The prior for the effect of centred speech rate is a normal distribution with mean -25 ms and SD = 10, and is based on results from Coretta (2019b). For the random effects, a half Cauchy distribution (location = 0, scale = 25) was used for the standard deviation and the residual standard deviation, and a LKJ(2) distribution for the correlation among the random terms.

Table 5.2 gives the posterior mean, posterior standard deviation, 2.5 and 97.5 quantiles (lower and upper bounds of the 95% credible interval), and the credible interval's width of the fixed effects of model 1. According to the hypotheses H1a-b set out in Section 5.1.3, the effect of the C2 voicing predictor (which refers to disyllabic words) should be 0 (i.e. the posterior distribution of the effect should be entirely within the ROPE), and the interaction between C2 voicing (= voiced) and number of syllables (= monosyllabic) should be positive. Based on H2a, the effect of the number of syllables (= monosyllabic) should be positive. However, note that the precision goal (CI width

Table 5.2: Summary of the Bayesian regression fitted to release-to-release duration (model 1, see Section 5.3.1).

Predictor	Mean	SD	Q2.5	Q97.5	CI width
Intercept	263.71	9.64	244.17	283.00	38.84
Voicing = voiced	-4.43	10.03	-23.86	15.45	39.30
Num. syll. = monosyllabic	17.34	9.76	-1.58	36.53	38.11
Speech rate (cntr.)	-36.10	2.06	-40.14	-32.13	8.01
voiced × monosyll.	16.53	12.72	-8.41	41.41	49.83

≤ 20 ms, based on the ROPE) was reached only for centred speech rate (CI width = 8.14 ms), so that these results come with a high degree of uncertainty. The posterior distribution of the estimated effect of C2 voicing on the release-to-release duration in monosyllabic words has a 95% credible interval (95% CI) between -23.86 and 15.45 ms (the mean is -4.43 ms, SD = 10.03). The 95% CI of the estimated interaction between C2 voicing (= voiced) and number of syllables (= monosyllabic) tends towards positive values, between -8.41 and 41.41 ms ($\bar{\theta}$ = 16.53 ms, SD = 12.72). The difference in duration of the release-to-release interval between monosyllabic and disyllabic words is more clearly positive, between -1.58 and 36.53 ms (95% CI, $\bar{\theta}$ = 17.34, SD = 9.76). Speech rate has a strong negative effect on the release-to-release duration with 95% CI = [-40.14, -32.13].

A second Bayesian regression (model 2) was fitted with the release-to-release duration as the outcome variable to test the effects of vowel and C2 place of articulation, which were entered as terms in the model without interactions. Centred speech rate was also included. The random effects structure was the same as with the first model. The relevant priors from the first model were kept. For the effects of vowel (/ɜ:/, /i:/) and place of articulation (labial), the very weakly informative prior used is a normal distribution with mean = 0 ms and SD = 30. This prior was based on duration differences depending on vowel height (Heffner 1937; House & Fairbanks 1953; Hertrich & Ackermann 1997) and labial place of articulation (Sharf 1962), which both range between 10 and 30 ms.

The summary of the fixed effects of model 2 are given in Table 5.3. As with model

Table 5.3: Summary of the Bayesian regression fitted to release-to-release duration (model 2, see Section 5.3.1).

Predictor	Mean	SD	Q2.5	Q97.5	CI width
Intercept	289.05	8.14	273.01	305.09	32.08
Vowel = /ɜ:/	-8.58	6.90	-21.90	4.87	26.78
Vowel = /i:/	-36.94	6.96	-50.10	-22.26	27.84
C2 place = labial	2.46	5.68	-9.15	13.28	22.44
Speech rate (cntr.)	-37.48	2.05	-41.51	-33.37	8.14

1, only the CI width of speech rate reached the intended precision. Based on the hypotheses H2b-c in Section 5.1.3, the effects of vowel = /ɜ:/ and vowel = /i:/ should be negative, and the latter effect should be more negative than the former. The posterior distribution of the effect of the vowel /ɜ:/ shows that this vowel tends to a negative effect, with a 95% CI between −21.90 and 4.87 ms ($\bar{\theta}$ = −8.58 ms, SD = 6.9). The vowel /i:/ has a more robust negative effect on release-to-release duration, with a 95% CI between −50.10 and −22.26 ($\bar{\theta}$ = −36.94 ms, SD = 6.96). Hypothesis H2c states that the effect of C2 place of articulation when the consonant is labial is positive (i.e. the interval is longer when C2 is labial). The robustness of the effect of C2 place of articulation (velar vs labial stop) is less compared to the other effects: The mean of the posterior is 2.46 ms (SD = 5.68), and the 95% CI is [−9.15, 13.28].

The credible intervals of the effects in the models reported above have widths which are greater than the chosen ROPE width of 20 ms. The wide credible intervals indicate that the estimated posterior distributions of the effects have a somewhat high degree of uncertainty with them. This uncertainty is potentially due to not controlling for vowel and number of syllables in the first and second model respectively. An exploratory model (model 3) was thus fitted to the data, in which all the terms from the two models above were included. The same priors of the two separate models were used in the combined model.

Including all the relevant terms in the model (C2 voicing and place, vowel, number of syllables in interaction with C2 voicing) reduces the width of the credible intervals substantially. Figure 5.2 shows the posterior distributions of the model terms with a va-

riety of credible intervals. Hypothesis H1a states that the effect of C2 voicing (= voiced) on the release-to-release duration of disyllabic words should be null (i.e., the posterior should be contained within the ROPE). The posterior distribution of the C2 voicing effect on release-to-release duration in this aggregated model (95% CI = [−10.45, 5.65]) is tighter than that of model 1 (95% CI = [-23.86, 15.45]) while the mean (−2.43 ms, SD = 4.06) is virtually unchanged (−4.43 ms, only a 2 ms difference). According to hypothesis H2a, the effect of number of syllables (= monosyllabic) on the release-to-release should be positive. The estimated effect of syllable number is robustly positive (95% CI = [9.17, 22.48]), with a mean (16.03 ms, SD = 3.32) similar to that in model 1. Based on hypothesis H1b, the interaction between C2 voicing (= voiced) and number of syllables (= monosyllabic) should be positive. The posterior distribution of the interaction between number of syllables and C2 voicing (95% CI = [2.65, 20.98]) suggests a positive and medium-sized interaction effect ($\bar{\theta}$ = 11.67 ms, SD = 4.71). This result indicates that the duration of the release-to-release is greater in monosyllabic words with voiced C2 than in monosyllabic words with voiceless C2, compatibly with H1b. The effects of vowel and place of articulation have similar means as in model 2, but the credible intervals are smaller. According to hypotheses H2b-c, the effects of vowel = /ɜː/ and vowel = /iː/ should be negative, the later effect should be more negative than the former, and the effect of C2 place of articulation when the consonant is labial should be positive. The release-to-release is on average 10.05 ms (SD = 2.95, 95% CI = [−15.92, −4.24]) shorter if the vowel is /ɜː/ and 39.3 ms (SD = 2.99, 95% CI = [−45.03, −32.76]) shorter if the vowel is /iː/. C2 place of articulation (labial) has a negligible positive mean effect (2.6 ms, SD = 2.39, 95% CI = [−2.29, 7.28]).

5.3.2 Vowel duration

A Bayesian regression model was fitted to test vowel duration (model 4). The following terms were entered: C2 voicing (voiceless, voiced), vowel (/ɑː/, /ɜː/, /iː/), number of syllables (disyllabic, monosyllabic), centred speech rate, all possible interactions between C2 voicing, vowel, and number of syllables. The same random structure as in the previous models was used (a by-speaker and by-word random intercept, and a by-speaker random coefficient for C2 voicing).

Table 5.4: Summary of the Bayesian regression fitted to release-to-release duration and predictors from model 1 and 2 (model 3, see Section 5.3.1).

Predictor	Mean	SD	Q2.5	Q97.5	CI width
Intercept	280.81	6.99	266.72	294.37	27.66
Voicing = voiced	-2.43	4.06	-10.45	5.65	16.10
Num. syll. = monosyllabic	16.03	3.32	9.17	22.48	13.31
Vowel = /ɜ:/	-10.05	2.95	-15.92	-4.24	11.68
Vowel = /i:/	-39.03	2.99	-45.03	-32.76	12.27
C2 place = labial	2.46	2.39	-2.29	7.28	9.57
Speech rate (cntr.)	-36.10	1.99	-39.96	-32.24	7.72
voiced × monosyll.	11.67	4.71	2.65	20.98	18.33

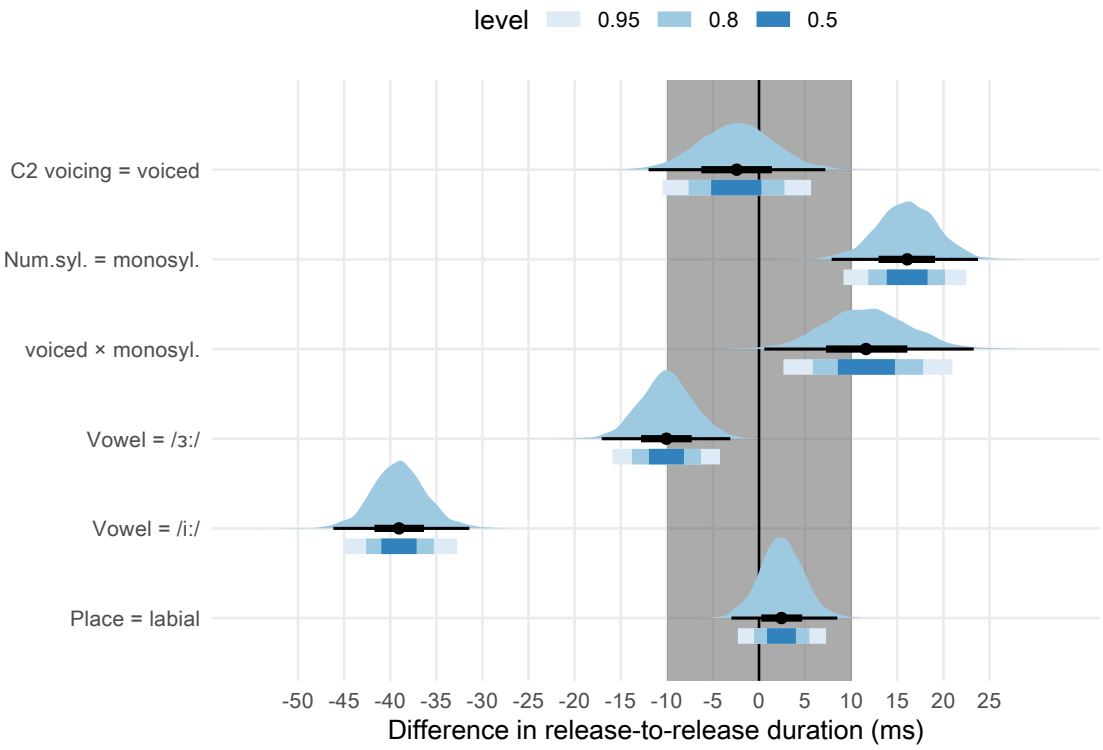


Figure 5.2: Posterior distributions and Bayesian credible intervals of the effects on release-to-release duration (model 3). For each effect, the thick blue-coloured bars indicate (from darker to lighter) the 50%, 80%, and 95% CI. The black point with bars are the posterior median (the point), the 98% (thin bar) and 66% (thicker bar) CI. The shaded grey area around 0 is the ROPE.

Table 5.5: Summary of the Bayesian regression fitted to vowel duration (model 4, see Section 5.3.2).

Predictor	Mean	SD	Q2.5	Q97.5	CI width
Intercept	124.91	5.96	112.94	136.77	23.83
Voicing = voiced	13.65	5.16	3.73	24.09	20.36
Vowel = /ɜː/	-9.03	5.13	-19.08	1.63	20.71
Vowel = /iː/	-36.77	5.00	-46.42	-26.67	19.74
Num. syll. = monosyllabic	14.91	5.07	5.15	25.14	19.99
Speech rate (cntr.)	-18.03	1.48	-20.93	-15.29	5.63
voiced × /ɜː/	0.24	6.83	-13.70	13.94	27.64
voiced × /iː/	6.73	6.59	-6.54	19.26	25.80
voiced × monosyll.	4.03	6.70	-8.98	17.69	26.67
/ɜː/ × monosyll.	0.53	7.07	-13.57	14.57	28.15
/iː/ × monosyll.	-16.07	6.93	-30.03	-2.68	27.35
voiced × /ɜː/ × monosyll.	-2.94	9.46	-21.37	15.77	37.14
voiced × /iː/ × monosyll.	14.46	9.18	-3.59	31.99	35.58

For the prior of the intercept of vowel duration, a normal distribution with mean 145 ms and standard deviation 30 was used (Heffner 1937; House & Fairbanks 1953; Peterson & Lehiste 1960; Sharf 1962; Chen 1970; Klatt 1973; Davis & Summers 1989; Laeufer 1992; Ko 2018). A normal distribution with mean 50 ms and standard deviation 20 was used as the prior for the effect of voicing on vowel duration (based on the above studies). A normal prior with mean 50 and standard deviation 25 was chosen instead for the effect of number of syllables and the interaction C2 voicing/number of syllables. For the effects of vowel, vowel/number of syllables interaction, and the three-way interaction vowel/number of syllables/C2 voicing, the prior was a normal distribution with mean 0 and standard deviation 30, based on differences reported in the studies above. A slightly more informative prior was used for the interaction between C2 voicing and vowel (mean = 0, SD = 20). The same priors as in the previous models were included for the random effects.

Table 5.5 reports the summary of model 4, while Figure 5.3 shows the posterior distributions and credible intervals. The precision target was reached in the non-interacting

predictors (permitting a few milliseconds above 20), with the exception of the intercept. All the interactions terms have CI widths above 25 ms. The 95% CI of the posterior distribution of the duration of /ɑː/ is included in the range 112.94–136.77 ms ($\bar{\theta}$ = 124.91 ms, SD = 5.96). The vowel /ɜː/ is 9.03 ms shorter (SD = 5.16) with CI = [−19.08, 1.63], while /iː/ is 36.77 ms shorter (SD = 5, 95% CI = [−46.42, −26.67]). C2 voicing has a small but robust positive effect on vowel duration in disyllabic words. The posterior distribution of the effect of voicing on /ɑː/ has mean 13.65 ms (SD = 5.16) and 95% CI = [3.73, 24.09]. The posterior of the interaction of voicing with vowel when the vowel is /ɜː/ is quite spread out around 0, with the 95% CI between −13.70 and 13.94 ms. This indicates that /ɑː/ and /ɜː/ are similar in their behaviour of voicing-driven durational differences. On the other hand, the effect of voicing is on average 6.73 ms greater (SD = 6.59, 95% CI = [−6.54, 19.26]) when the vowel is /iː/.

According to hypothesis H3, the interaction effect between C2 voicing (= voiced) and number of syllables (= monosyllabic) should be positive (i.e. the effect should be greater in monosyllabic than in disyllabic words). The magnitude of the voicing effect in disyllabic vs monosyllabic words is modulated by the identity of the vowel. The posterior distribution for the interaction C2 voicing/number of syllables when the vowel is /ɑː/ has mean 4.03 ms (SD = 6.7) and 95% CI [−8.98, 17.69]. This distribution indicates the possibility for a very small increase of the effect from disyllabic to monosyllabic words with /ɑː/. The three-way interaction C2 voicing/vowel/number of syllables suggests that the effect of voicing in monosyllabic words with /ɜː/ is very similar to that of monosyllabic /ɑː/-words ($\bar{\theta}$ = −2.94, SD = 9.46, 95% CI = [−21.37, 15.77]). On the other hand, the effect increases by 14.46 ms (SD = 9.18, CI = [−3.59, 31.99]) in monosyllabic words with /iː/ relative to disyllabic /iː/-words. Note that the credible intervals of these interaction effects are quite large, so that a wide range of values are probable at 95% confidence.

5.3.3 Consonant closure duration

To test various effects on C2 closure duration, model 5 was fit with closure duration as the outcome variable and the following predictors: C2 voicing (voiceless, voiced), C2 place of articulation (velar, labial), number of syllables (disyllabic, monosyllabic),

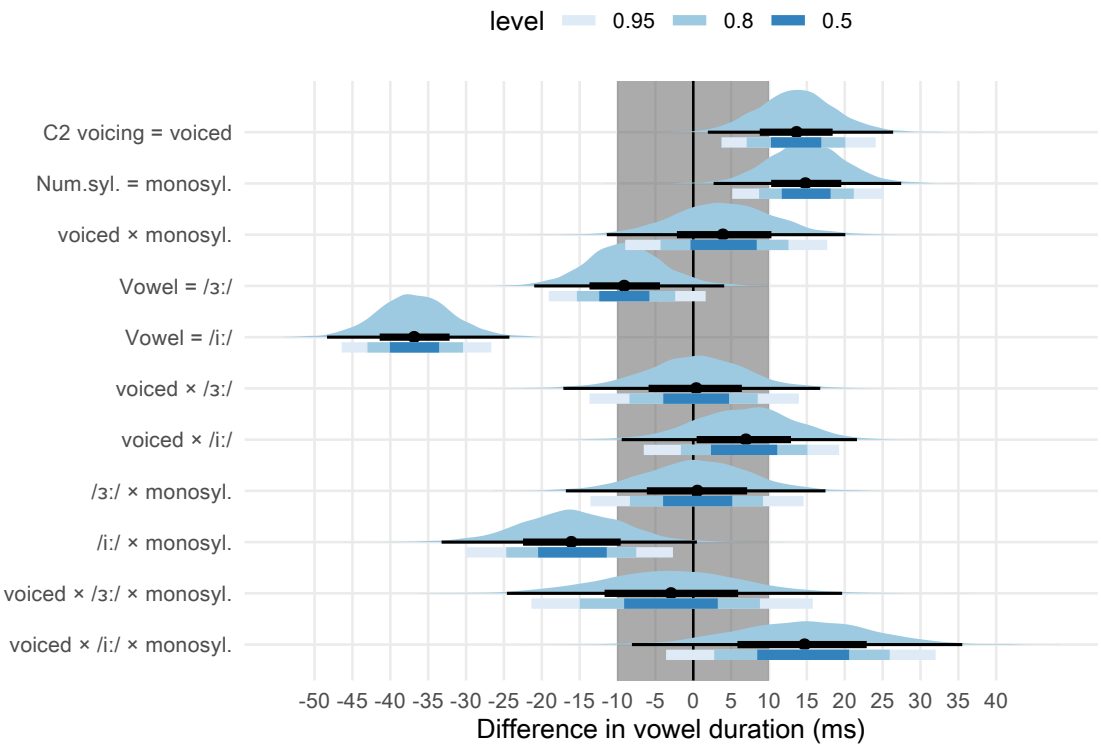


Figure 5.3: Posterior distributions and Bayesian credible intervals of the effects on vowel duration (model 4). For each effect, the thick blue-coloured bars indicate (from darker to lighter) the 50%, 80%, and 95% CI. The black point with bars are the posterior median (the point), the 98% (thin bar) and 66% (thicker bar) CI. The shaded grey area around 0 is the ROPE.

Table 5.6: Summary of the Bayesian regression fitted to closure duration (model 5, see Section 5.3.3).

Predictor	Mean	SD	Q2.5	Q97.5	CI width
Intercept	74.75	2.86	69.07	80.59	11.52
Voicing = voiced	-20.79	3.06	-26.77	-14.74	12.03
C2 place = labial	5.19	2.77	-0.03	10.76	10.79
Num. syll. = monosyllabic	2.98	2.90	-2.80	8.77	11.58
Speech rate (cntr.)	-9.21	1.26	-11.71	-6.74	4.97
voiced × labial	1.37	3.94	-6.79	8.93	15.72
voiced × monosyll.	1.82	4.06	-6.08	9.70	15.78
labial × monosyll.	-0.74	4.02	-8.95	6.88	15.83
voiced × labial × monosyll.	6.41	5.66	-4.72	17.45	22.17

all interactions between these predictor terms, and centred speech rate. The random effects were again a by-speaker and a by-word random intercept, and a by-speaker random coefficient for C2 voicing.

As priors, a normal distribution with mean 90 ms (SD = 20) was used for the intercept, based on Sharf (1962) and Luce & Charles-Luce (1985). The means reported in these studies also indicate that the closure of the stop in monosyllabic words is 10-30 ms shorter when the stop is voiced. A normal distribution with mean -20 ms (SD = 10) was chosen as the prior of the effect of C2 voicing on closure duration. The same studies indicate that labial stops have a closure which is 10-20 ms longer than the closure of velar stops. For the effect of C2 place, a normal distribution with mean 15 ms (SD = 10) was used.

The summary of model 5 is shown in Table 5.6. See Figure 5.4 for the posteriors and credible intervals of the effects. The 96% CI width of all the terms, with the exception of the three-way interaction (voicing/place/number of syllables), is below 20 ms (the precision goal has been reached). As explained in Section 5.1.3, no specific hypothesis addressing closure duration was formulated. The posterior distribution of the intercept for closure duration (corresponding to the duration of voiceless velar stops in disyllabic words) has mean 74.75 ms (SD = 2.86) and 95% CI = [69.07, 80.59]. The effect of C2

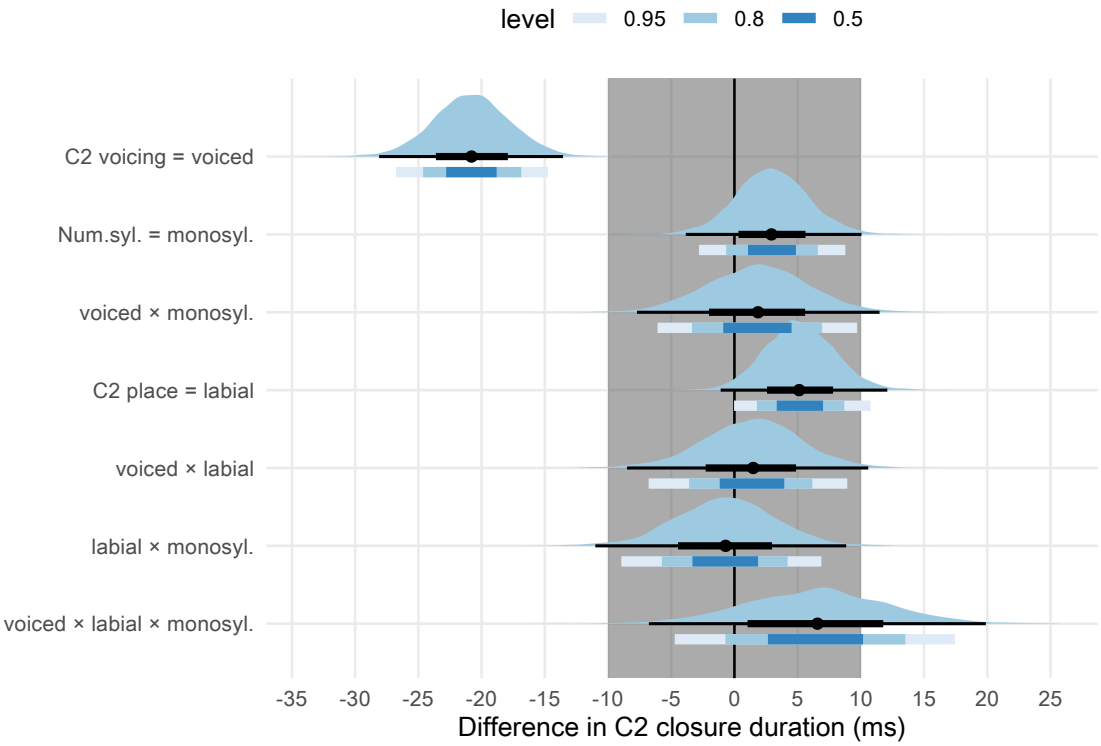


Figure 5.4: Posterior distributions and Bayesian credible intervals of the effects on closure duration (model 5). For each effect, the thick blue-coloured bars indicate (from darker to lighter) the 50%, 80%, and 95% CI. The black point with bars are the posterior median (the point), the 98% (thin bar) and 66% (thicker bar) CI. The shaded grey area around 0 is the ROPE.

voicing on closure duration is certainly negative, between −26.77 and −14.74 ms (95% CI). The posterior mean of this effect is −20.79 ms (SD = 3.06). A very small positive effect of place of articulation (labial) is suggested by the 95% CI from −0.03 to 10.76 ms ($\bar{\theta}$ = 5.19 ms, SD = 2.77). A possibly even smaller effect of number of syllables or no effect at all can be inferred from the posterior distribution which has mean 2.98 ms and SD 2.9 (95% CI = [−2.8, 8.77]). Note that the 95% CIs of the posterior distributions of all the effects, with the exception for the effect of voicing, are within the ROPE around 0.

5.4 Discussion

This study set out to build on the results discussed in Coretta (2019b) by investigating durational properties of the release-to-release interval in English monosyllabic and disyllabic words. It was expected that the release-to-release interval would not be affected by C2 voicing in disyllabic words but it would in monosyllabic words. Moreover, a conceptual replication of studies on the effect of consonant voicing on vowel and closure durations was sought, with a focus on comparing the effect in mono- vs disyllabic words. This section discusses in turn the results in relation to the release-to-release interval duration (Section 5.4.1) and to vowel and closure durations (Section 5.4.2) by comparing them with the hypotheses of this study. Section 5.4.3 synthesises and links these findings back to the articulatory grounding of the temporal properties of the release-to-release interval in mono- and disyllabic words (Section 5.1.1). Limitations and future work are also discussed.

5.4.1 Release-to-release interval

The first question (see Section 5.1.3) asked whether the voicing of C2 in disyllabic and monosyllabic words in English influences the duration of the release-to-release interval. Coretta (2019b) showed that the release-to-release interval duration is not affected by C2 voicing in disyllabic words of Italian and Polish. The hypotheses were that, in English, the interval is not affected in disyllabic words, like in Italian and Polish, but that it is in monosyllabic words. In sum, the results of this study indicate that the release-to-release duration of disyllabic words in English is relatively stable independent of whether C2 is voiceless (like in /tɑːpəs/) or voiced C2 (/tɑːbəs/). On the other hand, the release-to-release in monosyllabic words is longer if C2 is voiced (like in /tɑːb/ vs /tɑːp/).

Two pre-registered Bayesian regression models were fitted to the release-to-release duration (model 1-2). The established ROPE target has not been achieved (see Section 5.2.1). An exploratory model (model 3) including all predictors from model 1 and 2 resulted in higher estimate precision (CI widths below 20 ms). The results of model 3 suggest a negligible effect of C2 voicing on the interval duration in disyllabic words

(hypothesis 1a), with a 95% probability that the true effect is between −10 and +5 ms. At lower levels of probability, the posterior distribution indicates an effect between −6 and 1 ms (60% probability). If the voicing of C2 is conditioning the duration of the release-to-release interval, this effect is very small.

The possible small effect of C2 voicing in disyllabic words could be related to an annotation bias which affects the identification of stop releases. English voiceless stops are generally followed by aspiration, and the glottal friction that makes up aspiration could mask the burst of the release. If the release of the post-vocalic voiceless stops is annotated later than the actual release (by mistaking peaks in the aspiration noise for the release burst), this could lead to longer release-to-release durations when C2 is voiceless compared to when it is voiced. Such annotation bias could explain the quite small negative effect of voicing on the interval duration, and why it is in the opposite direction of the one predicted for monosyllabic words (i.e. *longer* release-to-release when C2 is voiced).

On the other hand, the release-to-release interval in monosyllabic words is longer when C2 is voiced (for example, /tɑːb/) vs when it is voiceless (/tɑːp/). The interaction term between number of syllables in the word and C2 voicing is positive, between +2.5 and +21 ms (at 95% probability), which means that the effect of C2 voicing increases by 2.5 to 21 ms in monosyllabic words relative to the effect in disyllabic words. This result is compatible with hypothesis 1b that the release-to-release interval is longer in monosyllabic words with a voiced C2 than in monosyllabic words with a voiceless C2. As discussed in Section 5.1, the absence of release-to-release isochrony in monosyllabic words is possibly due to the absence of a second vowel which would constitute the left articulatory anchor for vowel isochrony, which in turn is argued to be the necessary element for the release-to-release temporal stability.

The second question posed at the beginning of the paper was about other effects on the release-to-release duration. As expected by hypothesis 2a, the release-to-release is longer in monosyllabic than in disyllabic words. At 95% probability, the effect of number of syllables (from di- to monosyllabic) is between 9 and 22.5 ms. As for hypothesis 2b, the results are more robust for /iː/ than for /ɜː/. When the vowel is /iː/, the release-to-release interval is 33 to 45 ms shorter compared with an interval with /ɑː/.

The posterior distribution of the effect when the vowel is /ɜː/ substantially overlaps with the ROPE, although it tends towards the negative side. If there is an effect with this vowel compared to /ɑː/, it is negative and possibly around −10 ms. Finally, hypothesis 2c is not unequivocally corroborated. The posterior distribution of the effect of C2 place of articulation (labial) has very high precision (9.5 ms) and it is between 0 and 5 ms (at somewhat less than 80% probability). However, it lies within the ROPE and it is very close to 0.

5.4.2 Vowel and closure duration

Question 3 addressed the effect of voicing on vowel and closure duration, and the possible differences between disyllabic and monosyllabic words. The effect of voicing on vowel duration found in this study was estimated to lie between 4 and 25 ms. This range of values is very similar to that reported in Coretta (2019b) for Italian and Polish disyllabic words (the 95% confidence interval for the effect in these languages is [8, 25]), monosyllabic words were not tested). When compared to the values in previous studies that investigated disyllabic words (Sharf 1962; Klatt 1973; Davis & Summers 1989), the effect size found in this study tends towards smaller values. However, note that the posterior distribution of the effect in the current study is entirely contained in the meta-analytical posterior distribution of the effect in the other studies, which roughly ranges between −15 and +65 ms (see Appendix B). Thus, we can assume that the deviation of this study from previous ones is not substantial. As for the effect of number of syllables on vowel duration, a similar effect to that of voicing was found, whereby vowel durations increase by 5 to 25 ms in monosyllabic words compared to disyllabic words. This relation corresponds to what has previously been reported in the literature. Finally, given that the 95% CIs of the effects of voicing and number of syllables overlap with the right side of the ROPE without including 0, the data supports positive effects, but inference on their magnitude should be carefully weighted.

It was expected that the voicing effect on vowels would be stronger in monosyllabic than in disyllabic words (hypothesis 3). The credible intervals of the posterior distributions from model 4, which are larger than the ROPE, make interpretation less straightforward. At 80% probability, the difference in voicing effect between mono-

and disyllabic words is between −5 and +12.5 ms. The distribution is skewed towards the positive side, and this is compatible with results from previous studies, although the CI includes 0. The magnitude, however, is considerably lower than what previously reported. More data is needed to reach a sensible estimate precision and reduce uncertainty.

The three-way interaction between C2 voicing, vowel, and number of syllables reveals that the effect in monosyllabic words with the vowel /ɜː/ is similar to that with /ɑː/. On the other hand, the effect is larger if the vowel is /iː/. Model 4 estimates an effect increase of about 14.5 ms ([−4.27, 33.41]). Note that the credible interval is very wide (38 ms) and it spans over both negative and positive values, although tends more towards the latter. Moreover, the vowel /iː/ followed by a voiceless stop has, according to the model, the same duration in monosyllabic and disyllabic words. While it is not clear why the vowel should have the same duration in these contexts, this pattern suggest a possible process of /iː/ shortening in monosyllabic words. More research is warranted in relation to the observed patterns.

Turning now to consonants, there was no specific hypothesis concerning the effect of voicing on closure durations. C2 voicing has a robust negative effect on closure duration, so that voiced closures are 14.6-26.8 ms shorter than voiceless closures. The effects of number of syllables, place, and interactions all have credible intervals that are narrower than 20 ms (the ROPE width) but they lie entirely within the ROPE around 0. If these variables do have an effect on closure duration, the present analysis suggests that the means of these effects are between 0 and 5 ms. These values are smaller than those in Sharf (1962), which indicate a difference of 15 ms between velar and labial closure durations.

As a general trend, the differences in vowel and closure duration found in this study are smaller than those known from the literature, and considerably so in the case of vowels. A possible reason for this discrepancy could be found in problems arising from Type M errors (as briefly discussed in Section 5.1), and in differences of speech rate, as evidenced by comparing average segment durations. While the model's intercept of vowel duration in this study is approximately 125 ms (SD = 5.89), the mean vowel duration in the studies surveyed in the meta-analysis (Appendix B) is 150 ms (SD =

36). These longer durations may indicate lower speech rates in older studies and so the effect of voicing may have been greater there than at higher speech rates, assuming a linear increase of the effect. However, the ratio between vowel duration and the effect of voicing differs (a third in this study vs half in previous work). Ko's findings 2018 support the idea that the voicing effect (and the vowel-to-consonant ratio) are not stable across speaking rates, with the consequence that differences are enhanced at decreased speaking rates. More studies like Ko (2018) are needed to settle the issue of the diverging results.

5.4.3 General discussion

Coretta (2019b) proposes that the voicing-related adjustments in the relative timing of the closure onset within an isochronous speech interval (acoustically identified as the release-to-release interval) is the diachronic precursor of the cross-linguistically widespread effect of voicing on vowel duration.[3] Given that the duration of the release-to-release interval in Italian, Polish, and English disyllabic words is not affected by the voicing of the post-vocalic consonant, the relative durations of vowel and closure are thought to depend on the timing of the VC boundary within that interval. A later VC boundary implies a longer vowel and a shorter closure, while, vice versa, an earlier boundary produces a shorter vowel and a longer closure. Behind the differential timing of the VC boundary within the release-to-release interval, several other accounts can be envisaged, like accounts relating to laryngeal and supraglottal adjustments (Halle & Stevens 1967; Beguš 2017; Coretta 2019c).

As discussed in Section 5.1.1, a prerequisite of the articulatory account proposed here for the emergence of the voicing effect is the temporal stability of the acoustic release-to-release interval and of the related articulatory gestures. However, it was expected that English mono-syllabic words do not show such temporal stability, even though the voicing effect is present in this context and allegedly even grater than in disyllabic words (although cf. Appendix B).

As mentioned at the end of Section 5.1.1, the absence of temporal stability and presence of the voicing effect in monosyllabic words can be reconciled by drawing

[3]Note that isochrony here is intended as pertaining the context of voiceless vs voices stops only.

from known mechanisms of perceptual enhancement. Perceptual biases, like the ones proposed by perceptual accounts of the voicing effect (Javkin 1976; Kluender et al. 1988; Sanker 2019), can contribute to the increase of the effect of voicing, for example as a means to enhance the perceptual difference of voiceless vs voiced stops (Lisker 1974, 1986; Stevens & Keyser 1989). In particular, vowels can be further lengthened when followed by voiced stops and/or further shortened when followed by voiceless stops, so as to produce a greater and more perceptible difference.

In the case of disyllabic words, movements of the VC boundary within the isochronous interval will logically affect both vowel duration and closure duration. On the other hand, the absence of a second vowel acting as temporal articulatory anchor (as per vowel-to-vowel isochrony) in monosyllabic words would allow articulatory stretching or compression to operate independently on the vocalic and the consonantal gestures. In the monosyllabic context, the gestural duration of vowels and following consonants can be modified in such a way that could result in a change in timing of the onset of the consonant closing gesture and in the disruption of the release-to-release isochrony.

The presence of a voicing effect with absence of release-to-release temporal stability could be obtained, for example, by keeping the vocalic gesture when the following stop is voiced active for a longer time relative to when the following stop is voiceless. Although more research is needed in this area, the articulatory studies in Raphael (1972) and de Jong (1991) do suggest that the vocalic gesture in monosyllabic words is executed for a prolonged time when the following consonant is voiced. While differences in magnitude of the voicing effect should be further investigated, the potentially greater effect of voicing in monosyllabic words (albeit by a small fraction) could be ascribed to the fact that, while vowel-to-vowel isochrony constraints how vowels and consonants can be produced in disyllabic words, mechanisms affecting the VC boundary (articulatory and/or perceptual) in monosyllabic words are less constrained due to the non-application of vowel-to-vowel isochrony.

5.5 Conclusion

This paper set out to investigate temporal properties of the so-called “voicing effect”, by which vowels are shorter when followed by voiceless stops and longer when followed by voiced stops. Coretta (2019b) proposes that the voicing effect emerges via a mechanism of relative timing of the VC boundary within a temporally stable interval. Such interval was argued to be the interval between two consecutive releases, as evidenced by acoustic data from Italian and Polish disyllabic words. The temporal stability of the release-to-release in relation to consonantal voicing is thought to derive from two properties of gestural phasing, namely the isochrony of the distance between the vowels in a VCV sequence, and in-phase alignment of onset consonants and the following vowel. On the other hand, the lack of an articulatory anchor (a second vowel) in monosyllabic words would allow the release-to-release duration to be affected by C2 voicing and differ in the monosyllabic context.

This study adds to the current status of knowledge on temporal aspects of the voicing effect by showing that the release-to-release interval is not affected by C2 voicing in English disyllabic words, as in Italian and Polish, and that, instead, it is longer in monosyllabic words when C2 is voiced. While the timing of the VC boundary within the release-to-release in disyllabic words affects both vowel and closure durations in a logically dependent way, vowel and closure durations can be modulated more independently in monosyllabic words. The less constrained operation of production and perceptual mechanisms affecting the timing of the VC boundary was argued to be the reason for the seemingly greater effect of voicing reported for monosyllabic words. The data in this study, and the cumulative evidence from previous studies as evinced by a Bayesian meta-analysis, however, do not equivocally provide support for a difference in the effect between mono- and disyllabic words, and future work is necessary to shed light on the matter.

To conclude, the results of this study suggest some directions of research. Future studies should further investigate the articulatory temporal patterns of vocalic and consonantal gestures in disyllabic words. In particular, a complete assessment of the isochrony (or lack thereof) of consecutive vocalic gestures should include a variety of

oppositions, involving voicing, place of articulation, number of consonants, syllabic affiliation, and prosodic contexts. Moreover, work is needed to shed light on the timing of the consonant closing gesture relative to the articulatory gesture of the preceding vowel in voiceless vs voiced stops. Finally, the scenario of emergence of the voicing effect offered here should be examined in relation to other consonantal effects on vowel duration, like other laryngeal effects and effects of manner of articulation.

Chapter 6

Longer vowel duration correlates with greater tongue root displacement: Acoustic and articulatory data from Italian and Polish [Paper III]

This paper has been published in the Journal of the Acoustical Society of America as: Coretta, Stefano. 2020. Longer vowel duration correlates with greater tongue root displacement: Acoustic and articulatory data from Italian and Polish. *The Journal of the Acoustical Society of America*(147). 245–259. DOI: https://doi.org/10.1121/10.0000556.
When citing, please refer to the published version.

Abstract

Voiced stops tend to be preceded by longer vowels and produced with a more advanced tongue root than voiceless stops. The duration of a vowel is modulated by the voicing of the stop that follows and in many languages vowels are longer when followed by voiced stops. Tongue root advancement is known to be an articulatory mechanism which ensures the right pressure conditions for the maintenance of voicing during closure as dictated by the Aerodynamic Voicing Constraint. In this paper, it is argued that vowel duration and tongue root advancement enter in a direct statistical relation. Draw-

ing from acoustic and ultrasound tongue imaging data from 17 speakers of Italian and Polish, it is shown that tongue root advancement is initiated during the vowel, and that vowel duration and tongue root position at vowel offset are positively correlated. Longer vowel durations correspond to greater tongue root advancement. It is further proposed that the later closure onset of voiced stops within a temporally stable interval is responsible for both greater root advancement and shorter closure durations in the context of voiced stops.

6.1 Introduction

It is well known that voiced stops are almost universally characterised by two phonetic correlates: advanced tongue root and increased duration of the preceding vowel (Westbury 1983; Lisker 1974; Fowler 1992). While a lot of work has been done on each of these aspects separately, less is known about their relation. In this paper, I propose a link between the position of the tongue root at the onset of a post-vocalic stop and the duration of the vowel preceding that stop. In an exploratory study of the articulatory correlates of stop voicing, it was found that tongue root advancement—a mechanism known to facilitate voicing during stop closure—is initiated during the production of the vowel preceding the stop. This replicates previous work on tongue root position. Furthermore, the results of this study indicate that the acoustic duration of the vowel is positively correlated with tongue root position, such that longer vowel durations correspond to greater tongue root advancement. Such correlation is shown to derive from the timing of the consonantal closure relative to the preceding vowel.

6.1.1 Tongue root position and voicing

One of the differences in supra-glottal articulation between voiced and voiceless stops concerns the position of the tongue root relative to the front-back dimension of the oral tract. It has been repeatedly observed that the tongue root is in a more front position in voiced stops compared to voiceless stops (Kent & Moll 1969; Perkell 1969; Westbury 1983). This has been attributed to the fact that the initiation and maintenance of vocal fold vibration (i.e. voicing) requires a difference in air pressure between the cavities

below and above the glottis. Specifically, the sub-glottal pressure needs to be higher than the supra-glottal pressure. In other words, there must be a positive trans-glottal air pressure differential (van den Berg 1958; Rothenberg 1967). This property of voicing is formally known as the Aerodynamic Voicing Constraint (Ohala 2011). When the oral tract is completely occluded during the production of a stop closure, the supra-glottal pressure quickly increases, due to the incoming airstream from the lungs. Such pressure increase can hinder the ability to sustain vocal fold vibration during closure, to the point voicing ceases.

An articulatory solution to counterbalance the increased pressure is to enlarge the supra-glottal cavity by advancing the root of the tongue. In the context of articulatory adjustments, a distinction between passive and active gestures is generally drawn (see for example Rothenberg 1967). A passive enlargement of the oral cavity is the product of the incoming airflow, the pressure of which expands the pliable soft tissues of the cavity walls. On the other hand, active expansion is achieved by muscular activity, which can in turn be purposive (produced with the goal of cavity expansion) or non-purposive. While Rothenberg (1967) recognises that the distinction between purposive and non-purposive active gestures can be at times blurry, it is nonetheless important to note that the qualification of a gesture as active does not automatically implies a speaker's intention to produce the obtained result.

Rothenberg (1967) further calculates that the walls of the oral tract can absorb the incoming airflow for 20 to 30 ms by passive expansion, after which the sub- and supra-glottal pressures would equalise and voicing cease. Based on these estimates, a passive expansion of the pharyngeal walls is thus not generally sufficient to maintain voicing during the closure of a stop. Reaching a complete ballistic forward gesture would require the tongue root about 70 to 90 ms (Rothenberg 1967). Given that voiced stop closures are on average shorter than that (the mean duration is about 64 ms in Luce & Charles-Luce 1985), it is expected that the movement could be initiated during the production of the vowel, so that an appreciable amount of advancement is obtained when closure is achieved. Furthermore, Westbury (1983) finds that tongue root advancement is initiated before the achievement of full closure and that there is a forward movement even in some cases of voiceless stops, although the rate and magnitude of the advance-

ment are consistently higher in voiced stops. Finally, tongue root adjustments seem to target more specifically lingual consonants, while the tongue body is more involved in labials (Perkell 1969; Westbury 1983).

However, the relation between tongue root advancement and voicing is a complex one. First, tongue root advancement is not the only mechanism for sustaining voicing during a stop (Rothenberg 1967; Westbury 1983; Ohala 2011) and it has a certain degree of idiosyncrasy (Ahn 2018). For example, a cross-lingusitically common difference between voiceless and voiced stops concerns their respective closure durations. The closure of voiced stops is generally longer than that of voiceless stops (Lisker 1957; Umeda 1977; Summers 1987; Davis & Summers 1989; de Jong 1991). A shorter closure favours maintenance of vocal fold vibration by ensuring that the pressure build-up in the oral cavity does not equalise the sub-glottal and supra-glottal pressures (at which point voicing would stop). Other solutions which can help sustaining voicing during closure include larynx lowering (Riordan 1980), slackening of the vocal folds (Halle & Stevens 1967), opening of the velopharyngeal port (Yanagihara & Hyde 1966), and producing a retroflex occlusion (Sprouse et al. 2008). Moreover, (Ahn 2018) finds that not all the speakers she surveyed did show tongue root advancement, and a few had rather the reverse pattern.

Second, implementation of tongue root advancement can be decoupled from the actual presence of vocal fold vibration. In Westbury (1983), advancement of the tongue root is found in some productions of voiceless stops. This is counterintuitive, since tongue root advancement is generally considered to be a feature of voiced stops which require voicing-related pressure adjustments. Moreover, Ahn (2015, 2018) and Ahn & Davidson (2016) looked at utterance-initial stops and found that the tongue root is more advanced in the phonologically voiced stops independent of whether they are implemented with vocal fold vibration or not.

To summarise, tongue root advancement is a common articulatory solution employed to counterbalance the increase in supra-glottal pressure and maintaining voicing during the production of at least lingual voiced stops. While this gesture is not exclusive of voiced stops and it can be implemented even in the absence of vocal fold vibration, tongue root advancement seems to be a robust correlate of voicing.

6.1.2 Vowel duration and voicing

The results discussed here are part of a larger study which focusses on the effect of consonant voicing on preceding vowel durations. A great number of studies shows that, cross-linguistically, vowels tend to be longer when followed by voiced obstruents than when they are followed by voiceless ones (House & Fairbanks 1953; Peterson & Lehiste 1960; Chen 1970; Klatt 1973; Lisker 1974; Farnetani & Kori 1986; Fowler 1992; Hussein 1994; Esposito 2002; Lampp & Reklis 2004; Durvasula & Luo 2012). This so-called "voicing effect" has been reported in a variety of languages, including (but not limited to) English, German, Hindi, Russian, Arabic, Korean, Italian, and Polish (see Maddieson & Gandour 1976 and Beguš 2017 for a more comprehensive list).

Italian and Polish offer an opportunity to study the articulatory aspects of the voicing effect, given their reported differences in magnitude/presence of the effect and the relative ease of comparison. While Italian has been consistently reported as a voicing-effect language (Magno Caldognetto et al. 1979; Farnetani & Kori 1986; Esposito 2002), some studies found an effect in Polish (Slowiaczek & Dinnsen 1985; Nowak 2006; Malisz & Klessa 2008; Coretta 2019b) while others did not (Keating 1984b; Jassem & Richter 1989).

Coretta (2019b) argues, based on the acoustics of the same data reported here, that the stressed vowels of disyllabic (CV́CV) words in Italian and Polish are 16 ms longer (SE = 4.4) when followed by a voiced stop. The high degree of intra-speaker variation, backed up by statistical modelling, also indicates that these languages possibly behave similarly in regards to the voicing effect. Finally, the temporal distance between two consecutive stop releases in CV́CV words is not affected by the voicing of the second consonant. The duration of the release-to-release interval is stable across voicing contexts. Within this interval, the timing of VC boundary (the vowel offset/onset of stop closure) produces differences in the respective durations of vowel and closure, following a mechanism of temporal compensation (Lindblom 1967; Slis & Cohen 1969a,b; Lehiste 1970a,b). A later closure onset results in a long vowel and a short closure, while an earlier closure onset corresponds to a short vowel and a long closure. Since the closure of voiceless stops is longer than that of voiced stops, it follows that vowels are shorter when followed by the former than when followed by the latter.

6.1.3 This study

Previous research has established that tongue root advancement and longer vowel durations are two common correlates of voicing. In particular, voicing during closure can be maintained by advancing the tongue root during the production of voiced stops (which is possibly initiated earlier than the closure onset) and vowels followed by voiced stops tend to be longer than vowels followed by voiceless stops. The acoustic data further revealed that the duration of the stop closure bears on the duration of the preceding vowel, by means of a kind of compensatory mechanism.

The results from the articulatory data of this study, which will be discussed in the following sections, offer new insights on the link between closure and vowel duration. We will see that the relative timing of the closure also modulates the degree of tongue root advancement found at closure onset, thus creating a three-way network of relations with vowel duration and tongue root position. More specifically, the timing of the closure onset within the release-to-release interval determines the duration of the vowel, the duration of the closure, and the degree of tongue root advancement. Finally, it will be argued that a later closure onset as in the case of voiced stops has the double advantage of producing both a short closure duration and greater tongue root advancement, features both known to comply with the Aerodynamic Voicing Constraint.

6.2 Methodology

Following recent practices which encourage scientific transparency and data attribution (Crüwell et al. 2018; Berez-Kroeker et al. 2018; Roettger 2019), data (Coretta 2018a) and analysis code (`https://osf.io/d245b/`) are available on the Open Science Framework.

6.2.1 Participants

Participants were recruited in Manchester (UK), and Verbania (Italy). Eleven native speakers of Italian (5 females, 6 males) and 6 native speakers of Polish (3 females, 3 males) participated in this study. Most speakers of Italian are originally from the North

of Italy, while 3 are from Central Italy. The Polish speakers came from different parts of Poland (2 from the west, 3 from the centre, and 1 from the east). This study has been approved by the School of Arts, Languages, and Culture Ethics committee of the University of Manchester (REF 2016-0099-76). The participants signed a written consent and received a monetary compensation of £10.

6.2.2 Equipment

Simultaneous recordings of audio and ultrasound tongue imaging were obtained in the Phonetics Laboratory at the University of Manchester (UK) or in a quiet room in Verbania (Italy). An Articulate Instruments Ltd™ system was used for this study. The system is made of a TELEMED Echo Blaster 128 unit, an Articulate Instruments Ltd™ P-Stretch synchronisation unit, and a FocusRight Scarlett Solo pre-amplifier. A TELEMED C3.5/20/128Z-3 ultrasonic transducer (20mm radius, 2-4 MHz) and a Movo LV4-O2 Lavalier microphone were used respectively for the acquisition of ultrasonic and audio data. The ultrasonic probe was placed in contact with the sub-mental triangle, aligned with the mid-sagittal plane. A metallic headset designed by Articulate Instruments Ltd™ (2008) was used to hold the probe in a fixed position and inclination relative to the head. The acquisition of the mid-sagittal ultrasonic and audio signals was achieved with the software Articulate Assistant Advanced (AAA, v2.17.2) running on a Hewlett-Packard ProBook 6750b laptop with Microsoft Windows 7. The synchronisation of the ultrasonic and audio signals was performed by AAA after recording by means of a synchronisation signal produced by the P-Stretch unit. The ranges of the ultrasonic settings were: 43-68 frames per second, 88-114 number of scan lines, 980-988 pixel per scan line, field of view 71-93°, pixel offset 109-263, depth 75-180 mm. The audio signal was sampled at 22050 Hz (16-bit).

6.2.3 Materials

Disyllabic words of the form $C_1V_1C_2V_2$ were used as targets, where C_1 = /p/, V_1 = /a, o, u/, C_2 = /t, d, k, g/, and V_2 = V_1 (e.g. *pata*, *pada*, *poto*, etc.), giving a total of 12

Table 6.1: The list of Italian and Polish target words. An asterisk indicates a real word.

Italian			Polish		
pata	poto*	putu	pata	poto	putu
pada	podo	pudu	pada*	podo	pudu
paca*	poco*	pucu	paka*	poko	puku
paga*	pogo	pugu	paga	pogo	pugu

target words, used both for Italian and Polish.[1] The resulting words are nonce words, with a few exceptions, and they were presented in the languages' respective writing conventions (see Table 6.1). A labial stop was chosen as the first consonant to reduce possible coarticulation with the following vowel.[2] Central/back vowels only were included in the target words for two reasons. First, high and mid front vowels tend to be difficult to image with ultrasound, given their greater distance from the ultrasonic probe when compared with back vowels. Second, high and mid front vowels usually produce less tongue displacement from and to a stop consonant. This characteristic can make it more difficult to identify gestural landmarks using the methodology discussed in Section 6.2.5. Since the focus of the study was to explore differences in the closing gesture of voiceless and voiced stops, only lingual consonants have been included (the closure of labial stops cannot of course be imaged with ultrasound). The sentence *Dico X lentamente* 'I say X slowly' in Italian, and *Mówię X teraz* 'I say X now' for Polish functioned as frames for the test words. Speakers were instructed to read the sentences without pauses and to speak at a comfortable pace.

6.2.4 Procedure

The participants familiarised themselves with the sentence stimuli at the beginning of the session. Headset and probe were then fitted on the participant's head. The partici-

[1] Note that stressed vowels in open syllables in Italian are long (Renwick & Ladd 2016). Moreover, /o/ is used here for typographical simplicity to indicate the mid-back vowels of Italian and Polish, although they do differ in quality. See Krämer (2009), Renwick & Ladd (2016), and Gussmann (2007).

[2] However, note that Westbury (1983) and Vazquez-Alvarez & Hewlett (2007) report tongue body lowering in the context of labial stops.

pant read the sentence stimuli, which were presented on the computer screen in a random order, while the audio and ultrasonic signals were acquired simultaneously. The random list of sentences was read 6 times consecutively (with the exception of IT02, who repeated the sentences 5 times only). Due to software constraints, the order of the sentences within participant was kept the same for each of the six repetitions. The participant could optionally take breaks between one repetition and the other. Sentences with hesitations or speech errors were immediately discarded and re-recorded. A total of 1212 tokens (792 from Italian, 420 from Polish) were obtained.

6.2.5 Data processing and statistical analysis

The audio data was subject to forced alignment using the SPeech Phonetisation Alignment and Syllabification software (SPPAS, Bigi 2015). The outcome of the automatic alignment was then manually corrected, according to the recommendations in Machač & Skarnitzl (2009). The onset and offset of V1 in the $C_1V_1C_2V_2$ test words were respectively placed in correspondence of the appearance and disappearance of higher formant structure in the spectrogram. Vowel duration was calculated as the duration of the V1 onset to V1 offset interval. Speech rate was measured as the number of syllables in the sentence (8 in Italian and 6 in Polish) divided by the duration of the sentence in seconds.

The displacement of the tongue root was obtained from the ultrasonic data according to the procedure used in Kirkham & Nance (2017). Smoothing splines were automatically fitted to the visible tongue contours in AAA. Manual correction was then applied in cases of clear tracking errors. A fan-like frame consisting of 42 equidistant radial lines superimposed on the ultrasonic image was used as the coordinate system. The origin of the 42 fan-lines coincides with the (virtual) origin of the ultrasonic beams, such that each fan-line is parallel to the direction of the nearest ultrasonic scan lines. Tongue root displacement was thus calculated as the displacement of the fitted spline along a selected vector (Strycharczuk & Scobbie 2015), see Figure 6.1. For each participant, the fan-line with the highest standard deviation of displacement within the area corresponding to the speaker's tongue root was chosen as the tongue root displacement vector. A Savitzky–Golay smoothing filter (second-order, frame length 75 ms) was

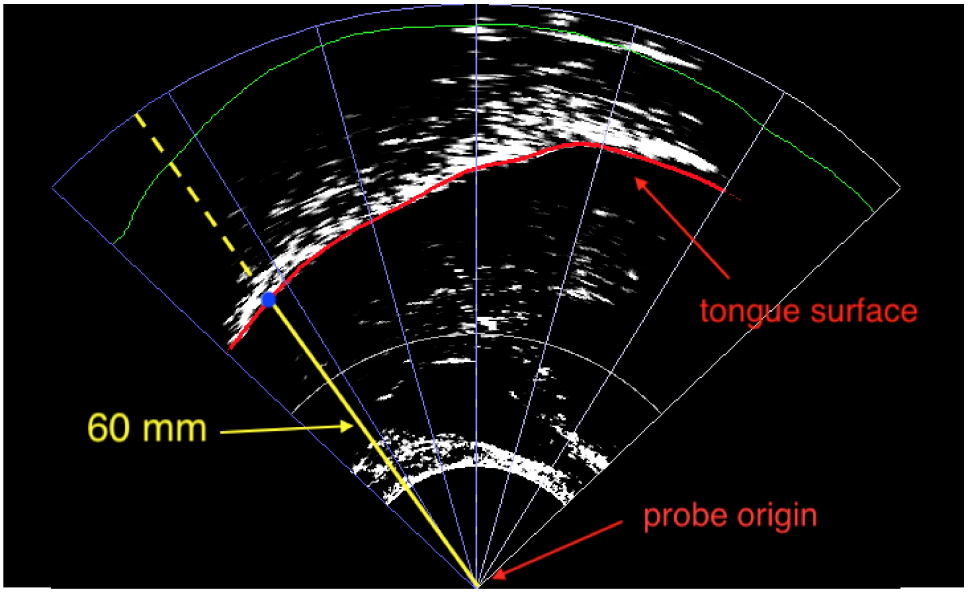


Figure 6.1: Schematics of the operationalisation of tongue root position, based on Kirkham & Nance (2017). The tongue root surface corresponds to the lower edge of the white band in the image. The tongue tip is on the right side. The outline of the fan-like coordinate systems is shown. The yellow line starting from the probe origin is the selected fan-line from which tongue root position is calculated (see text for the method of fan-line selection). Tongue root position thus corresponds to the distance (in millimetres) between the probe origin and the intersecting point of the tongue surface with the selected fan-line.

applied to the raw displacement. Displacement values for analysis are taken from the smoothed displacement signal. Tongue root displacement was obtained from a static time point (the onset of the closure of C2) and along the duration of the vowel. The displacement values along the vowel duration were extracted at time points corresponding to real ultrasonic video frames. Given the average frame rate is 55 frames per second, values are sampled about every 20 ms.

Statistical analysis was performed in R v3.5.2 (R Core Team 2018). Linear mixed-effects models were fitted with lme4 v1.1-19 (Bates et al. 2015). Factor terms were coded with treatment contrasts (the reference level is the first listed for each factor): C2 voicing (voiceless, voiced), vowel (/a/, /o/, /u/). Speech rate was centred for inclusion in the statistical models, by subtracting the mean speech rate across all speakers from the

calculated speech rate values. Centring ensures the intercepts are interpretable. *t*-tests with Satterthwaite's approximation to degrees of freedom on the individual terms were used to obtain *p*-values using lmerTest v3.0-1 (Kuznetsova et al. 2017; Luke 2017). An effect is considered significant if the *p*-value is below the alpha level ($\alpha = 0.05$). Generalised additive mixed models were fitted with mgcv v1.8-26 (Wood 2011, 2017). The smooths used thin plate regression splines as basis (Wood 2003). The ordered factor difference smooths method described in Sóskuthy (2017) and Wieling (2018) was used to model the effect of factor terms in GAMs. The models were fitted by maximum likelihood (ML) and autoregression in the residuals was controlled with a first-order autoregressive model.

Significance testing of the relevant predictors was achieved by comparing the ML score of the full model with the score of a null model (in which the relevant predictor is dropped), using the `compareML()` function of the itsadug package (van Rij et al. 2017). A preliminary analysis indicated that including either language or C2 place of articulation as predictors produced respective *p*-values above the alpha level, without affecting the estimates of the other terms. Section 6.4.3 further discusses the idiosyncratic behaviour of the tongue root observed between speakers, which does not seem to pattern in any way with their native language. For these reasons, these variables were not included in the models reported here and will not be discussed. Future research is warranted to ascertain language-related differences and possible effects of place of articulation.

6.3 Results

6.3.1 Tongue root position at C2 closure onset

Figure 6.2 shows raw data points and boxplots of the position of the tongue root at C2 closure onset when C2 is voiceless (left) and voiced (right). Since the position of the tongue root in millimetres depends on the speaker's anatomy and on the probe location, scaled tongue root position is used in this plot (note though that the unscaled data is used in statistical modelling). As a trend, the position of the tongue root is more advanced if

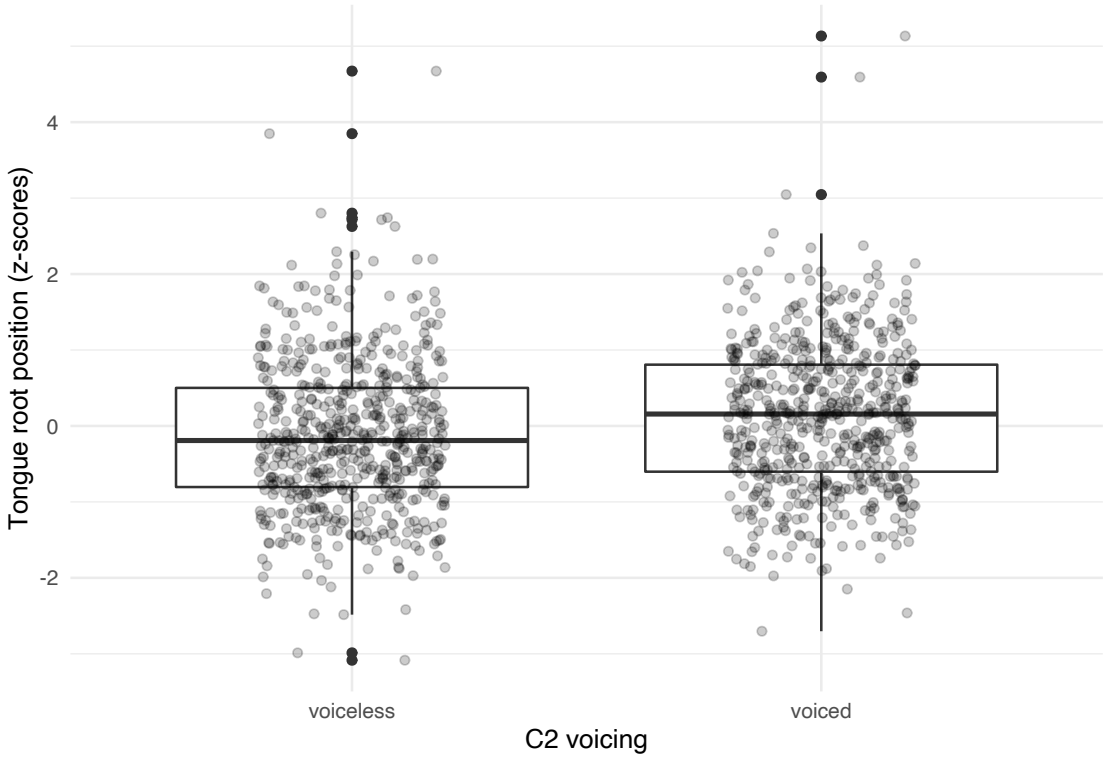


Figure 6.2: Raw data and boxplots of tongue root position in voiceless and voiced stops at closure onset. Higher values indicate advancement.

Table 6.2: Summary of the linear mixed-effects model fitted to tongue root position at vowel offset (see Section 6.3.1)

Predictor	Estimate	SE	CI low	CI up	df	t-value	p-value	< α
Intercept	-62.1396	1.8113	-65.6898	-58.5895	17.1188	-34.3058	0.0000	*
Voicing = voiced	0.7689	0.3473	0.0881	1.4497	19.3947	2.2137	0.0390	*
Speech rate (centr.)	0.4114	0.2793	-0.1360	0.9588	1168.1100	1.4732	0.1410	
Vowel = /o/	-1.8742	0.4249	-2.7069	-1.0414	19.2874	-4.4112	0.0003	*
Vowel = /u/	0.0865	0.4270	-0.7503	0.9233	19.6974	0.2027	0.8415	

C2 is voiced compared to its position when C2 is voiceless.

A linear mixed-effects model with tongue root position as the outcome variable was fitted with the following predictors (Table 6.2): fixed effects for C2 voicing (voiceless, voiced), centred speech rate (as number of syllables per second, centred), vowel (/a/, /o/, /u/); by-speaker and by-word random intercepts (a by-speaker random coefficient for C2 voicing led to singular fit, so it was not included in the final model). The effects of C2 voicing and vowel are significant according to *t*-tests with Satterthwaite's approximation to degrees of freedom. The tongue root at C2 closure onset is 0.77 mm (SE = 0.35) more front when C2 is voiced, and it is 1.87 mm (SE = 0.42) more retracted if V1 is /o/.

6.3.2 Tongue root position during V1

The position of the tongue root during the articulation of V1 was assessed with generalised additive mixed models (GAMM). A GAMM was fitted to tongue root position with the following terms (Table 6.3): C2 voicing as a parametric term; a smooth term over centred speech rate, a smooth term over V1 proportion with a by-C2 voicing difference smooth, a tensor product interaction over V1 proportion and centred speech rate; a factor random smooth over V1 proportion by speaker (penalty order = 1). A chi-squared test on the ML scores of the full model and a model excluding C2 voicing indicates that C2 voicing significantly improves fit ($\chi(3) = 7.758$, $p = 0.001$). Figure 6.3 shows that the root advances during the production of the vowel, relative to its position at V1 on-

Table 6.3: Summary of the GAM model fitted to tongue root position during V1 (see Section 6.3.2)

Predictor	Estimate	SE	EDF	Ref.DF	Statistic	p-value	< α
Intercept	-63.3328	1.7562			-36.0623	0.0000	*
Voicing = voiced	0.3311	0.1432			2.3122	0.0208	*
s(Speech rate (centr.))			7.5310	8.5159	4.4781	0.0000	*
s(Proportion)			3.6906	4.3631	10.4450	0.0000	*
s(Proportion): voiced			1.0121	1.0233	9.8423	0.0015	*
ti(Proportion, Speech Rate (c.))			2.1298	2.7632	2.9030	0.0429	*
s(Proportion, Speaker)			62.2802	152.0000	57.3447	0.0000	*

set. This forward movement is observed both in the context of a following voiced stop and in that of a following voiceless stop. However, the magnitude of the movement is greater in the former. At V1 offset (= C2 closure onset), the graph suggests a difference in tongue root position of about 1 mm.

6.3.3 Correlation between tongue root position and V1 duration

A second linear mixed regression was fitted to tongue root position to assess the effect of V1 duration on root position (Table 6.4). The following terms were included: centred V1 duration (in milliseconds), centred speech rate (as number of syllables per second), vowel (/a/, /o/, /u/), C2 place of articulation (coronal, velar); an interaction between centred V1 duration and vowel; by-speaker and by-word random intercept (a by-speaker random coefficient for V1 duration led to non-convergence, so it was not included in the final model). All predictors and the V1 duration/vowel interaction are significant. V1 duration and tongue root position are positively correlated: The longer the vowel, the more advanced the tongue root is at V1 offset ($\hat{\beta}$ = 0.065 mm, SE = 0.007). The effect is stronger with /a/ than with /o/ and /u/ (see Figure 6.4).

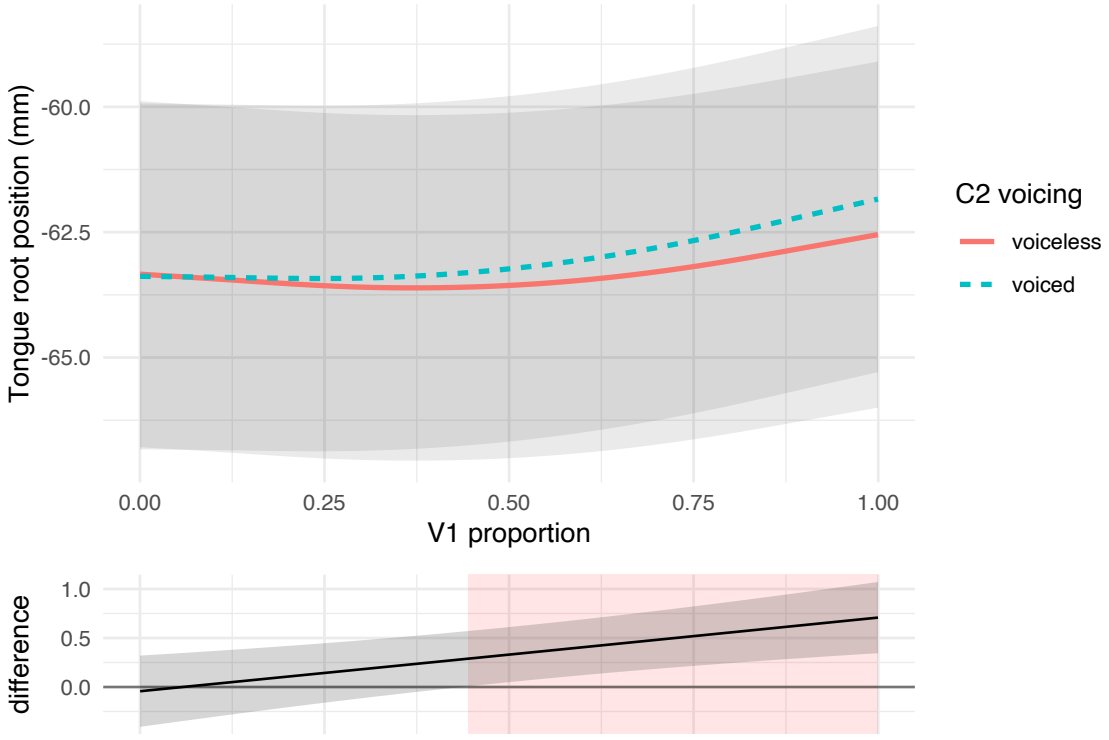


Figure 6.3: Predicted tongue root position (top figure) during vowels preceding voiceless and voiced stops, with 95% confidence intervals, and difference smooth (bottom figure). Higher values of tongue root position indicate a more advanced root. The shaded red area in the difference smooth indicates where the two curves are different. Predictions from a GAMM (see Section 6.3.2).

Table 6.4: Summary of the linear mixed-effects model for testing the correlation between tongue root position and V1 duration (see Section 6.3.3)

Predictor	Estimate	SE	CI low	CI up	df	t-value	p-value	< α
Intercept	-62.5793	1.7818	-66.0716	-59.0870	17.0874	-35.1212	0.0000	*
V1 duration (centr.)	0.0651	0.0073	0.0507	0.0795	955.6436	8.8558	0.0000	*
Speech rate (centr.)	1.2412	0.2903	0.6722	1.8102	1169.6885	4.2755	0.0000	*
Vowel = /o/	-1.3031	0.4597	-2.2040	-0.4021	18.3761	-2.8348	0.0108	*
Vowel = /u/	1.5863	0.5049	0.5967	2.5759	25.8255	3.1419	0.0042	*
V1 duration × /o/	-0.0303	0.0079	-0.0457	-0.0149	736.2314	-3.8504	0.0001	*
V1 duration × /u/	-0.0227	0.0090	-0.0403	-0.0052	751.2493	-2.5345	0.0115	*

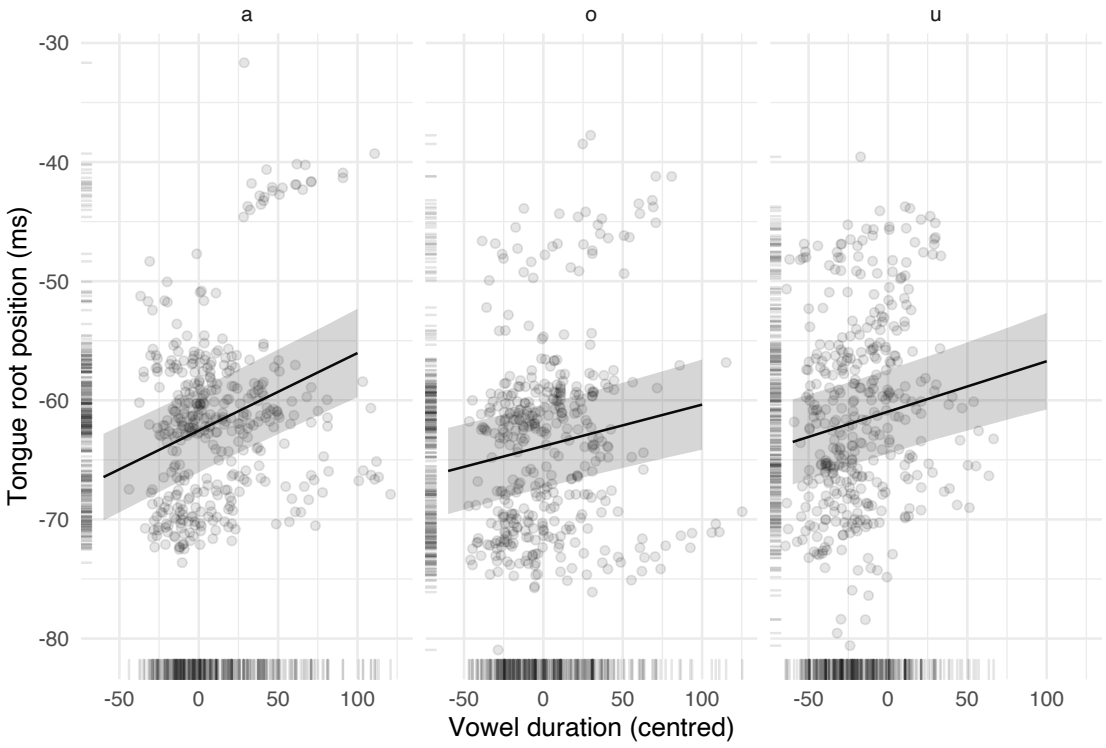


Figure 6.4: Raw data, regression lines, and 95% confidence intervals of the correlation between vowel duration and tongue root position for each vowel (/a/, /o/, and /u/). The regression line and confidence intervals are from a mixed-effects model (see Section 6.3.3).

Table 6.5: Summary of the GAM model fitted to tongue root position during V1 as a function of V1 duration (see Section 6.3.4)

Predictor	Estimate	SE	EDF	Ref.DF	Statistic	p-value	< α
Intercept	-63.0629	1.7397			-36.2484	0	*
s(V1 duration)			8.2418	8.8459	7.7096	0	*
s(Proportion)			3.9629	4.7052	17.9985	0	*
ti(Proportion, V1 duration)			2.8556	3.3236	8.9782	0	*
s(Proportion, Speaker)			59.9508	152.0000	65.7394	0	*

6.3.4 Tongue root position during V1 as a function of V1 duration

The effect of V1 duration on tongue root position during V1 was modelled by fitting a GAMM with the following terms (Table 6.5): tongue root position as the outcome variable, smooth terms over V1 duration and V1 proportion, a tensor product interaction over V1 proportion and V1 duration; a factor random smooth over V1 proportion by speaker (penalty order = 1). The full model with the tensor product interaction over V1 proportion and V1 duration has better fit according to model comparison with a model without the interaction ($\chi(3) = 12.559$, $p < 0.001$). Figure 6.5 shows the estimated root trajectories at four values of vowel duration. The general trend is that the forward movement of the root during the vowel is greater the longer the duration of the vowel (Figure 6.5). Moreover, the trajectory curvature increases with vowel duration: Shorter vowels have a flatter trajectory of tongue root advancement.

6.4 Discussion

6.4.1 Voicing, tongue root position and vowel duration

The results of this study of voicing and vowel duration in Italian and Polish revealed a few patterns in the relation between consonant voicing, tongue root position, and vowel duration. Unsurprisingly, the position of the tongue root at vowel offset is more front when the following stop is voiced than when the following stop is voiceless in both surveyed languages. This finding aligns with the results of previous work on English

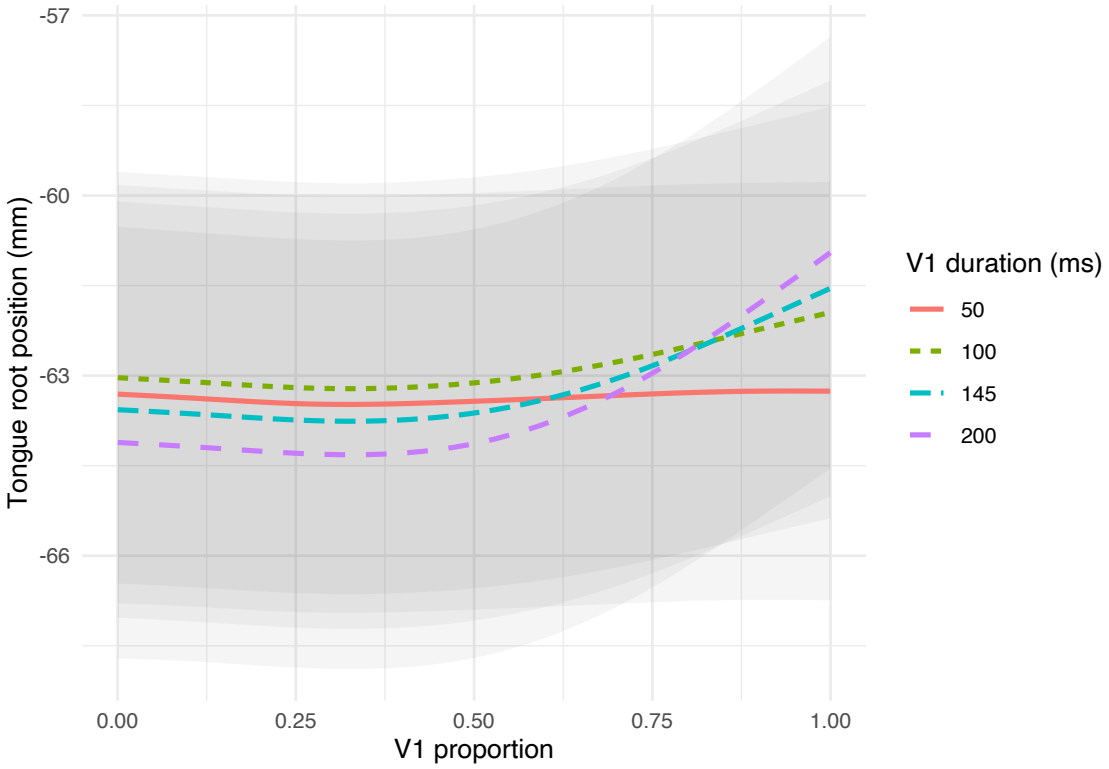


Figure 6.5: Predicted tongue root position during vowels at 4 exemplifying values of vowel duration, with 95% confidence intervals. Predictions from a GAMM (see Section 6.3.4).

(Kent & Moll 1969; Perkell 1969; Westbury 1983; Ahn 2018). When looking at the position of the tongue root during the vowel, it was found that the root starts advancing during the articulation of the vowel. Westbury (1983) found the same pattern in English. Moreover, similarly to the results in Westbury (1983), some tongue root advancement during the production of the vowel is found even when C2 is voiceless.

A possible reason for the presence of such a small degree of advancement in voiceless lingual stops is offered by arguments in relation to the absence of advancement in labials (voiced or voiceless). Westbury (1983) proposes that the articulation of the closure of lingual stops mechanically involves movements of the tongue root, so that, in order to keep a constant oral cavity volume, the root moves forward while the tongue body moves upward. On the other hand, the tongue can move freely in labial stops since their closure involves the lips. This idea is supported by the "trough effect" (Vazquez-Alvarez & Hewlett 2007), i.e. VCV sequences involving a labial stop show tongue body lowering, and by the fact that voiced labials tend to resort to tongue body lowering rather than tongue root advancement as a mechanism for voicing maintenance (Perkell 1969; Westbury 1983; Ahn 2018). The small degree of advancement in voiceless lingual stops could then as well be a mechanic consequence of the tongue moving upward for producing the stop closure.

The data discussed here also suggest that tongue root position is positively correlated with vowel duration, such that longer vowels show a more advanced tongue root at vowel offset (= closure onset) than shorter vowels. Said correlation exists independent of the voicing status of the consonant following the vowel (compatible with the finding that even voiceless stops have some degree of advancement). The correlation between tongue root and vowel duration could be interpreted as to indicate that the onset of the forward gesture of the root is timed relative to a landmark preceding the closure, independent of the duration of the vowel. The timing of the stop closure along the advancement movement would sanction the degree of advancement found at closure onset.

The dynamic data of tongue root advancement during the articulation of the vowel indicates that vowels followed by voiced stops have greater tongue root advancement at vowel offset than vowels followed by voiceless stops, in accordance with the results

from the static analysis at vowel offset. Moreover, a significant interaction between vowel duration and the trajectory shape was found. Shorter vowels have a flatter trajectory, while the curvature of the trajectory in longer vowels is greater.

When comparing the effects of vowel duration and speech rate on tongue root position, though, we are faced with a paradox. Both variables have a positive effect on tongue root position, so that longer vowels and higher speech rates imply a more advanced root. However, speech rate has a negative effect on vowel duration (and segments duration in general), such that higher speech rates are correlated with shorter vowel durations (this holds for this data). If higher speech rates mean shorter vowels and shorter vowels imply a less advanced root, we should also find less advancement with higher speech rates. However, the results indicate the opposite, and higher speech rates are correlated with more root advancement. A regression model on the position of the tongue root at *vowel onset* suggests that speech rate is positively correlated with tongue root position at vowel onset. The tongue root is already in a more advanced position at vowel onset when the speech rate is high, so that, if vowel duration is held constant, more advancement is expected at vowel offset with higher speech rates even when higher speech rate has a negative effect on vowel duration.

The articulatory patterns observed in this paper contribute to the understanding of the acoustic patterns discussed in previous work. If we take the release of the consonant preceding the vowel as a reference point, a delayed consonant closure could ensure that, by the time closure is made, an appreciable amount of tongue root advancement is achieved. Other things being equal, an increase in cavity volume increases the time required to reach trans-glottal pressure equalisation, which would cause cessation of voicing. This mechanism thus contributes to the maintenance of voicing during the stop closure.

The closure of voiced stops is achieved later (relative to the preceding consonant release) compared to the closure of voiceless stops. Moreover, the temporal distance between the releases of the two consecutive stops in CV́CV words is not affected by the voicing category of the second stop. Given the stability of the release-to-release interval duration, the delay in producing a full closure seen in the context of voiced stops has thus a double advantage: (1) A greater degree of tongue root advancement is

achieved at vowel offset/closure onset, and (2) the stop closure is shorter. Both of these articulatory features are compliant with the requirements dictated by the Aerodynamic Voicing Constraint. A more advanced tongue root ensures that the trans-glottal pressure differential is sufficient for voicing to be sustained, and a shorter closure reduces the pressure build-up during the stop closure. To conclude, it is proposed that the combined action of a temporally stable release-to-release interval and the differential timing of the VC boundary in the context of voiceless vs voiced stops contribute to both the acoustic patterns of vowel and closure duration and the articulatory patterns of tongue root position.

6.4.2 Estimates of tongue root displacement

It is worth to briefly discuss the estimated difference in tongue root position between voiceless and voiced stops and its significance. The estimated magnitude of such difference is 0.77 mm (SE = 0.35). The 95% confidence interval for the difference is approximately within the range 0-1.5 mm. Rothenberg (1967) argues that the anterior wall of the lower pharynx (corresponding to the tongue root) can move by 5 mm along the antero-posterior axis. Figure 1 in Kirkham & Nance (2017) suggests that the tongue root of one of the Twi speakers recorded is about 4 mm more front in /e/ (a +ATR vowel) than in /ɛ/ (a -ATR vowel). Given that a difference of 4 mm in root position can produce a substantially distinct acoustic output in vowels (like the two different phonemes of Twi), it makes sense to expect that differences in tongue root position as driven by consonantal factors should be of some magnitude smaller, like the ones found in this study. Moreover, the data presented here indicates that for every millisecond increase in vowel duration there is a 0.065 mm increase in tongue root advancement (see Section 6.3.3). If a maximal ballistic forward movement of the tongue root takes between 70 and 90 ms (Rothenberg 1967), we can calculate the maximum plausible displacement to be between 4.55 to 5.85 mm (0.065 mm times 70–90 ms). These values are in agreement with the maximum root displacement of 5 mm estimated by Rothenberg.

The results of this study also shed some light on timing aspects of tongue root advancement. As mentioned in the previous section, the correlation between tongue root position and vowel duration could be a consequence of the timing of the advancement

gesture. In order to obtain such correlation, the onset of the gesture (during the articulation of the vowel) should be at a fixed distance from an earlier reference point (like the vowel onset or the preceding consonant offset) such that the timing of consonant closure will create the correlation seen in the data. Although ideally the timing of the onset of the advancing gesture should be fixed, the velocity of the gesture itself could be different depending on the voicing of the following consonant. It is possible that the velocity will be greater in the context of voiced stops, especially if the advancing gesture in this context is executed with greater muscular force. Unfortunately, a preliminary screening of the current data was inconclusive as to whether timing and velocity are similar or differ in the voiceless and voiced contexts, due to the difficulty in identifying the onset of the advancing gesture. Further data should be collected with the aim of testing the hypothesis that the timing of the gesture onset is the same in voiceless and voiced contexts, while the velocity of the gesture should differ.

Although the results of this study are in agreement with previous work, the correlation between tongue root position and vowel duration needs to be replicated by expanding the enquired contexts to other types of consonants and vowels, and with other languages. Investigating the relative phasing of tongue root and body gestures in lingual and labial consonants is also necessary to clarify the mechanisms that could underlie the gestural timing of stop closure and tongue root advancement. Moreover, while the paper so far has focussed on group-level trends, it should be noted that, as found in other studies on the tongue root, individual speakers show a somewhat high degree of variability. The following section discusses this point.

6.4.3 Individual differences

The results presented in Section 6.3 and discussed in Section 6.4 are group-level patterns of the population sampled in the present study. However, the data is characterised by a certain degree of individual-level differences. Figure 6.6 shows two slope plots of mean tongue root position depending on C2 voicing for each speaker. In each plot, the two means of each speaker are linked by a line that shows the difference (or lack thereof) in means. Solid lines are Italian speakers, while dashed lines are Polish speakers. The *y*-axis of the left plot is the raw mean position in millimetres, while that of the

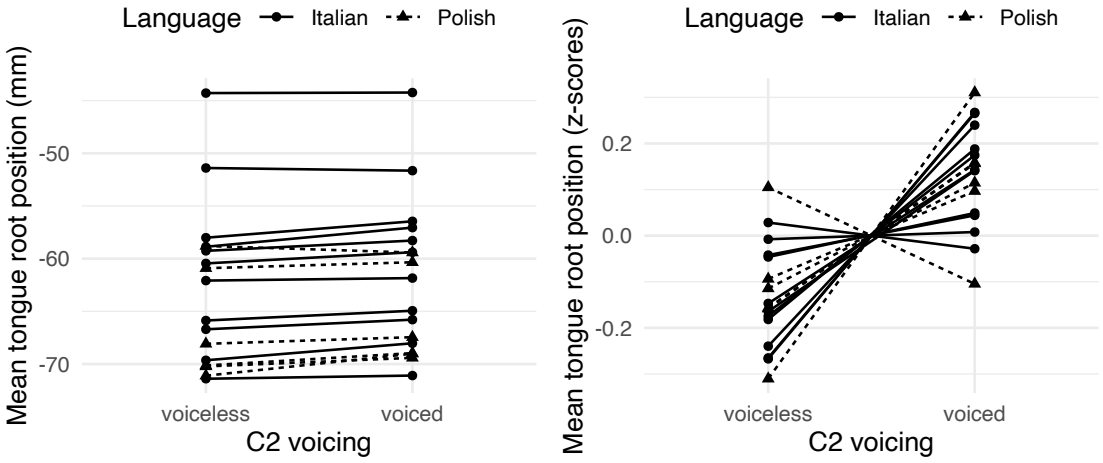


Figure 6.6: Slope plots of mean torngue root position in voiceless and voiced stops at closure onset, by-speaker. The plot on the left has raw position values in millimetres, while the plot on the right shows standardised values (z-scores) by speaker. See text for details.

right plot is the standardised values (z-scores) of the mean position. An upward-slanted slope line indicates that the mean tongue root position in the voiced condition is higher, while a downward-slanted slope is interpreted as a decrease in mean root position. A flat slope suggests there is no difference in means between the voiceless and voiced condition.

This plot shows that all three possibilities of slope direction are found in the data. The mean value of tongue root position of a voiced C2 relative to that of a voiceless stop is greater in some speakers, smaller in others, and similar in yet other speakers. Moreover, no discernible pattern can be found between speakers of Italian and Polish. Speakers of both languages show more or less the same range of variation. However, as we have seen in Section 6.3, the estimated overall effect of C2 voicing is robust and it implies a more advanced tongue root in voiced stops. The right plot of Figure 6.6 confirms this point visually. Two speakers show a declining slope (one is Italian and the other Polish), one speaker has a virtually flat slope, while all the others have a slope increasing at varying degrees. Note that the individual variation across speakers found in this data is qualitatively comparable to that in Ahn (2018).

The mean difference in tongue root position in voiceless vs voiced stops has been calculated for each speaker from the raw data. Figure 6.7 plots the speakers' mean dif-

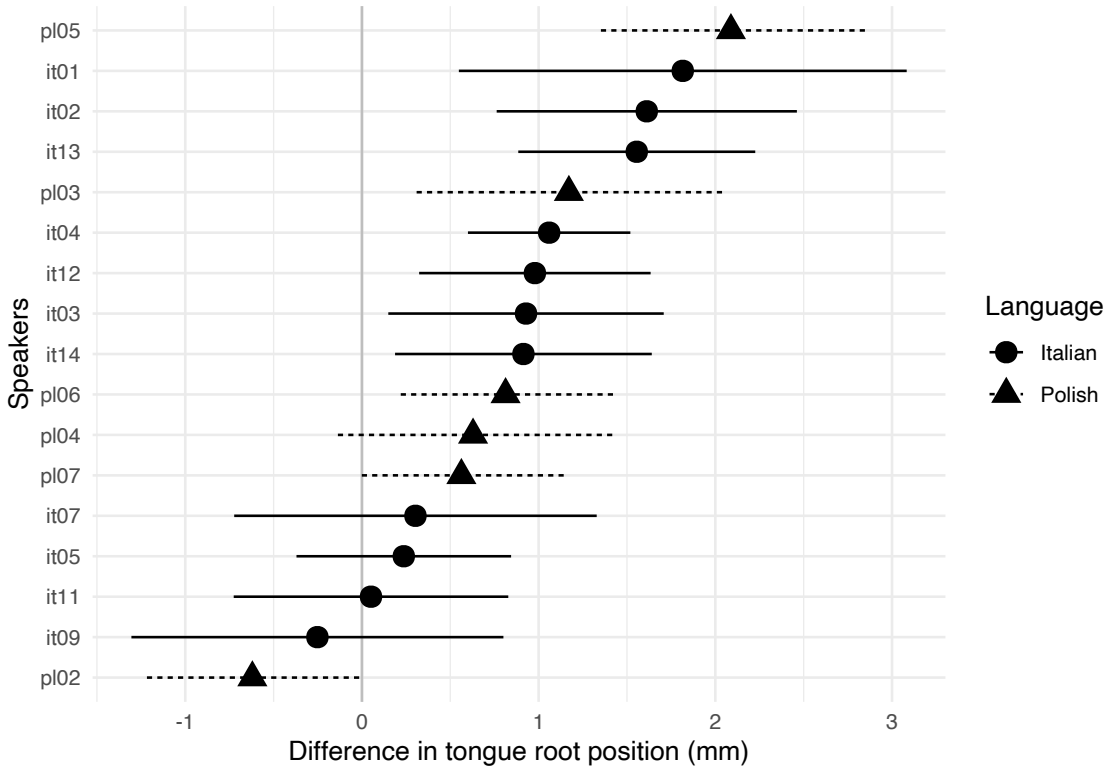


Figure 6.7: By-speaker raw mean difference in tongue root position between voiceless and voiced stops at closure onset (in millimetres). The horizontal segments are the standard errors of the mean differences.

ferences, with the respective standard error bars. The bottom 7 speakers (3 Polish, 4 Italian) show either a weak negative difference (the tongue root is slightly more advanced in voiceless stops) or a weak positive difference with wide standard errors which include 0. The remaining 11 speakers have a more robust positive difference (the tongue root is more advanced in voiced stops). Finally, speakers of each language do not cluster together, reiterating the observation made above that language does not seem to be an informative parameter.

Finally, interesting individual patterns can also be seen in the trajectories of tongue root position. Figure 6.8 shows these trajectories for all the speakers (note that the *y*-axis of each plot is on a different scale, so magnitude comparisons should not be made visually). Speakers IT01, IT03, and PL04 in particular have a somewhat categorical distinction in tongue root position during vowels followed by voiceless vs voiced stops. Such tongue root distinction is implemented across the total duration of the vowel, rather than towards the end (as suggested by the results from the aggregated data, see Section 6.3.2). The phonological literature reports cases in which the difference in tongue root position in vowels is enhanced, leading to phonological alternations or diachronic loss of the voicing distinction with maintenance of the tongue root distinction (see Vaux 1996 and references therein). The ultrasound data from this study offers articulatory evidence for a possible precursor of said phonological patterns.[3]

6.5 Conclusion

The maintenance of voicing during the closure of stops can achieved through a variety of articulatory mechanisms. Among these, shorter closure durations and cavity expansion by tongue root advancement are commonly observed solutions. Another robust correlate of consonant voicing is longer preceding vowel duration. This paper discussed articulatory data from an exploratory study of the effect of voicing on vowel duration first introduced in Coretta (2019b). Similarly to what previously found for English, the

[3]All the examples in Vaux (1996) are on vowels *following* voiceless vs voiced stops, rather than preceding, as in the current study. While beyond the scope of this paper, whether this is a systematic gap or not and how this relates to the present findings should be examined in future work.

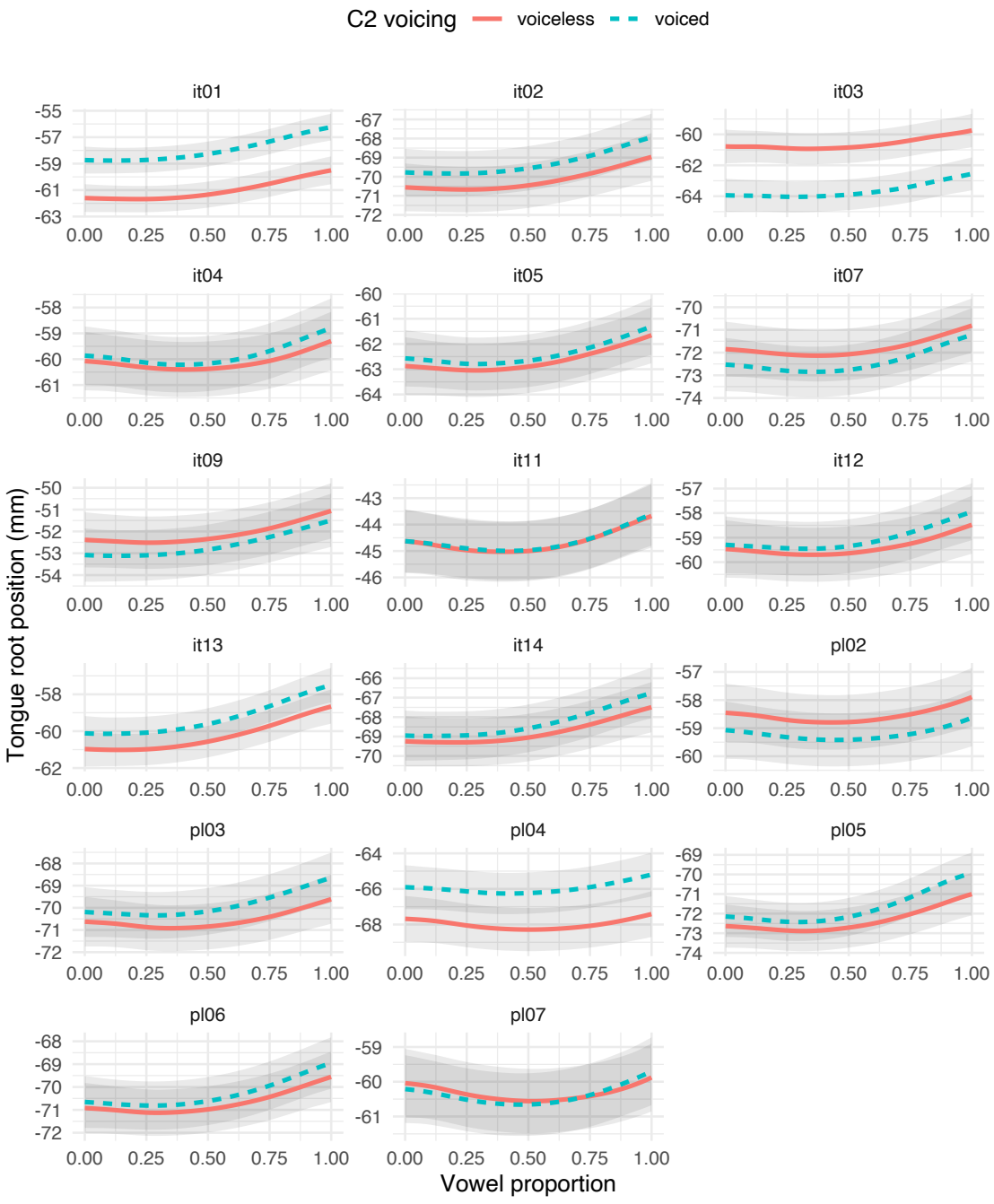


Figure 6.8: Predicted tongue root position during vowels followed by voiceless and voiced stops for each speaker. Predicted from a GAMM (see text). Note the different scales on the y-axis.

tongue root at stop closure onset is more advanced in voiced than in voiceless stops in Italian and Polish. The average difference in tongue root position is 0.77 mm (SE = 0.35). By modelling the trajectory of the tongue root during the production of vowels preceding stops, it was found that the root starts advancing during the vowel, both preceding voiceless and voiced stops. The magnitude of the advancing gesture was however greater in the voiced context. Moreover, tongue root position and vowel duration were found to be positively correlated. Longer vowel durations correspond to greater tongue root advancement.

It was argued that the combined action of two factors contribute to the patterns observed: (1) The duration of the interval between two consecutive releases, and (2) the timing of the C2 closure onset within such interval. The release-to-release interval duration has been found not to be affected by the voicing of the second consonant. The later closure onset of voiced stops within the release-to-release interval (compared to voiceless stops) has the double advantage of producing a shorter closure duration and ensuring that enough tongue root advancement is reached by the time the stop closure is achieved. Both of these aspects comply with the Aerodynamic Voicing Constraint (Ohala 2011) by delaying trans-glottal pressure equalisation (which would prevent vocal fold vibration). Future studies will need to test whether these findings replicate in Italian and Polish, and if they extend to other languages and contexts. In particular, further work on the relative differences in timing and velocity of the closing gesture and the root advancement gesture will be necessary to obtain a more in-depth understanding of the relation between consonant voicing, tongue root position, and vowel duration.

Chapter 7

Modelling electroglottographic data with wavegrams and generalised additive mixed models [Paper IV]

Coretta, Stefano. 2019. Modelling electroglottographic data with wavegrams and generalised additive mixed models. Manuscript. DOI: https://doi.org/10.31219/osf.io/m623d.

Abstract

While electroglottography is a practical and safe technique for obtaining articulatory data on voicing, statistical analysis of the signal it returns poses a few challenges given the highly dimensional nature of the signal. The wavegram has been proposed as a visualisation method which overcomes the limitations of reducing the complex electroglottographic signal to a single measure like the contact quotient. This paper introduces a method for modelling dynamic electroglottographic data based on the wavegram using generalised additive models (GAMs). Results from a pilot study which assesses the reliability of this method by comparing sustained modal and breathy phonation are presented. The applicability of wavegram GAMs is exemplified with the discussion of an exploratory study on the dynamical properties of voicing in vowels followed by voiceless and voiced stops in Italian and Polish. Increasing average glottal opening can be observed in the second half of vowels, although the timing and magnitude of the

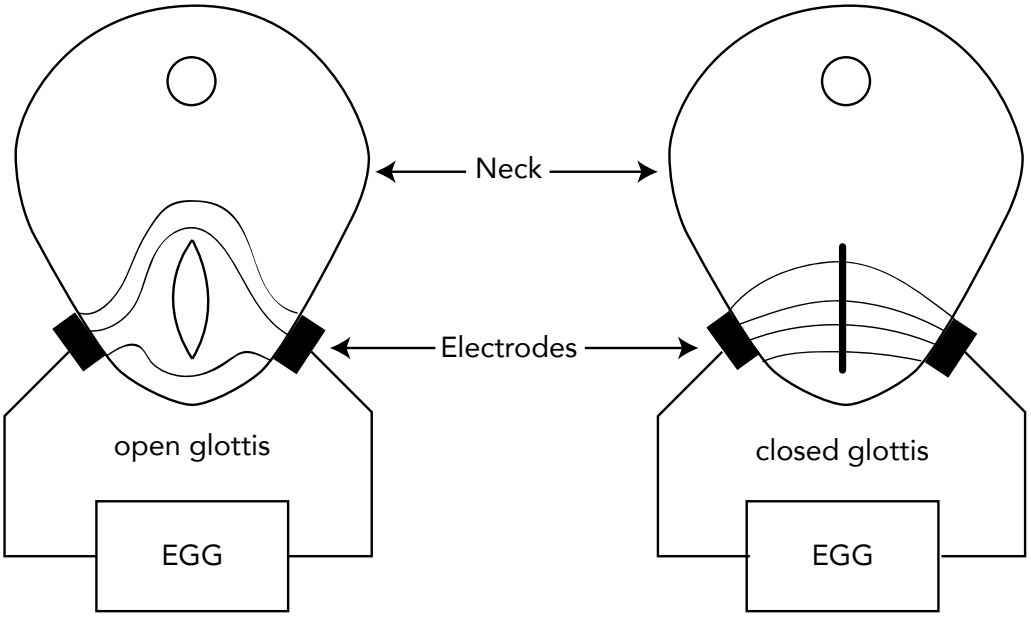


Figure 7.1: A schematics of the electroglottograph. A transverse section of the neck is shown with open glottis (on the left) and closed glottis (on the right). The electric field passing through the neck is represented by lines. When the vocal folds are apart, the opening distorts the electric field and impedance increases.

increase differs depending on the voicing of the following stop and on the language. Insights on the diachronic development of pre-aspiration based on these results are also discussed.

7.1 Introduction

The location of the vocal folds within the oral tract makes investigation of glottal activity difficult. Direct observation of the the larynx via invasive methods like laryngoscopy has the practical drawbacks of being of great discomfort to the participant and of requiring medical expertise to be performed. Electroglottography, on the other hand, is a non-invasive and safe technique which enables researchers to obtain an indirect account of glottal activity. However, the complexity of the electroglottographic signal and the imperfect mapping between the signal and the laryngeal activity it measures

pose some analytical challenges deriving from the reduction of the signal to a single measure. The wavegram technique has been proposed as a visualisation form of electroglottographic data which maintains the multi-dimensionality of the signal (Herbst et al. 2010). In this paper, generalised additive modelling (Hastie & Tibshirani 1986) is proposed as a means to statistically investigate glottal activity as derived from wavegrams.

Electroglottography, or EGG (Fabre 1957), is a technique that measures the degree of contact between the vocal folds (the Vocal Folds Contact Area, VFCA). A high frequency low voltage electrical current is sent through two electrodes which are in contact with the surface of the neck, one on each side of the thyroid cartilage (Figure 7.1). Impedance of this current is modulated by the VFCA, and greater vocal folds contact creates less impedance. The amplitude of the current is inversely correlated with VFCA and impedance, so that higher amplitude values indicate a greater contact area (Titze 1990). The EGG unit registers the current impedance and converts it to relative amplitude values. The time-developing amplitude signal thus provides us with information on the changes in VFCA, i.e. on properties of vocal folds vibration (voicing).

A glottal cycle can be divided in two phases (Childers & Krishnamurthy 1985; Hampala et al. 2016): (a) a contacting phase, in which the vocal folds are approaching each other, and (b) a de-contacting phase, in which the vocal folds move apart from each other. Transversal to this two-phase representation, the glottal cycle can be described in terms of whether the glottis is closed or not. According to this classification, the cycle can be divided into (1) a closed phase, in which the glottis is completely closed and glottal flow is 0 (in some contexts this phase could be absent, like in breathy voicing), and (2) an open phase, in which there is no complete contact between the vocal folds. The timing of these phases can be approximated from the EGG signal, as demonstrated by both experimental and modelling work (Hampala et al. 2016). An example EGG signal is provided in Figure 7.2.

Two important landmarks of glottal movement are the closing instant (the timepoint of glottal complete closure) and the opening instant (the moment in which the glottis first opens). These points delimit the open and closed phases of a glottal cycle. The ratio of the closed phase relative to the total cycle duration, the closed quotient, has

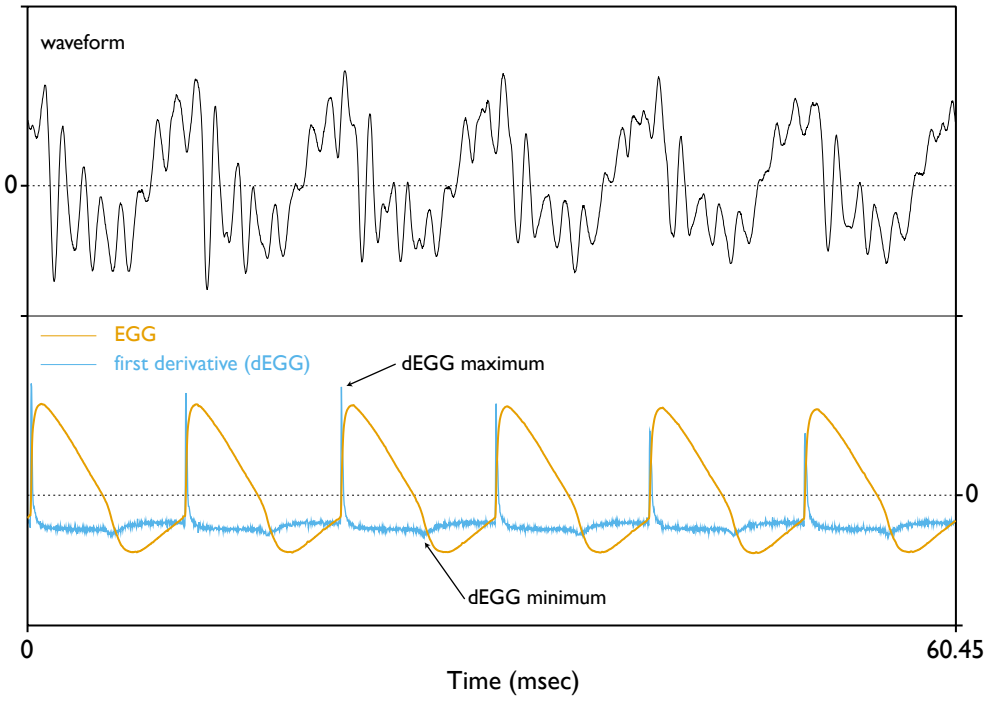


Figure 7.2: The electroglottographic signal (EGG) with corresponding first derivative (dEGG).

been used as an index of phonation type (Scherer & Titze 1987). Modal voice has higher closed quotient values than breathy voice, and lower values than creaky voice. One method for the detection of the closing and opening instants is based on the first derivative of the EGG signal (the dEGG, see Figure 7.2). Herbst et al. (2017), however, showed that this method returns values that are just a surrogate of the actual articulatory movements, due to the complex anatomy of the vocal folds, and that there are no clear contacting and decontact instants, but rather intervals. Herbst et al. (2017) call this EGG-based closed quotient the "contact quotient" and recommend to keep it distinct from the closed quotient obtained from direct observation of the vocal folds.

As an alternative to the contact quotient, Herbst et al. (2010) propose the wavegram, a visualisation method which does not reduce the EGG signal to a single value and thus suffers less from the limitations of the contact quotient. The wavegram incorporates information from the whole signal to obtain an image of vocal folds activity. A wavegram is a 3D representation of the EGG signal developing in time. Its structure is similar to that of a classical phonetic spectrogram. The *x*-axis indicates the temporal sequence of

individual glottal cycles. The *y*-axis represents the time within each glottal cycle, normalised between 0 and 1. Finally, the normalised amplitude of the signal corresponds to different colour intensities. Differences in intensity along the *x*-axis indicate changes in glottal activity. The procedure for constructing a wavegram is given in Figure 7.3. A wavegram can be produced for the EGG signal and for any of its transformations, like the dEGG.

The wavegram method was originally intended for a qualitative analysis based on visual inspection. However, wavegram data can be modelled using generalised additive models (GAMs, Hastie & Tibshirani 1986; Zuur 2012; Wood 2017). GAMs are a family of generalised models which can fit non-linear effects by additive combinations of smoothing splines. The flexibility of GAMs allows researchers to generate a fitted wavegram based on data from multiple repetitions of a single speaker and from multiple speakers. Random effects can also be included to account for idiosyncratic differences. Moreover, the potential for overfitting is reduced by a smoothing penalty parameter, which constraints the maximum number of basis functions used to construct the smoothing splines. This paper introduces wavegram GAMs as a way to quantitatively assess wavegram data. First, Section 7.2 presents results from a pilot study which informally evaluates the reliability of the proposed method. Section 7.3 illustrates how to conduct a dynamical wavegram GAM analysis of dEGG data through a practical example in which the wavegrams of vowels followed by voiceless and voiced stops in Italian and Polish are compared. This analysis indicates the presence of a pattern of glottal spreading in the second half of vowels followed by voiceless stops and of cross-linguistic differences. Insights on the diachronic emergence of pre-aspiration based on these results are discussed. Finally, Section 7.4 concludes and discusses the limitations of the current implementation of the method and future directions. The data (Coretta 2017, 2018a) and analysis scripts of this paper (`https://osf.io/3w8gh/`) can be found on the Open Science Framework.

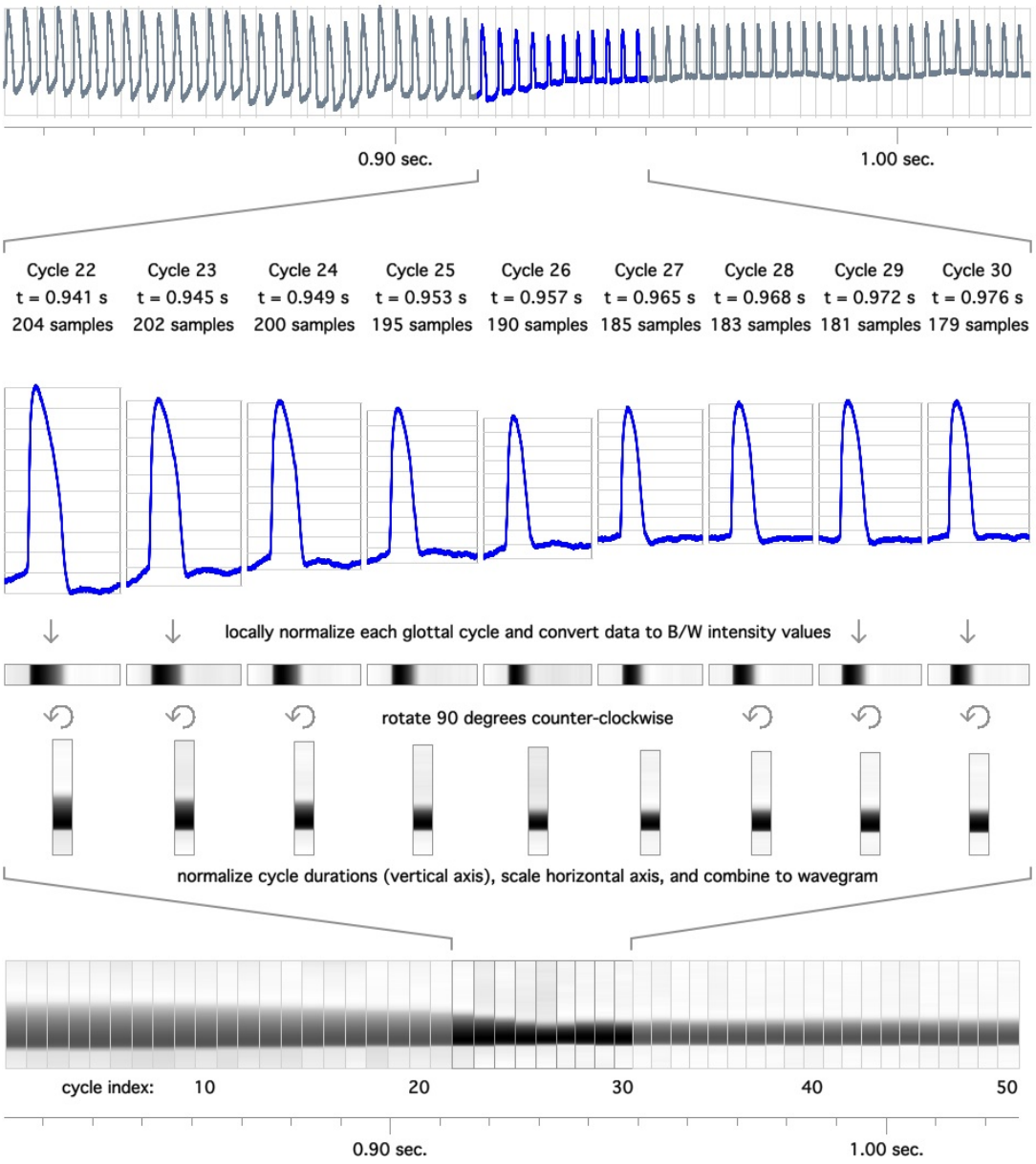


Figure 7.3: The wavegram. Created by Christian T. Herbst under a CC BY-SA 3.0 license.

7.2 Pilot study

Synchronised audio and EGG data were obtained from 5 trained phoneticians, who were asked to produce sustained tokens of /a/ with modal and breathy voice. The data was collected using a Glottal Enterprises EG2-PCX2 electroglottograph and a Movo LV4-O2 Lavalier microphone, at a sample rate of 44100 Hz (16-bit). The acquisition of the signals was controlled with Audacity running on a MacBook Pro (Retina, 13-inch, Mid 2014). The placement of the electrodes strap was checked with the height indicator on the EGG unit. Each participant uttered 10 consecutive tokens of a sustained /a/ in modal voice, followed by 10 tokens of a sustained breathy /a/. All subsequent data processing was performed in Praat (Boersma & Weenink 2018). The onset and offset of each token were detected with an automatic procedure which finds voiced and unvoiced intervals (`To TextGrid (vuv)`). The dEGG wavegram data was extracted from the first 8 glottal cycles of a 500 ms window, centred around the mid-point of each token. A glottal cycle was arbitrarily defined as the interval between two consecutive EGG minima (see Herbst et al. 2010 for an alternative algorithm). From each glottal cycle, the relative amplitude of the dEGG signal was extracted every 10 samples. The reader is referred to Coretta (2017) for the documentation of the algorithms and the research data.

A generalised additive mixed model (GAMM) was fitted to the data to statistically assess differences in vocal fold activity between modal and breathy voicing (see Sóskuthy 2017 and Wieling 2018 for a practical introduction to fitting GAMs in R). The following terms were included: the amplitude of the dEGG signal as the outcome variable, an interaction factor with language and phonation as a parametric term, a smooth over the glottal cycle index to model average changes of the dEGG signal across glottal cycles, and a smooth over the normalised time of the sample within the glottal cycle (as the proportion of the time relative to the duration of the glottal cycle) to model average changes of the dEGG signal within the glottal cycle; two difference smooths over normalised time of the glottal cycle onset and normalised sample time using a by-variable with the phonation factor; a tensor product interaction between normalised cycle time and normalised sample time to model changes of glottal activity through

time, and the same tensor product interaction with a language/phonation by-variable to model phonation-driven differences in changes of glottal activity. Finally, inter-speaker differences were modelled with a by-speaker factor smooth over normalised cycle time. A first-order autoregressive (AR1) model was included to deal with the relatively high auto-correlation in the residuals.

Figure 7.4 shows the modelled wavegrams of modal and breathy tokens. Since the tokens were produced with sustained phonation, no appreciable change within each wavegram can be observed. However, the comparison of the wavegram of modal voice with that of breathy voice reveals differences between the two phonation types. As a general trend, the dEGG maximum and dEGG minimum are achieved later within the glottal cycle in breathy voicing relative to modal voicing. Moreover, differences in velocity of closing and opening movements of the vocal folds are signalled by the relative widths of the purple-coloured bands (around the dEGG maximum) and the green-coloured bands (around the dEGG minimum). While in modal voicing the green band is wider, the purple band is in breathy voicing, indicating that the velocity into and out of the beginning of the closed phase is slower in breathy voicing. According to the approximate significance of the smooth terms, phonation has an effect on the shape of the wavegram as expected ($F(14.681) = 3.187$, Ref.EDF = 19.027, $p < 0.001$).

7.3 Wavegram GAM analysis of vowels followed by voiceless vs voiced stops

This section further illustrates the use of wavegram GAMs by discussing a dynamic analysis of changes in vocal folds activity during the production of vowels followed by voiceless vs voiced stops in Italian and Polish. EGG data was obtained from 9 Italian speakers and 6 Polish speakers. A detailed description of the experimental design is given in Coretta (2019b). Trochaic words with the form $C_1V_1C_2V_2$ were used, where C_1 = /p/, V_1 = /a, o, u/, C_2 = /t, d, k, g/, and V_2 = V_1 (e.g. /pata/, /pada/, /poto/, etc.), embedded in the frame sentences *Dico X lentamente* ‘I say X slowly’ in Italian, and *Mówię X teraz* ‘I say X now’ in Polish. The participants repeated each of the twelve

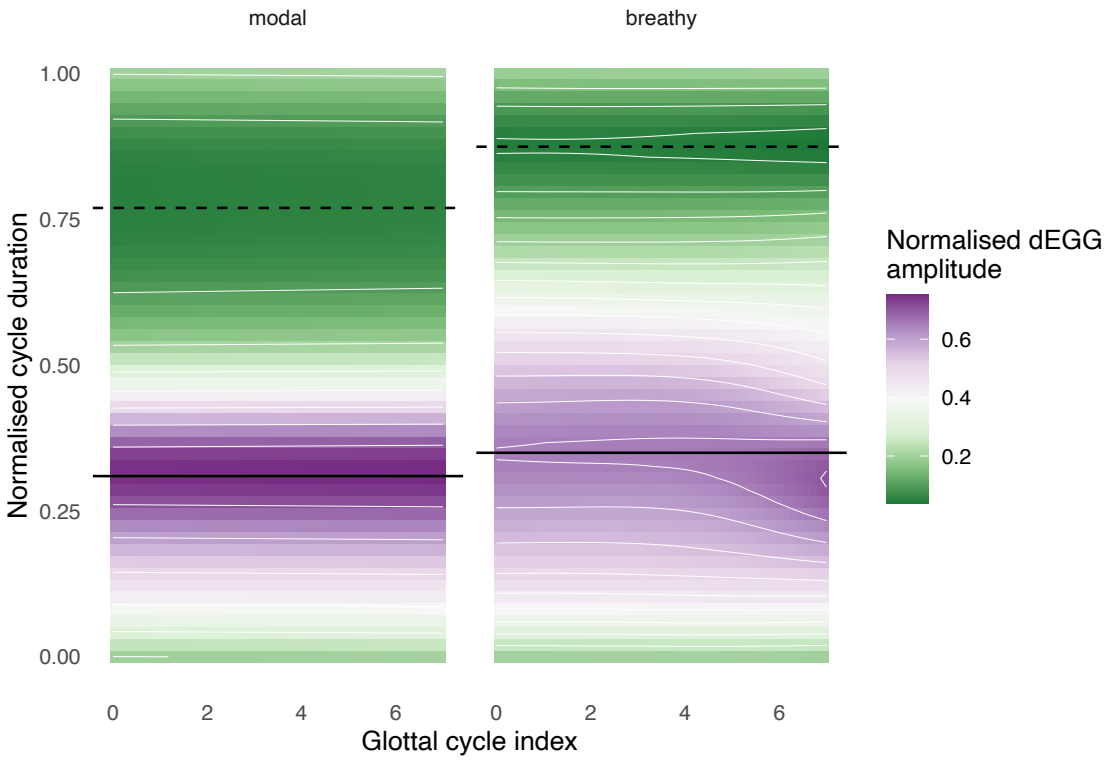


Figure 7.4: Fitted wavegram of modal and breathy phonation (Section 2). The horizontal lines represent the dEGG maximum (solid line) and minimum (dashed line).

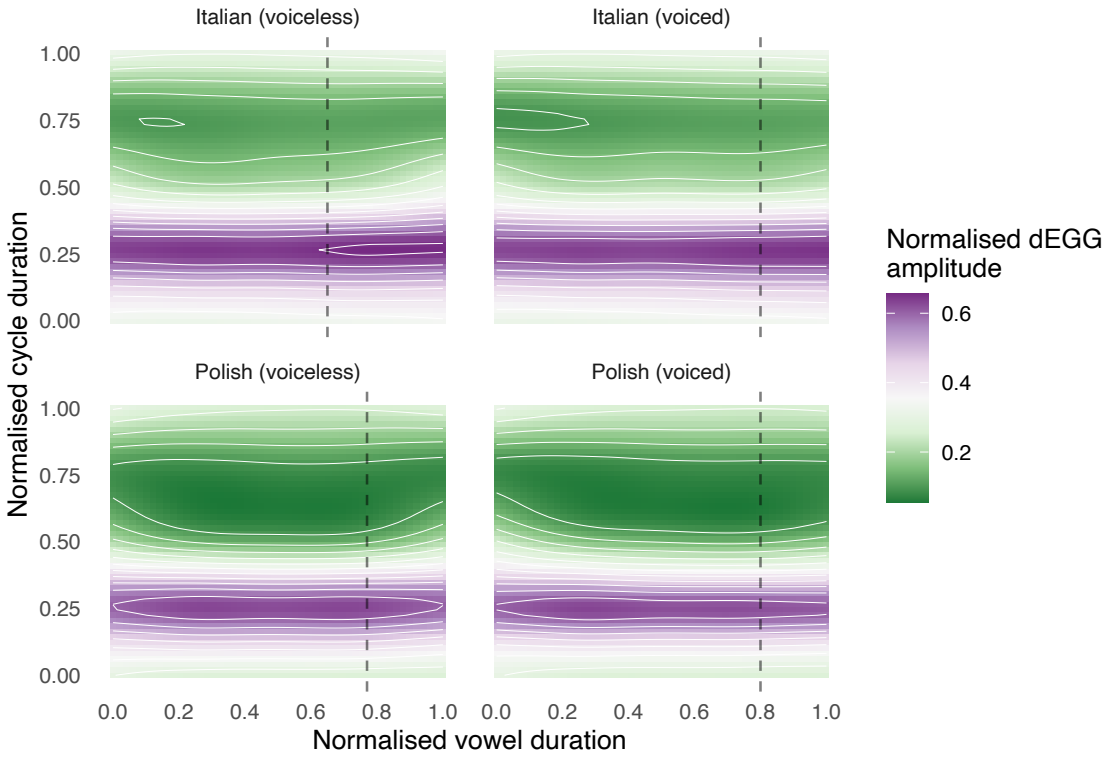


Figure 7.5: Fitted wavegram of vowels followed by voiceless and voiced stops in Italian and Polish (Section 3).

sentence stimuli six times. Processing and analysis of the EGG data were the same as with the pilot study (Section 7.2), with the exception that data was extracted from every glottal cycle within the whole duration of the first vowel of the word stimuli. The vocalic onset and offset were identified as the appearance and disappearance of higher formant structure respectively (Machač & Skarnitzl 2009). Vowel duration was then normalised between 0 and 1 for analysis. The data of this study is available on the Open Science Framework (Coretta 2018a).

The same GAM specification as in the pilot study was used to model changes in glottal activity. Normalised vowel duration was used instead of glottal cycle index. Figure 7.5 shows the modelled wavegrams of vowels followed by voiceless (left) and voiced stops (right), in Italian (top) and Polish (bottom). The pilot study showed that a widening of the wavegram dEGG maximum band (purple) with concomitant shrinkage of the dEGG minimum band (green) signals greater glottal opening. The change in band width corresponds to changes in velocity of the execution of the contacting and decon-

tacting movements. An interesting aspect of the modelled glottal activity concerns the first half of the vowels. The change in the wavegram indicates a process of decreasing glottal opening (from a breathier to a more modal phonation). The greater glottal spread observed at vowel onset could be related to the residual glottal spread of the preceding voiceless stop /p/. This means that the phonation at vowel onset is breathier and becomes more modal during the production of the vowel, stabilising itself at about 20% of the vowel duration.

Focussing now on the second half of the vowel, the wavegrams in Figure 7.5 show a pattern that is symmetrical to that observed in the first half. Namely, glottal opening increases towards the end of the vowel (also see Halle et al. 1967). The magnitude and timing of the change, however, differs in the voiceless and voiced contexts. The change is greater and is implemented earlier in vowels followed by voiceless stops (left panels) than those followed by voiced stops (right panels). The earlier and greater glottal spreading in vowels followed by voiceless stops could be implemented in anticipation of the open glottis required in the production of voiceless stops as a mechanism to cease/attenuate vocal fold vibration. In the case of voiced stops, Halle et al. (1967) propose that increased glottal width can facilitate voicing while the oral tract is constricted (cf. Westbury 1983, who rather argues that decreased glottal width favours voicing).

The wavegrams of vowels followed by voiceless stops also suggest an effect of language (the GAM terms with a by-language factor return *p*-values less than 0.001). The change in activity before voiceless stops is initiated earlier in Italian (at around 65% into the vowel) than in Polish (at about 80%). The approximate time of the change onset is represented by the vertical dashed lines in Figure 7.5. On the other hand, activity in vowels followed by voiced stops is similar in the two languages.

The observed greater increase in glottal opening during the production of vowels followed by voiceless stops in Italian is compatible with the reported presence of pre-aspiration (breathy or voiceless) in Italian geminate stops (Ní Chasaide & Gobl 1993; Stevens & Hajek 2004a,b, 2010; Stevens 2010; Stevens & Reubold 2014). Increased glottal spreading during vocal fold vibration can be interpreted as a precursor of voiceless pre-aspiration. An enough opened glottis can generate enough glottal airflow so as to equalise sub-glottal and supra-glottal pressure, at which point vocal fold vibration

cannot be supported any longer (van den Berg 1958; Rothenberg 1967; Ohala 2011). The outcome is voiceless glottal frication, or, in other words, voiceless pre-aspiration.

The patterns of glottal spreading observed here fit with proposed mechanisms of emergence of voiceless pre-aspiration or lack thereof. Pre-aspiration (whether normative or not), now argued to be more common than previously thought (Helgason 2002), is found in several Nordic languages (Helgason 1999, 2002), English (Gordeeva & Scobbie 2007; Nance & Stuart-Smith 2013; Hejná 2015), and, as mentioned above, Italian, among others. An interesting question is how glottal spreading in vowels in the context of voiceless stops can lead to the emergence of voiceless pre-aspiration in some cases and not in others.

Ní Chasaide (1985) argues that pre-aspiration develops as a means to enhance discriminability in geminate stops by increasing their overall duration. Under this account, closure shortening is a later development, rather than the cause of the emergence of pre-aspiration. Stevens et al. (2014) present further experimental evidence that the duration of pre-aspiration and that of closure are not correlated in Italian. In other words, pre-aspiration and the closure shortening process are independent. In agreement with Ní Chasaide's hypothesis, the presence of pre-aspiration increases the total duration of the VC sequence. Pre-closure glottal spreading can thus lead to the emergence of voiceless pre-aspiration, which can in turn be enhanced by delaying the onset of stop closure.

Lisker (1974) proposes that the laryngeal gesture of glottal spreading can determine the onset of the stop closure. While some varieties of English have been reported to show pre-aspiration, others lack it. Lisker argues that stop closure in voiceless stops occurs not long after the spreading gesture is initiated in order to avoid the emergence of voiceless pre-aspiration. The onset of the closure would be temporally attracted towards the time of glottal opening onset. Speculatively, this could be one of the mechanisms responsible for longer closure durations of voiceless stops relative to that of voiced stops (Lisker 1957; Umeda 1977; Summers 1987; Davis & Summers 1989; de Jong 1991), other things being equal.

To summarise, glottal spreading, which is a typical feature of the production of voiceless stops, is initiated during the articulation of the vowel preceding the stop. If the

degree of spreading surpasses a particular threshold, soon enough a percept of breathy phonation can arise. At this point, two possible scenarios can be envisaged. According to one pathway, breathiness can lead to voiceless pre-aspiration and subsequent enhancement by closure shortening. In such case, pre-aspiration could ultimately develop into normative pre-aspiration like that found in Icelandic. In the alternative scenario, on the other hand, pre-aspiration is prevented from arising. This can be achieved by shifiting the timing of the stop closure towards the onset of the glottal spreading gesture, while keeping the timing of the latter unchanged. As a consequence, stop closure duration is increased, resulting in the known pattern of the differential closure durations of voiceless vs voiced stops.

7.4 Conclusion

This paper introduced a method for modelling glottal activity as obtained from wavegram data (Herbst et al. 2010) using generalised additive (mixed) models (Hastie & Tibshirani 1986; Zuur 2012; Wood 2017). A pilot study of the wavegram GAM method showed that the modelled wavegrams are affected by the global changes in glottal activity depending on the phonation mode of voicing. In particular, the ability of the model to distinguish between modal and breathy voicing was tested, and the results indicated that breathy voicing corresponds to delayed dEGG maxima and minima within the glottal cycle and decreased velocities of the contacting phase (with concomitant increased velocity of the decontacting phase). Building on the results of the pilot study, a second study was run to investigate glottal activity in vowels followed by either a voiceless or a voiced stop in Italian and Polish. The results suggest a change of glottal activity during the second half of the vowel in anticipation of the glottal configuration of the following stop (in accordance with Halle et al. 1967). The change corresponds to increasing glottal width in both contexts. However the increase is greater and implemented earlier in the voiceless than in the voiced context.

Moreover, the change in glottal activity in vowels followed by voiceless stops starts earlier in Italian than in Polish. It was argued that the earlier and grater increase in glottal width in Italian is compatible with reports of pre-aspiration in some varieties of the

language (Ní Chasaide & Gobl 1993; Stevens & Reubold 2014). It was further speculated that once an appreciable degree of glottal spreading is implemented during the vowel, either spreading can result in voiceless pre-aspiration with subsequent stop closure shortening, or voiceless pre-aspiration can be avoided by anticipating oral closure while keeping the glottal gesture (with the consequence of increased closure duration).

A limitation of the proposed wavegram GAM analysis is that, while the model can statistically assess global differences in the wavegram, more localised variation still needs to be assessed qualitatively. For example, while the modelled wavegrams show differences in timing of change in glottal activity, there is no straightforward way of obtaining a unique and robust measure of such timing, and statements about this differences rely on visual inspection of the modelled wavegram. A wavegram GAM analysis, however, is still useful in providing a method to model data from multiple repetitions and multiple speakers in a flexible way. Future work will have to investigate different phonation types (like creaky voice) and phonetic contexts. Moreover, further tests should be conducted to assess the reliability of the method. Finally, ways to obtain quantitative data on timing and degree of changes in glottal activity will be necessary to extend the applicability of the method.

Part III

Chapter 8

General discussion

> Rather than trying to disprove a given hypothesis or showing that one is primary, we argue that it is useful to consider the inter-relationships between these different hypotheses.
>
> ——Winter and Röttger (2011)

This dissertation investigated the influence of consonant voicing on vowel duration by focussing on production aspects of vowel-consonant sequences. As a cross-linguistic tendency, known as the "voicing effect", vowels are shorter when followed by voiceless consonants and longer when followed by voiced ones. Several proposals as to what mechanisms underlie this tendency have been put forward, but no account has won consensus. ~~Two studies were carried out~~ to investigate durational and articulatory properties of vowels and consonants in Italian, Polish, and English. Four original publications presented and discussed the results from these studies. Chapter 4 and **??** show that the duration of the interval between the releases of the stops of a disyllabic word is not affected by the voicing of the second stop in Italian, Polish, and English. On the other hand, the release-to-release interval of English monosyllabic words is longer when the second consonant is voiced. The results in Chapter 6 suggest the existence of a statistical correlation between vowel duration and degree of tongue root advancement, such that longer vowel duratio(S)orrespond to greater tongue root advancement. Finally, in Chapter 7 it was speculated that the timing of the stop closure is modulated

by the presence of emerging voiceless pre-aspiration, so that either enhancement of pre-aspiration delays closure or its prevention anticipates it. This chapter provides an over-arching synthesis of the account proposed here in light of the observed patterns, and discusses the implications for theories of speech production.

8.1 A pluralist view

One of the questions posed at the beginning of the dissertation (Section 2.1) concerned the diachronic pathway to the emergence of the voicing effect. Previous articulatory accounts of the voicing effect, reviewed in Section 1.5, ascribe the driving force behind this phenomenon each to an individual property of speech. According to the compensatory temporal adjustment account (Lindblom 1967; Slis & Cohen 1969a,b; Lehiste 1970a,b), there is a trading relationship between the duration of vowels and that of the following consonant. The account of rate of stop closure transition states that the rate of the closing gesture of voiceless stops is higher than that of voiced stops (Öhman 1967b; Chen 1970). As a consequence, full closure is achieved earlier relative to the onset of the closing gesture when the stop is voiceless. A third notable account, that of laryngeal adjustment (Halle & Stevens 1967; Halle et al. 1967; Chomsky & Halle 1968), holds that the achievement of stop closure in voiced stops is slower than that of voiceless stops to allow for enough time to produce a glottal configuration that is suitable for voicing during closure.

The results of Study I (Chapter 4, Chapter 6, Chapter 7) and II (Chapter 5) bring together aspects of these three accounts, and allow us to formulate a holistic account of the voicing effect, discussed in the following paragraphs, that incorporates different components. Two important aspects of the account proposed here are: (1) the temporal stability of the interval spanning the vowel-consonant sequence, and (2) the timing of the vowel-consonant (VC) boundary within that interval. While (1) draws on the compensatory temporal adjustment account, (2) is informed by the account of rate of closure and that of laryngeal adjustment.

In Chapter 4, we saw that Italian and Polish disyllabic words show properties of temporal stability, as expected by an account of compensatory adjustment. In partic-

ular, it was observed that the duration of the interval between the release of the first consonant and the release of the second in CVCV words (the release-to-release interval) is not affected by the voicing category of the second consonant. However, while the original compensatory account states that the stop closure duration *determines* the duration of the preceding vowel, the formulation of the account proposed here takes a more neutral position. More specifically, I argue that the timing of the VC boundary within the release-to-release interval determines the durations of *both* vowel and stop closure, as discussed in Section 4.4.2.

The second fundamental aspect of the account put forward here is that the timing of the VC boundary is modulated by aspects that would fall under the rate of closure and the laryngeal adjustments accounts. The ultrasound and EGG data from Study I, as discussed in Chapter 6 and Chapter 7, suggest that mechanisms independent from closure duration *per se* can act upon the timing of the VC boundary and, indirectly, on the duration of stop closure. For one, tongue root position can influence the timing of the boundary by delaying it to allow for greater tongue root advancement, which is known to facilitate voicing during the production of the stop closure (Chapter 6). Second, the development of pre-aspiration can influence the timing of the VC boundary by delaying or anticipating it depending on whether pre-aspiration is enhanced or prevented (Chapter 7).

In Study II, discussed in Chapter 5, it was found that the temporal properties of English disyllabic words ~~reflect~~ those of Italian and Polish. As in the latter languages, the duration of the release-to-release interval is not affected by the voicing of the post-vocalic stop. However, the situation is different in English monosyllabic words. In this context, voicing does affect the duration of the interval, and the interval is longer when the post-vocalic consonant is voiced. The fact that voicing affects release-to-release duration was ascribed to mechanisms of contrast enhancement driven by perceptual factors (Section 5.1.1). In Chapter 5, I stipulate that the acoustic temporal stability of the release-to-release interval can be derived from an articulatory account of gestural phasing. While Section 5.1.1 only briefly outlined the main components of this articulatory account, the following section contains a more detailed description.

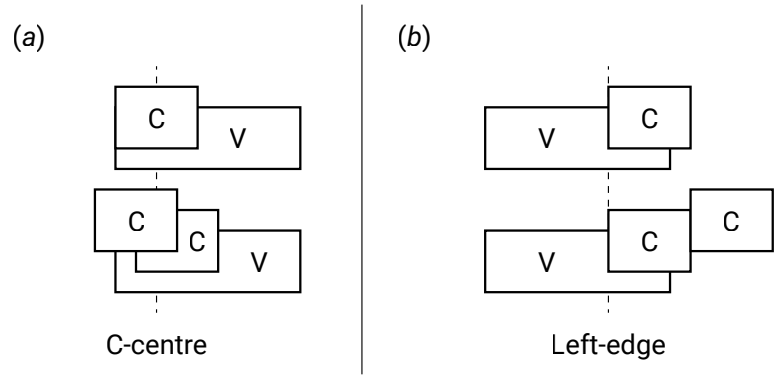


Figure 8.1: Gestural organisation patterns for onsets (a) and codas (b). C = consonant, V = vowel (design based on Marin & Pouplier 2010).

8.1.1 Gestural phasing

As proposed in Chapter 5, the temporal stability of the release-to-release interval is compatible with a certain view of the gestural organisation of vowel and consonants within the domains of the syllable and the word. In this section, I review the principles of Articulatory Phonology and vowel-to-vowel isochrony, and I show how the combination of these two frameworks in a single gestural phasing account can shed light on the durational patterns of the voicing effect discussed in this dissertation.

Within the framework of Articulatory Phonology (Browman & Goldstein 1986, 1988, 2000; Goldstein et al. 2006; Goldstein & Pouplier 2014), speech gestures can be implemented according to two coupling modes: in-phase or anti-phase mode. When two or more gestures are in an in-phase relation, they are initiated in synchrony. If two or more gestures follow an anti-phase coupling mode, the gestures are implemented sequentially, and one gesture starts when the preceding one has reached its target. These two coupling modes can account for temporal aspects observed in the relative phasing of consonants and vowels.

Marin & Pouplier (2010) show that onset consonants in American English are in-phase with respect to the vowel nucleus and anti-phase with each other. Such phasing pattern establishes a stable relationship between the centre of the consonant (or consonants in a cluster) and the following vowel. Independent of the number of onset consonants, the temporal midpoint of the onset (the so-called “C-centre”) is maintained at a fixed distance from the vowel, such that an increasing number of consonants in the on-

set does not change the distance between the vowel and the C-centre (Figure 8.1a). On the other hand, coda consonants are in an anti-phase relation with the preceding vowel and between themselves. When consonants are added to the coda, they are sequentially implemented. Temporal stability in codas is found in the lag between the vowel and the left-most edge of the coda, which is not affected by the number of coda consonants (Figure 8.1b). Other studies found further evidence for the synchronous and sequential coupling modes (see extensive review in Marin & Pouplier 2010 and Marin & Pouplier 2014), although the use of one mode over the other depends on the language and the consonants under study (Pouplier 2012). Consonants can thus be said to follow either a C-centre or a left-edge organisation pattern depending on whether they are in-phase or anti-phase with the vocalic gesture.

Phasing modes are defined within a unit that corresponds to the traditional syllable, or, in other words, relative to the following vowel for onsets and the preceding vowel for codas. Less is known about the relation between segments belonging to different syllables and how syllables are timed and phased within words. Öhman (1967a) and Fowler (1983) propose that vocalic gestures are implemented according to a rhythmic programme and that consonantal (constriction gestures) are superimposed ~~to~~ the vocalic gestural stream. Furthermore, the authors argue that the timing of vocalic gestures follows a regular cyclic pattern, which is in turn responsible for the rhythmic patterns of speech. Fowler (1983) reviews a collection of findings from speech production, perception, and phonological patterns that support the idea of a cyclic production of vowels.

A consequence of the cyclic production of vowels and the independence of the vocalic and consonantal gestures is that two consecutive vowels within a word would be at a stable temporal distance, independent of the nature and number of the intervening consonants. This hypothesis, however, is not borne out by the empirical evidence in Zmarich et al. (2011) and Zeroual et al. (2015). Using electromagnetic articulography, both studies find that the distance between two vowels is greater when the intervening consonant is a geminate compared to when it is a singleton consonant. Furthermore, de Jong (1991) finds only partial support for the independence of vowel and consonant gestures, based on the fact that the tail end of the opening gesture of the vowel is affected by the following consonant. These studies also find substantial inter-speaker variation

in the particulars of the gestural implementation of vowel-consonant sequences.

While the strong prediction of vowel-to-vowel isochrony is not confirmed by data comparing singleton and geminate consonants, a weaker formulation of isochrony might still ~~be appropriate.~~ A possible reason for why the isochrony breaks in the geminate context is that geminate consonants are a blend of two phasing patterns. For example, Zeroual et al. (2015) argue that their findings support the interpretation of geminates as two identical consonants produced sequentially. This would mean that the first part of the geminate is implemented anti-phase with the preceding vowel, while the second part is articulated in-phase with the following. The presence of an anti-phase gesture intervening between the vowels could be responsible for the disruption of the vowel-to-vowel isochrony. If this is the case, then isochrony would apply only in those cases where the intervening consonants are in-phase with the second vowel (see Chapter 9 for a set of hypotheses). This scenario is shown in Figure 8.2(a).

Turning now to the voicing effect, since CVCV words differing in C2 voicing include a singleton consonant, we can expect VV isochrony to apply. This is illustrated in Figure 8.2(b). Since onset consonants are in-phase with the following vowel, the timing of the gestural onset of voiceless and voiced stops in the second syllable of *pata* and *pada* respectively should also be identical. This is in part confirmed by the ultrasound tongue imaging data of Study I (see Appendix D). The duration of the interval between the acoustic release of C1 and the gestural onset time of C2 in CVCV words is not affected by the voicing status of C2, as discussed in Chapter 4 for Italian and Polish and Chapter 5 for English.

On the other hand, the velocity of the closing gesture is known to differ in voiceless vs voiced stops (Summers 1987).[1] While the timing of gestural onset is identical in both voicing contexts, the difference in closing velocity produces the observed acoustic pattern of the later timing of the acoustic VC boundary when the consonant is voiced than when it is voiceless. In Figure 8.2(b), the time of full oral closure is signalled by a dot on the displacement trajectory. Faster closing velocity in voiceless stops creates

[1] As mentioned in Section 3.1.4, tongue movement velocity was not analysed as part of the current investigation, and future work on this aspect is warranted.

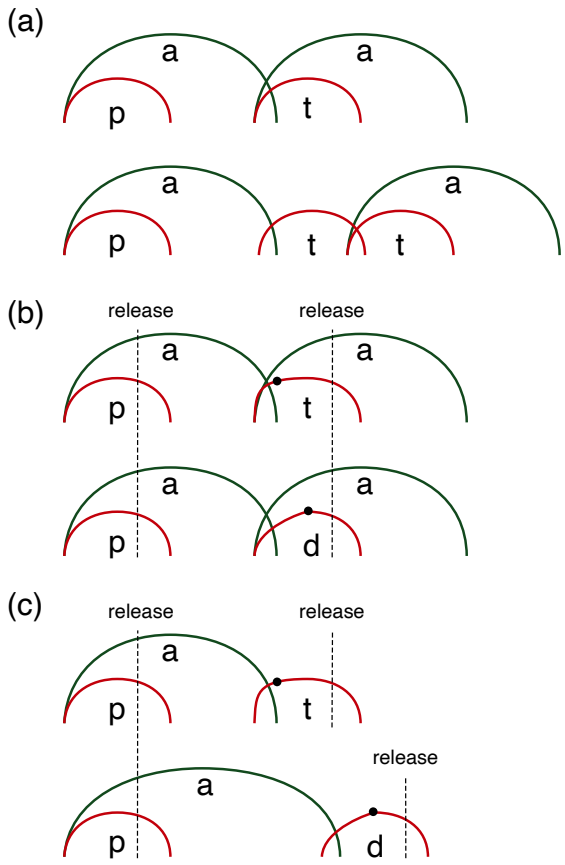


Figure 8.2: Schematics of the gestural phasing of vocalic and consonantal gestures in different contexts, which partially repeats Figure 5.1 from Chapter 5. The *x*-axis is time, while the *y*-axis can be interpreted as oral aperture for vowels and oral constriction for consonants. (a) shows singleton vs geminate stops, (b) voiceless and voiced stops in disyllabic words, and (c) voiceless and voiced stops in monosyllabic words. Note that in (a) the distance between the vowels increases in the geminate context, while it is stable in (b) and (c). The dots in (b) on the consonant gesture lines indicate the time of acoustic closure onset.

an early VC boundary, while a slow closing gesture generates a later VC boundary.

However, in tautosyllabic VC sequences, the consonant gesture is implemented anti-phase with the preceding vowel, meaning that the vocalic and consonantal gestures are produced sequentially. In such case, VV isochrony is broken and the duration of V1 can be modulated freely without proportionally affecting closure duration. This scenario is depicted in Figure 8.2(c). This is what it is argued to have happened in English monosyllabic words, where the temporal distance between the releases of C1 and C2 differs depending on C2 voicing, as shown in Chapter 5. Raphael (1972) and de Jong (1991) indeed find that, in English monosyllabic words, the vocalic gesture is held for longer when the following consonant is voiced and that the gesture onset of voiced stops is timed later during the production of the vowel relative to that of voiceless stops.

As a final note, in Section 4.4.2 I mentioned Tilsen's selection-coordination model as a promising one in that it provides us with theoretical machinery to describe word-holistic patterns of gestural phasing (Tilsen 2013, 2016). This aspect sets the selection-coordination model apart when compared with classical Articulatory Phonology, the focus of which is generally restricted to the level of syllables, as we have seen in the outline of the coupling modes above. However, a detailed development of a selection-coordination interpretation of the gestural phasing account proposed here is beyond the scope of this dissertation, and it is left to future research.

8.1.2 Diachrony, production, and perception

The composite account proposed in the previous sections is diachronic in nature, in ~~as~~ much as it reveals a possible historical pathway to the development of the voicing effect, rooted in production aspects of vowels and consonants. In particular, I argue that the voicing effect can emerge because of the temporal stability of VC sequences and the differences in timing of the VC boundary depending on the voicing of the post-vocalic consonant. Such an account assumes that the original scenario is one in which the duration of vowels and that of closures do not differ across voicing contexts. Durational differences can emerge via developmental learning and historical change in individual speakers and spread across the population.

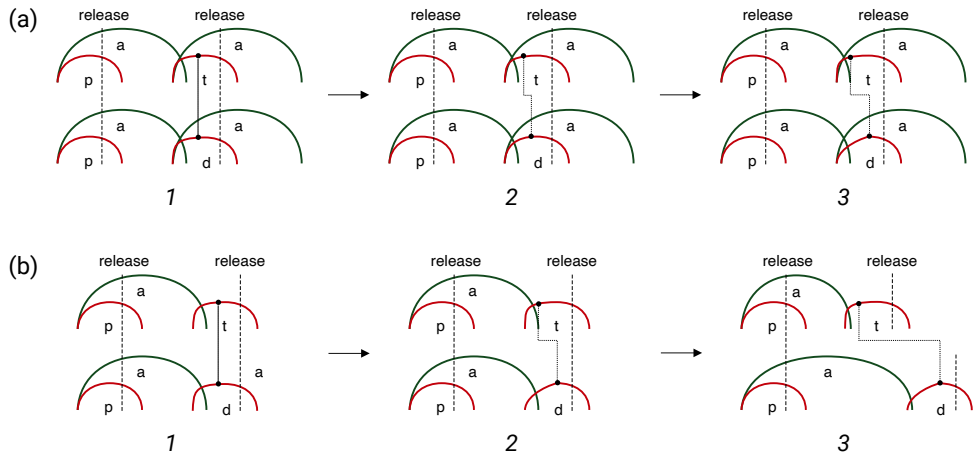


Figure 8.3: Schematics of the developmental/historical emergence of the voicing effect in disyllabic (a) and monosyllabic (b) words. See text for details.

Figure 8.3 illustrates how the emergence of the voicing effect in disyllabic (a) and monosyllabic (b) words with either a voiceless or a voiced post-vocalic stop is envisioned to work. For each type of word, the figure shows three stages that can be thought of as three stages in a continuum of developmental learning and/or the historical change of a language across generations (future work is needed to work out the full details of these scenarios). Starting with disyllabic words in (a) stage 1 (top left panel), the words *pata* and *pada* show an identical temporal profile. The release of the stops are indicated by a dashed line, while full oral closure (the acoustic VC boundary) is signalled by a black dot on the profile of the second consonant. The timing of the vocalic and consonantal gestures (including the timing of the release and the VC boundary) coincide, as illustrated by the straight vertical lines indicating releases and connecting oral closures. In other words, there is no voicing effect, and both the duration of vowels and that of stop closures is not affected by voicing.

At stage 2, the timing of the VC boundary in the two voicing conditions has shifted. There is now a voicing effect: vowels are shorter when followed by voiceless stops and longer when followed by voiced stops, and vice versa voiceless stop closures are longer and voiced stop closures shorter. This change is brought about by the differing closing velocity of voiceless and voiced stops. In relation to the VC boundary shift, three sce-

narios are possible: (1) the VC boundary shifts leftwards in the voiceless context, (2) the VC boundary shifts rightwards in the voiced context, (3) both (1) and (2) are implemented. Only diachronic data will be able to indicate which of these three scenarios is the correct one, but possibly (3) is compatible with the finding of two mechanisms that can lead to the closure velocity differential and hence to a VC boundary shift. As discussed in Chapter 6 and Chapter 7, these mechanisms are tongue root advancement in voiced stops and the emergence of pre-aspiration in voiceless stops. Tongue root advancement would be responsible for a delayed VC boundary in voiced stops, while prevention of pre-aspiration for an earlier VC boundary in voiceless stops. It goes without saying that other mechanisms not investigated here might also play a role, so I do not wish to attribute boundary shifts exclusively to these two mechanisms.

Finally, in (a) stage 3, the temporal distance between the VC boundaries in *pata* and *pada* has increased through time and has now reached a stable state. The absolute difference in vowel and closure durations will depend on the language and on how other factors, like perceptual ones, further modulate it. Note that throughout stages 1 to 3, the temporal stability of the release-to-release interval is maintained due to the preservation of vowel-to-vowel isochrony.

Turning to monosyllabic words in Figure 8.3(b), the bottom left panel in 1 illustrates the initial stage of *pat* and *pad* words where the temporal profiles are, as in (a)1, identical. There is no voicing effect. As in (a)2-3, the VC boundary in the voiceless and voiced contexts shifts in (b)2, and a voicing effect emerges. Now, as discussed in Chapter 5, the duration of the vowels can be modulated for contrast enhancement purposes by temporal shortening in the voiceless context and/or lengthening in the voiced one (the schematics in (b)3 shows the case where both shortening and lengthening occur). The temporal stability of the release-to-release interval is now disrupted. This is possible due to the absence of vowel-to-vowel isochrony (since there is no second vowel to temporally lock in monosyllabic words). Finally, note that the same contrast-enhancing perceptual biases that operate in the context of monosyllabic words can also operate in disyllabic ones, with the exception that vowel-to-vowel isochrony (and release-to-release temporal stability) is preserved in the latter.

The discussion in Section 4.4.2 hinted at an exemplar-theoretic account of the di-

achronic scenario outlined here. Under exemplar-based theories, speech processing operates on parameters the values of which are obtained from statistical distributions. At the early stages of the emergence of the voicing effect, the timing of the VC boundary (as produced by gestural phasing) would come from a distribution shared across voiceless and voiced stops. Deviation from this distribution are a consequence of the application of VC boundary shifting driven by physiological factors. The boundary shifting can then accumulate through time via a perception-production feedback-loop. Statistical sub-distributions in the timing of the VC boundary would thus start emerging for voiceless and voiced stops from the prior unique distribution. These sub-distributions can then act as distributions from which timing information is directly obtained during speech processing. At this point, while the original physiological biases might still be in place, the presence of independent sub-distributions blurs the statistical relationship among durational measures, and, as argued in Section 4.4.2, the true underlying mechanisms are difficult if not impossible to be recovered from synchronic data.

An exemplar-based view was chosen as an illustrative example of the cognitive mechanisms behind the emergence of the voicing effect, although other frameworks might as readily account for the patterns discussed in this dissertation. Future work is warranted to weight the goodness of each individual framework in doing so.

While the account proposed here is based on production mechanisms, it cannot be completely excluded that the production differences observed and stipulated here are a consequence of perceptual biases. In such a scenario, there could be a design feature of the perceptual system that generates the differential duration percept, even when there is not such difference in the acoustic output/input. An example from the domain of vision illustrates this possibility. When looking at a wheel spinning in a clockwise direction, the observer will see the wheel rotating counter-clockwise when the rotation speed exceeds a threshold. There is nothing in the mechanics of a wheel spinning around its axis that can explain this perceptual fact. Rather, visual processing has a design feature (for example, vision refresh rate) that creates the illusion of the wheel spinning in the opposite direction. In this example, the wheel and its percept are two clearly separable ontological entities, so that the percept cannot change how the wheel is spinning. In the case of speech, however, a bias in the perceptual system can (and very often does)

result in differences in production.

Teasing apart production and perceptual mechanisms in speech is more difficult than in the spinning wheel case, since production and perception operate within a single agent, i.e. the speaker/hearer. On the other hand, when we can find independent physical explanations behind production biases, we can consider them design features of the production system, and not just a consequence of perceptual biases. Since in the case of voicing there are independent production reasons for the observed patterns to exist, we can assume that, while perceptual biases can hook on the vowel duration differences as a cue to voicing and enhance such contrast, these differences emerge due to a production mechanism in the first place.

8.2 On cross-linguistic differences

The second aim of the dissertation was concerned with cross-linguistic differences in the development and implementation of the voicing effect. As discussed in Chapter 1, the degree of the voicing effect is generally thought to vary depending on the language. Allegedly, English is the language ~~possessing~~ the biggest effect, while the effect is smaller in Italian and possibly absent in Polish. However, Papers I and II indicate a somewhat different scenario. A direct comparison of the data from Study I and II is not straightforward, given the different materials used in the two studies, but it still provides us with some directions as to what difference, if any, there are between the languages in question. A Bayesian analysis of the effect of voicing in disyllabic words comparing English, Italian, and Polish (see Appendix C) suggests that, when controlling for differences in average baseline vowel duration and speech rate, there is no strong evidence for a difference in the effect of voicing across these languages. Note, however, that no conclusive statement can be done in this regard, due to the low precision of the relevant estimates obtained in the meta-analysis.

As thoroughly discussed in Section 4.4.1, the magnitude of the voicing effect found in Study I for Italian aligns with previous work on this language, which is unanimously regarded as a voicing-effect language. Polish, on the other hand, has been claimed both to show a voicing effect (Slowiaczek & Dinnsen 1985; Nowak 2006; Malisz & Klessa

2008) or not (Jassem & Richter 1989; Keating 1984b; Strycharczuk 2012). In Chapter 4, it was argued that the Polish speakers surveyed in Study I do show the voicing effect, and that this effect is similar in magnitude to that of Italian. As mentioned in Section 4.4.1, it is possible that the null results reported in some of the literature are due to low statistical power, rather than absence of the effect (based on what discussed in Section 3.3).

Previous work (Sharf 1962; Klatt 1973), and in some part the results in Chapter 5, crucially indicate there is a tendency for the effect in English to be greater in monosyllabic than disyllabic words. It is important, then, that differences across languages are tested directly and within the same phonological contexts. An example of this approach is the study in Laeufer (1992), who shows that the effect in English and French is comparable when the vowel baseline duration is analogous.

Moreover, the results of the Bayesian meta-analysis of the English voicing effect (Appendix B) indicate that, while there is a clear positive effect of voicing on vowel duration, less can be said about the magnitude of such effect. Although the estimated range of values is between 55 and 95 ms, the analysis revealed a possibility for publication bias in favour of larger effects, meaning that the obtained values might be overestimated. Note also that there could be differences in speech rate across studies which cannot be controlled for, and older studies might have obtained data based on lower speech rates (which could explain the greater baseline vowel duration in these studies). The smaller effect found in Study II is indicative of such differences. For example, the intercept estimate of vowel duration before voiceless stops is about 125 ms in Study II, but the average vowel duration in the meta-analytical data is 150 ms. Although the voicing effect does not scale linearly with vowel duration, as suggested by Ko (2018), slower speech rates (i.e. longer vowel durations) would correspond to a greater effect of voicing.

Tanner et al. (2019) argue that the effect they found in spontaneous speech is smaller than that of laboratory speech reported in previous studies. However, when their results are compared to those of Study II, this apparent difference is substantially reduced. The ratio of the difference in vowel duration in Tanner et al. (2019) is estimated to be between 1 and 1.16. The ratio in Study II is between 1.03 and 1.17, a range that is

virtually identical to that of Tanner et al. (2019).

I am not claiming, however, that the voicing effect is necessarily universal. Especially in the case of monosyllabic words, as argued above, language-specific perceptual enhancement can modulate the magnitude of the effect. For example, Tanner et al. (2019) demonstrate that, although not unambiguously, the effect of voicing on vowel duration in monosyllabic words differs across varieties of English. More work is needed to assess cross-linguistic differences with a sufficiently sampled dataset and by experimental procedures designed to reduce phonological differences.

8.3 Embracing variation and accepting uncertainty

The third objective of this dissertation focussed on research methods and practices in light of the principles of Open Science (Section 3.3). The planning and execution of research were carried out with openness and transparency in mind. Following state-of-the art recommendations on how to curate and share research objects (Marwick et al. 2017; Berez-Kroeker et al. 2018; Nüst et al. 2018), the data, code, documentation, and software information have been made available as a research compendium on the Open Science Framework (`https://osf.io/w92me/`, Coretta 2020). Data and code have been released under a Creative Commons Attribution 4.0 International license and a MIT licence respectively, so as to facilitate and encourage re-use and attribution. Three R packages were developed while conducting this research as complementary software: speakr (Coretta 2019d), tidymv (Coretta 2019e), and rticulate (Coretta 2018b).

Attention has been given in reporting the results from the statistical analyses in a way commensurate with the weight of the evidence provided. As (Vasishth & Gelman 2019) put it, "conclusions based on data are *always* uncertain, and this is regardless of whether the outcome of the statistical test is statistically significant or not." Rather than focussing on a dichotomous distinction between "significant" and "non-significant", greater emphasis has been put on parameter estimation and quantification of uncertainty. This approach culminated in Study II with the full adoption of the framework of Bayesian statistics. What this dissertation as a whole hopefully highlights is the importance of embracing variation and accepting uncertainty, not as a detrimental aspect

for science, but rather as a necessary step in the cumulative enterprise of knowledge acquisition.

One aspect not directly touched upon here is how Open Science practices affect collaboration among multiple researchers. In fact, the workflow followed here has a ready application in collaborative environments. For example, the Git versioning software has built-in functionalities which can streamline collaboration, like the ability to cleverly merge changes applied by different team members on the same file. The same workflow can be used not only for data collection and analysis, but also for collaborative writing. Although most institutions nowadays are equipped with self-hosted servers, platforms like GitHub, GitLab, and the Open Science Framework provide free servers and software solutions that facilitates research teamwork. Finally, sharing data and code can further foster scientific collaboration (McKiernan et al. 2016; Klein et al. 2018). In this regard, it is hoped that the linguistic community at large will be able to appreciate the innumerable benefits of Open Science, and that researchers in this discipline will decide to adopt and cultivate its principles.

Chapter 9

Implications and future research

This dissertation set out to investigate properties of speech production that could illuminate the debate on the origin of the effect of consonant voicing on vowel duration. While the results of this endeavour contributed to answering the questions introduced in Section 2.1, some open issues remain and new questions are brought to light. The following paragraphs review these and offer some final thoughts on future directions of research.

The use of nonce words in the present experiments might be seen as problematic due to the fact that they might have encouraged unnatural speech. However, using nonce words has the perk of facilitating experimental design and control over phonological factors. Note that phonetic effects like the ones under discussion can operate according to analogical processes based on stored exemplars and/or abstract representations. Hence, even if unnatural speech materials were used here, these studies should have tapped into the speakers' knowledge, even though indirectly. Moreover, nonce words eliminate issues related to usage factors like lexical frequency, which can have a substantial influence (Hay et al. 2015; Sóskuthy & Hay 2017; Sóskuthy et al. 2018; Todd et al. 2019). On the other hand, neighbourhood density was not controlled for, a factor which is also known to play a role (Baese-Berk & Goldrick 2009; Goldrick et al. 2013; Seyfarth et al. 2016; Glewwe 2018). It is desirable that future work investigates the patterns observed here using real words, while carefully controlling for lexical frequency and neighbourhood density, among other usage factors.

A limitation of the studies in this work is related to statistical precision of the ef-

fect estimates. As discussed throughout this dissertation, some of the effects have quite large confidence/credible intervals. In some cases, like in the cross-linguistic comparative analysis of the voicing effect, making unambiguous inferential statements becomes difficult. Possibly, much of this uncertainty derives from the sample sizes employed in these studies. Although the number of speakers included here is generally similar to or greater than average (see Appendix E), the number of observations might not have been sufficient enough to reach an appreciable degree of precision. The results discussed in this dissertation stress how important obtaining a sufficient sample is, especially when dealing with small effect sizes. Much of the phonetic literature relies on small sample sizes, but most work is done on data which is typically statistically noisy.

Moreover, arguments of effect size are very often used to make statements about what constitute a theoretically relevant effect. However, much of phonetics and phonological theory makes qualitative rather than precise quantitative predictions, and argumentations on the theoretical relevance of effect sizes at present are probably biased due to the issues discussed in Section 3.3. In any case, precision targets based on just noticeable differences have restricted scope. For example, Huggins (1972) shows that the perceptual threshold for segment durations varies depending on the type and baseline duration of the segment (cf. the Weber–Fechner law, Fechner 1966). Speakers can reliably detect differences of down to 5 ms with vowels that are 90 ms long (Nooteboom & Doodeman 1980). Furthermore, these perceptual thresholds might be relevant only within the task they have been elicited in, and in more natural contexts even smaller differences could be perceptible in conjunction with other, possibly more robust, cues. In sum, our current knowledge of perceptible differences is still limited, and future work should focus on investigating this matter both in light of theoretical and practical considerations. Until we can establish with certainty what differences in which contexts are physiologically impossible to be perceived, it is probably wise to report even very small, seemingly irrelevant effects while aiming at the highest level of precision possible.

This dissertation focusses on the voicing effect of stop consonants, but other manners of articulation participate in the effect. According to the compensatory account presented here, a greater difference in consonant duration should correspond to a greater

voicing effect. For example, Crystal & House (1988) report a greater difference in fricative duration than in stop duration, which would be compatible with a greater voicing effect in the former. However, while House & Fairbanks (1953) and Peterson & Lehiste (1960) find that the effect is greater in fricatives than in stops, Tanner et al. (2019) observe the opposite trend. Future work should directly test the relation between differences in consonant duration and the voicing effect with consonants belonging to different manners of articulation.

The compensatory account presented here rests on the cyclic production of vocalic gestures (Öhman 1967a; Fowler 1983). Zmarich et al. (2011) and Zeroual et al. (2015) demonstrate that the temporal distance between two vowels with an intervening stop differs depending on whether the stop is a singleton or a geminate (e.g. *aba* vs *abba*, the distance is greater in the latter case). However, I argue that the absence of vowel-to-vowel isochrony across these contexts follows from the gestural organisation of the segments involved. In the case of singleton stops, the vocalic gestures are executed consecutively and the intervocalic consonant is executed synchronously with the second vowel (Figure 9.1a top). In the case of geminates, on the other hand, the intervening consonant is the outcome of the double selection of a single gesture where the first selection is executed anti-phase with the first vowel and the second in-phase with the second vowel (Figure 9.1a bottom). The gesture of the second vowel is initiated after the completion of the intervening anti-phase gesture and in synchrony with the in-phase gesture. As a consequence, the gestural onset of the second consonant is delayed in the geminate context relative to the singleton context.

Future work should investigate vowel-to-vowel isochrony within contexts that have a comparable phasing profile (for example, within CV.CV words and within CVC.CV, but not across the two). The acoustic patterns discussed in this dissertation assume that the distance between the vowels in CV.CV words is not affected by the second consonant and that the timing of the gestural onset and release are the same independent of voicing, while the velocity profiles of the closing gesture differs (Figure 9.1b). By extension, I propose that the vowel-to-vowel interval should be isochronous when comparing different manners of articulation, like in pairs such as Italian /kadi/ ‘you fall’ and /kazi/ ‘cases’. Jaw displacement could be employed as an index of the gestural timing

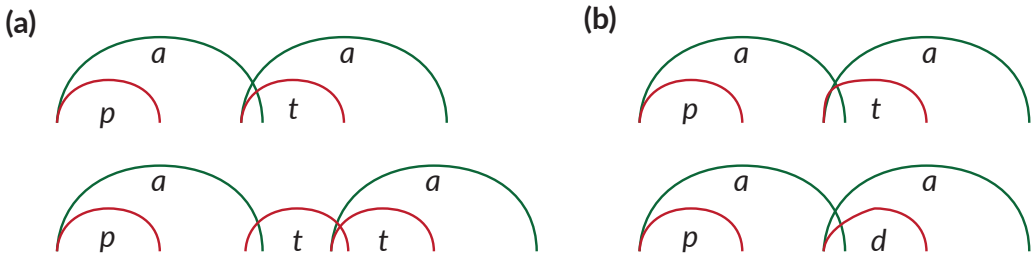


Figure 9.1: Schematics of the gestural phasing of vocalic and consonantal gestures in four different contexts. (a) shows singleton vs geminate stops, while in (b) voiceless and voiced stops are contrasted. Note that in (a) the distance between the vowels increases in the geminate context, while it is stable in (b). The different slopes of the closing part of the gesture in /t/ vs /d/ accounts for the difference in acoustic closure onset.

of the vocalic gestures in plurisyllabic words (Menezes & Erickson 2013; Erickson & Kawahara 2016; Kawahara et al. 2017). The temporal distance between the onsets or maxima of jaw displacement should be identical or very similar across voicing contexts, place of articulation, and manner of the intervening consonant.

To conclude, this work has drawn from acoustic data, ultrasound tongue imaging, and electroglottography, and a diverse set of related but contrasting languages (namely Italian, English, and Polish) with the aim of shedding new light on the widespread phenomenon known as the voicing effect. This investigation led to the development of a holistic account of the voicing effect that combines durational and articulatory aspects of previous research, complemented with diachronic considerations of the pathways that can lead to the emergence of this phenomenon. While contributing to our understanding of the voicing effect, the proposed account also opens up promising and exciting new avenues for research, which can further our knowledge in the domain of speech and language.

Part IV

Appendices

Appendix A

Assessing mid-sagittal tongue contours in polar coordinates using generalised additive (mixed) models

Coretta, Stefano. 2019. Assessing mid-sagittal tongue contours in polar coordinates using generalised additive (mixed) models. Manuscript. DOI: https://doi.org/10.31219/osf.io/q6vzb.

Abstract

Statistical modelling of whole tongue contours has been mostly dominated by the use of Smoothing Splines Analysis of Variance (SSANOVA), although the quantitative analysis of UTI data remains a challenge. Recently, a variety of research disciplines witnessed an increased use of Generalised Additive Models (GAMs) and their mixed-effects counterpart. This family of models is a highly flexible solution which extends standard generalised linear mixed regressions to model non-linear effects. This paper offers a review of GAMs fitted to tongue contours in polar coordinates, as an alternative to polar SSANOVA, given the increasing popularity of these models among linguists. Polar GAMs fitting, significance testing, and model plotting are illustrated by means of an example study that compares tongue contours of voiceless and voiced stops of 12 speakers of Italian and Polish. A brief tutorial illustrates fitting and plotting of polar GAMs with the R package rticulate. The series of polar GAMs indicates a high

degree of idiosyncrasy in tongue root position in voiceless and voiced stops, within and across speakers. Limitations of the current implementation of polar GAMs (such as across-speaker normalisation) and future directions are also briefly discussed.

A.1 Introduction

Since the publication of the seminal paper by Davidson (2006), statistical modelling of whole tongue contours obtained with ultrasound imaging has been dominated by the use of Smoothing Splines Analysis of Variance (SSANOVA, Gu 2013). These models have greatly advanced our understanding of tongue articulation and speech modelling. On the other hand, Generalised Additive Models (GAMs) and their mixed-effects counterpart (GAMMs, Wood 2006) are increasingly adopted in linguistics as a means to model complex data. This paper introduces an implementation of GAMs fitted to tongue contours using a polar coordinate system. The implementation of polar GAMs is illustrated with ultrasound tongue imaging data of an example study that compares voiceless and voiced stops. A brief tutorial shows how to fit polar GAMs with the R package rticulate, developed to facilitate the fitting procedure. Among the advantages of polar GAMs over the current implementation of polar tongue SSANOVA is the possibility of modelling the effect of multiple predictors and that of controlling for autocorrelation in the residuals with the inclusion of autoregressive models.

A.1.1 Ultrasound tongue imaging

Ultrasound imaging is a non-invasive technique for obtaining an image of internal organs and other body tissues. 2D ultrasound imaging has been successfully used for imaging sections of the tongue surface (for a review, see Gick 2002 and Lulich et al. 2018). An image of the (2D) tongue surface can be obtained by placing the transducer in contact with the sub-mental triangle (the area under the chin), aligned either with the mid-sagittal or the coronal plane. The ultrasonic waves propagate from the transducer in a radial fashion through the aperture of the mandible and get reflected when they hit the air above the tongue surface. This “echo” is captured by the transducer and translated into an image like the one shown in Figure A.1 (for a technical description, see

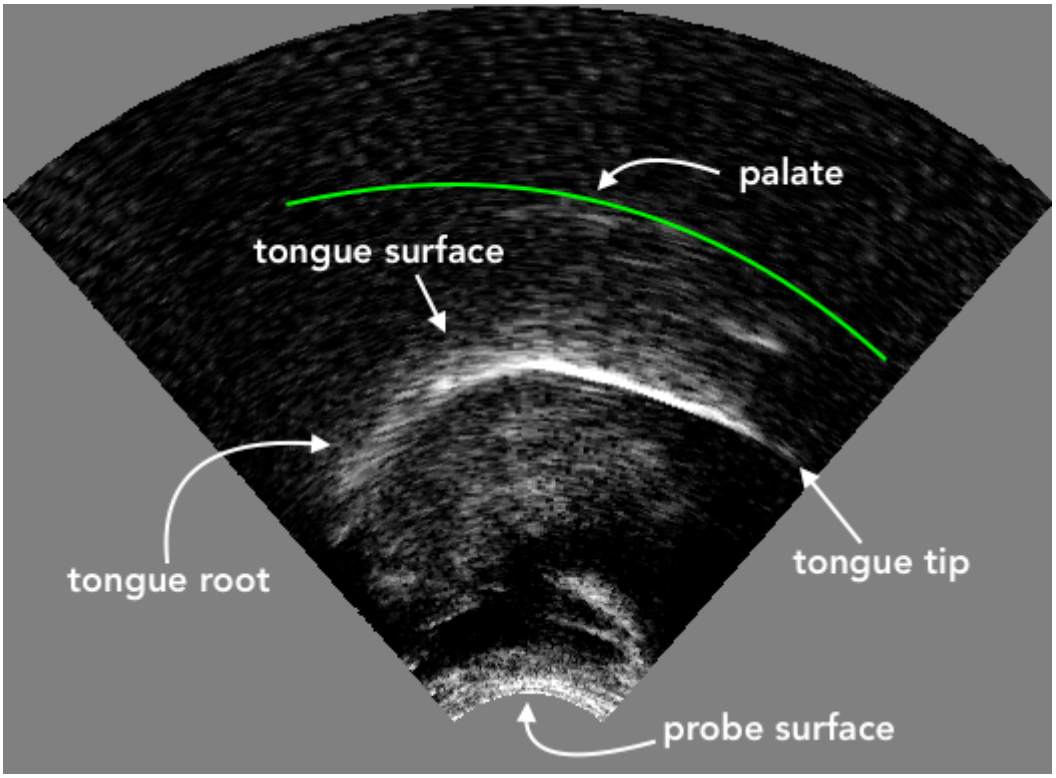


Figure A.1: An ultrasound image showing a mid-sagittal view of the tongue. The white curved stripe in the image indicates where the ultrasonic waves have been reflected by the air above the tongue. The tongue surface corresponds to the lower edge of the white stripe. In this image, the tongue tip is located on the right. The green curve approximates the location of the palate.

Stone 2005).

A.1.2 Generalised Additive models

Generalised additive models, or GAMs, are a more general form of non-parametric modelling that allows fitting non-linear as well as linear effects, and combine properties of linear and additive modelling (Hastie & Tibshirani 1986). GAMs are built with smoothing splines (like SSANOVA, see Helwig & Ma 2016), which are defined piecewise with a set (the *basis*) of polynomial functions (the *basis functions*). When fitting GAMs, the smoothing splines try to maximise the fit to the data while being constrained by a smoothing penalty (usually estimated from the data itself). Such penalisation constitutes a guard against overfitting. GAMs are thus powerful and flexible models that

can deal with non-linear data efficiently.

Moreover, GAMs have a mixed-effect counterpart, Generalised Additive Mixed Models (GAMMs), in which random effects can be included (for a technical introduction to GAM(M)s, see Zuur 2012 and Wood 2017). GAMs can offer relief from issues of autocorrelation between points of a tongue contour (given that points close to each other are not independent from one another). For example, GAMs can fit separate smooths to individual contours, or a first-order autoregression model can be included which tries to account for the autocorrelation between each point in the contour and the one following it. Tongue contours obtained from ultrasound imaging lend themselves to be efficiently modelled using GAM(M)s.

A.1.3 Polar coordinates

Mielke (2015) and Heyne & Derrick (2015a,b) have shown that using polar coordinates of tongue contours rather than Cartesian coordinates brings several benefits, among which reduced variance at the edges of the tongue contour. Points in a polar coordinate system are defined by pairs of radial and angular values. A point is described by a radius, which corresponds to the radial distance from the origin, and by the angle from a reference radius. Tongue contours, due to their shape, tend to have increasing slope at the left and right edges, in certain cases tending to become almost completely vertical. The verticality of the contours has the effect of increasing the variance of the fitted contours (and hence the confidence intervals), and in some cases it can even generate uninterpretable curves.

This issue is illustrated in Figure A.2. The *x* and *y* axes are the *x* and *y* Cartesian coordinates in millimetres. The plot shows LOESS smooths superimposed on the points of the individual tongue contours of an Italian speaker (IT01, see Appendix A.2.1). These contours refer to the mid-sagittal shape of the tongue during the closure of four consonants (/t, d, k, g/) preceded by one of three vowels (/a, o, u/). The tip of the tongue is on the right-end side of each panel. Focussing on the smooths, it can be noticed that the smooths in the contexts of the vowel /u/ diverge substantially from the true contours (as inferred by the points). In the contexts of velar consonants and the other two vowels, the back/root of the tongue is somewhat flattened out relative to the actual contours.

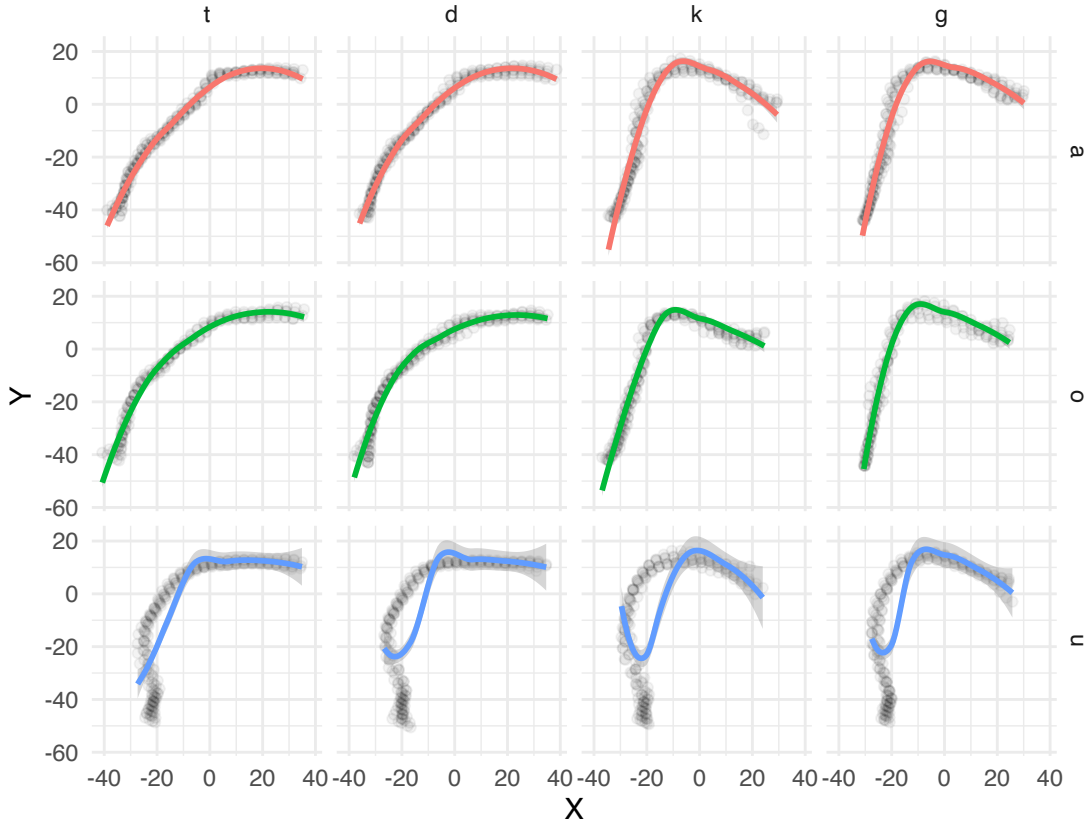


Figure A.2: Estimated tongue contours of IT01 depending on C2 place, vowel and C2 voicing.

These smoothing artefacts arise because, especially at the left-edge of these particular contours, the slope of the curve increases in such a way that the curve bends under itself (see for example the context /ug/, when *x* is between -30 and -20). Since those points on the bend share the same *x* value, the smooth just averages across the *y* values of those points. Figure A.3 shows a more appropriate (artefact-free) way of representing individual tongue contours. In these plots, the points of each contour are connected sequentially by a line, rather than smoothed over. The parts in which the contours bend over themselves are kept as such and visualised correctly.

```
## `geom_smooth()` using formula 'y ~ x'
```

These figures illustrate that using Cartesian coordinates for modelling tongue contours can introduce smoothing artefacts which can in turn negatively affect the model output. When tongue contours are expressed with polar coordinates, on the other hand,

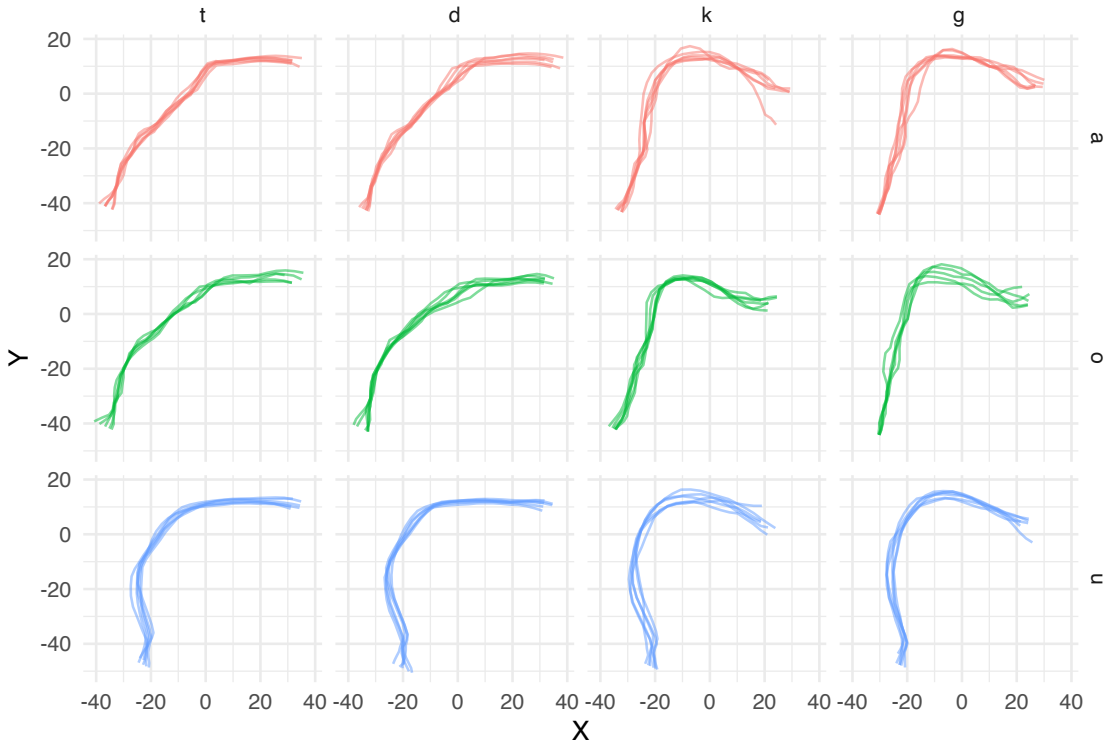


Figure A.3: Estimated tongue contours of IT01 depending on C2 place, vowel and C2 voicing.

the variance is reduced and the fitted contours generally reflect more closely the underlying tongue shape. Mielke has implemented a series of R (R Core Team 2018) functions for fitting polar SSANOVAs to tongue contours (`http://phon.chass.ncsu.edu/manual/tongue_ssanova.r`). While model fitting is achieved using polar coordinates, plotting is done by reconverting the coordinates to a Cartesian system. This same procedure is used in the polar GAM modelling presented here.

A.2 Polar GAM(M)s

GAMs fitted to tongue contours in polar coordinates are introduced here. A polar GAM is constructed as follows. The outcome variable of the model are the radial coordinates, while a smooth term over the angular coordinates is the predictor which takes care of modelling the curved shape of the contour. Other predictors, such as consonant or vowel type, speech rate, or random effects, can also be included. The model returns fitted smooths in polar coordinates. The predicted polar coordinates of the smooths can

be derived from the model and converted into a Cartesian coordinate system (centred on the origin of the polar system) for plotting. A simple example with data from one speaker will illustrate how to fit polar GAMs with the R package rticulate. The following section gives information on the ultrasonic system used for data collection and on how the data has been processed, before moving onto model fitting itself.

A.2.1 Data collection and processing

Synchronised audio and ultrasound tongue imaging data have been recorded from 11 speakers of Italian and 6 speakers of Polish while reading a series of controlled sentences (only 6 of the 11 Italian speakers are analysed here, see Appendix A.3). An Articulate Instruments Ltd™ set-up was used for this study. The ultrasonic data was collected through a TELEMED Echo Blaster 128 unit with a TELEMED C3.5/20/128Z-3 ultrasonic transducer (20mm radius, 2-4 MHz). A synchronisation unit (P-Stretch) was plugged into the Echo Blaster unit and used for automatic audio/ultrasound synchronisation. A FocusRight Scarlett Solo pre-amplifier and a Movo LV4-O2 Lavalier microphone were used for audio recording. The acquisition of the ultrasonic and audio signals was achieved with the software Articulate Assistant Advanced (AAA, v2.17.2) running on a Hawlett-Packard ProBook 6750b laptop with Microsoft Windows 7. Stabilisation of the ultrasonic transducer was ensured by using the metallic headset produced by Articulate Instruments Ltd™ (2008).

Before the reading task, the participant's occlusal plane was obtained using a bite plate (Scobbie et al. 2011). The participants read nonce words embedded in the frame sentence *Dico __ lentamente* 'I say __ slowly' (Italian) and *Mówię __ teraz* 'I say __ now' (Polish). The words follow the structure $C_1\acute{V}_1C_2V_2$, where C_1 = /p/, V_1 = /a, o, u/, C_2 = /t, d, k, g/, and $V_2 = V_1$. Each speaker repeated the stimuli six times.

Spline curves were fitted to the visible tongue contours using the AAA automatic tracking function. Manual correction was applied in those cases that showed clear tracking errors. The time of maximum tongue displacement within consonant closure was then calculated in AAA following the method in Strycharczuk & Scobbie (2015). A fan-like frame consisting of 42 equidistant radial lines was used as the coordinate system. The origin of the 42 fan-lines coincides with the centre of the ultrasonic probe, such that

each fan-line is parallel to the direction of the ultrasonic signal. Tongue displacement was thus calculated as the displacement of the fitted splines along the fan-line vectors. The time of maximum tongue displacement was the time of greater displacement along the fan-line vector with the greatest standard deviation. The vector standard deviation search area was restricted to the portion of the contour corresponding to the tongue tip for coronal consonants, and to the portion corresponding to the tongue dorsum for velar consonants.

The Cartesian coordinates of the tongue contours were extracted from the ultrasonic data at the time of maximum tongue displacement (always within C2 closure). The contours were subsequently normalised within speaker by applying offsetting and rotation relative to the participant's occlusal plane (Scobbie et al. 2011). Each participants' dataset is thus constituted by x and y coordinates of the tongue contours that define respectively the horizontal and vertical axes. The horizontal plane is parallel to the speaker's occlusal plane.

A.2.2 Fitting a polar GAM

GAMs can be fitted in R with the `gam()` function from package mgcv (Wood 2011, 2017). `bam()` is a more efficient function when the dataset has several hundreds observations. The package rticulate has been developed as a wrapper of the `bam()` function to be used with tongue contours. The special function `polar_gam()` can fit a variety of GAM models to tongue contours coordinates, using the same syntax of mgcv. The function accepts tongue contours either in Cartesian or polar coordinates. In the first case, the coordinates can be transformed into polar before fitting. If the data is in the AAA fan-like coordinate system, the origin is automatically estimated with the method in Heyne & Derrick (2015b). If the data is not exported from AAA, the user can specify the known coordinates of the probe origin. The function `plot_polar_smooths()`, used for plotting the estimated contours, converts the coordinates back into Cartesian using the same origin as with GAM fitting.

A GAM in R can be specified with a formula that uses the same syntax of lme4, a commonly used package for linear mixed-effects models (Bates et al. 2015). The mgcv package allows to specify smoothing spline terms with the function `s()`. This function

takes the term along which a spline is created (for example, time in a time series, or *x*-coordinates in a Cartesian system). Among the arguments of `s()`, the user can select the type of spline (the `bs` argument) and the grouping factor used for comparison (the `by` argument). For a more in-depth introduction to GAMs in R for linguistics, see Sóskuthy (2017) and Wieling (2017).

As means of illustration, the following paragraphs will show how to fit a polar GAM with data from one of the Italian speakers. Due to differences in the placement of the probe and in the speakers' anatomy, different portions of the tongue are likely to be imaged across speakers, so that scaling might not be possible (or wise). For this reason, it is recommended to fit separate models for each participant, rather than aggregate all of the data in a single model.

We can start from a simple model in which we test the effect of C2 place, vowel, and voicing on tongue contours. `vc_voicing` is an ordered factor that specifies the combination of C2 place, vowel, and voicing. Modelling different contours for each combination of the three predictors can be achieved by using `vc_voicing` with the `by` argument of the difference smooth, and by including `vc_voicing` as a parametric term. The following code fits the specified model to the contour data of IT01. When running the code, the coordinates of the estimated origin used for the conversion to polar coordinates are returned. The model is fitted by Maximum Likelihood (ML) here to allow model comparison below.

```
it01_gam <- polar_gam(
  Y ~
    vc_voicing +             # parametric term
    s(X) +                   # reference smooth
    s(X, by = vc_voicing),   # difference smooth
  data = tongue_it01,
  method = "ML"
)
```

```
## The origin is x = 14.3901068810439, y = -65.2314851170583.
```

The function `plot_polar_smooths()` can be used to plot the estimated con-

tours. The shaded areas around the estimated contours are 95% confidence intervals. Note that, differently from SSANOVA, statistical significance can't be assessed from the overlapping (or lack thereof) of the confidence intervals. The output of `plot_polar_smooths()` is shown in Figure A.4. For more details on fitting and plotting more complex models (for example models with multiple predictors or tensor product smooths), see the package vignette `polar-gams` (accessible with `vignette("polar-gams", package = "rticulate")`).

```
plot_polar_smooths(
  it01_gam,
  X,
  voicing,
  facet_terms = c2_place + vowel,
  # the following splits the factor interaction into the individual terms,
  # so that they can be called in the plotting arguments
  split = list(vc_voicing = c("vowel", "c2_place", "voicing"))
) +
  coord_fixed() +
  theme(legend.position = "top")
```

One way to assess significance of model terms is to compare the ML score of the full model against one without the relevant predictor, using the function `compareML()` from the itsadug package (van Rij et al. 2017). Both the parametric term and the difference smooth need to be removed in the null model.

```
it01_gam_0 <- polar_gam(
  Y ~
    # vc_voicing +            # remove parametric term
    s(X),                     # keep reference smooth
    # s(X, by = vc_voicing),  # remove difference smooth
  data = tongue_it01,
  method = "ML"
```

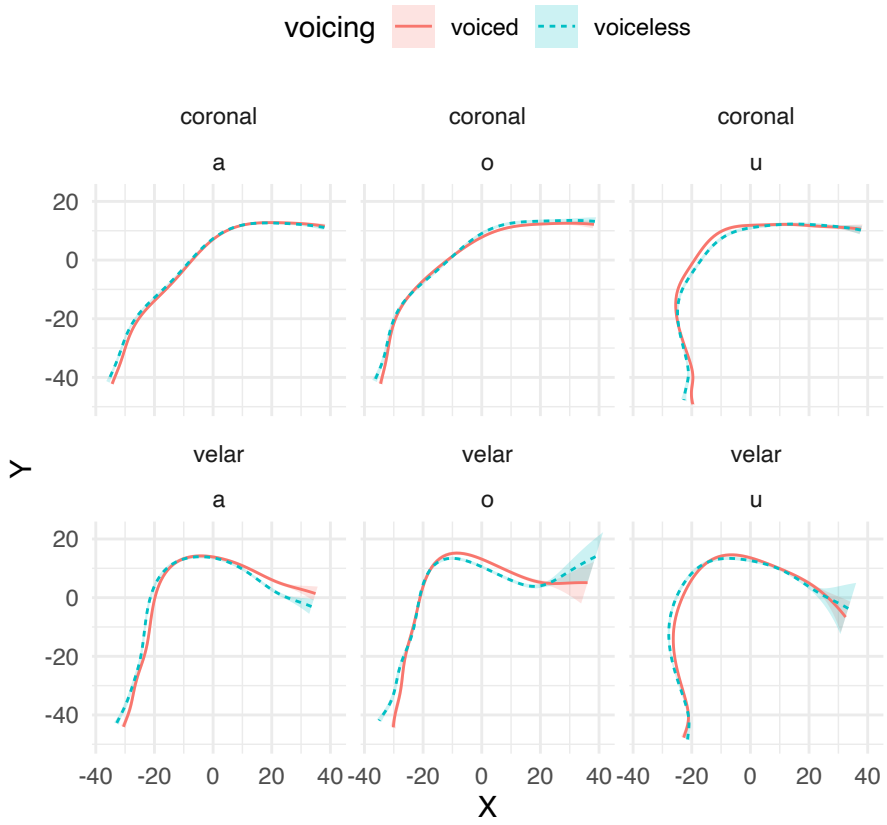


Figure A.4: Estimated tongue contours of IT01 depending on C2 place, vowel and C2 voicing.

```
)
```

```
## The origin is x = 14.3901068810439, y = -65.2314851170583.
```

```
compareML(it01_gam_0, it01_gam)
```

```
## it01_gam_0: Y ~ s(X)
##
## it01_gam: Y ~ vc_voicing + s(X) + s(X, by = vc_voicing)
##
## Chi-square test of ML scores
## -----
##        Model    Score Edf Difference     Df  p.value Sig.
## 1 it01_gam_0 7085.633   3
## 2   it01_gam 4417.563  36   2668.070 33.000  < 2e-16  ***
##
## AIC difference: 5602.30, model it01_gam has lower AIC.
```

To check which part of the contour differs among conditions, the method recommended in Sóskuthy (2017) is to plot the difference smooth and check the confidence interval. The parts of the confidence interval that don't include 0 indicate that the difference between contours in that part is significant. Figure A.5 illustrates the use of difference smooths with the difference smooths of voiceless vs voiced coronal stops when the vocalic context is /a/ or /u/. As per usual, the tongue tip is on the right-end side of each plot. The difference smooths indicate that there is a significant difference along the posterior part of the tongue (the root and dorsum). Based on the predicted smooths sown in Figure A.4, we can argue that, in the context of coronal consonants, the root is more advanced in voiced relative to voiceless stops (when the vowel is either /a/ or /u/), and that the dorsum is also somewhat retracted in voiced stops if the vowel is /u/.

As mentioned in the introduction, autocorrelation in the data can produce unwanted patterns in the residuals, which in turn can affect the estimated smooths (and falsely increase certainty about them). A first-order autoregressive (AR1) model can be included

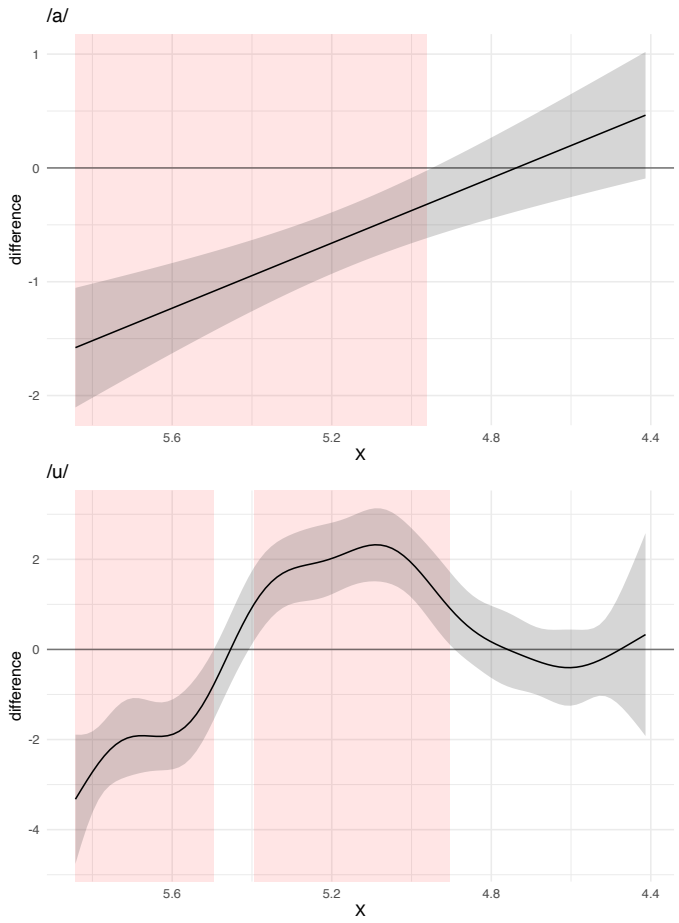


Figure A.5: Difference smooth of voiceless vs voiced stops in the context of /a/ (top) and /u/ (bottom).

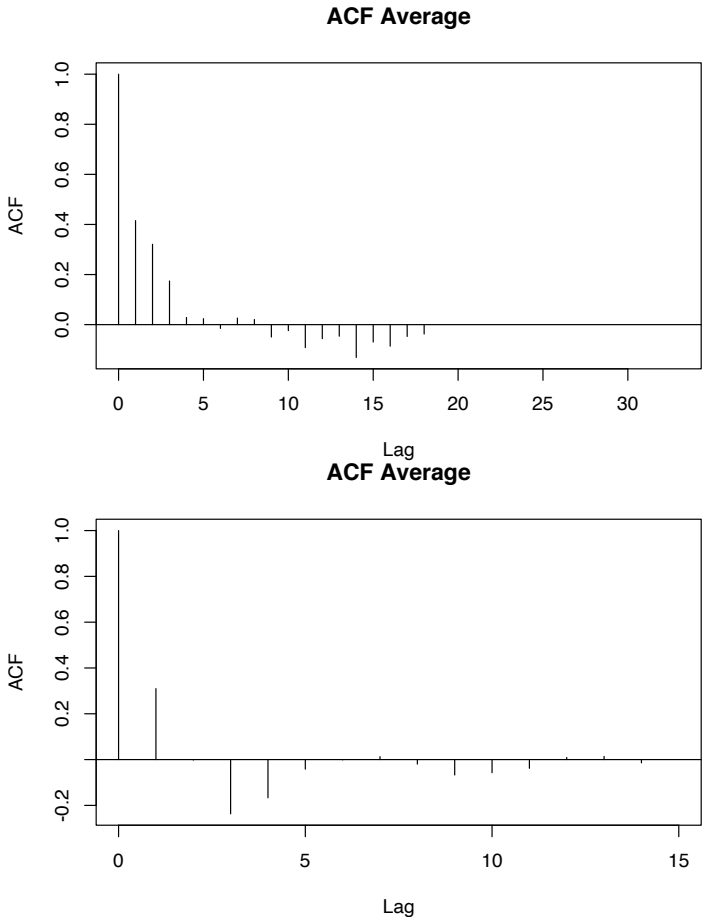


Figure A.6: Autocorrelation plots of a model fitted without (top) and with (bottom) a first-order autoregressive model (AR1).

to reduce autocorrelation at lag 1. Figure A.6 show the autocorrelations in the residuals without (top) and with (bottom) an AR1 model. The GAM model with the AR1 correction has lower autocorrelation values. In this case, it is thus advisable to perform ML comparison and smooths plotting with models in which an AR1 model has been included. For a more in-depth treatment of issues related to autocorrelation, see Sóskuthy (2017).

A.3 Comparing tongue root position in voiceless and voiced stops

Tongue root advancement is a well-known mechanism employed to keep intra-oral pressure below the threshold required for voicing (Ohala 2011; Kent & Moll 1969; Perkell 1969; Westbury 1983; Ahn 2018). Among the languages reported to show tongue root position differences between phonation categories in stops there are English (Westbury 1983; Ahn 2018), Brazilian Portuguese (Ahn 2018), and Hindi (Ahn 2016). Tongue root advancement is one of several mechanisms employed by speakers to enlarge the oral cavity during the production of a stop closure. The decrease in pressure that follows from such expansion ensures that voicing can be maintained during the closure.

Mid-sagittal tongue contours at maximum tongue displacement of voiceless and voiced stops have been compared using polar GAMs. To exemplify how polar GAMs can be used to model articulatory differences within and between speakers, data from 6 speakers of Italian and 6 speakers of Polish will be discussed. Note that the 6 Italian speakers are representative of the general trends found in the entire sample of 11 speakers. Figure A.7 to Figure A.18 show an appreciable degree of variation across speakers and phonological contexts in relation to the differences in tongue shapes between voiceless and voiced stops (IT07 and Pl05 both miss data from /u/ due to the poor quality of the ultrasonic image for this vowel). In some speakers and contexts, the tongue root (the left part of the tongue contours) is more advanced in voiced stops than in voiceless stops.

In particular, IT01, IT02, PL05, and PL06 show a robust pattern in which the tongue root in voiced stops is more advanced than in voiceless stops in most vowel/place contexts. The other speakers, however, either don't have any tongue root advancement (like Pl03), or they have advancement in only some of the phonological contexts (like PL07). Moreover, IT11 has the opposite pattern, especially with velar stops, such that voiced stops have a retracted tongue root compared to voiceless stops. IT04 is a clear example of tongue body lowering (another cavity expansion mechanism), as it can be seen in coronal stops. This level of idiosyncrasy (both within and between speakers) is not surprising, and it qualitatively resembles the degree of variability found, for exam-

ple, in Ahn (2018) for English and Brasilian Portuguese. Finally, no clear patterns can be discerned between speakers of Italian and Polish that could point to cross-linguistic differences.

As for the magnitude of the difference in tongue root position, such difference is about 2 mm in the data reported here. Kirkham & Nance (2017) find that the tongue root in +ATR vowels is on average 4 mm more advanced than the respective −ATR vowels. Rothenberg (1967) argues, based on modelling, that the tongue root can move forward by a maximum of about 5 mm mid-sagittally. This movement corresponds to an average volume increase of 18 cm^2. Given these estimates, it can be argued that a 2 mm change in root position along the mid-sagittal plane contributes to an appreciable oral cavity volume increase. Considering that other volume expansion mechanisms can operate along with the advancement of the tongue root (like larynx lowering, tongue body lowering, etc.), the tongue root driven volume increase found here, although at first sight small, seems to be sufficient to allow for voicing to be maintained during the closure of voiced stops.

A.4 Conclusions

Generalised additive (mixed) models (GAMs) can be efficiently used to statistically assess differences in tongue contour shapes as obtained from ultrasound tongue imaging. This paper showed how GAMs can be fitted to tongue contours in polar coordinates in R with the specialised package rticulate. An example of how GAMs can help modelling differences in tongue contours has been illustrated with data from 12 speakers of Italian and Polish in which the mid-sagittal tongue contours of voiceless and voiced stops where compared. The advantages of polar GAMs over the current implementation of polar tongue SSANOVA include: the ability to specify multiple predictors and random effects; control over the autocorrelation in the residuals which could otherwise make the model overconfident; separate methods for assessing statistical significance at the level of the predictor (with model comparison) and for identifying which part of the tongue differs significantly (by visualising the difference smooths). The same general issues noted in Davidson (2006) for SSANOVA apply to polar GAMs. In particular,

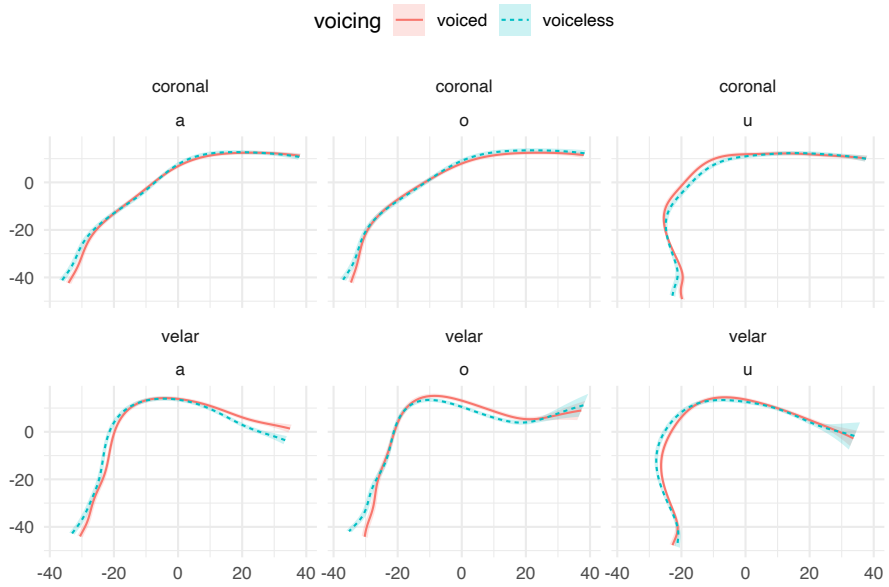


Figure A.7: Tongue contours of voiceless and voiced stops in IT01.

while within-speaker normalisation can be achieved by rotation and offsetting of the data relative to a bite plate (as done here), across-speaker normalisation represents a bigger challenge. Since we can't deduced with sufficient certainty from the ultrasonic image which part of the tongue is being actually imaged, it is not possible to define fixed anatomical landmarks across speakers that can be used in normalisation. For this reason it has been recommended here to fit separate models for each speaker. Future work will explore ways of allowing the user to use data aggregated from multiple speakers while accounting for the uncertainty in which parts of the tongue are imaged. Finally, polar GAMs can also be readily extended to model 3D tongue surfaces and whole tongue contours differences over time (in other words, how the sectional shape of the tongue changes over time).

A.5 Data Accessibility Statement

The data and code used in this paper can be viewed and downloaded at the Open Science Framework link `https://osf.io/q7hyz/?view_only=f9d9f865619848f4bc575bc86cb07282`.

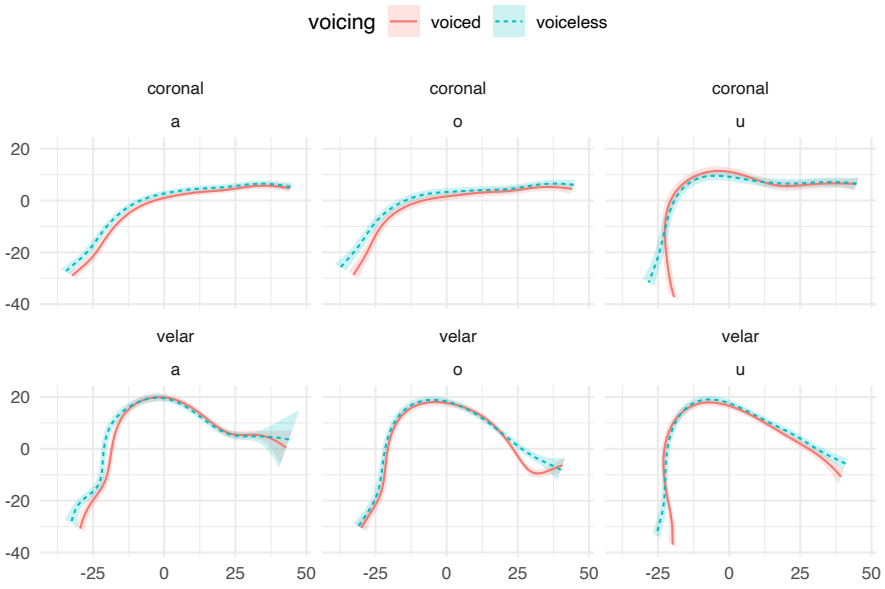


Figure A.8: Tongue contours of voiceless and voiced stops in IT02.

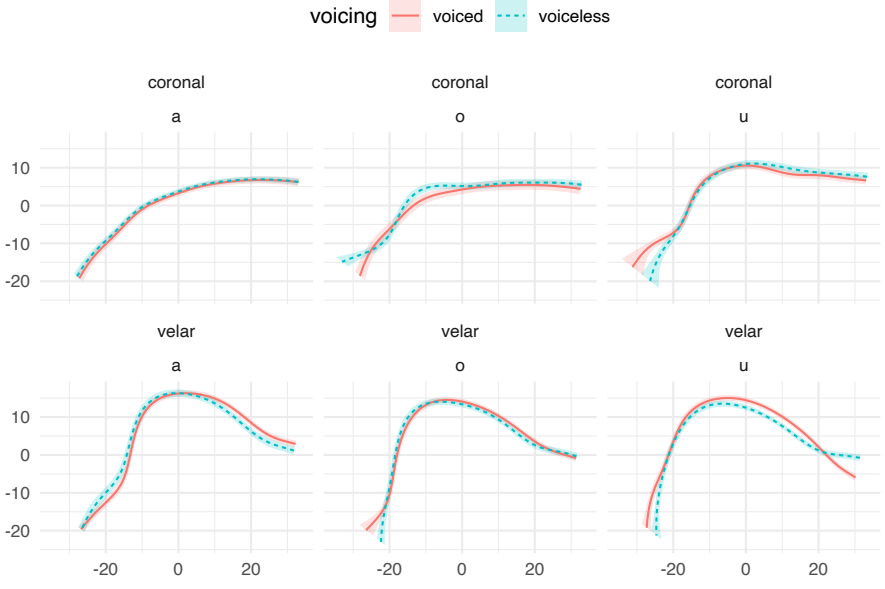


Figure A.9: Tongue contours of voiceless and voiced stops in IT03.

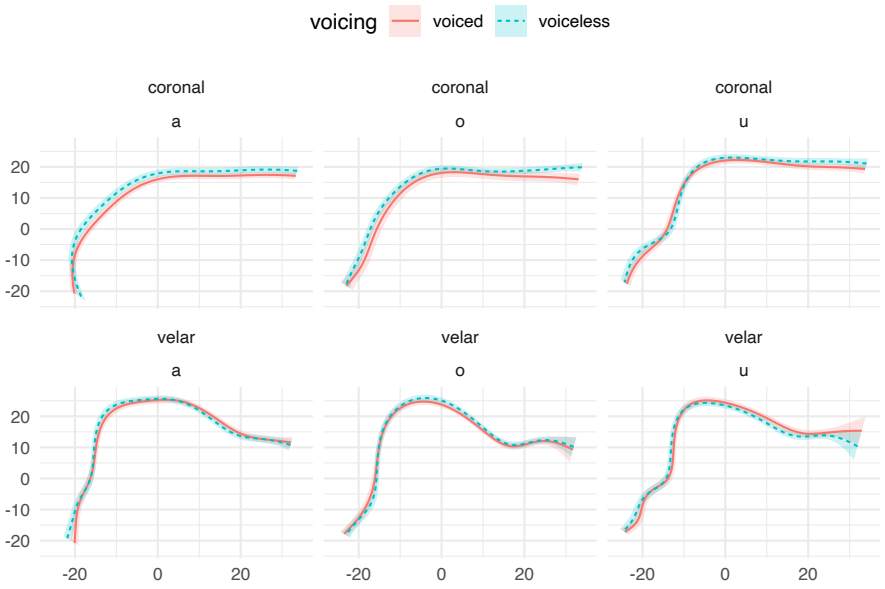


Figure A.10: Tongue contours of voiceless and voiced stops in IT04.

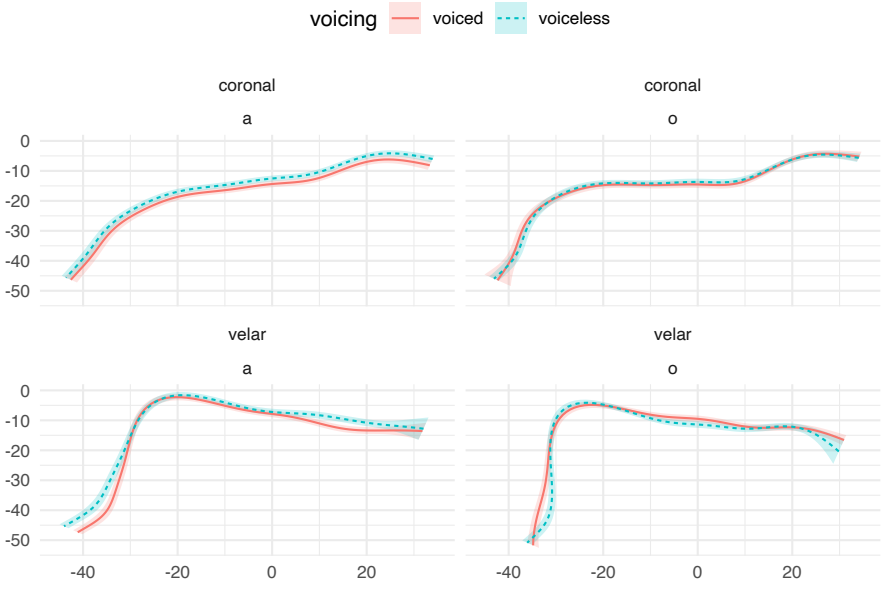


Figure A.11: Tongue contours of voiceless and voiced stops in IT07.

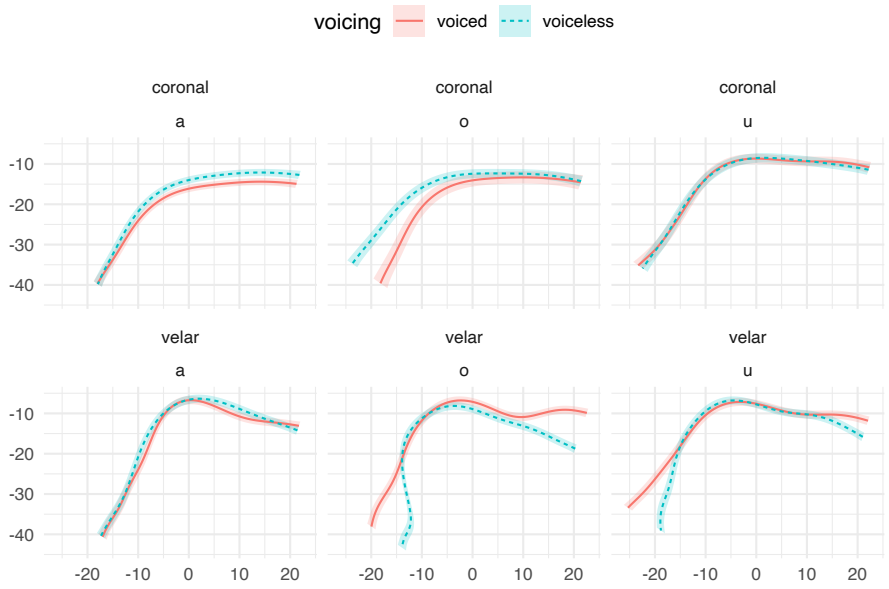


Figure A.12: Tongue contours of voiceless and voiced stops in IT11.

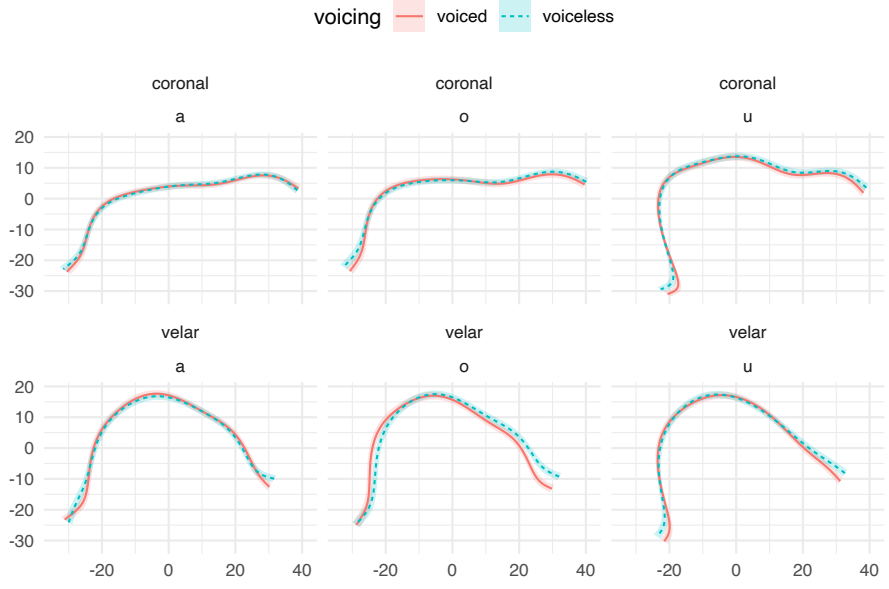


Figure A.13: Tongue contours of voiceless and voiced stops in PL02.

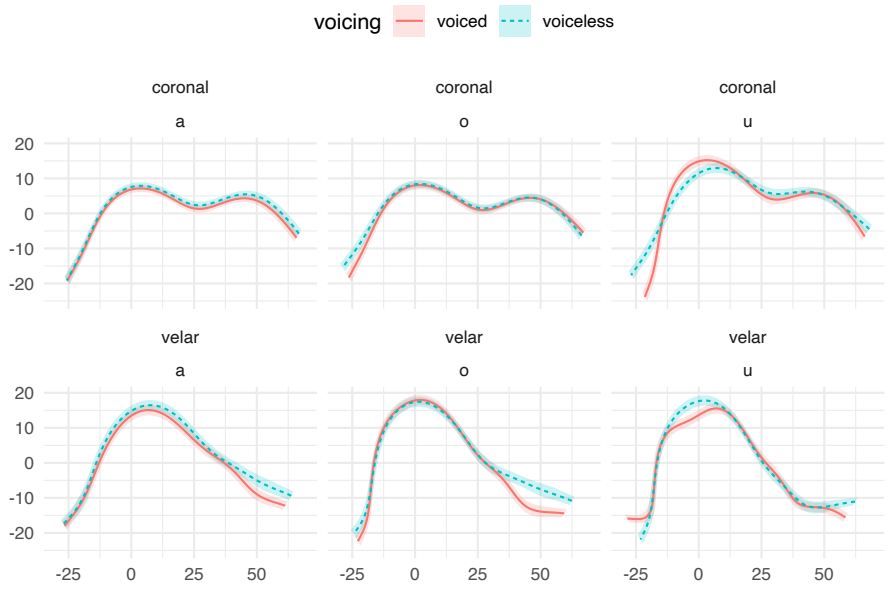


Figure A.14: Tongue contours of voiceless and voiced stops in PL03.

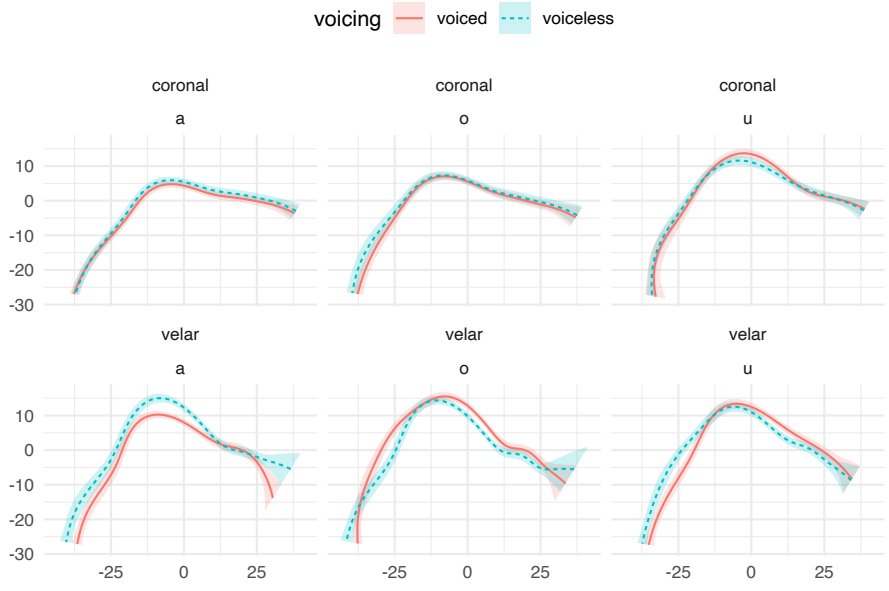


Figure A.15: Tongue contours of voiceless and voiced stops in PL04.

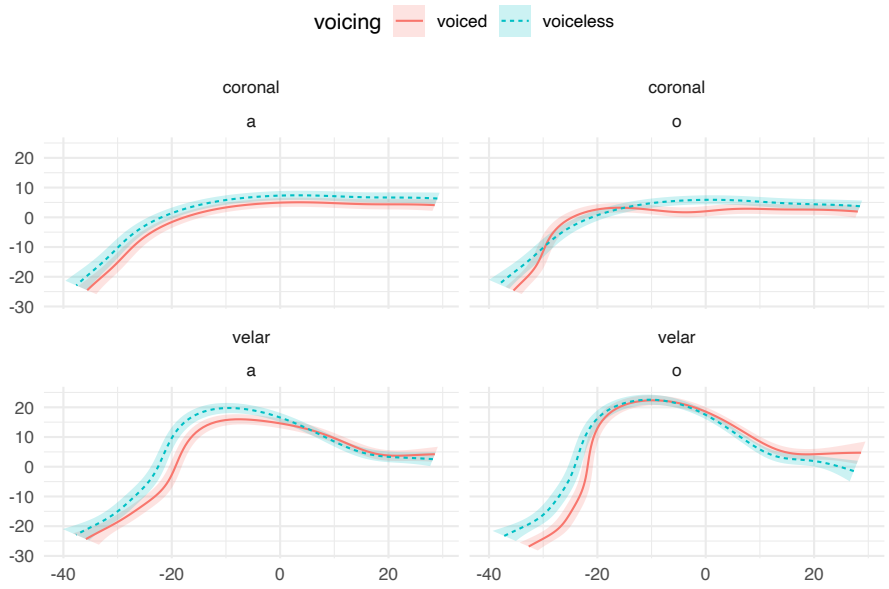


Figure A.16: Tongue contours of voiceless and voiced stops in PL05.

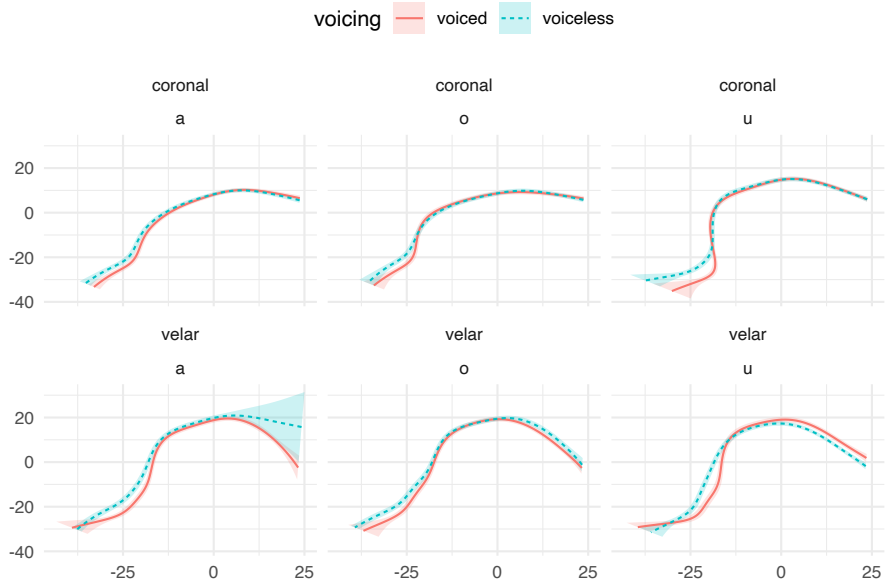


Figure A.17: Tongue contours of voiceless and voiced stops in PL06.

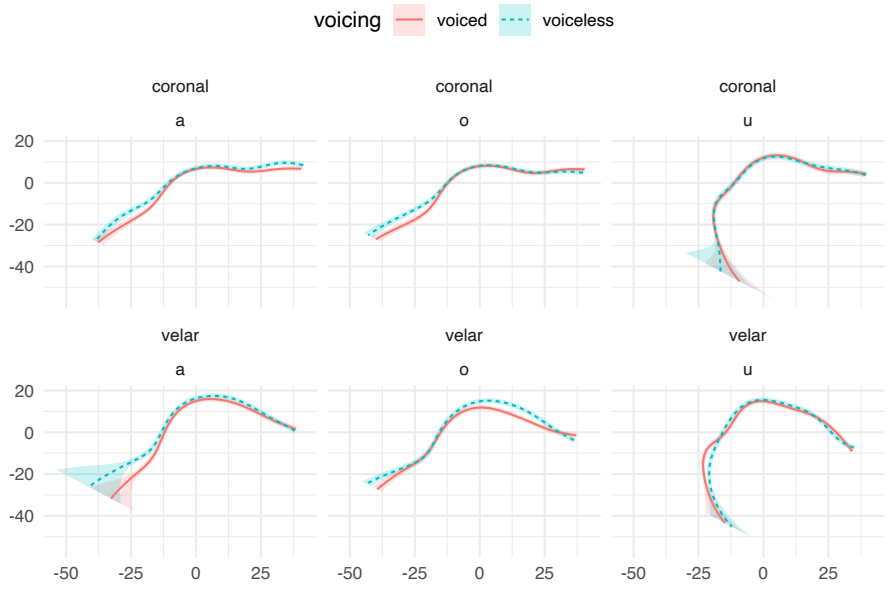


Figure A.18: Tongue contours of voiceless and voiced stops in PL07.

Appendix B

Bayesian meta-analysis of the voicing effect in English

A Bayesian meta-analysis of the English voicing effect was run on the basis of 11 estimated posterior distributions extracted from 9 different publications, following the procedures discussed in Nicenboim et al. (2018). The studies were selected by scraping the first 100 results on Google Scholar with the keywords “vowel duration voicing English.” Other studies which were known to the author but not present among the Google Scholar results were also included. Since two publications (Sharf 1962 and Klatt 1973) tested both monosyllabic and disyllabic words, two separate posterior distributions were estimated for each word type. This leads to a total of 11 posterior distribution of the effect of consonant voicing on vowel duration in English (7 estimated posteriors from 7 publications plus 2 each from 2 publications).

The posterior distributions of each study have been obtained by fitting a Bayesian linear model to the summary data (the means of vowel duration before voiceless and voiced stops) provided by the respective publications. These models had the mean vowel durations as outcome and consonant voicing (voiceless vs voiced) as the only predictor. Three studies, Luce & Charles-Luce (1985), Davis & Summers (1989), and Ko (2018), reported measures of dispersion along with the means. Measurement error models were used to obtain the posterior distributions from these studies. The measurement error term in such models allows us to include information of the dispersion of the mean vowel durations, and hence of the uncertainty that comes with them. All the

Table B.1: Bayesian estimates of the voicing effect in indvidual studies.

Study	Estimate	Est.Error	Q2.5	Q97.5	Syllable position	N. speakers
Heffner (1937)	61.66	19.68	21.85	100.28	final	1
House & Fairbanks (1953)	81.72	14.97	52.08	111.11	final	10
Zimmerman & Sapon (1958)	86.38	29.13	19.56	139.83	final	2
Peterson & Lehiste (1960)	103.52	29.39	43.60	161.22	final	5
Sharf (1962)	24.57	14.02	-3.54	53.22	non-final	1
Sharf (1962)	53.45	34.43	-19.58	119.51	final	1
Chen (1970)	152.41	25.32	94.46	195.87	final	1
Klatt (1973)	21.17	42.44	-85.48	102.46	non-final	3
Klatt (1973)	52.88	45.02	-63.60	126.78	final	3
Mack (1982)	125.20	21.17	83.11	165.33	final	3
Luce & Charles-Luce (1985) final	77.19	9.82	57.26	95.99	final	3
Luce & Charles-Luce (1985) medial	40.72	8.78	24.08	58.67	final	3
Davis & Van Summers (1989)	18.43	4.38	9.86	27.19	non-final	3
Laeufer (1992)	72.69	42.08	-15.44	154.12	final	5
Ko (2018)	35.89	35.66	-34.37	105.06	final	7

models for estimating the posterior of the individual studies were fitted with the following priors: a normal distribution with mean = 0 ms and SD = 300 for the intercept, and a normal distribution with mean = 0 ms and SD = 100 for the effect of consonant voicing. The simple models (without an error term) also included a prior for the residual variance as a half Cauchy distribution with location = 0 ms and scale = 25.

A data set with the mean estimates and estimated standard errors from these 11 posterior distributions (Table B.1) has then been used to fit a further Bayesian measurement error model. In this model, the mean estimates with the estimated standard errors were included as the outcome, while a by-study random intercept was the only predictor. The models were fitted in R with brms using Markov Chain Monte Carlo simulations, with 4 chains, 2000 iterations of which 1000 for warm-up.

The following is the summary of the meta-analytical model (as output by `summary()` function). The population-level effects are the ones of interest. Figure B.1 is a visual aid to the summury, and shows a variety of credible intervals of the estimates from the model. The blue-coloured bars represent (from darker to lighter blue) the 50%, 80%, and 95% credible intervals (CIs). The black lines are the 66% (thick) and 98% (thin) CIs.

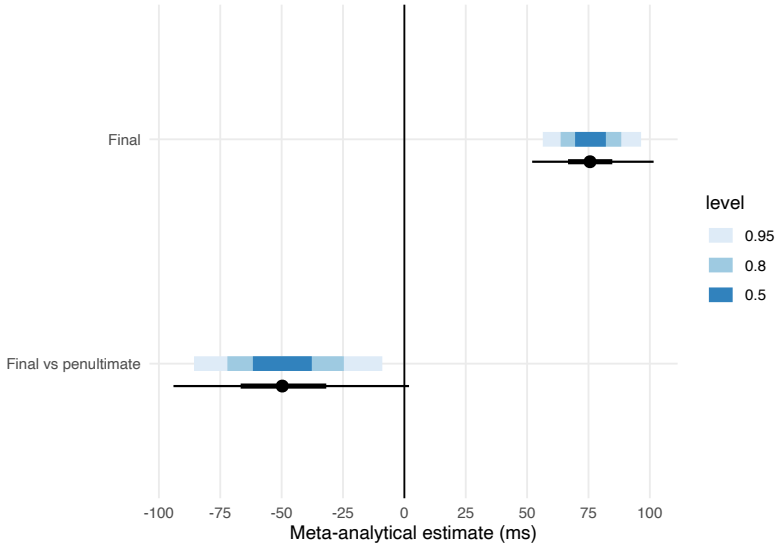


Figure B.1: Credible intervals of the meta-analytical posterior distributions.

```
## Group-Level Effects:
## ~study (Number of levels: 15)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)    23.36      8.88     9.80    44.34 1.00     1234     2026
##
## Population-Level Effects:
##                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept           75.83     10.01    56.39    96.43 1.00     1645     1315
## syl_posnonMfinal   -49.14     19.10   -85.69    -8.98 1.00     1698     1831
```

The 95% credible interval (CI) of the model intercept (which corresponds to the estimated voicing effect in word-final syllables) is between 56.39 and 96.43 ms. This means that there is a 95% probability that the true effect lies between about 56 and 96 ms. The mean of the posterior distribution is 75.83 ms (SD = 10.01). Given the 95% CI of the meta-analytical posterior distribution, it can be inferred that the true effect of voicing in word-final syllables in English is positive and between 50 and 100 ms. However, note that the meta-analytical estimate might suffer from publication bias (cf. below).

The posterior mean of the coefficient when the target syllable is in penultimate position is -49.14 ms (SD = 19.10, 95% CI = [-85.69, -8.98]). Note that the estimated error is double compared to that of the intercept, which means the there is greater uncertainty

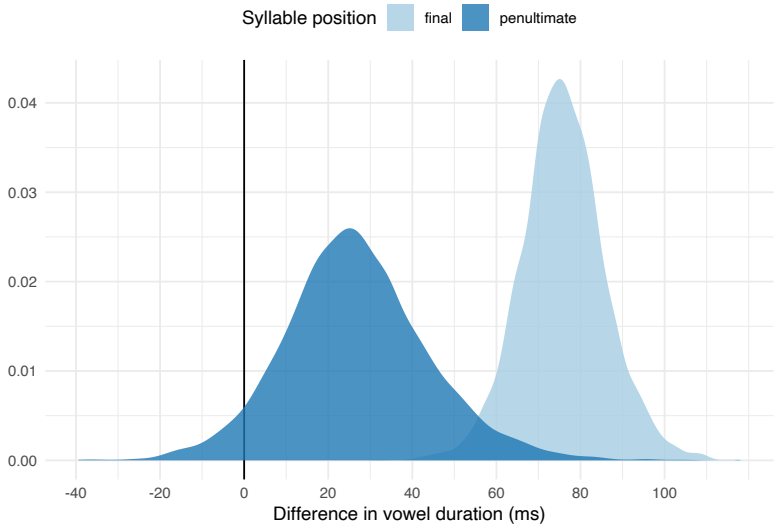


Figure B.2: Meta-analytical posterior distributions of the voicing effect in syllable-final and penultimate position.

in this than the other estimate. We can argue that, on average, the mean voicing effect in penultimate syllables is about 50 ms smaller than the mean effect in monosyllabic words in the surveyed studies. The mean of the voicing effect in disyllabic words can thus be estimated to be around 25 ms (75 - 50 ms).

A visual representation of the meta-analytical distributions is given in Figure B.2. The plot shows the full posterior distributions of the voicing effect in the word-final and penultimate contexts. Note how the posterior distribution in penultimate position is wider than the other.

Figure B.3 shows the mean estimates (the points) of the voicing effect with 95% CIs (the horizontal segments) for each of the 11 studies. For each study, the plot gives both the original estimate (as obtained from the raw data summary of the study) and the estimate shrunk by the random effects in the meta-analytical model. The vertical lines indicate the meta-analytical 95% CI of the voicing effect in final (solid) and penultimate syllable position (dashed). Original estimates further away from the meta-analytical mean effect and those with greater uncertainty (wider errors) show greater shrinkage to the mean.

Figure B.4 is a funnel plot, which can be used to visually check whether the sample suffers from publication bias. In this plot, the x-axis corresponds to the original esti-

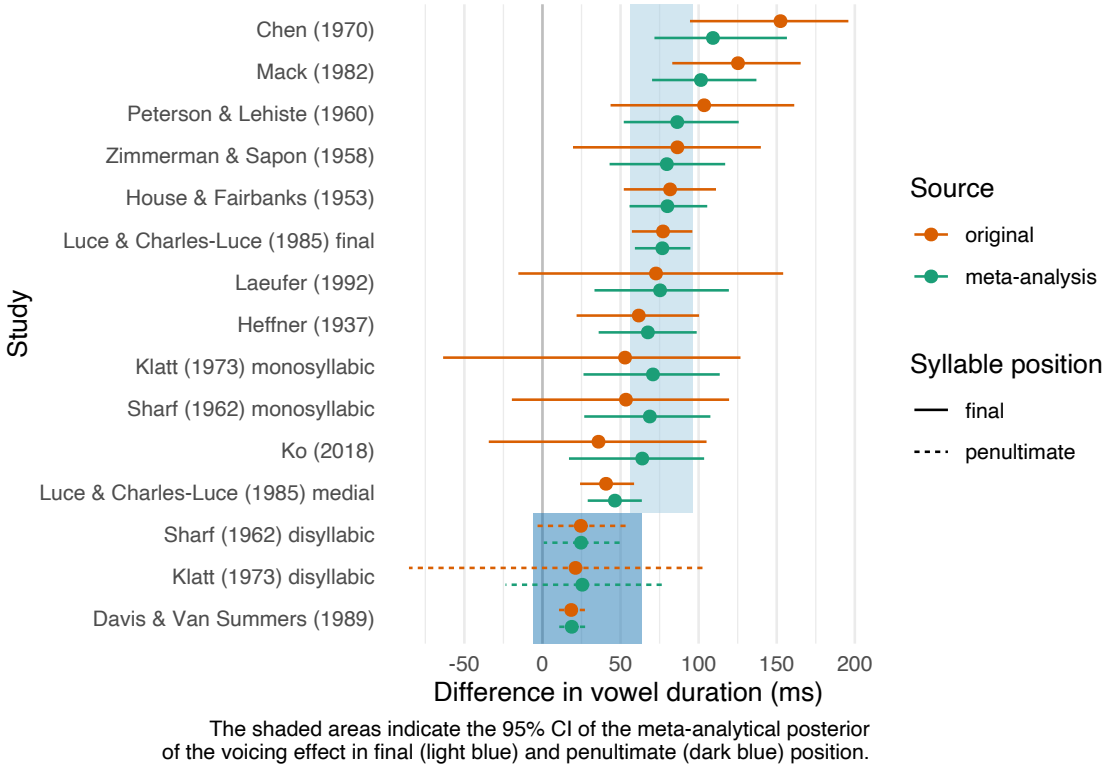


Figure B.3: Estimated voicing effect from the original source and from the meta-analysis.

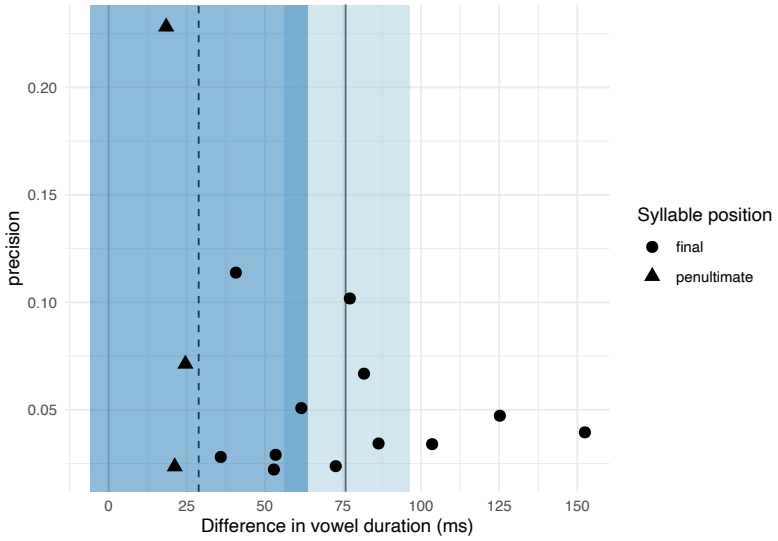


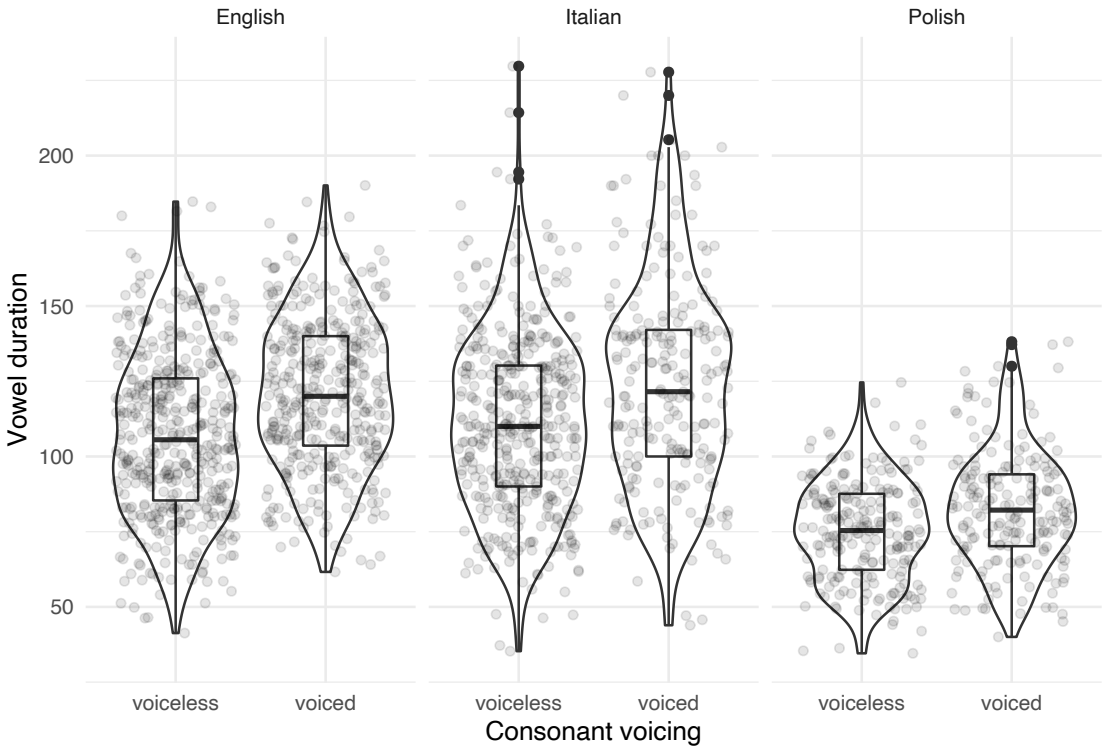
Figure B.4: By-study funnel plot showing the estimate against the precision. The vertical thick and dashed lines are the meta-analytical means of the effect in final and penultimate position.

mated difference in vowel duration, while the y-axis is a measure of precision (calculated as 1 divided by the estimated error of the difference). The meta-analytical means are indicated by the thick and dashed vertical lines for syllable-final and penultimate postion respectively. The shaded areas indicate the 95% CI of the meta-analytical posterior of the voicing effect in final (light blue) and penultimate (dark blue) position. When there is no bias, the points with lower precision should be more spread out and symmetrically placed around the meta-analytical mean, while points with higher precision should cluster around the mean. This ideal situation is clearly not the case for the final syllable context. There seems to be a bias towards bigger effects (which also happen to have lower precision). This indicates that the estimate probably suffers from publication bias (i.e. bias towards publicating positive and significant results) and it is not representative of the true effect. It is not possible to assess bias with the effect in penultimate syllable position given the low number of studies.

Appendix C

Cross-linguistic comparison of the voicing effect in English, Italian, and Polish

A Bayesian analysis was run to statistically test differences in the voicing effect in disyllabic (CV́CV) words of English (Study II), Italian, and Polish (Study I). Note that the experimental design differs between the two studies (see Chapter 3), so results should be interpreted with caution. The following graph shows violin and box plots of the raw vowel duration data, by voicing of C2 and language. English and Italian have similar vowel durations and a similar effect of voicing, while Polish has generally shorter vowels and a somewhat smaller effect.

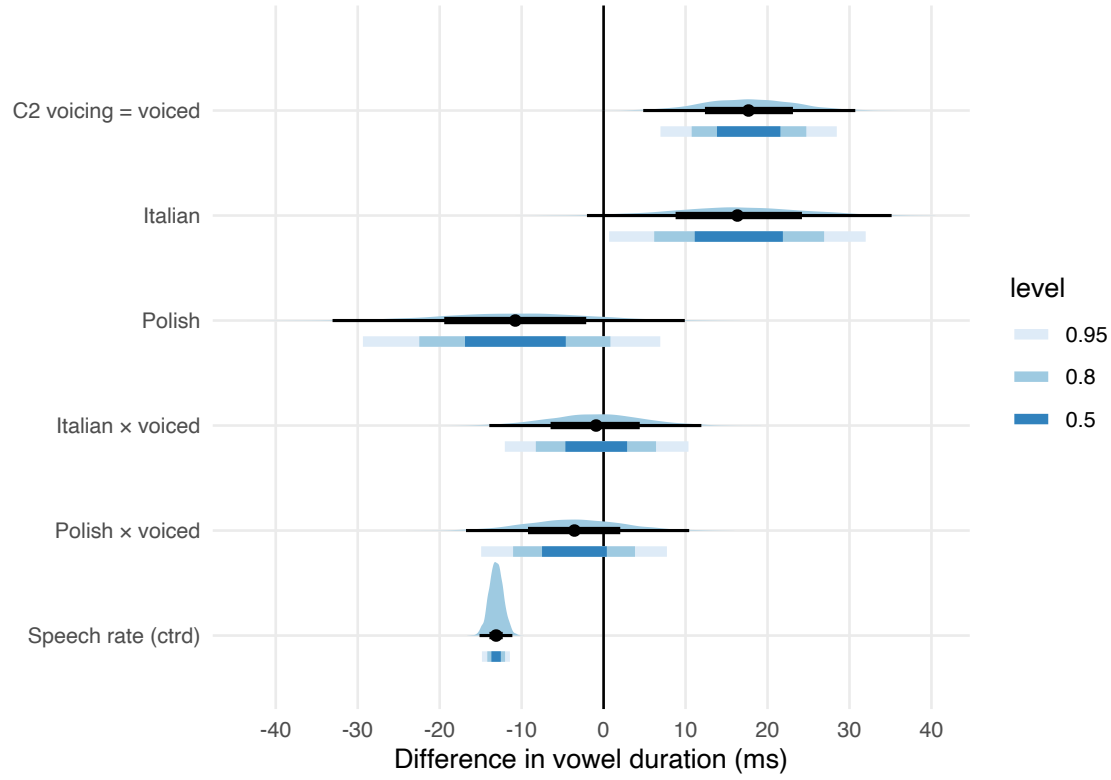


A Bayesian mixed-effects regression was fitted to V1 duration with brms (Bürkner 2017, 2018) in R (R Core Team 2019). Language, C2 voicing, centred speech rate, and an interaction between language and voicing were included as predictors. Random intercepts for speaker and word were used, together with by-speaker and by-word random coefficients for voicing.

The following priors were used: for the intercept of vowel duration, a normal distribution with mean 145 and SD 30, for the effect of language a normal distribution with mean 0 and SD 50, for the effect of voicing a normal distribution with mean 25 and SD 10, for the interaction between language and voicing a normal distribution with mean 0 and SD 10, and for centred speech rate a normal distribution with mean -25 and SD 10.

```
## Group-Level Effects:
## ~speaker (Number of levels: 32)
##                               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)                    14.84      2.09    11.42    19.57 1.01     1076
## sd(voicingvoiced)                 5.25      1.04     3.44     7.56 1.00     1972
## cor(Intercept,voicingvoiced)      0.13      0.21    -0.30     0.51 1.00     2499
##                               Tail_ESS
## sd(Intercept)                     1887
```

```
## sd(voicingvoiced)                2755
## cor(Intercept,voicingvoiced)     2937
##
## ~word (Number of levels: 39)
##                               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)                    14.71      2.24    10.84    19.63 1.00     1177
## sd(voicingvoiced)                 9.59      7.12     0.41    25.76 1.00      389
## cor(Intercept,voicingvoiced)     -0.07      0.42    -0.77     0.78 1.00      933
##                               Tail_ESS
## sd(Intercept)                     2141
## sd(voicingvoiced)                  798
## cor(Intercept,voicingvoiced)      1304
##
## Population-Level Effects:
##                                Estimate Est.Error l-95% CI u-95% CI Rhat
## Intercept                         96.78      6.16    84.51   108.63 1.00
## languageItalian                   16.44      8.00     0.67    31.98 1.00
## languagePolish                   -10.82      9.13   -29.36     6.91 1.00
```

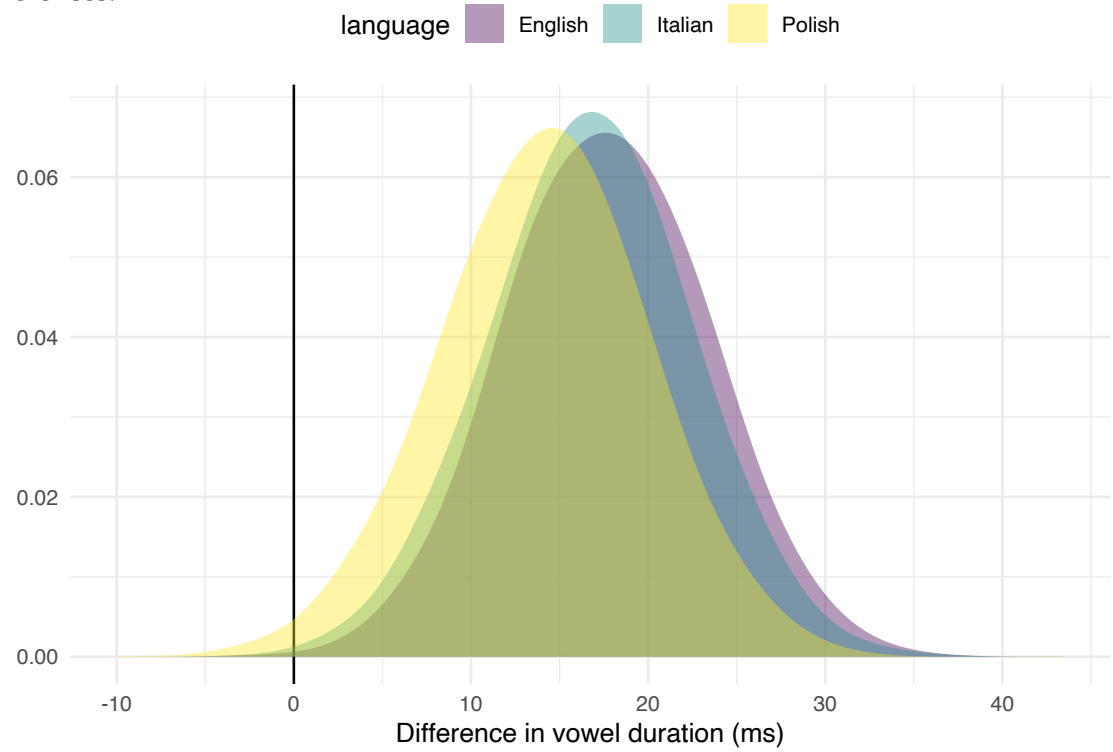


The plot above shows the posterior distributions for the effects of language, voicing, centred speech rate, and language × voicing interaction. The effect of voicing in English

is between 7 and 28.5 ms at 95% probability. There is weak evidence for longer vowels in Italian and shorter vowels in Polish compared to English, but the credible intervals of these effects are very wide. Speech rate has a strong and robust negative effect on vowel duration: for each syllable per second unit increase, vowels get 11.5-15 ms shorter. The posterior distributions for the interaction between language and voicing indicate that probably the effect of voicing in Italian is very similar to that of English, while there is some extremely weak indication for a slightly smaller effect in Polish. Note, however, that the posterior distributions of the interactions are very wide (more than 20 ms).

In sum, the present data does not offer robust evidence neither for or against cross-lingustic differences in voicing effect. If there is a difference, it will likely be within the range ±10 ms.

The following plot shows the posterior probability distributions of the effect of voicing marginalised over language. The great overlap among the distributions is indicative of the high uncertainty regarding the presence vs absence of cross-lingusitic differences.



Appendix D

Gesture onset timing of voiceless and voiced stops in Italian and Polish

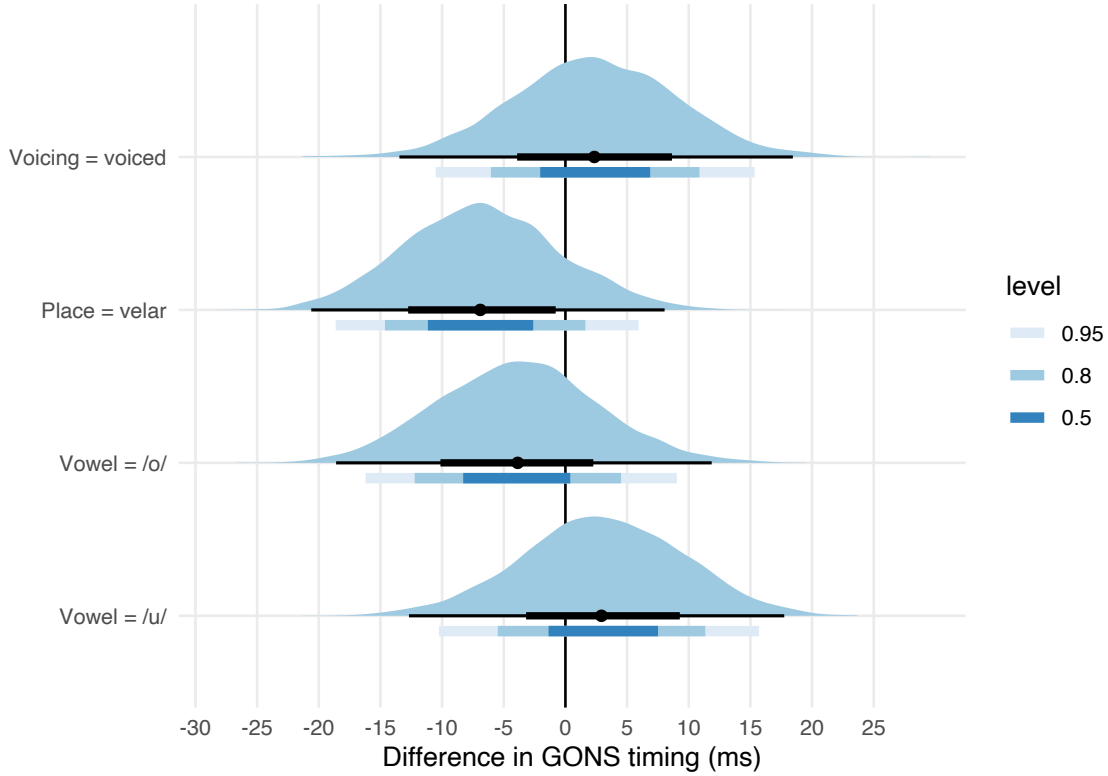
A consequence of the gestural organisation proposed to account for the stability of the release-to-release interval duration in disyllabic words is that the timing of the gestural onset should not be affected by the voicing status of the consonant in disyllabic words. In other words, the interval between the release of the the consonant preceding the vowel and the onset of the closing gesture of the post-vocalic consonant should be the same whether the consonant is voiceless or voiced. The difference in vowel duration (and closure duration) would be a consequence of the different velocity of the closing gesture in voiceless vs voiced stops, rather than of a difference in gestural onset.

The ultrasound tongue imaging data from Study I partially suggests that the temporal distance between C1 release and C2 gestural onset is not affected by C2 voicing. A Bayesian regression was fit to the duration of the C1 release to C2 gesture onset (GONS) interval, with C2 voicing, vowel, C2 place of articulation, interactions between voicing and vowel and voicing and place, and centred speech rate as predictors. By speaker and by-word random intercepts were also included. A normal distribution with mean 0 ms and SD 200 was used as prior for the intercept, while a distribution with mean 0 and SD 10 was used for vowel, place and the interactions. For speech rate, the prior was a normal distribution with mean 0 and SD 50.

```
## Group-Level Effects:
## ~item (Number of levels: 24)
```

```
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)    13.89      3.34     8.21    21.57 1.00     1647     2112
##
## ~speaker (Number of levels: 16)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)    54.28     10.30    38.16    78.59 1.00     1067     1818
##
## Population-Level Effects:
##                                   Estimate Est.Error l-95% CI u-95% CI Rhat
## Intercept                            85.34     14.81    56.38   114.50 1.00
## c2_phonationvoiced                    2.37      6.68   -10.51    15.35 1.00
## c2_placevelar                        -6.74      6.26   -18.63     5.94 1.00
## vowelo                               -3.84      6.50   -16.20     9.04 1.00
## vowelu                                2.96      6.54   -10.26    15.71 1.00
## speech_rate_c                       -14.58      5.69   -25.69    -3.43 1.00
## c2_phonationvoiced:c2_placevelar     -0.82      7.52   -15.97    13.30 1.00
## c2_phonationvoiced:vowelo             3.16      7.57   -11.65    17.99 1.00
## c2_phonationvoiced:vowelu            -0.04      7.67   -14.98    15.02 1.00
##                                   Bulk_ESS Tail_ESS
## Intercept                              746      952
## c2_phonationvoiced                    3307     3002
## c2_placevelar                         2738     2607
```

The following plot shows the posterior probabilities of the effects of voicing, place of articulation, and vowel on gestural onset timing. The credible intervals are quite large (> 25 ms). At 80% probability, the effect of voicing is between −5 and +10 ms, while at 95% probability it is between −10.5 and +15.5.



The present data does not offer unambiguous support for isochronous timing of C2 gestural onset, but it suggests that the difference is smaller than 15 ms. The gestural literature does not explicitly posit a lower limit as to what range of values would indicate gestural isochrony. Hermes et al. (2019) measure the lag between the gestures of an onset consonant and the vocalic nucleus and report that, in the standard population, the mean lag is 32 ms (SD 66). If a 32 ms lag in implementation of two gestures can be interpreted as indicating a relation of synchrony between these gestures, than a difference below 15 ms could be interpreted as suggesting an isochronous production of voiceless and voiced consonantal gestures. Note that while in Hermes et al. (2019) the temporal lag refers to a syntagmatic relation between two gestures, the case of the voiceless/voiced consonants is paradigmatic. Future work should: (1) identify a minimum theoretical value below which two gestures can be considered to be paradigmatically isochronous, and (2) investigate the temporal relation of gestural onsets in VCV sequences using a bigger sample.

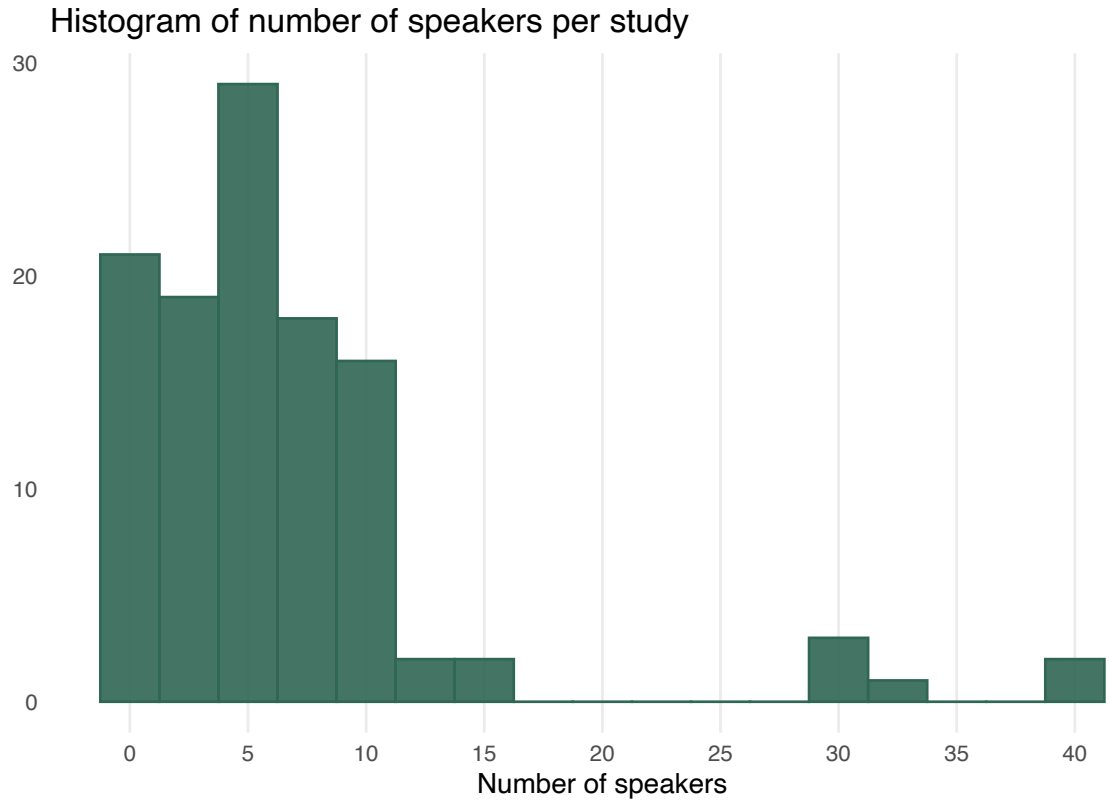
Appendix E

An informal analysis of number of speakers per phonetic study by year and endangerment status

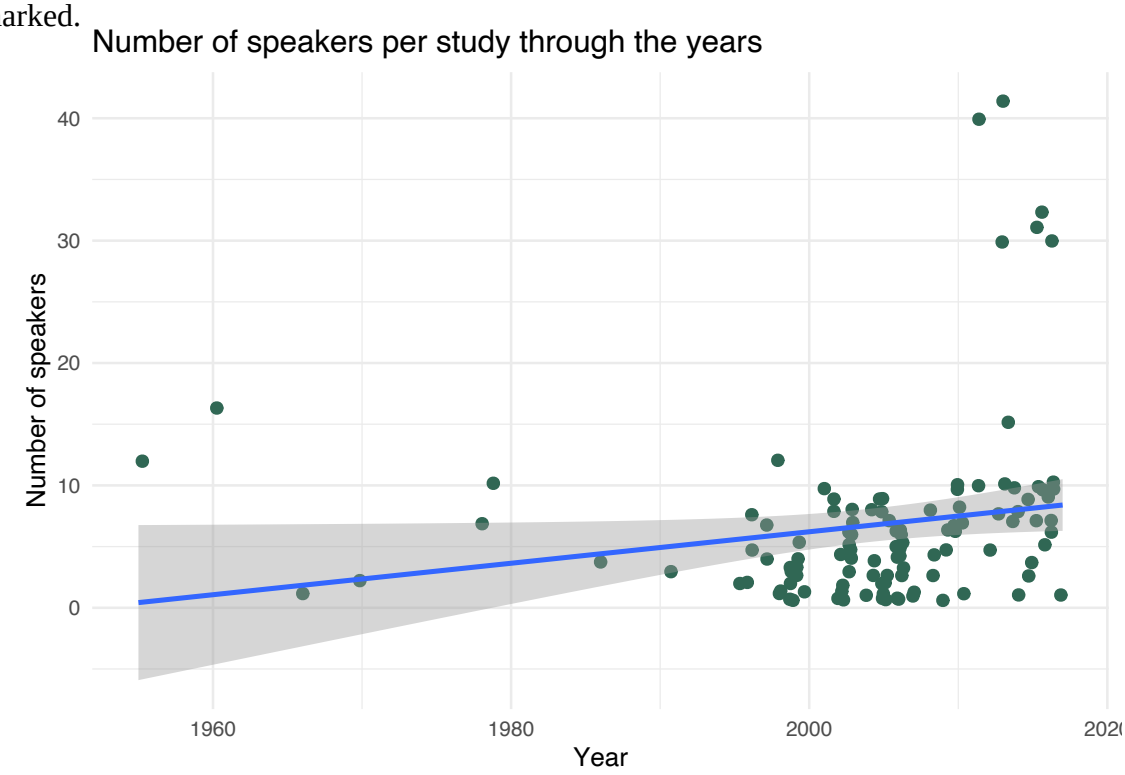
This analysis is based on the dataset used in Roettger & Gordon (2017) and Gordon & Roettger (2017) (Gordon & Roettger 2018).[1] The dataset contains information on number of participants from 113 studies, published between 1955 and 2017 (the majority of the studies are within the range 1990–2017).

The median number of speakers per study across the entire dataset is 5. The histogram below shows that most studies have 10 speakers or less, and that there are a few outliers with 30-40 speakers.

[1] A previous version of this appendix appeared as a blog post at `https://stefanocoretta.github.io/post/an-estimate-of-number-of-speakers-per-study-in-phonetics/`.

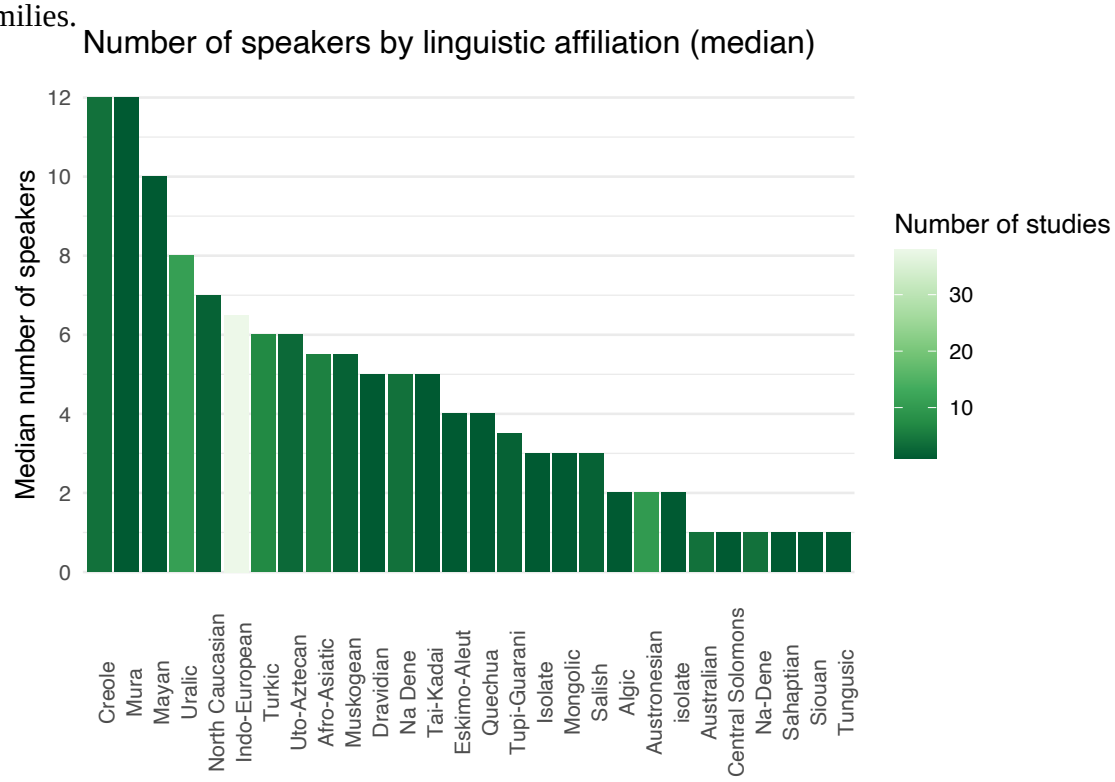


The following plot shows the number of speakers across publication year. There is a tendency for an increase in number of speakers, although the trend is not particularly marked.



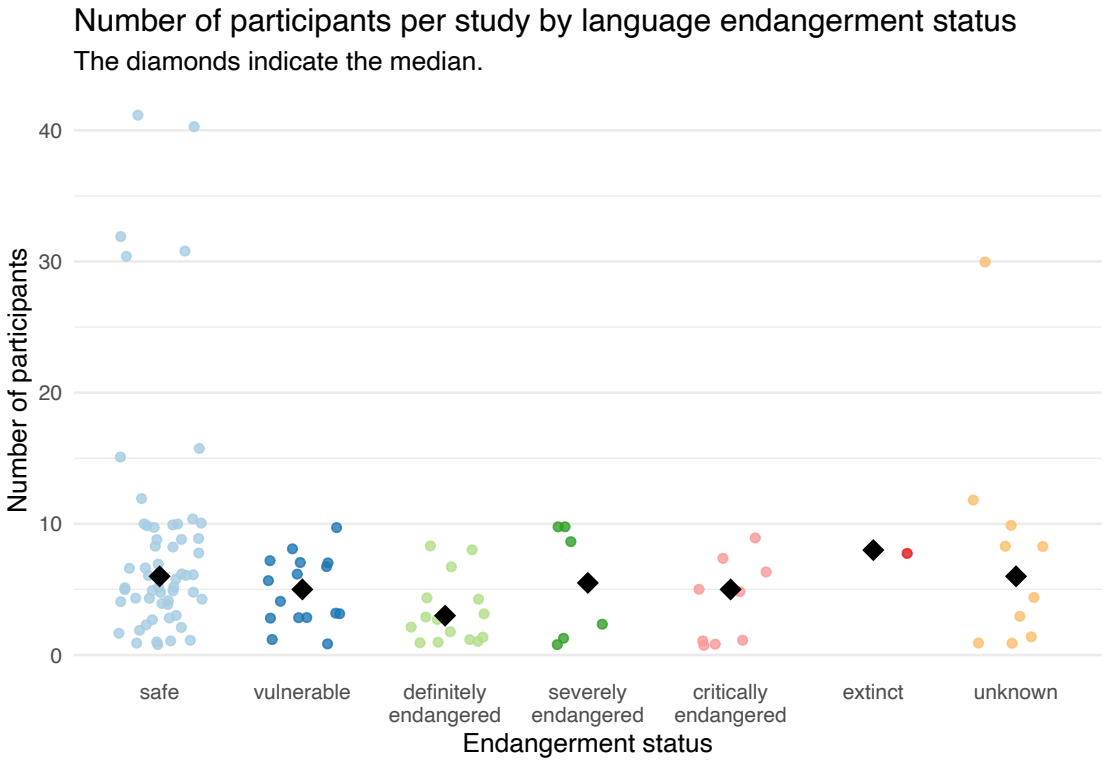
The following bar chart shows the median number of speakers in studies grouped

by linguistic affiliation. The colour of the bars indicates the number of studies. Indo-European languages stand out in terms of number of studies (> 30), but the median number of speakers in this family does not fare much better than other less-reachable families.



Information on the endangerment status of the languages in the dataset was obtained from GlottoLog.[2] The following strip chart shows the number of speakers for each of the studies (each point) categorised by the endangerment of the target language. With the caveat that there are more studies on safe languages, there is a trend of decreasing number of speakers from safe, to vulnerable, to definitely endangered languages. The very low number of studies on languages of greater endangerment status makes it harder to establish patterns. Note also that the decreasing trend is in fact small (1/2 speakers).

[2] `https://glottolog.org/meta/downloads`.



While generalisations based on this cursory analysis would not be wise, there seems to be a tendency for studies to have a very low number of speakers (median 5 speakers per study). The majority of studies analysed data from 10 speakers or less. This estimate is independent of publication year and endangerment status of the language enquired.

Bibliography

Abari, Kálmán. 2012. Reproducible research in speech sciences. *International Journal of Computer Science Issues* 9(6). 43–52.

Abdelli-Beruh, Nassima B. 2004. The stop voicing contrast in French sentences: Contextual sensitivity of vowel duration, closure duration, voice onset time, stop release and closure voicing. *Phonetica* 61(4). 201–219.

Abercrombie, David. 1967. *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Ahn, Suzy. 2015. The role of the tongue root in phonation of American English stops. Paper presented at Ultrafest VII `http://www.ultrafest2015.hku.hk/docs/S_Ahn_ultrafest.pdf`.

Ahn, Suzy. 2016. An ultrasound study of tongue position during Hindi laryngeal contrasts. *The Journal of the Acoustical Society of America* 140(4). 3221–3221.

Ahn, Suzy. 2018. The role of tongue position in laryngeal contrasts: An ultrasound study of English and Brazilian Portuguese. *Journal of Phonetics* 71. 451–467.

Ahn, Suzy & Lisa Davidson. 2016. Tongue root positioning in English voiced obstruents: Effects of manner and vowel context. *The Journal of the Acoustical Society of America* 140(4). 3221–3221.

Ambridge, Ben. 2018. Against stored abstractions: A radical exemplar model of language acquisition. Pre-print available at PsyArXiv.

Ananthapadmanabha, Tirupattur V., Aragulla Prasad Prathosh & Angarai Ganesan Ramakrishnan. 2014. Detection of the closure-burst transitions of stops and affricates

in continuous speech using the plosion index. *The Journal of the Acoustical Society of America* 135(1). 460–471.

Antoniou, Mark, Catherine T. Best, Michael D. Tyler & Christian Kroos. 2010. Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of Phonetics* 38(4). 640–653.

Articulate Instruments Ltd™. 2008. Ultrasound stabilisation headset users manual: Revision 1.4. Edinburgh, UK: Articulate Instruments Ltd.

Articulate Instruments Ltd™. 2011. Articulate Assistant Advanced user guide. Version 2.16.

Arvaniti, Amalia. 2009. Rhythm, timing and the timing of rhythm. *Phonetica* 66. 46–63.

Baese-Berk, Melissa & Matthew Goldrick. 2009. Mechanisms of interaction in speech production. *Language and cognitive processes* 24(4). 527–554.

Bailey, George. 2019a. Emerging from below the social radar: Incipient evaluation in the North West of England. *Journal of Sociolinguistics* 23(1). 3–28.

Bailey, George. 2019b. Ki(ng) in the north: Effects of duration, boundary, and pause on post-nasal [g]-presence. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1–26.

Bakker, Marjan, Annette van Dijk & Jelte M. Wicherts. 2012. The rules of the game called psychological science. *Perspectives on Psychological Science* 7(6). 543–554.

Baranowski, Maciej, Ricardo Bermúdez-Otero, George Bailey & Danielle Turton. 2016. Ahead but not faster: the effect of high token frequency on sound change. Paper presented at NWAV 45, Simon Fraser University, Vancouver, 4th November.

Baranowski, Maciej & Danielle Turton. 2015. Manchester English. In Raymond Hickey (ed.), *Researching northern english*, 293–316. Amsterdam/Philadelphia: John Benjamins.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Beckman, Jill, Michael Jessen & Catherine Ringen. 2013. Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics* 49(02). 259–284.

Beguš, Gašper. 2017. Effects of ejective stops on preceding vowel duration. *The Journal of the Acoustical Society of America* 142(4). 2168–2184.

Belasco, Simon. 1953. The influence of force of articulation of consonants on vowel duration. *The Journal of the Acoustical Society of America* 25(5). 1015–1016.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah & Stanley Dubinsky. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

van den Berg, Janwillem. 1958. Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research* 1(3). 227–244.

Bermúdez-Otero, Ricardo. 2007. Diachronic phonology. In *The Cambridge handbook of phonology*, 517. Cambridge: Cambridge University Press.

Bermúdez-Otero, Ricardo. 2013. An amphichronic approach to English syllabification. Rutger-UMass-MIT Phonology Workshop, `http://www.bermudez-otero.com/amphichronic.pdf`.

Bermúdez-Otero, Ricardo. 2015. Amphichronic explanation and the life cycle of phonological processes. In *The Oxford handbook of historical phonology*, 374–399. Oxford: Oxford University Press.

Bermúdez-Otero, Ricardo. 2017. Stratal phonology. In S. J. Hannahs & Anna R. K. Bosch (eds.), *The Routledge handbook of phonological theory*, 100–134. Routledge.

Bermúdez-Otero, Ricardo, Maciej Baranowski, George Bailey & Danielle Turton. 2016. A constant rate effect in Manchester /t/-glottalling: high-frequency words

are ahead of, but change at the same rate as, low-frequency words. Paper presented at OCP 13, Budapest, January 2016.

Bertinetto, Pier Marco & Michele Loporcaro. 2005. The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *Journal of the International Phonetic Association* 35(2). 131–151. doi:10.1017/S0025100305002148.

Betancourt, Michael. 2018. Calibrating model-based inferences and decisions. arXiv preprint arXiv:1803.08393.

Bigi, Brigitte. 2015. SPPAS - Multi-lingual approaches to the automatic annotation of speech. *The Phonetician* 111–112. 54–69.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 557–582.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Blevins, Juliette. 2006. A theoretical synopsis of Evolutionary Phonology. *Theoretical linguistics* 32(2). 117–166.

Boersma, Paul & David Weenink. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.40.

Bolker, Ben & David Robinson. 2019. broom.mixed: Tidying methods for mixed models. R package version 0.2.4.

Bonett, Douglas G. 2006. Confidence interval for a coefficient of quartile variation. *Computational Statistics & Data Analysis* 50(11). 2953–2957.

Bortolini, Umberta, Claudio Zmarich, Renato Fior & Serena Bonifacio. 1995. Word-initial voicing in the productions of stops in normal and preterm Italian infants. *International Journal of Pediatric Otorhinolaryngology* 31. 191–206.

Browman, Catherine P. & Louis Goldstein. 1988. Some notes on syllable structure in articulatory phonology. *Phonetica* 45(2-4). 140–155.

Browman, Catherine P. & Louis Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49. 155–180.

Browman, Catherine P. & Louis Goldstein. 2000. Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la communication parlée* (5). 25–34.

Browman, Catherine P. & Louis M. Goldstein. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3. 219–252. doi:10.1017/S0952675700000658.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28.

Bürkner, Paul-Christian. 2018. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* 10(1). 395–411.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5). 365.

Campos-Astorkiza, Judit Rebeka. 2007. *Minimal contrast and the phonology-phonetics interaction*: University of Southern California dissertation.

Celata, Chiara & Paolo Mairano. 2014. On the timing of V-to-V intervals in Italian: a review, and some new hypotheses. *Revista de Filología Románica* 31. 37.

Chambers, Christopher D., Zoltan Dienes, Robert D. McIntosh, Pia Rotshtein & Klaus Willmes. 2015. Registered reports: realigning incentives in scientific publishing. *Cortex* 66. A1–A2.

Chen, Matthew. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22(3). 129–159.

Childers, Donald G. & Ashok K. Krishnamurthy. 1985. A critical review of electroglottography. *Critical reviews in biomedical engineering* 12(2). 131–161.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York, Evanston, and London: Harper & Row.

Classe, André. 1939. *The rhythm of English prose*. Blackwell.

Coretta, Stefano. 2017. Pilot study on EGG data analysis [Data]. Open Science Framework. https://osf.io/r94v8/.

Coretta, Stefano. 2018a. An exploratory study of the voicing effect in Italian and Polish [Data]. Open Science Framework. https://osf.io/8zhku/.

Coretta, Stefano. 2018b. rticulate: Ultrasound Tongue Imaging in R. R package version 1.3.2, https://github.com/stefanocoretta/rticulate.

Coretta, Stefano. 2019a. Compensatory aspects of the effect of voicing on vowel duration in English [Data]. Open Science Framework. https://osf.io/ep8wb/. doi: 10.17605/OSF.IO/EP8WB.

Coretta, Stefano. 2019b. An exploratory study of voicing-related differences in vowel duration as compensatory temporal adjustment in Italian and Polish. OSF pre-print.

Coretta, Stefano. 2019c. Longer vowel duration correlates with greater tongue root displacement: Acoustic and articulatory data from Italian and Polish. OSF pre-print.

Coretta, Stefano. 2019d. speakr: A wrapper for the phonetic software praat. R package version 2.1.0, https://github.com/stefanocoretta/speakr.

Coretta, Stefano. 2019e. tidymv: Tidy model visualisation for generalised additive models. R package version 2.2.0, https://github.com/stefanocoretta/tidymv.

Coretta, Stefano. 2020. Vowel duration and consonant voicing: A production study [Research compendium]. Open Science Framework https://osf.io/w92me. doi: 10.17605/OSF.IO/W92ME.

Coretta, Stefano & Massimiliano Canzi. 2018. The effect of lexical frequency on vowel phonation as a correlate of /t/-glottaling. Talk presented at LAGB 2018.

Cristofaro, Sonia. 2012. Cognitive explanations, distributional evidence, and diachrony. *Studies in Language* 36(3). 645–670.

Cristofaro, Sonia. 2014. Competing motivation models and diachrony: What evidence for what motivations? In Brian MacWhinney, Andrej Malchukov & Edith A. Moravcsik (eds.), *Competing motivations in grammar and usage*, Oxford: Oxford University Press.

Crüwell, Sophia, Johnny van Doorn, Alexander Etz, Matthew Makel, Hannah Moshontz, Jesse Niebaum, Amy Orben, Sam Parsons & Michael Schulte-Mecklenbeck. 2018. 8 easy steps to open science: An annotated reading list. PsyArXiv.

Crystal, Thomas H. & Arthur S. House. 1988. Segmental durations in connected speech signals: Current results. *The Journal of the Acoustical Society of America* 83(4). 1553–1573.

Cumming, Ruth. 2011. The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics* 39(3). 375–387.

Cyran, Eugeniusz. 2011. Laryngeal realism and laryngeal relativism: Two voicing systems in Polish? *Studies in Polish Linguistics* 6(1). 45–80.

Cysouw, Michael. 2015. Accountable and reproducible research.

Cysouw, Michael & Jeff Good. 2013. Languoid, doculect, and glossonym: Formalizing the notion 'language'. *Language Documentation & Conservation* 7. 331–359.

Dale, Rick, Eric Dietrich & Anthony Chemero. 2009. Explanatory pluralism in cognitive science. *Cognitive Science* 33(5). 739–742.

Dauer, Richard M. 1987. Phonetic and phonological components of language rhythm. In *Proceedings of the xith international congress of phonetic sciences*, 447–450.

Davidson, Lisa. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America* 120(1). 407–415.

Davidson, Lisa. 2016. Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics* 54. 35–50.

Davis, Stuart & W. Van Summers. 1989. Vowel length and closure duration in word-medial VC sequences. *Journal of Phonetics* 17. 339–353.

d'Imperio, Mariapaola & Sam Rosenthall. 1999. Phonetics and phonology of main stress in Italian. *Phonology* 16(1). 1–28. doi:10.1017/S0952675799003681.

Docherty, Gerard J. 1992. *The timing of voicing in British English obstruents*. Berlin, New York: Foris.

Dumas-Mallet, Estelle, Katherine S. Button, Thomas Boraud, Francois Gonon & Marcus R. Munafò. 2017. Low statistical power in biomedical science: a review of three human research domains. *Royal Society open science* 4(2). 160254.

Durvasula, Karthik & Qian Luo. 2012. Voicing, aspiration, and vowel duration in Hindi. *Proceedings of Meetings on Acoustics* 18. 1–10.

Easterbrook, Phillipa J., Ramana Gopalan, J. A. Berlin & David R. Matthews. 1991. Publication bias in clinical research. *The Lancet* 337(8746). 867–872.

Einarsson, Stefán. 1927. *Beiträge zur Phonetik der isländischen Sprache*. Oslo : A.W. Brøggers.

Elert, Claes-Christian. 1970. Phonologic studies of quantity in Swedish. Skriptor: Uppsala.

Erickson, Donna & Shigeto Kawahara. 2016. Articulatory correlates of metrical structure: Studying jaw displacement patterns. *Linguistics Vanguard* 2(1).

Esposito, Anna. 2002. On vowel height and consonantal voicing effects: Data from Italian. *Phonetica* 59(4). 197–231.

Etz, Alexander, Quentin F. Gronau, Fabian Dablander, Peter A. Edelsbrunner & Beth Baribault. 2018. How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review* 25(1). 219–234.

Fabre, P. 1957. Un procede electrique percutane d'inscrition de l'accolement glottique au cours de la phonation: glottographie de haute frequence. Premiers resultats. *Bulletin de l'Académie nationale de médecine* 141. 66.

Farnetani, Edda & Shiro Kori. 1986. Effects of syllable and word structure on segmental durations in spoken Italian. *Speech Communication* 5(1). 17–34.

Fechner, Gustav Theodor. 1966. *Elements of psychophysics [Elemente der Psychophysik]*, vol. 1. Adler, H. E. (translator), United States of America: Holt, Rinehart and Winston.

Ferrero, Franco E., Emanuela Magno Caldognetto, Kiryaki Vagges & Carlo Lavagnoli. 1978. Some acoustic characteristics of Italian vowels. *Journal of Italian Linguistics Amsterdam* 3(1). 87–94.

Fintoft, Knut. 1961. The duration of some Norwegian speech sounds. *Phonetica* 7(1). 19–39.

Flake, Jessica Kay & Eiko I. Fried. 2019. Measurement schmeasurement: Questionable measurement practices and how to avoid them. Pre-print available at PsyArXiv.

Fomel, Sergey & Jon Claerbout. 2009. Guest editors' introduction: Reproducible research. *Computing in Science and Engineering* 11(1). 5–7.

Fowler, Carol A. 1983. Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General* 112(3). 386.

Fowler, Carol A. 1992. Vowel duration and closure duration in voiced and unvoiced stops: There are no contrast effects here. *Journal of Phonetics* 20(1). 143–165.

Fox, John. 2003. Effect displays in R for generalised linear models. *Journal of Statistical Software* 8(15). 1–27.

Fox, John & Sanford Weisberg. 2019. *An R companion to applied regression*. Thousand Oaks, CA 3rd edn.

Gafos, Adamantios, Christo Kirov & Jason Shaw. 2010. Guidelines for using Mview.

Gahl, Susanne & R. Harald Baayen. 2019. Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood. *Journal of Phonetics* 74. 42–54.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11.

Gelman, Andrew & John Carlin. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9(6). 641–651.

Gelman, Andrew & Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University, `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`.

Gelman, Andrew & Francis Tuerlinckx. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15(3). 373–390.

Gick, Bryan. 2002. The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association* 32(02). 113–121.

Giegerich, Heinz J. 1992. *English phonology: An introduction*. Cambridge: Cambridge University Press.

Giordano, Rosa & Leandro D'Anna. 2010. A comparison of rhythm metrics in different speaking styles and in fifteen regional varieties of Italian. In *Proceedings of the 5th international conference on speech prosody*, .

Glewwe, Eleanor. 2018. The effect of lexical competition on vowel duration before voiced and voiceless English stops. *The Journal of the Acoustical Society of America* 143(3). 1968–1968.

Goldrick, Matthew, Charlotte Vaughn & Amanda Murphy. 2013. The effects of lexical neighbors on stop consonant articulation. *The Journal of the Acoustical Society of America* 134(2). EL172–EL177.

Goldstein, Louis & Catherine P Browman. 1986. Representation of voicing contrasts using articulatory gestures. *Journal of Phonetics* 14(2). 339–342.

Goldstein, Louis, Dani Byrd & Elliot Saltzman. 2006. The role of vocal tract gestural action units in understanding the evolution of phonology. In Michael A. Arbib (ed.), *Action to language via the mirror neuron system*, 215–249. Cambridge: Cambridge University Press.

Goldstein, Louis & Marianne Pouplier. 2014. The temporal organization of speech. In V. Ferreira, M. Goldrick & M. Miozzo (eds.), *The Oxford handbook of language production*, Oxford: Oxford University Press.

Gordeeva, Olga B. & James M. Scobbie. 2007. Non-normative preaspirated voiceless fricatives in Scottish English: Phonetic and phonological characteristics. *QMU Speech Science Research Centre Working Papers* .

Gordeeva, Olga B. & James M. Scobbie. 2010. Preaspiration as a correlate of word-final voice in Scottish English fricatives. In Susanne Fuchs, M. Toda & M. Żygis (eds.), *Turbulent sounds: An interdisciplinary guide*, 167–208. De Gruyter Mouton.

Gordeeva, Olga B. & James M. Scobbie. 2011. Laryngeal variation in the Scottish English voice contrast: Glottalisation, ejectivisation and aspiration. In *Speech Science Research Centre Working Papers*, vol. 19, Queen Margaret University.

Gordon, Matt & Timo B. Roettger. 2018. Studies on acoustic correlates of word stress - an online corpus. (Last updated: 2018 April 17).

Gordon, Matthew & Timo B. Roettger. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard* 3(1).

Grabe, Esther & Ee Ling Low. 2002. Durational variability in speech and the rhythm class hypothesis. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology*, vol. 7 515-546, De Gruyter Mouton.

Greenland, Sander. 2017. Invited commentary: the need for cognitive science in methodology. *American Journal of Epidemiology* 186(6). 639–645.

Gregoire, A. 1911. Influences des consonnes occlusives sur la durée des syllabes precedents. *Revue de Phonétique* 1. 260–292.

Grimaldi, Mirko, B. Gili Fivela, Francesco Sigona, Michele Tavella, Paul Fitzpatrick, Laila Craighero, Luciano Fadiga, Giulio Sandini & Giorgio Metta. 2008. New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph. *Proceedings of LangTech* 1–5.

Gu, Chong. 2013. *Smoothing spline ANOVA models*. Springer Science & Business Media.

Gussmann, Edmund. 2007. *The phonology of Polish*. Oxford: Oxford University Press.

Hajek, John & Mary Stevens. 2008. Vowel duration, compression and lengthening in stressed syllables in Central and Southern varieties of standard Italian. In *Proceedings of the 9th annual conference of the International Speech Communication Association*, 516–519.

Halle, Morris & Kenneth Noble Stevens. 1967. Mechanism of glottal vibration for vowels and consonants. *The Journal of the Acoustical Society of America* 41(6). 1613–1613.

Halle, Morris, Kenneth Noble Stevens & Alan Victor Oppenheim. 1967. On the mechanism of glottal vibration for vowels and consonants. In *Quarterly progress report*, vol. 85, 267–277.

Hampala, Vít, Maxime Garcia, Jan G. Švec, Ronald C. Scherer & Christian T. Herbst. 2016. Relationship between the electroglottographic signal and vocal fold contact area. *Journal of Voice* 30(2). 161–171.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Hastie, Trevor & Robert Tibshirani. 1986. Generalized additive models. *Statistical Science* 1(3). 297–310.

Hay, Jennifer B., Janet B. Pierrehumbert, Abby J. Walker & Patrick LaShell. 2015. Tracking word frequency effects through 130 years of sound change. *Cognition* 139. 83–91.

Heffner, R.-M.S. 1937. Notes on the length of vowels. *American Speech* 12. 128–134.

Hejná, Michaela. 2015. *Pre-aspiration in Welsh English: A case study of Aberystwyth*: The University of Manchester dissertation.

Helgason, Pétur. 1999. Phonetic preconditions for the development of normative preaspiration. In *Proceedings of the 14th International Congress of the Phonetic Sciences*, vol. 99, 1–7.

Helgason, Pétur. 2002. *Preaspiration in the Nordic languages: synchronic and diachronic aspects*: Institutionen för lingvistik dissertation.

Helwig, Nathaniel E & Ping Ma. 2016. Smoothing spline ANOVA for super-large samples: Scalable computation via rounding parameters. arXiv.org preprint, arXiv:1602.05208 [stat.CO].

Herbst, Christian T., W. Tecumseh S. Fitch & Jan G. Švec. 2010. Electroglottographic wavegrams: A technique for visualizing vocal fold dynamics noninvasively. *The Journal of the Acoustical Society of America* 128(5). 3070–3078.

Herbst, Christian T., Harm K. Schutte, Daniel L. Bowling & Jan G. Svec. 2017. Comparing chalk with cheese—the EGG contact quotient is only a limited surrogate of the closed quotient. *Journal of Voice* 31(4). 401–409.

Hermes, Anne, Doris Mücke & Martine Grice. 2013. Gestural coordination of italian word-initial clusters: the case of 'impure s'. *Phonology* 30(01). 1–25.

Hermes, Anne, Doris Mücke, Tabea Thies & Michael T. Barbe. 2019. Coordination patterns in essential tremor patients with deep brain stimulation: Syllables with low and

high complexity. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1–20.

Hertrich, Ingo & Hermann Ackermann. 1997. Articulatory control of phonological vowel length contrasts: Kinematic analysis of labial gestures. *The Journal of the Acoustical Society of America* 102(1). 523–536.

Heyne, Matthias & Donald Derrick. 2015a. Benefits of using polar coordinates for working with ultrasound midsagittal tongue contours. *The Journal of the Acoustical Society of America* 137(4). 2302–2302.

Heyne, Matthias & Donald Derrick. 2015b. Using a radial ultrasound probe's virtual origin to compute midsagittal smoothing splines in polar coordinates. *The Journal of the Acoustical Society of America* 138(6). EL509–EL514.

Hirose, Hajime. 1977. Laryngeal adjustments in consonant production. *Phonetica* 34(4). 289–294.

Hirose, Hajime & Thomas Gay. 1972. The activity of the intrinsic laryngeal muscles in voicing control. *Phonetica* 25(3). 140–164.

Hock, Hans Henrich. 1991. *Principles of historical linguistics*. Berlin: Mouton de Gruyter.

House, Arthur S. 1961. On vowel duration in English. *The Journal of the Acoustical Society of America* 33(9). 1174–1178.

House, Arthur S. & Grant Fairbanks. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America* 25(1). 105–113.

Hualde, José Ignacio & Marianna Nadeu. 2011. Lenition and phonemic overlap in Rome Italian. *Phonetica* 68(4). 215–242.

Huggins, A. William F. 1972. Just noticeable differences for segment duration in natural speech. *The Journal of the Acoustical Society of America* 51(4B). 1270–1278.

Hussein, Lutfi. 1994. *Voicing-dependent vowel duration in Standard Arabic and its acquisition by adult American students*: Columbus, OH: The Ohio State University dissertation.

Huszthy, Bálint. 2016. Italian as a voice language without voice assimilation. In *Proceedings of ConSOLE XXIV*, 428–452.

Ioannidis, John P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2(8). e124.

Jacewicz, Ewa, Robert Allen Fox & Samantha Lyle. 2009. Variation in stop consonant voicing in two regional varieties of American English. *Journal of the International Phonetic Association* 39(3). 313–334.

Janda, Richard D. 1999. Accounts of phonemic split have been greatly exaggerated—but not enough. In *Proceedings of the 14th international congress of phonetic sciences*, vol. 14, 329–332.

Jansen, Wouter. 2004. *Laryngeal contrast and phonetic voicing: a laboratory phonology approach to English, Hungarian, and Dutch*: University Library Groningen dissertation.

Jarosz, Andrew F & Jennifer Wiley. 2014. What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving* 7(1). 2–9.

Jassem, Wiktor. 2003. Polish. *Journal of the International Phonetic Association* 33(1). 103–107.

Jassem, Wiktor & Lutoslawa Richter. 1989. Neutralization of voicing in Polish obstruents. *Journal of Phonetics* 17(4). 317–325.

Javkin, Hector R. 1976. The perceptual basis of vowel duration differences associated with the voiced/voiceless distinction. *Report of the Phonology Laboratory, UC Berkeley* 1. 78–92.

John, Leslie K., George Loewenstein & Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* 23(5). 524–532.

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson & John W. Mullenix (eds.), *Talker variability in speech processing*, 145–165. San Diego, CA: Academic Press.

de Jong, Kenneth. 1991. An articulatory study of consonant-induced vowel duration changes in English. *Phonetica* 48(1). 1–17.

de Jong, Kenneth. 2004. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics* 32(4). 493–516.

de Jong, Kenneth & Bushra Zawaydeh. 2002. Comparing stress, lexical focus, and segmental focus: Patterns of variation in Arabic vowel duration. *Journal of Phonetics* 30(1). 53–75.

Kagaya, Ryohei & Hajime Hirose. 1975. Fiberoptic electromyographic and acoustic analyses of Hindi stop consonants. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics* 9. 27–46.

Kass, Robert E. & Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430). 773–795.

Kawahara, Shigeto, Donna Erickson & Atsuo Suemitsu. 2017. The phonetics of jaw displacement in Japanese vowels. *Acoustical Science and Technology* 38(2). 99–107.

Kay, Matthew. 2019. tidybayes: Tidy data and geoms for Bayesian models. R package version 1.1.0.

Keating, Patricia A. 1984a. Phonetic and phonological representation of stop consonant voicing. *Language* 60(2). 286–319.

Keating, Patricia A. 1984b. Universal phonetics and the organization of grammars. *UCLA Working Papers in Phonetics* 59. 35–49. `https://escholarship.org/uc/item/2497n8jq`.

Kent, Raymond D. & Kenneth L. Moll. 1969. Vocal–tract characteristics of the stop cognates. *Journal of the Acoustical Society of America* 46(6B). 1549–1555.

Kerr, Norbert L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2(3). 196–217.

Kicinski, Michal. 2013. Publication bias in recent meta-analyses. *PLoS ONE* 8(11). e81823.

Kingston, John & Randy L. Diehl. 1994. Phonetic knowledge. *Language* 419–454.

Kiparsky, Paul. 1988. Phonological change. In Frederick J. Newmeyer (ed.), *Linguistics: the Cambridge survey*, vol. 1 Linguistic theory: foundations, 363–415. Cambridge: Cambridge University Press.

Kiparsky, Paul. 2000. Opacity and cyclicity. *The linguistic review* 17(2-4). 351–366.

Kiparsky, Paul. 2015. Phonologization. In *The Oxford handbook of historical phonology*, 563–579: Oxford: Oxford University Press.

Kirby, James & Morgan Sonderegger. 2018. Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics* 70. 70–85.

Kirby, James P. 2016. Obstruent voicing, aspiration, and tone: implications for laryngeal phonology. Poster presented at LabPhon 15.

Kirby, James P. & D. Robert Ladd. 2016. Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages. *The Journal of the Acoustical Society of America* 140(4). 2400–2411.

Kirkham, Sam & Claire Nance. 2017. An acoustic-articulatory study of bilingual vowel production: Advanced tongue root vowels in Twi and tense/lax vowels in Ghanaian English. *Journal of Phonetics* 62. 65–81.

Klatt, Dennis H. 1973. Interaction between two factors that influence vowel duration. *The Journal of the Acoustical Society of America* 54(4). 1102–1104.

Klein, Olivier, Tom E. Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel & Michael C. Frank. 2018. A practical guide for transparency in psychological science. *Collabra: Psychology* 4(1). 20. doi:10.1525/collabra.158.

Kluender, Keith R., Randy L. Diehl & Beverly A. Wright. 1988. Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics* 16. 153–169.

Knuth, Donald E. 1984. Literate programming. *The Computer Journal* 27(2). 97–111.

Ko, Eon-Suk. 2018. Asymmetric effects of speaking rate on the vowel/consonant ratio conditioned by coda voicing in English. *Phonetics and Speech Sciences* 10(2). 45–50.

Krämer, Martin. 2009. *The phonology of Italian*. Oxford: Oxford University Press.

Kroos, Christian, Philip Hoole, Barbara Kühnert & Hans G. Tillmann. 1997. Phonetic evidence for the phonological status of the tense-lax distinction in German. *Forschungsberichte-Institut für Phonetik und Sprachliche Kommunikation der Universität München* (35). 17–25.

Kruschke, John. 2015. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd edition)*. Amsterdam, The Netherlands: Academic Press.

Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Haubo Bojesen Christensen. 2017. `lmerTest` package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26.

Laeufer, Christiane. 1992. Patterns of voicing-conditioned vowel duration in French and English. *Journal of Phonetics* 20(4). 411–440.

Lampp, Claire & Heidi Reklis. 2004. Effects of coda voicing and aspiration on Hindi vowels. *The Journal of the Acoustical Society of America* 115(5). 2540–2540.

Lehiste, Ilse. 1970a. Temporal organization of higher-level linguistic units. *The Journal of the Acoustical Society of America* 48(1A). 111.

Lehiste, Ilse. 1970b. Temporal organization of spoken language. In *OSU Working Papers in Linguistics*, vol. 4, 96–114. https://linguistics.osu.edu/sites/linguistics.osu.edu/files/workingpapers/osu_wpl_04.pdf.

Lehiste, Ilse. 1976. Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics* 4(2). 113–117.

Lieber, Rochelle. 2009. *Introducing morphology*. Cambridge University Press.

Lindblom, Björn. 1967. Vowel duration and a model of lip mandible coordination. *Speech Transmission Laboratory Quarterly Progress Status Report* 4. 1–29. http://www.speech.kth.se/prod/publications/files/qpsr/1967/1967_8_4_001-029.pdf.

Lisker, Leigh. 1957. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language* 33(1). 42–49.

Lisker, Leigh. 1974. On "explaining" vowel duration variation. In *Proceedings of the Linguistic Society of America*, 225–232.

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

Löfqvist, Anders, Thomas Baer, Nancy S. McGarr & Robin Seider Story. 1989. The cricothyroid muscle in voicing control. *The Journal of the Acoustical Society of America* 85(3). 1314–1321.

Luce, Paul A. & Jan Charles-Luce. 1985. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America* 78(6). 1949–1957.

Łukaszewicz, Beata. 2018. Phonetic evidence for an iterative stress system: The issue of consonantal rhythm. *Phonology* 35(1). 115–150.

Luke, Steven G. 2017. Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods* 49(4). 1494–1502.

Lulich, Steven M., Kelly H. Berkson & Kenneth de Jong. 2018. Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion. *Journal of Phonetics* 71. 410–424.

Machač, Pavel & Radek Skarnitzl. 2007. Temporal compensation in Czech. In *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, 6–10 August*, 537–540.

Machač, Pavel & Radek Skarnitzl. 2009. *Principles of phonetic segmentation*. Praha: Epocha.

Mack, Molly. 1982. Voicing dependent vowel duration in English and French: Monolingual and bilingual production. *The Journal of the Acoustical Society of America* 71(1). 173–178.

Maddieson, Ian. 1976. Further studies on vowel length before aspirated consonants. Paper presented at the 91st meeting of the Acoustical Society of America.

Maddieson, Ian & Jack Gandour. 1976. Vowel length before aspirated consonants. In *UCLA Working papers in Phonetics*, vol. 31, 46–52. `https://escholarship.org/uc/item/31f5j8m7`.

Magno Caldognetto, Emanuela, Franco Ferrero, Kyriaki Vagges & Maria Bagno. 1979. Indici acustici e indici percettivi nel riconoscimento dei suoni linguistici (con applicazione alle consonanti occlusive dell'italiano). *Acta Phoniatrica Latina* 2. 219–246.

Malisz, Zofia & Katarzyna Klessa. 2008. A preliminary study of temporal adaptation in Polish VC groups. In *Proceedings of Speech Prosody*, 383–386. `https://www.isca-speech.org/archive/sp2008/papers/sp08_383.pdf`.

Marin, Stefania & Marianne Pouplier. 2010. Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control* 14(3). 380–407.

Marin, Stefania & Marianne Pouplier. 2014. Articulatory synergies in the temporal organization of liquid clusters in Romanian. *Journal of Phonetics* 42. 24–36.

Marszalek, Jacob M., Carolyn Barber, Julie Kohlhart & B. Holmes Cooper. 2011. Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills* 112(2). 331–348.

Marwick, Ben, Carl Boettiger & Lincoln Mullen. 2017. Packaging data analytical work reproducibly using R (and friends). *The American Statistician* 72(1).

Maxwell, Michael. 2013. A system for archivable grammar documentation. In Georg Rehm, Cerstin Mahlow & Michael Piotrowski (eds.), *Systems and Frameworks for Computational Morphology. Third International Workshop*, 72–91. Springer.

Maxwell, Michael & Jonathan D. Amith. 2005. Language documentation: the Nahuatl grammar. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, 474–485. Berlin Heidelberg: Springer-Verlag.

McElreath, Richard. 2015. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

McKiernan, Erin C., Philip E. Bourne, C. Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A. Nosek, Karthik Ram & Courtney K. Soderberg. 2016. Point of view: How open science helps researchers succeed. *eLife* 5. e16800. doi:10.7554/eLife.16800.001.

Menezes, Caroline & Donna Erickson. 2013. Intrinsic variations in jaw deviations in English vowels. In *Proceedings of Meetings on Acoustics*, vol. 19, 060253.

Meyer, Ernst Alfred. 1903. *Englische Lautdauer: Eine experimentalphonetische Untersuchung*. Akademiska Bokhandeln, Uppsala, Sweden.

Meyer, Ernst Alfred. 1904. Zur Vokaldauer im Deutschen. In *Nordiska studier tillegnade A. Noreen*, 347–356. Uppsala: K.W. Appelbergs Boktryckeri.

Meyer, Ernst Alfred & Z. Gombocz. 1909. Zur Phonetik der ungarischen Sprache. *Le Monde Oriental* 3. 122–197.

Mielke, Jeff. 2015. An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons. *The Journal of the Acoustical Society of America* 137(5). 2858–2869.

Mitleb, Fares. 1982. Voicing effect on vowel duration is not an absolute universal. *The Journal of the Acoustical Society of America* 71(S1). S23–S23.

Morin, Olivier. 2015. A plea for "shmeasurement" in the social sciences. *Biological Theory* 10(3). 237–245.

Mortensen, Johannes & John Tøndering. 2013. The effect of vowel height on Voice Onset Time in stop consonants in CV sequences in spontaneous Danish. In *Proceedings of Fonetik 2013*, Linköping, Sweden: Linköping University.

Moslin, Barbara J. & Patricia A. Keating. 1977. Voicing distinction in Polish word initial stop consonants. *The Journal of the Acoustical Society of America* 62. S27.

Motulsky, Harvey J. 2014. Common misconceptions about data analysis and statistics. *Naunyn-Schmiedeberg's Arch Pharmacol* 387. 1017–1023.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1(1). 0021.

Nance, Claire & Jane Stuart-Smith. 2013. Pre-aspiration and post-aspiration in Scottish Gaelic stop consonants. *Journal of the International Phonetic Association* 43(02). 129–152.

Navarro Tomás, Tomás. 1916. *Cantidad de las vocales acentuadas*. Madrid Sucesores de Hernando.

Nespor, Marina. 1990. On the rhythm parameter in phonology. In Iggy Roca (ed.), *Logical issues in language acquisition*, 157–175. Dordrecht: Foris.

Ní Chasaide, Ailbhe. 1985. *Preaspiration in phonological stop contrasts: an instrumental phonetic study*: University of Wales dissertation.

Ní Chasaide, Ailbhe & Christer Gobl. 1993. Contextual variation of the vowel voice source as a function of adjacent consonants. *Language and Speech* 36(2-3). 303–330.

Nicenboim, Bruno, Timo B. Roettger & Shravan Vasishth. 2018. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics* 70. 39–55.

Nickerson, Raymond S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2(2). 175–220.

Nissen, Silas Boye, Tali Magidson, Kevin Gross & Carl T. Bergstrom. 2016. Publication bias and the canonization of false facts. *Elife* 5. e21451.

Nooteboom, Sieb G. & Gert J. N. Doodeman. 1980. Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America* 67(1). 276–287.

Nowak, Pawel. 2006. *Vowel reduction in Polish*: Berkeley, CA: University of California, Berkeley dissertation.

Nüst, Daniel, Carl Boettiger & Ben Marwick. 2018. How to read a research compendium. *arXiv preprint arXiv:1806.09525* .

O'Dell, Michael L. & Tommi Nieminen. 2008. Coupled oscillator model for speech timing: Overview and examples. In *Nordic prosody: Proceedings of the Xth conference*, 179–190.

Ohala, John J. 1989. Sound change is drawn from a pool of synchronic variation. In Leiv Breivik & Ernst Jahr (eds.), *Language change: Contributions to the study of its causes*, 173–198. New York: Mouton de Gruyter.

Ohala, John J. 2011. Accommodation to the aerodynamic voicing constraint and its phonological relevance. In *Proceedings of the 17th International Congress of Phonetic Sciences*, 64–67.

Ohala, Manjari & John J. Ohala. 1992. Phonetic universals and Hindi segment duration. In John J. Ohala, T. Nearey, B. Derwing, M. Hodge & G. Wiebe (eds.), *Proceedings of the International Conference on Spoken Language Processing, Banff*, 831–834.

Öhman, Sven E. G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America* 39(1). 151–168.

Öhman, Sven E. G. 1967a. Numerical model of coarticulation. *The Journal of the Acoustical Society of America* 41(2). 310–320.

Öhman, Sven E. G. 1967b. Peripheral motor commands in labial articulation. *Speech Transmission Laboratory Quarterly Progress Status Report* 8. 30–63.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251). aac4716.

Pamies Bertrán, Antonio. 1999. Prosodic typology: on the dichotomy between stress-timed and syllable-timed languages. *Language design: Journal of theoretical and experimental linguistics* 2. 103–130.

Pape, Daniel & Luis M. T. Jesus. 2014. Production and perception of velar stop (de)voicing in European Portuguese and Italian. *EURASIP Journal on Audio, Speech, and Music Processing* 2014(1). 6.

Pashler, Harold & Eric-Jan Wagenmakers. 2012. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* 7(6). 528–530.

Peirce, Jonathan W. 2009. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics* 2(10).

Peng, Roger D. 2009. Reproducible research and biostatistics. *Biostatistics* 10(3). 405–408.

Peng, Roger D. 2015. *Report writing for data science in R.* Lulu.

Perezgonzalez, Jose D. 2015. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology* 6(223).

Perkell, Joseph S. 1969. *Physiology of speech production: Results and implication of quantitative cineradiographic study*. Cambridge, MA: MIT Press.

Peterson, Gordon E. & Ilse Lehiste. 1960. Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America* 32(6). 693–703.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: word frequency, lenition and contrast. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–157. Amsterdam Philadelphia: John Benjamins Publishing Company.

Pike, Kenneth L. 1945. The intonation of American English. In Dwight Bolinger (ed.), *Intonation*, 53–83. Harmondsworth: Penguin.

Plug, Leendert & Rachel Smith. 2018. Segments, syllables and speech tempo perception. In *Proceedings of the 9th International Conference on Speech Prosody 2018*, 279–283.

Port, Robert F. 1981. Linguistic timing factors in combination. *The Journal of the Acoustical Society of America* 69(1). 262–274.

Port, Robert F. & Jonathan Dalby. 1982. Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics* 32(2). 141–152.

Port, Robert F., Jonathan Dalby & Michael O'Dell. 1987. Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America* 81(5). 1574–1585.

Port, Robert F. & Rosemarie Rotunno. 1979. Relation between voice onset time and vowel duration. *The Journal of the Acoustical Society of America* 66(3). 654–662.

Pouplier, Marianne. 2012. The gestural approach to syllable structure: Universal, language-and cluster-specific aspects. In *Speech planning and dynamics*, 63–96. New York, Oxford, Wien: Peter Lang.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team. 2019. R: A language and environment for statistical computing. `https://www.R-project.org/`.

Raftery, Adrian E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25. 111–163.

Raftery, Adrian E. 1999. Bayes factors and BIC: Comment on "A critique of the Bayesian information criterion for model selection". *Sociological Methods & Research* 27(3). 411–427.

Raphael, Lawrence J. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristic of word final consonants in American English. *The Journal of the Acoustical Society of America* 51(4B). 1296–1303.

Raphael, Lawrence J. 1975. The physiological control of durational differences between vowels preceding voiced and voiceless consonants in English. *Journal of Phonetics* 3(1). 25–33.

Ratnikova, E. I. 2017. Towards a log-normal model of phonation units lengths distribution in the oral utterances. *International Research Journal* 3(57). 46–49.

Renwick, Margaret & Robert D. Ladd. 2016. Phonetic distinctiveness vs. lexical contrastiveness in non-robust phonemic contrasts. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1). 1–29.

Riordan, Carol J. 1980. Larynx height during english stop consonants. *Journal of Phonetics* 8. 353–360.

Roese, Neal J. & Kathleen D. Vohs. 2012. Hindsight bias. *Perspectives on psychological science* 7(5). 411–426.

Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic sciences. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1–27.

Roettger, Timo B. & Matthew Gordon. 2017. Methodological issues in the study of word stress correlates. *Linguistics Vanguard* 3(1).

Rogers, Henry. 2004. *Writing systems: A linguistic approach*. Oxford: Blackwell.

Rosen, Kristin M. 2005. Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison. *Journal of Phonetics* 33(4). 411–426.

Rosenthal, Robert. 1979. The file drawer problem and tolerance for null results. *Psychological bulletin* 86(3). 638.

Rossi, Joseph S. 1990. Statistical power of psychological research: What have we gained in 20 years? *Journal of consulting and clinical psychology* 58(5). 646.

Rothenberg, Martin. 1967. *The breath-stream dynamics of simple-released-plosive production*, vol. 6. Basel: Biblioteca Phonetica.

Rothenberg, Martin & James J. Mahshie. 1988. Monitoring vocal fold abduction through vocal fold contact area. *Journal of Speech, Language, and Hearing Research* 31(3). 338–351.

Saltzman, Elliot, Hosung Nam, Jelena Krivokapic & Louis Goldstein. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the 4th International Conference on Speech Prosody*, 175–184.

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor & Eivind Hovig. 2013. Ten simple rules for reproducible computational research. *PLoS Computational Biology* 9(10). 1–4.

Sanker, Chelsea. 2018. Effects of laryngeal features on vowel duration: implications for Winter's Law. *Papers in Historical Phonology* 3. 180–205.

Sanker, Chelsea. 2019. Influence of coda stop features on perceived vowel duration. *Journal of Phonetics* 75. 43–56.

Scherer, Ronald C. & Ingo R. Titze. 1987. The abduction quotient related to vocal quality. *Journal of Voice* 1(3). 246–251.

Schooler, Jonathan W. 2014. Metascience could rescue the 'replication crisis'. *Nature News* 515(7525). 9.

Schwab, Matthias, N. Karrenbach & Jon Claerbout. 2000. Making scientific computations reproducible. *Computing in Science & Engineering* 2(6). 61–67.

Schwartz, Geoffrey. 2016. On the evolution of prosodic boundaries–parameter settings for Polish and English. *Lingua* 171. 37–73.

Schwartz, Geoffrey & Daria Arndt. 2018. Laryngeal Realism vs. Modulation theory – evidence from VOT discrimination in Polish. *Language Sciences* 69. 98–112.

Schwartz, Geoffrey, Anna Balas & Arkadiusz Rojczyk. 2015. Phonological factors affecting L1 phonetic realization of proficient Polish users of English. *Research in Language* 13(2). 181–198.

Scobbie, James M., Eleanor Lawson, Steve Cowen, Joanne Cleland & Alan A. Wrench. 2011. A common co-ordinate system for mid-sagittal articulatory measurement. In *QMU CASL Working Papers*, 1–4.

Seyfarth, Scott, Esteban Buz & T. Florian Jaeger. 2016. Dynamic hyperarticulation of coda voicing contrasts. *The Journal of the Acoustical Society of America* 139(2). EL31–EL37.

Sharf, Donald J. 1962. Duration of post-stress intervocalic stops and preceding vowels. *Language and Speech* 5(1). 26–30.

Sharf, Donald J. 1964. Vowel duration in whispered and in normal speech. *Language and Speech* 7(2). 89–97.

Silberzahn, Raphael, Eric L. Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard & Evelina Bonnier. 2018. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science* 1(3). 337–356.

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22(11). 1359–1366.

Slis, Iman Hans & Antonie Cohen. 1969a. On the complex regulating the voiced-voiceless distinction I. *Language and Speech* 12(2). 80–102.

Slis, Iman Hans & Antonie Cohen. 1969b. On the complex regulating the voiced-voiceless distinction II. *Language and Speech* 12(3). 137–155.

Slowiaczek, Louisa M. & Daniel A. Dinnsen. 1985. On the neutralizing status of Polish word-final devoicing. *Journal of Phonetics* 13(3). 325–341.

Solé, Maria-Josep, Patrice Speeter Beddor & Manjari Ohala. 2007. *Experimental approaches to phonology*. Oxford University Press.

Song, Fujian, Sheetal Parekh, Lee Hooper, Yoon K. Loke, J. Ryder, Alex J. Sutton, C. Hing, Chun Shing Kwok, Chun Pang & Ian Harvey. 2010. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess* 14(8). 1–193.

Sóskuthy, Márton. 2013. *Phonetic biases and systemic effects in the actuation of sound change*: Edinburgh: University of Edinburgh dissertation.

Sóskuthy, Márton. 2017. Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. arXiv.org preprint, arXiv:1703.05339.

Sóskuthy, Márton, Paul Foulkes, Vincent Hughes & Bill Haddican. 2018. Changing words and sounds: The roles of different cognitive units in sound change. *Topics in Cognitive Science* 10(4). 1–16.

Sóskuthy, Márton & Jennifer B. Hay. 2017. Changing word usage predicts changing word durations in New Zealand English. *Cognition* 166.

Sprouse, Ronald L., Maria-Josep Solé & John J. Ohala. 2008. Oral cavity enlargement in retroflex stops. *Proceedings of the 8th International Seminar on Speech Production, Strasbourg* 425–428.

Stan Development Team. 2017. Stan: A C++ library for probability and sampling, version 2.14.0. `http://mc-stan.org/`.

Stevens, Kenneth N. & Samuel Jay Keyser. 1989. Primary features and their enhancement in consonants. *Language* 81–106.

Stevens, Kenneth N., Samuel Jay Keyser & Haruko Kawasaki. 2014. Toward a phonetic and phonological theory of redundant features. In Joseph S. Perkell & Dennis H. Klatt (eds.), *Invariance and variability in speech processes*, 426–463. Psychology Press.

Stevens, Mary. 2010. How widespread is preaspiration in Italy? A preliminary acoustic phonetic overview. In *Proceedings of FONETIK 2010*, 97–102. Lund.

Stevens, Mary & John Hajek. 2004a. Comparing voiced and voiceless geminates in Sienese Italian: what role does preaspiration play? In *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, 340–345.

Stevens, Mary & John Hajek. 2004b. Preaspiration in Sienese Italian and its interaction with stress in /VC:/ sequences. Paper presented at the International Conference Speech Prosody, Japan, March 23-26.

Stevens, Mary & John Hajek. 2010. Preaspirated /pp tt kk/ in Standard Italian: a sociophonetic v. phonetic analysis. In *Proceedings of the 2010 Speech Science and Technology Association Conference*, Melbourne, Australia.

Stevens, Mary & Ulrich Reubold. 2014. Pre-aspiration, quantity, and sound change. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 5(4). 455–488.

Stone, Maureen. 2005. A guide to analysing tongue motion from ultrasound images. *Clinical linguistics & phonetics* 19(6-7). 455–501.

Strycharczuk, Patrycja. 2012. Sonorant transparency and the complexity of voicing in Polish. *Journal of Phonetics* 40(5). 655–671.

Strycharczuk, Patrycja & James M. Scobbie. 2015. Velocity measures in ultrasound data. Gestural timing of post-vocalic /l/ in English. In *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5.

Summers, W. Van. 1987. Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *The Journal of the Acoustical Society of America* 82(3). 847–863.

Tanner, James, Morgan Sonderegger, Jane Stuart-Smith & SPADE Data Consortium. 2019. Vowel duration and the voicing effect across English dialects. Pre-print available at `https://ling.auf.net/lingbuzz/004640`.

Thieberger, Nicholas. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In *Language documentation and description*, vol. 2, London: SOAS.

Tilsen, Sam. 2013. A dynamical model of hierarchical selection and coordination in speech planning. *PLoS ONE* 8(4). e62800.

Tilsen, Sam. 2016. Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics* 55. 53–77.

Titze, Ingo R. 1990. Interpretation of the electroglottographic signal. *Journal of Voice* 4(1). 1–9.

Todd, Simon, Janet B. Pierrehumbert & Jennifer Hay. 2019. Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition* 185. 1–20.

Toivonen, Ida, Lev Blumenfeld, Andrea Gormley, Leah Hoiting, John Logan, Nalini Ramlakhan & Adam Stone. 2015. Vowel height and duration. In Ulrike Steindl, Thomas Borer, Huilin Fang, Alfredo García Pardo, Peter Guekguezian, Brian Hsu, Charlie O'Hara & Iris Chuoying Ouyang (eds.), *Proceedings of the 32nd West Coast Conference on Formal Linguistics*, vol. 32, 64–71. Somerville, MA: Cascadilla Proceedings Project.

Tressoldi, Patrizio E. & David Giofré. 2015. The pervasive avoidance of prospective statistical power: major consequences and practical solutions. *Frontiers in psychology* 6(726).

Trubetzkoy, Nikolai Sergeevich. 1969. *Principles of phonology*. Berkley and Los Angeles: University of California Press. Translated by Christiane A.M. Baltaxe.

Tukey, John W. 1980. We need both exploratory and confirmatory. *The American Statistician* 34(1). 23–25.

Umeda, Noriko. 1975. Vowel duration in American English. *The Journal of the Acoustical Society of America* 58(2). 434–445.

Umeda, Noriko. 1977. Consonant duration in American English. *The Journal of the Acoustical Society of America* 61(3). 846–858.

Vagges, Kyriaki, Franco E. Ferrero, Emanuela Magno-Caldognetto & Cristina Lavagnoli. 1978. Some acoustic characteristics of Italian consonants. *Journal of Italian Linguistics* 3(1). 69–84.

Van Heuven, W. J. B., P. Mandera, E. Keuleers & M. Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* 67. 1176–1190.

van Rij, Jacolien, Martijn Wieling, R. Harald Baayen & Hedderik van Rijn. 2017. itsadug: Interpreting time series and autocorrelated data using GAMMs. R package version 2.3.

Vasishth, Shravan, M. Beckman, B. Nicenboim, Fangfang Li & Eun Jong Kong. 2018a. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics* 71. 147–161.

Vasishth, Shravan & Andrew Gelman. 2019. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data. PsyArXiv.

Vasishth, Shravan, Daniela Mertzen, Lena A. Jäger & Andrew Gelman. 2018b. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175.

Vaux, Bert. 1996. The status of ATR in feature geometry. *Linguistic Inquiry* 27(1). 175–182.

Vazquez-Alvarez, Yolanda & Nigel Hewlett. 2007. The ‘trough effect’: an ultrasound study. *Phonetica* 64. 105–121.

van ’t Veer, Anna Elisabeth & Roger Giner-Sorolla. 2016. Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology* 67. 2–12.

Wagenmakers, Eric-Jan. 2007. A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review* 14(5). 779–804.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas & Rogier A. Kievit. 2012. An agenda for purely confirmatory research. *Perspectives on Psychological Science* 7(6). 632–638.

Waniek-Klimczak, Ewa. 2011. Aspiration in Polish: A sound change in progress? In Mirosław Pawlak & Jakub Bielak (eds.), *New perspectives in language, discourse and translation studies*, 3–11. Heidelberg, Dordrecht, London, New York: Springer.

Warren, Paul & Jen Hay. 2006. Using sound change to explore the mental lexicon. In M. Claire Fletcher-Flinn & G. M. Haberman (eds.), *Cognition and language: Perspectives from New Zealand*, chap. 8, 105–126. Brisbane, QLD, AUS: Australian Academic Press.

Warren, Willis & Adam Jacks. 2005. Lip and jaw closing gesture durations in syllable final voiced and voiceless stops. *The Journal of the Acoustical Society of America* 117(4). 2618–2618.

Weigel, William Frederick. 2002. The Yokuts canon: A case study in the interaction of theory and description. Paper presented at the annual meeting of the Linguistics Society of America, January 2002, San Francisco.

Weigel, William Frederick. 2005. *Yowlumne in the Twentieth century*: University of California, Berkley dissertation.

Wells, John C. 1982. *Accents of English*, vol. 1. Cambridge: Cambridge University Press.

Wells, John C. 1990. Syllabification and allophony. In Susan Ramsaran (ed.), *Studies in the pronunciation of English: A commemorative volume in honour of A. C. Gimson*, 76–86. New York: Routledge.

Westbury, John R. 1983. Enlargement of the supraglottal cavity and its relation to stop consonant voicing. *The Journal of the Acoustical Society of America* 73(4). 1322–1336.

White, Laurence, Elinor Payne & Sven L. Mattys. 2009. Rhythmic and prosodic contrast in Venetan and Sicilian italian. In M. Vigario, S. Frota & M. J. Freitas (eds.), *Phonetics and phonology: Interactions and interrelations*, 137–158. Amsterdam: John Benjamins.

Wickham, Hadley. 2017. tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2.1.

Wieling, Martijn. 2017. Generalized additive modeling to analyze dynamic phonetic data: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *The Mind Research Repository (beta)* (1).

Wieling, Martijn. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70. 86–116.

Winter, Bodo. 2015. The other N: The role of repetitions and items in the design of phonetic experiments. In *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow: The University of Glasgow. `https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0181.pdf`.

Winter, Bodo & Timo B. Roettger. 2011. The nature of incomplete neutralization in German: Implications for Laboratory Phonology. *Grazer Linguistische Studien* 76. 55–74.

Wood, Simon. 2006. *Generalized additive models: An introduction with R*. CRC Press.

Wood, Simon. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1). 3–36.

Wood, Simon. 2017. *Generalized additive models: An introduction with R*. Chapman and Hall/CRC 2nd edn.

Wood, Simon N. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(1). 95–114.

Xie, Yihui. 2014. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch & Roger D. Peng (eds.), *Implementing reproducible computational research*, Chapman and Hall: CRC Press.

Xie, Yihui. 2016. *bookdown: Authoring books and technical documents with r markdown*. Chapman and Hall/CRC.

Xie, Yihui. 2019. bookdown: Authoring books and technical documents with r markdown. R package version 0.11.

Xie, Yihui, Joseph J. Allaire & Garrett Grolemund. 2018. *R markdown: The definitive guide*. CRC Press.

Yanagihara, Naoaki & Charlene Hyde. 1966. An aerodynamic study of the articulatory mechanism in the production of bilabial stop consonants. *Studia Phonologica* 4. 70–80.

Yu, Alan C. L. 2010. Tonal effects on perceived vowel duration. In C. Fougeron, B. Kuehnert, M. Imperio & N. Vallee (eds.), *Laboratory phonology*, vol. 10 4, 151–168. New York: Mouton de Gruyter.

Zeroual, Chakir, Philip Hoole, Adamantios I. Gafos & John H. Esling. 2015. Gestural coordination differences between intervocalic simple and geminate plosives in Moroccan Arabic: An EMA investigation. In *Proceedings of ICPhS*, 1–5.

Zimmerman, Samuel A. & Stanley M. Sapon. 1958. Note on vowel duration seen cross linguistically. *The Journal of the Acoustical Society of America* 30(2). 152–153.

Zmarich, Claudio, Barbara Gili Fivela, Pascal Perrier, Christophe Savariaux & Graziano Tisato. 2011. Speech timing organization for the phonological length contrast in Italian consonants. In *Twelfth annual conference of the international speech communication association*, 401–404.

Zuur, Alain F. 2012. *A beginner's guide to generalized additive models with R*. Highland Statistics Limited: Newburgh.

Zuur, Alain F., Elena N. Ieno & Chris S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1(1). 3–14.